Revista Investigación Educativa - Vol. 8 - n.º 16 - 1990 (P. 509-516)

INFERENCIA BAYESIANA SOBRE UNA PROPORCIÓN

por
José Serrano Ángulo
Dpto. de Didáctica y Organización Escolar
Universidad de Málaga

1. RESUMEN

La investigación en las ciencias sociales y en particular en CC.EE. la mayoría de las variables no son cuantitativas, por lo que uno de los análisis más frecuentes es el estudio de las frecuencias y el de las proporciones. Aunque el tamaño de la muestra sea considerable, en la mayoría de las investigaciones, este se divide por una o más variables, quedando submuestras pequeñas.

En esta comunicación se estudia el porcentaje de centros con biblioteca, según los datos de la investigación «Evaluación de la Reforma en el Ciclo Superior de la E.G.B. en Andalucía». En este caso se dispone de una muestra de 48 centros con reforma divididos en zona rural (38) y zona urbana (10) y una muestra de control de 8 centros igualmente divididos en zona rural (4) y zona urbana (4). Según la teoría clásica no se puede estimar la proporción de centros con biblioteca con muestras pequeñas, por lo que propongo un análisis bayesiano de la proporción. En la sección 2 se expone brevemente el modelo y la función de verosimilitud. En la sección 3 se enumera algunos inconvenientes y ventajas de este análisis frente al clásico. Por último, en la sección 4, se calcula las funciones de verosimilitud de la proporción en los distintos casos. Los cálculos y las representaciones gráficas se han hecho con ayuda del programa Mathematica.

PALABRAS CLAVES:

Inferencia Bayesiana; Reforma; Materiales.

2. EL MODELO Y LA FUNCIÓN DE VEROSIMILITUD

El proceso que genera los datos es conocido como proceso de Bernoulli y los resultados son dos: «éxito» y «fracaso». El proceso de Bernoulli tiene un parámetro π que es la probabilidad de que ocurra un «éxito». Si repetimos el mismo proceso de Bernoulli n veces, tendremos un proceso binomial en el que hay dos parámetros n y π. La probabilidad de que ocurra x «éxitos» en una prueba binomial de parámetro n π p es

$$P(X|\pi,n) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}$$

y la función de verosimilitud es

$$l(\pi|x,n) \propto \pi^x(1-\pi)^{n-x}.$$

Así, para la distribución a priori de π se puede tomar una beta de parámetros a y b con $a > 0$ y $b > 0$ que viene dada por

$$P(\pi) = \frac{1}{Be(a,b)}\pi^{a-1}(1-\pi)^{b-1}$$

donde

$$Be(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

siendo $\Gamma(a) = (a—1)!$ si a es un entero. La distribución beta tiene distintas formas para los distintos valores de a y b. Cuando $a = 1$ y $b = 1$ se tiene la distribución uniforme en el intervalo [0,1] y cuando $a = b$ es simétrica respecto de 1/2. La media es

$$\frac{a}{a+b},$$

la moda es

$$\frac{a-1}{a+b-2}$$

y la varianza es

$$\frac{\left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)}{a+b+1}.$$

Cuando a y b crecen se aproxima a la normal,

$$N\left(\frac{a}{a+b}, \frac{\left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)}{a+b+1}\right)$$

