Capítulo 9 Propiedades de los estimadores puntuales y métodos de estimación

9.2 Eficiencia relativa

En el capítulo anterior se vieron distintos tipos de estimadores y se evaluaron algunas propiedades que estos pueden tener. En este capítulo se hará un análisis más forma de algunas propiedades matemáticas de estimadores puntuales. Primero se explorara la eficiencia relativa. En la sección 8.3 se vio como un estimador con varianza pequeña es mejor que uno con varianza grande. Esto se puede decir con distintas palabras, si se tienen dos estimadores insesgados $\hat{\theta}_1$ y $\hat{\theta}_2$, se dice que $\hat{\theta}_1$ es *relativamente más eficiente que* $\hat{\theta}_2$ si $V(\hat{\theta}_2) > V(\hat{\theta}_1)$. Se puede usar la razón de varianzas para definir la *eficiencia relativa* de dos estimadores insesgados.

Definición 9.1

Dados dos estimadores insesgados $\hat{\theta}_1$ y $\hat{\theta}_2$ de un parámetro θ, con varianzas $V(\hat{\theta}_1)$ y $V(\hat{\theta}_2)$, respectivamente, entonces la *eficiencia* $\hat{\theta}_1$ con respecto a $\hat{\theta}_2$, denotada $eff(\hat{\theta}_1, \hat{\theta}_2)$, se define como la razón

$$eff(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)}$$

Entonces se puede decir que si $\hat{\theta}_1$ y $\hat{\theta}_2$ son estimadores insesgados y $eff(\hat{\theta}_1, \hat{\theta}_2) > 1$ entonces significa que la varianza del segundo estimador es mayor que la varianza del primer estimador y por lo tanto el primer estimador $\hat{\theta}_2$ es un mejor estimador insesgado que $\hat{\theta}_2$. De la misma manera si lo opuesto sucede entonces la varianza del segundo estimador es menor y por lo tanto este estimador es el mejor.

9.3 Consistencia

El siguiente temas que se revisara es el de *consistencia*, cuando se quiere estimar un parámetro objetivo de una población hay algunas características que necesita tener el estimador usado. Una de ellas es que existe una convergencia entre el estimador y el parámetro objetivo. En otras palabras se quiere que la distancia entre el estimador y el parámetro objetivo, $|\hat{\theta} - \theta|$, sea menor que algún número real ε positivo arbitrario. Se puede expresar esta cercanía en término probabilísticos, intuitivamente si aumenta el tamaño de la muestra se esperaría que la probabilidad de que esta cercanía sea lo más pequeña posible aumente.

$$P\left(\left|\hat{\theta} - \theta\right| \leq \varepsilon\right)$$

Si esta probabilidad de hecho tiende a 1 cuando $n \to \infty$, entonces se dice que $\hat{\theta}$ es un *estimador consistente* de θ o que $\hat{\theta}$ converge en probabilidad en θ. Con esto se llega a la siguiente definición.

Definición 9.2

Se dice que el $\hat{\theta}_n$ es un *estimador consistente* de θ si, para cualquier número positivo ε,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1$$

o bien de forma equivalente,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

La notación $\hat{\theta}_n$ expresa que el estimador para θ se calcula usando una muestra de tamaño n.

Teorema 9.1

Un estimador insesgado $\hat{\theta}_n$ para θ es un estimador consistente de θ si

$$\lim_{n\to\infty} V(\hat{\theta}_n) = 0$$

Demostración

Si Y es cualquier variable aleatoria con $E(Y) = \mu$ y $V(Y) = \sigma^2 < \infty$ y si k es cualquier constante no negativa, el teorema de Tchebysheff (véase el Teorema 4.13) implica que

$$P(|Y - \mu| > k\sigma) \leq \frac{1}{k^2}$$

Sea n cualquier tamaño muestral fijo. Para cualquier número positivo ε,

$$k = \frac{\varepsilon}{\sigma_{\hat{\theta}_n}}$$

es un número positivo. La aplicación del teorema de Tchebysheff para esta n fijo y esta selección de k muestra que

$$P(|\hat{\theta}_n - \theta| > \varepsilon) \leq \frac{1}{\left(\varepsilon/\sigma_{\hat{\theta}_n}\right)^2}$$

$$P\left(|\hat{\theta}_n - \theta| > \left[\frac{\varepsilon}{\sigma_{\hat{\theta}_n}}\right]\sigma_{\hat{\theta}_n}\right) \leq \frac{V(\hat{\theta}_n)}{\varepsilon^2}$$

