Notas de Estadística

Autores: Graciela Boente- Víctor Yohai

Contenido

Chapter 2

Introducción a la Inferencia Estadística

2.1 Poblaciones finitas

Frecuentemente en los problemas de las diferentes disciplinas se estudia el comportamiento de varias variables definidas sobre un conjunto de objetos. El conjunto de objetos será denominado *población* y será representado por $\mathcal{P} = \{a_1, a_2, \ldots, a_n\}$; $a_1, a_2, \ldots, a_n$ serán denominados los *elementos* de la población $\mathcal{P}$. Sobre esos elementos se observan variables, indicadas $X_1, X_2, \ldots, X_k$, que son características que cambian de individuo a individuo. Luego para cada elemento a en $\mathcal{P}$, estará definido $X_1(a), X_2(a), \ldots, X_k(a)$.

Ejemplo 1: Consideremos una población $\mathcal{P}$ formada por un conjunto de 1000 parcelas que constituyen una explotación agrícola y donde se cultiva solamente trigo. Sea $X(a)$ la cosecha en la parcela a durante un determinado año medida en kilogramos.

Ejemplo 2: Consideremos el conjunto $\mathcal{P}$ de votantes en una determinada elección donde se presentan 3 candidatos, que denominamos 1, 2 y 3. Definimos $X(a)$ como el número del candidato votado por a.

Ejemplo 3: Supongamos que la población $\mathcal{P}$ consiste de todos los pájaros de una especie determinada que habitan en una región determinada. Para

cada pájaro se define $X(a)$ como el largo del pájaro y $Y(a)$ el área de las alas.

Distribución de una variable en la población. Llamaremos *distribución de una variable X en la población $\mathcal{P}$* a la distribución que se obtiene cuando se elige al azar un elemento de la población, es decir, cuando se le da a todo elemento de la población la misma probabilidad. Luego se tiene

$$F_X(x) = \frac{\#\{a \in \mathcal{P}, \quad X(a) \leq x\}}{\#\mathcal{P}}$$

donde $\#A$ indica el número de elementos de A. Del mismo modo se define distribución conjunta de dos o más variables en la población $\mathcal{P}$. Luego si X e Y son variables definidas sobre la población $\mathcal{P}$ será

$$F_{XY}(x,y) = \frac{\#\{a \in \mathcal{P} \ : \ X(a) \leq x, \ Y(a) \leq y\}}{\#\mathcal{P}}$$

Obsérvese que la distribución de una variable definida en una población finita es necesariamente discreta, ya que la variable correspondiente toma sólo un número finito de valores.

2.2 Poblaciones infinitas

En muchos problemas interesa la distribución de una variable aleatoria X (o de varias variables $X_1, X_2, \ldots, X_k$) que se observan cada vez que se repite un mismo experimento perfectamente definido. En estos casos, cada elemento a estudiar corresponde al resultado de un experimento, pero no existe un conjunto finito fijo de experimentos definido de antemano, ya que al menos teóricamente se puede repetir el experimento tantas veces como se quiera. Se puede pensar entonces en una *población infinita* compuesta por los infinitos posibles experimentos que teóricamente se pueden realizar, aunque tal población no tiene existencia real.

Ejemplo 1: El experimento consiste en tirar una moneda y X vale 0 ó 1 según caiga ceca o cara.

Ejemplo 2: El experimento consiste en repartir 10 cartas elegidas al azar de un mazo de 52. X es el número de corazones, e Y el número de sietes.

Ejemplo 3: El experimento consiste en fabricar y probar una lámpara; X es el tiempo de duración de la misma.

Ejemplo 4: Se desea medir una magnitud física, cuyo valor verdadero μ es desconocido. Cada medición está afectada de un error aleatorio. Luego lo que se observa al hacer una medición es una variable $X = \mu + \varepsilon$, donde ε es el error. La medición se puede repetir tantas veces como se quiera.

Lo que hace que una población sea infinita es que el experimento pueda repetirse infinitas veces y no el número de posibles resultados que puede ser finito como puede verse en los ejemplos 1 y 2.

Distribución de una variable en una población infinita. En el caso de *población infinita* se puede suponer que cada vez que se repite el experimento se observa una variable aleatoria X (o varias variables $X_1, X_2, \ldots, X_k$) con una cierta distribución $F(x)$ (o distribución conjunta $F(x_1, x_2, \ldots, x_k)$), y que a diferentes experimentos corresponden variables aleatorias independientes. De acuerdo a la ley de los grandes números, $F(x)$ puede verse como el límite en casi todo punto de la distribución empírica asociada a n repeticiones independientes del experimento. Es decir, si se realiza una sucesión de experimentos y los valores observados son $x_1, x_2, \ldots, x_n, \ldots$, entonces si $F_n(x) = \#\{x_i : x_i \leq x,\ 1 \leq i \leq n\} \ / \ n$ se tendrá $F_n(x) \to F(x)$ en c.t.p. La distribución $F(x)$ será denominada *distribución de la variable X en la población infinita considerada.*

2.3 Modelos para la distribución de una variable en una población

Tanto en el caso de poblaciones finitas como en el de poblaciones infinitas, la distribución F puede ser muy complicada e irregular. Sin embargo, frecuentemente puede ser aproximada por una distribución de forma relativamente sencilla. Consideremos el ejemplo 1 de 2.1. Como la población es finita, la distribución real de X es discreta. Sin embargo, como el número de parcelas es muy grande, 1000, y como es muy probable que los valores $X(a_i)$ sean todos diferentes (pueden diferir muy poco, pero es muy difícil que haya 2 exactamente iguales), resulta que la probabilidad de cada uno de los valores es muy pequeña (1/1000). Por lo tanto, se puede pensar que la distribución real puede aproximarse por una distribución continua de forma

sencilla, por ejemplo una distribución normal. Esto sugiere la introducción del concepto de *modelo.*

Llamaremos *modelo de la distribución de una variable en una población* a un conjunto de hipótesis que se suponen válidas para la distribución de una variable en una población. Más formalmente, supongamos que la variable tiene distribución F perteneciente a una familia $\mathcal{F}$. Al fijar el modelo, se establecen hipótesis sobre la familia $\mathcal{F}$ que, en general, se cumplirán en forma aproximada. La *bondad* de un modelo para describir la distribución de una población estará dada por el grado de aproximación que tengan las hipótesis del modelo con la distribución real.

Por lo tanto, de acuerdo a lo que dijimos anteriormente, se podría usar un modelo continuo para la distribución de variables en poblaciones finitas.

Clasificaremos los modelos en *paramétricos* y *no paramétricos.*

Modelos paramétricos: Consisten en suponer que la distribución $F(x)$ de la variable en la población pertenece a una familia de distribuciones que depende de un número finito de parámetros reales. Así, ejemplos de modelos paramétricos son los siguientes:

(a) $F(x)$ pertenece a la familia $N(\mu, \sigma^2)$,

(b) $F(x)$ pertenece a la familia $B_i(\theta, n)$,

(c) $F(x)$ pertenece a la familia $P(\lambda)$,

(d) $F(x)$ pertenece a la familia $\varepsilon(\lambda)$,

(e) Si $F(x, y)$ es la distribución de dos variables, un modelo puede ser $F(x, y)$ pertenece a la familia $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$,

(f) Si $F(x_1, x_2, \ldots, x_k)$ es la distribución de k variables un modelo puede ser $F(x_1, \ldots, x_k)$ pertenece a la familia $M(\theta_1, \theta_2, \ldots, \theta_k, n)$.

En general, un modelo paramétrico tendrá la siguiente forma. Si $F(x)$ es la distribución de una variable X, entonces $F(x)$ pertenece a la familia $\mathcal{F} = \{F(x, \theta_1, \theta_2, \ldots, \theta_k) \quad \boldsymbol{\theta} \in \Theta\}$, donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$ es el vector de parámetros que toma valores en un conjunto $\Theta \subset R^k$. Esto significa que existe algún valor $\boldsymbol{\theta} \in \Theta$, digamos $\boldsymbol{\theta}_0$ tal que $F(x, \boldsymbol{\theta}_0)$ coincide con la distribución $F(x)$ (aunque en la realidad no coincidirá, sino que resultará parecida).

Ejemplo 1: Para el ejemplo 1 de 2.1, podemos usar el modelo definido por la familia de distribuciones $N(\mu, \sigma^2)$.

Ejemplo 2: Para el ejemplo 2 de 2.1 , podemos usar el modelo $M(\theta_1, \theta_2, \theta_3, 1)$. En este caso, el modelo será exacto con

$$\theta_i = \frac{\#\{a \in P; \quad X(a) = i\}}{\#P}, \quad i = 1, 2, 3.$$

Ejemplo 3: Para el ejemplo 3 de 2.1, podemos usar para la distribución $F(x, y)$ el modelo $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

Ejemplo 4: Para el ejemplo 3 de 2.2 podemos usar el modelo $\varepsilon(\lambda)$.

Ejemplo 5: Para el ejemplo 4 de 2.2 se puede usar el modelo $N(\mu, \sigma^2)$.

Modelos no paramétricos: En los modelos no paramétricos se supone que la distribución $F(x)$ de la variable (o de las variables si hay más de una) en la población, pertenece a una familia $\mathcal{F}$, pero esta familia no puede ser indicada con un número finito de parámetros reales.

Ejemplo 6: Consideremos nuevamente el ejemplo 4 de 2.2. Un modelo no paramétrico razonable sería el siguiente. Sea μ el valor verdadero que se quiere medir, luego la distribución de X (el valor observado en una medición pertenece a la familia $\mathcal{F}$ de todas las distribuciones tales que:

(i) Son continuas con densidad $f(x)$,

(ii) $f(\mu + x) = f(\mu - x)$ es decir son simétricas alrededor del verdadero valor μ, por lo tanto la "probabilidad" de un error positivo es la misma que de uno de igual valor absoluto pero negativo.

(iii) Si $\mu > x > x'$, entonces $f(x') < f(x) < f(\mu)$. Es decir, a medida que se alejan del verdadero valor los posibles resultados tiene menor "probabilidad".

Esta familia de distribuciones $\mathcal{F}$ descripta por (i), (ii) y (iii) no puede ser indicada por un número finito de parámetros.

Ventajas relativas de los modelos paramétricos y no paramétricos

La ventaja fundamental de los modelos paramétricos, consiste en que la distribución que se elige para representar a la distribución de la variable en la población puede ser descripta por un número finito de parámetros. Esto permite inclusive la posibilidad de tabulación. Por ejemplo en el caso de la familia $N(\mu, \sigma^2)$ basta tabular la distribución $N(0, 1)$. Para obtener otra distribución de la familia basta con realizar una transformación lineal. En el caso de la familia $P(\lambda)$ basta tabularla para algunos valores de λ. Por ejemplo, para valores de λ escalonados de 0.1 en 0.1. Para otros valores de λ, la distribución se puede obtener por interpolación.

Además, como la descripción del modelo tiene una formulación analítica relativamente simple, su tratamiento matemático es más sencillo y las conclusiones a las que se pueden arribar más fuertes.

Los modelos no paramétricos carecen de estas ventajas, pero en recompensa tienen mucha mayor flexibilidad. Esto se debe a que la familia de posibles distribuciones para la población es más numerosa y por lo tanto mayor es la posibilidad que haya en esta familia una distribución muy próxima a la real.

Por ejemplo, en el caso del ejemplo 6 de 2.3 μ ya no representa el valor esperado de la variable X, que podría no existir. Por lo tanto, su valor aproximado no podría conocerse promediando los valores observados como en el caso paramétrico, en el que se supone, por ejemplo, que X tiene distribución $N(\mu, \sigma^2)$.

Elección del modelo: La elección del modelo puede ser hecha en base a consideraciones teóricas, o porque la experiencia indica que ajusta bien. Por ejemplo, si F es la distribución del tiempo de espera hasta que un determinado mecanismo falle, y por consideraciones teóricas podemos suponer que el mecanismo tiene "falta de desgaste", podemos suponer como modelo para F la familia exponencial $\varepsilon(\lambda)$. En otros problemas puede suceder que no se pueda elegir el modelo en base a consideraciones teóricas, pero si la experiencia indica a través de estudios anteriores, por ejemplo, que puede ser bien aproximada por una distribución normal, entonces se usaría como modelo la familia $N(\mu, \sigma^2)$.

Veremos en el transcurso del curso, métodos para poner a prueba el modelo elegido, es decir métodos para determinar si el modelo elegido puede describir dentro de una aproximación aceptable la distribución de la variable (o variables) en la población. Esto se hará en el capítulo 6.

2.4 Muestra de una distribución. Inferencia estadística

Supongamos que hemos definido un modelo para la distribución F de una variable en una población, y para fijar ideas supongamos que hemos elegido un modelo paramétrico $F(x, \boldsymbol{\theta})$ con $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta$, donde $\Theta \in R^k$. En general, va a interesar saber sobre F algo más que el hecho de pertenecer a la familia $F(x, \boldsymbol{\theta})$. Puede interesar conocer totalmente la distribución, es decir, el valor de $\boldsymbol{\theta}$, o algunas características de la misma.

Ejemplo 1: Volvamos al ejemplo 1 de 2.1 y supongamos que hemos elegido para la distribución de X en la población la familia $N(\mu, \sigma^2)$. Consideremos tres problemas diferentes.

(a) Interesa conocer la distribución F completamente. En este caso hace falta conocer los valores de ambos parámetros, μ y σ^2.

(b) Se requiere sólo el conocimiento de la producción total. Como hay 1000 parcelas la producción total sería 1000 μ y por lo tanto bastaría con conocer μ.

(c) Se ha fijado una meta de producir al menos 200 toneladas de trigo y lo único que interesa es saber si se cumple o no la meta. Luego en este caso lo único que interesa es saber si $\mu < 200$ o $\mu \geq 200$, aunque no interesa el valor exacto de μ.

Volvamos al problema general, la característica numérica que interesa de la distribución puede ser expresada como $q(\theta_1, \theta_2, \ldots, \theta_k)$, donde $q(\theta_1, \theta_2, \ldots, \theta_k)$ es una función de Θ en R si interesa una sola característica numérica, o en R^h si interesan h características. En el ejemplo 1, tendríamos para (a) $q(\mu, \sigma^2) = (\mu, \sigma^2)$; para (b) $q(\mu, \sigma^2) = 1000\mu$ y para (c)

$$q(\mu, \sigma^2) = \begin{cases} 0, & \text{si} \quad \mu < 200 \\ 1, & \text{si} \quad \mu \geq 200 \end{cases} .$$

Así, en este último caso $q(\mu, \sigma^2) = 0$ nos indica que no se cumplió la meta y $q(\mu, \sigma^2) = 1$ indica que se cumplió.

Para conocer el valor de $q(\theta_1, \theta_2, \ldots, \theta_k)$ exactamente, deberíamos conocer el valor de la variable X en toda la población. Así, en el ejemplo 1,

deberíamos conocer la producción de todas las parcelas. Observar el valor de la variable para todos los elementos de la población puede ser muy costoso, o aún imposible, como en el caso de poblaciones infinitas. Inclusive en el caso de poblaciones finitas puede ser imposible si se quiere la información con cierta premura. En el ejemplo 1, si se pueden cosechar sólo 20 parcelas por día, se necesitarían 50 días para conocer cuál es la producción de cada una de las 1000 parcelas. Si se quisiera el primer día de la cosecha hacer una estimación de la producción total, ésta debería hacerse en base a los resultados de las 20 parcelas cosechadas ese día.

Se puede definir la *Estadística* como la ciencia que estudia los procedimientos para determinar el valor de una o varias características $q(\theta_1, \ldots, \theta_k)$ de una distribución de una variable en una población que se supone pertenece a una familia $F(x, \theta_1, \theta_2, \ldots, \theta_k)$ observando sólo unos pocos elementos si se trata de una población finita o realizando unos pocos experimentos en el caso de una población infinita. Al conjunto de estas pocas observaciones en base a las cuales se determinará $q(\theta_1, \theta_2, \ldots, \theta_k)$ se denomina *muestra.* Si el modelo es no paramétrico esta formulación cambiará ligeramente, como se verá más adelante.

Los procedimientos estadísticos pueden clasificarse en dos grandes tipos: procedimientos de diseño y procedimientos de inferencia.

Procedimientos de diseño: Son los procedimientos para elegir las observaciones que componen la muestra, de manera que con pocas observaciones se pueda obtener la mayor información posible sobre $q(\theta_1, \theta_2, \ldots, \theta_k)$.

Procedimientos de inferencia: Son los procedimientos que permiten a partir de la muestra inferir la característica de la distribución de la variable en la población que interesa, es decir $q(\theta_1, \theta_2, \ldots, \theta_k)$.

Para ejemplificar, volvemos nuevamente al Ejemplo 1. En este caso un posible diseño, no necesariamente el óptimo, para la selección de la muestra de 20 observaciones puede ser el siguiente. Se elige la primera parcela al azar. El rendimiento de esta parcela será una variable aleatoria que llamaremos X_1 y que tendrá distribución $N(\mu, \sigma^2)$. La segunda parcela se elige al azar entre todas las que quedan. El rendimiento de esta parcela será una variable aleatoria que llamaremos X_2. Como la población de parcelas es grande (hay 1000 parcelas), la distribución de la variable X prácticamente no se modificará después de la extracción de la primera parcela, por lo tanto a los efectos prácticos, X_2 puede ser considerada como una variable aleatoria

independiente de X_1 y con la misma distribución $N(\mu, \sigma^2)$. Repitiendo este procedimiento tendremos variables aleatorias $X_1, X_2, \ldots, X_{20}$ que podemos considerar independientes y cada una con una distribución $N(\mu, \sigma^2)$. Denominaremos a $X_1, X_2, \ldots, X_{20}$ *muestra aleatoria* de tamaño 20 de la distribución $N(\mu, \sigma^2)$.

En general, se dirá que $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ es una *muestra aleatoria de tamaño n de una distribución* $F(\mathbf{x})$ si $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ son variables aleatorias (o vectores aleatorios) independientes e idénticamente distribuídas con distribución $F(\mathbf{x})$. Es decir si

$$F_{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = F(\mathbf{x}_1)\, F(\mathbf{x}_2) \ldots F(\mathbf{x}_n) \tag{2.1}$$

y en el caso que $F(x)$ sea una distribución discreta o continua con función de frecuencia o de probabilidad p, (2.1) será equivalente a

$$p_{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = p(\mathbf{x}_1)\, p(\mathbf{x}_2) \ldots p(\mathbf{x}_n)$$

En el caso de poblaciones finitas, una muestra aleatoria de tamaño n se obtendrá observando n elementos de la población elegidos al azar. Para que las variables fuesen estrictamente independientes los elementos deberían elegirse uno a uno y ser restituídos en la población antes de elegir el próximo. Sin embargo si el tamaño de la muestra es relativamente pequeño respecto al total de la población, aunque no se haga la restitución las variables observadas serán aproximadamente independientes, y a los fines prácticos podemos considerarla una muestra aleatoria.

En el caso de poblaciones infinitas, la muestra aleatoria se obtendrá simplemente repitiendo el experimento n veces y observando cada vez el vector de variables correspondiente.

Consideremos ahora cómo a partir de la muestra $X_1, X_2, \ldots, X_{20}$ que hemos obtenido, utilizando procedimientos de inferencia resolvemos los problemas (a), (b) y (c) que hemos planteado.

El problema (a) consistía en encontrar aproximadamente la distribución de la variable X en la población, es decir, estimar μ y σ^2.

Definamos $\overline{X}_n = (1/n) \sum_{i=1}^{n} X_i$; luego para estimar μ se puede usar $\overline{X}_{20}$. Es de esperar que $\overline{X}_{20}$ se aproxima a μ ya que de acuerdo a la ley de los grandes números $\lim_{n\to\infty} \overline{X}_n = \mu$ c.t.p.

El procedimiento estadístico para estimar μ a partir de la muestra, es formar el promedio de los valores que la componen; es decir $\overline{X}_{20}$. Esto es un *procedimiento de inferencia estadística*, ya que a partir de una muestra

de 20 observaciones, *inferimos* el valor μ característico de la distribución de la variable en la población.

Similarmente se puede estimar σ^2. Partimos de $\sigma^2 = \text{Var } X_i = E(X_i^2) - (E(X_i))^2$. Dado que $E(X_i^2)$ puede estimarse por $(1/20)\sum_{i=1}^{20} X_i^2$, σ^2 puede estimarse por

$$\widehat{\sigma}_{20}^2 = \frac{1}{20}\sum_{i=1}^{20} X_i^2 - \overline{X}_{20}^2$$

Haciendo manipulaciones algebraicas, se obtiene

$$\widehat{\sigma}_{20}^2 = \frac{1}{20}\sum_{i=1}^{20} (X_i - \overline{X}_{20})^2$$

En general, si se tuviese una muestra aleatoria de tamaño n, σ^2 podría estimarse por

$$\widehat{\sigma}_{n}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X}_{n})^2$$

En el problema (b), cuando se quiere conocer la producción total, es decir $q(\mu, \sigma^2) = 1000\mu$, podemos usar para esta estimación 1000 $\overline{X}_{20}$. Es decir, el procedimiento de inferencia sería el siguiente. Se hace el promedio de las observaciones que componen la muestra, y se lo multiplica por 1000.

En el problema (c), es decir el problema de decidir si $\mu < 200$ o $\mu \geq 200$, el procedimiento de inferencia puede ser el siguiente: se decidirá que $\mu < 200$ si $\overline{X}_{20} < 200$ y se decidirá que $\mu \geq 200$ si $\overline{X}_{20} \geq 200$.

Los problemas (a) y (b) son los que se denominan de *estimación puntual,* mientras que el problema (c) es un problema de *test de hipótesis,* ya que en base a la muestra se desea decidir entre dos opciones y determinar las probabilidades de error. Como veremos más adelante, las dos hipótesis no se considerarán en forma simétrica y se determinará cuál de los dos errores a cometer es más grave, para poder controlar su probabilidad.

Los procedimientos que hemos propuesto no son los únicos posibles, ni necesariamente los mejores; solamente fueron introducidos para ejemplificar la naturaleza de los procedimientos estadísticos. Podemos formular una *primera* generalización de la situación descripta en el Ejemplo 1 diciendo que un *problema de inferencia estadística paramétrica* consistirá en: dada una muestra aleatoria de tamaño n, $X_1, X_2, \ldots, X_n$ de la distribución de una variable en una población de la cual se conoce solamente que pertenece a una familia $\mathcal{F} = \{F(x, \theta_1, \theta_2, \ldots, \theta_k)$ con $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta\}$, donde

$\Theta \subseteq R^k$, se quiere inferir conocimiento de algunas características de esta distribución, definidas por una función $q(\boldsymbol{\theta})$ que va de Θ en R^h, siendo h el número de características en las que se está interesado.

Ejemplo 2: Volvamos al ejemplo 6 de 2.3. Supongamos que se quiere conocer μ. Observemos que si F es la distribución de la variable X, entonces de acuerdo con las hipótesis del modelo para toda $F \in \mathcal{F}$ se tiene que μ es la esperanza correspondiente a la distribución F, si es que esta existe (puede no existir) y también μ es la mediana correspondiente a F (la mediana siempre existe). Luego μ es una cierta función de F, digamos $\mu = q(F)$. Si queremos estimar μ, debemos tomar una muestra aleatoria de F, digamos de tamaño n; $X_1, X_2, \ldots, X_n$. Esto se logrará repitiendo n veces la medición de μ. Consideremos ahora el procedimiento para inferir μ. Si estuviésemos seguros que F tiene esperanza podríamos usar para estimar μ, $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$, ya que de acuerdo a la ley de los grandes números debería converger a $E(X_i) = \mu$. Sin embargo la existencia de esperanza no es una hipótesis que hemos requerido para que $F \in \mathcal{F}$. En caso que F no tenga esperanza, se puede mostrar que $\overline{X}_n$ no converge a μ y por lo tanto no será un buen estimador.

En este caso, podemos usar el siguiente procedimiento: ordenamos las X_i, obteniendo $X^{(1)} < X^{(2)} < X^{(3)} < \cdots < X^{(n)}$, donde $X^{(1)}$ es la menor de las X_i, $X^{(2)}$ la siguiente, hasta llegar a $X^{(n)}$, que sería la mayor de todas. Supongamos que $n = 2p + 1$, luego estimamos μ por $\widehat{\mu} = X^{(p+1)}$, es decir por la observación central. Si $n = 2p$ podemos tomar como $\widehat{\mu} = (X^{(p)} + X^{(p+1)})/2$. Por ejemplo, si tuviésemos 7 mediciones y estas resultasen 6.22; 6.25; 6.1; 6.23; 6.18; 6.15; 6.29, se tendría $X^{(1)} = 6.1$; $X^{(2)} = 6.15$; $X^{(3)} = 6.18$; $X^{(4)} = 6.22$; $X^{(5)} = 6.23$; $X^{(6)} = 6.25$ y $X^{(7)} = 6.29$. Estimaríamos μ por $\widehat{\mu} = X^{(4)} = 6.22$. Se puede mostrar que este procedimiento da resultados razonables para una familia $\mathcal{F}$ como la estudiada.

El ejemplo 2 nos sugiere la siguiente formulación del *problema de inferencia estadística no paramétrica:* Dada una muestra aleatoria de tamaño n, $X_1, \ldots, X_n$ de la distribución F de una variable en una población, y de la cual se sabe solamente que pertenece a una familia $\mathcal{F}$ que no puede ser indicada por un número finito de parámetros reales, interesa conocer algunas características de F expresadas como una función $q(F)$ que va de $\mathcal{F}$ a R^h, siendo h el número de características que interesan.

El siguiente ejemplo nos permitirá formular un tipo de problemas de inferencia estadística más general que el estudiado hasta ahora.

Ejemplo 3: Supongamos que el rendimiento por hectárea de un cierto cultivo depende de la cantidad de fertilizante que se usa y que la relación es de la forma

$$X = aG + b + \varepsilon$$

donde G es la cantidad de fertilizante usado por hectárea, X el rendimiento por hectárea y ε un término aleatorio que tiene en cuenta todos los otros factores que intervienen en la determinación de los rendimientos, a y b son parámetros desconocidos.

Supongamos que se cultivan n parcelas usando respectivamente $G_1, G_2, \ldots, G_n$ cantidad de fertilizante por hectárea y sean los rendimientos respectivos observados $X_1, X_2, \ldots, X_n$. Luego se tendrá:

$$X_i = aG_i + b + \varepsilon_i \qquad\qquad 1 \leq i \leq n$$

Supongamos que las ε_i son variables aleatorias independientes igualmente distribuídas con distribución $N(0, \sigma^2)$, donde σ^2 es desconocido. Los valores $G_1, G_2, \ldots, G_n$ son valores numéricos conocidos (no variables aleatorias).

Luego en este caso las variables aleatorias $X_i, \quad 1 \leq i \leq n$, serán independientes con distribución $N(aG_i + b, \sigma^2)$ y por lo tanto no son igualmente distribuídas. En este caso estamos interesados en conocer los parámetros a y b que establecen la relación entre G y X quizás también en σ^2 que establece la varianza de ε, es decir del término residual.

Estos parámetros deben ser estimados a partir del vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Sin embargo, el vector $\mathbf{X}$ tiene componentes con diferentes distribuciones. Se podrían dar ejemplos donde las variables no sean tampoco independientes.

Esto nos sugiere un concepto más amplio de problema estadístico que los vistos anteriormente.

Un *problema de inferencia estadística paramétrica general* consistirá en: dado un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de cuya *distribución conjunta* se conoce solamente que pertenece a una familia $\mathcal{F} = \{F(x_1, x_2, \ldots, x_n, \theta_1, \theta_2, \ldots, \theta_k) \quad \text{con } \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta \subset R^k\}$, inferir conocimiento sobre una función $q(\boldsymbol{\theta})$ de Θ en R^h.

En el ejemplo 3, $\boldsymbol{\theta} = (a, b, \sigma^2)$ y la densidad correspondiente a la distribución es

$$p(x_1, x_2, \ldots, x_n; a, b, \sigma^2) = \frac{1}{(2\pi)^{n/2}}\, e^{-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(x_i - a\,G_i - b)^2}$$

La función $q(\boldsymbol{\theta})$ dependerá del problema que interesa. Si se quiere conocer la relación entre G y X lo que interesará será $q(\boldsymbol{\theta}) = (a, b)$. Si interesa saber

cuál es el rendimiento promedio cuando se utilizan 200 kg por hectárea, lo que interesará conocer será $q(\boldsymbol{\theta}) = 200\,a + b$. Si interesa saber solamente si el fertilizante tiene un efecto positivo, la función $q(\boldsymbol{\theta})$ estará dada por $q(\boldsymbol{\theta}) = \begin{cases} 0 & \text{si} \quad a \leq 0 \\ 1 & \text{si} \quad a > 0 \end{cases}$.

Un procedimiento de inferencia estadística para este problema se verá en el ejemplo 1 de la sección 3.4. Una teoría general que abarca este problema se verá en el capítulo 7.

De la misma forma se podría formular el concepto de *problema de inferencia estadística no paramétrica general.*

Concepto de estadístico

Supongamos dado un problema de inferencia estadística donde se observa un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ con distribución en la familia $F(x_1, x_2, \ldots, x_n; \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$ y donde se quiera inferir acerca de $q(\boldsymbol{\theta})$. Esta inferencia se tendrá que hacer a partir de $\mathbf{X}$, es decir, por funciones de $\mathbf{X}$. Luego se define como *estadístico* a cualquier función medible que tenga como argumento a $\mathbf{X}$ y que tome valores en un espacio euclideo de dimensión finita. En el ejemplo 1, hemos visto que la estimación de μ y σ^2 se hacía mediante el estadístico

$$\mathbf{T} = r(\mathbf{X}) = \left(\sum_{i=1}^{n} \frac{X_i}{n} \,, \quad \sum_{i=1}^{n} \frac{(X_i - \overline{X}_n)^2}{n} \right)$$

En el ejemplo 3, se usó el estadístico $\mathbf{T} = r(\mathbf{X}) = X^{(p+1)}$.

Hasta ahora, hemos supuesto que el parámetro de existir es fijo. Existe otra aproximación, en la cual, el parámetro es una variable aleatoria. Los procedimientos estadísticos bayesianos suponen que $\boldsymbol{\theta}$ es una variable aleatoria no observable, a valores en un espacio Θ con distribución τ. La distribución *a priori* τ establecida antes de tomar la muestra, se modifica en base a los datos para determinar la distribución *a posteriori,* que resume lo que se puede decir del parámetro $\boldsymbol{\theta}$ en base a las suposiciones hechas y a los datos.

Los métodos estadísticos, que van desde el análisis de datos hasta el análisis bayesiano, permiten sacar en forma creciente conclusiones cada vez más fuertes, pero lo hacen al precio de hipótesis cada vez más exigentes y, por lo tanto, menos verificables.

Chapter 3

Estimación puntual

3.1 Introducción

En este capítulo introduciremos algunos conceptos de la teoría de estimación puntual. Los resultados que se desarrollarán, se aplican al problema de ajustar distribuciones de probabilidad a los datos. Muchas familias de distribuciones, como la normal, $N(\mu, \sigma^2)$, o la Poisson, $P(\lambda)$, dependen de un número finito de parámetros y salvo que éstos se conozcan de antemano, deben ser estimados para conocer aproximadamente la distribución de probabilidad.

Consideremos el siguiente problema de inferencia estadística paramétrica. Supongamos se ha observado un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de cuya distribución sólo se conoce que pertenece a una familia $\mathcal{F} = \{F(x_1, x_2, \ldots, x_n, \boldsymbol{\theta})$ donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta \subset I\!R^p\}$. Supongamos que interese conocer *aproximadamente* $q(\boldsymbol{\theta})$, donde $q(\boldsymbol{\theta})$ es una función de Θ en $I\!R$. La única información que se tiene sobre $\boldsymbol{\theta}$ es el vector $\mathbf{X}$, por lo tanto cualquier estimación que se haga de $\boldsymbol{\theta}$, deberá estar basada en $\mathbf{X}$. Un *estimador puntual* de $q(\boldsymbol{\theta})$ será cualquier estadístico $\delta(\mathbf{X})$ de $I\!R^n$ en $I\!R$.

Un buen estimador $\delta(\mathbf{X})$ deberá tener la propiedad de que cualquiera sea el valor de $\boldsymbol{\theta}$, que es desconocido, la diferencia $\delta(\mathbf{X}) - q(\boldsymbol{\theta})$ sea pequeña. En qué sentido esta diferencia es pequeña será especificado más adelante.

Así en el ejemplo 1 de 2.4 se tenía para el problema (a) necesidad de estimar $q_1(\mu, \sigma^2) = \mu$ y $q_2(\mu, \sigma^2) = \sigma^2$, para el problema (b) se requería estimar $q(\mu, \sigma^2) = 1000\ \mu$. En cambio el problema (c) no era de estimación, ya que lo que se buscaba no era aproximar $q(\mu, \sigma^2)$ que vale 0 ó 1 según $\mu < 200$ ó $\mu \geq 200$, sino decidir si $q(\mu, \sigma^2)$ era 0 ó 1.

También podemos considerar problemas de estimación puntual no paramétrica. En este caso sólo se conoce que el vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ tiene una distribución $F(x_1, x_2, \ldots, x_n)$ perteneciente a una familia $\mathcal{F}$, pero esta familia no puede indicarse con un número finito de parámetros, y quiere estimarse una función $q(F)$ que va de $\mathcal{F}$ en $\mathbb{R}$. El ejemplo 2 de 2.4 es un ejemplo de este tipo.

El ejemplo 3 de 2.4 es otro ejemplo de estimación puntual paramétrica.

Comenzaremos describiendo distintos métodos de estimación que intuitivamente parecen razonables, su justificación queda diferida para más adelante.

3.2 Método de los momentos

Sea $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria de una familia de distribuciones $F(x, \theta)$, donde $\theta \in \Theta \subset \mathbb{R}$, y supongamos que se quiera estimar θ.

Sea g una función de $\mathbb{R}$ en $\mathbb{R}$, luego el método de los momentos estima θ, por el valor $\widehat{\theta} = \delta(\mathbf{X})$ que satisface la ecuación

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) = E_{\widehat{\theta}}(g(X_1)), \tag{3.1}$$

donde $E_\theta(X)$ significa la esperanza de X cuando X tiene la distribución $F(x, \theta)$. La justificación heurística de este método se basa en el hecho que de acuerdo a la ley de los grandes números

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to E_\theta(g(X_1)) \quad \text{c.t.p.}$$

y por lo tanto, si θ puede expresarse como una función continua de $E_\theta(g(X_1))$, se puede esperar que cuando n es grande el valor $\widehat{\theta}$ que satisface la ecuación (3.1) estará cerca de θ.

En general, se toman como funciones g las funciones generadoras de momentos, ya que se supone que los parámetros de la distribución se relacionan con los momentos a través de alguna función continua.

Ejemplo 1: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución de la cual sólo se conoce que está en la familia $N(\mu, 1)$. Usando el método

de los momentos y usando $g(x) = x$ se obtiene

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E_{\widehat{\mu}}(X_1) = \widehat{\mu}\ .$$

Luego $\widehat{\mu} = (1/n)\sum_{i=1}^{n} X_i$ es el estimador de μ resultante.

Ejemplo 2: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $N(0, \sigma^2)$. Usando el método de los momentos con $g(x) = x^2$ se obtiene

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = E_{\widehat{\theta}}(X_1^2) = \widehat{\sigma}^2\ .$$

Luego $\widehat{\sigma}^2 = \delta(X_1, \ldots, X_n) = (1/n)\sum_{i=1}^{n} X_i^2$ es el estimador de σ^2 resultante.

Ejemplo 3: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $P(\lambda)$, usando la función $g_1(x) = x$ se obtiene como estimador de λ

$$\frac{1}{n}\sum_{i=1}^{n} X_i = E_{\widehat{\lambda}}(X_i) = \widehat{\lambda}\ .$$

Luego el estimador de los momentos resultantes usando la función g_1 resulta

$$\widehat{\lambda}_1 = \delta_1(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i\ .$$

También podemos usar la función $g_2(x) = x^2$. Recordando que

$$E_\lambda(X_1^2) = \mathrm{Var}_\lambda(X_1) + (E_\lambda(X_1))^2 = \lambda + \lambda^2\ ,$$

obtenemos

$$\frac{1}{n}\sum_{i=1}^{n} X_1^2 = E_{\widehat{\lambda}}(X_1^2) = \widehat{\lambda} + \widehat{\lambda}^2\ ,$$

y resolviendo esta ecuación de segundo grado el valor resulta

$$\widehat{\lambda} = -\frac{1}{2} \pm \sqrt{\frac{1}{4} + \sum_{i=1}^{n} \frac{X_i^2}{n}}\ .$$

Como el parámetro λ es positivo, la solución que interesa es la positiva. Luego el estimador correspondiente a g_2 vendrá dado por

$$\widehat{\lambda}_2 = \delta_2(X_1, X_2, \ldots, X_n) = -\frac{1}{2} + \sqrt{\frac{1}{4} + \sum_{i=1}^{n} \frac{X_i^2}{n}}$$

Luego observamos que eligiendo distintas funciones g, obtenemos diferentes estimadores. Todavía no estamos en condiciones de comparar uno con otro, por lo que dejamos este punto sin resolver hasta más adelante.

Generalización cuando hay varios parámetros: Supongamos que se tiene una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución perteneciente a la familia $\mathcal{F} = \{F(x, \theta_1, \theta_2, \ldots, \theta_p)$ con $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p) \in \Theta \subset I\!R^p\}$.

Para estimar $\theta_1, \theta_2, \ldots, \theta_p$ por el método de los momentos se procede como sigue: Se consideran k funciones $g_1, g_2, \ldots, g_p$ de $I\!R$ en $I\!R$ y se resuelve el siguiente sistema

$$\frac{1}{n}\sum_{i=1}^{n} g_j(X_i) = E_{\widehat{\boldsymbol{\theta}}}(g_j(X_1)) \qquad j = 1, 2, \ldots, p\ .$$

Ejemplo 4: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Consideremos $g_1(x) = x$ y $g_2(x) = x^2$. Como se tiene

$$E_{\mu,\sigma^2}(g_1(X_1)) = \mu \quad \text{y} \quad E_{\mu,\sigma^2}(g_2(X_1)) = \sigma^2 + \mu^2\ ,$$

para estimar μ y σ^2 se deberá resolver el sistema

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i &= \widehat{\mu} \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \widehat{\mu}^2 + \widehat{\sigma}^2\ . \end{aligned}$$

Luego, se tiene

$$\widehat{\mu} = \delta_1(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

y

$$\widehat{\sigma}^2 = \delta_2(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{\mu})^2$$

que coinciden con los estimadores que habíamos propuesto en el ejemplo 1 de 2.4.

