

UNA ALTERNATIVA BAYESIANA A LOS CONTRASTES DE LA BONDAD DEL AJUSTE

F. J. Girón
C. R. Ortiz
Departamento de Estadística
Universidad de Málaga

RESUMEN

Se considera el problema de hacer inferencias acerca del modelo beta simétrico, desde un punto de vista bayesiano. Los resultados se aplican posteriormente al contraste de la bondad de ajuste en el caso de una hipótesis nula simple (sin parámetros marginales) y en el caso de que la hipótesis nula conste de un número finito de modelos. Caso de que el test se acepte, se dan expresiones para las probabilidades a posteriori de los diferentes modelos. En el caso de que se rechaze la hipótesis nula el test da cierta información acerca de las posibles desviaciones hacia otras alternativas.

ABSTRACT

Inferences on the parameter of a symmetric beta distribution are considered from a Bayesian perspective. These results are then applied to goodness of fit testing for the case of simple null hypothesis (no nuisance parameters) and for the case of a composite null hypothesis when a finite number of models is postulated. In the case of acceptance posterior probabilities of the different models in the null hypothesis are derived. When the null hypothesis is rejected the test provides some information about possible departures.

1. Introducción

Los contrastes de la bondad del ajuste que, como en el caso de los basados en la distribución Chi-cuadrado, utilizan los datos agrupados en vez de los originales, suelen presentar, además de éste, el inconveniente

niente de la elección del número de intervalos en que se subdivide el espacio muestral y el del número mínimo de observaciones a incluir en cada intervalo (en la literatura existente se recomienda un número de cinco observaciones por celda o intervalo).

El problema de la elección óptima de la partición del espacio muestral es bastante complicado y no se ha resuelto satisfactoriamente (véase, por ejemplo, Breiman (1973)).

Ante éste doble problema —el que acabamos de considerar— y el señalado al principio, de la no utilización de los datos originales (raw data) sino de los datos agrupados (incluso en el caso en que la hipótesis nula sea compuesta, los parámetros de ésta han de estimarse por el método de la máxima verosimilitud, basados en los datos agrupados), es conveniente considerar otras alternativas.

La solución obvia suele ser utilizar contrastes tipo Kolmogorov-Smirnov, que utilizan los datos originales a través de la función de distribución empírica.

En éste artículo proponemos una solución de tipo paramétrico al problema de contrastar una hipótesis nula arbitraria (sin mencionar alternativas) en los casos en que ésta sea simple o compuesta de un número finito de modelos. El método puede en principio extenderse al caso de una hipótesis nula compuesta. Para ello adoptamos el punto de vista bayesiano.

El método se basa, esencialmente en las dos consideraciones siguientes:

1.^a Supongamos que F_0 es la función de distribución unidimensional continua de una variable aleatoria X_0 , que describe el procedimiento probabilístico por el cual se va a generar la muestra aleatoria observada. Para simplificar la exposición supongamos que F_0 está perfectamente estipulada, es decir, no contiene parámetros desconocidos.

Sea F la función de distribución, que también suponemos continua, de una posible alternativa. Entonces es bien sabido que la variable $Y_0 = F_0(X_0)$ se distribuye uniformemente en $[0, 1]$, y que $Y = F(X_0)$ tiene una cierta distribución, que dependerá de F , concentrada también en $[0, 1]$, cualquiera que sea la alternativa F .

Mediante esta consideración pasamos al problema de contrastar la hipótesis nula de que unos ciertos datos transformados provienen de una distribución uniforme en $[0, 1]$. Esto nos conduce a lo siguiente:

2.^a La elección de una familia paramétrica en $[0, 1]$, que contenga como caso particular a la distribución uniforme, y que sea lo suficientemente rica para contener alternativas interesantes.

Bernardo (1982) al tratar un problema análogo al nuestro, considera la familia de distribuciones «alfa» de dos parámetros, mientras que nosotros consideramos la familia «beta» simétrica.