Demostración

Entonces, para cualquier n fija,

$$0 \leq P(|\hat{\theta}_n - \theta| > \varepsilon) \leq \frac{V(\hat{\theta}_n)}{\varepsilon^2}$$

Si $\lim_{n\to\infty} V(\hat{\theta}_n) = 0$ y tomamos el límite cuando $n \to \infty$ de la sucesión de probabilidades anterior,

$$\lim_{n\to\infty}(0) \leq \lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) \leq \lim_{n\to\infty} \frac{V(\hat{\theta}_n)}{\varepsilon^2} = 0$$

Entonces, $\hat{\theta}_n$ es un estimador consistente para θ.

Teorema 9.2

Suponga que $\hat{\theta}_n$ converge en probabilidad en θ y que $\hat{\theta}'_n$ converge en probabilidad en θ'.

- **a** $\hat{\theta}_n + \hat{\theta}'_n$ converge en probabilidad en $\theta + \theta'$.
- **b** $\hat{\theta}_n \times \hat{\theta}'_n$ converge en probabilidad en $\theta \times \theta'$.
- **c** Si $\theta' \neq 0$, $\hat{\theta}_n / \hat{\theta}'_n$ converge en probabilidad en θ/θ'.
- **d** Si $g(\cdot)$ es una función de valor real que es continua en θ, entonces $g(\hat{\theta}_n)$ converge en probabilidad en $g(\theta)$.

Teorema 9.3

Suponga que U_n tiene una función de distribución que converge en una función de distribución normal estándar cuando $n \to \infty$. Si W_n converge en probabilidad en 1, entonces la función de distribución de U_n/W_n converge en una función de distribución normal estándar.

9.4 Suficiencia

En esta sección se vera el tema de suficiencia, es importante saber si los estadísticos que se están usando resumen toda la información de una muestra acerca de un parámetro objetivo. Se dice que estos estadísticos tienen la propiedad de suficiencia. Para ilustrar esta noción de un estadístico suficiente se realizara el siguiente ejemplo.

Ejemplo

Consideremos los resultados de n intentos de un experimento binomial, $X_1, X_2, \ldots, X_n$, donde

$$X_i = \begin{cases} 1, & \text{si el } i\text{-ésimo intento es un éxito} \\ 0, & \text{si el } i\text{-ésimo intento es un fracaso} \end{cases}$$

Si p es la probabilidad de éxito en cualquier intento, entonces, para $i = 1, 2, \ldots, n$,

$$X_i = \begin{cases} 1, & \text{con probabilidad } p, \\ 0, & \text{con probabilidad } q = 1 - p \end{cases}$$

Suponga que nos dan un valor de $Y = \sum_{i=1}^{n} X_i$, el número de éxitos entre los n intentos. Entonces nos podemos preguntar si Y resume toda la información de el parámetro p o si hay otras funciones de $X_1, X_2, \ldots, X_n$ que proporcionen más información. Para responder se puede ver la distribución condicional de $X_1, X_2, \ldots, X_n$, dada Y la cual es,

$$P(X_1 = x_1, \ldots, X_n = x_n | Y = y) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, Y = y)}{P(Y = y)}$$

El numerador del lado derecho de esta expresión es 0 si $\sum_{i=1}^{n} x_i \neq y$, y es la probabilidad de una sucesión independiente de número 0 y 1 con un total de y números 1 y $(n-y)$ números 0 si $\sum_{i=1}^{n} x_i = y$. De la misma manera, el denominador es la probabilidad binomial exactamente y éxitos en n intentos. Por lo tanto, si $y = 0, 1, 2, \ldots, n$,

$$P(X_1 = x_1, \ldots, X_n = x_n | Y = y) = \begin{cases} \frac{p^y(1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \frac{1}{\binom{n}{y}}, & \text{si } \sum_{i=1}^{n} x_i = y \\ 0, & \text{en cualquier otro punto} \end{cases}$$

Es importante observar que la distribución de X_1,X_2,...,X_n, dada Y, n depende de p. Esto nos dice que una vez que se conozca Y no habrá otra función de las variables X_i que proporcione más información sobre el posible valor de p. Como consecuencia se dice que el estadístico Y es *suficiente* para p.