Ejemplo 5: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $\Gamma(\alpha, \lambda)$. Consideremos $g_1(x) = x$ y $g_2(x) = x^2$. Como se tiene

$$E_{\alpha,\lambda}(g_1(X_1)) = \frac{\alpha}{\lambda} \quad \text{y} \quad E_{\alpha,\lambda}(g_2(X_1)) = \frac{\alpha(\alpha+1)}{\lambda^2}\ ,$$

para estimar α y λ se deberá resolver el sistema

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} X_i &= \frac{\widehat{\alpha}}{\widehat{\lambda}} \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= \frac{\widehat{\alpha}(\widehat{\alpha}+1)}{\widehat{\lambda}^2} \ . \end{aligned}$$

Indiquemos por $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ y por $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Entonces, despejando del sistema anterior, los estimadores de los momentos para λ y α resultan ser

$$\widehat{\lambda} = \delta_1(X_1, X_2, \ldots, X_n) = \frac{\overline{X}}{\widehat{\sigma}^2}$$

y

$$\widehat{\alpha} = \delta_2(X_1, X_2, \ldots, X_n) = \frac{\overline{X}^2}{\widehat{\sigma}^2} \ .$$

Estimación de $q(\boldsymbol{\theta})$. Si lo que interesa estimar es una función de $\boldsymbol{\theta}$, $q(\boldsymbol{\theta})$ y esta función es continua, el método de los momentos consistirá en estimar primero $\boldsymbol{\theta}$ por $\widehat{\boldsymbol{\theta}}$ y luego $q(\boldsymbol{\theta})$ se estimará por $q(\widehat{\boldsymbol{\theta}})$. La justificación de esto reside en que si $\widehat{\boldsymbol{\theta}}$ está próximo a $\boldsymbol{\theta}$, entonces como q es continua, $q(\widehat{\boldsymbol{\theta}})$ estará próxima a $q(\boldsymbol{\theta})$.

3.3 Método de máxima verosimilitud

Supongamos que se observa un vector muestra $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ discreto o continuo cuya función de densidad discreta o continua pertenezca a una familia $p(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ y se quiera estimar $\boldsymbol{\theta}$.

En el caso discreto $p(\mathbf{x}, \boldsymbol{\theta})$ representa la probabilidad de observar el vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, cuando el valor del parámetro es $\boldsymbol{\theta}$. Es razonable pensar que si hemos observado el vector $\mathbf{x}$, este tendrá alta probabilidad. Luego se podría estimar $\boldsymbol{\theta}$ como el valor que hace máxima $p(\mathbf{x}, \boldsymbol{\theta})$. Un razonamiento análogo se puede hacer en el caso continuo, recordando que la probabilidad de un hipercubo con centro en $\mathbf{x}$ y de arista Δ, cuando Δ es pequeño tiene probabilidad aproximadamente igual $p(\mathbf{x}, \boldsymbol{\theta})\ \Delta^n$. Esto sugiere la siguiente definición:

Definición 1: Diremos $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ es un estimador de máxima verosimilitud (E.M.V.) de $\boldsymbol{\theta}$, si se cumple

$$p(\mathbf{X}, \widehat{\boldsymbol{\theta}}(\mathbf{X})) = \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{X}, \boldsymbol{\theta})$$

Ejemplo 1: Supongamos que θ puede tomar valores $\theta = 1$ ó $\theta = 0$ y que $p(x, \theta)$ viene dado por

x	θ	
	0	1
0	0.3	0.6
1	0.7	0.4
Σ	1	1

Supongamos que se observe una muestra de tamaño 1 con valor X. Luego el estimador de máxima verosimilitud viene dado por

$$\widehat{\boldsymbol{\theta}}(\mathbf{X}) = \begin{cases} 1 & \text{si } \mathbf{X} = 0 \\ 0 & \text{si } \mathbf{X} = 1 \end{cases}$$

Cómputo del E.M.V.: Supongamos ahora que Θ es un subconjunto abierto de $I\!R^p$, que el soporte de $p(\mathbf{x}, \boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$ y que $p(\mathbf{x}, \boldsymbol{\theta})$ tiene derivadas parciales respecto a todas las componentes θ_i.

Como la función $\ln(\mu)$ (logaritmo natural) es monótona creciente, maximizar $p(\mathbf{x}, \boldsymbol{\theta})$ será equivalente a maximizar $\ln p(\mathbf{x}, \boldsymbol{\theta})$. Luego el E.M.V. $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ debe verificar:

$$\frac{\partial \ln p(\mathbf{X}, \boldsymbol{\theta})}{\partial \theta_i} = 0 \qquad\qquad i = 1, 2, \ldots, p \; . \tag{3.2}$$

Hasta ahora hemos supuesto que $\mathbf{X}$ es un vector con una distribución arbitraria. Supongamos ahora que $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ es una muestra aleatoria de una distribución discreta o continua con densidad $p(\mathbf{x}, \boldsymbol{\theta})$. Luego se tiene

$$p(\mathbf{x}, \boldsymbol{\theta}) = p(x_1, x_2, \ldots, x_n, \boldsymbol{\theta}) = \prod_{j=1}^{n} p(x_j, \boldsymbol{\theta})$$

y bajo las condiciones dadas anteriormente, el sistema de ecuaciones (3.2) se transforma en

$$\sum_{i=1}^{n} \frac{\partial \ln p(x_i, \widehat{\boldsymbol{\theta}})}{\partial \theta_j} = 0 \qquad j = 1, 2, \ldots, p \; . \tag{3.3}$$

Supongamos que indicamos por $\psi_j(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial \ln p(x, \widehat{\boldsymbol{\theta}})}{\partial \theta_j}$, entonces (3.3) puede escribirse como

$$\sum_{i=1}^{n} \psi_j(\mathbf{x}_i, \boldsymbol{\theta}) = 0 \qquad j = 1, 2, \ldots, p \; .$$

Esta ecuación corresponde a la forma general de los denominados $M-$estimadores, que veremos más adelante.

Por supuesto que tanto (3.2) como (3.3) son condiciones necesarias pero no suficientes para que $\boldsymbol{\theta}$ sea un máximo. Para asegurarse que $\widehat{\boldsymbol{\theta}}$ es un máximo deberían verificarse las condiciones de segundo orden respectivas. Además debe verificarse que no se trata de un máximo relativo sino absoluto.

Ejemplo 2: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $Bi(\theta, k)$, con k conocido, luego cada variable X_i tiene función de densidad

$$p(x, \theta) = \left(\begin{array}{c} k \\ x \end{array} \right) \theta^x \, (1 - \theta)^{k-x}$$

y

$$\frac{\partial \ln p(x, \theta)}{\partial \theta} = \frac{x}{\theta} - \frac{k - x}{1 - \theta} = \frac{x - k\theta}{\theta(1 - \theta)} \; .$$

Luego (3.3) se transforma en la ecuación

$$\sum_{i=1}^{n} \frac{X_i - k\widehat{\theta}}{\widehat{\theta}(1 - \widehat{\theta})} = 0 \; ,$$

y despejando $\widehat{\theta}$ resulta

$$\widehat{\theta}(X_1, X_2, \ldots, X_n) = \frac{1}{nk} \sum_{i=1}^{n} X_i \; .$$

Ejemplo 3: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Busquemos los E.M.V. de μ y σ^2. La función de densidad de cada variable X_i es

$$p(x, \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) e^{-\frac{1}{2\,\sigma^2} (x - \mu)^2} \; .$$

Por lo tanto,

$$\frac{\partial \ln p(x,\mu,\sigma^2)}{\partial \mu} = \frac{x-\mu}{\sigma^2}$$

y

$$\frac{\partial \ln p(x,\mu,\sigma^2)}{\partial \sigma^2} = -\frac{1}{2\,\sigma^2} + (\sigma^2)^2\,\frac{1}{2}\,(x-\mu)^2\ .$$

Luego el sistema (3.3) se transforma en el sistema

$$\begin{aligned} \sum_{i=1}^{n}(X_i-\widehat{\mu})/\widehat{\sigma}^2 &= 0 \\ \sum_{i=1}^{n} -\frac{1}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}(X_i-\widehat{\mu})^2 &= 0 \end{aligned}$$

que tiene como solución

$$\begin{aligned} \widehat{\mu}(X_1,X_2,\ldots,X_n) &= \sum_{i=1}^{n} X_i/n = \overline{X} \\ \widehat{\sigma}^2(X_1,X_2,\ldots,X_n) &= \sum_{i=1}^{n}(X_i-\overline{X})^2/n \end{aligned}$$

que son los mismos estimadores que encontramos por el método de los momentos.

Ejemplo 4: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $\Gamma(\alpha,\lambda)$. La densidad de X_i está dada por

$$p(x,\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\,\lambda^{\alpha}\,x^{\alpha-1}\,e^{-\lambda\,x}\ ,$$

con lo cual

$$\frac{\partial \ln p(x,\alpha,\lambda)}{\partial \alpha} = \ln\lambda + \ln x - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

y

$$\frac{\partial \ln p(x,\alpha,\lambda)}{\partial \lambda} = \frac{\alpha}{\lambda} - x\ ,$$

donde $\Gamma'(\alpha)$ indica la derivada de la función $\Gamma(\alpha)$. Luego el sistema (3.3) se transforma en el sistema

$$\begin{aligned} n \ln \widehat{\lambda} + \sum_{i=1}^{n} \ln(X_i) - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} &= 0 \\ \frac{n \widehat{\alpha}}{\widehat{\lambda}} - n \overline{X} &= 0 \,, \end{aligned}$$

con $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Luego $\widehat{\lambda} = \frac{\widehat{\alpha}}{\overline{X}}$. Pero, este sistema no tiene una solución explícita ya que al reemplazar el valor de $\widehat{\lambda}$ obtenemos la ecuación no lineal

$$n \left(\ln \widehat{\alpha} - \ln(\overline{X}) \right) + \sum_{i=1}^{n} \ln(X_i) - n \frac{\Gamma'(\widehat{\alpha})}{\Gamma(\widehat{\alpha})} = 0 \,,$$

que puede resolverse, por ejemplo mediante, el algoritmo de Newton-Raphson. Para iniciar el proceso, se puede tomar como estimador inicial el estimador de los momentos, por ejemplo.

En este caso, el estimador de máxima verosimilitud no coincide con el estimador de los momentos.

Invarianza de los E.M.V. Supongamos que $\boldsymbol{\lambda} = q(\boldsymbol{\theta})$ es una función *biunívoca* de Θ sobre Λ, donde $\Lambda \subset I\!R^p$. Luego la densidad $p(\mathbf{x}, \boldsymbol{\theta})$ se puede expresar en función de $\boldsymbol{\lambda}$ ya que $\boldsymbol{\theta} = q^{-1}(\boldsymbol{\lambda})$. Denominemos a la densidad de $\mathbf{X}$ como función de $\boldsymbol{\lambda}$ por $p^*(\mathbf{x}, \boldsymbol{\lambda})$. Claramente se tiene

$$p^*(\mathbf{x}, \boldsymbol{\lambda}) = p(\mathbf{x}, q^{-1}(\boldsymbol{\lambda}))$$

Luego se definen los E.M.V. $\widehat{\boldsymbol{\theta}}$ y $\widehat{\boldsymbol{\lambda}}$ por

$$p(\mathbf{x}, \widehat{\boldsymbol{\theta}}) = \max_{\boldsymbol{\theta} \in \Theta} p(\mathbf{x}, \boldsymbol{\theta}) \tag{3.4}$$

y

$$p^*(\mathbf{x}, \widehat{\boldsymbol{\lambda}}) = \max_{\lambda \in \Lambda} p^*(\mathbf{x}, \boldsymbol{\lambda}) \tag{3.5}$$

El siguiente teorema muestra que los estimadores de máxima verosimilitud son invariantes por transformaciones biunívocas.

Teorema 1: *Si $\widehat{\boldsymbol{\theta}}$ es E.M.V. de $\boldsymbol{\theta}$, entonces $\widehat{\boldsymbol{\lambda}} = q(\widehat{\boldsymbol{\theta}})$ es E.M.V. de $\boldsymbol{\lambda}$.*

DEMOSTRACIÓN: Como $\widehat{\boldsymbol{\theta}}$ es E.M.V. de $\boldsymbol{\theta}$ se tendrá que (3.4) vale. Como $\widehat{\lambda} = q(\widehat{\boldsymbol{\theta}})$, (3.4) se puede escribir como

$$p(\mathbf{x}, q^{-1}(\widehat{\boldsymbol{\lambda}})) = \max_{\boldsymbol{\lambda} \in \Lambda} p(\mathbf{x}, q^{-1}(\boldsymbol{\lambda}))$$

pero, esta ecuación de acuerdo a la definición de p^* es equivalente a

$$p^*(\mathbf{x}, \widehat{\boldsymbol{\lambda}}) = \max_{\boldsymbol{\lambda} \in \Lambda} p^*(\mathbf{x}, \boldsymbol{\lambda}) \ ,$$

luego $\widehat{\boldsymbol{\lambda}}$ satisface (3.5) y por lo tanto es un E.M.V. de $\boldsymbol{\lambda}$.

Ejemplo 5: De acuerdo al Teorema 1, en el ejemplo 2, el E.M.V. de $\lambda = q(\theta) = \ln \theta$ será

$$\widehat{\lambda} = \ln \widehat{\theta} = \ln \Big(\frac{\overline{X}}{k} \Big) \ .$$

En general, si $\boldsymbol{\lambda} = q(\boldsymbol{\theta})$, aunque q no sea bunívoca, se define el estimador de máxima verosimilitud de $\boldsymbol{\lambda}$ por

$$\widehat{\boldsymbol{\lambda}} = q(\widehat{\boldsymbol{\theta}}) \ .$$

Ejemplo 6: Supongamos que en el ejemplo 3 interese encontrar el E.M.V. de $\lambda = q(\mu, \sigma^2) = \mu/\sigma^2$. Aunque esta transformación no es biunívoca, el E.M.V. de λ será

$$\widehat{\lambda} = q(\widehat{\mu}, \widehat{\sigma}^2) = \frac{\overline{X}}{\sum_{i=1}^n (X_i - \overline{X})^2/n} = \frac{\sum_{i=1}^n X_i}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

pues basta completar la transformación dada a una transformación biunívoca, tomando por ejemplo, $q_1(\mu, \sigma^2) = \mu$.

3.4 Método de cuadrados mínimos

Supongamos que $X_1, X_2, \ldots, X_n$ son variables aleatorias de la forma

$$X_i = S_i(\theta_1, \ldots, \theta_p) + \varepsilon_i \qquad\qquad 1 \leq i \leq n \tag{3.6}$$

donde $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ es un vector de parámetros desconocido, del cual lo único que se conoce es que está en un conjunto $\Theta \subset I\!R^p$ y ε_i son variables aleatorias con

(i) $E(\varepsilon_i) = 0$

(ii) $\mathrm{Var}(\varepsilon_i) = \sigma^2$

(iii) $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son variables aleatorias independientes.

Ejemplo 1: Consideremos el ejemplo 3 de 2.4. Luego, en este caso, poniendo θ_1 en lugar de a y θ_2 en lugar de b, se tiene

$$X_i = \theta_1 G_i + \theta_2 + \varepsilon_i \qquad 1 \leq i \leq n$$

donde las variables ε_i satisfacen (i), (ii) y (iii).

Luego si llamamos:

$$S_i(\theta_1, \theta_2) = \theta_1 G_i + \theta_2 \qquad 1 \leq i \leq n$$

estamos en la situación descripta por la ecuación (3.6).

Ejemplo 2: Podemos generalizar el ejemplo 1 por la siguiente situación. Supongamos que la variable X depende de otras dos variables G y H y que la forma de la dependencia es

$$X = u(G, H, \theta_1, \theta_2, \ldots, \theta_p) + \varepsilon$$

donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ se conoce que pertenece a un conjunto $\Theta \subset I\!R^p$, y donde ε es una variable aleatoria que aglutina todos los otros factores que determina X y que son desconocidos.

Por ejemplo se pueden tener

$$u_1(G, H, \boldsymbol{\theta}) = \theta_1 G + \theta_2 H + \theta_3$$

o

$$u_2(G, H, \boldsymbol{\theta}) = \theta_1 G^2 + \theta_2 H^2 + \theta_3 HG + \theta_4 H + \theta_5 G + \theta_6$$

o

$$u_3(G, H, \boldsymbol{\theta}) = \theta_1 e^{\theta_2 G} + \theta_3 e^{\theta_4 H} \ .$$

Supongamos que se hagan n experimentos. En el experimento i-ésimo se fijan G y H iguales respectivamente a G_i y H_i y se observa un valor X_i. Luego se tendrá

$$X_i = u(G_i, H_i, \theta_1, \theta_2, \ldots, \theta_p) + \varepsilon_i \qquad 1 \leq i \leq n$$

donde se puede suponer que las ε_i satisfacen (i), (ii) y (iii). Luego, si llamamos

$$S_i(\theta_1, \theta_2, \ldots, \theta_p) = u(G_i, H_i, \theta_1, \theta_2, \ldots, \theta_p)$$

obtenemos que las variables X_i satisfacen (3.6).

Llamaremos *estimador de cuadrados mínimos* (E.C.M.) al valor $\widehat{\boldsymbol{\theta}}(X_1, X_2, \ldots, X_n)$ que hace mínima la expresión $\sum_{i=1}^n (X_i - S_i(\theta_1, \theta_2, \ldots, \theta_p))^2$, es decir si

$$\sum_{i=1}^n (X_i - S_i(\widehat{\boldsymbol{\theta}}))^2 = \min_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^n (X_i - S_i(\boldsymbol{\theta}))^2 \ . \tag{3.7}$$

Este estimador tiene la siguiente justificación intuitiva: Se desea que $S_i(\theta_1 \ldots \theta_p)$ "ajuste" bien a X_i, y por lo tanto los términos residuales ε_i deberían ser pequeños. Esto se logra minimizando la suma de los cuadrados de las desviaciones respectivas.

Se puede demostrar que si además de satisfacer (i), (ii) y (iii), los ε_i tienen distribución normal, entonces el E.M.C. coincide con el E.M.V. Esto se verá en el problema 3 de 3.4.

Computación de los E.C.M.: Si Θ es abierto y si las funciones $S_i(\theta_1, \theta_2, \ldots, \theta_p)$ son derivables respecto a cada θ_i, $\widehat{\boldsymbol{\theta}}$ deberá satisfacer el sistema de ecuaciones siguiente

$$\frac{\partial \sum_{i=1}^n (X_i - S_i(\widehat{\boldsymbol{\theta}}))^2}{\partial \theta_j} = 0 \qquad\qquad j = 1, 2, \ldots, p \ ,$$

que es equivalente a:

$$\sum_{i=1}^n (X_i - S_i(\widehat{\boldsymbol{\theta}})) \frac{\partial S_i(\widehat{\boldsymbol{\theta}})}{\partial \theta_j} = 0 \qquad\qquad j = 1, 2, \ldots, p \ .$$

Igual que en el caso de los E.M.V. estas condiciones son necesarias para el E.M.C. pero no son suficientes. También se deberán cumplir las condiciones de segundo orden, y se deberá verificar que se trata de un mínimo absoluto y no local.

Ejemplo 3: Volvemos al ejemplo 1. Luego se tiene

$$\frac{\partial S_i(\boldsymbol{\theta})}{\partial \theta_1} = G_i \qquad \text{y} \qquad \frac{\partial S_i(\boldsymbol{\theta})}{\partial \theta_2} = 1 \ .$$

Luego (3.7) se transforma en

$$\begin{aligned}\sum_{i=1}^{n}(X-\widehat{\theta}_1 G_i-\widehat{\theta}_2)G_i &= 0\\ \sum_{i=1}^{n}(X_i-\widehat{\theta}_1 G_i-\widehat{\theta}_2) &= 0\,.\end{aligned}$$

Es fácil ver la que la solución de este sistema viene dada por

$$\begin{aligned}\widehat{\theta}_1 &= \sum_{i=1}^{n}(X_i-\overline{X})(G_i-\overline{G})\Big/\sum_{i=1}^{n}(G_i-\overline{G})^2\,,\\ \widehat{\theta}_2 &= \overline{X}-\widehat{\theta}_1\overline{G}\,,\end{aligned}$$

donde

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad \text{y} \qquad \overline{G}=\frac{1}{n}\sum_{i=1}^{n}G_i\,.$$

Geométricamente la recta $X=\widehat{\theta}_1 G+\widehat{\theta}_2$ tiene la propiedad siguiente: Minimaza la suma de los cuadrados de las distancias de los puntos (G_i, X_i) a la recta, si esta distancia se la mide paralelamente al eje de las X. Es decir si $X_i^*=\theta_1 G_1+\theta_2$, la recta $X=\widehat{\theta}_1 G+\widehat{\theta}_2$ hace mínimo $\sum_{i=1}^{n}(X_i-X_i^*)^2$.

Para un mayor desarrollo de los métodos de cuadrados mínimos, consultar Draper y Smith [2].

3.5 Criterios para medir la bondad de un estimador

Supongamos que se tenga una muestra $\mathbf{X}=(X_1, X_2, \ldots, X_n)$ de cuya distribución sólo se conoce que pertenece a la familia $\mathcal{F}=\{F(\mathbf{x},\boldsymbol{\theta}) \text{ donde } \boldsymbol{\theta}\in\Theta\subset I\!R^p\}$. Supongamos además que se está interesado en estimar una función real $q(\boldsymbol{\theta})$. Para poder elegir el estimador $\delta(\mathbf{X})$ que se utilizará, se deberá dar un criterio para comparar dos estimadores cualesquiera. Esto se hará como sigue:

Es razonable pensar que dado un estimador $\delta(\mathbf{X})$ de $q(\boldsymbol{\theta})$, el error $\delta(\mathbf{X})-q(\boldsymbol{\theta})$ producirá un perjuicio o pérdida dado por un real no negativo, que dependerá por un lado del valor del estimador $\delta(\mathbf{X})$ y por otro del valor verdadero del vector $\boldsymbol{\theta}$ de parámetros.

Así llamaremos *función de pérdida* a una función $\ell(\boldsymbol{\theta}, d)$ no negativa que nos indica cuánto se pierde cuando el valor del estimador es "d" y el valor verdadero del vector de parámetros es $\boldsymbol{\theta}$. Entonces si usamos el estimador $\delta(\mathbf{X})$ la pérdida será

$$\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))$$

y esta pérdida será una variable aleatoria ya que depende de $\mathbf{X}$. Para evaluar *globalmente* el estimador $\delta(\mathbf{X})$ se puede utilizar el valor medio de esta pérdida, que indicará de acuerdo a la ley de los grandes números aproximadamente la pérdida promedio, si estimamos $q(\boldsymbol{\theta})$ muchas veces con vectores $\mathbf{X}$ independientes. Luego, definimos la función de pérdida media del estimador δ o *función de riesgo* $R(\delta, \boldsymbol{\theta})$ a

$$R(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\ell(\boldsymbol{\theta}, \delta(\mathbf{X})))$$

Un primer ejemplo de función de pérdida puede obtenerse tomando el error absoluto, es decir

$$\ell_1(\boldsymbol{\theta}, d) = |d - q(\boldsymbol{\theta})|$$

y en este caso, la pérdida media corresponde a un estimador $\delta(\mathbf{X})$ viene dada por

$$R_1(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(|\delta(\mathbf{X} - q(\boldsymbol{\theta})|)$$

Si consideramos como función de pérdida el cuadrado del error tenemos

$$\ell_2(\boldsymbol{\theta}, d) = (d - q(\boldsymbol{\theta}))^2$$

que es una función que desde el punto de vista matemático es más sencilla que ℓ_1, ya que es derivable en todo punto.

La función de pérdida cuadrática fue la primera utilizada en Estadística, y aún hoy la más difundida. De ahora en adelante, salvo mención en contrario supondremos que la función de pérdida es ℓ_2. La pérdida media, o riesgo, correspondiente está dada por

$$R_2(\delta, \boldsymbol{\theta}) = E(\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2$$

y será llamada en adelante error cuadrático medio, e indicada por $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta)$. Luego

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) = R_2(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2 \tag{3.8}$$

La función $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta)$ nos proporciona un criterio para determinar si un estimador $\delta_1(\mathbf{X})$ de $q(\boldsymbol{\theta})$ es mejor que otro $\delta_2(\mathbf{X})$, basta verificar

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta_1) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta_2) \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

En este orden de ideas, un estimador óptimo δ^* podría definirse mediante la siguiente condición: Para cualquier otro estimador δ se tiene

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta) \qquad \forall\, \theta \in \Theta \tag{3.9}$$

Sin embargo, salvo en casos triviales no existirán tales estimadores óptimos. Para mostrar esto definamos para cada posible valor $\boldsymbol{\theta} \in \Theta$, el estimador constante $\delta_{\boldsymbol{\theta}}(\mathbf{X}) = q(\boldsymbol{\theta})$ que no depende del valor de la muestra. Luego si δ^* satisface (3.9), debe cumplirse:

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta_{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}((q(\boldsymbol{\theta}) - q(\boldsymbol{\theta}))^2) = 0 \quad \forall \boldsymbol{\theta} \in \Theta$$

Pero como $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \geq 0$ y $\ell_2(\boldsymbol{\theta}, d) = 0$ implica que $d = q(\boldsymbol{\theta})$, se obtiene

$$P_{\boldsymbol{\theta}}(\delta^*(\mathbf{X}) = q(\boldsymbol{\theta})) = 1 \quad \forall \boldsymbol{\theta} \in \Theta \tag{3.10}$$

(donde $P_{\boldsymbol{\theta}}(\Lambda)$ indica la probabilidad del evento Λ cuando el valor de los parámetros está dado por el vector $\boldsymbol{\theta}$). La ecuación (3.10) significa que a partir de la muestra se puede estimar sin error $q(\boldsymbol{\theta})$. Esta situación sólo se da muy raramente, por ejemplo, cuando $q(\boldsymbol{\theta})$ es constante.

Otro ejemplo algo diferente de función de pérdida, corresponde a la función

$$\ell_3(\boldsymbol{\theta}, d) = I_{\{|q(\boldsymbol{\theta}) - d| > c\}}$$

donde $I_{\{|q(\boldsymbol{\theta}) - d| > c\}}$ es la función que vale 1 si $|q(\boldsymbol{\theta}) - d| > c$ y 0 en caso contrario. Esta pérdida da origen a la función de riesgo

$$R_3(\delta, \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(|\delta(\mathbf{X}) - q(\boldsymbol{\theta})| > c)\ .$$

A diferencia de las anteriores, en este caso, $\ell_3(\boldsymbol{\theta}, d) = 0$ no implica implica $q(\boldsymbol{\theta}) = d$. Por otra parte, esta pérdida no es convexa como función de d mientras que ℓ_1 y ℓ_2 lo son. En muchas situaciones, se podrán obtener procedimientos de estimación más efectivos para pérdidas convexas.

El estimador δ^* con E.C.M. mínimo uniformemente en $\boldsymbol{\theta}$ como se indica en (3.9) no existe, salvo en casos excepcionales, debido a que la clase de todos los posibles estimadores es muy amplia y contiene estimadores poco razonables como los $\delta_{\boldsymbol{\theta}}(\mathbf{X})$ definidos anteriormente. Por lo tanto, una manera de obtener estimadores óptimos consistirá en restringir primero la clase de los estimadores δ considerados, y luego buscar aquél con E.C.M. uniformemente menor dentro de esta clase. Otra forma de obtener estimadores óptimos consistirá en minimizar algún criterio general basado en la función de riesgo, como el máximo riesgo.

Antes de empezar el estudio de las clases de estimadores daremos una noción importante.

Definición 1: Se dice que un estimador $\delta(\mathbf{X})$ de $q(\boldsymbol{\theta})$ es *inadmisible* respecto de la pérdida $\ell(\boldsymbol{\theta}, d)$, si existe otro estimador $\delta'(\mathbf{X})$ mejor que él, es decir, si existe $\delta'(\mathbf{X})$ tal que

$$R(\delta', \boldsymbol{\theta}) \leq R(\delta, \boldsymbol{\theta}) \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

El estimador $\delta(\mathbf{X})$ se dirá *admisible* si no es inadmisible, es decir, si no existe ningún otro estimador que sea uniformemente mejor que él.

El siguiente Teorema muestra la ventaja de utilizar pérdidas convexas.

Teorema 1. *Supongamos que $\ell(\boldsymbol{\theta}, d)$ es una pérdida estrictamente convexa en d y que $\delta(\mathbf{X})$ es admisible para $q(\boldsymbol{\theta})$. Si $\delta'(\mathbf{X})$ es otro estimador de $q(\boldsymbol{\theta})$ con el mismo riesgo que $\delta(\mathbf{X})$ entonces $P_{\boldsymbol{\theta}}(\delta(\mathbf{X}) = \delta'(\mathbf{X})) = 1$.*
Demostración. Supongamos que $P_{\boldsymbol{\theta}}(\delta(\mathbf{X}) = \delta'(\mathbf{X})) < 1$ y sea $\delta^*(\mathbf{X}) = (\delta(\mathbf{X}) + \delta'(\mathbf{X}))/2$. Luego, por ser $\ell(\boldsymbol{\theta}, d)$ convexa se cumple

$$\ell(\boldsymbol{\theta}, \delta^*(\mathbf{X})) < \frac{\ell(\boldsymbol{\theta}, \delta(\mathbf{X})) + \ell(\boldsymbol{\theta}, \delta'(\mathbf{X}))}{2} \tag{3.11}$$

salvo si $\delta(\mathbf{X}) = \delta'(\mathbf{X})$. Luego, tomando esperanza en ambos miembros de (3.11) se obtiene

$$R(\delta^*, \boldsymbol{\theta}) < \frac{R(\delta, \boldsymbol{\theta}) + R(\delta', \boldsymbol{\theta})}{2} = R(\delta, \boldsymbol{\theta}) \tag{3.12}$$

lo que contradice el hecho de que $\delta(\mathbf{X})$ es admisible.

3.6 Estimadores insesgados

Una propiedad "razonable" que se puede exigir a un estimador está dada por la siguiente definición:

Definición 1: Se dice que $\delta(\mathbf{X})$ es un *estimador insesgado* para $q(\boldsymbol{\theta})$ si $E_{\boldsymbol{\theta}}(\delta(\mathbf{X})) = q(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta$.

Esto significa que si calculamos el estimador δ para varias muestras independientes, y luego promediamos los valores así obtenidos, entonces de

acuerdo a la ley de los grandes números el promedio converge al valor $q(\boldsymbol{\theta})$ que queremos estimar.

Definición 2: Si un estimador no es insesgado, se dice *sesgado*, definiéndose el *sesgo* del estimador como $E_{\boldsymbol{\theta}}(\delta(\mathbf{X})) - q(\boldsymbol{\theta})$.

Cuando $\delta(\mathbf{X})$ es un estimador insesgado, su ECM coincide con su varianza ya que

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) = E_{\boldsymbol{\theta}}[(\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2] = E_{\boldsymbol{\theta}}[(\delta(\mathbf{X}) - E_{\boldsymbol{\theta}}(\delta(\mathbf{X})))^2] = \mathrm{Var}_{\boldsymbol{\theta}}(\delta(\mathbf{X})).$$

Para ilustrar estas definiciones veremos algunos ejemplos.

Ejemplo 1: Supongamos tener una variable X de cuya distribución F en la población sólo se sabe que tiene esperanza finita, es decir sólo se conoce que pertenece a $\mathcal{F}$, donde $\mathcal{F}$ es la familia de todas las distribuciones con esperanza finita. Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de F y supongamos que se quiere estimar $q_1(F) = E_F(X)$. Estamos frente a un problema de estimación no paramétrica, ya que la familia no puede indicarse con un número finito de parámetros. Un posible estimador para $q_1(F)$ es $\overline{X} = (1/n)\sum_{i=1}^n X_i$. El estimador $\overline{X}$ es insesgado ya que

$$E_F(\overline{X}) = E_F\Big(\frac{1}{n}\sum_{i=1}^n X_i\Big) = \frac{1}{n}\sum_{i=1}^n E_F(X_i) = E_F(X) = q_1(F)$$

$\overline{X}$ se denomina *media muestral.*

Ejemplo 2: Supongamos ahora que se conoce que la distribución F de X en la población pertenece a la familia $\mathcal{F}$ de todas las distribuciones que tienen segundo momento finito, es decir tales que $E_F(X^2) < \infty$. Supongamos que se quiere estimar $q_2(F) = \mathrm{Var}_F(X)$ a partir de una muestra aleatoria $X_1, X_2, \ldots, X_n$. Ya hemos visto que un estimador adecuado podría ser

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$$

Veremos que $\widehat{\sigma}^2$ no es un estimador insesgado de $q_2(F)$. Desarrollando el cuadrado del segundo miembro en la definición obtenemos

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^n X_i^2 - n\overline{X}^2}{n}.$$

Luego, se tiene

$$E_F(\widehat{\sigma}^2) = E_F(X^2) - E_F(\overline{X}^2) \tag{3.13}$$

Por otro lado, se tiene

$$\mathrm{Var}_F(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}_F(X_i) = \frac{1}{n}\mathrm{Var}_F(X) \ .$$

Como

$$\mathrm{Var}_F(\overline{X}) = E_F(\overline{X}^2) - (E_F(\overline{X}))^2 \ ,$$

resulta

$$E_F(\overline{X}^2) = \mathrm{Var}_F(\overline{X}) + (E_F(\overline{X}))^2 = \frac{1}{n}\mathrm{Var}_F(X) + (E_F(X))^2 \tag{3.14}$$

y reemplazando (3.14) en (3.13) resulta

$$\begin{aligned} E_F(\widehat{\sigma}^2) &= E_F(X^2) - (E_F(X))^2 - \frac{1}{n}\mathrm{Var}_F(X) = \mathrm{Var}_F(X)(1-1/n) \\ &= \frac{n-1}{n}\ \mathrm{Var}_F(X) = \frac{n-1}{n} q_2(F). \end{aligned}$$

Esto prueba que $\widehat{\sigma}^2$ no es un estimador insesgado para $\mathrm{Var}_F(X)$, aunque el sesgo es $-\mathrm{Var}_F(X)/n$, y por lo tanto, tiende a 0 cuando n tiende a infinito. El sesgo puede corregirse dividiendo $\widehat{\sigma}^2$ por $(n-1)/n$, obteniendo así el estimador insesgado

$$s^2 = \frac{n}{n-1}\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

que denominaremos *varianza muestral.*

Ejemplo 3: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución de la cual se conoce únicamente que pertenece a la familia $N(\mu, \sigma^2)$ y supongamos que se quieran estimar μ y σ^2. Como se tiene

$$\mu = E_{\mu,\sigma^2}(X) \ ; \quad \sigma^2 = \mathrm{Var}_{\mu,\sigma^2}(X)$$

por lo visto en Ejemplos 1 y 2, resulta que $\overline{X}$ y s^2 son estimadores insesgados de μ y σ^2 respectivamente.

Si nos restringimos a la clase de los estimadores insesgados, se podrá encontrar frecuentemente, estimadores óptimos. Daremos la siguiente definición:

Definición 2: Se dirá que $\delta(\mathbf{X})$ es un *estimador insesgado de mínima varianza* para $q(\boldsymbol{\theta})$, uniformemente en $\boldsymbol{\theta} \in \Theta$ (IMVU) si:

(a) $\delta(\mathbf{X})$ es insesgado para $q(\boldsymbol{\theta})$

(b) dado otro estimador insesgado para $q(\boldsymbol{\theta})$, $\delta^*(\mathbf{X})$, se cumple $\mathrm{Var}_{\boldsymbol{\theta}}(\delta(\mathbf{X})) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta^*(\mathbf{X})) \quad \forall\, \boldsymbol{\theta} \in \Theta$.

3.7 Estadísticos suficientes

Consideremos un vector aleatorio $\mathbf{X}$ de dimensión n cuya distribución pertenece a una familia $\mathcal{F} = \{F(\mathbf{x}, \boldsymbol{\theta}) \text{ con } \boldsymbol{\theta} \in \Theta \subset I\!R^p\}$. El vector $\mathbf{X}$ interesa en cuanto nos provee información sobre el valor verdadero de $\boldsymbol{\theta}$. Puede ocurrir que una parte de la información contenida en $\mathbf{X}$ carezca de interés para el conocimiento de $\boldsymbol{\theta}$, y por consiguiente convenga eliminarla simplificando así la información disponible.

Al realizar esta simplificación, eliminando de $\mathbf{X}$ toda la información irrelevante, se obtendrá otro vector $\mathbf{T}$ que puede ser de dimensión menor que n.

Llamaremos *estadístico* a cualquier función medible $\mathbf{T} = r(\mathbf{X})$ con valores en un espacio euclídeo de dimensión finita.

Si la función r no es biunívoca, del conocimiento de $\mathbf{T}$ no se podrá reconstruir el valor de $\mathbf{X}$, por lo que $\mathbf{T}$ conservará sólo una parte de la información que hay en $\mathbf{X}$. El estadístico $\mathbf{T}$ será llamado *suficiente* cuando conserve toda la información relevante para el conocimiento de $\boldsymbol{\theta}$. Esto se formalizará en la siguiente definición.

Definición 1: Sea $\mathbf{X}$ un vector aleatorio de dimensión n cuya distribución es $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Se dice que un *estadístico* $\mathbf{T} = r(\mathbf{X})$ *es suficiente* para $\boldsymbol{\theta}$ si la distribución de $\mathbf{X}$ condicional a que $\mathbf{T} = \mathbf{t}$ es independiente de $\boldsymbol{\theta}$ para todo $\mathbf{t}$.