Si la distribución de π es una beta de parámetros a y b

$$p(\pi) \propto \pi^{(a-1)}(1-\pi)^{(b-1)}$$

y si en n observaciones ha ocurrido x «éxitos» la distribución a posteriori es

$$p(\pi|x,n) \propto p(\pi) \times l(\pi|x,n)$$

es decir

$$p(\pi|x,n) \propto \pi^{(a-1)}(1-\pi)^{(b-1)}\pi^{x}(1-\pi)^{(n-x)}$$

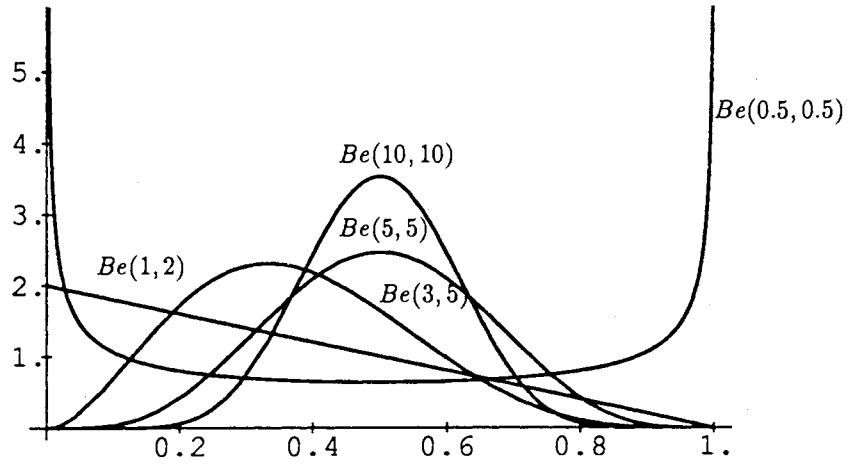
reagrupando términos se tiene

$$p(\pi|x,n) \propto \pi^{(a+x-1)}(1-\pi)^{(b+n-x-1)}$$

esto es una distribución beta de parámetros $a + x$ y $b + n - x$. Luego las distribuciones betas constituyen la familia de distribuciones conjugadas para el parámetro π.

3. INCONVENIENTES Y VENTAJAS DEL ANÁLISIS BAYESIANO DE LAS PROPORCIONES

Una de las mayores controversia en el análisis bayesiano es la elección de la distribución a priori. En el caso de la proporción, la manera en que se combina los parámetros sugiere que la distribución a priori puede considerarse como la distribución de una muestra previa de a + b observaciones. Si se considera esto, entonces la distribución a priori sería la Be(0,0), que es una distribución impropia. Otra opción podría ser la distribución uniforme Be(1,1), que supone no tener información previa en la línea que propone Jeiffreys; aunque hay autores que sugieren la distribución *Be*(1/2, 1/2), como la mejor distribución a priori para el parámetro π(véase Bernardo Girón). Con una muestra de tamaño moderado, tomando una distribución a priori u otra, las distribuciones a posteriori que se obtienen son prácticamente las mismas. No ocurre lo mismo con muestras pequeñas, que con diferentes distribuciones a priori se tiene distintos resultados. Con la distribución a priori Be(0,0) se obtiene una interpretación de los resultados más obvia, pero al ser una distribución impropia en algunos casos se obtienen distribuciones a posteriori también impropias por lo que tomaré como distribución a priori la Be(0.5,0.5). En la figura 1 se puede observar las gráficas de varias distribuciones betas.



Otra dificultad a la hora de aplicar el análisis bayesiano es el del cálculo de las probabilidades, sin embargo esto no es un problema hoy día gracias a los ordenadores y a los programas informáticos disponibles. Como hemos visto para la proporción, los cálculos que hay que hacer son sobre la distribución beta: si se toma una distribución a priori con parámetros enteros se puede tabular las probabilidades extremas de esta distribición, como se hace por ejemplo con la F de Snedecor. Tanto si se toma o no una distribución a priori con parámetros enteros se puede calcular las probabilidades de la distribución a posteriori con la ayuda del ordenador y del programa adecuado, por ejemplo el Mathematica.

En el estudio clásico de las proporciones se distingue entre muestras grandes y pequeñas, utilizando distribuciones distintas en un caso y en otro. Para el caso de muestras pequeñas se utiliza la distribución binomial B(n, π), en la que no se puede dar un intervalo de confianza para la proporción de la población. En este caso se comprueba si el número de «éxitos» «x» en n veces es extremo o no. Para muestras grandes se utiliza la distribución normal N (π, π (1-π)/n)), ya que la distribución binomial B(n, π) se puede aproximar por la normal N (nπ, nπ (1-π)). En este caso se comprueba si la proporción de la muestra es significativamente distinta o no pudiéndose dar intervalos de confianza para la proporción de la población. Al necesitarse la varianza de la distribución muestral para la estimación, se sustituye el valor de π por el de la proporción de la muestra p. Por lo que para muestras grandes las dos teorías dan el mismo resultado numérico. En cambio, para el caso de muestras pequeñas, el análisis bayesiano trabaja con la proporción al igual que con las muestras grandes, pudiéndose dar intervalos de confianza en un caso y en otro, pues la distribución de π es continua en ambos casos.