Definición 9.3

Sea $Y_1, Y_2, \ldots, Y_n$ una muestra aleatoria de una distribución de probabilidad con parámetro desconocido θ. Entonces se dice que el estadístico $U = g(Y_1, \ldots, Y_n)$ es *suficiente* para θ si la distribución condicional de $Y_1, Y_2, \ldots, Y_n$ dada U, no depende de θ.

Se ha visto en capítulos anteriores distintas funciones de distribución de varias variables aleatorias. A menudo la distribución, asociada con una variable aleatoria Y, depende del valor de un parámetro θ. Entonces se denota de la siguiente manera para el caso discreto, $p(y|\theta)$ y $f(y|\theta)$ para el caso continuo.

Definición 9.4

Sean $y_1, y_2, \ldots, y_n$ observaciones muestrales tomadas de variables aleatorias correspondientes $Y_1, Y_2, \ldots, Y_n$ cuya distribución depende de un parámetro θ. Entonces, si $Y_1, Y_2, \ldots, Y_n$ son variables aleatorias discretas, la *verosimilitud de la muestra*, $L(y_1, y_2, \ldots, y_n|\theta)$, se define como la probabilidad conjunta de $y_1, y_2, \ldots, y_n$.
Si $Y_1, Y_2, \ldots, Y_2$ son variables aleatorias continuas, la verosimilitud $L(y_1, y_2, \ldots, y_n|\theta)$ se define como la densidad conjunta evaluada en $y_1, y_2, \ldots, y_n$.

Otra forma de decir esto es que la función da la probabilidad o *verosimilitud* de observar el evento $(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n)$ cuando el valor del parámetro es θ. Además, si el conjunto de variables aleatorias $Y_1, Y_2, \ldots, Y_n$ denota una muestra aleatoria de una distribución discreta con función de probabilidad $p(y|\theta)$, entonces

$$L(y_1, \ldots, y_n|\theta) = p(y_1, \ldots, y_n|\theta) = p(y_1|\theta) \times \cdots \times p(y_n|\theta)$$

mientras que si $Y_1, Y_2, \ldots, Y_n$ tienen una distribución continua con función de densidad $f(y|\theta)$, entonces

$$L(y_1, \ldots, y_n|\theta) = f(y_1, \ldots, y_n|\theta) = f(y_1|\theta) \times \cdots \times f(y_n|\theta)$$

Teorema 9.4

Sea U un estadístico basado en la muestra aleatoria $Y_1, Y_2, \ldots, Y_n$. Entonces U es un *estadístico suficiente* para la estimación de un parámetro θ si y sólo si la verosimilitud $L(\theta) = L(y_1, y_2, \ldots, y_n|\theta)$ se puede factorizar en dos funciones no negativas,

$$L(y_1, y_2, \ldots, y_n|\theta) = g(u, \theta) \times h(y_1, y_2, \ldots, y_n)$$

donde $g(u, \theta)$ es una función sólo de u y θ y $h(y_1, y_2, \ldots, y_n)$ no es una función de θ. Este teorema también se conoce como el *criterio de factorización*

9.5 Teorema de Rao- Blackwell y estimación insesgada de varianza mínima

Los estimadores suficientes desempeñas un papel importante para determinar buenos estimadores para parámetros. Si $\hat{\theta}$ es un estimador insesgado para θ y si U es un estadístico suficiente para θ, entonces hay una función de U que también es un estimador insesgado para θ y tiene una varianza *no mayor* que $\hat{\theta}$. La base teórica de esto se muestra en el siguiente teorema.