Esto puede interpretarse como afirmando que una vez conocido el valor $\mathbf{t}$ de $\mathbf{T}$, la distribución de $\mathbf{X}$ es independiente de $\boldsymbol{\theta}$ y por lo tanto no contiene información suplementaria sobre $\boldsymbol{\theta}$. En otros términos: una vez conocido el valor de $\mathbf{T}$ podemos olvidarnos del valor $\mathbf{X}$, ya que en $\mathbf{T}$ está toda la información que $\mathbf{X}$ tiene sobre $\boldsymbol{\theta}$.

Ejemplo 1: Supongamos que una máquina produce cierto artículo, existiendo la probabilidad θ de que lo produzca defectuoso. Supongamos además que se observa un lote de n artículos producidos sucesivamente por la máquina,

de manera que la aparición de uno defectuoso resulte independiente del resultado obtenido para los restantes artículos del lote.

Consideremos las variable aleatorias X_i, $1 \leq i \leq n$, que valen 1 ó 0 según el i-ésimo artículo observado sea o no defectuoso. Entonces cada una de las variables $X_1, X_2, \ldots, X_n$ sigue una ley binomial $\mathrm{Bi}(\theta, 1)$, de modo que la función de probabilidad puntual conjunta es igual a

$$p(x_1, x_2, \ldots, x_n, \theta) = \theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{n-\Sigma_{i=1}^n x_i}$$

donde x_i vale 0 ó 1.

Si queremos estimar el parámetro θ, parece razonable pensar que sólo se deberá utilizar la cantidad total de artículos defectuosos del lote, ya que el orden en que han aparecido los mismos parece irrelevante para el conocimiento de θ. Por lo tanto, es de esperar que el estadístico $T = \sum_{i=1}^n X_i$ sea suficiente.

Para ver si esta conjetura es correcta, calculemos la distribución de $\mathbf{X} = (X_1, \ldots, X_n)$ dado $T = t$:

$$p_{\mathbf{X}|T}(x_1, \ldots, x_n, \theta|t) = \frac{p_{\mathbf{X},T}(x_1, x_2, \ldots, x_n, t, \theta)}{p_T(t, \theta)} \tag{3.15}$$

El numerador de este cociente es la probabilidad conjunta:

$$\begin{aligned} & P_\theta(X_1 = x_1, \ldots, X_n = x_n, r(X_1, \ldots, X_n) = t) \\ = & \begin{cases} \theta^t(1-\theta)^{n-t} & \text{si } r(x_1, \ldots, x_n) = t \\ 0 & \text{si } r(x_1, \ldots, x_n) \neq t \end{cases} \end{aligned}$$

y como el estadístico $T = \sum_{i=1}^n X_i$ sigue una ley binomial $\mathrm{Bi}(\theta, n)$ el denominador de (3.15) vale

$$p_T(t, \theta) = \binom{n}{t} \theta^t(1-\theta)^{n-t}$$

Así resulta

$$p_{\mathbf{X}|T}(x_1, \ldots, x_n, \theta|t) = \begin{cases} 1/\binom{n}{t} & \text{si } r(x_1, \ldots, x_n) = t \\ 0 & \text{si } r(x_1, \ldots, x_n) \neq t\ . \end{cases}$$

De esta manera $p_{\mathbf{X}/T}$ es independiente de θ y por lo tanto el estadístico $T = \Sigma\, X_i$ es suficiente para θ.

Una caracterización útil de los estadísticos suficientes es la proporcionada por el siguiente teorema:

Teorema 1 (de factorización): *Sea* $\mathbf{X}$ *un vector aleatorio con función de densidad o función de probabilidad puntual* $p(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. *Entonces, el estadístico* $\mathbf{T} = r(\mathbf{X})$ *es suficiente para* $\boldsymbol{\theta}$ *si y sólo si existen dos funciones* g *y* h *tales que*

$$p(\mathbf{x}, \boldsymbol{\theta}) = g(r(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}) \tag{3.16}$$

DEMOSTRACIÓN: La haremos sólo para el caso discreto. Supongamos primero que existen dos funciones g y h tales que $p(\mathbf{x}, \boldsymbol{\theta})$ se factoriza según (3.16). Entonces la función de densidad conjunta vale

$$p_{\mathbf{XT}}(\mathbf{x}, \mathbf{t}, \boldsymbol{\theta}) = \begin{cases} g(\mathbf{t}, \boldsymbol{\theta}) h(\mathbf{x}) & \text{si } r(\mathbf{x}) = \mathbf{t} \\ 0 & \text{si } r(\mathbf{x}) \neq \mathbf{t} \end{cases}$$

y la densidad marginal $p_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\theta})$ está dada por

$$\begin{aligned} p_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\theta}) &= \sum_{r(\mathbf{x})=\mathbf{t}} p_{\mathbf{XT}}(\mathbf{x}, \mathbf{t}, \boldsymbol{\theta}) = \sum_{r(\mathbf{x})=\mathbf{t}} g(r(\mathbf{x}), \boldsymbol{\theta}) h(\mathbf{x}) \\ &= g(\mathbf{t}, \boldsymbol{\theta}) \sum_{r(\mathbf{x})=\mathbf{t}} h(\mathbf{x}) = g(\mathbf{t}, \boldsymbol{\theta}) h^*(\mathbf{t}) \end{aligned}$$

donde las sumatorias se realizan sobre todos los $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ tales que $r(\mathbf{x}) = \mathbf{t}$. Así resulta la función de densidad condicional

$$p_{\mathbf{X}|\mathbf{T}}(\mathbf{x}, \boldsymbol{\theta}|\mathbf{t}) = \begin{cases} h(\mathbf{x})/h^*(\mathbf{t}) & \text{si } r(\mathbf{x}) = \mathbf{t} \\ 0 & \text{si } r(\mathbf{x}) \neq \mathbf{t} \end{cases}$$

y por lo tanto la distribución de $\mathbf{X}$ dado $\mathbf{T} = \mathbf{t}$ es independiente de $\boldsymbol{\theta}$ para todo $\mathbf{t}$.

Recíprocamente, si suponemos que $\mathbf{T} = r(\mathbf{X})$ es suficiente para $\boldsymbol{\theta}$, se tiene

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}) &= P_{\boldsymbol{\theta}}(\mathbf{X} = \mathbf{x},\, \mathbf{T} = r(\mathbf{x})) = p_{\mathbf{XT}}(\mathbf{x}, r(\mathbf{x}), \boldsymbol{\theta}) \\ &= p_{\mathbf{X}|\mathbf{T}}(\mathbf{x}, \boldsymbol{\theta}|r(\mathbf{x})) p_{\mathbf{T}}(r(\mathbf{x}), \boldsymbol{\theta}) \end{aligned}$$

El primero de los factores del último miembro es por hipótesis independiente de $\boldsymbol{\theta}$ y por eso podemos llamarlo $h(\mathbf{x})$; mientras que el segundo –que depende de $\mathbf{x}$ a través de $\mathbf{t}$– puede denominarse $g(r(\mathbf{x}), \boldsymbol{\theta})$. El teorema queda demostrado. Para una demostración general, ver Teorema 8 y Corolario 1 de Lehmann [4]. También se puede ver Bahadur [1].

Ejemplo 2: Supongamos que las variables aleatorias $X_1, X_2, \ldots, X_n$ son independientes y que están uniformemente distribuídas en el intervalo $[\theta_1, \theta_2]$ de manera que su función de densidad conjunta vale

$$p(x_1, \ldots, x_n, \theta_1, \theta_2) = \begin{cases} (\theta_2 - \theta_1)^{-n} & \text{si } \theta_1 \leq x_i \leq \theta_2,\ \forall i,\ 1 \leq i \leq n \\ 0 & \text{en el resto de } I\!R^n \end{cases}$$

Si definimos los estadísticos

$$r_1(\mathbf{X}) = \min\{X_i :\ 1 \leq i \leq n\} \qquad \text{y} \qquad r_2(\mathbf{X}) = \max\{X_i :\ 1 \leq i \leq n\}$$

y si denotamos con $I_{[\theta_1,\theta_2]}(y)$ a la función característica del intervalo $[\theta_1, \theta_2]$ (que vale 1 para todo y del intervalo y 0 fuera del mismo), resulta:

$$p(x_1, \ldots, x_n, \theta_1, \theta_2) = (\theta_2 - \theta_1)^{-n} I_{[\theta_1,\theta_2]}(r_1(x_1, \ldots, x_n)) I_{[\theta_1,\theta_2]}(r_2(x_1, \ldots, x_n))$$

Por lo tanto la función de densidad $p(\mathbf{x}, \boldsymbol{\theta})$ se factoriza como en (3.16) con $h(\mathbf{x}) = 1$. La función g que depende de $\mathbf{X}$ a través de $r_1(\mathbf{x})$ y $r_2(\mathbf{x})$ vale en este caso

$$g(r_1(\mathbf{x}), r_2(\mathbf{x}), \boldsymbol{\theta}) = (\theta_2 - \theta_1)^{-n} I_{[\theta_1,\theta_2]}(r_1(\mathbf{x})) I_{[\theta_1,\theta_2]}(r_2(\mathbf{x}))$$

Esto demuestra que el estadístico

$$\mathbf{T} = (r_1(\mathbf{X})\ ,\ r_2(\mathbf{X}))$$

es suficiente para θ_1 y θ_2.

El siguiente resultado es Corolario inmediato del Teorema 1.

Corolario. *Sea* $\mathbf{X}$ *un vector aleatorio con función de densidad o función de probabilidad puntual* $p(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. *Supongamos que la familia* $\{p(\mathbf{x}, \boldsymbol{\theta})\}$ *tiene soporte común, independiente de* $\boldsymbol{\theta}$. *Entonces, una condición necesaria y suficiente para que* $\mathbf{T}$ *sea suficiente para* $\boldsymbol{\theta}$ *es que fijados* $\boldsymbol{\theta}_1$ *y* $\boldsymbol{\theta}_2$ *el cociente* $\frac{p(\mathbf{x},\boldsymbol{\theta}_1)}{p(\mathbf{x},\boldsymbol{\theta}_2)}$ *sea función de* $\mathbf{T}$.

El siguiente Teorema muestra que una función biunívoca de un estadístico suficiente es también un estadístico suficiente. Esta propiedad es intuitivamente razonable: si $\mathbf{T}$ contiene toda la información relevante acerca de $\boldsymbol{\theta}$, y $\mathbf{T}^*$ es una función biunívoca de $\mathbf{T}$, entonces también $\mathbf{T}^*$ la contiene ya que el vector $\mathbf{T}$ puede reconstruirse a partir del vector $\mathbf{T}^*$.

Teorema 2: *Si* $\mathbf{X}$ *es un vector aleatorio con una distribución* $F(\mathbf{x}, \boldsymbol{\theta})$*, con* $\boldsymbol{\theta} \in \Theta$ *si* $\mathbf{T} = r(\mathbf{X})$ *es un estadístico suficiente para* $\boldsymbol{\theta}$ *y si* m *es una función biunívoca de* $\mathbf{T}$ *entonces el estadístico* $\mathbf{T}^* = m(\mathbf{T})$ *también es suficiente para* $\boldsymbol{\theta}$*.*

DEMOSTRACIÓN: Apliquemos el teorema de factorización a la función de densidad del vector $\mathbf{X}$:

$$p(\mathbf{x}, \boldsymbol{\theta}) = g(r(\mathbf{x}), \boldsymbol{\theta})h(\mathbf{x}) = g(m^{-1}(m(r(\mathbf{x})), \boldsymbol{\theta})h(\mathbf{x})$$

El primer factor del último miembro es una función $g^*(r^*(\mathbf{x}), \boldsymbol{\theta})$, donde $r^*(\mathbf{x}) = m(r(\mathbf{x}))$, y esto prueba que $\mathbf{T}^* = r^*(\mathbf{X})$ es suficiente para $\boldsymbol{\theta}$.

3.8 Estadísticos minimales suficientes

De la noción intuitiva de suficiencia, se deduce que si $\mathbf{T}$ es suficiente para $\boldsymbol{\theta}$ y $\mathbf{T} = H(\mathbf{U})$ entonces $\mathbf{U}$ es suficiente para $\boldsymbol{\theta}$, ya que el conocimiento de $\mathbf{U}$ permite conocer $\mathbf{T}$ que es el que contiene toda la información relevante sobre $\boldsymbol{\theta}$. Más aún, salvo que H sea biunívoca $\mathbf{T}$ da una mayor reducción de la muestra original que $\mathbf{U}$. Este hecho motiva la siguiente definición.

Definición 1: Sea $\mathbf{X}$ un vector aleatorio de dimensión n cuya distribución es $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Se dice que un *estadístico* $\mathbf{T} = r(\mathbf{X})$ *es minimal suficiente* para $\boldsymbol{\theta}$ si dado cualquier otro estadístico $\mathbf{U} = g(\mathbf{X})$ suficiente para $\boldsymbol{\theta}$ existe una función H tal que $\mathbf{T} = H(\mathbf{U})$.

En muchas situaciones, es fácil construir estadísticos minimal suficientes. Sea $S(\boldsymbol{\theta}) = \{\mathbf{x} : p(\mathbf{x}, \boldsymbol{\theta}) > 0\}$, $S(\boldsymbol{\theta})$ se llama el soporte de la densidad o de la probabilidad puntual $p(\mathbf{x}, \boldsymbol{\theta})$, según corresponda. Para simplificar, supondremos que las posibles distribuciones del vector $\mathbf{X}$ tienen todas el mismo soporte, es decir, que el conjunto $S(\boldsymbol{\theta})$ no depende de $\boldsymbol{\theta}$.

Teorema 1. *Supongamos que* $\mathbf{X}$ *tiene una distribución perteneciente a una familia finita de distribuciones* $\mathcal{F} = \{F(\mathbf{x}, \boldsymbol{\theta}_i) \quad 1 \leq i \leq k\}$ *con densidades o probabilidades puntuales* $p(\mathbf{x}, \boldsymbol{\theta}_i)$*,* $1 \leq i \leq k$ *todas con el mismo soporte. Entonces el estadístico*

$$\mathbf{T} = r(\mathbf{x}) = \left(\frac{p(\mathbf{x}, \boldsymbol{\theta}_2)}{p(\mathbf{x}, \boldsymbol{\theta}_1)}, \ldots, \frac{p(\mathbf{x}, \boldsymbol{\theta}_k)}{p(\mathbf{x}, \boldsymbol{\theta}_1)} \right)$$

es minimal suficiente.

DEMOSTRACIÓN. Obviamente, para todo $1 \leq i < j \leq k$ el cociente $p(\mathbf{x}, \boldsymbol{\theta}_i)/p(\mathbf{x}, \boldsymbol{\theta}_j)$ es función de $\mathbf{T}$. Por lo tanto, por el Corolario del teorema de Factorización, $\mathbf{T}$ es suficiente.
Sea ahora $\mathbf{U}$ un estadístico suficiente para $\boldsymbol{\theta}$. Entonces, utilizando el Corolario anterior se cumple que para todo $2 \leq i \leq k$, el cociente $\frac{p(\mathbf{x}, \boldsymbol{\theta}_i)}{p(\mathbf{x}, \boldsymbol{\theta}_1)}$ es una función de $\mathbf{U}$. Luego, $\mathbf{T}$ es función de $\mathbf{U}$ y $\mathbf{T}$ es minimal suficiente.

En muchas situaciones, se pueden obtener estadísticos minimales suficientes combinando el Teorema 1 con el siguiente Teorema.

Teorema 2. *Supongamos que* $\mathbf{X}$ *tiene una distribución perteneciente a una familia de distribuciones* $\mathcal{F} = \{F(\mathbf{x}, \boldsymbol{\theta}) \quad \boldsymbol{\theta} \in \Theta\}$ *con densidades o probabilidades puntuales* $p(\mathbf{x}, \boldsymbol{\theta})$*, todas con el mismo soporte. Sea*

$$\mathcal{F}_0 = \{F(\mathbf{x}, \boldsymbol{\theta}) \quad \boldsymbol{\theta} \in \Theta_0 \subset \Theta\} \subset \mathcal{F}.$$

Supongamos además que $\mathbf{T} = r(\mathbf{X})$ *es un estadístico minimal suficiente para* $\boldsymbol{\theta} \in \Theta_0$ *y suficiente para* $\boldsymbol{\theta} \in \Theta$*, entonces* $\mathbf{T}$ *es minimal suficiente para* $\boldsymbol{\theta} \in \Theta$*.*
DEMOSTRACIÓN. Sea $\mathbf{U}$ un estadístico suficiente para $\boldsymbol{\theta}$, entonces $\mathbf{U}$ es suficiente para $\boldsymbol{\theta} \in \Theta_0$. Por lo tanto, $\mathbf{T}$ es función de $\mathbf{U}$, con lo cual $\mathbf{T}$ es minimal suficiente.

Ejemplo 1. Sean $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $Bi(\theta, 1)$, $0 < \theta < 1$. Hemos visto que $T = \sum_{i=1}^{n} X_i$ es suficiente para $\theta \in (0, 1)$. Queremos ver que es minimal suficiente.

Para ello consideremos la familia finita $\mathcal{F}_0 = \{Bi(1/4, 1), Bi(3/4, 1)\}$. Luego, un estadístico minimal suficiente para esta familia está dado por

$$U = g(\mathbf{x}) = \frac{p(\mathbf{x}, \frac{3}{4})}{p(\mathbf{x}, \frac{1}{4})} = 3^{2T-n}$$

que es una función biunívoca de T. Por lo tanto, T es un estadístico minimal suficiente para $\mathcal{F}_0$ y suficiente para $\theta \in (0, 1)$, con lo cual es minimal suficiente para $\theta \in (0, 1)$.

3.9 Estimadores basados en estadísticos suficientes

Supongamos que $\mathbf{X}$ es un vector correspondiente a una muestra de una distribución que pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Supongamos que $\mathbf{T} = r(\mathbf{X})$ es un estadístico suficiente para $\boldsymbol{\theta}$. Luego de acuerdo al concepto

intuitivo que tenemos de estadístico suficiente, para estimar una función $q(\boldsymbol{\theta})$ deberán bastar estimadores que dependan sólo de $\mathbf{T}$, ya que en $\mathbf{T}$ está toda la información que $\mathbf{X}$ contiene sobre el parámetro $\boldsymbol{\theta}$. Esto es justamente lo que afirma el siguiente teorema.

Teorema 1 (Rao–Blackwell): *Sea* $\mathbf{X}$ *un vector de una distribución perteneciente a la familia* $F(\mathbf{x}, \boldsymbol{\theta})$ *con* $\boldsymbol{\theta} \in \Theta$. *Sea* $\mathbf{T}$ *un estadístico suficiente para* $\boldsymbol{\theta}$ *y* $\delta(\mathbf{X})$ *un estimador de* $q(\boldsymbol{\theta})$. *Definamos un nuevo estimador*

$$\delta^*(\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T}).$$

Luego se tiene

(i) $\mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \leq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta), \ \forall\, \theta \in \Theta$

(ii) *La igualdad en* (i) *se satisface si y sólo si*

$$P_{\boldsymbol{\theta}}(\delta^*(\mathbf{T}) = \delta(\mathbf{X})) = 1 \quad \forall\, \boldsymbol{\theta} \in \Theta$$

(iii) *Si* $\delta(\mathbf{X})$ *es insesgado, entonces* $\delta^*(\mathbf{T})$ *también lo es.*

Demostración: Podemos escribir

$$\begin{aligned} \mathrm{ECM}_{\boldsymbol{\theta}}(\delta) &= E_{\boldsymbol{\theta}}((\delta(\mathbf{X}) - q(\boldsymbol{\theta}))^2) \\ &= E_{\boldsymbol{\theta}}([(\delta^*(\mathbf{T}) - q(\boldsymbol{\theta})) + (\delta(\mathbf{X}) - \delta^*(T))]^2) \\ &= E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))^2) + E_{\boldsymbol{\theta}}((\delta(\mathbf{X}) - \delta^*(\mathbf{T}))^2) \\ &+ 2\, E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) \end{aligned} \tag{3.17}$$

Luego, usando

$$\begin{aligned} E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) &= E_{\boldsymbol{\theta}}[E((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))|\mathbf{T}] \\ &= E_{\boldsymbol{\theta}}[(\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))E(\delta(\mathbf{X}) - \delta^*(\mathbf{T})|\mathbf{T}] \end{aligned}$$

y

$$E_{\boldsymbol{\theta}}(\delta(\mathbf{X}) - \delta^*(\mathbf{T})|\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T}) - \delta^*(\mathbf{T}) = \delta^*(\mathbf{T}) - \delta^*(\mathbf{T}) = 0\,,$$

se obtiene

$$E_{\boldsymbol{\theta}}((\delta^*(\mathbf{T}) - q(\boldsymbol{\theta}))(\delta(\mathbf{X}) - \delta^*(\mathbf{T}))) = 0\,.$$

Luego (3.17) se transforma en

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) = \mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) + E_{\boldsymbol{\theta}}((\delta(\mathbf{X}) - \delta^*(\mathbf{T}))^2)$$

y resulta

$$\mathrm{ECM}_{\boldsymbol{\theta}}(\delta) \geq \mathrm{ECM}_{\boldsymbol{\theta}}(\delta^*) \ .$$

Además igualdad se cumple sólo si $P_{\boldsymbol{\theta}}(\delta(\mathbf{X}) = \delta^*(\mathbf{T})) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta$.

Luego ya se ha demostrado (i) y (ii). Para mostrar (iii) supongamos que δ es insesgado, luego se tiene

$$E_{\boldsymbol{\theta}}(\delta^*(\mathbf{T})) = E_{\boldsymbol{\theta}}(E(\delta(\mathbf{X})|\mathbf{T})) = E_{\boldsymbol{\theta}}(\delta(\mathbf{X})) = q(\boldsymbol{\theta})$$

Luego se cumple (iii).

Observación: El estimador $\delta^*(\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T})$ es realmente un estimador ya que depende sólo de $\mathbf{T}$ (y por lo tanto de $\mathbf{X}$) y no de $\boldsymbol{\theta}$, ya que por ser $\mathbf{T}$ un estadístico suficiente la distribución de $\delta(\mathbf{X})$ condicional $\mathbf{T} = \mathbf{t}$ es independiente de $\boldsymbol{\theta}$, por lo tanto lo mismo sucede con la esperanza condicional.

Ejemplo 1: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución $Bi(\theta, 1)$. Luego $\delta(X_1, \ldots, X_n) = X_1$ es un estimador insesgado de θ. Un estadístico suficiente para θ es $T = \sum_{i=1}^n X_i$ (ver ejemplo 1 de 3.7). Por lo tanto, de acuerdo al teorema de Rao–Blackwell, $\delta^*(T) = E(\delta(X_1, \ldots, X_n)|T)$ será otro estimador insesgado de θ y $\mathrm{Var}_\theta(\delta^*) \leq \mathrm{Var}_\theta(\delta)$. Vamos a calcular entonces $\delta^*(T)$.

Por ser $X_1, X_2, \ldots, X_n$ idénticamente distribuídas y como T es invariante por permutaciones entre $X_1, X_2, \ldots, X_n$, la distribución conjunta de (X_i, T) es la misma para todo i. Por lo tanto, $E(X_i|T)$ será independiente de i (ver Problema 1 de 3.9). Luego

$$E(X_i|T) = E(X_1|T) = \delta^*(T) \qquad 1 \leq i \leq n \ .$$

Sumando en i se tiene

$$\sum_{i=1}^n E(X_i|T) = n\, \delta^*(T) \ .$$

Pero además vale que

$$\sum_{i=1}^n E(X_i|T) = E\Big(\sum_{i=1}^n X_i|T\Big) = E(T|T) = T \ ,$$

luego

$$\delta^*(T) = \frac{T}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Es fácil ver que

$$\mathrm{Var}_\theta(\delta^*(T)) \leq \mathrm{Var}_\theta(\delta(X))$$

ya que

$$\mathrm{Var}_\theta(\delta^*(T)) = \theta(1-\theta)/n \quad \text{y} \quad \mathrm{Var}_\theta(\delta(X)) = \theta(1-\theta) \ .$$

3.10 Familias exponenciales

Definición: Se dice que una familia de distribuciones continuas o discretas en $I\!R^q$, $F(\mathbf{x}, \boldsymbol{\theta})$, donde $\mathbf{x} = (x_1, \ldots, x_q)$ y $\boldsymbol{\theta} \in \Theta \subset I\!R^p$ es una *familia exponencial a k parámetros* si la correspondiente función de densidad discreta o continua se puede escribir como

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{\sum_{i=1}^{k} c_i(\boldsymbol{\theta}) r_i(\mathbf{x})} h(\mathbf{x}) \tag{3.18}$$

donde $c_1(\boldsymbol{\theta}), \ldots, c_k(\boldsymbol{\theta})$ son funciones de Θ en $I\!R$, $A(\boldsymbol{\theta})$ es una función de Θ en $I\!R^+$ (reales no negativos), $r_1(\mathbf{x}), \ldots, r_k(\mathbf{x})$ son funciones de $I\!R^q$ en $I\!R$ y $h(\mathbf{x})$ es una función de $I\!R^q$ en $I\!R^+$.

Ejemplo 1: Sea la familia $Bi(\theta, n)$ con n fijo y θ en (0,1). Luego

$$\begin{aligned} p(x,\theta) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} = (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^n \binom{n}{x} \quad x = 0, 1, \ldots, n \\ &= (1-\theta)^n e^{x \ln(\theta/(1-\theta))} \binom{n}{x} \end{aligned}$$

Luego esta familia es exponencial a un parámetro con $A(\theta) = (1-\theta)^n$; $r(x) = x$; $c(\theta) = \ln(\theta/(1-\theta))$ y $h(x) = \binom{n}{x}$.

Ejemplo 2: Sea la familia $N(\mu, \sigma^2)$ con $\mu \in I\!R$ y σ^2 real positivo. Luego, su densidad viene dada por

$$p(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\,\sigma^2}x^2+(\frac{\mu}{\sigma^2})x-\frac{\mu^2}{2\sigma^2}}$$

$$= \frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2)}} e^{(-\frac{1}{2\,\sigma^2})x^2+\frac{\mu}{\sigma^2}x} \tag{3.19}$$

Luego esta es una familia exponencial a dos parámetros con $A(\mu,\sigma^2) = e^{-\mu^2/2\sigma^2}/\sqrt{2\pi\sigma^2}$; $c_1(\mu,\sigma^2) = (-1/2\,\sigma^2)$; $c_2(\mu,\sigma^2) = \mu/\sigma^2$; $r_1(x) = x^2$; $r_2(x) = x$; $h(x) = 1$.

Ejemplo 3: Sea la familia $P(\lambda)$. Se puede mostrar que es exponencial a un parámetro. Ver problema 2.i) de 3.10.

Ejemplo 4: Sea la familia $\varepsilon(\lambda)$. Se puede mostrar que es exponencial a un parámetro. Ver problema 2.ii) de 3.10.

Ejemplo 5: Sea la familia de distribuciones normales bivariadas $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$. Es exponencial a 5 parámetros. Ver problema 2.iii) de 3.10.

Teorema 1: *Una familia exponencial a k parámetros cuya función de densidad viene dada por (3.18) tiene como estadístico suficiente para* $\boldsymbol{\theta}$ *el vector* $\mathbf{T} = \mathbf{r}(\mathbf{X}) = (r_1(\mathbf{X}),\ldots,r_k(\mathbf{X}))$.

Demostración. Inmediata a partir del Teorema 1 de 3.9.

El siguiente teorema establece la propiedad más importante de las familias exponenciales.

Teorema 2: *Sea* $\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n$ *una muestra aleatoria de una distribución que pertenece a una familia exponencial a k parámetros, cuya función de densidad viene dada por (3.18). Luego la distribución conjunta de* $\mathbf{X}_1,\ldots,\mathbf{X}_n$ *también pertenece a una familia exponencial a k parámetros y el estadístico suficiente para* $\boldsymbol{\theta}$ *es el vector*

$$\mathbf{T}^* = (T_1^*,\ldots,T_k^*),\ \textit{donde}\ T_i^* = \sum_{j=1}^{n} r_i(\mathbf{X}_j), \quad 1 \le i \le k$$

Demostración: Es inmediata, ya que por (3.18) se tiene

$$p(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n,\boldsymbol{\theta}) = \textstyle\prod_{j=1}^{n} p(\mathbf{x}_i,\boldsymbol{\theta})$$
$$= (A(\boldsymbol{\theta}))^n e^{c_1(\boldsymbol{\theta})\Sigma_{i=1}^n r_1(\mathbf{x}_j)+\cdots+c_k(\boldsymbol{\theta})\Sigma_{i=1}^n r_k(\mathbf{x}_j)} \textstyle\prod_{j=1}^{n} h(\mathbf{x}_j)$$
$$= A^*(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta})r_1^*(\mathbf{x}_1,\ldots,\mathbf{x}_n)+\cdots+c_k(\boldsymbol{\theta})r_k^*(\mathbf{x}_1,\ldots,\mathbf{x}_n)} h^*(\mathbf{x}_1,\ldots,\mathbf{x}_n)$$

donde $A^*(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^n$; $r_i^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{j=1}^n r_i(\mathbf{x}_j)$, $h^*(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^n h(\mathbf{x}_j)$, y por lo tanto el Teorema 2 queda demostrado.

Este último Teorema nos afirma que para familias exponenciales de k parámetros, cualquiera sea el tamaño de la muestra, siempre existe un estadístico suficiente de sólo k componentes. Es decir, que toda la información se puede resumir en k variables aleatorias. Se puede mostrar que esta propiedad bajo condiciones generales caracteriza a las familias exponenciales. Para esta caracterización se puede consultar Sección 2.5 de Zacks [7] y Dynkin [3].

Ejemplo 3: Volvamos al ejemplo 1. Supongamos que tomamos una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución $Bi(\theta, n)$ con n fijo. Luego la distribución conjunta de la muestra pertenecerá a una familia exponencial a un parámetro con estadístico suficiente $T = \sum_{i=1}^n X_i$.

Ejemplo 4: Sea $X_1, \ldots, X_n$ una muestra de una distribución perteneciente a la familia $N(\mu, \sigma^2)$. Luego, de acuerdo a lo visto en el ejemplo 2 y al teorema 2, la distribución conjunta de $X_1, X_2, \ldots, X_n$ pertenece a una familia exponencial a dos parámetros y con estadístico suficiente $T = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i \right)$.

El siguiente teorema establece que las familias de distribuciones de los estadísticos suficientes de una familia exponencial a k parámetros también forma una familia exponencial a k parámetros.

Teorema 3: *Sea* $\mathbf{X}$ *un vector cuya distribución pertenece a una familia exponencial a k parámetros cuya función de densidad satisface (3.18). Luego la función de densidad de los estadísticos suficientes* $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ *es de la forma*

$$p_{\mathbf{T}}(t_1, t_2, \ldots, t_k, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) t_1 + \cdots + c_k(\boldsymbol{\theta}) t_k} h^*(t_1, \ldots, t_k)$$

Por lo tanto la familia de distribuciones de $\mathbf{T}$ *también forma una familia exponencial a k parámetros.*

DEMOSTRACIÓN: Sólo se hará para el caso discreto. Para el caso general se puede consultar Lema 8 de 2.7 en Lehmann [4]. En el caso particular elegido se tiene:

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{\sum_{j=1}^k c_j(\boldsymbol{\theta}) r_j(\mathbf{x})} h(\mathbf{x})$$

Luego si $\mathbf{T} = r(\mathbf{x}) = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ y si $\mathbf{t} = (t_1, \ldots, t_k)$, se tendrá

$$\begin{aligned} p_{\mathbf{T}}(\mathbf{t}, \boldsymbol{\theta}) &= \sum_{\{\mathbf{x}:\, r(\mathbf{x})=\mathbf{t}\}} p(\mathbf{x}, \boldsymbol{\theta}) = \sum_{\{\mathbf{x}:\, r(\mathbf{x})=\mathbf{t}\}} A(\boldsymbol{\theta}) e^{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) r_j(\mathbf{x})} h(\mathbf{x}) \\ &= A(\boldsymbol{\theta}) e^{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) t_j} \sum_{\{\mathbf{x}:\, r(\mathbf{x})=\mathbf{t}\}} h(\mathbf{x}) = A(\boldsymbol{\theta}) e^{\sum_{j=1}^{k} c_j(\boldsymbol{\theta}) t_j} h^*(\mathbf{t}) \end{aligned}$$

con $h^*(\mathbf{t}) = \sum_{\{\mathbf{x}: r(\mathbf{x})=\mathbf{t}\}} h(\mathbf{x})$.

El siguiente lema es de carácter técnico y nos será útil en lo que sigue.

Lema 1: *Sea $\mathbf{X} = (X_1, \ldots, X_q)$ un vector aleatorio cuya distribución pertenece a una familia exponencial a un parámetro discreta o continua con densidad dada por $p(\mathbf{x}, \boldsymbol{\theta}) = A(\theta) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x})$; con $\theta \in \Theta$, donde Θ es un abierto en $\mathbb{R}$ y $c(\theta)$ infinitamente derivable. Luego, si $m(\mathbf{x})$ es un estadístico tal que*

$$\int \ldots \int |m(\mathbf{x})| p(\mathbf{x}, \theta) dx_1 \ldots dx_q < \infty \quad \forall\, \theta \in \Theta$$

o

$$\Sigma_{x_1} \ldots \Sigma_{x_q} |m(\mathbf{x})| p(\mathbf{x}, \theta) < \infty$$

según sea $\mathbf{X}$ continua o discreta, entonces las expresiones

$$\int \ldots \int m(\mathbf{x}) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \ldots dx_q \quad o \quad \Sigma_{x_1} \ldots \Sigma_{x_q} m(\mathbf{x}) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x})$$

según corresponda, son infinitamente derivables y se puede derivar dentro de los signos integral o sumatoria, respectivamente.

Demostración: No se dará en este curso, puede consultarse en el Teorema 9 de 2.7 de Lehmann [4].

Teorema 4: *Sea $\mathbf{X} = (X_1, \ldots, X_q)$ un vector aleatorio cuya distribución pertenece a una familia exponencial a un parámetro con densidad dada por $p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x})$ con $\theta \in \Theta$, donde Θ es un abierto en $\mathbb{R}$ y $c(\theta)$ es infinitamente derivable. Luego se tiene:*

(i) $A(\theta)$ es infinitamente derivable.

(ii)

$$E_\theta(r(\mathbf{X})) = -\frac{A'(\theta)}{A(\theta) c'(\theta)}$$

(iii)

$$Var_\theta(r(\mathbf{x})) = \frac{\frac{\partial E_\theta(r(\mathbf{x}))}{\partial \theta}}{c'(\theta)}$$

DEMOSTRACIÓN: Supongamos que $\mathbf{X}$ sea continuo. El caso discreto es totalmente similar. Como

$$\int \dots \int A(\theta) e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \dots dx_q = 1$$

se tiene

$$\frac{1}{A(\theta)} = \int \dots \int e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \dots dx_q$$

Como el segundo miembro de esta igualdad satisface las condiciones del Lema 1 con $m(\mathbf{x}) = 1$, resulta infinitamente derivable y luego también $A(\theta)$, con lo cual queda demostrado (i).

Por otro lado se tiene

$$A(\theta) \int \dots \int e^{c(\theta) r(x)} h(x) dx_1 \dots dx_q = 1 \quad \forall\, \theta \in \Theta$$

y usando el Lema 1 que nos permite derivar dentro del signo integral resulta

$$A'(\theta) \int \dots \int e^{c(\theta) r(\mathbf{x})} h(\mathbf{x}) dx_1 \dots dx_q +$$
$$A(\theta) c'(\theta) \int \dots \int r(\mathbf{x}) e^{c(\theta) r(\mathbf{x})} dx_1 \dots dx_q = 0$$

y esta última ecuación se puede escribir

$$\frac{A'(\theta)}{A(\theta)} + c'(\theta) E_\theta(r(\mathbf{x})) = 0$$

y luego

$$E_\theta(r(\mathbf{x})) = -\frac{A'(\theta)}{c'(\theta) A(\theta)}$$

y se ha demostrado (ii).

(iii) se deja para resolver en el Problema 3 de 3.10.

3.11 Estadísticos completos

Sea $\mathbf{X}$ un vector aleatorio cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Hasta ahora hemos visto que tomando estimadores insesgados de una función $q(\boldsymbol{\theta})$ basados en estadísticos suficientes se logra mejorar la estimación. Lo que no conocemos es si puede haber más de un estimador insesgado, basado en un estadístico suficiente $\mathbf{T}$ dado. Veremos que bajo ciertas condiciones hay uno solo.

Definición 1: Sea $\mathbf{X}$ un vector aleatorio cuya distribución pertenece a una familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Un estadístico $\mathbf{T} = r(\mathbf{X})$ se dice completo si $E_\theta(g(\mathbf{T})) = 0$ para todo $\boldsymbol{\theta}$ implica que $P_{\boldsymbol{\theta}}(g(\mathbf{T}) = 0) = 1$ para todo $\boldsymbol{\theta} \in \Theta$

Ejemplo 1: Sea X una variable aleatoria con distribución $Bi(\theta, k)$ con k fijo y $0 \leq \theta \leq 1$. Sea g tal que $E_\theta(g(X)) = 0$, para todo θ. Mostraremos que $g(x) = 0$, $x = 0, 1, \ldots, k$. Tenemos

$$E_\theta(g(X)) = \sum_{x=0}^{k} g(x) \binom{k}{x} \theta^x (1-\theta)^{k-x} = 0 \quad \forall\, \theta \in [0,1] \tag{3.20}$$

Sea $\lambda = \theta/(1-\theta)$; luego cuando $\theta \in [0,1]$, λ toma los valores en $\mathbb{R}^+$ (reales no negativos).

Poniendo (3.20) en función de λ resulta

$$(1-\theta)^k \sum_{x=0}^{k} g(x) \binom{k}{x} \lambda^x = 0 \quad \forall\, \lambda \in \mathbb{R}^+$$

Luego

$$Q(\lambda) = \sum_{x=0}^{k} g(x) \binom{k}{x} \lambda^x = 0 \quad \forall\, \lambda \in \mathbb{R}^+$$

Pero $Q(\lambda)$ es un polinomio de grado k con infinitas raíces, luego todos sus coeficientes deben ser 0. Por lo tanto,

$$g(x) \binom{k}{x} = 0 \qquad x = 0, 1, \ldots, k\ ,$$

y entonces

$$g(x) = 0 \qquad x = 0, 1, \ldots, k\ .$$

Con lo que queda probado que $T(X) = X$ es un estadístico completo.

