CONSEJO SUPERIOR DE INVESTIGACIONES CIENTIFICAS

TAMAÑO OPTIMO DE UNA MUESTRA: SOLUCION BAYESIANA

por
José-Miguel Bernardo Herranz
Universidad de Valencia

PUBLICADO EN LA REVISTA
TRABAJOS DE ESTADISTICA
Y DE
INVESTIGACION OPERATIVA

TAMAÑO OPTIMO DE UNA MUESTRA: SOLUCION BAYESIANA

José-Miguel Bernardo Herranz
Universidad de Valencia

I. Sumario.

El método Bayesiano de diseño de experimentos se aplica a situaciones en las que el investigador está interesado en hacer inferencias, más bien que en tomar decisiones, respecto a un parámetro $\theta \epsilon \Theta$, i.e. el espacio de decisiones es el conjunto de las posibles densidades de probabilidad respecto de una adecuada medida definida sobre Θ.

Se determinan las condiciones bajo las que la utilidad terminal esperada de una muestra debe ser proporcional a medida Shannon-Lindley de la información que se espera proporcione sobre θ

Se propone el único método coherente bajo tales condiciones de seleccionar el tamaño óptimo de la muestra. Se obtienen expresiones asintóticas válidas para situaciones en las que existe una gran información a priori sobre el parámetro θ.

II. Especificación del problema.

Supóngase que se pretende realizar inferencias sobre el valor de un vector $\theta \epsilon \Theta$, Θ subconjunto de un espacio euclídeo de k dimensiones Con tal objeto se deben realizar un cierto número de observaciones independientes de un cierto proceso $p_x(\cdot/\theta)$, i e se debe extraer una muestra de una cierta población X cuya distribución depende del parámetro objeto de la investigación.

Nuestro objetivo es producir un método, esencialmente único, de determinar el tamaño óptimo de tal muestra.

Se han publicado numerosos trabajos sobre este tema pero, en la medida de nuestros conocimientos, los resultados ofrecidos son procedimientos ad hoc únicamente sostenidos mediante razonamientos empíricos. En este trabajo se especifican las condiciones bajo las que la solución ofrecida es la única compatible con un comportamiento coherente del investigador, es decir un comportamiento sujeto a unos pocos, fuertemente intuitivos, axiomas de consistencia o racionalidad. El trabajo se inscribe íntegramente en la estructura Bayesiana.

Para un comportamiento coherente ante la elección de un experimento es necesario asignar una distribución de probabilidad sobre Θ que describa las opiniones iniciales del investigador sobre el parámetro θ, especificar el conjunto de las posibles decisiones, construir una función de utilidad que evalúe las preferencias del investigador entre sus posibles consecuencias y seleccionar el experimento que proporcione la mayor utilidad esperada (Ramsey, 1926; De Finetti, 1937; Savage, 1954; Pratt et al., 1964; Lindley, 1971).

Formalmente, sea $\overset{\circ}{p}_\theta(\cdot)$ una densidad de probabilidad respecto a la medida de Lebesgue que describe las opiniones iniciales del investigador respecto de θ. Si se extrae una muestra de tamaño n se obtiene un cierto vector $z=(x_1, x_2, \ldots, x_n)$, $x_i \epsilon X$, relacionado con θ mediante una densidad de probabilidad

$$p_z(\cdot/\theta, n) = \prod_{i=1}^{n} p_x(\cdot/\theta)$$

respecto a una cierta medida σ-finita dominante definida sobre X^n, integración con respecto a la cual denotaremos por $\int \ldots dz$.

Una vez obtenido z el investigador debe modificar sus opiniones sobre θ a (digamos) $p_\theta(\cdot/z, n)$ de acuerdo con el teorema de Bayes.

En este trabajo la atención se centra en una situación en la que el investigador no tiene en mente un conjunto específico de decisiones: simplemente quiere hacer inferencias sobre θ. Esto es típicamente cierto en la investigación científica o social. Las acciones finales del investigador son, en consecuencia, afirmaciones probabilísticas sobre θ. Tales afirmaciones sólo dependen de la densidad de probabilidad sobre θ que describe las opiniones finales del investigador sobre el parámetro de interés θ. Argüimos, por lo tanto, que el espacio de decisiones adecuado a un problema de inferencia es el conjunto de las densidades de probabilidad sobre Θ.