En las dos útimas décadas las soluciones otorgadas a los inconvenientes han hecho que éstos no sean considerables comparados con las siguientes ventajas:

— La posibilidad de incluir probabilidades subjetivas, si hay una sospecha de que un resultado es más posible que otro.

— No se hacen suposiciones a cerca del valor del parámetro, como se hacen en el análisis clásico, sino que se trabaja con funciones de verosimilitud, expresando los resultados en tales funciones.

— Se trabaja con la misma familia de distribuciones para muestras grandes y pequeñas. Con muestras grandes los resultados son numericamente lo mismo para ambos análisis.

— La posibilidad de añadir nuevos datos al análisis sin necesidad de repetir los cálculos.

— La regla de parada en el caso de las proporciones no influye en los resultados, es decir, da lo mismo elegir n resultados en los que hay x «éxitos», que elegir resultados hasta conseguir x «éxitos».

— Los cálculos necesarios para hallar la función de verosimilitud en el caso particular de las proporciones son muy fáciles, sólo hay que sumar los «éxitos» al primer parámetro de la distribución beta y el número de fracaso al segundo parámetro.

4. ANÁLISIS DE LAS PROPORCIONES DE CENTROS CON BIBLIOTECA

Los datos que se diponía en la investigación de «Evaluación de la Reforma en el Ciclo Superior de la E.G.B. en Andalucía», era una muestra de 48 centros con reforma divididos en zona rural (38) y zona urbana (10), y una muestra de control de 8 centros igualmente divididos en zona rural (4) y zona urbana (4) (como se muestra en la tabla 1).

TABLA 1

	sí	no	totales
muestra urbana	3	7	10
muestra rural	21	17	38
control urbana	0	4	4
control rural	1	3	4

De los 10 centros con reforma en zonas urbanas, 3 de ellos tenían biblioteca y los 7 restantes no. Tomando como distribución a priori del parámetro π la Be(0.5,0.5) se obtiene una distribución a posteriori Be(3.5, 7.5) (véase figura 2 con media igual a 3.5/11. El intervalo de máxima verosimilitud para π al 95% es [0.089,0.598].

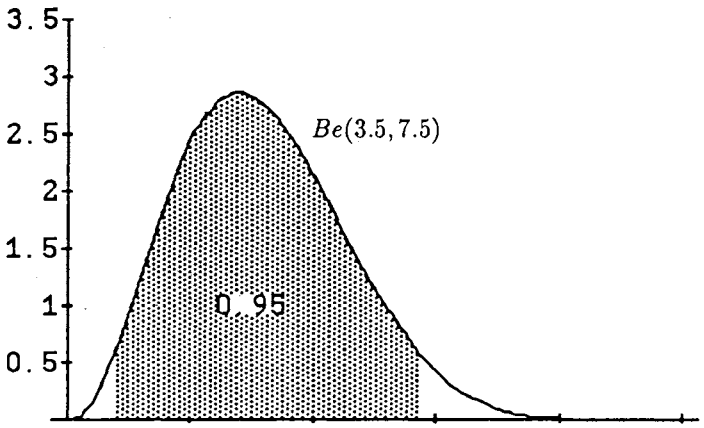
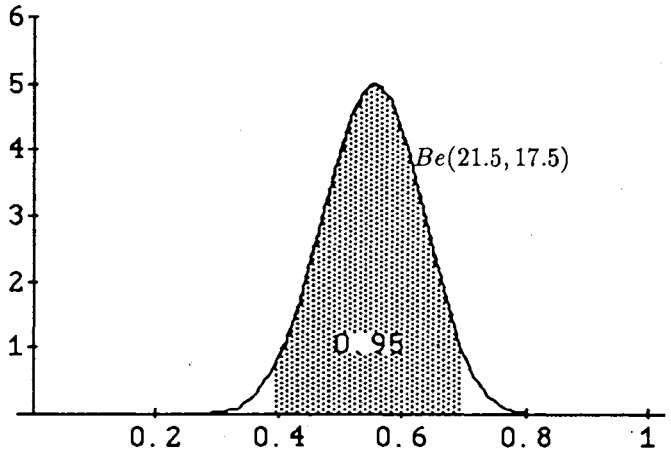
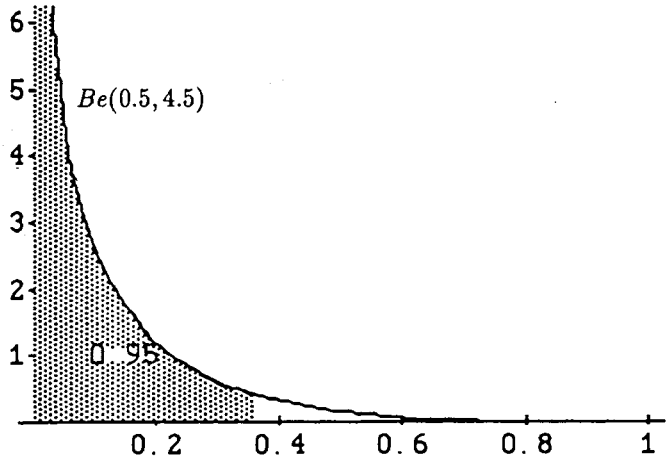