Los resultados del test, que ofrecemos en la sección 2, serán óptimos frente a alternativas F tales que $F(X_0)$ se distribuyese como una beta simétrica.

2. Análisis bayesiano del modelo beta simétrico

Supongamos que X_1, \dots, X_n es una muestra aleatoria de una distribución beta de parámetros α y β , con $\alpha = \beta$, que la representaremos por $Be(x/\alpha)$. Su función de densidad vendrá dada por

$$f(x/\alpha) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} \{x(1-x)\}^{\alpha-1} \quad ; \quad 0 < x < 1 \quad ; \quad \alpha > 0$$

La función de verosimilitud del parámetro α , para la muestra X_1, \dots, X_n será por tanto

$$L(\alpha; x_1, \dots, x_n) = \left\{ \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} (G(\mathbf{x}))^{\alpha-1} \right\}^n$$

donde $G(\mathbf{x})$ es la media geométrica de las $x_i(1-x_i)$, es decir,

$$G(\mathbf{x}) = \left(\prod_{i=1}^n x_i(1-x_i) \right)^{1/n}$$

que es un estadístico suficiente para el parámetro α . Como además

$$A(G, n) = \int_0^\infty \left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} G^{\alpha-1} \right]^n d\alpha < +\infty$$

para $0 < G < 0,25$, $n \geq 1$, existe una familia conjugada para α , (De Groot, 1970) dada por la familia de densidades

$$g(\alpha/G, n) = \frac{1}{A(G, n)} \left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} G^{\alpha-1} \right]^n$$

Así, si podemos representar la información a priori sobre α por un miembro de esta familia, digamos, por $g(\alpha/G_0, n_0)$, la distribución a posteriori vendrá dada por $g(\alpha/G_1, n_1)$, donde $n_1 = n_0 + n$ y $G_1 = G_0^{n_0 \cdot n_1} \cdot G^n$.

El aspecto de esta familia de densidades es similar al de la familia gamma de dos parámetros.

Cuando no se disponga de información a priori o bien ésta no pudiese representarse por un miembro de la familia, hemos de recurrir a encontrar la distribución de referencia o no informativa.

Si utilizamos el criterio de Bernardo (1979) de maximizar la información desconocida, como nos encontramos en el caso regular, éste es equivalente a aplicar la regla de Jeffreys y por lo tanto la distribución a priori de referencia es proporcional a la raíz cuadrada de la cantidad de información de Fisher.

En nuestro caso tenemos que

$$\log L(\alpha; \mathbf{x}) = n \{ \log \Gamma(2\alpha) - 2 \log \Gamma(\alpha) + (\alpha - 1) \log G(\mathbf{x}) \}$$

Derivando esta expresión dos veces respecto al parámetro α obtenemos:

$$\frac{\hat{\alpha} \log L(\alpha; \mathbf{x})}{\hat{\alpha}} = n \{ 2\Psi(2\alpha) - 2\Psi(\alpha) + \log G(\mathbf{x}) \}$$

y

$$i(\alpha) = -E \left\{ \frac{\hat{\alpha}^2 \log L(\alpha; \mathbf{x})}{\hat{\alpha}^2} \right\} = 2n \{ \Psi'(\alpha) - 2\Psi'(2\alpha) \}$$

y por lo tanto, la distribución de referencia $g(\alpha)$ es proporcional a $\sqrt{\Psi'(\alpha) - 2\Psi'(2\alpha)}$, donde Ψ es la función digamma y Ψ' su derivada, conocida generalmente como función trigamma.

Es fácil comprobar que la distribución de referencia

$$g(\alpha) = \sqrt{\Psi'(\alpha) - 2\Psi'(2\alpha)}$$

es impropia y que para valores pequeños de α se comporta como la función $1/\alpha$.