La conveniencia de terminar una investigación presentando la densidad final sobre Θ en lugar de limitarse a presentar algún tipo de estimador de θ ya ha sido reconocida (Box & Tiao, 1973; Dickey, 1973). Pondremos de manifiesto que el uso del conjunto de posibles densidades finales de decisiones de una satisfactoria solución, en el marco de la teoría de la decisión, a los problemas de inferencia. Específicamente, definimos el espacio de decisiones D como el conjunto de densidades de probabilidad respecto a la medida de Lebesgue definida en Θ, estrictamente positivas:

$$D = \Big\{ p_\theta(\cdot) \text{ funciones medibles-Lebesgue reales, definidas sobre } \Theta;$$

$$p_\theta(\theta) > 0 \ \forall \theta \epsilon \Theta, \int_\Theta p_\theta(\theta)\, d\theta = 1 \Big\}$$

Si, finalmente, suponemos que:

a) Las preferencias del investigador ante los posibles resultados finales de la investigación pueden describirse mediante una función de utilidad $u: D x \Theta \to R$ de forma que $u(p_\theta(\cdot), \theta)$ es la utilidad que para el investigador tiene terminar la investigación presentando como densidad final $p_\theta(\cdot)$ cuando θ es el verdadero valor del parámetro.

b) $c(z, n)$ es el coste, medido en unidades de utilidad, de extraer una muestra de tamaño n y obtener el vector de resultados z. Entonces, (Raiffa & Schlaifer, 1961):

a) Si no se extrae muestra alguna, manteniendo las opiniones iniciales sobre θ, el investigador puede esperar una utilidad u_0 dada por

(1) $$u_0 = \int_\Theta u(\overset{\circ}{p}_\theta(\cdot), \theta)\, \overset{\circ}{p}_\theta(\theta)\, d\theta$$

b) Si se extrae una muestra de tamaño n el investigador puede esperar una utilidad terminal u_n a un coste $c(n)$ dados por

$$u_n = \int_{X^n} \max_{p_\theta(\cdot)\in D} \left\{\int_\Theta u(p_\theta(\cdot),\theta)\, p_\theta(\theta/z,n)\,d\theta\right\} p_z(z/n)\,dz \tag{2}$$

$$c(n) = \int_{X^n} c(z,n)\, p_z(z/n)\,dz \tag{3}$$

donde $p_z(z/n) = \int_\Theta p_z(z/\theta, n)\, \overset{\circ}{p}_\theta(\theta)\,d\theta$ (densidad predictiva)

Consecuentemente, el aumento de utilidad que se espera proporcione la extracción de una muestra de tamaño n, $u^*(n)$, viene dado por

$$u^*(n) = u_n - u_0 - c(n) \tag{4}$$

Y el tamaño muestral óptimo es el que maximiza (4).

La discusión precedente proporciona la solución al problema propuesto una vez la función u de utilidad ha sido especificada.

El objeto de este trabajo es encontrar la forma que tal función debe adoptar si las preferencias del investigador satisfacen determinados axiomas frecuentemente verificados en la investigación científica, y obtener entonces soluciones explícitas al problema.

III. La utilidad de una muestra.

Caracterizaremos a continuación ciertas posibles propiedades de la función de utilidad:

*Definición 1. $u: D x \Theta \to R$ es una función de utilidad **propia** si para todo $p'_\theta(\cdot) \in D$*

$$\max_{p_\theta(\cdot)\in D} \int_\Theta u(p_\theta(\cdot),\theta)\, p'_\theta(\theta)\,d\theta = \int_\Theta u(p'_\theta(\cdot),\theta)\, p'_\theta(\theta)\,d\theta \tag{5}$$

y $p'_\theta(\cdot)$ es la única densidad para la que el máximo es alcanzado.

Si es investigador acepta los principios de coherencia y quiere decir la verdad, esto es, terminar la investigación con afirmaciones sobre el parámetro θ que reflejan el estado final de sus opiniones, entonces sus preferencias son descritas por una función de utilidad propia. En efecto, si la función de utilidad u no satisface (5) el investigador podría maximizar u_n en (2) presentando como resultado final una densidad $p_\theta(\cdot)$ distinta de la $p_\theta(\cdot/z, n)$ que por coherencia debe reflejar sus verdaderas opiniones finales.

*Definición 2. $u: D x \Theta \to R$ es una función de utilidad **local** si para todo $p_\theta(\cdot) \in D$ y $\theta \in \Theta$.*

$$u(p_\theta(\cdot),\theta) = h(p_\theta(\theta),\theta) \tag{6}$$

Resulta conveniente suponer que $h: (0,\infty) x \Theta \to R$ es continuamente derivable respecto a su primera variable. Si la recompensa que el investigador espera obtener terminando su trabajo con unas opiniones finales sobre el parámetro de interés descritas por $p_\theta(\cdot)$ depende sólo de la probabilidad que tal densidad asigna a un pequeño intervalo que contenga a su verdadero valor, entonces sus preferencias son adecuadamente descritas por una función local de utilidad.