Ejemplo 2: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución que pertenece a la familia $Bi(\theta, k)$. Sea $T = r(X_1, \ldots, X_n) = X_1 + X_2 + \cdots + X_n$. Luego T es un estadístico suficiente y tiene distribución $Bi(\theta, nk)$, por lo tanto de acuerdo a lo visto en el ejemplo 1 es completo.

Ejemplo 3: Consideremos una variable X con distribución $U[0, \theta]$, $\theta \in I\!R^+$. Sea $T = X$. Luego se puede demostrar que T es un estadístico completo. La demostración de este hecho está fuera de los alcances de este curso. De todos modos, veremos una proposición más débil relacionada con completitud. Sea g de $I\!R^+$ en $I\!R$ una función continua. Luego veremos que si $E_\theta(g(X)) = 0$ para todo θ en $I\!R^+$, entonces $g(x) = 0$

$$E_\theta(g(X)) = \frac{1}{\theta)} \int_0^\theta g(x)dx = 0, \quad \forall\, \theta \geq 0,$$

luego

$$\int_0^\theta g(x)dx = 0, \quad \forall\, \theta \in I\!R^+$$

Sea $G(\theta) = \int_0^\theta g(x)dx$, entonces se tiene

$$G(\theta) = 0 \qquad \forall\, \theta \in I\!R^+$$

Usando el Teorema Fundamental del Cálculo Integral se tiene que

$$\frac{\partial G(\theta)}{\partial \theta} = g(\theta) = 0 \quad \forall\, \theta \in I\!R^+$$

Lo que faltaría ver es que en el caso en que g no es continua, $E_\theta(g(X)) = 0 \quad \forall\, \theta \in I\!R^+$ implica $g(x) = 0$ con probabilidad 1.

El siguiente teorema muestra que bajo condiciones muy generales el estadístico suficiente correspondiente a una familia exponencial es completo.

Teorema 1: *Sea una familia exponencial a k parámetros, discreta o continua con función de densidad dada por*

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) r_1(\mathbf{x}) + \ldots + c_k(\boldsymbol{\theta}) r_k(\mathbf{x})} h(\mathbf{x})$$

y sea $\Lambda = \{\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k) : \lambda_i = c_i(\boldsymbol{\theta});\ \boldsymbol{\theta} \in \Theta\}$.

a) Si Λ *contiene* $k + 1$ *puntos* $\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(k+1)}$ *tales que* $\{\boldsymbol{\lambda}^{(j)} - \boldsymbol{\lambda}^{(1)}\ , 2 \leq j \leq k + 1\}$ *son linealmente independientes, entonces el estadístico suficiente* $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ *es minimal suficiente.*

b) Si Λ un conjunto que contiene una esfera en $\mathbb{R}^k$, entonces estadístico suficiente $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$ es completo.

DEMOSTRACIÓN: a) Como $\mathbf{T}$ es suficiente para $\mathcal{F} = \{p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) r_1(\mathbf{x}) + \ldots + c_k(\boldsymbol{\theta}) r_k(\mathbf{x})} h(\mathbf{x}) \quad \boldsymbol{\theta} \in \Theta\}$, de acuerdo al Teorema 2 de la sección 3.8 bastará probar que $\mathbf{T}$ es minimal suficiente para una subfamilia finita de $\mathcal{F}$. Sean $\boldsymbol{\theta}^{(j)}$, $1 \leq j \leq k+1$, tales que

$$\boldsymbol{\lambda}^{(j)} = (\lambda_1^{(j)}, \ldots, \lambda_k^{(j)}) = \left(c_1(\boldsymbol{\theta}^{(j)}), \ldots, c_k(\boldsymbol{\theta}^{(j)}) \right) .$$

Consideremos la subfamilia

$$\begin{aligned} \mathcal{F}_0 = \{p(\mathbf{x}, \boldsymbol{\theta}^{(j)}) &= A(\boldsymbol{\theta}^{(j)}) e^{\sum_{i=1}^k c_i(\boldsymbol{\theta}^{(j)}) r_i(\mathbf{x})} h(\mathbf{x}) \\ &= A(\boldsymbol{\theta}^{(j)}) e^{\sum_{i=1}^k \lambda_i^{(j)} r_i(\mathbf{x})} h(\mathbf{x}) \; 1 \leq j \leq k+1\} . \end{aligned}$$

Luego, por el Teorema 1 de la sección 3.8 un estadístico minimal suficiente para $\mathcal{F}_0$ está dado por

$$\begin{aligned} \mathbf{T}^* &= r^*(\mathbf{x}) = \left(\frac{p(\mathbf{x}, \boldsymbol{\theta}^{(2)})}{p(\mathbf{x}, \boldsymbol{\theta}^{(1)})}, \ldots, \frac{p(\mathbf{x}, \boldsymbol{\theta}^{(k+1)})}{p(\mathbf{x}, \boldsymbol{\theta}^{(1)})} \right) \\ &= \left(\frac{A(\boldsymbol{\theta}^{(2)}) e^{\lambda_1^{(2)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(2)} r_k(\mathbf{x})}}{A(\boldsymbol{\theta}^{(1)}) e^{\lambda_1^{(1)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(1)} r_k(\mathbf{x})}}, \ldots, \frac{A(\boldsymbol{\theta}^{(k+1)}) e^{\lambda_1^{(k+1)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(k+1)} r_k(\mathbf{x})}}{A(\boldsymbol{\theta}^{(1)}) e^{\lambda_1^{(1)} r_1(\mathbf{x}) + \ldots + \lambda_k^{(1)} r_k(\mathbf{x})}} \right) \end{aligned}$$

que es equivalente a

$$\mathbf{T}^{**} = r^{(**)}(\mathbf{x}) = \left(\sum_{i=1}^k [\lambda_i^{(2)} - \lambda_i^{(1)}] r_i(\mathbf{x}), \ldots, \sum_{i=1}^k [\lambda_i^{(k+1)} - \lambda_i^{(1)}] r_i(\mathbf{x}) \right) .$$

Como $\mathbf{T}^{**} = M\mathbf{T}$ donde la matriz $M \in \mathbb{R}^{k \times k}$ es no singular, ya que su $j-$ésima columna es el vector $\boldsymbol{\lambda}^{(j+1)} - \boldsymbol{\lambda}^{(1)}$, $\mathbf{T}$ es equivalente a $\mathbf{T}^{**}$ y por lo tanto, es minimal suficiente para $\mathcal{F}_0$, de donde se obtiene el resultado.
b) Para una demostración general se puede ver Teorema 1 de Sección 4.3 de Lehmann [4]. En este curso sólo se demostrará para el caso que $k = 1$, y que $T = r(\mathbf{X})$ toma un número finito de valores racionales. De acuerdo al teorema 3, en este caso la función de densidad de T será de la forma:

$$p(t, \theta) = A(\boldsymbol{\theta}) e^{c(\boldsymbol{\theta}) t} h(t)$$

Supongamos que los posibles valores de T que tienen probabilidad positiva es el conjunto $A = \{t_1, t_2, \ldots, t_r\} \cup \{-t'_1, -t'_2, \ldots, -t'_s\}$ donde los t_i y los t'_j son racionales no negativos.

Sea v un múltiplo común de los denominadores de todos los racionales t_i y t'_j y sean $w_i = vt_i \quad 1 \leq i \leq r$ y $w'_i = vt'_i$, $1 \leq i \leq s$. Luego los w_i y los w'_i son naturales. Finalmente sea $w = \max_{1 \leq i \leq s} w'_i$, $z_i = w_i + w, \quad 1 \leq i \leq r$ y $z'_i = -w'_i + w$, $1 \leq i \leq s$. Luego los z_i y los z'_i son naturales y todos diferentes.

Supongamos que

$$E_{\boldsymbol{\theta}}(g(T)) = 0 \qquad \forall\, \boldsymbol{\theta} \in \Theta$$

luego

$$\sum_{i=1}^{r} g(t_i)p(t_i, \boldsymbol{\theta}) + \sum_{i=1}^{s} g(-t'_i)p(-t'_i, \boldsymbol{\theta}) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta$$

con lo cual

$$\sum_{i=1}^{r} g(t_i)A(\boldsymbol{\theta})e^{c(\boldsymbol{\theta})t_i}h(t_i) + \sum_{i=1}^{s} g(-t'_i)A(\boldsymbol{\theta})e^{-c(\boldsymbol{\theta})t'_i}h(-t'_i) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta\,,$$

de donde se obtiene

$$\sum_{i=1}^{r} g(t_i)h(t_i)(e^{c(\boldsymbol{\theta})/v})^{t_i v} + \sum_{i=1}^{s} g(-t'_i)h(-t'_i)(e^{c(\boldsymbol{\theta})/v})^{-t'_i v} = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta\,.$$

Llamando $\lambda = e^{c(\boldsymbol{\theta})/v}$ resulta que como hay infinitos posibles valores de $c(\boldsymbol{\theta})$, el conjunto Λ de posibles valores de λ, también es infinito. Luego tenemos

$$\sum_{i=1}^{r} g(t_i)h(t_i)\lambda^{w_i} + \sum_{i=1}^{s} g(-t'_i)h(-t'_i)\lambda^{-w'_i} = 0 \quad \forall\, \lambda \in \Lambda$$

Multiplicando por λ^w la última ecuación resulta

$$P(\lambda) = \sum_{i=1}^{r} g(t_i)h(t_i)\lambda^{z_i} + \sum_{i=1}^{s} g(t'_i)h(-t'_i)\lambda^{z'_i} = 0 \quad \forall\, \lambda \in \Lambda$$

Luego el polinomio $P(\lambda)$ tiene infinitas raíces y por lo tanto, todos los coeficientes deben ser 0, es decir, $g(t_i)h(t_i) = 0, \quad 1 \leq i \leq r$ y $g(-t'_i)h(-t'_i) = 0, \quad 1 \leq i \leq s$. Como $h(t_i) > 0, \quad 1 \leq i \leq r$ y $h(-t'_i) > 0, \quad 1 \leq i \leq s$,

resulta que $g(t_i) = 0 \quad 1 \leq i \leq r$ y $g(-t_i') = 0 \quad 1 \leq i \leq s$. Con lo cual, $P_{\boldsymbol{\theta}}(g(T) = 0) = 1$ para todo $\boldsymbol{\theta} \in \Theta$.

Ejemplo 4: Sea X_1 una variable $N(\mu, \sigma_1^2)$ y X_2 independiente de X_1 una variable $N(\mu, \sigma_2^2)$, luego si $\boldsymbol{\theta} = (\mu, \sigma_1^2, \sigma_2^2)$ la densidad de $\mathbf{X} = (X_1, X_2)$ puede escribirse como

$$p(x_1, x_2, \boldsymbol{\theta}) = \frac{1}{2\,\pi\sigma_1\sigma_2} e^{-\mu^2(\frac{1}{2\,\sigma_1^2}-\frac{1}{2\sigma_2^2})} e^{(-\frac{1}{2\,\sigma_1^2})x_1^2+(-\frac{1}{2\,\sigma_2^2})x_2^2+(\frac{\mu}{\sigma_1^2})x_1+(\frac{\mu}{\sigma_2^2})x_2}$$

Por lo tanto es una familia exponencial a 4 parámetros, pero no satisface la condición del Teorema 1 ya que el conjunto

$$\Lambda = \{\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4) \text{ con } \lambda_1 = -\frac{1}{2\,\sigma_1^2}\,, \lambda_2 = -\frac{1}{2\,\sigma_2^2}\;\lambda_3 = \frac{\mu}{\sigma_1^2}\;\lambda_4 = \frac{\mu}{\sigma_2^2}\}\,,$$

está en una superficie de dimensión 3, ya que depende de 3 parámetros, σ_1^2, σ_2^2 y μ, y por lo tanto no contiene ninguna esfera de $\mathbb{R}^4$. Como el Teorema 1 de la sección 3.11 da un condición suficiente pero no necesaria para completitud, no se deduce que $\mathbf{T} = (X_1, X_2, X_1^2, X_2^2)$ no sea completo. Sin embargo, dado que $E_{\mu,\sigma_1^2,\sigma_2^2}(X_1 - X_2) = \mu - \mu = 0$ y $X_1 - X_2$ no es igual a 0 resulta que $\mathbf{T}$ no es completo.

El Teorema 1 nos permite, sin embargo, deducir que $\mathbf{T}$ es minimal suficiente.

Por lo tanto, hemos visto un estadístico minimal suficiente no necesariamente es completo. El siguiente resultado establece la recíproca.

Teorema 2: *Sea* $\mathbf{T}$ *un estadístico suficiente y completo para* $\boldsymbol{\theta}$*. Si existe un estadístico minimal suficiente para* $\boldsymbol{\theta}$ *entonces* $\mathbf{T}$ *es minimal suficiente.*
DEMOSTRACIÓN. La haremos sólo en el caso en que el estadístico minimal suficiente y el estadístico suficiente y completo T tienen dimensión 1. Sea U el estadístico minimal suficiente para $\boldsymbol{\theta}$, luego por ser T suficiente se cumple que $U = m(T)$. Queremos ver que m es biunívoca.

Sea $\psi(t)$ la función arcotangente. Luego $\psi : \mathbb{R} \to [0, 2\pi]$ es una función estrictamente creciente y acotada. Por lo tanto, $E_{\boldsymbol{\theta}}(\psi(T)) < \infty$ y bastará mostrar que $\psi(T)$ es función de U.

Definamos $\eta(U) = E(\psi(T)|U)$. Como U es suficiente $\eta(U)$ es un estadístico. Luego, si

$$g(T) = \psi(T) - \eta[m(T)] = \psi(T) - \eta(U)$$

se cumple que $E_{\boldsymbol{\theta}}[g(T)] = 0$ para todo $\boldsymbol{\theta} \in \Theta$. Por lo tanto, $P_{\boldsymbol{\theta}}(\psi(T) = \eta(U)) = 1$ para todo $\boldsymbol{\theta} \in \Theta$, y entonces T es equivalente a U.

El siguiente Teorema es útil en muchas situaciones, donde probar independencia entre estadísticos puede resultar laborioso.

Teorema 3: (Teorema de Basu) *Sea* $\mathbf{T}$ *un estadístico suficiente y completo para* $\boldsymbol{\theta}$. *Sea* $\mathbf{U} = g(\mathbf{X})$ *un estadístico cuya distribución no depende de* $\boldsymbol{\theta}$ *entonces* $\mathbf{U}$ *es independiente de* $\mathbf{T}$.
DEMOSTRACIÓN. Sea A un suceso, como $\mathbf{U}$ tiene distribución independiente de $\boldsymbol{\theta}$, $p_A = P(\mathbf{U} \in A)$ no depende de $\boldsymbol{\theta}$.
Sea $\eta_A(\mathbf{t}) = P(\mathbf{U} \in A|\mathbf{T} = \mathbf{t})$. Como $\mathbf{T}$ es suficiente $\eta_A(\mathbf{T})$ es un estadístico. Por otra parte, $E_{\boldsymbol{\theta}}(\eta_A(\mathbf{T}) - p_A) = 0$ para todo $\boldsymbol{\theta} \in \Theta$, con lo cual la completitud de $\mathbf{T}$ implica que $P_{\boldsymbol{\theta}}(\eta_A(\mathbf{T}) = p_A) = 1$ para todo $\boldsymbol{\theta} \in \Theta$ y por lo tanto, $\mathbf{U}$ es independiente de $\mathbf{T}$.

3.12 Estimadores insesgados de mínima varianza uniformemente

El siguiente teorema nos da un método para construir estimadores IMVU cuando se conoce un estadístico que es a la vez suficiente y completo.

Teorema 1 (Lehmann-Scheffé): *Sea* $\mathbf{X}$ *un vector aleatorio de cuya distribución pertenece a la familia* $F(\mathbf{x}, \boldsymbol{\theta})$ *con* $\boldsymbol{\theta} \in \Theta$. *Sea* $\mathbf{T}$ *un estadístico suficiente y completo. Luego dada una función* $q(\boldsymbol{\theta})$ *de* Θ *en* $\mathbb{R}$, *se tiene que*

(i) *Existe a lo sumo un estimador insesgado de* $q(\boldsymbol{\theta})$, *basado en* $\mathbf{T}$.

(ii) *Si* $\delta(\mathbf{T})$ *es un estimador insesgado de* $q(\boldsymbol{\theta})$, *entonces* $\delta(\mathbf{T})$ *es IMVU.*

(iii) *Si* $\delta(\mathbf{X})$ *es un estimador insesgado para* $q(\boldsymbol{\theta})$, *luego* $\delta^*(\mathbf{T}) = E(\delta(\mathbf{X})|\mathbf{T})$ *es un estimador IMVU para* $q(\boldsymbol{\theta})$.

DEMOSTRACIÓN:

(i) Sean $\delta_1(\mathbf{T})$ y $\delta_2(\mathbf{T})$ dos estimadores insesgados de $q(\boldsymbol{\theta})$. Luego

$$E_{\boldsymbol{\theta}}(\delta_1(\mathbf{T}) - \delta_2(\mathbf{T})) = q(\boldsymbol{\theta}) - q(\boldsymbol{\theta}) = 0 \quad \forall\, \boldsymbol{\theta} \in \Theta$$

luego como $\mathbf{T}$ es completo

$$P_{\boldsymbol{\theta}}(\delta_1(\mathbf{T}) - \delta_2(\mathbf{T}) = 0) = 1, \ \forall \boldsymbol{\theta} \in \Theta$$

(ii) Sea $\delta(\mathbf{T})$ un estimador insesgado de $q(\boldsymbol{\theta})$, y sea $\delta_1(\mathbf{X})$ otro estimador insesgado. Si llamamos $\delta_1^*(\mathbf{T}) = E(\delta_1(\mathbf{X})|\mathbf{T})$ sabemos por el Teorema 1 de la sección 3.9 que $\delta_1^*(\mathbf{T})$ es insesgado y

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta_1^*) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1) \quad \forall\, \boldsymbol{\theta} \in \Theta \tag{3.21}$$

Pero de acuerdo a (i) se tiene que $\delta_1^*(\mathbf{T}) = \delta(\mathbf{T})$ con probabilidad 1. Luego

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta_1^*) = \mathrm{Var}_{\theta}(\delta)$$

y luego de 3.21 resulta que

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1)$$

y (ii) queda demostrado.

(iii) Como $\delta^*(\mathbf{T})$ es por el Teorema 1 de la sección 3.9 insesgado, de (ii) se deduce que es un estimador IMVU para $q(\boldsymbol{\theta})$.

De acuerdo al punto (ii) de este teorema, en el caso de tener un estadístico suficiente y completo $\mathbf{T}$, cualquier estimador insesgado basado en $\mathbf{T}$ es un estimador IMVU. El punto (iii) nos indica cómo construir un estimador IMVU de $q(\boldsymbol{\theta})$ a partir de cualquier estimador insesgado.

Teorema 2: *Sea* $\mathbf{X}$ *un vector aleatorio cuya distribución pertenece a una familia exponencial a* k *parámetros con función de densidad dada por*

$$p(\mathbf{x}, \boldsymbol{\theta}) = A(\boldsymbol{\theta}) e^{c_1(\boldsymbol{\theta}) r_1(\mathbf{x}) + \cdots + c_k(\boldsymbol{\theta}) r_k(\mathbf{x})} h(\mathbf{x})$$

donde $\boldsymbol{\theta}$ *toma valores en el conjunto* Θ*. Supongamos además que*

$$\Lambda = \{\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_k) : \lambda_i = c_i(\boldsymbol{\theta}); \ \boldsymbol{\theta} \in \Theta\}$$

contiene una esfera en $\mathbb{R}^k$*. Sea* $\mathbf{T} = (r_1(\mathbf{X}), \ldots, r_k(\mathbf{X}))$*, luego si* $\delta(\mathbf{T})$ *es un estimador insesgado de* $q(\boldsymbol{\theta})$*, entonces* $\delta(\mathbf{T})$ *es un estimador IMVU para* $q(\boldsymbol{\theta})$*.*

Demostración: Inmediata a partir de los Teoremas 3 de sección 3.10 y 1 de sección 3.12.

Ejemplo 1: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $Bi(\theta, k)$ con k fijo. Luego, la distribución conjunta de la muestra viene dada por

$$\begin{aligned} p(x_1, x_2, \ldots, x_n, \theta) &= \binom{k}{x_1}\binom{k}{x_2}\ldots\binom{k}{x_n}\theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{nk-\Sigma_{i=1}^n x_i} \\ &= (1-\theta)^{nk}\, e^{(\Sigma_{i=1}^n x_i)\ln(\theta/(1-\theta))}\binom{k}{x_1}\binom{k}{x_2}\ldots\binom{k}{x_n} \end{aligned}$$

Esta familia constituye una familia exponencial, con estadístico suficiente $T = \sum_{i=1}^n X_i$. Por otro lado $c(\theta) = \ln\theta/(1-\theta)$ toma todos los posibles valores de $\mathbb{R}$ cuando θ varía en el intervalo (0,1). Luego T es un estadístico suficiente y completo. Como $\delta(T) = T/nk$ es un estimador insesgado de θ, resulta un estimador IMVU de θ.

Ejemplo 2: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $N(\mu, \sigma^2)$. Luego usando (3.19) resulta que la distribución conjunta de la muestra viene dada por

$$p(x_1, \ldots, x_n, \mu, \sigma^2) = \frac{1}{2\pi\sigma^2)^{n/2}} e^{-\frac{n\,\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n X_i}$$

Luego constituye una familia exponencial a dos parámetros con estadístico suficiente $\mathbf{T} = \left(\sum_{i=1}^n X_i^2,\ \sum_{i=1}^n X_i\right)$. Mostraremos ahora que $\mathbf{T}$ es completo. Bastará mostrar que

$$\Lambda = \{(\lambda_1, \lambda_2) : \lambda_1 = -\frac{1}{2\,\sigma^2},\ \lambda_2 = \frac{\mu}{\sigma^2},\ \lambda \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^+\}$$

contiene una esfera.

Mostraremos que Λ contiene todo $(\lambda_1, \lambda_2) \in \mathbb{R}^2$ con $\lambda_1 < 0$.

Sea (λ_1, λ_2) con $\lambda_1 < 0$, tenemos que mostrar que viene de un par (μ, σ^2) con $\sigma^2 > 0$. Para ver esto basta tomar $\sigma^2 = -1/2\,\lambda_1$ y $\mu = \lambda_2\sigma^2 = -\lambda_2/2\lambda_1$. Luego $\mathbf{T}$ es completo.

Como $\overline{X}$ es un estimador insesgado de μ, y como depende de $\mathbf{T}$, resulta que es IMVU de μ.

Por otro lado $s^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1) = \left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right)/(n-1)$ es un estimador insesgado de σ^2 y además depende de $\mathbf{T}$, luego es IMVU para σ^2.

Ejemplo 3: Sea X_1 una variable $N(\mu, \sigma_1^2)$ y X_2 independiente de X_1 una variable $N(\mu, \sigma_2^2)$. Vimos en el Ejemplo 4 de la sección 3.11 que

$\mathbf{T} = (X_1, X_2, X_1^2, X_2^2)$ era minimal suficiente pero no era completo. Se puede mostrar que en este caso no hay ningún estimador IMVU (ver Problema 7 de 3.11).

Ejemplo 4: El siguiente ejemplo muestra que no siempre existen estimadores IMVU. Volvamos al ejemplo 1 y supongamos que se quiera estimar $q(\theta)$. Como $T = \sum_{i=1}^{n} X_i$ es un estadístico suficiente, un estimador IMVU de $q(\theta)$ deberá estar basado en T. Supongamos que $\delta(T)$ es un estimador IMVU para $q(\theta)$. Como T tiene distribución $Bi(\theta, kn)$ y $\delta(T)$ es insesgado se tendrá

$$q(\theta) = E_\theta(\delta(T)) = \sum_{i=0}^{kn} \delta(t) \binom{kn}{t} \theta^r (1-\theta)^{kn-t}$$

Luego una condición necesaria para que $q(\theta)$ tenga un estimador IMVU es que sea un polinomio de grado menor o igual a kn. Se puede mostrar que es también una condición suficiente aunque no lo demostraremos.

Por lo tanto no existen estimadores IMVU, por ejemplo, para e^θ, $\ln\theta$, sen θ. Esto no quiere decir que no existen buenos estimadores. Si $q(\theta)$ es continua, un buen estimador será $\delta(T) = q(T/nk)$ ya que T/nk es un estimador IMVU de θ.

Ejemplo 5: En este ejemplo veremos que un estimador IMVU puede ser mejorado en su error cuadrático medio por otro estimador no insesgado. Volvamos al ejemplo 2 y supongamos que se desea estimar σ^2. Hemos visto que un estimador IMVU para σ^2 es $s^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$, sin embargo veremos que s^2 no es admisible.
Sea $\widehat{\sigma}_c^2 = c\,U$ donde $U = \sum_{i=1}^{n}(X_i - \overline{X})^2$. Luego, $s^2 = \widehat{\sigma}^2_{\frac{1}{n-1}}$. Sabemos que U/σ^2 tiene distribución χ^2_{n-1}, por lo tanto, $E_{\sigma^2}(U) = (n-1)\,\sigma^2$ y $\mathrm{Var}_{\sigma^2}(U) = 2\,(n-1)\,\sigma^4$. Con lo cual,

$$\begin{aligned}
\mathrm{ECM}_{\sigma^2}(\widehat{\sigma}_c^2) &= E_{\sigma^2}\left[(\widehat{\sigma}_c^2 - \sigma^2)^2\right] \\
&= \mathrm{Var}_{\sigma^2}(\widehat{\sigma}_c^2) + \left[E_{\sigma^2}(\widehat{\sigma}_c^2) - \sigma^2\right]^2 \\
&= c^2 \mathrm{Var}_{\sigma^2}(U) + \left[c\,E_{\sigma^2}(U) - \sigma^2\right]^2 \\
&= 2\,c^2(n-1)\,\sigma^4 + \left[c\,(n-1)\,\sigma^2 - \sigma^2\right]^2 \\
&= \sigma^4\left[c^2(n+1)(n-1) - 2(n-1)c + 1\right]
\end{aligned}$$

Luego, el ECM de $\widehat{\sigma}_c^2$ es un polinomio de grado 2 en c que alcanza su mínimo cuando $c = 1/(n+1)$. Por lo tanto, $U/(n+1)$ tiene menor ECM que el estimador IMVU s^2.

Cómo caraterizamos los estimadores IMVU cuando no existe un estadístico suficiente y completo?

Lema 1: *Sea δ_0 un estimador insesgado de $q(\theta)$. Dado cualquier otro estimador δ insesgado de $q(\theta)$, se cumple que $\delta = \delta_0 - U$ con $E_\theta(U) = 0$ $\forall \theta \in \Theta$.*

Luego como $ECM_\theta(\delta) = Var_\theta(\delta) = Var_\theta(\delta_0 - U) = E_\theta\{(\delta_0 - U)^2\} - q(\theta)^2$, para encontrar el estimador IMVU basta minimizar $E_\theta\{(\delta_0 - U)^2\}$, o sea, basta encontrar la proyección de δ_0 sobre el espacio de los estimadores del 0.

Teorema 3: *Supongamos que $\mathbf{X}$ es un vector aleatorio de cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Sea $\Delta = \{\delta(\mathbf{X}) : E_\theta \delta^2(\mathbf{X}) < \infty\}$. Sea $\mathcal{U} = \{\{\delta(\mathbf{X}) \in \Delta : E_\theta \delta(\mathbf{X})\} = 0 \ \forall \theta \in \Theta$. Una condición necesaria y suficiente para que $\delta \in \Delta$, insesgado, sea IMVU para $q(\theta)$ es que $E_\theta(\delta U) = 0$, $\forall \theta \in \Theta$, $\forall U \in \mathcal{U}$.*

3.13 Desigualdad de Rao–Cramer

En esta sección mostraremos que bajo hipótesis muy generales, la varianza de un estimador insesgado no puede ser inferior a cierta cota.

Supongamos que $\mathbf{X} = (X_1, \ldots, X_n)$ es un vector aleatorio de cuya distribución pertenece a la familia de distribuciones discreta o continua con densidad $p(\mathbf{x}, \theta)$, con $\theta \in \Theta$; donde Θ es un conjunto abierto de $\mathbb{R}$. Supongamos además que se cumplen las siguientes condiciones (en lo que sigue suponemos que $\mathbf{X}$ es continuo, para el caso discreto habrá que reemplazar todos los signos $\int$ por $\sum$):

(A) El conjunto $S = \{\mathbf{x} : p(\mathbf{x}, \theta) > 0\}$ es independiente de θ.

(B) Para todo $\mathbf{x}$, $p(\mathbf{x}, \theta)$ es derivable respecto de θ.

(C) Si $h(\mathbf{X})$ es un estadístico tal que $E_\theta[|h(\mathbf{X})|] < \infty$ para todo $\theta \in \Theta$ entonces se tiene

$$\frac{\partial}{\partial \theta}\left[\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x}\right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(\mathbf{x}) \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} d\mathbf{x}$$

donde $d\mathbf{x} = (dx_1, \ldots, dx_n)$ (o sea se puede derivar dentro del signo integral)

(D)
$$0 < I(\theta) = E_\theta \left[\left(\frac{\partial \ln p(\mathbf{X}, \theta)}{\partial \theta} \right)^2 \right] < \infty$$

$I(\theta)$ se denomina *número de información de Fisher.*

Lema 1: *Supongamos que se cumplan las condiciones A, B, C y D. Sea* $\psi(\mathbf{x}, \theta) = \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta}$. *Entonces,*

(i) $E_\theta \psi(\mathbf{X}, \theta) = 0$ *y* $Var_\theta \psi(\mathbf{X}, \theta) = I(\theta)$.

(ii) Si además existe la derivada segunda de $p(\mathbf{x}, \theta)$ *respecto de* θ *y si para todo estadístico* $h(\mathbf{X})$ *tal que,* $E_\theta[|h(\mathbf{X})|] < \infty$ *para todo* $\theta \in \Theta$, *se cumple que*

$$\frac{\partial^2}{\partial \theta^2} \left[\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} \right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} h(\mathbf{x}) \frac{\partial^2 p(\mathbf{x}, \theta)}{\partial \theta^2} d\mathbf{x} \tag{3.22}$$

entonces
$$I(\theta) = -E_\theta \frac{\partial^2 \ln p(\mathbf{X}, \theta)}{\partial \theta^2} = -E_\theta \frac{\partial \psi(\mathbf{X}, \theta)}{\partial \theta}$$

DEMOSTRACIÓN: (i) Por ser $p(\mathbf{x}, \theta)$ una densidad, si S es el conjunto definido en la condición (A) se tiene

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(\mathbf{x}, \theta) d\mathbf{x} = \int_S \ldots \int p(\mathbf{x}, \theta) d\mathbf{x} = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} p(\mathbf{x}, \theta) I_S(\mathbf{x}) d\mathbf{x} = 1$$

donde I_S es la función indicadora del conjunto S.
Luego aplicando la condición (C) a $h(\mathbf{x}) = I_S(\mathbf{x})$ se obtiene derivando ambos miembros que

$$\int_{\infty}^{\infty} \ldots \int_{\infty}^{\infty} \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} I_S(\mathbf{x}) d\mathbf{x} = 0,$$

y por lo tanto

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \left[\frac{\partial p(\mathbf{x}, \theta)}{\partial \theta} \Big/ p(\mathbf{x}, \theta) \right] I_S(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} = 0\,.$$

Esta última ecuación es equivalente a

$$\int_{\infty}^{\infty} \cdots \int_{\infty}^{\infty} \left[\frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta} \right] I_S(\mathbf{x}) p(\mathbf{x}, \theta) d\mathbf{x} = 0$$

la cual implica

$$E_\theta \psi(\mathbf{X}, \theta) = E_\theta \left(\frac{\partial \ln p(\mathbf{X}, \theta)}{\partial \theta} \right) = 0 \tag{3.23}$$

Como $I(\theta) = E_\theta \psi^2(\mathbf{X}, \theta)$, (3.23) implica que $\mathrm{Var}_\theta \psi(\mathbf{X}, \theta) = I(\theta)$
(ii) De la igualdad

$$\frac{\partial^2 \ln p(\mathbf{x}, \theta)}{\partial \theta^2} = \frac{\frac{\partial^2 p(\mathbf{x}, \theta)}{\partial \theta^2}}{p(\mathbf{x}, \theta)} - \psi^2(\mathbf{x}, \theta)$$

se obtiene que

$$E_\theta \frac{\partial^2 \ln p(\mathbf{X}, \theta)}{\partial \theta^2} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial^2 p(\mathbf{x}, \theta)}{\partial \theta^2} d\mathbf{x} - E_\theta \psi^2(\mathbf{X}, \theta)\,. \tag{3.24}$$

Utilizando (3.22) con $h(\mathbf{x}) = I_S(\mathbf{x})$ se obtiene que el primer término del lado derecho de (3.24) es igual a cero, de donde el resultado.

Teorema 1 (Rao–Cramer): *Bajo las condiciones A, B, C y D si* $\delta(\mathbf{X})$ *es un estimador insesgado de* $q(\theta)$ *tal que* $E_\theta \delta^2(\mathbf{X}) < \infty$ *se tiene*

(i)
$$\mathrm{Var}_{\,\theta}(\delta(\mathbf{X})) \geq \frac{|q'(\theta)|^2}{I(\theta)}$$

(ii) (i) *vale como igualdad si y sólo si* $\delta(\mathbf{X})$ *es estadístico suficiente de una familia exponencial, es decir si y sólo si*

$$p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta)\delta(\mathbf{x})} h(\mathbf{x}) \tag{3.25}$$

Demostración: (i) Sea $\psi(\mathbf{x}, \theta) = \frac{\partial p(\mathbf{x}, \theta)}{\partial \theta}$. Por el Lema 1 tenemos que

$$E_\theta \psi(\mathbf{X}, \theta) = 0 \quad \text{y} \quad \mathrm{Var}_\theta \psi(\mathbf{X}, \theta) = I(\theta). \tag{3.26}$$

Por otro lado, como $\delta(\mathbf{X})$ es insesgado se tiene

$$E_\theta(\delta(\mathbf{X})) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \delta(\mathbf{x}) p(\mathbf{x}, \theta) I_S(\mathbf{x}) d\mathbf{x} = q(\theta)$$

y luego aplicando la hipótesis C, tomando $h(\mathbf{X}) = \delta(\mathbf{X})I_S(\mathbf{X})$ se obtiene derivando ambos miembros que

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \delta(\mathbf{x}) \frac{\partial p(\mathbf{x},\theta)}{\partial \theta} \, I_S(\mathbf{x}) d\mathbf{x} = q'(\theta)$$

de donde

$$\begin{aligned} E_\theta \left[\delta(\mathbf{X})\psi(\mathbf{X},\theta)\right] &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \delta(\mathbf{x})\psi(\mathbf{x},\theta)p(\mathbf{x},\theta)d\mathbf{x} \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \delta(\mathbf{x}) \frac{\partial \ln p(\mathbf{x},\theta)}{\partial \theta} \, I_S(\mathbf{x})p(\mathbf{x},\theta)d\mathbf{x} \\ &= q'(\theta) \end{aligned} \tag{3.27}$$

Teniendo en cuenta (3.26), (3.27) se puede escribir como

$$\mathrm{Cov}\left(\delta(\mathbf{X}),\, \psi(\mathbf{X},\theta)\right) = q'(\theta) \tag{3.28}$$

De acuerdo a la desigualdad de Cauchy–Schwartz, $\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X), \mathrm{Var}(Y)$, y vale la igualdad si y sólo si $P(Y = aX + b) = 1$ para algunas constantes a y b. Por lo tanto, usando (3.28) resulta

$$\left[q'(\theta)\right]^2 \leq \mathrm{Var}_\theta(\delta(\mathbf{X})) \cdot \mathrm{Var}_\theta\left(\psi(\mathbf{X},\theta)\right) \tag{3.29}$$

y la igualdad vale si y sólo si

$$\frac{\partial \ln p(\mathbf{X},\theta)}{\partial \theta} = \psi(\mathbf{x},\theta) = a(\theta)\delta(\mathbf{x}) + b(\theta) \text{ con probabilidad 1.} \tag{3.30}$$

Usando (3.26) y (3.29) resulta

$$\mathrm{Var}_\theta(\delta(\mathbf{X}) \geq \frac{q'(\theta)^2}{I(\theta)} \tag{3.31}$$

que es lo que se afirma en (i).

(ii) (3.31) valdrá como igualdad si y sólo si cumple (3.30). Mostraremos que (3.30) se cumple si y sólo si se cumple (3.25).

Integrando respecto de θ en (3.30), se obtiene

$$\ln p(\mathbf{x},\theta) = \delta(\mathbf{x}) \int a(\theta)d\theta + g(\mathbf{x}) + \int b(\theta)d\theta$$

que se puede escribir como

$$\ln p(\mathbf{x},\theta) = \delta(\mathbf{x})c(\theta) + g(\mathbf{x}) + B(\theta)$$

donde $c(\theta) = \int a(\theta)d\theta$ y $B(\theta) = \int b(\theta)d\theta$. Luego, despejando $p(\mathbf{x}, \theta)$ resulta

$$p(\mathbf{x}, \theta) = e^{B(\theta)} e^{\delta(\mathbf{x})c(\theta)} e^{g(\mathbf{x})}$$

y llamando $A(\theta) = e^{B(\theta)}$ y $h(\mathbf{x}) = e^{g(\mathbf{x})}$; resulta (3.25).

Supongamos ahora que se cumple (3.25), mostraremos que se cumple (3.30).

Si se cumple (3.25), tomando logaritmos se tiene

$$\ln p(\mathbf{x}, \theta) = \ln A(\theta) + c(\theta)\delta(\mathbf{x}) + \ln h(\mathbf{x})$$

y derivando se obtiene

$$\frac{\partial \ln p(\mathbf{x}, \theta)}{\partial \theta} = \frac{A'(\theta)}{A(\theta)} + c'(\theta)\delta(\mathbf{x})$$

y por lo tanto se cumple (3.30). Esto prueba el punto (ii).

Observación 1: Si $\delta(\mathbf{X})$ es un estimador insesgado de $q(\theta)$ y $\mathrm{Var}_\theta(\delta(\mathbf{x})) = [q'(\theta)]^2/I(\theta)$ para todo $\theta \in \Theta$. Entonces del punto (i) del Teorema 1 resulta que $\delta(\mathbf{X})$ es IMVU. Por lo tanto esto da otro criterio para verificar si un estimador insesgado dado es IMVU.