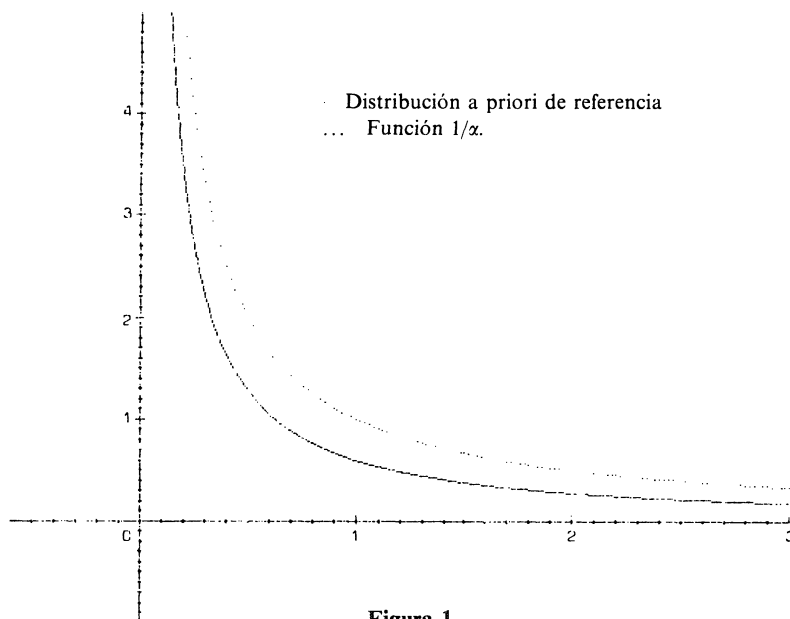


Figura 1.

La distribución a posteriori de referencia será por tanto proporcional a

$$\left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} G^{\alpha-1} \right]^n \sqrt{\Psi'(\alpha) - 2\Psi'(2\alpha)}$$

Como

$$S(G; n) = \int_0^\infty \left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} G^{\alpha-1} \right]^n \sqrt{\Psi'(\alpha) - 2\Psi'(2\alpha)} d\alpha < +\infty$$

para $0 < G < 0,25$; $0 < \alpha < +\infty$, la distribución a posteriori de referencia será

$$g(\alpha/G; n) = \frac{1}{S(G, n)} \left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} G^{\alpha-1} \right]^n \sqrt{\Psi'(\alpha) - 2\Psi'(2\alpha)}$$

Esta distribución puede aproximarse por una distribución gamma $G(\alpha/a, p)$ para muestras muy pequeñas ($n \simeq 5$). Incluso para tamaños de 1 y 2 la aproximación es excelente. Esta se basa en el resultado empírico de que la media de la distribución a posteriori es aproximadamente igual al estimador de máxima verosimilitud y su precisión se aproxima muy bien por la precisión asintótica que viene dada por

$$2n(\Psi'(\hat{\alpha}) - 2\Psi'(2\hat{\alpha}))$$

donde $\hat{\alpha}$ es el estimador de máxima verosimilitud. Es importante hacer notar que el estimador de máxima verosimilitud, solución de la ecuación

$$2\Psi(2\alpha) - 2\Psi(\alpha) + \log G(x) = 0$$

no depende del tamaño muestral sino de la media geométrica $G(x)$. En la figura 2 se muestra la variación del estimador máximo-verosímil con la media geométrica.

Si representamos por $\hat{\alpha}(G)$ el estimador de máxima verosimilitud, que es función del estadístico suficiente G , la distribución asintótica