En tales condiciones, pueden establecerse los siguientes resultados (Bernardo, 1974):

Teorema 1. Si $u: D x \Theta \to R$ es una función de utilidad propia y local, entonces existen una constante $A > 0$ y una función real $B(\cdot)$ tales que

(i) $\quad u(p_\theta(\cdot),\theta) = A \cdot \log\{p_\theta(\theta)\} + B(\theta)$

(ii) $\quad u_n - u_0 = g \cdot I_\theta(n)$

donde $g = A/\log(2)$ e $I_\theta(n)$ es la información (Shannon, 1948; Lindley, 1956) que puede esperarse sobre θ mediante la extracción de una muestra de tamaño n y viene dada por

$$I_\theta(n) = \int_{X^n}\int_\Theta \log_2\left\{\frac{p_\theta(\theta/z,n)}{\overset{\circ}{p}_\theta(\theta)}\right\} p_\theta(\theta/z,n)\,d\theta\; p_z(z/n)\,dz \tag{7}$$

Consecuentemente, si las preferencias del investigador son descritas por una función de utilidad propia y local, entonces el incremento de utilidad que puede esperar de la extracción de una muestra de tamaño n viene dado por $g \cdot I_\theta(n)$, por lo que:

a) La constante g debe interpretarse como la utilidad esperada que el investigador asigna a una unidad de información, i.e. (Renyi, 1966) a una afirmación del tipo

$$\theta \in S \subset \Theta \text{ con } \int_S \overset{\circ}{p}_\theta(\theta)\, d\theta = \frac{1}{2}$$

b) El tamaño óptimo de la muestra es el que maximiza la expresión

(8) $$g \cdot I_\theta(n) - c(n)$$

III. Información proporcionada por una muestra

La puesta en práctica de la solución obtenida en (8) exige el cálculo de $I_\theta(n)$ para el proceso $p_x(\cdot/\theta)$ que represente la situación experimental considerada. El cálculo directo mediante (7) conduce a resultados complicados, difíciles de interpretar, y a menudo penosos de obtener. Parece, en consecuencia, adecuado buscar una aproximación válida en las condiciones experimentales más habituales.

Dada la tratabilidad y amplio espectro de situaciones reales que cubre la familia exponencial, nos limitaremos aquí a tratar el problema en el caso en que el proceso $p_x(\cdot/\theta)$ pertenece a la familia exponencial. El método, sin embargo, puede generalizarse a otros procesos (Bernardo, 1974).

Definición 3. Dado un proceso $p_x(\cdot/\theta)$ de la familia exponencial, i.e. de la forma

$$p_x(x_i/\theta) = F(x_i) \cdot G(\theta) \cdot \exp\{t(x_i) \cdot \phi(\theta)\}$$

*definiremos su **parámetro natural** ψ mediante*

$$\psi = \psi(\theta) = \phi(\theta) \quad \text{si} \quad \phi(\Theta) = R = (-\infty, \infty)$$

$$\psi = \psi(\theta) = \log\{\phi(\theta) - a\} \quad \text{si} \quad \phi(\Theta) = (a, \infty)$$

$$\psi = \psi(\theta) = \log\{b - \phi(\theta)\} \quad \text{si} \quad \phi(\Theta) = (-\infty, b)$$

$$\psi = \psi(\theta) = \log\{(\phi(\theta) - a)/(b - \phi(\theta))\} \quad \text{si} \quad \phi(\Theta) = (a \cdot b)$$

(Se sabe –Lehmann, 1959– que $\phi(\Theta)$ debe ser un intervalo finito o infinito).

Definición 4. Considérese un proceso $p_x(\cdot/\theta)$ de la familia exponencial y sea ψ su parámetro natural

*(i) definiremos una **densidad inicial no informativa** para θ, y la representamos por $\pi_\theta(\cdot)$, mediante*

$$\pi_\theta(\theta) \propto \left| \frac{d\psi(\theta)}{d\theta} \right|, \quad \forall \theta \in \Theta$$

*(ii) para cualquier densidad inicial $\overset{\circ}{p}_\theta(\cdot)$, definimos su **tamaño muestral equivalente** n_0 como la solución, si existe, de*

$$\overset{\circ}{p}_\theta(\theta) \propto \{p_x(\cdot/\theta)\}^{n_0} \cdot \pi_\theta(\theta), \quad \forall \theta \in \Theta$$

Consecuentemente, una densidad inicial no informativa respecto al parámetro de un proceso de la familia exponencial, equivale a una densidad inicial (impropia) uniforme sobre R para el parámetro natural.

Si $\overset{\circ}{p}_\theta(\cdot)$ pertenece a la familia de distribuciones conjugada del proceso (Raiffa & Schlaifer, 1961), su tamaño muestral equivalente n_0 existe siempre, y suele poder determinarse por simple inspección.