Observación 2: Si $p(\mathbf{x}, \theta) = A(\theta)e^{c(\theta)\delta(\mathbf{x})}h(\mathbf{x})$, y si $\delta(\mathbf{X})$ es un estimador insesgado de $q(\theta)$, entonces $\delta(\mathbf{X})$ es un estimador IMVU de $q(\theta)$. Esto resulta de (i) y (ii).

Observación 3: Si $\delta(\mathbf{X})$ es un estimador de θ, su varianza debe ser mayor o igual que $1/I(\theta)$. Luego se puede esperar que cuanto mayor sea $I(\theta)$ (como $1/I(\theta)$ será menor) existe la posibilidad de encontrar estimadores con menor varianza y por lo tanto más precisos. De ahí el nombre de "número de información" que se le da a $I(\theta)$. Es decir cuanto mayor es $I(\theta)$, mejores estimadores de θ se pueden encontrar, y por lo tanto se puede decir que más información da el vector $\mathbf{X}$ sobre θ. El hecho de que se pueden encontrar estimadores con varianza aproximadamente igual a $1/I(\theta)$ será cierto para n grande. Para esto consultar sección 3.13 y el apéndice (B) de este capítulo. Para una generalización del Teorema de Rao–Cramer el caso en que θ es un vector puede consultarse el Teorema 4.3.1 de Zacks [7] y el Teorema 7.3 de Lehmann [5].

El siguiente teorema nos indica que una muestra aleatoria de tamaño n $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ de una familia con densidad $p(\mathbf{X}, \boldsymbol{\theta})$ nos da n veces más información que una sola observación.

Teorema 2: *Sea $\mathbf{X}_1, \ldots, \mathbf{X}_n$ una muestra aleatoria de una distribución con densidad $p(\mathbf{x}, \theta)$ con $\theta \in \Theta \subset I\!R$. Luego, si se denomina $I_n(\theta)$ al número de información de $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ y $I_1(\theta)$ al número de información de $\mathbf{X}_1$, entonces se tiene $I_n(\theta) = nI_1(\theta)$.*

DEMOSTRACIÓN: Se tiene que

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n, \theta) = \prod_{i=1}^{n} p(\mathbf{x}_i, \theta)$$

y entonces

$$\ln p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \theta) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i, \theta) \; .$$

Por lo tanto,

$$\frac{\partial \ln p(\mathbf{x}_1, \ldots, \mathbf{x}_n, \theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \ln p(\mathbf{x}_i, \theta)}{\partial \theta}$$

Con lo cual, por ser $\mathbf{X}_1, \ldots, \mathbf{X}_n$ independientes, se tiene

$$I(\theta) = \mathrm{Var}_\theta \left(\frac{\partial \ln p(\mathbf{X}_1, \ldots, \mathbf{X}_n, \theta)}{\partial \theta} \right) = \sum_{i=1}^{n} \mathrm{Var}_\theta \left(\frac{\partial \ln p(\mathbf{X}_i, \theta)}{\partial \theta} \right) = nI_1(\theta) \, .$$

Ejemplo 1: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $Bi(\theta, 1)$. Luego se tiene

$$p(x, \theta) = \theta^x (1 - \theta)^{1-x}$$

luego

$$\ln p(x, \theta) = x \ln \theta + (1 - x) \ln(1 - \theta)$$

y por lo tanto

$$\frac{\partial \ln p(x, \theta)}{\partial \theta} = \frac{x}{\theta} - \frac{1 - x}{1 - \theta} = \frac{x - \theta}{\theta(1 - \theta)}$$

luego

$$\begin{aligned} I_1(\theta) &= E\left(\Big(\frac{\partial \ln p(X_1, \theta)}{\partial \theta} \Big)^2 \right) \\ &= \Big[\frac{1}{\theta(1 - \theta)} \Big]^2 E_\theta (X - \theta)^2 \\ &= \Big[\frac{1}{\theta(1 - \theta)} \Big]^2 \mathrm{Var}_\theta (X - \theta)^2 \\ &= \frac{1}{\theta(1 - \theta)} \end{aligned}$$

y por lo tanto,

$$I_n(\theta) = \frac{n}{\theta(1-\theta)} \, .$$

Consideremos el estimador insesgado de θ, $\overline{X} = (1/n)\sum_{i=1}^{n} X_i$. Se tiene que

$$\mathrm{Var}_\theta(\overline{X}) = \frac{\theta(1-\theta)}{n} = \frac{1}{I(\theta)}$$

y por lo tanto, de acuerdo con la observación 2 es IMVU. Esto es un ejemplo donde el estimador IMVU satisface la desigualdad de Rao–Cramer como igualdad. Esto podríamos haberlo visto directamente mostrando que $\overline{X}$ es el estadístico suficiente de una familia exponencial.

Veremos ahora un ejemplo donde el estimador IMVU satisface la desigualdad de Rao–Cramer estrictamente.

Sea $q(\theta) = \theta(1-\theta) = \mathrm{Var}_\theta(X_1)$. Conocemos por el ejemplo 2 de la sección 3.3, que

$$\delta(X_1, X_2, \ldots, X_n) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

es un estimador insesgado de $q(\theta)$. Además se tiene

$$\begin{aligned} \delta(X_1, \ldots, X_n) &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i - n\overline{X}^2\right) \\ &= \frac{n}{n-1}\overline{X}(1-\overline{X}) \end{aligned}$$

Luego $\delta(X_1, \ldots, X_n)$ *depende* del estadístico suficiente y completo $\overline{X}$ y por lo tanto es IMVU.

Sin embargo se tendrá que

$$\mathrm{Var}_\theta(\delta(X_1, \ldots, X_n)) > \frac{q'(\theta)^2}{nI_1(\theta)} \tag{3.32}$$

ya que $\delta(X_1, \ldots, X_n)$ no es el estadístico suficiente de una familia exponencial.

Para la verificación directa de (3.32) ver Problema 11 de 3.13.

3.14 Consistencia de estimadores

La teoría asintótica estudia las propiedades de los procedimientos de inferencia estadística cuando el tamaño de la muestra n que se utiliza es grande, más precisamente, en el límite cuando n tiende a infinito.

Una propiedad deseable para un estimador, es que cuando n es grande la sucesión $\delta_n(X_1, \ldots, X_n)$ se aproxime en algún sentido al valor que queremos estimar. Para precisar estas ideas introduciremos el concepto de consistencia.

Sea $\mathcal{F} = \{F(x, \boldsymbol{\theta}) \text{ con } \boldsymbol{\theta} \in \Theta\}$ una familia de distribuciones y supongamos que para cada n se tiene un estimador $\delta_n(X_1, \ldots, X_n)$ de $q(\boldsymbol{\theta})$ basado en una muestra aleatoria de tamaño n. Daremos la siguiente definición:

Definición 1: $\delta_n(X_1, \ldots, X_n)$ es una *sucesión fuertemente consistente de estimadores* de $q(\boldsymbol{\theta})$ si

$$\lim_{n\to\infty} \delta_n(X_1, \ldots, X_n) = q(\boldsymbol{\theta}) \qquad c.t.p.$$

o sea si $P_{\boldsymbol{\theta}}(\delta_n(X_1, \ldots, X_n) \to q(\boldsymbol{\theta})) = 1$ para todo $\boldsymbol{\theta} \in \Theta$.

Definición 2: $\delta_n(X_1, \ldots, X_n)$ es una *sucesión débilmente consistente de estimadores* de $q(\boldsymbol{\theta})$ si

$$\lim_{n\to\infty} \delta_n(X_1, \ldots, X_n) = q(\boldsymbol{\theta}) \quad \text{en probabilidad.}$$

Es decir, para todo $\varepsilon > 0$ y $\boldsymbol{\theta} \in \Theta$

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}(|\delta_n(X_1, \ldots, X_n) - q(\boldsymbol{\theta})| > \varepsilon) = 0\,.$$

Observación 1: Puesto que convergencia en c.t.p. implica convergencia en probabilidad, entonces toda sucesión fuertemente convergente también lo será débilmente.

Ejemplo 1: Sea $X_1, \ldots, X_n$ una muestra aleatoria de una función de distribución $F(x)$ totalmente desconocida, tal que $E_F(X_1)$ existe. Llamemos $q(F)$ a $E_F(X_1)$. Si

$$\delta_n(X_1, \ldots, X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i\,,$$

por la ley fuerte de los grandes números este estimador resulta fuertemente consistente para $q(F)$.

Si además $E_F(X^2) < \infty$, entonces

$$\delta_n(X_1, \ldots, X_n) = s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 = \frac{1}{n-1} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \overline{X}_n^2$$

es fuertemente consistente para $q(F) = \mathrm{Var}_F X_1$. En efecto,

$$s_n^2 = \frac{n}{n-1} \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{n}{n-1} \overline{X}_n^2 \, .$$

Por la ley fuerte de los grande números

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \to E_F(X_1^2) \quad \text{c.t.p.} \qquad \text{y} \qquad \frac{1}{n} \sum_{i=1}^{n} X_i \to E_F(X_1) \quad \text{c.t.p.}$$

Luego, $\overline{X}_n^2 \to E_F(X_1)^2$ c.t.p. y como $n/(n-1)$ converge a 1 se tiene que

$$\lim_{n \to \infty} s_n^2 = \mathrm{Var}_F(X_1) \quad \text{c.t.p.}$$

Observación 2: Si $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución $N(\mu, \sigma^2)$ se tiene que $\overline{X}_n$ es fuertemente consistente para μ y s_n^2 es fuertemente consistente para σ^2, ya que por lo visto recién

$$\lim_{n \to \infty} \overline{X}_n = E(X_1) \quad \text{c.t.p.}$$

y

$$\lim_{n \to \infty} s_n^2 = \mathrm{Var}\ (X_1) \quad \text{c.t.p.}$$

y sabemos que $E(X_1) = \mu$ y $\mathrm{Var}(X_1) = \sigma^2$.

El siguiente teorema nos da una condición suficiente para que una sucesión de estimadores sea débilmente consistente.

Teorema 1: *Sea, para todo n, $\delta_n = \delta_n(X_1, \ldots, X_n)$ un estimador de $q(\boldsymbol{\theta})$ basado en una muestra aleatoria de tamaño n. Si $\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n) \to 0$ y $E_{\boldsymbol{\theta}}(\delta_n) \to q(\boldsymbol{\theta})$, entonces $\delta_n(X_1, \ldots, X_n)$ es débilmente consistente.*

Demostración: Debemos ver que

$$\lim_{n \to \infty} P_{\boldsymbol{\theta}}(|\delta_n(X_1, \ldots, X_n) - q(\boldsymbol{\theta})| > \varepsilon) = 0 \quad \forall \varepsilon > 0.$$

Por la desigualdad de Markov se tiene

$$\begin{aligned} P_{\boldsymbol{\theta}}(|\delta_n(X_1,\ldots,X_n) - q(\boldsymbol{\theta})| \geq \varepsilon) &\leq \frac{E_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n) - q(\boldsymbol{\theta}))^2}{\varepsilon^2} \\ &\leq \frac{\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n) + \left[E_{\boldsymbol{\theta}}(\delta_n) - q(\boldsymbol{\theta})\right]^2}{\varepsilon^2} \end{aligned}$$

Como por hipótesis $E_{\boldsymbol{\theta}}(\delta_n) - q(\boldsymbol{\theta})) \to 0$ y $(\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n)) \to 0$ se obtiene el resultado.

El siguiente teorema muestra que si $\delta_n(X_1,\ldots,X_n)$ es una sucesión de estimadores IMVU para $q(\boldsymbol{\theta})$ entonces cumple la hipótesis del Teorema 1.

Teorema 2: *Sea $\delta_n(X_1,\ldots,X_n)$ una sucesión de estimadores IMVU para $q(\boldsymbol{\theta})$, donde $X_1,\ldots,X_n$ es una muestra aleatoria de una distribución perteneciente a la familia $F(x,\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Luego $\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n))$ tiende a cero si n tiende a infinito.*

Demostración: Sea

$$\delta_n^*(X_1,\ldots,X_n) = \frac{\sum_{i=1}^n \delta_1(X_i)}{n}$$

luego $E_{\boldsymbol{\theta}}(\delta_n^*) = E_{\boldsymbol{\theta}}(\delta_1) = q(\boldsymbol{\theta})$, es decir δ_n^* es un estimador insesgado de $q(\boldsymbol{\theta})$.

Por otro lado, $\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n^*(X_1,\ldots,X_n)) = \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1(X_1))/n$. Por ser $\delta_n(X_1,\ldots,X_n)$ IMVU de $q(\boldsymbol{\theta})$ se cumple

$$\mathrm{Var}_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n)) \leq \mathrm{Var}_{\boldsymbol{\theta}}(\delta_n^*(X_1,\ldots,X_n)) = \mathrm{Var}_{\boldsymbol{\theta}}(\delta_1(X_1))/n$$

y por lo tanto,

$$\lim_{n\to\infty} \mathrm{Var}_{\boldsymbol{\theta}}(\delta_n(X_1,\ldots,X_n)) = 0.$$

Corolario 1: *Si $\delta_n(X_1,\ldots,X_n)$ es una sucesión de estimadores IMVU para $q(\boldsymbol{\theta})$ donde $X_1,\ldots,X_n$ es una muestra aleatoria de una distribución perteneciente a la familia $\mathcal{F} = \{F(x,\boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta\}$ entonces $\delta_n(X_1,\ldots,X_n)$ es una sucesión de estimadores débilmente consistentes.*

Demostración: Resulta inmediatamente de los teoremas 1 y 2.

3.15 Consistencia de los estimadores de los momentos

En este párrafo demostraremos la consistencia de los estimadores de los momentos.

Teorema 3: *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $\mathcal{F} = \{F(x,\theta)$ con $\theta \in \Theta \subset I\!R\}$, $h(x)$ una función continua con valores en $I\!R$ y supongamos que $E_\theta(h(X_1)) = g(\theta)$ es, como función de θ, continua y estrictamente monótona. Sea el estimador de momentos $\widehat{\theta}_n$ definido como la solución de*

$$\frac{1}{n}\sum_{i=1}^{n} h(X_i) = E_\theta(h(X_1)) = g(\theta).$$

Luego con probabilidad 1 existe n_0 tal que para todo $n \geq n_0$ la ecuación que define $\widehat{\theta}_n$ tiene solución y es fuertemente consistente para θ.

DEMOSTRACIÓN: Sea $\varepsilon > 0$. Hay que demostrar que, con probabilidad 1, existe n_0 tal que

$$|\widehat{\theta}_n - \theta| < \varepsilon \quad \text{para } n \geq n_0.$$

Supongamos que $g(\theta)$ es estrictamente creciente. El caso contrario se demuestra en forma análoga. Luego, se tiene,

$$g(\theta - \varepsilon) < g(\theta) < g(\theta + \varepsilon)\,.$$

Sea $\delta = \min(g(\theta + \varepsilon) - g(\theta),\ g(\theta) - g(\theta - \varepsilon))$; luego

$$g(\theta - \varepsilon) \leq g(\theta) - \delta < g(\theta) < g(\theta) + \delta \leq g(\theta + \varepsilon)\,.$$

Por otro lado, por la ley fuerte de los grandes números

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} h(X_i) = g(\theta) \quad \text{c.t.p.}$$

Luego, con probabilidad 1, dado $\delta > 0$ existe n_0 tal que para todo $n \geq n_0$ se tiene

$$g(\theta) - \delta \leq \frac{1}{n}\sum_{i=1}^{n} h(X_i) \leq g(\theta) + \delta\,.$$

De esta desigualdad se infiere que

$$g(\theta-\varepsilon) \leq \frac{1}{n}\sum h(X_i) \leq g(\theta+\varepsilon) \quad \text{para } n \geq n_0$$

y como $g(\theta)$ es continua y estrictamente creciente, para $n \geq n_0$ existe un único valor $\widehat{\theta}_n$ que satisface

$$\frac{1}{n}\sum h(X_i) = E_{\widehat{\theta}_n}(h(X_1)) = g(\widehat{\theta}_n)$$

Además dicho valor debe estar entre $\theta-\varepsilon$ y $\theta+\varepsilon$, es decir que $\theta-\varepsilon \leq \widehat{\theta}_n \leq \theta+\varepsilon$ para $n \geq n_0$ que es lo que queríamos demostrar.

3.16 Consistencia de los estimadores de máxima verosimilitud

En esta sección enunciaremos un teorema que establece la consistencia de los estimadores de máxima verosimilitud para el caso de un solo parámetro. La demostración se dará en el Apéndice A.

$$\max_{\boldsymbol{\theta}\in\Theta} \prod_{i=1}^{n} p(x_i, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i, \widehat{\boldsymbol{\theta}}_n) \tag{3.33}$$

Se puede demostrar que bajo condiciones muy generales $\widehat{\boldsymbol{\theta}}_n$ definido por (3.33) es fuertemente consistente.

Teorema 1: *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución discreta o continua con densidad en la familia $p(x, \theta)$ con $\theta \in \Theta$, donde Θ es un intervalo abierto de $\mathbb{R}$. Supongamos que $p(x, \theta)$ es derivable respecto de θ y que el conjunto $S = \{x : p(x, \theta) \neq 0\}$ es independiente de θ para todo $\theta \in \Theta$. Sea $\widehat{\theta}_n$ el estimador de máxima verosimilitud de θ, que satisface*

$$\sum_{i=1}^{n} \frac{\partial \ln p(x_i, \widehat{\theta}_n)}{\partial \theta} = 0 \tag{3.34}$$

Supongamos finalmente que la ecuación (3.34) tiene a lo sumo una solución y que $\theta \neq \theta'$ implica que $p(x, \theta) \neq p(x, \theta')$. Entonces $\lim_{n\to\infty} \widehat{\theta}_n = \theta$ c.t.p., es decir, $\widehat{\theta}_n$ es una sucesión de estimadores fuertemente consistente.

Con el objetivo de simplificar la demostración, la condiciones utilizadas en el Teorema 1 son más fuertes que las estrictamente necesarias para que el teorema sea válido. El teorema también vale en el caso de que haya más de un parámetro. Para una demostración más general se puede consultar el Teorema 5.3.1 de Zacks [7] y en Wald [6].

3.17 Estimadores asintóticamente normales y eficientes

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con densidad perteneciente a la familia $p(x, \theta)$ con $\theta \in \Theta$, donde Θ es un intervalo abierto de $\mathbb{R}$, y sea $\delta_n(X_1, \ldots, X_n)$ un estimador insesgado de $q(\theta)$. Luego suponiendo las condiciones A, B, C y D del Teorema 1 de la sección 3.13 se tiene

$$E_\theta[\delta_n(X_1, \ldots, X_n)] = q(\theta) \tag{3.35}$$

$$\mathrm{Var}_\theta(\delta_n(X_1, \ldots, X_n)) \geq \frac{[q'(\theta)]^2}{nI_1(\theta)} \, . \tag{3.36}$$

(3.35) y (3.36) son equivalentes a:

$$E_\theta[\sqrt{n}(\delta_n(X_1, \ldots, X_n) - q(\theta))] = 0 \tag{3.37}$$

$$\mathrm{Var}_\theta[\sqrt{n}(\delta_n(X_1, \ldots, X_n) - q(\theta))] \geq \frac{[q'(\theta)]^2}{I_1(\theta)} \, . \tag{3.38}$$

El mismo Teorema 1 de 3.13, establece que sólo excepcionalmente habrá estimadores que satisfagan simultaneamente (3.37), y la igualdad en (3.38) para n finito. En efecto, esto sucede unicamente si se cumplen

$$q(\theta) = E_\theta(\delta_n(X_1, \ldots, X_n)) \quad \text{y} \quad p(\mathbf{x}, \theta) = A(\theta) e^{c(\theta)\delta_n(x_1, \ldots, x_n)} h(x_1, \ldots, x_n)$$

Sin embargo, bajo condiciones muy generales, existen estimadores (por ejemplo, los de máxima verosimilitud), que para n grande satisfacen aproximadamente (3.37) y la igualdad en (3.38). Para precisar estas propiedades daremos la siguiente definición:

Definición 1: Se dice que $\delta_n(X_1, \ldots, X_n)$ es una sucesión de *estimadores asintóticamente normal y eficiente* (A.N.E.) si $\sqrt{n}(\delta_n(X_1, \ldots, X_n) - q(\theta))$ converge en distribución a una normal con media cero y varianza $[q'(\theta)]^2 / I_1(\theta)$.

Es decir que si $\delta_n(X_1, \ldots, X_n)$ es A.N.E., para n grande se comporta aproximadamente como si tuviese distribución $N(q(\theta), [q'(\theta)]^2 / nI_1(\theta))$, es decir como si fuera insesgado con varianza $[q'(\theta)]^2 / nI_1(\theta)$, que es la menor varianza posible de acuerdo con el Teorema de Rao–Cramer.

El siguiente Teorema, demostrado en el Apéndice B, establece que bajo condiciones muy generales los estimadores de máxima verosimilitud son A.N.E.

Teorema 1: *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución discreta o continua con densidad perteneciente a la familia $p(x, \theta)$ con $\theta \in \Theta$ y Θ un abierto en $\mathbb{R}$. Supongamos que $p(\mathbf{x}, \theta)$ tiene derivada tercera respecto de θ continua y que satisface las condiciones A, C y D del Teorema 1 de 3.13. Sea $\psi(x, \theta) = \frac{\partial \ln p(x,\theta)}{\partial \theta}$ y supongamos además que*

$$\Big|\frac{\partial^3 \ln p(x, \theta)}{\partial \theta^3}\Big| = \Big|\frac{\partial^2 \psi(x, \theta)}{\partial \theta^2}\Big| \leq K$$

para todo $x \in S$ y para todo $\theta \in \Theta$ (S es el mismo que en la condición A). Sea $\widehat{\theta}_n$ un estimador de máxima verosimilitud de θ consistente y sea $q(\theta)$ derivable con $q'(\theta) \neq 0$ para todo θ. Entonces $q(\widehat{\theta}_n)$ es A.N.E. para estimar $q(\theta)$.

Las hipótesis que se han supuesto en este teorema son más fuertes que las estrictamente necesarias con el objetivo de simplificar la demostración. También se puede demostrar un teorema similar para el caso de más de un parámetro. Una demostración más general se puede ver en la sección 5.5 de Zacks [7].

3.18 Apéndice A: Demostración de la consistencia de los estimadores de máxima verosimilitud

Comenzaremos probando algunas propiedades de funciones convexas.

Definición 1: Sea $f(x)$ una función definida sobre un intervalo de $\mathbb{R}$ y que toma valores en $\mathbb{R}$. Diremos que $f(x)$ es *convexa* si:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) \quad \text{con } 0 \leq \lambda \leq 1$$

y diremos que $f(x)$ es *estrictamente convexa* si:

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) \qquad 0 < \lambda < 1.$$

Teorema 1: *Sea $f(x) : I\!R \to I\!R$ una función convexa. Sean $\lambda_1, \ldots, \lambda_n$ tales que $0 \leq \lambda_i \leq 1$ y $\sum_{i=1}^{n} \lambda_i = 1$. Entonces se tiene:*

$$f\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

Además, si $f(x)$ es estrictamente convexa y hay al menos un λ_i que cumple $0 < \lambda_i < 1$ (esto es equivalente a que haya por lo menos dos $\lambda_i > 0$), entonces:

$$f\Big(\sum_{i=1}^{n} \lambda_i x_i\Big) < \sum_{i=1}^{n} \lambda_i f(x_i)$$

DEMOSTRACIÓN: Por inducción (para $n = 2$ se obtiene la definición 1).

Teorema 2 (Desigualdad de Jensen): *Sea Y una variable aleatoria y $h : I\!R \to I\!R$ una función convexa, luego se tiene*

$$E(h(Y)) \geq h(E(Y))$$

Además si h es estrictamente convexa y Y no es constante con probabilidad 1 se tiene:

$$E(h(Y)) > h(E(Y))$$

DEMOSTRACIÓN: Sólo haremos el caso en que Y es discreta y toma un número finito de valores.

Supongamos que Y toma los valores $y_1, y_2, \ldots, y_k$ con probabilidades $p_1, p_2, \ldots, p_k$. Luego aplicando el Teorema 1 se obtiene:

$$h(E(Y)) = h\Big(\sum_{i=1}^{k} y_i p_i\Big) \leq \sum_{i=1}^{n} h(y_i)p_i = E(h(Y))$$

En el caso en que h sea estrictamente convexa y Y no sea constante, hay al menos dos p_i mayores que cero, luego también por el Teorema 1 obtenemos:

$$h(E(Y)) = h\Big(\sum_{i=1}^{k} y_i p_i\Big) < \sum_{i=1}^{k} h(y_i)p_i = E(h(Y))$$

Teorema 3: *Sea $f : \mathbb{R} \to \mathbb{R}$ tal que $f"(x) > 0$ para todo x; luego $f(x)$ es convexa.*

DEMOSTRACIÓN: Puede verse en cualquier libro de cálculo.

Teorema 4: *Sean p y q dos densidades o dos funciones de densidad discretas o continuas distintas. Luego se tiene:*

$$E_p\Big(\ln \frac{q(X)}{p(X)}\Big) < 0$$

(donde E_p significa que se calcula la esperanza considerando que X tiene una distribución discreta o continua cuya densidad o probabilidad puntual es p).

DEMOSTRACIÓN: Primero veremos que $q(X)/p(X)$ no es constante con probabilidad 1. La demostración se hará suponiendo que X es continua. El caso discreto es totalmente análogo. Supongamos que $q(X)/p(X) = k$ c.t.p., donde k es una constante. Luego $E_p(q(X)/p(X)) = k$. Esto es:

$$\int_{-\infty}^{+\infty} (q(x)/p(x))p(x)dx = k \tag{3.39}$$

pero

$$\int_{-\infty}^{+\infty} (q(x)/p(x))p(x)dx = 1 \tag{3.40}$$

pues $q(x)$ es una densidad. Luego, de (3.39) y (3.40) resulta $k = 1$. Entonces $p(X) = q(X)$ c.t.p. y esto contradice la hipótesis. Por lo tanto $q(X)/p(X)$ no es constante.

Por otro lado $-\ln(x)$ es una función estrictamente convexa ya que:

$$\frac{d^2(-\ln x)}{dx^2} = \frac{1}{x^2} > 0\ .$$

Luego, estamos en condiciones de aplicar la desigualdad de Jensen (Teorema 2), con $Y = q(X)/p(X)$ y $h(x) = -\ln x$. En estas condiciones obtenemos

$$E_p\left[-\ln \frac{q(X)}{p(X)}\right] > -\ln\left[E_p \frac{q(X)}{p(X)}\right] = -\ln \int_{-\infty}^{+\infty} \frac{q(x)}{p(x)} p(x)dx = -\ln 1 = 0.$$

Luego $E_p[-\ln(q(X)/p(X))] > 0$ y $E_p[\ln(q(X)/p(X))] < 0$ con lo que obtenemos la tesis.

Demostración del Teorema 1 de Sección 3.16

Sea $L_n(X_1,\ldots,X_n,\theta) = (1/n)\sum_{i=1}^n \ln p(X_i,\theta)$. Luego $\widehat{\theta}_n$ satisface $L_n(X_1,\ldots,X_n,\widehat{\theta}_n) = \max_{\theta\in\Theta} L_n(X_1,\ldots,X_n,\theta)$ y

$$\frac{\partial L_n(X_1,\ldots,X_n,\widehat{\theta}_n)}{\partial\theta} = 0\ .$$

Además se tiene

$$L_n(X_1,\ldots,X_n,\theta+\delta) - L_n(X_1,\ldots,X_n,\theta) = \frac{1}{n}\sum_{i=1}^n \ln\Big(\frac{p(X_i,\theta+\delta)}{p(X_i,\theta)}\Big) \quad (3.41)$$

$$L_n(X_1,\ldots,X_n,\theta-\delta) - L_n(X_1,\ldots,X_n,\theta) = \frac{1}{n}\sum_{i=1}^n \ln\Big(\frac{p(X_i,\theta-\delta)}{p(X_i,\theta)}\Big) \quad (3.42)$$

Como $\theta \neq \theta'$ implica $p(X_1,\theta) \neq p(X_1,\theta')$, aplicando el Teorema 4 resulta que

$$E_\theta\left(\ln\left[\frac{p(X_1,\theta+\delta)}{p(X_1,\theta)}\right]\right) < 0 \quad (3.43)$$

$$E_\theta\left(\ln\left[\frac{p(X_1,\theta-\delta)}{p(X_1,\theta)}\right]\right) < 0 \quad (3.44)$$

Entonces, de (3.41), (3.42), (3.43) y (3.44) y de la ley fuerte de los grandes números resulta que con probabilidad igual a 1 existe un n_0 tal que $n > n_0$ implica:

$$L_n(X_1,\ldots,X_n,\theta-\delta) < L_n(X_1,\ldots,X_n,\theta)$$

y

$$L_n(X_1,\ldots,X_n,\theta+\delta) < L_n(X_1,\ldots,X_n,\theta)\ .$$

Luego, para $n > n_0$ en el intervalo $(\theta-\delta,\theta+\delta)$ existe un máximo relativo, digamos θ_n^*, que satisface

$$\frac{\partial L_n(X_1,\ldots,X_n,\theta_n^*)}{\partial\theta} = 0\ ,$$

pero hemos supuesto que $\widehat{\theta}_n$ era el único que satisfacía esta igualdad. Luego, $\widehat{\theta}_n = \theta_n^*$ y por lo tanto $\widehat{\theta}_n \in (\theta-\delta,\theta+\delta)$.

3.19 Apéndice B: Demostración de la normalidad y eficiencia asintótica de los estimadores de máxima verosimilitud

Demostraremos previamente un lema.

Lema 1: *Sea $X_1, \ldots, X_n$ una sucesión de variables aleatorias tales que $\sqrt{n}(X_n - \mu)$ converge en distribución a $N(0, \sigma^2)$. Sea $g(x)$ una función definida en $\mathbb{R}$ tal que $g'(\mu) \neq 0$ y $g'(x)$ es continua en $x = \mu$. Luego se tiene que $\sqrt{n}(g(X_n) - g(\mu))$ converge en distribución a una distribución $N(0, \sigma^2(g'(\mu))^2)$.*

Demostración: Primero demostraremos que $X_n \to \mu$ en probabilidad. Sean $\varepsilon > 0$ y $\delta > 0$ arbitrarios y X una variable aleatoria con distribución $N(0, \sigma^2)$. Luego existe K suficientemente grande tal que $P(|X| > K) < \delta$. Por otro lado, $P(|X_n - \mu| > \varepsilon) = P(\sqrt{n}|X_n - \mu\| \geq \sqrt{n}\varepsilon)$. Sea n_0 tal que $\sqrt{n_0}\varepsilon \geq K$. Luego si $n \geq n_0$:

$$P(|X_n - \mu| \geq \varepsilon) \leq P(\sqrt{n}|X_n - \mu| \geq K) \; .$$

Como $\sqrt{n}(X_n - \mu)$ converge en distribución a una variable con distribución $N(0, \sigma^2)$, se tiene

$$\lim_{n\to\infty} P(|X_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} P(\sqrt{n}|X_n - \mu| \geq K) = P(|X| \geq K) < \delta \; .$$

Luego

$$\lim_{n\to\infty} (P|X_n - \mu| \geq \varepsilon) < \delta \quad \text{para todo } \delta > 0 \; ,$$

por lo tanto, $\lim_{n\to\infty} P(|X_n - \mu| \geq \varepsilon) = 0$ y resulta $X_n \to \mu$ en probabilidad.

Por otra parte, el teorema del valor medio implica que

$$\sqrt{n}(g(X_n) - g(\mu)) = \sqrt{n}g'(\xi_n)(X_n - \mu) \tag{3.45}$$

con ξ_n un punto intermedio entre X_n y μ. Luego, $\xi_n \to \mu$ en probabilidad y como $g'(x)$ es continua en μ, $g'(\xi_n) \to g'(\mu)$ en probabilidad.

Por lo tanto, como por hipótesis $\sqrt{n}(X_n - \mu)$ converge en distribución a una $N(0, \sigma^2)$ y $g'(\xi_n) \to g'(\mu)$ en probabilidad, aplicando la propiedad 5 de 1.8, resulta que $\sqrt{n}(g(X_n) - g(\mu))$ converge en distribución a una $N(0, \sigma^2(g'(\mu))^2)$.

Demostración del Teorema 1 de la sección 3.17. Indiquemos por

$$\psi'(x,\theta) = \frac{\partial \psi(x,\theta)}{\partial \theta} \quad \text{y} \quad \psi''(x,\theta) = \frac{\partial^2 \psi(x,\theta)}{\partial \theta^2} \ .$$

El estimador de máxima verosimilitud satisface:

$$\sum_{i=1}^{n} \psi(X_i, \widehat{\theta}_n) = 0 \ .$$

Desarrollando en serie de Taylor alrededor de θ se obtiene:

$$\sum_{i=1}^{n} \psi(X_i,\theta) + \left(\sum_{i=1}^{n} \psi'(X_i,\theta)\right)(\widehat{\theta}_n - \theta) + \frac{1}{2}\left(\sum_{i=1}^{n} \psi''(X_i,\xi_n)\right)(\widehat{\theta}_n - \theta)^2 = 0 \ ,$$

donde ξ_n es un punto intermedio entre $\widehat{\theta}_n$ y θ. Despejando $(\widehat{\theta}_n - \theta)$ y multiplicando ambos miembros por $\sqrt{n}$ se obtiene:

$$\sqrt{n}(\widehat{\theta}_n - \theta) = \frac{-\sum_{i=1}^{n} \psi(X_i,\theta)/\sqrt{n}}{(1/n)\sum_{i=1}^{n} \psi'(X_i,\theta) + (1/2n)\Big(\sum_{i=1}^{n} \psi''(X_i,\xi_n)\Big)(\widehat{\theta}_n - \theta)}$$

Sea $D(X_1,\ldots,X_n,\theta)$ el denominador de esta última expresión. Vamos a demostrar que:

(a)

$$D(X_1,\ldots,X_n,\theta) \to -I_1(\theta) = -E_\theta\left[\psi(X,\theta)\right]^2 \quad \text{en probabilidad.}$$

(b) $\sum_{i=1}^{n} \psi(X_i,\theta)/\sqrt{n}$ converge en distribución a una distribución $N(0, I_1(\theta))$

Probemos (a). Como $|\psi''(X_i,\theta)| \le K$ para todo θ, se tiene que

$$\left|\frac{1}{2n}\sum_{i=1}^{n} \psi''(X_i,\xi_n)(\widehat{\theta}_n - \theta)\right| \le \frac{K}{2}|(\widehat{\theta}_n - \theta)|$$

y luego como $\widehat{\theta}_n$ es consistente se deduce que:

$$\frac{1}{n}\sum_{i=1}^{n} \psi''(X_i,\xi_n)(\widehat{\theta}_n - \theta) \to 0 \quad \text{en probabilidad.} \tag{3.46}$$

Por otro lado, como $\psi'(X_i,\theta)$ son n variables aleatorias, independientes igualmente distribuidas, por la ley de los grandes números implica que

$$\frac{1}{n}\sum_{i=1}^{n} \psi'(X_i,\theta) \to E(\psi'(X_1,\theta) \quad \text{en probabilidad.} \tag{3.47}$$

Pero de acuerdo con el Lema 1 de la sección 3.13

$$E_\theta(\psi'(X_1, \theta)) = -I_1(\theta) \ .$$

Luego, usando (3.46) y (3.47) se obtiene:

$$D(X_1, \ldots, X_n, \theta) \to -I_1(\theta) \quad \text{en probabilidad,}$$

con lo que queda probado (a). Para probar (b) observemos que, como las variables aleatorias

$$\psi(X_i, \theta) = \frac{\partial \ln p(X_i, \theta)}{\partial \theta}$$

son independientes e igualmente distribuidas con esperanza 0 y varianza $I_1(\theta)$ (ver Lema 1 de la sección 3.13), por el Teorema Central del límite

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, \theta)$$

converge en distribución a $\mathrm{N}(0, I_1(\theta))$.

Luego $\sqrt{n}(\widehat{\theta}_n - \theta)$ converge en distribución a una ley $N(0, I_1(\theta)/(I_1(\theta)^2)$ o sea $N(0, 1/I_1(\theta))$.

Consideremos ahora el estimador de máxima verosimilitud $q(\theta)$ dado por $q(\widehat{\theta}_n)$.

Por el Lema 1 se tendrá que $\sqrt{n}(q(\widehat{\theta}_n) - q(\theta))$ converge en distribución a una $N(0, (q'(\theta))^2/I_1(\theta))$.

REFERENCIAS DEL CAPITULO 3

[1] Bahadur, R.R. (1954). Sufficiency and Statistical Decision Functions. *Annals of Mathematical Statistics* **25**, 423–462.

[2] Draper, N. and Smith, H. (1966). *Applied Regression Analysis.* J. Wiley & Sons.

[3] Dynkin, E.B. (1961). Necessary and Sufficient Statistics for Families of Distributions. *Selected Translations of Mathematical Statistics and Probability* **1**, 23–41.

[4] Lehmann, E.L. (1994). *Testing Statistical Hypothesis.* Chapman & Hall.

[5] Lehmann, E.L. (1983). *Theory of Point Estimation.* J. Wiley & Sons.

[6] Wald, A.N. (1949). Note on the Consistency of the Maximum Likelihood Estimates. *Annals of Mathematical Statistics* **20**, 595–601.

[7] Zacks, S. (1971). *The Theory of Statistical Inference.* J. Wiley & Sons.

Chapter 4

Estimadores Bayesianos y Minimax

4.1 Enfoque Bayesiano del problema de la estimación puntual

Consideremos nuevamente un problema estadístico de estimación paramétrico. Se observa un vector $\mathbf{X} = (X_1, ..., X_n)$, que puede ser, por ejemplo, aunque no necesariamente, una muestra aleatoria de cierta distribución) con densidad discreta o continua en la familia $f(\mathbf{x}, \boldsymbol{\theta})$, con $\boldsymbol{\theta} = (\theta_1, ..., \theta_p) \in \Theta \subset I\!R^p$.