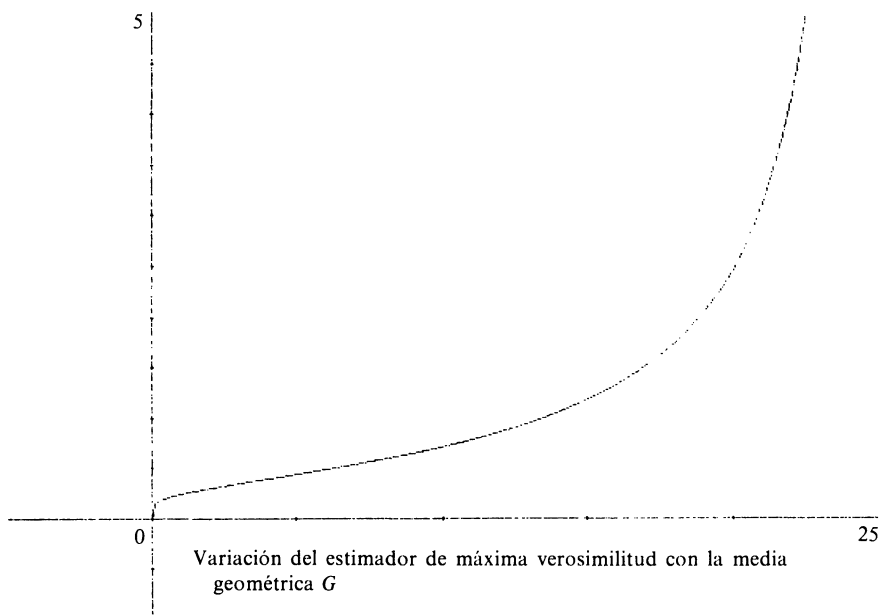


Figura 2.

del parámetro α , es aproximadamente normal con media y precisión dadas por

$$\alpha/G, n \sim N(\alpha; \hat{\alpha}(G), 2nF(\hat{\alpha}(G)))$$

donde

$$F(\alpha) = \Psi'(\alpha) - 2\Psi'(2\alpha)$$

Para tamaños muestrales menores que 100, la distribución a posteriori puede aproximarse, como acabamos de indicar, por una distribución gamma de parámetros

$$\alpha/G, n \sim G(\alpha; a, p)$$

siendo

$$a = 2n\hat{\alpha}F(\hat{\alpha}) \quad ; \quad p = 2n\hat{\alpha}^2F(\hat{\alpha})$$

A partir de la distribución a posteriori, o de su aproximación, pueden realizarse todo tipo de inferencias. En particular el contrastar la hipótesis de que las observaciones provienen de una distribución uniforme sería equivalente a determinar un intervalo de máxima densidad a posteriori (de probabilidad predeterminada) y comprobar si el valor de $\alpha = 1$ pertenece a dicho intervalo.

3. Contrastes de significación puros: aplicación a la discriminación de modelos

En este apartado aplicamos los resultados del análisis bayesiano del modelo beta simétrico a lo que Cox y Hinkley (1974, p. 64) denominan «contrastes de significación puros». Consideramos el caso en que la hipótesis nula sea simple y el caso en que esta conste de un número finito de alternativas. El caso de que la hipótesis nula contenga parámetros marginales será considerado en otro artículo posterior.

La idea es totalmente análoga a la seguida por Bernardo (1982), con la diferencia de que aquí utilizamos la familia «beta» en lugar de la «alfa».

3.1. CONTRASTE DE UNA HIPOTESIS NULA SIMPLE

Supongamos que se desea contrastar la hipótesis nula H_0 : la muestra aleatoria simple X_1, \dots, X_n proviene de una distribución unidimensional continua completamente especificada $F_0(x)$ (es decir, sin parámetros marginales).

Este problema se reduce a contrastar la hipótesis nula H'_0 : la muestra aleatoria simple Y_1, \dots, Y_n , donde $Y_i = F(X_i)$, $i = 1, 2, \dots, n$, proviene de una distribución uniforme en $[0, 1]$.

Como la distribución uniforme en $[0, 1]$ es un caso particular de la distribución beta simétrica cuando $\alpha = 1$, el problema anterior lo podemos asimilar a contrastar la hipótesis nula $H''_0: \alpha = 1$ en el modelo beta simétrico, para lo cual se aplican las técnicas de la sección anterior.