fig. 2

De los 38 centros con reforma en zonas rurales, 21 de ellos tenían biblioteca y los 17 restantes no, por lo que distribución a posteriori es una Be(21.5, 17.5) (véase figura 2 con media igual a 21.5/39. El intervalo de máxima verosimilitud para π al 95% es [0.369,0.702].



De los 4 centros de control en zonas urbanas, ninguno de ellos tenían biblioteca, por lo que la distribución a posteriori es una Be(0.5, 4.5) (véase figura 2 con media igual a 0,5/5. El intervalo de máxima verosimilitud para π al 95% es [0,0.362].



De los 4 centros de control en zonas rurales, 1 de ellos tenía biblioteca y los 3 restantes no, por lo que distribución a posteriori es una Be(1.5,3.5) (véase figura 2 con media igual a 1.5/5. El intervalo de máxima verosimilitud para π al 95% es [0.022,0.691].

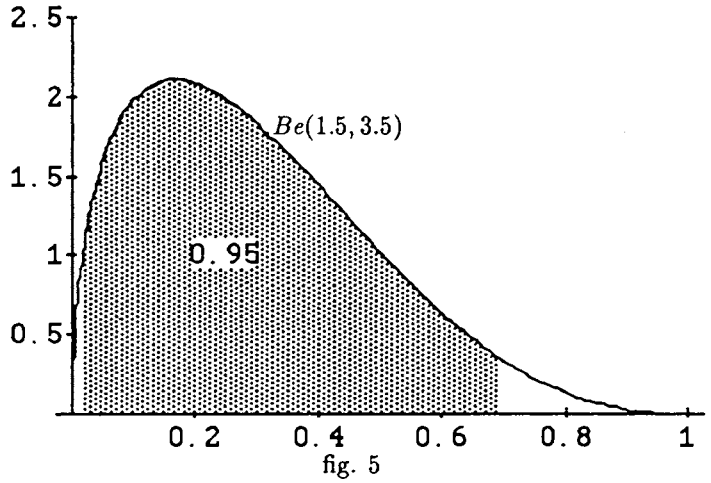


fig. 5

Se puede observar un mayor porcentaje de centros con biblioteca, entre los centros con reforma, que entre los centro de control.

5. CONCLUSIÓN

El análisis que podemos hacer, una vez aplicado éste procedimento, es el siguiente: Entre los centros con reforma hay mayor porcentaje con biblioteca que entre los centros de control.

6. REFERENCIA

BERNARDO, J. M. Y GIRÓN, F. J. (1988): A Bayesian Analysis of Simple Mixture Problems. *Bayesian Statistics* 3. pp. 61-72. J. M. Bernardo, D. V. Lindley, M. Il. DeGroot y A. F. M. Smith, eds. Valencia: University Press.