El enfoque llamadado frecuentista que hemos estudiado no supone ningún conocimiento previo de $\boldsymbol{\theta}$. El enfoque bayesiano, por lo contrario, supone que se tiene alguna información previa sobre $\boldsymbol{\theta}$. Esta información está expresada por medio de una distribución sobre $\boldsymbol{\theta}$, denominada distribución *a priori.* Aquí supondremos que esta distribución a priori, indicada por τ, tiene una densidad $\gamma(\boldsymbol{\theta})$. La distribución a priori puede tener distintas interpretaciones según el problema. Se pueden dar las siguientes alternativas

- La distribución a priori está basada en experiencias previas similares.
- La distribución a priori expresa una creencia subjetiva.

El hecho de que el enfoque bayesiano considere una distribución de probabilidades sobre $\boldsymbol{\theta}$, supone tratar a $\boldsymbol{\theta}$ como una variable aleatoria, y por lo tanto a esta variable la denominaremos Θ para distinguirla del valor que toma $\boldsymbol{\theta}$. Esta notación puede llevar a confusión dado que también llamamos

Θ al conjunto de valores de $\boldsymbol{\theta}$. Sin embargo, por el contexto quedará claro el significado de este símbolo en cada caso.

Dado que consideramos ahora el valor del parámetro como el valor de una variable aleatoria, la interpretación de la familia de densidades $f(\mathbf{x}, \boldsymbol{\theta})$ en el enfoque bayesiano también cambia. En el enfoque bayesiano $f(\mathbf{x}, \boldsymbol{\theta})$ se interpreta como la distribución condicional de la muestra $\mathbf{X}$ dado que la variable Θ toma el valor $\boldsymbol{\theta}$.

Una vez observada la muestra $\mathbf{X}$ se puede calcular la distribución condicional de Θ dada $\mathbf{X}$. Esta distribución se denomina distribución *a posteriori* y está dada por

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}, \boldsymbol{\theta})\gamma(\boldsymbol{\theta})}{\int ... \int f(\mathbf{x}, \mathbf{t})\gamma(\mathbf{t})d\mathbf{t}}. \tag{4.1}$$

En efecto el numerador de (4.1) corresponde a la densidad conjunta de $\mathbf{X}$ y Θ, y el denominador a la densidad marginal de $\mathbf{X}$.

Si la distribución de $\boldsymbol{\theta}$ fuese discreta, habría que reemplazar las integrales del denominador por las correspondientes sumatorias. En lo sucesivo, supondremos que las distribuciones de $\mathbf{X}$ y de $\boldsymbol{\theta}$ son continuas, pero el tratamiento en el caso discreto es similar.

Una de las ventajas del enfoque bayesiano es que se pueden definir en forma natural estimadores óptimos, sin necesidad de restricciones poco naturales como la de estimadores insesgados a la que debimos recurrir en el enfoque frecuentista. Para ver esto supongamos que queremos estimar $\lambda = q(\boldsymbol{\theta})$ y consideremos una función de pérdida $\ell(\boldsymbol{\theta}, d)$ que indica el costo de estimar $\lambda = q(\boldsymbol{\theta})$ utilizando del valor d. Supongamos que se tiene un estimador $\hat{\lambda} = \delta(\mathbf{x})$. Luego la pérdida será una variable aleatoria $\ell(\Theta, \delta(\mathbf{X}))$, y la pérdida esperada que llamaremos *riesgo de Bayes* está dada por

$$r(\delta, \tau) = E(\ell(\Theta, \delta(\mathbf{X}))), \tag{4.2}$$

donde aquí la esperanza se toma con respecto a la distribución conjunta de $\mathbf{X}$ y Θ. Por lo tanto, dada la distribución priori τ, un estimador óptimo será aquel que minimice $r(\delta, \tau)$. Este estimador se denomina *estimador de Bayes* correspondiente a la distribución a priori τ y será representado por δ_τ.

Luego, la función de riesgo de la teoría frecuentista, $R(\delta, \boldsymbol{\theta})$, estará dada por

$$\begin{aligned} R(\delta, \boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}(\ell(\boldsymbol{\theta}, \delta(\mathbf{X}))) \\ &= E(\ell(\Theta, \delta(\mathbf{X}))|\Theta = \boldsymbol{\theta}) = \int \ell(\boldsymbol{\theta}, \delta(\mathbf{x}))f(\mathbf{x}, \boldsymbol{\theta})d\mathbf{x}. \end{aligned} \tag{4.3}$$

Con lo cual,

$$r(\delta,\tau) = E_\tau\left(E(\ell(\Theta,\delta(\mathbf{X}))|\Theta)\right) = \int ... \int R(\delta,\boldsymbol{\theta})\gamma(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4.4}$$

Consideremos como caso particular la función de pérdida cuadrática, es decir,

$$\ell(\boldsymbol{\theta},d) = (q(\boldsymbol{\theta}) - d)^2.$$

En este caso, el estimador de Bayes será la función $\delta_\tau(X)$ que minimiza el error cuadrático medio

$$E((\delta(\mathbf{X}) - q(\Theta))^2)$$

y por lo tanto, de acuerdo a la teoría de esperanza condicional, éste será único y estará dado por

$$\delta_\tau(\mathbf{x}) = E(q(\Theta)|\mathbf{X} = \mathbf{x}) = \int ... \int q(\boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta},$$

es decir, será la esperanza condicional de $q(\Theta)$ con respecto a la distribución a posteriori de Θ.

Ejemplo 1. Sea $\mathbf{X} = (X_1, ..., X_n)$ una muestra independiente de una distribución $\text{Bi}(\theta, 1)$, y supongamos que la distribución a priori τ de θ sea una distribución $\beta(a, b)$, es decir, con una densidad

$$\gamma(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}I_{[0,1]}(\theta). \tag{4.5}$$

Es conocido que esta distribución tiene la siguiente esperanza y varianza

$$E(\theta) = \frac{a}{a+b}, \tag{4.6}$$

$$var(\theta) = \frac{ab}{(a+b)^2(a+b+1)} = \frac{E(\theta)\left(1-E(\theta)\right)}{a+b+1}. \tag{4.7}$$

Luego si se conoce la media y la varianza de la distribución a priori de Θ, se pueden determinar a y b. La fórmula (4.7) muestra que para un dado valor de la esperanza, la varianza depende de $a + b$, tendiendo a 0 cuando $a + b \to +\infty$.

La distribución de la muestra $X_1, X_2, ...X_n$ dado el valor de θ tiene una función de probabilidad puntual dada por

$$f(x_1, ..x_n, \theta) = \theta^{\sum_{i=1}^n x_i}\left(1-\theta\right)^{n-\sum_{i=1}^n x_i}. \tag{4.8}$$

Luego usando (4.1) se tiene que la distribución a posteriori de θ tiene una densidad dada por

$$f(\theta|x_1,...,x_n) = \frac{\theta^{\sum_{i=1}^n x_i+a-1}(1-\theta)^{n-\sum_{i=1}^n x_i+b-1}}{\int_0^1 t^{\sum_{i=1}^n x_i+a-1}(1-t)^{n-\sum_{i=1}^n x_i+b-1}dt}. \tag{4.9}$$

Ahora bien, el denominador de esta expresión es igual a

$$\frac{\Gamma(n+a+b)}{\Gamma(a+\sum_{i=1}^n x_i)\Gamma(n-\sum_{i=1}^n x_i+b)}$$

por lo tanto, la distribución a posteriori de θ dado $\mathbf{X} = \mathbf{x}$ es $\beta(a+\sum_{i=1}^n x_i, n-\sum_{i=1}^n +b)$.

Supongamos que la función de pérdida es cuadrática. Luego el estimador de Bayes, que indicaremos por $\delta_{a,b}$, está dado por $E(\Theta|\mathbf{X})$, y de acuerdo a (4.6) tendremos que

$$\delta_{a,b} = \frac{T+a}{a+b+n} = \frac{n}{n+a+b}\,\frac{T}{n} + \frac{a+b}{a+b+n}\,\frac{a}{a+b}, \tag{4.10}$$

donde $T = \sum_{i=1}^n X_i$. Por lo tanto, el estimador de Bayes se comporta como un promedio ponderado de dos estimadores: el IMVU $\delta_1 = T/n$ que no usa la información de la distribución a priori y $\delta_2 = a/(a+b)$ que corresponde a la esperanza de la distribución a priori y que se usaría si no se hubiese observado la muestra. También vemos que el peso asignado a δ_2 tiende a 0 cuando el tamaño de la muestra n aumenta.

De acuerdo a (4.10), el estimador de Bayes correspondiente a una distribución a priori $\beta(a,b)$ puede interpretarse como el estimador frecuentista correspondiente a una muestra de tamaño $n+a+b$ con $\sum_{i=1}^n X_i + a$ éxitos.

Observación 1. En el ejemplo anterior hemos partido de una distribución a priori $\beta(a,b)$, y hemos obtenido que la distribución a posteriori también está en la misma familia, ya que es $\beta(a+\sum_{i=1}^n x_i, n-\sum_{i=1}^n x_i+b)$. Se dice entonces que la familia de distribuciones beta es la *conjugada* de la familia de distribuciones binomial.

Ejemplo 2. Sea $\mathbf{X} = (X_1,...,X_n)$ una muestra independiente de una distribución $\mathrm{N}(\theta,\sigma^2)$, con σ^2 conocido, y supongamos que la distribución a priori de θ sea $\mathrm{N}(\mu,\rho^2)$.

Luego la densidad de la muestra $\mathbf{X} = (X_1,...,X_n)$ dado θ está dada por

$$f(\mathbf{x},\theta) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^n (x_i-\theta)^2\right), \tag{4.11}$$

donde $\exp(x) = e^x$. La densidad de la distribución a priori está dada por

$$\gamma(\theta) = \frac{1}{(2\pi)^{1/2}\rho}\exp\left(\frac{-(\theta-\mu)^2}{2\rho^2}\right) \tag{4.12}$$

Luego multiplicando (4.11) y (4.12), desarrollando los cuadrados y haciendo algún manipuleo algebraico, se obtiene que distribución conjunta de $\mathbf{X}$ y Θ está dada por

$$f_{\mathbf{X},\Theta}(\mathbf{x},\theta) = C_1(\mathbf{x},\sigma^2,\mu,\rho^2)\exp\left(-\frac{\theta^2}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\rho^2}\right)+\theta\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\rho^2}\right)\right),$$

donde $C_1(\mathbf{x},\sigma^2,\mu,\rho^2)$ no depende de θ. Completando cuadrados, se obtiene

$$f_{\mathbf{X},\Theta}(\mathbf{x},\theta) = C_2(\mathbf{x},\sigma^2,\mu,\rho^2)\exp\left(-\frac{1}{2D}\left(\theta - D\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\rho^2}\right)\right)^2\right), \tag{4.13}$$

donde

$$D = \frac{1}{(n/\sigma^2)+(1/\rho^2)}. \tag{4.14}$$

Finalmente, usando (1) obtenemos

$$f(\theta|\mathbf{x}) = C_3(\mathbf{x},\sigma^2,\rho^2,\mu)\exp\left(-\frac{1}{2D}(\theta - D\left(\frac{\bar{x}}{\sigma^2}+\frac{\mu}{\rho^2}\right)\right)^2 \tag{4.15}$$

Luego, esta densidad, excepto una función que depende sólo de $\mathbf{x}$, corresponde a una distribución

$$\mathrm{N}\left(D\left(\frac{n\bar{x}}{\sigma^2}+\frac{\mu}{\rho^2}\right), D\right). \tag{4.16}$$

Como se trata de la distribución condicional de Θ dado $\mathbf{X} = \mathbf{x}$, podemos considerar a C_3 como constante. Luego la distribución a posteriori de Θ está dada por (4.16).

Supongamos nuevamente que consideramos una función de pérdida cuadrática. El estimador de Bayes estará dado, en ese caso, por la esperanza condicional de Θ dado $\mathbf{X}$, y por lo tanto de acuerdo a (4.16) y (4.14) estará dado por

$$\delta_\tau(\mathbf{X}) = D\left(\frac{n\bar{X}}{\sigma^2}+\frac{\mu}{\rho^2}\right) = w\bar{X} + (1-w)\mu, \tag{4.17}$$

donde

$$w = \frac{n/\sigma^2}{(n/\sigma^2) + (1/\rho^2)}.$$

Por lo tanto, nuevamente, el estimador de Bayes es un promedio ponderado del estimador IMVU de la teoría frecuentista $\bar{X}$ y la media de la distribución a priori μ. Los pesos son inversamente proporcionales a las varianzas σ^2/n y ρ^2 de ambos estimadores. A medida que el tamaño de la muestra n crece, el peso del estimador basado en la información a priori tiende a 0. Es decir, a medida que el tamaño de la muestra crece, la información a priori tiene menos relevancia para determinar el estimador de Bayes.

Observación 2. En este ejemplo partimos de una distribución a priori en la familia $\mathrm{N}(\mu, \rho^2)$, y obtenemos que la distribución a posteriori está dada por (4.16), y por lo tanto está en la misma familia normal. Luego la familia de distribuciones conjugadas de la familia normal con varianza conocida es la familia normal.

Veamos algunas propiedades de los estimadores Bayes para funciones de pérdida arbitrarias.

Teorema 1. *Sea δ_τ un estimador Bayes respecto de la distribución a priori τ y supongamos que δ_τ es único, entonces δ_τ es admisible.*
DEMOSTRACIÓN. Supongamos que existe otro estimador δ^* tan bueno como δ_τ, es decir, $R(\delta^*, \boldsymbol{\theta}) \leq R(\delta_\tau, \boldsymbol{\theta})$ para todo $\boldsymbol{\theta} \in \Theta$. Integrando respecto a la distribución a priori de $\boldsymbol{\theta}$ en ambos miembros de la desigualdad, obtenemos $r(\delta^*, \tau) \leq r(\delta_\tau, \tau)$. Con lo cual, por la unicidad $\delta^* = \delta_\tau$.

Se puede obtener un resultado de admisibilidad para reglas Bayes sin pedir unicidad, siempre y cuando, Θ sea abierto, la distribución a priori tenga una densidad positiva para todo $\boldsymbol{\theta} \in \Theta$ y la función de riesgo $R(\delta, \boldsymbol{\theta})$ sea continua en $\boldsymbol{\theta}$ para todo estimador δ.

Hemos visto que en el caso de la pérdida cuadrática, el estimador Bayes podía obtenerse como la esperanza de la distribución a posteriori de Θ. El siguiente Teorema da una manera de obtener el estimador Bayes para el caso de otras funciones de pérdida.

Teorema 2. *Sea τ la distribución de Θ y $F_{\boldsymbol{\theta}}(\mathbf{x})$ la distribución condicional de $\mathbf{X}$ dado $\boldsymbol{\theta}$. Supongamos que se cumplen las siguientes condiciones para estimar $q(\boldsymbol{\theta})$ utilizando la pérdida $\ell(\boldsymbol{\theta}, d)$*

a) Existe un estimador δ_0 con riesgo finito.

b) Para cada valor de $\mathbf{x}$ existe un valor, que indicaremos $\delta_\tau(\mathbf{x})$, que minimiza $E\left(\ell(\boldsymbol{\theta}, d)|\mathbf{X} = \mathbf{x}\right)$.

Entonces, $\delta_\tau(\mathbf{x})$ es un estimador de Bayes respecto a τ.
DEMOSTRACIÓN. Sea $\delta(\mathbf{X})$ un estimador con riesgo Bayes finito. Luego, como la pérdida es nonegativa, $E\left(\ell(\boldsymbol{\theta}, \delta(X))|\mathbf{X} = \mathbf{x}\right)$ es finita para casi todo $\mathbf{x}$. Por lo tanto, tenemos

$$E\left(\ell\left(\boldsymbol{\theta}, \delta(\mathbf{x})\right)|\mathbf{X} = \mathbf{x}\right) \geq E\left(\ell\left(\boldsymbol{\theta}, \delta_\tau(\mathbf{x})\right)|\mathbf{X} = \mathbf{x}\right)$$

de donde, tomando esperanza respecto a la distribución marginal de $\mathbf{X}$, obtenemos $r(\delta, \tau) \geq r(\delta_\tau, \tau)$ y por lo tanto, δ_τ es un estimador Bayes.

Corolario *Sea τ una distribución a priori para $\boldsymbol{\theta}$ y supongamos que se cumplen las condiciones del Teorema 2.*

a) Para la pérdida $\ell(\boldsymbol{\theta}, d) = w(\boldsymbol{\theta})(q(\boldsymbol{\theta}) - d)^2$, donde $w(\boldsymbol{\theta}) > 0$ y $E(w(\boldsymbol{\theta})) < \infty$, la regla Bayes δ_τ está dada por

$$\delta_\tau(\mathbf{x}) = \frac{E\left(q(\boldsymbol{\theta})w(\boldsymbol{\theta})|\mathbf{X} = \mathbf{x}\right)}{E\left(w(\boldsymbol{\theta})|\mathbf{X} = \mathbf{x}\right)}$$

b) Para la pérdida $\ell(\boldsymbol{\theta}, d) = |q(\boldsymbol{\theta}) - d|$, la regla Bayes $\delta_\tau(\mathbf{x})$ es la mediana de la distribución a posteriori de $q(\boldsymbol{\theta})$ condicional a $\mathbf{x}$

c) Para la pérdida $\ell(\boldsymbol{\theta}, d) = I_{|q(\boldsymbol{\theta})-d|>c}$, la regla Bayes δ_τ es el punto medio del intervalo $\mathcal{I}$ de longitud $2c$ que maximiza $P(q(\boldsymbol{\theta}) \in \mathcal{I}|\mathbf{X} = \mathbf{x})$

4.2 Utilización de métodos bayesianos para resolver problemas frecuentistas

En esta sección vamos a mostrar como los resultados de la teoría bayesiana pueden ser útiles, aunque no se comparta ese punto de vista. Es decir, veremos que los resultados de esta teoría se pueden usar para resolver problemas que surgen de la teoría frecuentista.

Consideremos una muestra $\mathbf{X} = (X_1, ..., , X_n)$ con distribución conjunta $f(\mathbf{x}, \boldsymbol{\theta})$ donde el vector de parámetros $\boldsymbol{\theta} \in \Theta$. Supongamos que queremos

estimar $\lambda = q(\boldsymbol{\theta})$ y que tenemos una función de pérdida $\ell(\boldsymbol{\theta}, d)$. En el enfoque frecuentista un estimador $\delta(\mathbf{X})$ de λ queda caracterizado por su función de riesgo

$$R(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\ell(\boldsymbol{\theta}, \delta(\mathbf{X})) = \int \ell(\boldsymbol{\theta}, \delta(\mathbf{x}))f(\mathbf{x}, \boldsymbol{\theta})d\mathbf{x}. \tag{4.18}$$

Como θ es desconocido, lo ideal sería encontrar un estimador $\delta^*(\mathbf{X})$ tal que, dado cualquier otro estimador $\delta(\mathbf{x})$ se tuviese

$$R(\delta^*, \boldsymbol{\theta}) \leq R(\delta, \boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta.$$

Como ya hemos visto al comienzo del curso estos estimadores no existen excepto en casos muy poco interesantes.

Una alternativa es comparar los estimadores a través del máximo riesgo. Dado un estimador $\delta(\mathbf{X})$ de λ su máximo riesgo se define por

$$MR(\delta) = \sup_{\boldsymbol{\theta} \in \Theta} R(\delta, \boldsymbol{\theta}). \tag{4.19}$$

El criterio de comparar los estimadores por su máximo riesgo es pesimista, ya que actua como si el parámetro fuese a tomar el valor más desfavorable para el estimador. Un estimador óptimo de acuerdo a este criterio es un estimador δ^* tal que dado cualquier otro estimador δ se tiene

$$MR(\delta^*) \leq MR(\delta). \tag{4.20}$$

Definición 1. Un estimador satisfaciendo (4.20) se denomina *minimax*.

Vamos a ver como la teoría bayesiana nos ayuda a encontrar estimadores minimax. Para ello, consideremos una distribución a priori τ con densidad $\gamma(\theta)$. El correspondiente estimador de Bayes δ_τ verifica que, dado cualquier otro estimador δ, se tiene

$$r(\delta_\tau, \tau) \leq r(\delta, \tau). \tag{4.21}$$

Luego, de acuerdo a (4.4) se tendrá entonces que para cualquier estimador δ

$$\int R(\delta_\tau, \boldsymbol{\theta})\gamma(\boldsymbol{\theta})d\boldsymbol{\theta} \leq \int R(\delta, \boldsymbol{\theta})\gamma(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{4.22}$$

Sea $r_\tau = r(\delta_\tau, \tau)$, es decir, el mínimo riesgo de Bayes cuando la distribución a priori es τ.

Definición 2. Se dirá que una distribución a priori τ_0 es *menos favorable* si, para cualquier otra distribución τ, se tiene $r_\tau \leq r_{\tau_0}$.

Naturalmente uno se puede preguntar para qué distribuciones a priori τ el estimador Bayes δ_τ será minimax. Un procedimiento minimax, al minimizar el máximo riesgo, trata de comportarse lo mejor posible en la peor situación. Por lo tanto, uno puede esperar que el estimador minimax sea Bayes para la peor distribución posible que es la distribución menos favorable.

El siguiente Teorema nos permite usar la teoría bayesiana para encontrar estimadores minimax.

Teorema 1. *Supongamos que se tiene una distribución a priori τ_0 tal que el estimador de Bayes δ_{τ_0} tiene función de riesgo, $R(\delta_\tau, \boldsymbol{\theta})$, constante en $\boldsymbol{\theta}$. Entonces:*

a) δ_{τ_0} es un estimador minimax,

b) si δ_{τ_0} es el único estimador Bayes respecto de τ_0, δ_{τ_0} es el único estimador minimax,

c) τ_0 es la distribución menos favorable.

Demostración. Como el riesgo de δ_{τ_0} es constante se cumple que

$$r(\delta_{\tau_0}, \tau_0) = \int R(\delta_{\tau_0}, \boldsymbol{\theta})\gamma_0(\boldsymbol{\theta})d\boldsymbol{\theta} = \sup_{\boldsymbol{\theta}\in\Theta} R(\delta_{\tau_0}, \boldsymbol{\theta}). \tag{4.23}$$

a) Consideremos un estimador $\delta \neq \delta_{\tau_0}$, luego como

$$MR(\delta) = \sup_{\boldsymbol{\theta}\in\Theta} R(\delta, \boldsymbol{\theta}) \geq R(\delta, \boldsymbol{\theta})$$

tomando esperanza respecto a la distribución a priori τ_0 obtenemos

$$MR(\delta) = \sup_{\boldsymbol{\theta}\in\Theta} R(\delta, \boldsymbol{\theta}) \geq \int R(\delta, \boldsymbol{\theta})\gamma_0(\boldsymbol{\theta})d\boldsymbol{\theta} = r(\delta, \tau_0). \tag{4.24}$$

Como δ_{τ_0} es Bayes respecto de τ_0, se cumple que

$$r(\delta, \tau_0) \geq r(\delta_{\tau_0}, \tau_0). \tag{4.25}$$

Con lo cual, a partir de (4.23), (4.24) y (4.25) obtenemos

$$MR(\delta) \geq r(\delta_{\tau_0}, \tau_0) = \sup_{\boldsymbol{\theta}\in\Theta} R(\delta_{\tau_0}, \boldsymbol{\theta}) = MR(\delta_{\tau_0})$$

y por lo tanto, δ_{τ_0} es minimax.

b) Supongamos ahora que δ_{τ_0} es el único estimador Bayes, luego se cumple

$$r(\delta, \tau_0) > r(\delta_{\tau_0}, \tau_0). \tag{4.26}$$

Con lo cual, utilizando ahora (4.23), (4.24) y (4.26) obtenemos

$$MR(\delta) \geq r(\delta, \tau_0) > r(\delta_{\tau_0}, \tau_0) = MR(\delta_{\tau_0})$$

y por lo tanto, δ_{τ_0} es el único estimador minimax.

c) Sea τ otra distribución a priori y δ_τ el estimador Bayes respecto de τ. Luego, por ser δ_τ Bayes se cumple

$$r(\delta_\tau, \tau) \leq r(\delta_{\tau_0}, \tau). \tag{4.27}$$

Por otra parte, como el riesgo de δ_{τ_0} es constante se verifica

$$\begin{aligned} r(\delta_{\tau_0}, \tau) &= \int R(\delta_{\tau_0}, \boldsymbol{\theta})\gamma(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= \sup_{\boldsymbol{\theta}\in\Theta} R(\delta_{\tau_0}, \boldsymbol{\theta}) = r(\delta_{\tau_0}, \tau_0), \end{aligned} \tag{4.28}$$

Por lo tanto, (4.27) y (4.28) nos permiten concluir que

$$r(\delta_\tau, \tau) \leq r(\delta_{\tau_0}, \tau_0)$$

con lo cual, τ_0 es la distribución menos favorable.

Ejemplo 3. Consideremos el Ejemplo 1 de estimación bayesiana para la familia binomial, usando distribuciones a priori en la familia $\beta(a, b)$ y como fúnción de pérdida la función $\ell(\theta, d) = (\theta - d)^2$. Luego hemos visto que el único estimador de Bayes está dado por

$$\delta_{a,b} = \frac{T + a}{n + a + b},$$

con $T = \sum_{i=1}^n X_i$.

Si encontramos a y b tales que $R(\delta_{a,b}, \theta)$ es constante, ese estimador será minimax y la distribución a priori correspondiente será la distribución menos favorable. Como $E_\theta(T) = n\theta$ y $var(T) = n\theta(1 - \theta)$ se tiene

$$E_\theta(\delta_{a,b}) = \frac{n\theta + a}{n + a + b}, \tag{4.29}$$

y

$$var_\theta(\delta_{a,b}) = \frac{n\theta(1-\theta)}{(n+a+b)^2}, \tag{4.30}$$

Luego, usando (4.29) y (4.30) se deduce que

$$\begin{aligned} R(\delta_{a,b},\theta) &= E((\delta_{a,b}-\theta)^2) \\ &= var_\theta(\delta_{a,b}) + (\theta - E_\theta(\delta_{a,b}))^2 \\ &= \frac{n\theta(1-\theta)}{(n+a+b)^2} + \left(\theta - \frac{n\theta+a}{n+a+b}\right)^2 \\ &= \frac{n\theta(1-\theta)+(a+b)^2\theta^2 - 2a(a+b)\theta + a^2}{(n+a+b)^2} \\ &= \frac{(-n+(a+b)^2)\theta^2 + (n-2a(a+b))\theta + a^2}{(n+a+b)^2}. \end{aligned} \tag{4.31}$$

Para que (4.31) sea constante en θ, los coeficientes en θ y θ^2 del numerador deben ser 0. Por lo tanto, se debe cumplir

$$-n + (a+b)^2 = 0, \qquad n - 2a(a+b) = 0$$

La solución de este sistema de ecuaciones es $a = b = \sqrt{n}/2$, y por lo tanto el estimador de Bayes correspondiente, que será minimax, estará dado por

$$\delta_{\text{mmax}} = \frac{T + (\sqrt{n}/2)}{n + \sqrt{n}}. \tag{4.32}$$

La correspondiente función de riesgo está dada por

$$R(\delta_{\text{mmax}},\theta) = \frac{n/4}{(n+\sqrt{n})^2} = \frac{1}{4(\sqrt{n}+1)^2}.$$

El Teorema 1 no nos permite obtener un estimador minimax en el caso de la media de una distribución normal. El siguiente Teorema resultará útil en esa situación.

Teorema 2. *Supongamos que $\delta(\mathbf{X})$ sea un estimador tal que*

(i) $R(\delta,\boldsymbol{\theta}) = C \ \forall \boldsymbol{\theta} \in \Theta$,

(ii) existe una sucesión de distribuciones a priori τ_k tales que

$$\lim_{k\to\infty} r(\delta_{\tau_k},\tau_k) = C.$$

Entonces δ es minimax.
DEMOSTRACIÓN: Sea δ' otro estimador para $q(\boldsymbol{\theta})$. Se cumple entonces que

$$\sup_{\boldsymbol{\theta}} R(\delta', \boldsymbol{\theta}) \geq \int R(\delta', \boldsymbol{\theta}) \gamma_k(\boldsymbol{\theta}) d\boldsymbol{\theta} = r(\delta', \tau_k) \geq r(\delta_{\tau_k}, \tau_k). \tag{4.33}$$

Con lo cual, tomando límite en ambos miembros de (4.33), y usando (ii) se obtiene

$$MR(\delta') = \sup_{\boldsymbol{\theta}} R(\delta', \boldsymbol{\theta}) \geq C = MR(\delta),$$

y por lo tanto, δ es minimax.

Ejemplo 4. Consideremos una muestra aleatoria $\mathbf{X} = (X_1, ..., X_n)$ de una distribución $\mathrm{N}(\theta, \sigma^2)$, donde σ^2 conocida. El estimador $\delta(\mathbf{X}) = \bar{X}$ tiene como función de riesgo $R(\delta, \theta) = \sigma^2/n$, y por lo tanto se cumple la condición (i) del Teorema 2. Por otro lado, consideremos una sucesión de distribuciones a priori τ_k=$\mathrm{N}(0, \rho_k^2)$ con $\rho_k^2 \to +\infty$. Usando la función de pérdida cuadrática, de acuerdo a lo visto en el ejemplo 2, los estimadores de Bayes son

$$\delta_{\tau_k} = w_k \bar{X},$$

donde

$$w_k = \frac{n/\sigma^2}{(n/\sigma^2) + (1/\rho_k^2)}. \tag{4.34}$$

Es fácil ver que

$$\lim_{k \to \infty} w_k = 1 \tag{4.35}$$

y que

$$\lim_{k \to \infty} \rho_k^2 \, (1 - w_k)^2 = \lim_{k \to \infty} \rho_k^2 \frac{1/\rho_k^4}{\left((n/\sigma^2) + (1/\rho_k^2)\right)^2} = 0 \tag{4.36}$$

Por otro lado, se tiene

$$R(\delta_{\tau_k}, \theta) = var_\theta(\delta_{\tau_k}) + (\theta - E_\theta(\delta_{\tau_k}))^2 = w_k^2 \frac{\sigma^2}{n} + (1 - w_k)^2 \theta^2. \tag{4.37}$$

Luego

$$r(\delta_{\tau_k}, \tau_k) = E_{\tau_k}(R(\delta_{\tau_k}, \boldsymbol{\theta})) = w_k^2 \frac{\sigma^2}{n} + (1 - w_k)^2 \rho_k^2.$$

Con lo cual, usando (4.35) y (4.36) se concluye que

$$\lim_{k \to \infty} r(\delta_{\tau_k}, \tau_k) = \frac{\sigma^2}{n}$$

Por lo tanto se cumple la condición (ii) del Teorema 2, y el estimador $\delta(\mathbf{X}) = \bar{X}$ es minimax. El Teorema 2 no nos permite obtener la unicidad del estimador minimax.

Chapter 5

Intervalos y Regiones de Confianza

5.1 Regiones de confianza – Definición y Ejemplos

Consideremos nuevamente el problema de estimación. Dado un vector $\mathbf{X}$ con distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$, un estimador puntual de $\boldsymbol{\theta}$ es una función $\widehat{\boldsymbol{\theta}} = \delta(\mathbf{X})$ que representa un *único* valor que aproxima a $\boldsymbol{\theta}$. Si se da solamente ese valor no se tendrá ninguna idea de la precisión de dicha aproximación, es decir de las posibles diferencias entre $\boldsymbol{\theta}$ y $\widehat{\boldsymbol{\theta}}$. Una forma de obtener información sobre la precisión de la estimación, en el caso de que $\boldsymbol{\theta}$ sea unidimensional, es proporcionar un intervalo $[a(\mathbf{X}), b(\mathbf{X})]$ de manera que la probabilidad de que dicho intervalo contenga el verdadero valor $\boldsymbol{\theta}$ sea alta, por ejemplo, 0.95.

En este caso, la precisión con que se conoce $\boldsymbol{\theta}$ depende de la longitud del intervalo, es decir, de $b(\mathbf{X}) - a(\mathbf{X})$, cuanto más pequeña sea esa longitud, más determinado quedará $\boldsymbol{\theta}$.

Si $\boldsymbol{\theta}$ es un vector de $I\!R^p$, en vez de dar un intervalo para estimarlo, se deberá dar una cierta región de $I\!R^p$, por ejemplo, esférica o rectangular.

La siguiente definición formaliza estos conceptos.

Definición 1: Dado un vector $\mathbf{X}$ con distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$, *una región de confianza* $S(\mathbf{X})$ *para* $\boldsymbol{\theta}$ *con nivel de confianza* $1 - \alpha$ será una función que a cada $\mathbf{X}$ le hace corresponder un subconjunto de Θ de manera que $P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S(\mathbf{X})) = 1 - \alpha$ para todo $\boldsymbol{\theta} \in \Theta$.

Es decir, $S(\mathbf{X})$ cubre el valor verdadero del parámetro con probabilidad

$1-\alpha$. El valor de α debe ser fijado de acuerdo al grado de seguridad con que se quiere conocer $\boldsymbol{\theta}$; generalmente se toma $\alpha = 0.05$ ó $\alpha = 0.01$.

Como caso particular, cuando $\boldsymbol{\theta}$ sea unidimensional se dirá que $S(\mathbf{X})$ es un intervalo de confianza si $S(\mathbf{X})$ es de la forma

$$S(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$$

La longitud de $S(\mathbf{X})$

$$L = b(\mathbf{X}) - a(\mathbf{X})$$

dependerá del nivel α elegido, cuanto más chico sea α, o sea, cuanto más grande sea la probabilidad con que el intervalo cubra al verdadero valor del parámetro, más grande será la longitud de aquél, o sea, menos precisa la estimación de $\boldsymbol{\theta}$.

Ejemplo 1: Sea $X_1, \ldots, X_n$ una muestra de una población con distribución $N(\mu, \sigma_0^2)$ donde μ es desconocido y σ_0^2 conocido. Supongamos que se necesite un intervalo de confianza para μ de nivel $1-\alpha$.

Consideremos $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$. Sabemos que $\overline{X}_n$ tiene distribución $N(\mu, \sigma_0^2/n)$. Luego $V = \sqrt{n}(\overline{X}_n - \mu)/\sigma_0$, tendrá distribución $N(0,1)$. La ventaja de la variable aleatoria V sobre $\overline{X}_n$ es que tiene distribución independiente de μ.

Definimos z_α tal que $P(V \geq z_\alpha) = \alpha$; y por simetría $P(V \leq -z_\alpha) = \alpha$. Luego

$$\begin{aligned} P(-z_{\frac{\alpha}{2}} \leq V \leq z_{\frac{\alpha}{2}}) &= 1 - P(V \leq z_{\frac{\alpha}{2}}) - P(V \leq -z_{\frac{\alpha}{2}}) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \, . \end{aligned}$$

Si reemplazamos V por $\sqrt{n}(\overline{X}_n - \mu)/\sigma_0$ se tendrá

$$P_\mu\left(-z_{\frac{\alpha}{2}} \leq \sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma_0}\right) \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha \, ,$$

con lo cual, despejando resulta

$$P_\mu(\overline{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}) = 1 - \alpha \, .$$

Por lo tanto, un intervalo de confianza para μ de nivel $1-\alpha$ será

$$S(\mathbf{X}) = \left[\overline{X}_n - z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}} \, , \quad \overline{X}_n + z_{\frac{\alpha}{2}}\frac{\sigma_0}{\sqrt{n}}\right]$$

ya que

$$P_\mu[\mu \in S(\mathbf{X})] = 1 - \alpha.$$

Conviene precisar nuevamente el significado de esta igualdad. Para fijar ideas, supongamos $\alpha = 0.05$. La expresión "$S(\mathbf{X})$ cubre a μ con probabilidad 0.95", indica que si un experimentador extrayese un número suficientemente grande de muestras $\mathbf{X}$ de tamaño n de una distribución $N(\mu, \sigma_0^2)$ y construyese las regiones $S(\mathbf{X})$ correspondientes a cada una de ellas, aproximadamente el 95% de estas regiones $S(\mathbf{X})$ contendrán el parámetro μ. Esto es, dada $\mathbf{X}$, la afirmación "$S(\mathbf{X})$ cubre a μ" tiene probabilidad 0.95 de ser correcta y probabilidad 0.05 de ser falsa.

Ejemplo 2: Un físico hace 16 mediciones de cierta magnitud (a determinar), dichas mediciones X_i serán $X_i = \mu + \epsilon_i$ donde ϵ_i son los errores de medición.

Supongamos que los ϵ_i son variables aleatorias independientes con distribución $N(0, 4)$ (dato que se conoce por experimentos anteriores).

Supongamos que el promedio de las 16 observaciones obtenidas es $\overline{X}_{16} = 20$ y consideremos el problema de encontrar un intervalo de confianza para μ con nivel 0.95; luego $\alpha = 0.05$ y de las tablas normales se obtiene $z_{\alpha/2} = z_{0.025} = 1.96$.

Luego el intervalo de confianza será:

$$\left[20 - \frac{1.96\sqrt{4}}{\sqrt{16}}\,, \quad 20 + \frac{1.96\sqrt{4}}{\sqrt{16}}\right] = [19.02\,,\, 20.98],$$

y su longitud es 1.96.

Supongamos ahora que se quiere conocer cuál deberá ser el número de observaciones para que el intervalo sea de longitud 0.1. Entonces

$$0.1 = 1.96\frac{2}{\sqrt{n}} \quad \text{o sea} \quad \sqrt{n} = 1.96\frac{2}{0.1} = 39.2$$

de donde, $n = (39.2)^2 = 1536.64$. Por lo tanto, se necesitan 1537 observaciones para obtener un intervalo con la longitud deseada.

5.2 Procedimientos generales para obtener regiones de confianza

Teorema 1: *Sea* $\mathbf{X}$ *un vector aleatorio cuya distribución pertenece a la familia* $F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$, *y sea* $U = G(\mathbf{X}, \boldsymbol{\theta})$ *una variable aleatoria cuya*

distribución es independiente de $\boldsymbol{\theta}$. *Sean* A *y* B *tales que* $P(A \leq U \leq B) = 1-\alpha$. *Luego, si se define* $S(\mathbf{X}) = \{\boldsymbol{\theta} : A \leq G(\mathbf{X}, \boldsymbol{\theta}) \leq B\}$, *se tiene que* $S(\mathbf{X})$ *es una región de confianza a nivel* $(1-\alpha)$ *para* $\boldsymbol{\theta}$.