Teorema 9.5

El Teorema de Rao- Blackwell
Sea $\hat{\theta}$ un estimador insesgado para θ tal que $V(\hat{\theta}) < \infty$. Si U es un estadístico suficiente para θ, definamos $\hat{\theta}^* = E(\hat{\theta}|U)$. Entonces, para toda θ,

$$E(\hat{\theta}^*) = \theta \qquad \text{y} \qquad V(\hat{\theta}^*) \leq V(\hat{\theta})$$

Demostración

Como U es suficiente para θ, la distribución condicional de cualquier estadístico (incluyendo $\hat{\theta}$), dada U, no depende de θ. Entonces, $\hat{\theta}^* = E(\hat{\theta}|U)$ no es una función de θ y es por lo tanto un estadístico.
Recuerde los Teoremas 5.14 y 5.15, donde consideramos la forma de hallar medias y varianzas de variables aleatorias con el uso de medias y varianzas condicionales. Como $\hat{\theta}$ es un estimador insesgado para θ, el Teorema 5.14 implica que

$$E(\hat{\theta}^*) = E[E(\hat{\theta}|U)] = E(\hat{\theta}) = \theta$$

Entonces, $\hat{\theta}^*$ es un estimador insesgado para θ.
El Teorema 5.15 implica que

$$V(\hat{\theta}) = V[E(\hat{\theta}|U)] + E[V(\hat{\theta}|U)] = V(\hat{\theta}^*) + E[V(\hat{\theta}|U)]$$

Como $V(\hat{\theta}|U = u) \geq 0$ para toda u, se deduce que $E[V(\hat{\theta}|U)] \geq 0$ y por lo tanto que $V(\hat{\theta}) \geq V(\hat{\theta}^*)$, como dijimos.

Este teorema implica que un estimador insesgado para θ con varianza pequeña es, o puede hacerse que sea, una función de un estadístico suficiente. Para las distribuciones que se han estudiado, el *criterio de factorización* de manera típica identifica un estadístico U que mejor resume la información de los datos acerca del parámetro θ. Tales estadísticos reciben el nombre de *estadísticos suficientes mínimos*
Ahora si empezamos con un estimador insesgado para un parámetro θ y el estadístico suficiente obtenido por medio del criterio de factorización y aplicamos el teorema de Rao-Blackwell en general lleva a un *estimador insesgado de varianza mínima* o MVUE

(Minimum Variance Unbiased Estimator) por sus siglas en ingles.
No obstante, si U es el estadístico suficiente que mejor resume los datos y alguna función de U, por ejemplo $h(U)$, se puede hallar de modo que $E[h(u)] = \theta$, se deduce que $h(U)$ es el MVUE para θ.

9.6 Método de momentos

El método de momentos es un procedimiento muy sencillo para hallar un estimador para uno o más parámetros poblacionales. Recuerde que el k-ésimo momento de una variable aleatoria tomado alrededor del origen, es

$$\mu'_k = E(Y^k)$$

El correspondiente k-ésimo momento muestral es el promedio

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

El método de momentos está basado en la idea de que los momentos muestrales deben dar buenas estimaciones de los momentos poblaciones correspondientes. Es decir, m'_k debe ser un buen estimador de μ'_k. Entonces el método se puede expresar como sigue.

Método de momentos
Escoja como estimaciones los valores de los parámetros que son soluciones de las ecuaciones $\mu'_k = m'_k$, para $k = 1, 2, \ldots, t$, donde t es el número de parámetros por estimar.

Ejemplo

Sea una muestra aleatoria de n observaciones, Y_1,Y_2,. . .,Y_n que se selecciona de una población en la que Y_i, para $i = 1, \ldots, n$, posee una función de densidad de probabilidad uniforma en el intervalo $(0, \theta)$ donde θ es desconocida. Usando el método de moementos se necesita encontrar el primer momento de una distribución uniforme, de resultados anteriores se tiene los siguiente.
El valor de μ'_1 para una variable aleatoria uniforme es

$$\mu'_1 = \mu = \frac{\theta}{2}$$

El correspondiente primer momento muestral es

$$m'_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$$

Igualando la población correspondiente y el momento muestral, obtenemos

$$\mu'_1 = \frac{\theta}{2} = \overline{Y} = m'_1$$

El estimador que se obtiene mediante el método de momentos para θ es la solución de la ecuación anterior. Esto es, $\hat{\theta} = 2\overline{Y}$.

