Estimador de máxima verosimilitud en la distribución hipergéometrica.

Luis Felipe Munguia Aca. Dr. Juárez Hernández Bulmaro[b]
[a,b,c] *Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias Físico Matemáticas, Puebla, Puebla, México.*
[a] felipe94buap@gmail.com, bjuarez@fcfm.buap.mx

Resumen
La distribución hipergeométrica es una distribución discreta que se define por 3 parámetros: tamaño de la población (N), conteo de eventos en la población (R) y tamaño de le muestra (n). Y el método de estimación de máxima verosimilitud que es un método habitual para ajustar el modelo y estimar sus parámetros, son dos de los conceptos más importantes que se llegan a ver en los cursos de probabilidad y estadística. Sin embargo, como menciona Hanwen Zhang (junio 2009) los estimadores de máxima verosimilitud de N y R de la distribución hipergeométrica $Hg(n; R; N)$ no son abordados en la mayoría de los libros de textos estadísticos. Por lo que el propósito de este trabajo es presentar una manera de obtener dichos estimadores utilizando el método de máxima verosimilitud.

Palabra claves: Distribución hipergeométrica, estimador de máxima verosimilitud.

Introducción

El método de estimación de máxima verosimilitud es uno de los procedimientos que más se usan en la estadística inferencial que consiste en estimar valores de parámetros desconocidos. En muchas de las distribuciones probabilísticas de uso común, estimar sus parámetros por el método de máxima verosimilitud no resulta complicado. Sin embargo, el estimar el tamaño de la población N en la distribución hipergeométrica no es claro en los textos de estadística inferencial o en varios no es abordado, por ejemplo, Mood & Boes (1974), Bickel & Doksum (2001) y Casella y Berger (2002). Por otro lado, Shao (2003) proporciona el estimador de máxima verosimilitud de N, y de R. Lo que se pretende hacer en esta exposición es mostrar como Hanwen Zhang proporciona una manera adecuada de estimar N y R.

Estimación de N

Supóngase que $X \sim H(n; R; N)$ y se requiere estimar el tamaño de la población N, en este caso, al tomar como parámetro objetivo a N, la función de verosimilitud se denotará por:

$$L(N) = L(x, N) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}} \tag{1}$$

Si $\max(n - N + R, 0) \leq x \leq \min(R, n)$. Ahora como la función $L(N)$ es discreta, no se puede determinar el máximo al tomar derivadas con respecto a N, sino, más bien hay que observar en donde $L(N)$ es creciente y decreciente. Así que, para hacer análisis, se calcula la relación $D(N) = L(N) / L(N-1)$, y se averigua donde $D(N) > 1$ y donde $D(N) < 1$. Esto es:

$$D(N) = \frac{\binom{N-R}{n-x}\binom{N-1}{n}}{\binom{N}{n}\binom{N-R-1}{n-x}} > 1$$

o bien,

$$D(N) = \frac{(N-R)!(N-1)!(N-n)!(N-R-1-n+x)!}{(N-R-1)!N!(N-n-1)!(N-R-n+x)!} > 1$$

Factorizando, nos lleva a $D(N) > 1$ si y sólo si $(N-R)(N-n) > N(N-R-n+x)$. Por lo tanto, los valores de N donde $D(N) > 1$ está dado por $N < Rn/x$. Análogamente, $D(N) < 1$ para $N > Rn/x$.

De acuerdo con Hanwen Zhang (junio 2009) algunos establecen que el estimador de máxima verosimilitud de N es Rn/x. Sin embargo, la afirmación no es correcta, puesto que, en la definición de máxima verosimilitud, la realización del estimador debe caer en el espacio paramétrico. Pero la realización de la estadística Rn/x puede ser fraccional y no califica para ser estimador de máxima verosimilitud de N.