DEMOSTRACIÓN:

$$\begin{aligned} P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S(\mathbf{X})) &= P_{\boldsymbol{\theta}}(A \leq G(\mathbf{X}, \boldsymbol{\theta}) \leq B) = \\ &= P_{\boldsymbol{\theta}}(A \leq U \leq B) = P(A \leq U \leq B) = 1-\alpha \end{aligned}$$

la penúltima igualdad es válida pues la distribución de U es independiente de $\boldsymbol{\theta}$.

Cabe preguntarse bajo qué condiciones $S(\mathbf{X})$ es un intervalo, en el caso en que $\boldsymbol{\theta}$ es unidimensional. Notemos que, en ese caso, si $G(\mathbf{X}, \theta)$ es monótona como función de θ, para cada valor de $\mathbf{X}$ dado, entonces

$$h_{\mathbf{X}}(\theta) = G(\mathbf{X}, \theta)$$

tiene inversa.

Supongamos $h_{\mathbf{X}}(\theta)$ creciente, resulta entonces

$$S(\mathbf{X}) = \{\theta : h_{\mathbf{X}}^{-1}(A) \leq \theta \leq h_{\mathbf{X}}^{-1}(B)\}$$

es decir, $S(\mathbf{X})$ es un intervalo.

Si $h_{\mathbf{X}}(\theta)$ es decreciente, resultará en forma análoga,

$$S(\mathbf{X}) = \{\theta : h_{\mathbf{X}}^{-1}(B) \leq \theta \leq h_{\mathbf{X}}^{-1}(A)\}$$

Nota: En el Ejemplo 1, consideramos $U = \sqrt{n}(\overline{X}_n - \mu)/\sigma_0$ y vimos que esta variable aleatoria tiene distribución $N(0,1)$, o sea, independiente de μ. En ese ejemplo tomamos $A = -z_{\alpha/2}$ y $B = z_{\alpha/2}$. También podríamos haber tomado $A = -z_\beta$ y $B = z_\gamma$ donde β y γ son arbitrarios tales que $\beta + \gamma = \alpha$. El hecho de tomar $\beta = \gamma = \alpha/2$ se debe a que de esta forma se obtiene el intervalo más pequeño posible (Ver problema 1 de 5.1).

Veamos que el procedimiento que hemos usado en dicho ejemplo es el que se deduce del Teorema 1.

De acuerdo al Teorema 1,

$$\begin{aligned} S(\mathbf{X}) &= \{\mu : -z_{\frac{\alpha}{2}} \leq G(\mathbf{X}, \mu) \leq z_{\frac{\alpha}{2}}\} = \\ &= \left\{\mu : -z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma_0} \leq z_{\frac{\alpha}{2}}\right\} = \\ &= \left\{\mu : -z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}} + \overline{X}_n \leq \mu \leq \overline{X}_n + z_{\frac{\alpha}{2}} \frac{\sigma_0}{\sqrt{n}}\right\} . \end{aligned}$$

Vamos a tratar de usar un procedimiento similar para el caso de tener una muestra $X_1, X_2, \ldots, X_n$ de una distribución $N(\mu, \sigma^2)$ donde ahora también σ^2 es desconocido. En este caso, parece natural reemplazar σ^2 por un estimador del mismo. Sabemos que el estimador IMVU para σ^2 es

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

y luego podríamos definir

$$U = \frac{\sqrt{n}(\overline{X} - \mu)}{s} \tag{5.1}$$

Para poder aplicar el método que nos proporciona el Teorema 1, debemos demostrar que U tiene una distribución que no depende de μ y σ^2 y, además, debemos conocer esa distribución. Esto se hará en el siguiente Teorema.

Teorema 2: *Sea $X_1, ..., X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Luego*

(i) $V = \sqrt{n}(\overline{X} - \mu)$ *tiene distibución* $N(0, 1)$

(ii) $W = \sum_{i=1}^{n}(X_i - \overline{X})^2/\sigma^2$ *tiene distribución* χ^2 *con n-1 grados de libertad*

(iii) V *y* W *son independientes*

(iv) U *dado por (5.1) tiene distribución* $\mathcal{T}_{n-1}$*, de Student con* $n-1$ *grados de libertad.*

DEMOSTRACIÓN: Sea $Y_i = (X_i - \mu)/\sigma,\ 1 \leq i \leq n$. Luego estas variables forman una muestra aleatoria de una distribución $N(0, 1)$. Además, es fácil verificar que

$$V = \sqrt{n}\,\overline{Y}, \quad W = \sum_{i=1}^{n} Y_i^2, \quad U = \frac{V}{\sqrt{W/(n-1)}}, \tag{5.2}$$

Sea $\mathbf{a}_1$ el vector fila n-dimensional con todas sus componentes iguales a $1/\sqrt{n}$. Como $\|\mathbf{a}_1\| = 1$, se puede completar una base ortonormal $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Sea A la matriz de $n \times n$ cuyas filas son $\mathbf{a}_1$,...$\mathbf{a}_n$. Como las filas de A son ortogonales y de norma 1, la matriz A resulta ortogonal. Consideremos

la transformación $\mathbf{Z} = A\mathbf{Y}$, donde $\mathbf{Y} = (Y_1, ..., Y_n)'$ y $\mathbf{Z} = (Z_1, ..., Z_n)'$. Luego, por una propiedad de los vectores normales respecto de transformaciones ortogonales, las variables $Z_1, ..., Z_n$ son también independientes con distribución $N(0, 1)$. Por otro lado, resulta

$$Z_1 = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}} = \sqrt{n}\,\overline{Y} = V \tag{5.3}$$

y el punto (i) queda demostrado.

Además, se tiene que:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 = \sum_{i=1}^{n} Y_i^2 - Z_1^2. \tag{5.4}$$

Como A es ortogonal se deduce que

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} Y_i^2,$$

y usando (5.2) y (5.4) obtenemos

$$W = \sum_{i=2}^{n} Z_i^2,$$

y por lo tanto queda demostrado (ii).

Como V depende de Z_1 y W de $Z_2, ..., Z_n$ también queda demostrado (iii).

Finalmente, (iv) se deduce de los puntos (i), (ii), (iii) del Teorema y de (5.2).

Estamos ahora en condiciones de encontrar intervalos de confianza para la media, en el caso de una muestra aleatoria con media y varianza desconocidas.

Definamos $t_{n,\alpha}$ por la ecuación

$$P(U > t_{n,\alpha}) = \alpha$$

donde U es una variable aleatoria $\mathcal{T}_n$. Luego, análogamente al caso normal, se tiene:

$$P(-t_{n,\frac{\alpha}{2}} \leq U \leq t_{n,\frac{\alpha}{2}}) = 1 - \alpha$$

Teorema 3: *Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria cuya distribución pertenece a la familia $N(\mu, \sigma^2)$ con μ y σ^2 desconocidos. Luego si*

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \quad y \quad s^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$$

se tiene que un intervalo de confianza con nivel $(1-\alpha)$ para μ está dado por:

$$\left[\overline{X} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\ , \quad \overline{X} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$$

DEMOSTRACIÓN: Por el Teorema 2 se tiene que $U = \sqrt{n}(\overline{X} - \mu)/s$ tiene distribución $\mathcal{T}_{n-1}$ y luego

$$P(-t_{n-1,\frac{\alpha}{2}} \leq U \leq t_{n-1,\frac{\alpha}{2}}) = 1 - \alpha\ .$$

Luego, por el Teorema 1

$$\left\{\mu : -t_{n-1,\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{s}\sqrt{n} \leq t_{n-1,\frac{\alpha}{2}}\right\}$$

es una región de confianza para μ con nivel $1 - \alpha$. Pero esta región es equivalente a

$$\left\{\mu : \overline{X} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right\}\ .$$

En el próximo Teorema encontraremos intervalos de confianza para la varianza, en el caso de una muestra normal, con media conocida o no.

Definamos $\chi^2_{n,\alpha}$ por la ecuación

$$P(U > \chi^2_{n,\alpha}) = \alpha$$

donde U es una variable aleatoria con distribución χ^2_n.

Teorema 4: *Sea $X_1, \ldots, X_n$ una muestra aleatoria cuya distribución pertenece a la familia $N(\mu, \sigma^2)$. Sean β y γ tales que $\beta + \gamma = \alpha$*

(i) *Si μ es conocido, un intervalo de confianza de nivel $1 - \alpha$ para σ^2 está dado por:*

$$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,\beta}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{n,1-\gamma}}\right]$$

(ii) *Si μ es desconocido, un intervalo de confianza de nivel $1-\alpha$ para σ^2 está dado por:*

$$\left[\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,\beta}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,1-\gamma}}\right]$$

DEMOSTRACIÓN: (i) Sea $W = \sum_{i=1}^n (X_i - \mu)^2/\sigma^2$. Como las variables $Y_i = (X_i - \mu)/\sigma$ son independientes, con distribución $N(0,1)$ y $W = \sum_{i=1}^n Y_i^2$ entonces W tiene distribución χ^2_n. Luego:

$$\begin{aligned} P(\chi^2_{n,1-\gamma} \leq W \leq \chi^2_{n,\beta}) &= P(W \geq \chi^2_{n,1-\gamma}) - P(W > \chi^2_{n,\beta}) = \\ &= 1 - \gamma - \beta = 1 - \alpha \end{aligned}$$

Entonces, una región de confianza a nivel $1-\alpha$ está dada por

$$\left\{\sigma^2 : \chi^2_{n,1-\gamma} \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \leq \chi^2_{n,\beta}\right\} =$$

$$= \left\{\sigma^2 : \frac{1}{\chi^2_{n,\beta}} \leq \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \leq \frac{1}{\chi^2_{n,1-\gamma}}\right\}$$

y esto es equivalente a la región definida en (i).

(ii) Definamos ahora

$$W = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2}$$

Sabemos por el Teorema 2 (ii) que W tiene distribución χ^2_{n-1}. Por lo tanto:

$$P(\chi^2_{n-1,1-\gamma} \leq W \leq \chi^2_{n-1,\beta}) = 1 - \alpha$$

Entonces, una región de confianza de nivel $1-\alpha$ está dada por:

$$\left\{\sigma^2 : \chi^2_{n-1,1-\gamma} \leq \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \leq \chi^2_{n-1,\beta}\right\}$$

$$= \left\{\sigma^2 : \frac{1}{\chi^2_{n-1,\beta}} \leq \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \leq \frac{1}{\chi^2_{n-1,1-\gamma}}\right\}$$

$$= \left\{\sigma^2 : \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,\beta}} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,1-\gamma}}\right\} .$$

5.3 Procedimiento en dos pasos para encontrar un intervalo de longitud prefijada para la media de una $N(\mu, \sigma^2)$, μ y σ desconocidos

Volvamos a considerar el intervalo de confianza para μ cuando σ^2 es desconocido, en el caso de una muestra con distribución $N(\mu, \sigma^2)$. La longitud de dicho intervalo, $L(X_1, \ldots, X_n)$, está dada por

$$L(X_1, \ldots, X_n) = 2t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Como se ve, este intervalo tiene longitud variable, ya que depende de s, que es una variable aleatoria dependiente de los valores que toma la muestra. Luego, es imposible calcular n de modo que la longitud del intervalo sea igual a un número prefijado. Esto es comprensible, ya que lógicamente cuanto más grande sea la varianza, más grande debe ser la muestra necesaria para obtener la misma precisión en la estimación de μ. Como σ^2 no es conocida, no se podrá asegurar con una muestra de tamaño fijo una determinada precisión, es decir, una determinada longitud del intervalo. Una manera de solucionar este problema es tomando dos muestras, una inicial para estimar σ^2, y en base a esta estimación, determinar el tamaño de otra muestra complementaria que nos permita obtener un intervalo con la longitud deseada.

Seguidamente describimos el método. Supongamos que se quiera obtener un intervalo de confianza de longitud L para la media μ, de una población normal con media y varianza desconocida. Se toma una muestra inicial de tamaño m : $X_1, \ldots, X_m$. Este valor m inicial, puede ser cualquier valor mayor que dos. A partir de este valor inicial estimamos σ^2 por:

$$s_m^2 = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \overline{X}_m)^2 \qquad \text{donde} \quad \overline{X}_m = \frac{1}{m} \sum_{i=1}^{m} X_i$$

Luego, la muestra complementaria se debe tomar de tamaño n donde n satisface

$$2t_{m-1,\frac{\alpha}{2}} \frac{s_m}{\sqrt{m+n}} \leq L \tag{5.5}$$

Sea $X_{m+1}, \ldots, X_{m+n}$ la muestra complementaria y

$$\overline{X}_{m+n} = \frac{1}{m+n} \sum_{i=1}^{m+n} X_i$$

El intervalo de confianza de nivel $1-\alpha$ estará dado por:

$$\left[\overline{X}_{m+n} - t_{m-1,\frac{\alpha}{2}} \frac{s_m}{\sqrt{m+n}}\ , \quad X_{m+n} + t_{m-1,\frac{\alpha}{2}} \frac{s_m}{\sqrt{m+n}}\right] \tag{5.6}$$

Este intervalo tiene longitud $2t_{m-1,\frac{\alpha}{2}}\ s_m/\sqrt{m+n}$ que por (5.5) es menor o igual que L.

El siguiente Teorema muestra que el intervalo dado por (5.6) es un intervalo de confianza para μ de nivel $1-\alpha$.

Teorema 5: *Sean $X_1, ...X_n$ variables aleatorias independientes con distribución $N(\mu, \sigma^2)$, donde n se elige satisfaciendo (5.5). Luego el intervalo dado por (5.6) es un intervalo de confianza de nivel 1-α de longitud menor o igual que L.*

DEMOSTRACIÓN: Comencemos por mostrar las siguientes proposiciones:

(i) $W = (m-1)s_m^2/\sigma^2$ tiene distribución χ^2_{m-1}

(ii) $V = \sqrt{m+n}(\overline{X}_{m+n} - \mu)/\sigma$ tiene distribución $N(0,1)$

(iii) V y W son independientes

(iv) $\sqrt{m+n}(\overline{X}_{m+n} - \mu)/s_m$ tiene distribución $\mathcal{T}_{m-1}$

En el Teorema 2 ya ha sido probado (i).
Podría parecer que (ii) fue demostrada en el mismo Teorema. Sin embargo, esto es no es cierto ya que lo que se demostró es que el promedio normalizado de observaciones $N(\mu, \sigma^2)$ tiene distribución $N(0,1)$, para un tamaño de muestra fijo. En nuestro caso, n es un número aleatorio, ya que depende del valor s_m, obtenido con las primeras m observaciones. Comencemos obteniendo la función de distribución conjunta de V y W, $F_{VW}(v,w) = P(V \leq v, W \leq w)$.
Llamemos A_i al evento $\{n = i\}$. Los sucesos A_i son obviamente disjuntos y además $\bigcup_{i=1}^{\infty} A_i = \Omega$, donde Ω es el espacio muestral.
Dado un evento cualquiera A, se tiene

$$\begin{aligned} A &= \bigcup_{i=1}^{\infty} (A \cap A_i) \\ P(A) &= \sum_{i=1}^{\infty} P(A \cap A_i), \end{aligned}$$

y por lo tanto,

$$\begin{aligned} F_{VW}(v,w) &= P(V \le v, W \le w) = \sum_{i=0}^{\infty} P(V \le v\,,\, W \le w\,,\, n=i) \\ &= \sum_{i=0}^{\infty} P(\sqrt{m+i}\frac{(\overline{X}_{m+i}-\mu)}{\sigma} \le v\,,\, W \le w\,,\, n=i)\,. \end{aligned}$$

En virtud del Teorema 2, se tiene que $\sum_{j=1}^{m} X_j$ es independiente de s_m y por otra parte, cada X_j con $j > m$ también es independiente de s_m. Por lo tanto, como $\overline{X}_{m+i} = (1/(m+i))(\sum_{j=1}^{m} X_j + \sum_{j=m+1}^{m+i} X_j)$ se deduce que $\overline{X}_{m+i}$ es independiente de s_m.

Por otro lado, de acuerdo con su definición, n depende sólo de s_m. Luego, el suceso

$$\{\sqrt{m+i}\frac{(\overline{X}_{m+i}-\mu)}{\sigma} \le v\}$$

es independiente de $\{W \le w\} \cap \{n = i\}$ y por lo tanto,

$$F_{VW}(v,w) = \sum_{i=1}^{\infty} P(\sqrt{m+i}\frac{(\overline{X}_{m+i}-\mu)}{\sigma} \le v) P(W \le w\,,\, n=i)\,.$$

Pero, por el Teorema 2, para cada i fijo $\sqrt{m+i}(\overline{X}_{m+i}-\mu)/\sigma$ tiene distribución $N(0,1)$. Luego si $\Phi(v)$ es la función de distribución de una variable $N(0,1)$, se tendrá

$$\begin{aligned} F_{VW}(v,w) &= \sum_{i=1}^{\infty} \Phi(v) P(W \le w\,,\, n=i) \\ &= \Phi(v) \sum_{i=0}^{\infty} P(W \le w\,,\, n=i)\,. \end{aligned}$$

Pero $\sum_{i=1}^{\infty} P(W \le w\,,\, n=i) = P(W \le w) = F_W(w)$. Por lo tanto, se tiene

$$F_{VW}(v,w) = \Phi(v) F_W(w) \tag{5.7}$$

y como

$$F_V(v) = \lim_{w\to\infty} F_{VW}(v,w) = \Phi(v) \lim_{w\to\infty} F_W(w) = \Phi(v)\,,$$

hemos demostrado (ii).

Para demostrar (iii) reemplacemos en (5.7) $\Phi(v)$ por $F_V(v)$ y obtenemos

$$F_{VW}(v,w) = F_V(v)F_W(w)$$

lo que implica que V y W son independientes.
(iv) se deduce inmediatamente de (i), (ii) y (iii), teniendo en cuenta que

$$\frac{\sqrt{m+n}(\overline{X}_{m+n} - \mu)}{s_m} = \frac{\sqrt{m+n}(\overline{X}_{m+n} - \mu)/\sigma}{((m-1)s_m^2/(m-1)\sigma^2)^{1/2}}$$

Llamemos U a esta última variable, de acuerdo a (iv) se tiene que U tiene distribución independiente de μ y σ^2 y además

$$P(-t_{m-1,\frac{\alpha}{2}} \leq U \leq t_{m-1,\frac{\alpha}{2}}) = 1 - \alpha$$

Luego, de acuerdo con el método general para obtener regiones de confianza, se tendrá que una región de confianza para μ de nivel $(1-\alpha)$ estará dada por:

$$\begin{aligned} & \left\{\mu : -t_{m-1,\frac{\alpha}{2}} \leq \frac{\sqrt{m+n}(\overline{X}_{m+n} - \mu)}{s_m} \leq t_{m-1,\frac{\alpha}{2}}\right\} \\ = & \left\{\mu : \overline{X}_{m+n} - t_{m-1,\frac{\alpha}{2}}\frac{s_m}{\sqrt{m+n}} \leq \mu \leq \overline{X}_{m+n} + t_{m-1,\frac{\alpha}{2}}\frac{s_m}{\sqrt{m+n}}\right\} . \end{aligned}$$

Nota: El tamaño de la muestra inicial, m, puede ser, en principio, cualquiera con la condición de que sea mayor que dos. El valor más conveniente a usar dependerá del conocimiento previo que se tenga sobre σ^2. Si m es muy pequeño, la estimación de σ^2 será poco confiable y habrá que tomar una segunda muestra grande, con lo cual aumentará el costo. Si se toma muy grande, es probable que se tomen más observaciones que las necesarias. Lo ideal sería elegir m cerca del número total de observaciones que serían necesarias si se conociera σ^2.

5.4 Intervalos de confianza para diferencia de medias de una distribución normal

5.4.1 Muestras independientes

Supongamos primero que se tienen dos muestras aleatorias $X_1, \ldots, X_{n_1}$ y $Y_1, \ldots, Y_{n_2}$ independientes entre sí, de distribuciones $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$

respectivamente con μ_1, μ_2 y σ^2 desconocidos, y se desea encontrar un intervalo de confianza para $\lambda = \mu_1 - \mu_2$. Observemos que $\hat{\lambda} = \overline{Y} - \overline{X}$ es un estimador insesgado de λ. Es fácil demostrar utilizando el Teorema 2 de la sección 3.12 que este estimador es IMVU.

La varianza de este estimador es

$$\sigma^2_{\hat{\lambda}} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \tag{5.8}$$

Por lo tanto,

$$U = \frac{\hat{\lambda} - \lambda}{\sigma_{\hat{\lambda}}} = \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \cdot \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma} \tag{5.9}$$

tiene distribución N(0,1).

Como σ es desconocido, no podemos utilizar U para encontrar un intervalo de confianza para λ. La solución a este problema es reemplazar σ por un estimador. Un estimador insesgado de σ^2 es

$$s^2 = \frac{1}{n_1 + n_2 - 2} \left(\sum_{i=1}^{n_1} (X_i - \overline{X})^2 + \sum_{i=1}^{n_2} (Y_i - \overline{Y})^2 \right)$$

Para demostrar que s^2 es insesgado basta recordar que de acuerdo a lo visto en el Capítulo 3 se tiene

$$E(\sum_{i=1}^{n_1} (X_i - \overline{X})^2) = (n_1 - 1)\sigma^2$$

y

$$E(\sum_{i=1}^{n_2} (Y_i - \overline{Y})^2) = (n_2 - 1)\sigma^2 \ .$$

También del Teorema 2 de la Sección 3.12 se puede deducir que s^2 es IMVU. Luego, definimos el estadístico T reemplazando en U el parámetro σ por el estimador s, es decir,

$$T = \frac{\hat{\lambda} - \lambda}{\hat{\sigma_{\hat{\lambda}}}} = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{s} \tag{5.10}$$

donde

$$\hat{\sigma}^2_{\hat{\lambda}} = s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \tag{5.11}$$

El siguiente Teorema prueba que T tiene distribución de Student con $n_1 + n_2 - 2$ grados de libertad

Teorema 1: *Sean $X_1, ..., X_{n_1}$ y $Y_1, ..., Y_{n_2}$ dos muestras aleatorias independientes de las distribuciones $N(\mu_1, \sigma^2)$ y $N(\mu_2, \sigma^2)$ respectivamente. Sean*

$$\begin{aligned} V &= \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2}{\sigma^2} \\ W &= \frac{\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2}{\sigma^2} \end{aligned}$$

Luego

(i) U definida en (5.9), V y W son variables aleatorias independientes con distribuciones $N(0,1)$, $\chi^2_{n_1-1}$ y $\chi^2_{n_2-1}$ respectivamente.

(ii) La variable

$$Z = V + W$$

tiene distribución $\chi^2_{n_1+n_2-2}$.

(iii) La variable T definida en (5.10) tiene distribución $\mathcal{T}_{n_1+n_2-2}$.

(iv) El intervalo

$$\left[\hat{\lambda} - t_{n_1+n_2-2,\frac{\alpha}{2}}\ \hat{\sigma}_{\hat{\lambda}},\ \hat{\lambda} + t_{n_1+n_2-2,\frac{\alpha}{2}}\ \hat{\sigma}_{\hat{\lambda}}\right]$$

es un intervalo de confianza a nivel $1 - \alpha$ para $\lambda = \mu_1 - \mu_2$.

DEMOSTRACIÓN: Ya hemos demostrado que U tiene distribución $N(0,1)$.

Por otra parte, en el Teorema 2 de la Sección 5.2, se demostró la independencia entre $\overline{X}$ y V y entre $\overline{Y}$ y W. Como además resulta $\overline{X}$ independiente de W (la primera depende de $X_1, \ldots, X_{n_1}$ y la segunda de $Y_1, \ldots, Y_{n_2}$) y $\overline{Y}$ independiente de V, resulta U independiente de V y W. En el mismo Teorema se demostró que V y W tienen distribuciónes $\chi^2_{n_1-1}$ y $\chi^2_{n_2-1}$, respectivamente. Resulta entonces claro que V y W son también independientes.

Para demostrar (ii) basta utilizar el hecho de que suma de variables χ^2 independientes tiene también distribución χ^2 con número de grados de libertad igual a la suma de los grados de libertad de los sumandos.

El resultado (iii) resulta inmediato de los puntos (i) y (ii). El resultado (iv) resulta de aplicar (ii) y el Teorema 1 de la Sección 5.2.

En el caso más simple en que σ^2 sea conocido, se puede también encontrar fácilmente un intervalo de confianza para λ utilizando el estadístico U.

Si $X_1, \ldots, X_{n_1}$ y $Y_1, \ldots, Y_{n_2}$ son muestras aleatorias independientes entre sí de distribuciones $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ con μ_1, μ_2, σ_1^2 y σ_2^2 desconocidos ($\sigma_1^2 \neq \sigma_2^2$), el problema de encontrar una región de confianza para $\mu_1 - \mu_2$ con nivel exacto $1-\alpha$ no tiene una solución totalmente satisfactoria. Este problema se conoce con el nombre de Behrens–Fisher. Sin embargo, es posible encontrar en forma sencilla un intervalo de confianza para $\mu_1 - \mu_2$ de nivel asintótico $1 - \alpha$ (ver definición 1 y problema 7 de 5.6).

Nota 1: Si $X_1, \ldots, X_{n_1}$ y $Y_1, \ldots, Y_{n_2}$ son muestras aleatorias independientes entre sí de distribuciones $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ respectivamente, con μ_1, μ_2 conocidos o no, entonces:

(1) Si $\sigma_1^2 = \sigma_2^2 = \sigma^2$ se pueden encontrar intervalos de confianza para σ^2 (o para σ) (ver problema 1 de 5.4).

(2) Si no se puede suponer $\sigma_1^2 = \sigma_2^2$ es posible encontrar intervalos de confianza para σ_2^2/σ_1^2 (o para σ_2/σ_1) (ver problema 2 de 5.4).

5.4.2 Muestras apareadas

Supongamos ahora que $(X_1, Y_1), \ldots, (X_n, Y_n)$ es una muestra aleatoria de una distribución normal bivariada $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ con $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho$ desconocidos y que se desea encontrar un intervalo de confianza para $\lambda = \mu_1 - \mu_2$.

En este caso podemos definir las variables $Z_i = X_i - Y_i$, $1 \leq i \leq n$. Estas variables forman una muestra de una distribución N(λ,σ_Z^2), con

$$\sigma_Z^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2,$$

y por lo tanto, de acuerdo a lo visto en el Teorema 3 de la Sección 5.2 tenemos que un intervalo de confianza de nivel $1 - \alpha$ está dado por

$$\left[\overline{Z} - t_{n-1,\frac{\alpha}{2}} \frac{s_Z}{\sqrt{n}}, \overline{Z} + t_{n-1,\frac{\alpha}{2}} \frac{s_Z}{\sqrt{n}}\right] \tag{5.12}$$

donde

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i, \quad s_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2 \, .$$

Nota 2: Muchas veces, en los casos reales, interesará decidir antes de tomar la muestra, si conviene usar un diseño de muestras aleatorias independientes entre sí provenientes de distribuciones $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2)$ o muestras apareadas provenientes de una distribución bivariada, $N(\mu_1, \mu_2, \sigma^2, \sigma^2, \rho)$.

Por ejemplo, si se quiere estimar la diferencia de rendimientos de dos variedades de un cereal, uno podría preguntarse cuál de los dos diseños siguientes proveerá más información sobre esta diferencia:

(i) Elegir al azar en el terreno considerado $2n$ parcelas de área A. En n de ellas elegidas al azar cultivar la variedad 1 y en en los restantes cultivar la variedad 2.

(ii) Elegir al azar n parcelas de área 2A y dividir cada una de ellas en dos mitades de la misma área. y luego éstas en dos mitades. En cada mitad de una parcela cultivar una variedad distinta.

En el primer caso, tendríamos un diseño correspondiente a muestras aleatorias normales independientes entre sí. En el segundo, uno correspondiente a muestras apareadas que podrían ser consideradas provenientes de una normal bivariada con un cierto cociente de correlación ρ.

Trataremos de determinar cuál de los dos diseños es mejor, comparando las longitudes de los intervalos de confianza respectivos. Para esto supondremos que las varianzas para los rendimientos de ambos cereales son los mismos.

Para el caso de muestras independientes tendremos $n_1 = n_2 = n$, y la longitud del intervalo viene dado por

$$L_1 = 2t_{2n-2,\frac{\alpha}{2}}\sqrt{2\frac{s^2}{n}}$$

donde

$$s^2 = \frac{1}{2n-2}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right)$$

y para el caso de muestras para muestras apareadas

$$L_2 = 2t_{n-1,\frac{\alpha}{2}}\sqrt{\frac{s_Z^2}{n}}$$

donde

$$s_Z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \overline{Z})^2 \ .$$

Como estas longitudes dependen de la muestra considerada, y por lo tanto son aleatorias, consideraremos cuál diseño nos provee el intervalo con menor longitud cuadrada esperada. Es decir, compararemos las esperanzas de los cuadrados de las longitudes. Se toman cuadrados por la única razón de simplificar el cálculo de las esperanzas. Como s^2 y s_Z^2 son estimadores insesgados de σ^2 y de $\sigma_Z^2 = 2(1-\rho)\sigma^2$, se tiene

$$E(L_1^2) = \frac{4 \times 2\sigma^2 t^2_{2n-2,\frac{\alpha}{2}}}{n}$$

y en el caso de muestras apareadas

$$E(L_2^2) = \frac{4 \times 2\sigma^2(1-\rho) t^2_{n-1,\frac{\alpha}{2}}}{n}$$

Luego resulta

$$\frac{E(L_2^2)}{E(L_1^2)} = (1-\rho)\frac{t^2_{n-1,\frac{\alpha}{2}}}{t^2_{2n-2,\frac{\alpha}{2}}}$$

Por lo tanto será mejor tomar muestras apareadas si

$$(1-\rho)\frac{t^2_{n-1,\frac{\alpha}{2}}}{t^2_{2n-2,\frac{\alpha}{2}}} < 1$$

o sea si

$$\rho > 1 - \frac{t^2_{2n-2,\frac{\alpha}{2}}}{t^2_{n-1,\frac{\alpha}{2}}} \tag{5.13}$$

Se puede mostrar que $t_{n,\frac{\alpha}{2}}$ tiende a $z_{\frac{\alpha}{2}}$ en forma monótona decreciente cuando $n \to \infty$. Luego se tendrá que

$$\lambda = 1 - \frac{t^2_{2n-2,\frac{\alpha}{2}}}{t^2_{n-1,\frac{\alpha}{2}}} > 0$$

tendiendo a 0 cuando $n \to \infty$.

Luego, para que sea más conveniente tomar muestras apareadas es una condición necesaria que $\rho > 0$. Para muestras grandes esta condición es prácticamente suficiente ya que λ se hace muy pequeño.

Sea, por ejemplo, $n = 20$ y $\alpha = 0.05$, luego $\lambda = 0.03$. Luego basta que $\rho > 0.03$ para que el diseño apareado sea más eficiente. Para un ejemplo práctico, ver ejercicio 3 de 5.4. Por otra parte, por (5.13) resulta que en caso de tomarse muestras apareadas convendrá elegir los pares de manera que ρ sea lo más grande posible.

5.5 Optimalidad de los intervalos de confianza

Sea $\mathbf{X}$ un vector cuya distribución pertenece a la familia $F(\mathbf{x}, \theta)$ con $\theta \in \Theta \subset \mathbb{R}$ y sea $S(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$ un intervalo de confianza con nivel $1 - \alpha$ para θ. Como ya lo hemos observado en 5.1, la precisión de nuestra estimación vendrá dada por la longitud del intervalo, es decir, por $L(\mathbf{X}) = b(\mathbf{X}) - a(\mathbf{X})$ y por lo tanto, será conveniente que ésta fuese lo menor posible. Como ya lo hemos visto, $L(\mathbf{X})$ es en general una variable aleatoria; luego parece razonable como criterio para medir la bondad de un intervalo de confianza considerar $E_\theta(L(\mathbf{X}))$.

Luego, un intervalo de confianza con nivel $1 - \alpha$, $[a(\mathbf{X}), b(\mathbf{X})]$, puede ser considerado óptimo si, para todo otro intervalo de confianza de nivel $1 - \alpha$, $[a'(\mathbf{X}), b'(\mathbf{X})]$ se tiene

$$E_\theta(b(\mathbf{X}) - a(\mathbf{X})) \leq E_\theta(b'(\mathbf{X}) - a'(\mathbf{X})) \quad \forall\, \theta \in \Theta\ .$$

Sin embargo, igual que en el caso de estimación puntual, es posible mostrar que salvo ejemplos triviales no existen intervalos con esta propiedad. La única forma de encontrar intervalos óptimos es restringir la clase de posibles intervalos.

Una forma de restringir los posibles intervalos de confianza o en general las regiones de confianza, es exigiendo la siguiente propiedad.

Definición 1: Se dirá que una región $S(\mathbf{X})$ es *insesgada* si

$$P_\theta(\theta \in S(\mathbf{X})) \geq P_\theta(\theta' \in S(\mathbf{X})) \quad \forall\, \theta, \theta' \in \Theta\ .$$

Es decir que $S(\mathbf{X})$ es insesgado si el valor verdadero θ tiene mayor probabilidad de estar en la región que cualquier otro valor θ'.

Luego parece natural buscar el intervalo de confianza de menor longitud entre los intervalos de confianza insesgados. Luego surge la siguiente definición:

Definición 2: Se dirá que un intervalo de confianza $S(\mathbf{X})$ es *insesgado de mínima longitud esperada uniformemente* en θ (IMLEU) con nivel $(1 - \alpha)$ si

a) $S(\mathbf{X})$ es insesgado y tiene nivel $(1 - \alpha)$.

b) Sea $S(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$. Luego si $S'(\mathbf{X}) = [a'(\mathbf{X}), b'(\mathbf{X})]$ es otro intervalo insesgado de nivel $1 - \alpha$, se tiene

$$E_\theta(b(\mathbf{X}) - a(\mathbf{X})) \leq E_\theta(b'(\mathbf{X}) - a'(\mathbf{X})) \quad \forall\, \theta \in \Theta\ .$$

Se puede mostrar que los intervalos obtenidos para μ cuando $X_1, \ldots, X_n$ es una muestra aleatoria de $N(\mu, \sigma^2)$ para el caso de σ^2 conocido o desconocido (en Ejemplo 1 de 5.1 y Teorema 3 de 5.2) son realmente IMLEU. También, los intervalos obtenidos para σ^2 cuando μ es conocido o desconocido en el Teorema 4 de 5.2 es IMLEU, si β y γ se eligen de manera que la longitud esperada sea mínima. Se puede mostrar que para n grande estos β y γ se aproximan a $\alpha/2$ (ver [3]). Los procedimientos desarrollados en 5.4 para encontrar intervalos de confianza para las diferencias de medias también son IMLEU.

El estudio detallado de la optimalidad de estos procedimientos puede verse en Pratt [2]. Estos resultados dependen de resultados relacionados con la teoría de tests óptimos que puede verse en Lehmann [1].

5.6 Regiones de confianza con nivel asintótico $(1 - \alpha)$

En muchos problemas estadísticos, es imposible o muy complicado encontrar regiones de confianza con un nivel dado. En su reemplazo se pueden construir regiones cuyo nivel sea aproximadamente el deseado, tendiendo a él a medida que el tamaño de la muestra aumenta. La siguiente definición formaliza esta idea.

Definición 1: Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $F(x, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Se dice que $S_n(X_1, \ldots, X_n)$ es una sucesión de regiones de confianza con nivel asintótico $1 - \alpha$ si:

$$\lim_{n \to \infty} P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S_n(X_1, \ldots, X_n)) = 1 - \alpha \quad \forall\, \theta \in \Theta\ .$$

El siguiente Teorema nos da un procedimiento para construir intervalos de confianza con nivel asintótico $(1 - \alpha)$.

Teorema 1: *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $F(x, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Supongamos que para cada n se tienen definidas funciones $U_n = G_n(X_1, \ldots, X_n, \boldsymbol{\theta})$ tales que U_n converge a U en distribución, donde U es una variable aleatoria con distribución independiente de $\boldsymbol{\theta}$. Sean A y B puntos de continuidad de F_U, tales que $P(A \leq U \leq B) = 1 - \alpha$. Definamos $S_n(X_1, \ldots, X_n) = \{\boldsymbol{\theta} : A \leq G_n(X_1, \ldots, X_n, \boldsymbol{\theta}) \leq B\}$. Luego, $S_n(X_1, \ldots, X_n)$ es una sucesión de regiones de confianza con nivel asintótico $(1 - \alpha)$.*

DEMOSTRACIÓN:

$$\begin{aligned} P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S_n(X_1,\ldots,X_n)) &= P_{\boldsymbol{\theta}}(A \le G_n(X_1,\ldots,X_n,\boldsymbol{\theta}) \le B) \\ &= P_{\boldsymbol{\theta}}(A \le U_n \le B) \end{aligned}$$

Luego, $\lim_{n\to\infty} P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S_n(X_1,\ldots,X_n)) = \lim_{n\to\infty} P_{\boldsymbol{\theta}}(A \le U_n \le B) = P_{\boldsymbol{\theta}}(A \le U \le B) = P(A \le U \le B) = 1-\alpha$.