El método de momentos permite generar estimadores de parámetros desconocidos de manera sencilla y proporciona estimadores *consistentes*, pero los estimadores obtenidos no son funciones de estadísticos suficientes y por lo tanto frecuentemente no son eficientes y en muchos casos son sesgados.

9.7 Método de máxima verosimilitud

En la sección 9.5 se vio un método para obtener un estimador insesgado de varianza mínima por un parámetro objetivo; usando el criterio de factorización junto con el teorema de Rao-Blackwell. Sin embargo este método requiere que encontremos alguna función de un estadístico suficiente mínimo que es un estimador insesgado para el parámetro objetivo. Esto es difícil de hacer por lo que se necesita recurrir a distintos métodos. Como se vio en la sección pasada el método de momentos por lo general no lleva a los mejores estimadores. El método de máxima verosimilitud a diferencia del anterior con frecuencia proporciona estimadores insesgados de varianza mínima.
La técnica, llamada *método de máxima verosimilitud*, logra esto al seleccionar como estimaciones los valores de los parámetros que maximizan la verosimilitud de la muestra observada.

Método de máxima verosimilitud
Suponga que la funición de verosimilitud depende de k parámetros θ_1, θ_2, . . ., θ_k. Escoja como estimaciones los valores de los parámetros que maximicen la verosimilitud $L(y_1, y_2, \ldots, y_n|\theta_1, \theta_2, \ldots, \theta_k)$.

Ejemplo

Suponga que se tiene un experimento binomial con n ensayos que resultaron en las observaciones $y_1, y_2, \ldots, y_n$, donde $y_i = 1$ si el i-ésimo intento fue un éxito y $y_i = 0$ en cualquier otro punto. Para encontrar el estimador de máxima verosimilitud primero necesitamos la verosimilitud.
La verosimilitud de la muestra obsrevada es la probabilidad de observar $y_1, y_2, \ldots y_n$. En consecuencia,

$$L(p) = L(y_1, y_2, \ldots y_n|p) = p^y(1-p)^{n-y}$$

donde $y = \sum_{i=1}^{n} y_i$. Ahora deseamos determinar el valor de p que maximice $L(p)$. Si $y = 0$, $L(p) = (1-p)^n$ y $L(p)$ se maximiza cuando $p = 0$. Análogamente, si $y = n$, $L(p) = p^n$ y $L(p)$ se maximiza cuando $p = 1$. Esto solo elimina dos valores de y aun se necesita saber que sucede cuando $y = 1, 2, \ldots, n-1$, si esto sucede entonces $L(p) = p^y(1-p)^{n-y}$ es cero cuando $p = 0$ y $p = 1$. Además es continua para valores de p entre 0 y 1. Para encontrar p que maximice $L(p)$ para estos valores de y se necesita derivar la verosimilitud e igualar a 0. En ocasiones esta derivada no es fácil de obtener, por lo que se recurre a el logaritmo natural de la función de verosimilitud. Como este, $\ln[L(p)]$ es una función creciente monotónica de $L(p)$ entonces ambas maximizan el mismo valor de p.

Ejemplo

Entonces se puede calcular la derivada del logaritmo de la verosimilitud para facilitar los cálculos. Se tiene lo siguiente.

$$\ln[L(p)] = \ln[p^y(1-p)^{n-y}] = y\ln p + (n-y)\ln(1-p)$$

Si $y = 1, 2, \ldots, n-1$, la derivada de $\ln[L(p)]$ con respecto a p, es

$$\frac{d\ln[L(p)]}{p} = y\left(\frac{1}{p}\right) - (n-y)\left(\frac{1}{1-p}\right) \qquad (1)$$

Para $y = 1, 2, \ldots, n-1$, el valor que maximice (o minimice) $\ln[L(p)]$ es la solución de la ecuación

$$\frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}}$$

Resolviendo, obtenemos la estimación $\hat{p} = y/n$. Entonces el estimador de máxima verosimilitud es $\hat{p} = Y/n$ o la fracción de éxitos en el número total de intentos n.