3.2. CONTRASTE DE MODELOS

Consideremos el caso en que la hipótesis nula que queremos contrastar conste de un número finito de modelos. Para ser más específicos, supongamos que queremos contrastar la hipótesis nula H_0 : la muestra aleatoria simple X_1, \dots, X_n , proviene de alguna de las distribuciones unidimensionales continuas $F_1(x), \dots, F_k(x)$, completamente especificadas.

De hecho, en el caso de que la hipótesis nula se acepte, pretendemos saber, además, cual es la distribución a posteriori de los modelos de que consta la hipótesis nula.

Antes de entrar en el análisis bayesiano de éste problema conviene estudiar previamente la función de verosimilitud conjunta del parámetro α y del parámetro G . Conviene recordar que G depende ahora del modelo considerado, que podemos tratar como parámetro marginal.

La función de verosimilitud conjunta de α y G es

$$L(n; \alpha, G) = \left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} G^{\alpha-1} \right]^n ; \quad 0 < \alpha < +\infty ; \quad 0 < G < 0,25$$

Esta función posee un punto de silla para $\alpha = 1$ y $G = \exp(-2)$ en el que se anulan las dos derivadas primeras y constituye, por tanto,

una solución de las ecuaciones de verosimilitud

$$\frac{\hat{\partial}L}{\hat{\partial}\alpha} = 0 \quad ; \quad \frac{\hat{\partial}L}{\hat{\partial}G} = 0$$

Sin embargo, esta solución no corresponde a un máximo de la función de verosimilitud. Un análisis más detallado revela que el punto $\alpha = 1$, $G = \exp(-2)$ es un punto de silla de la función $L(n; \alpha, G)$ cualquiera que sea el valor del tamaño muestral n . Además, la función de verosimilitud se hace infinita cuando $\alpha \rightarrow 0$ y $G \rightarrow 0$, o cuando $\alpha \rightarrow +\infty$ y $G \rightarrow 0,25$.

Desde un punto de vista intuitivo, este comportamiento es, sin embargo, completamente lógico. Pequeños valores de la media geométrica G indican una fuerte evidencia de que α es pequeño; por otra parte, valores de G próximos a 0,25 evidencian que la distribución está muy concentrada y, por tanto, que α es grande.

Este comportamiento anómalo de la verosimilitud hace imposible que la función de verosimilitud conjunta se distribuya asintóticamente como una normal bivariente.

La aplicación del método de Bernardo (1979) de maximizar la información desconocida al caso de considerar parámetros marginales, bien G si lo que nos interesa es la distribución del parámetro α , para aceptar o rechazar H_0 , o bien α si lo que nos interesa es la distribución a posteriori de los diferentes modelos de la hipótesis nula, parece complicada al no disponer de resultados asintóticos referentes a la función de verosimilitud conjunta. A la vista de la función de verosimilitud, la distribución de referencia en ambos casos, debería dar poca probabilidad a los entornos de $\alpha = 0$, $G = 0$ y de $\alpha = +\infty$, $G = 0,25$ y concentrarse alrededor del punto $\alpha = 1$, $G = \exp(-2)$.

Con el fin de solventar esta dificultad, proponemos la siguiente solución de tipo heurístico.

Para contrastar primero H_0 , sean $D = \{x_1, \dots, x_n\}$ los datos originales, y representemos por $D_i = \{F_1(x_1), \dots, F_i(x_n)\}$ los datos transformados correspondientes al modelo i -ésimo. La distribución a posteriori del parámetro α supuesto que el verdadero modelo es el i -ésimo y los datos observados son D vendrá dada por

$$g(\alpha/i; D) = \frac{1}{S(G_i; n)} \left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2} G_i^{\alpha-1} \right]^n \sqrt{F(\alpha)}$$

donde

$$G_i = \left\{ \prod_{j=1}^n F_i(x_j)[1 - F_i(x_j)] \right\}^{1/n}$$

Para cada uno de los modelos se contrasta si $\alpha = 1$. Si este valor no es soportado por ninguno de los modelos, la hipótesis nula será rechazada.