Con los elementos introducidos en las Definiciones 3 y 4, podemos construir la aproximación a $I_\theta(n)$ que nos proponíamos. En efecto (Bernardo, 1974)

Teorema 2. Si un investigador va a realizar n observaciones independientes de un proceso $p_x(\cdot/\theta)$ perteneciente a la familia exponencial, y sus opiniones iniciales respecto de θ pueden describirse mediante una densidad de tamaño equivalente n_0, entonces

$$\lim_{n_0 \to \infty} \frac{I_\theta(n)}{\frac{1}{2} \log_2 \left\{ \frac{n + n_0}{n_0} \right\}} = 1$$

La aproximación obtenida

$$I_\theta(n) \simeq \frac{1}{2} \log_2 \frac{n + n_0}{n_0}$$

resulta extraordinariamente precisa incluso para valores moderados de n_0. Por otra parte, en la investigación científica es habitual empezar la fase experimental con fuertes opiniones iniciales sobre el parámetro θ objeto de la investigación.

IV. Tamaño muestral óptimo.

Cuando se pretende hacer inferencias sobre el valor del parámetro de un proceso mediante la observación de un determinado número de sus realizaciones, el Teorema 2 permite obtener una sencilla expresión del tamaño muestral más adecuado.

Si las preferencias del investigador entre los posibles resultados de su trabajo son adecuadamente descritas por una función de utilidad propia y local, entonces la solución ofrecida es la única compatible con los axiomas de coherencia.

Teorema 3. Si se quiere hacer inferencias sobre el parámetro θ de un proceso $p_x(\cdot/\theta)$ de la familia exponencial, de forma que se está dispuesto a pagar una cantidad media g por una unidad de información sobre θ,

Si el coste esperado de extraer una muestra de tamaño n es $c(n)$ tal que $c(\cdot)$ es diferenciable,

Si las opiniones iniciales del investigador sobre θ pueden expresarse mediante una densidad de tamaño equivalente n_0,

Si las preferencias del investigador son descritas por una función de utilidad propia y local,

Entonces, el tamaño muestral óptimo es n^* donde $n^* = \max(0, n')$ y n' es la solución de la ecuación

$$n' + n_0 = \frac{1}{2 \log (2)} \frac{g}{c'(n')} \tag{9}$$

El resultado del Teorema 3 es particularmente atrayente si la función de coste esperado $c(n)$ es lineal en n, como frecuentemente ocurre. En tal caso, con $c(n) = c_0 + c \cdot n$ la ecuación (9) se escribe

$$n' + n_0 = \frac{1}{2 \log (2)} \frac{g}{c} \tag{10}$$

que expresa un resultado intuitivo: el número total de observaciones requeridas, i.e. la suma del tamaño muestral equivalente a la información inicial y del número de observaciones a realizar efectivamente, debe ser una cantidad fija que depende exclusivamente de la relación g/c, del valor unitario de la información (que mide el interés de la investigación) al coste unitario de observación.

Con una función de coste lineal, y una fuerte información a priori que garantice la validez de la aproximación encontrada en el teorema 2, el número total de observaciones debe ser constante; en consecuencia, en tales condiciones no hay ventaja apreciable en un muestreo secuencial.

V. Agradecimientos

Queremos agradecer sus valiosos comentarios sobre este trabajo al Profesor Dennis V. Lindley.

VI. Referencias.

BERNARDO, J.M. (1974): *Diseño Bayesiano de la experimentación científica*. Tesis doctoral. Universidad de Valencia

BOX, G.E.P. & TIAO, G.C. (1973): *Bayesian inference in statistical analysis*. Addison-Wesley.

DE FINETTI, B. (1937): *La prevision, ses lois logiques, ses sources subjectives*. Reeditado: *Studies in subjective probability*. (Kyburg & Smokler, eds.) Wiley, 1964 : 93-158.

DICKEY, J. (1973): Scientific reporting and personal probabilities: Student hypothesis. *J. Roy. Statist. Soc. Ser B 35*. 285-305

LEHMANN, E.L. (1959): *Testing Statistical Hypothesis*. Wiley.

LINDLEY, D.V. (1956): On a measure of the information provided by an experiment. *Ann. Math. Statist* 27 986-1005.

LINDLEY, D.V. (1971): *Bayesian Statistics a review*. S.I.A.M.

PRATT, J.W., RAIFFA, H. & SHLAIFER, R. (1964): The foundations of decision under uncertainty: an elementary exposition. *J. Amer. Statist. Assoc. 59* 353-375.

RAIFFA, H & SCHLAIFER, R. (1961): *Applied Statistical Decision Theory*. Division of Research. Harvard Business School.

RAMSEY, F.P. (1926): Truth and probability. Reeditado: *Studies in subjective probability*. (Kyburg & Smokler, eds.) Wiley, 1964: 61-92

RENYI, A. (1966): *Calcul des probabilités*. Dunod.

SAVAGE, L.J. (1954): *The foundations of statistics*. Wiley.

SHANNON, C.E. (1948): A Mathematical Theory of Communication *Bell Syst. Tech. J. 27* 379-423, 623-656.