Shao (2003, pág. 277) nos proporciona el estimador de máxima verosimilitud correcto de N, que es $\widehat{N_{Mv}} = [Rn/x]$ donde $[a]$ denota la parte entera de a. Además, al tener en cuenta la propiedad creciente y decreciente de $L(N)$ nos lleva a tener dos candidatos para ser el estimador de máxima verosimilitud de N: $[Rn/x]$ y $[Rn/x]+1$. Para verificar cuando el estimador de máxima verosimilitud es de hecho $[Rn/x]$ recuérdese que $D(N) < 1$ si $N > Rn/x$. Además, se puede ver que $[Rn/x]+1 > Rn/x$, entonces $D([Rn/x]+1) < 1$ por lo que $L([Rn/x]+1)/L([Rn/x]) < 1$. Así se concluye que:

$$L([Rn/x]+1) < L([Rn/x])\,. \qquad (2)$$

Además, se puede concluir que $[Rn/x]$ es el único estimador de máxima verosimilitud ya que en (2) nunca se puede tener la igualdad.

Hanwen Zhang proporciona un ejemplo al suponer $R = 5$, $n = 4$ y $x = 3$, para este caso $[Rn/x] = [6.67] = 6$ que es el estimador de máxima verosimilitud como se ve en la figura.

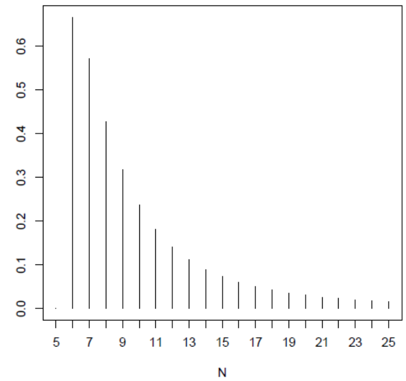

Figura 1.

Estimación de R

Ahora supóngase que el parámetro de interés es R, el número de individuos con alguna característica especifica en una población de tamaño N, y en una muestra de tamaño n, x individuos con esta característica son seleccionados. Nuevamente al considerar la función de verosimilitud como una función de R, al ser éste el parámetro de interés, esto es

$$L(R) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}.$$

Nuevamente se calcula el cociente $\boldsymbol{D(R) = L(R)/L(R+1)}$ y se observa en qué valores de $\boldsymbol{R}$, $\boldsymbol{D(R) > 1}$ y para qué valores, $\boldsymbol{D(R) < 1}$. Así

$$\boldsymbol{D(R) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{R+1}{x}\binom{N-R-1}{n-x}} > 1}.$$

Que es equivalente a

$$\boldsymbol{D(N) = \frac{(R+1-x)(N-R)}{(R+1)(N-R-n+x)}} > 1.$$

Que a su vez es equivalente a $\boldsymbol{(R+1-x)(N-R) > (R+1)(N-R-n+x)}$. Por lo que $\boldsymbol{D(R) > 1}$ si y sólo si $\boldsymbol{R > (x(N+1)-n)/n}$, es decir, $\boldsymbol{L(R)}$ disminuye para enteros mayores que $\boldsymbol{(x(N+1)-n)/n}$ e incrementa para enteros mayores que $\boldsymbol{(x(N+1)-n)/n}$, lo que llevaría a que $\widehat{\boldsymbol{R}_{\boldsymbol{Mv}}} = (\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})-\boldsymbol{n})/\boldsymbol{n}$, pero esto es verdad sólo si $(\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})-\boldsymbol{n})/\boldsymbol{n}$ es un entero y en el caso que $\boldsymbol{D}(\boldsymbol{R}) = \boldsymbol{1}$, entonces

$$\boldsymbol{L}\left(\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}\right) = \boldsymbol{L}\left(\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}+\boldsymbol{1}\right)$$

y el estimador de máxima verosimilitud puede ser $\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}$ o $\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}+\boldsymbol{1}$ y no es único.
Si $\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}$ no es un número entero, los valores de $\boldsymbol{R}$ que maximiza $\boldsymbol{L}(\boldsymbol{R})$ es $\left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}\right]$ o $\left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}\right]+\boldsymbol{1}$ considerando que el estimador de $\boldsymbol{R}$ debe tomar valores en el conjunto de los números enteros. Sin embargo, nuevamente por el argumento anterior, se puede afirmar que

$$\boldsymbol{L}\left(\left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}\right]\right) < \boldsymbol{L}\left(\left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}\right]+\boldsymbol{1}\right)$$

ya que

$$\left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}\right] < \frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}},$$

y en este caso,

$$\widehat{\boldsymbol{R}_{\boldsymbol{Mv}}} = \left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1}-\boldsymbol{n})}{\boldsymbol{n}}\right]+\boldsymbol{1} = \left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})}{\boldsymbol{n}}\right]$$