Caso contrario, el modelo que mejor soportase el valor $\alpha = 1$ sería el modelo probabilístico mejor adaptado a los datos. De modo que si $p(i)$ representa la probabilidad a priori del modelo i -ésimo, la probabilidad a posteriori de dicho modelo, dados los datos, $p(i/D)$, vendría dada por

$$p(i/D) \propto p(i)g(1/i; D)$$

de donde

$$p(i/D) = \frac{p(i)/S(G_i; n)}{\sum_{i=1}^n p(i)/S(G_i; n)}$$

A modo de ilustración del método se simuló la siguiente muestra de tamaño 100 (tabla 1), de una distribución normal $N(0, 1)$ y se contrastó la hipótesis nula H_0 : la muestra proviene de uno de los tres modelos siguientes:

$M1$: Normal $N(0, 1)$

$M2$: Cauchy $Ca(0, 0,2605)$

$M3$: Exponencial doble $ED(0, \sqrt{0,5})$

El primer modelo es el verdadero; el segundo se eligió también simétrico, centrado en cero y tal que el percentil 95 fuese igual que el de la normal $N(0, 1)$. El tercero también se eligió simétrico, de media cero y varianza unidad.

Los resultados del análisis se muestran en las figuras 3A, 3B y 3C que indican claramente que los modelos 1 y 3 explican bien la muestra, siendo el primer modelo mejor que el tercero. El segundo modelo

Tabla 1

1,2306132	-,29134232	,39220857	-,86437819	1,1113666
-,48058559	1,4442753	1,8257632	-,13631486E-02	,76291885
-,96992044E-02	,68747243	,20899657	,65372514	-,16224358
,42164983	-,19477349	,41928465	-,1,2159723	2,0521986
1,4403960	1,6854693	-,1,2451966	,81241214	-,1,6794818
1,4891025	,50670096	2,5745290	2,8145234E-03	-,40698112
-,89757464	,44837617	-,6,1762796E-02	,48797083	-,40300332
-,34017949	,69571090	,24650416	-,69058295	-,18345239
1,2255440	-,7,3278841E-02	,88070584	-,54251081	-,6,7432342E-02
,38865177	,55222293	2,8430182	-,8,8676747E-02	-,73724718
,10299988	-,1,0441678	-,60358264	-,2,3211448	-,79840089
,52724431	,73396504	,70467920	,83682114	-,1,3395975
-,1,1351589	-,1,0641626	1,4723798	,57075696	,50725904
,20404613	-,69653477	-,41510497	-,1,7850013	7,0646895E-03
-,64208448	,51166461	,74840103	1,2926824	,26151546
-,29569477	1,4080515	,12987375	,55574395	-,35089266
2,5871262E-02	-,74695809	-,2,9108123E-03	2,4919450E-02	-,73486365
,35006961	,95985185	-,64137665	1,0297449	-,63830152
-,1,2108437	,33449730	-,39614778	,24855621	-,1,2123103
-,1,6198127	1,3157943	,14295182	-,1,2515105	-,57119391