Si U es *cualquier* estadístico suficiente para la estimación de un parámetro θ, incluyendo el estadístico suficiente obtenido del uso óptimo del criterio de factorización, el estimador de máxima verosimilitud es siempre alguna función de U. Esto es, el EMV depende de las observaciones muestrales sólo mediante el valor de un estadístico suficiente.

Demostración

Para demostrar esto, sólo necesitmaos observar que si U es un estadístico suficiente para θ, el criterio de factorización (Teorema 9.4) implica que la verosimilitud puede ser factorizada como

$$L(\theta) = L(y_1, y_2, \ldots, y_n|\theta) = g(u, \theta)h(y_1, y_2, \ldots, y_n)$$

donde $g(u, \theta)$ es una función de sólo u y θ y $h(y_1, y_2, \ldots y_n)$ no depende de θ. Por lo tanto se deduce que

$$\ln[L(\theta)] = \ln[g(u, \theta)] + \ln[h(y_1, y_2, \ldots y_n)]$$

Observe que $\ln[h(y_1, y_2, \ldots y_n)]$ no depende de θ y por lo tanto maximizar $\ln[L(\theta)]$ con respecto a θ es equivalente a maximizar $\ln[g(u, \theta)]$ con respecto a θ. Como $\ln[g(u, \theta)]$ depende de los datos sólo mediante el valor del estadístico suficiente U, el EMV para θ es siempre una función de U. En consecuencia si un EMV para un parámetro se puede hallar y luego ajustar para ser insesgado, el estimador resultante es con frecuencia un MVUE del parámetro en cuestión.

Otra propiedad intersante de los estimadores de máxima verosimilitud es que tienen la *propiedad de invarianza*. Esto significa que si no

estamos interesados en un parámetro θ y nos interesa más alguna función de este parámetro, por ejemplo $t(\theta)$. Entonces si $t(\theta)$ es una función biunívoca de θ y si $\hat{\theta}$ es un estimador de máxima verosimilitud, entonces el EMV de $t(\theta)$ está dado por

$$\widehat{t(\theta)} = t(\hat{\theta})$$

9.8 Algunas propiedades de los estimadores de máxima verosimilitud con muestras grandes

Los estimadores de máxima verosimilitud también tienen propiedades interesantes cuando se trabaja con muestras grandes. Suponga que se tiene una función derivable de θ, $t(\theta)$. De la sección pasada se obtuvo que los EMV tienen la propiedad de invarianza, entonces el estimador de máxima verosimilitud de $t(\theta)$ esta dado por $t(\hat{\theta})$. Entonces para tamaños de muestra grandes se tiene que

$$Z = \frac{t(\hat{\theta}) - t(\theta)}{\sqrt{\dfrac{\left[\frac{\partial t(\theta)}{\partial \theta}\right]^2}{nE\left[-\frac{\partial^2 \ln f(Y|\theta)}{\partial \theta^2}\right]}}}$$

tiene aproximadamente una distribución normal estándar. En esta expresión, la cantidad $f(Y|\theta)$ del denominador es la función de densidad correspondiente a la distribución continua de interés, evaluada en el valor aleatorio Y. Z entonces tiene el siguiente intervalo de confianza aproximado de muestra grande $100(1-\alpha)\%$ para $t(\theta)$:

$$t(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\frac{\left[\frac{\partial t(\theta)}{\partial \theta}\right]^2}{nE\left[-\frac{\partial^2 \ln f(Y|\theta)}{\partial \theta^2}\right]}}$$

$$\approx t(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\left(\frac{\left[\frac{\partial t(\theta)}{\partial \theta}\right]^2}{nE\left[-\frac{\partial^2 \ln f(Y|\theta)}{\partial \theta^2}\right]}\right)\Bigg|_{\theta=\hat{\theta}}}$$