Resumiendo, se tiene que

$$\widehat{\boldsymbol{R}_{\boldsymbol{Mv}}} = \begin{cases} \frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})}{\boldsymbol{n}}-\boldsymbol{1}, & \boldsymbol{si}\ \frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})}{\boldsymbol{n}}\ \boldsymbol{es\ un\ entero} \\ \frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})}{\boldsymbol{n}}, & \boldsymbol{si}\ \frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})}{\boldsymbol{n}}\ \boldsymbol{es\ un\ entero} \\ \left[\frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})}{\boldsymbol{n}}\right], & \boldsymbol{si}\ \frac{\boldsymbol{x}(\boldsymbol{N}+\boldsymbol{1})}{\boldsymbol{n}}\ \boldsymbol{no\ es\ un\ entero} \end{cases}$$

Hanwen Zhang proporciona unos ejemplos al suponer $N = 19, n = 5$ y $x = 2$, para este caso $(x(N+1))/n = 8$ que es un entero, por lo que el estimador de máxima verosimilitud de R puede ser 7 u 8 como se ve en la figura, es decir, el estimador de máxima verosimilitud para R no es único en este caso.
Ahora al suponer que $N = 20, n = 15$ y $x = 2$, entonces $(x(N+1))/n = 8.4$, así que el estimador de máxima verosimilitud de R es $[8.4] = 8$ como se ve en la figura.

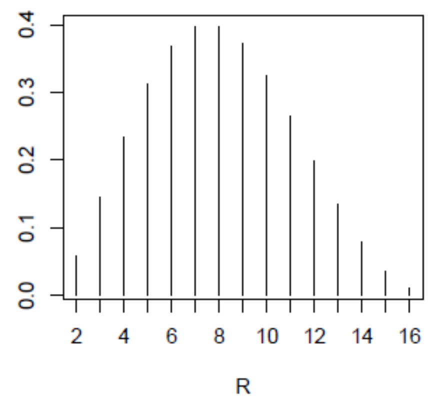

Figura 2.

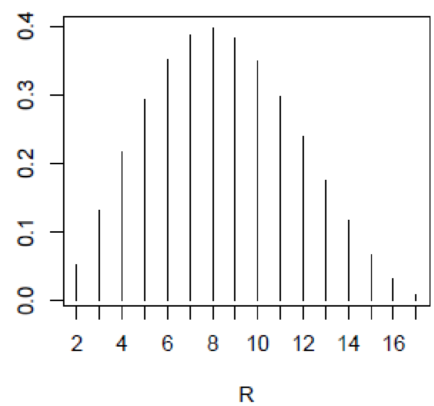

Figura 3.

Conclusiones

Lo que se pretendió con esta exposición fue presentar las estimaciones de máxima verosimilitud en la distribución hipergeométrica y observar que dado que los parámetros manejados $\boldsymbol{N}$ y $\boldsymbol{R}$ son números enteros se debe hacer un análisis alternativo al método de la segunda derivada, para determinar en qué valores de $\boldsymbol{N}$ y $\boldsymbol{R}$ la función de verosimilitud alcanza su máximo, puesto que la función de verosimilitud de la hipergeométrica es discreta en las variables $\boldsymbol{N}$ y $\boldsymbol{R}$.

Referencias

[1.] Bickel, P. & Doksum, K. (2001), **Mathematical Statistics. Basic Ideas and Selected Tipics**., Vol. I, second edn, Prentice-Hall.
[2.] Casella, G. & Berger, R. (2002), **Statistical Inference**., second edn, Duxbury Pres.
[3.] Mood, A.M, G. F. & Boes, D. (1974), **Introduction to the theory of statistics**., International edition. McGraw Hill.
[4.] Shao, J. (2003), **Mathematical statistics**., second edn, Springer.
[5.] Zhang Hanwen (2009) **A Note About Maximum Likelihood Estimator in Hypergeometric Distribution,** Comunicaciones en Estadística Junio 2009, Vol. 2, No. 1