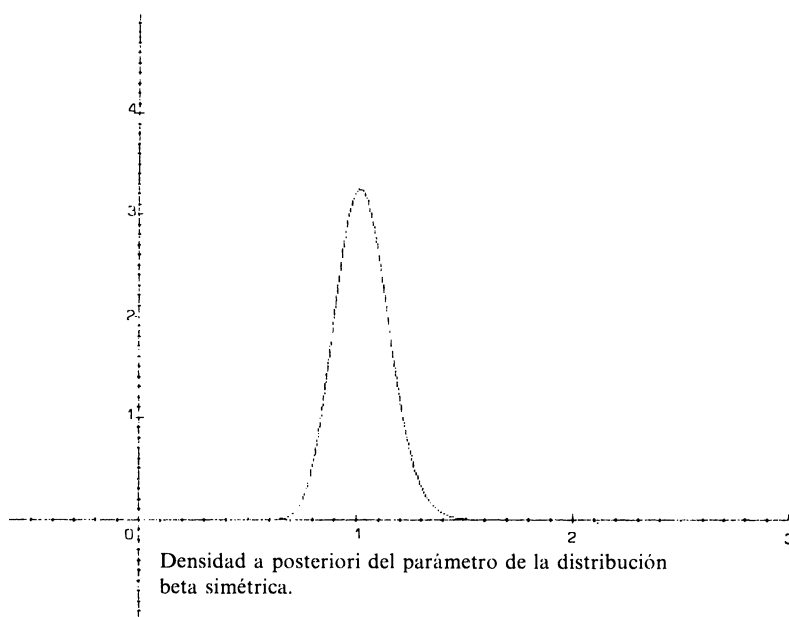


Figura 3A. Modelo Normal.

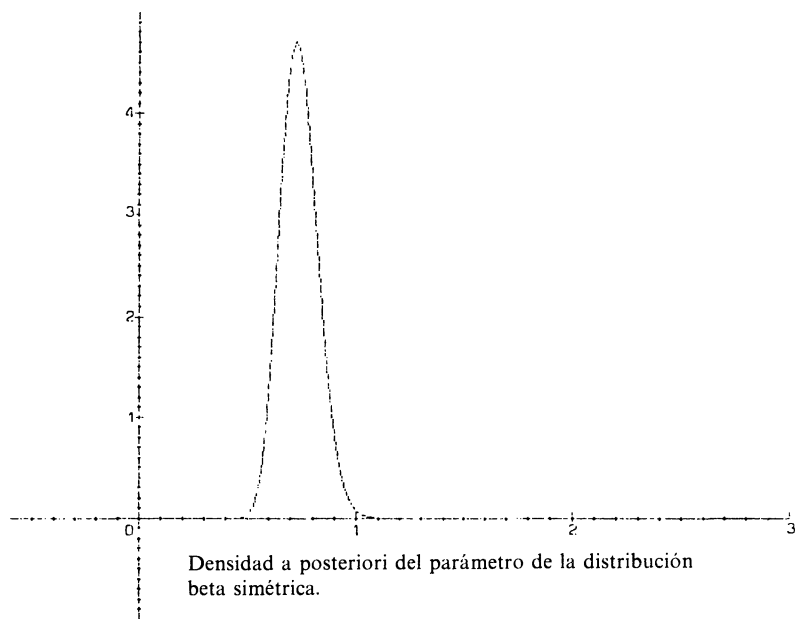


Figura 3B. Modelo de Cauchy.

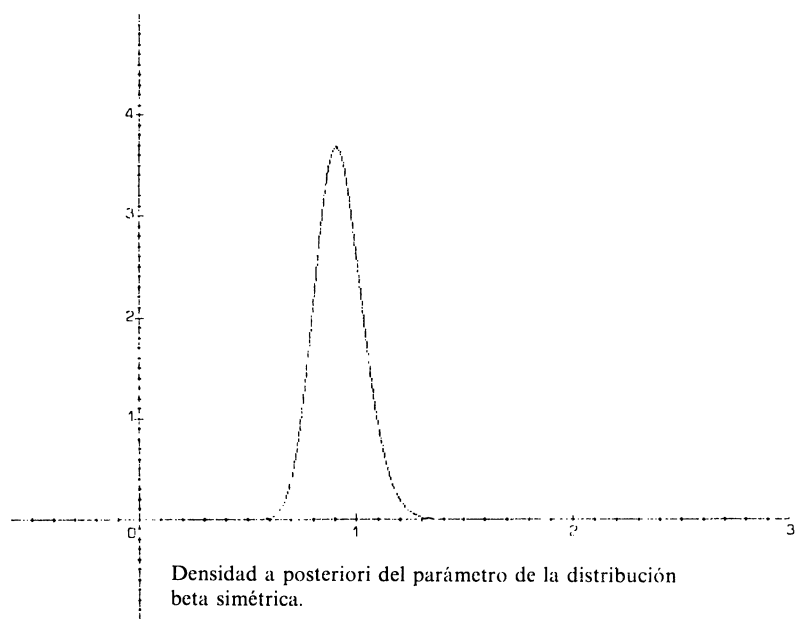


Figura 3C. Modelo exponencial doble.

da poca probabilidad a posteriori al valor $\alpha = 1$, por lo que este modelo se rechaza.

Los parámetros de interés referentes a los tres modelos se resumen en la tabla 2.

Obsérvese la igualdad hasta cuatro cifras decimales de la media verdadera con el estimador de máxima verosimilitud. La desviación típica asintótica no aparece en la tabla, puesto que también coincide, hasta cuatro cifras decimales con la verdadera, calculada por integración numérica.

Suponiendo a priori los tres modelos igualmente posibles, las probabilidades a posteriori de estos modelos son

$$p(M1/D) = 0,5381 \quad ; \quad p(M2/D) = 0,0143 \quad ; \quad p(M3/D) = 0,4476$$

Para poder discriminar entre los modelos 1 y 3, habría hecho falta una muestra mayor o bien, como se señala en los comentarios, una familia más rica de distribuciones en el intervalo unidad como podría ser la mixtura de distribuciones beta con dos parámetros.

Tabla 2

Modelos Parámetros	Normal	Cauchy	Exp. doble
Media geométrica G	0,1384	0,1048	0,1277
Est. de m.v. $\hat{\alpha}$	1,0332	0,7419	0,9254
Media	1,0332	0,7419	0,9254
Moda	1,0185	0,7321	0,9126
Desviación Típica	0,1230	0,0852	0,1089
Aproximación $Ga(\alpha/a, p)$	$a = 68,313$ $p = 70,579$	$a = 102,180$ $p = 75,809$	$a = 78,071$ $p = 72,240$
$S(G_i; n)$	0,1842	6,9127	0,2214

4. Comentarios

Debido a que la familia beta simétrica de un sólo parámetro α , no es lo suficientemente rica para incluir otras alternativas interesantes, no cabe esperar gran poder discriminatorio del contraste. Incluso la familia beta de dos parámetros, para la que pretendemos generalizar estos resultados, no sería suficientemente flexible, como se pone de manifiesto al comparar una distribución normal y una exponencial doble con los dos primeros momentos iguales. Debido a esto, pretendemos ampliar los resultados al caso de mixturas de distribuciones beta, así como a hipótesis nulas compuestas (es decir, con algún parámetro marginal).

Por otro lado, el test propuesto da cierta información, caso de rechazar la hipótesis nula, acerca de las posibles alternativas. Así, el que la media geométrica G sea sensiblemente mayor que $\exp(-2)$, indica que la verdadera hipótesis nula tiene menor dispersión que la postulada. Caso que la media geométrica fuese pequeña ($\exp(-2)$) indicaría que, o bien la verdadera distribución nula es más dispersa que la postulada o bien que la hipótesis nula está «desplazada» respecto de la verdadera. El test no da sin embargo indicación del signo del desplazamiento, cosa que se consigue con el modelo beta de dos parámetros.

REFERENCIAS

- BERNARDO, J. M. (1979): «Reference posterior distributions for Bayesian inference», *J. Roy. Statist. Soc.*, B, **41**, 113-147 (with discussion).
- BERNARDO, J. M. (1982): «Contraste de modelos probabilísticos desde una perspectiva bayesiana», *Trab. Estadist.*, **33,2**, 16-30.
- BREIMAN, L. (1973): *Statistics: With a View toward Applications*. Houghton Mifflin Co. Boston.
- COX, D. R., and HINKLEY, D. V. (1974): *Theoretical Statistics*. Chapman and Hall. London.
- DE GROOT, M. H. (1970): *Optimal Statistical Decisions*. McGraw-Hill Co. New York.