Ejemplo 1: Sea $X_1,\ldots,X_n$ una muestra independiente de una distribución $B_i(\theta,1)$. Definamos:

$$U_n = \frac{\sum_{i=1}^n X_i - n\theta}{\sqrt{n\theta(1-\theta)}}$$

Sabemos por el Teorema Central del Límite que U_n converge en distribución a una ley $N(0,1)$; por lo tanto, una sucesión de regiones de confianza con nivel asintótico $1-\alpha$ vendrá dada por:

$$\begin{aligned} & S_n(X_1,\ldots,X_n) = \left\{\theta : -z_{\frac{\alpha}{2}} \le \frac{\sum_{i=1}^n X_i - n\theta}{\sqrt{n\theta(1-\theta)}} \le z_{\frac{\alpha}{2}}\right\} \\ = & \left\{\theta : \left(\frac{\sum_{i=1}^n X_i - n\theta}{\sqrt{n\theta(1-\theta)}}\right)^2 \le z^2_{\frac{\alpha}{2}}\right\} \\ = & \left\{\theta : \left(\sum_{i=1}^n X_i\right)^2 + n^2\theta^2 - 2n\theta\sum_{i=1}^n X_i \le z^2_{\frac{\alpha}{2}}\, n\theta(1-\theta)\right\} \\ = & \left\{\theta : \theta^2(n^2 + nz^2_{\frac{\alpha}{2}}) - \theta\left(2n\sum_{i=1}^n X_i + z^2_{\frac{\alpha}{2}}\, n\right) + \left(\sum_{i=1}^n X_i\right)^2 \le 0\right\} \\ = & \ [\widehat{\theta}_1, \widehat{\theta}_2] \end{aligned}$$

donde $\widehat{\theta}_1$ y $\widehat{\theta}_2$ son las raíces de la ecuación

$$\theta^2(n^2 + nz^2_{\frac{\alpha}{2}}) - \theta\left(2n\sum_{i=1}^n X_i + z^2_{\frac{\alpha}{2}}\, n\right) + \left(\sum_{i=1}^n X_i\right)^2 = 0$$

La siguiente propiedad, que daremos sin demostración y que es equivalente a la propiedad 5 de 1.8, nos permitirá encontrar un intervalo de confianza más sencillo para θ en el ejemplo anterior.

Propiedad 1: *Sea X_n una sucesión de variables aleatorias, X una variable aleatoria y a una constante. Supongamos que $X_n \to X$ en distribución.*

Sea además, una sucesión de variables aleatorias Y_n tal que $Y_n \to a$ en probabilidad; luego $Y_n X_n \to aX$ en distribución.

Volvamos ahora al Ejemplo 1. U_n se puede escribir

$$U_n = \frac{\sqrt{n}(\overline{X} - \theta)}{\sqrt{\theta(1-\theta)}}$$

Por otro lado, sabemos que un estimador consistente de θ es $\overline{X}$. Luego

$$\frac{1}{\sqrt{\overline{X}(1-\overline{X})}} \to \frac{1}{\sqrt{\theta(1-\theta)}} \quad \text{en probabilidad.}$$

Con lo cual, usando la propiedad anterior y llamando

$$V_n = \frac{\sqrt{n}(\overline{X} - \theta)}{\sqrt{\overline{X}(1-\overline{X})}}$$

se tiene que $V_n \to N(0,1)$ en distribución.

Por lo tanto, un intervalo de confianza para θ de nivel $1-\alpha$, viene dado por

$$\begin{aligned} S_n(X_1, \ldots, X_n) &= \left\{ \theta : -z_{\frac{\alpha}{2}} \leq \frac{\sqrt{n}(\overline{X} - \theta)}{\sqrt{\overline{X}(1-\overline{X})}} \leq z_{\frac{\alpha}{2}} \right\} \\ &= \left[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}} \, , \, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sqrt{\overline{X}(1-\overline{X})}}{\sqrt{n}} \right] \end{aligned}$$

Ejemplo 2: Supongamos que se tiene una muestra aleatoria $X_1, \ldots, X_n$ de una distribución totalmente desconocida y sólo se sabe que $E(X_1) = \mu$ y $\mathrm{Var}(X_1) = \sigma^2$ son finitos. Se quiere encontrar un intervalo de confianza para μ con nivel asintótico $1-\alpha$. Sea

$$U_n = \sqrt{n}(\overline{X} - \mu)/\sigma$$

Por el Teorema Central del Límite, sabemos que $U_n \to N(0,1)$ en distribución.
Por otro lado,

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

es un estimador fuertemente consistente de σ^2. Luego, $s_n \to \sigma$ en probabilidad.
Con lo cual, utilizando la Propiedad 1, si

$$V_n = \sqrt{n}(\overline{X} - \mu)/s_n$$

se tendrá que

$$V_n \to N(0,1) \quad \text{en distribución.}$$

Luego, un intervalo de confianza para μ, con nivel asintótico $1-\alpha$ estará dado por

$$\begin{aligned} S_n(X_1,\ldots,X_n) &= \left\{\mu : -z_{\frac{\alpha}{2}} \le \frac{\sqrt{n}(\overline{X}-\mu)}{s_n} \le z_{\frac{\alpha}{2}}\right\} \\ &= \left[\overline{X} - z_{\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\ ,\ \overline{X} + z_{\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\right]\ . \end{aligned}$$

5.7 Regiones de confianza basadas en estimadores de máxima verosimilitud

Veamos ahora un procedimiento que nos permitirá, en condiciones bastante generales, encontrar regiones de confianza con nivel asintótico $(1-\alpha)$.

Sea $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una distribución con densidad $f(x, \boldsymbol{\theta})$. Sabemos, que bajo condiciones muy generales (ver Capítulo 3) el estimador de máxima verosimilitud,EMV, $\widehat{\boldsymbol{\theta}}$ tiene distribución asintóticamente normal. Más precisamente, cuando $\theta \in \mathbb{R}$ bajo condiciones de regularidad,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \to N(0, \frac{1}{I_1(\theta)}) \quad \text{en distribución,}$$

donde $I_1(\theta)$ es el número de información de Fisher de X_1.
Luego, si llamamos

$$U_n = \sqrt{n}\,\sqrt{I_1(\theta)}\,(\widehat{\theta}_n - \theta)$$

se tendrá que

$$U_n \to N(0,1) \quad \text{en distribución.}$$

Por lo tanto, una región de confianza para θ de nivel asintótico $1-\alpha$ estará dada por

$$S_n = \{\theta : -z_{\frac{\alpha}{2}} \le \sqrt{n}\sqrt{I_1(\theta)}(\widehat{\theta}_n - \theta) \le z_{\frac{\alpha}{2}}\}$$

Obsérvese que éste fue el procedimiento que se usó en el Ejemplo 1 de (5.6)(demostrarlo).

Esta región no tiene porqué ser un intervalo, y puede ser difícil de calcular. En el caso en que $I_1(\theta)$ sea continua, se podrá obtener un intervalo de confianza a nivel asintótico $(1-\alpha)$, de la siguiente forma relativamente simple: Sabemos que $\widehat{\theta}_n \to \theta$ en probabilidad, ya que el E.M.V. es consistente, entonces si $I_1(\theta)$ es continua, se tiene

$$\lim_{n\to\infty} I_1(\widehat{\theta}_n) = I_1(\theta) \quad \text{en probabilidad.}$$

Si llamamos $U_n^* = \sqrt{n}\sqrt{I_1(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta)$, resulta que

$$U_n^* \to N(0,1) \quad \text{en distribución.}$$

Por lo tanto, un intervalo de confianza para θ de nivel de confianza asintótico $1-\alpha$ vendrá dado por:

$$\begin{aligned} S_n &= \{\theta : -z_{\frac{\alpha}{2}} \le \sqrt{n}\sqrt{I_1(\widehat{\theta}_n)}(\widehat{\theta}_n - \theta) \le z_{\frac{\alpha}{2}}\} \\ &= \left[\widehat{\theta}_n - \frac{z_{\frac{\alpha}{2}}}{\sqrt{I_1(\widehat{\theta}_n)}\sqrt{n}}\ ,\ \widehat{\theta}_n + \frac{z_{\frac{\alpha}{2}}}{\sqrt{I_1(\widehat{\theta}_n)}\sqrt{n}}\right] . \end{aligned}$$

La longitud de estos intervalos es

$$L = 2z_{\frac{\alpha}{2}} \frac{1}{\sqrt{n}\sqrt{I_1(\widehat{\theta}_n)}} .$$

Luego, bajo condiciones en que vale el Teorema de consistencia del EMV se tiene

$$\lim_{n\to\infty} \sqrt{n}\, L = 2z_{\frac{\alpha}{2}} \frac{1}{\sqrt{I_1(\theta)}} \quad \text{c.t.p.}$$

y bajo condiciones muy generales, también se puede mostrar que

$$\lim_{n\to\infty} \sqrt{n}\, E_\theta(L) = 2z_{\frac{\alpha}{2}} \frac{1}{\sqrt{I_1(\theta)}} .$$

Puede demostrarse que bajo condiciones muy generales, para todo intervalo $\mathcal{I}$ insesgado, con nivel asintótico $1-\alpha$ se tendrá

$$\lim_{n\to\infty} \sqrt{n}\, E_\theta(L_{\mathcal{I}}) \ge 2z_{\frac{\alpha}{2}} \frac{1}{\sqrt{I_1(\theta)}}$$

donde, $L_{\mathcal{I}}$ indica la longitud del intervalo $\mathcal{I}$. Por lo tanto, los intervalos obtenidos a partir del estimador de máxima verosimilitud pueden considerarse asintóticamente insesgados de menor longitud esperada.

Para ver estas propiedades en detalle, consultar Wilks [4, pp. 374–376].

Luego de la descripción de los métodos para obtener intervalos de confianza a nivel asintótico, podría pensarse en los casos que es posible encontrarlos en lugar de los intervalos exactos. Sin embargo, la convergencia del nivel de confianza al valor deseado depende fuertemente de la distribución y podría ser necesario un tamaño de muestra grande para que la aproximación del nivel asintótico sea aceptable. En general, no se puede determinar el tamaño de muestra n para el cual la aproximación asintótica es suficientemente buena usando consideraciones teóricas. En la mayoría de los casos es necesario estudiar este problema por métodos de Monte Carlo que se estudiarán más adelante.

5.8 Regiones de confianza simultáneas

Supongamos que se tiene un vector aleatorio $\mathbf{X}$ cuya distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ y que $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Ocurre a veces que se tienen regiones de confianza para θ_1 y θ_2 por separado, es decir, se tienen $S_1(\mathbf{X})$ y $S_2(\mathbf{X})$, tales que:

$$P(\theta_1 \in S_1(\mathbf{X})) = 1 - \alpha \quad \text{y} \quad P(\theta_2 \in S_2(\mathbf{X})) = 1 - \alpha$$

pero $P(\theta_1 \in S_1(\mathbf{X}), \theta_2 \in S_2(\mathbf{X})) \leq 1 - \alpha$.

Luego, $S_1(\mathbf{X}) \times S_2(\mathbf{X})$ no es una región de confianza simultánea de nivel $(1 - \alpha)$ para (θ_1, θ_2).

Una forma de conseguir que la probabilidad simultánea de que θ_1 y θ_2 estén en $S_1(\mathbf{X})$ y $S_2(\mathbf{X})$ respectivamente, sea al menos $(1 - \alpha)$ se obtiene considerando regiones de confianza de nivel $(1 - \alpha/2)$ para θ_1 y θ_2, es decir, tales que:

$$P(\theta_1 \in S_1(\mathbf{X})) = 1 - \frac{\alpha}{2} \quad \text{y} \quad P(\theta_2 \in S_2(\mathbf{X})) = 1 - \frac{\alpha}{2} \, .$$

Luego, si A^c indica el complemento del conjunto A,

$$P(\theta_1 \in S_1(\mathbf{X}), \theta_2 \in S_2(\mathbf{X}) = 1 - P[(\theta_1 \in S_1(\mathbf{X}))^c \cup (\theta_2 \in S_2(\mathbf{X}))^c] \, .$$

Como $P(A \cup B) \leq P(A) + P(B)$, se deduce que

$$\begin{aligned} P(\theta_1 \in S_1(\mathbf{X}), \theta_2 \in S_2(\mathbf{X})) &\geq 1 - P(\theta_1 \notin S_1(\mathbf{X})) - P(\theta_2 \notin S_2(\mathbf{X})) \\ &= 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha \, . \end{aligned}$$

Es decir, tomando regiones de confianza para cada parámetro de nivel $1-\alpha/2$ nos aseguramos un nivel simultáneo mayor o igual que $1-\alpha$. Este procedimiento se puede generalizar inmediatamente para el caso que se requieran regiones simultáneas para $k-$parámetros. Bastará tomar para cada parámetro un región de nivel α/k.

Ejemplo 1: Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Hemos visto que un intervalo de confianza para μ de nivel $1-\alpha$ está dado por:

$$S_1 = \left[\overline{X} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \ , \ \overline{X} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right] \ ,$$

mientras que un intervalo de confianza para σ^2 de nivel $1-\alpha$ está dado por:

$$S_2 = \left[\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \ , \ \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}\right] \ .$$

Luego, si tomamos

$$\begin{aligned} S_1^* &= \left[\overline{X} - t_{n-1,\frac{\alpha}{4}} \frac{s}{\sqrt{n}} \ , \ \overline{X} + t_{n-1,\frac{\alpha}{4}} \frac{s}{\sqrt{n}}\right] \ , \\ \text{y} \quad S_2^* &= \left[\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,\frac{\alpha}{4}}} \ , \ \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{n-1,1-\frac{\alpha}{4}}}\right] \end{aligned}$$

$S_1^* \times S_2^*$ es una región de confianza simultánea para (μ, σ^2) de nivel mayor o igual que $1-\alpha$.

El inconveniente que tiene este método es que el nivel es mayor que el deseado, esto ofrece más seguridad que la deseada de que los valores de los parámetros estarán dentro de la región, pero por otra parte las regiones resultan más grandes que lo necesario y por lo tanto, será más imprecisa la determinación de los parámetros.

Obtendremos ahora en el caso normal una región de confianza simultánea para μ y σ^2 de nivel exactamente igual a $1-\alpha$.

Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$. Sabemos que $U = \sqrt{n}(\overline{X} - \mu)/\sigma$ y $V = S^2/\sigma^2$, donde $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2$ son independientes con distribución $N(0,1)$ y χ^2_{n-1} respectivamente. Luego, se tendrá

$$P\left(-z_{\frac{\beta}{2}} \leq \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \leq z_{\frac{\beta}{2}} \ , \ \chi^2_{n-1,1-\frac{\beta}{2}} \leq \frac{S^2}{\sigma^2} \leq \chi^2_{n-1,\frac{\beta}{2}}\right) =$$

$$
\begin{aligned}
&= P\left(-z_{\frac{\beta}{2}} \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq z_{\frac{\beta}{2}}\right) P\left(\chi_{n-1,1-\frac{\beta}{2}} \leq \frac{S^2}{\sigma^2} \leq \chi^2_{n-1,\frac{\beta}{2}}\right) \\
&= (1-\beta)(1-\beta) = (1-\beta)^2
\end{aligned}
$$

Tomemos $\beta = 1-(1-\alpha)^{1/2}$, entonces $(1-\beta)^2 = (1-\alpha)$; luego

$$
S_n = \left\{(\mu,\sigma^2) : -z_{\frac{\beta}{2}} \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq z_{\frac{\beta}{2}} \ , \ \chi^2_{n-1,1-\frac{\beta}{2}} \leq \frac{S^2}{\sigma^2} \leq \chi^2_{n-1,\frac{\beta}{2}}\right\}
$$

es una región de confianza simultánea para (μ,σ^2) de nivel $1-\alpha$. Para estudiar la forma de S_n podemos escribir

$$
S_n = \left\{(\mu,\sigma^2) : \frac{n(\overline{X}-\mu)^2}{z^2_{\frac{\beta}{2}}} \leq \sigma^2 \ , \ \frac{S^2}{\chi^2_{n-1,\frac{\beta}{2}}} \leq \sigma^2 \leq \frac{S^2}{\chi^2_{n-1,1-\frac{\beta}{2}}}\right\}
$$

La condición

$$
\sigma^2 \geq \frac{n(\overline{X}-\mu)^2}{z^2_{\frac{\beta}{2}}}
$$

nos indica la región del plano (μ,σ^2), por encima de la parábola $\sigma^2 = n(\overline{X}-\mu)^2/z^2_{\frac{\beta}{2}}$ y la condición

$$
\frac{S^2}{\chi^2_{n-1,\frac{\beta}{2}}} \leq \sigma^2 \leq \frac{S^2}{\chi^2_{n-1,1-\frac{\beta}{2}}}
$$

indica la franja horizontal comprendida entre las rectas horizontales $\sigma^2 = S^2/\chi^2_{n-1,\frac{\beta}{2}}$ y $S^2/\chi^2_{n-1,1-\frac{\beta}{2}}$.

5.9 Cotas superiores e inferiores de confianza

En los ejemplos vistos anteriormente interesaba conocer el parámetro desconocido con la mayor precisión posible y para este propósito lo más adecuado era construir intervalos de confianza de longitud tan pequeña como era posible. En esta sección, estudiaremos otro tipo de regiones de confianza que surgen naturalmente cuando se está interesado en conocer una cota superior o inferior del parámetro.

Consideremos el siguiente ejemplo. En el Departamento de Control de un laboratorio se recibe un frasco con cierta droga que puede contener alguna impureza indeseada.

Supongamos que se hagan n mediciones de la concentración de la impureza, las que están afectadas de un error, luego se observan $X_1, \ldots, X_n$ donde

$$X_i = \mu + \varepsilon_i \ , \qquad 1 \leq i \leq n$$

donde μ es el valor verdadero de la concentración de la impureza y los ε_i son variables aleatorias $N(0, \sigma^2)$ independientes. Luego $X_1, \ldots, X_n$ es una muestra de una distribución $N(\mu, \sigma^2)$.

En este caso, sólo se estará interesado en determinar si la droga es aceptable o no, y para esto más que un intervalo de confianza interesará tener una cota superior $\overline{\mu}(\mathbf{X})$, $(\mathbf{X} = (X_1, \ldots, X_n))$ tal que la probabilidad de que $\mu \leq \overline{\mu}(X_1, \ldots, X_n)$ sea alta. De esta manera se tendría acotada con probabilidad grande la concentración de impureza de la droga.

Esto sugiere la siguiente definición.

Definición 1: Sea $\mathbf{X}$ un vector cuya distribución pertenece a la familia $F(\mathbf{x}, \theta)$, donde $\theta \in \Theta \subset I\!R$. Se dirá que $\overline{\theta}(\mathbf{X})$ es una *cota superior de confianza con nivel de confianza* $(1 - \alpha)$ *para* θ si $P(\overline{\theta}(\mathbf{X}) \geq \theta) = 1 - \alpha$ o sea si $(-\infty, \overline{\theta}(\mathbf{X})]$ es una región de confianza de nivel $1 - \alpha$. A este tipo de región de confianza semirrecta izquierda se denomina también *intervalo de confianza unilateral izquierdo con nivel* $1 - \alpha$.

Definición 2: Sea $\mathbf{X}$ un vector cuya distribución pertenece a la familia $F(\mathbf{x}, \theta)$ con $\theta \in \Theta \subset I\!R$. Se dirá que $\underline{\theta}(\mathbf{X})$ es una cota inferior de confianza con nivel de confianza $1-\alpha$ si $P(\underline{\theta}(\mathbf{X}) \leq \theta) = 1-\alpha$, o sea si $[\underline{\theta}(\mathbf{X}), \infty)$ es una región de confianza de nivel $1 - \alpha$. A este tipo de región la denominaremos *intervalo de confianza unilateral derecho.*

El siguiente Teorema nos da un procedimiento general para obtener cotas superiores e inferiores de confianza con nivel $1 - \alpha$.

Teorema. *Sea* $\mathbf{X}$ *un vector aleatorio cuya distribución pertenece a la familia* $F(\mathbf{x}, \theta)$ *con* $\theta \in \Theta \subset I\!R$. *Sea* $G(\mathbf{x}, \theta)$ *una función estrictamente monótona en* θ *y tal que* $U = G(\mathbf{X}, \theta)$ *tiene distribución independiente de* θ. *Consideremos* A *y* B *tales que* $P(U \leq A) = \alpha$ *y* $P(U \geq B) = \alpha$.

(a) *Si* $G(\mathbf{x}, \theta)$ *es creciente y continua en* θ, *las cotas superiores e inferior con nivel de confianza* $1 - \alpha$ *vienen dadas respectivamente por las soluciones a las siguientes ecuaciones*

$$G(\mathbf{X}, \overline{\theta}(\mathbf{X})) = B \qquad y \qquad G(\mathbf{X}, \underline{\theta}(\mathbf{X})) = A \ .$$

(b) Si $G(\mathbf{X},\theta)$ es decreciente y continua en cambio $\overline{\theta}(\mathbf{X})$ y $\underline{\theta}(\mathbf{X})$ vienen dadas respectivamente por

$$G(\mathbf{X},\overline{\theta}(\mathbf{X})) = A \qquad y \qquad G(\mathbf{X},\underline{\theta}(\mathbf{X})) = B\ .$$

DEMOSTRACIÓN: La haremos sólo para el caso que $G(\mathbf{x},\theta)$ es creciente en θ y para la cota superior. En este caso $\overline{\theta}(\mathbf{X})$ está definida por

$$G(\mathbf{X},\overline{\theta}(\mathbf{X})) = B\ .$$

Luego,

$$\begin{aligned} P_\theta(\theta \le \overline{\theta}(\mathbf{X})) &= P_\theta(G(\mathbf{X},\theta) \le G(\mathbf{X},\overline{\theta}(\mathbf{X}))) \\ &= P_\theta(G(\mathbf{X},\theta) \le B) = P(U \le B) = 1-\alpha\ . \end{aligned}$$

Ejemplo 1: Supongamos que como en el ejemplo de la droga, donde se quería medir la concentración de impureza, $\mathbf{X} = (X_1,\ldots,X_n)$ es una muestra aleatoria de una distribución $N(\mu,\sigma^2)$ y supongamos que σ^2 sea conocido. Luego,

$$U = G(\mathbf{X},\mu) = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$$

tiene distribución $N(0,1)$. Por lo tanto, en este caso $A = -z_\alpha$ y $B = z_\alpha$.

Luego, como $G(\mathbf{x},\mu)$ es decreciente en μ se tendrá que las cotas superiores e inferiores de confianza de nivel de confianza $1-\alpha$ se obtendrán de la siguiente forma.
Sean $\overline{\mu}(\mathbf{X})$ y $\underline{\mu}(\mathbf{X})$) definidas por

$$\frac{\sqrt{n}(\overline{X}-\overline{\mu}(\mathbf{X}))}{\sigma} = -z_\alpha, \qquad \frac{\sqrt{n}(\overline{X}-\underline{\mu}(\mathbf{X}))}{\sigma} = z_\alpha$$

es decir, despejando se obtiene

$$\overline{\mu}(\mathbf{X}) = \overline{X} + z_\alpha\frac{\sigma}{\sqrt{n}} \qquad \underline{\mu}(\mathbf{X}) = \overline{X} - z_\alpha\frac{\sigma}{\sqrt{n}}$$

Ejemplo 2: Sea $X_1,\ldots,X_n$ una muestra aleatoria de una distribución $N(\mu,\sigma^2)$ y supongamos σ^2 desconocido; luego sabemos que

$$U = G(\mathbf{X},\mu) = \frac{\sqrt{n}(\overline{X}-\mu)}{s}$$

tiene distribución $\mathcal{T}_{n-1}$.

Luego, procediendo como en el Ejemplo 1, obtendremos como cota superior e inferior de confianza con nivel $1-\alpha$

$$\overline{\mu}(\mathbf{X}) = \overline{X} + t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \qquad \text{y} \qquad \underline{\mu}(\mathbf{X}) = \overline{X} - t_{n-1,\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

5.9.1 Comparación de cotas superiores e inferiores de confianza

Así como en el caso de intervalos de confianza interesaba que tuviesen longitud lo más corta posible, cabe preguntarse cómo sería deseable que fuesen las cotas superiores e inferiores. Planteado de otra manera, dadas por ejemplo dos cotas superiores $\overline{\theta}_1(\mathbf{X})$ y $\overline{\theta}_2(\mathbf{X})$, existe algún criterio para compararlas y concluir por ejemplo que una es más conveniente que otra? Análogamente en el caso de cotas inferiores.

Como en el caso de cota superior se tiene controlada la posibilidad que $\overline{\theta}(\mathbf{X})$ esté por debajo de θ, ya que esto sólo puede suceder con probabilidad α, el riesgo no controlado es que $\overline{\theta}(\mathbf{X})$ sobrevalúe θ muy por encima de lo necesario. Esta sobrevaluación que la llamaremos $C(\mathbf{X},\theta)$ estará dada por

$$C(\mathbf{X},\theta) = \begin{cases} \overline{\theta}(\mathbf{X}) - \theta & \text{si } \overline{\theta}(\mathbf{X}) > \theta \\ 0 & \text{si } \overline{\theta}(\mathbf{X}) \leq \theta \end{cases}$$

Luego parece razonable buscar cotas superiores que minimicen $E_\theta(C(\mathbf{X},\theta))$ uniformemente en θ.

Del mismo modo en el caso de cotas inferiores, se puede definir la subvaluación por

$$D(\mathbf{X},\theta) = \begin{cases} \theta - \underline{\theta}(\mathbf{X}) & \text{si } \theta > \underline{\theta}(\mathbf{X}) \\ 0 & \text{si } \theta \leq \underline{\theta}(\mathbf{X}) \end{cases}$$

y en este caso interesará minimizar $E_\theta(D(\mathbf{X},\theta))$ uniformemente en θ.

La teoría de la optimalidad de las cotas de confianza se deriva de la teoría de optimalidad de los tests y por lo tanto se pospone hasta el Capítulo 6.

Solamente diremos que contrariamente a lo que sucedía con intervalos de confianza, existen en casos no triviales cotas uniformemente óptimas. Por ejemplo, los procedimientos derivados en el Ejemplo 1 tienen esta propiedad. En el caso del Ejemplo 2, no existen procedimientos uniformemente óptimos.

De todos modos los procedimientos derivados en ese ejemplo son uniformemente óptimos si se restringe al conjunto de procedimientos insesgados. (Una cota es insesgada si su intervalo de confianza unilateral asociado es una región de confianza insesgada.)

REFERENCIAS

1. Lehmann, E.L. (1994) *Testing Statistical Hypothesis*. Chapman and Hall.

2. Pratt, E. (1961) Length of Confidence Intervals, *J. Amer. Statist. Assoc.* 16: 243–258.

3. Tate, R.F. y Klett, G.W. (1959) Optimal Confidence Intervals for the variance of a Normal Distribution, *J. Amer. Statist. Assoc.* 54: 674–682.

4. Wilks, S.S. (1962) *Mathematical Statistics*, J. Wiley and Sons.

Chapter 6

Tests de Hipótesis

6.1 Introducción

El test de hipótesis es una manera formal de decidir entre dos opciones, o sea, es una manera de distinguir entre distribuciones de probabilidad en base a variables aleatorias generadas por una de ellas. Veamos un ejemplo para tener una idea de lo que significan.

Ejemplo 1. Supongamos que un comerciante debe comprar un cargamento de N manzanas. El comerciante ignora qué parte del cargamento no se encuentra en buen estado. Como inspeccionar todo el cargamento es muy costoso, decide elegir al azar una muestra de n manzanas.

Sea X el número de manzanas en mal estado que observa en la muestra. Luego si D es el número de manzanas en mal estado que hay en el cargamento, se tiene que la distribución de X es hipergeométrica y su función de probabilidad puntual está dada por

$$p(x, D) = \frac{\begin{pmatrix} D \\ x \end{pmatrix}\begin{pmatrix} N-D \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \quad \text{si } \max(0, D-N+n) \leq x \leq \min(n, D)$$

y D puede tomar valores en el conjunto $\Theta = \{0, 1, 2, \ldots, N\}$.

Supongamos que se hubiese convenido que el cargamento debería tener no más de D_0 manzanas en mal estado. Luego, en base a la variable X, que el comerciante observa, debe decidir si el cargamento satisface los requisitos

convenidos. Es decir, debe decidir entre dos alternativas

$$D \in \Theta_1 = \{0, 1, \ldots, D_0\} \qquad \text{o} \qquad D \in \Theta_2 = \{D_0 + 1, \ldots, N\}$$

Esto puede ser expresado como que el comerciante debe decidir entre dos hipótesis:

$$\text{H} : D \in \Theta_1 \qquad \text{contra} \qquad \text{K} : D \in \Theta_2$$

y esta decisión debe hacerla a partir del valor observado X.

Un test será una regla de decisión basada en X. Esto puede ser expresado matemáticamente como una función $\varphi(X)$ que toma dos valores: 1 y 0. 1 significará que rechaza H y por lo tanto acepta K y 0 que acepta H.

Supongamos por ejemplo que $N = 1000$, $n = 100$ y $D_0 = 150$. Un posible test está dado por:

$$\varphi_1(X) = \begin{cases} 1 & \text{si } X > 15 \\ 0 & \text{si } X \leq 15 \, . \end{cases}$$

De acuerdo con este test se rechaza el cargamento, es decir, se decide que $D \in \Theta_2$ si se observa en la muestra más de 15 manzanas en mal estado. Si se quisiera usar un test más seguro para el comprador (en el sentido de que la probabilidad de aceptar un cargamento con más de 150 manzanas en mal estado sea menor) se podría usar, por ejemplo,

$$\varphi_2(X) = \begin{cases} 1 & \text{si } X > 5 \\ 0 & \text{si } X \leq 5 \, . \end{cases}$$

Por ahora, no tenemos ningún criterio para elegir entre dos tests, ni entre los muchos otros que podrían definirse. En los párrafos siguientes atacaremos el problema de definir criterios para comparar diferentes tests, y el de elegir un test óptimo.

Ejemplo 2. Supongamos que para curar una determinada enfermedad se emplea una droga que cura la enfermedad con una probabilidad θ_0 conocida. Se ha obtenido una nueva droga y se quiere determinar si vale la pena cambiar la droga. Para ello se prueba la nueva droga con n pacientes obteniéndose los resultados $X_1, \ldots, X_n$, donde $X_i = 1$ indica que el $i-$ésimo paciente se curó y $X_i = 0$, que no se curó. Sea θ la probabilidad de curar de la nueva droga, la cual no es conocida.

Se está dispuesto a cambiar de droga si la nueva droga es tal que $\theta \geq \theta_0 + 0.05$, es decir si esta última cura al menos un 5% más de pacientes que la vieja. Luego, se tiene que decidir entre dos hipótesis:

$$\mathrm{H} : \theta \leq \theta_0 + 0.05 \qquad \text{y} \qquad \mathrm{K} : \theta > \theta_0 + 0.05$$

Un test será una función $\varphi(X_1, \ldots, X_n)$ que toma valores 1 ó 0. $\varphi(X_1, \ldots, X_n) = 0$ indicará que aceptamos H, es decir, no se continúa usando la droga vieja.

Para ejemplificar, supongamos que $\theta_0 = 0.8$ y $n = 100$. Un posible test sería

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \sum_{i=1}^{100} X_i \geq 85 \\ \\ 0 & \text{si } \sum_{i=1}^{100} X_i < 85 \, . \end{cases}$$

Este test acepta K, es decir, cambia de droga si 85 pacientes o más resultan curados.

Si se quisiera ser más conservador, es decir, estar más seguro que la droga tiene la probabilidad de curar mayor que 0.85 antes de tomar la decisión de cambiarla, se podría usar el test

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \sum_{i=1}^{100} X_i \geq 90 \\ \\ 0 & \text{si } \sum_{i=1}^{100} X_i < 90 \, . \end{cases}$$

6.2 Formulación general del problema del test de hipótesis

Supongamos que se obtiene un vector aleatorio $\mathbf{X} = (X_1, ..., X_n)$ cuya función de distribución pertenece a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta \subset I\!R^p$. Sean Θ_1 y Θ_2 tales que $\Theta_1 \cap \Theta_2 = \emptyset$ y $\Theta_1 \cup \Theta_2 = \Theta$. Un test para este problema será una regla basada en $\mathbf{X}$ para decidir entre las dos hipótesis

$$\mathrm{H} : \theta \in \Theta_1 \qquad \text{contra} \qquad \mathrm{K} : \theta \in \Theta_2$$

Definición 1. Se llama *test* a una función $\varphi : I\!R^n \to [0, 1]$.

Se dice que un test φ es *no aleatorizado* si toma solamente los valores 0 ó 1.

Cuando $\varphi(\mathbf{X}) = 1$ se rechazará la hipótesis H y por lo tanto, se aceptará K. En cambio, $\varphi(\mathbf{X}) = 0$ indicará que se acepta H.

Si el test toma valores distintos de 0 y 1 se dice que es un *test aleatorizado*. En este caso, el valor $\varphi(\mathbf{x})$ indica con que probabilidad se rechaza H si se observa $\mathbf{X} = \mathbf{x}$, es decir, $\varphi(\mathbf{x}) = P(\text{rechazar H}|\mathbf{X} = \mathbf{x})$

Por ejemplo, $\varphi(\mathbf{X}) = 1/2$ indicará que si observamos el vector $\mathbf{X}$ debemos rechazar H con probabilidad $1/2$, es decir, podríamos tirar una moneda y si saliera ceca aceptarla, $\varphi(\mathbf{X}) = 1/6$ indicará que si observamos $\mathbf{X}$ debemos rechazar H con probabilidad $1/6$; en este caso podríamos tirar un dado; si saliese 1 rechazaríamos H y en los demás casos la aceptaríamos.

La aleatorización introduce en la decisión un elemento extraño al fenómeno estudiado, como el lanzamiento de una moneda o un dado, con que hemos ejemplificado. Por lo tanto, se evitan en lo posible los tests aleatorizados en los casos prácticos. Sin embargo, desde el punto de vista teórico, conviene como se verá, admitir la posibilidad de tests aleatorizados.

En la mayoría de las situaciones, los tests vienen dados como funciones de un estadístico, llamado *estadístico del test*, que, por ejemplo, como en el caso de la sección anterior, sirven para rechazar H para valores grandes. En general, el estadístico del test sirve para medir la diferencia entre los datos y lo que se espera de ellos bajo H.

Definición 2. La *región crítica* $\mathcal{R}$, de un test φ, es el conjunto de puntos $\mathbf{X}$ que llevan a la decisión de rechazar H y la *región de aceptación* $\mathcal{A}$ es el conjunto de puntos $\mathbf{X}$ que llevan a aceptar H.

Dado un test para un problema de test de hipótesis se podrá incurrir en dos tipos de error.

Definición 3. Se llamará *error de tipo 1* al que se comete al rechazar la hipótesis H, cuando es verdadera. Se llamará *error de tipo 2* al que se comete al aceptar H, cuando esta hipótesis es falsa.

Luego, para un test no aleatorizado, la probabilidad de cometer un error de tipo 1 será $P_{\boldsymbol{\theta}}(\mathcal{R})$, cuando $\boldsymbol{\theta} \in \Theta_1$. Mientras que la probabilidad de error de tipo 2, será $P_{\boldsymbol{\theta}}(\mathcal{A}) = 1 - P_{\boldsymbol{\theta}}(\mathcal{R})$, cuando $\boldsymbol{\theta} \in \Theta_2$.

Ejemplo 1 (donde se visualiza la necesidad de introducir tests aleatorizados). Supongamos que una empresa automotriz sostiene que domina la mitad del mercado, esto es que la mitad de los compradores de automóviles

se deciden por alguno de los modelos fabricados por ella. Se desea testear si la afirmación hecha por la empresa es exagerada o no.

Supongamos que se toma una muestra de compradores que, para facilidad en los cálculos, consideraremos de tamaño $n = 6$.

Las hipótesis en cuestión son:

$$\mathrm{H} : \theta = 1/2 \qquad \text{contra} \qquad \mathrm{K} : \theta < 1/2$$

donde θ es la probabilidad de que un comprador tomado al azar compre un automóvil de la empresa.

Consideremos para cada comprador i, la variable X_i tal que $X_i = 1$ si el comprador se decide por un auto fabricado por la empresa; $X_i = 0$ en caso contrario. Luego, cada X_i tendrá distribución $Bi(\theta, 1)$.

Supongamos también que se quiere tener una probabilidad de error de tipo 1 de 0.25, es decir que la probabilidad de rechazar H cuando es verdadera es del 25%.

Parecería intuitivo considerar un test de la forma

$$\varphi_k(\mathbf{X}) = \begin{cases} 1 & \text{si } X < k \\ 0 & \text{si } X \geq k \end{cases}$$

Consideremos los test φ_2 y φ_3. Veamos que ninguno de ellos satisface la exigencia planteada para el error de tipo 1.

Suponiendo que las decisiones de los compradores son independientes entre sí, $T = \sum_{i=1}^{6} X_i$, tiene distribución $Bi(\theta, 6)$.

Calculemos la probabilidad de error de tipo 1 para ambos tests. Para ello usaremos la tabla de la distribución $Bi(6, 1/2)$.

t	0	1	2	3	4	5	6
$P_{\frac{1}{2}}(T = t)$	1/64	6/64	15/64	20/64	15/64	6/64	1/64

Por lo tanto,

$$P_{\frac{1}{2}}(\varphi_2 = 1) = P_{\frac{1}{2}}(T < 2) = 7/64 < 0.25$$

y

$$P_{\frac{1}{2}}(\varphi_3 = 1) = P_{\frac{1}{2}}(T < 3) = 22/64 > 0.25$$

Resulta claro entonces que no podremos elegir un test en la familia de tests no aleatorizados φ_k con un error de tipo 1 igual a 0.25.

Tendría sentido, en esta situación, plantearse un test de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } T < 2 \\ \gamma & \text{si } T = 2 \\ 0 & \text{si } T > 2 \end{cases}$$

y tratar de elegir γ de forma tal que tenga el error de tipo I deseado. Para eso se requiere

$$P_{\frac{1}{2}}(\varphi(\mathbf{X}) = 1) = P_{\frac{1}{2}}(T < 2) + \gamma\, P_{\frac{1}{2}}(T = 2) = 0.25\ .$$

Luego, se deberá cumplir

$$\frac{7}{64} + \gamma\,\frac{15}{64} = 0.25,$$

o sea $\gamma = 3/5$.

Una forma de efectivizar el test, en este caso, podría ser la siguiente. Cuando se observa que $T < 2$, se rechaza H; cuando se observa que $T > 2$, se acepta H; cuando se observa $T = 2$ se colocan en una urna tres bolillas rojas y dos bolillas negras y se extrae una al azar; si resulta roja se rechaza H y si no se acepta.

Notemos que si en lugar de pedir que la probabilidad de error de tipo 1 sea 0.25 hubiésemos pedido que fuera 0.10; el test hubiera resultado de la forma

$$\varphi^*(\mathbf{X}) = \begin{cases} 1 & \text{si } T < 1 \\ 0.9 & \text{si } T = 1 \\ 0 & \text{si } T > 1 \end{cases}$$

O sea, cuanto más exigentes somos respecto del error de tipo 1, más estricta es la cota dada para el estadístico del test.

Debemos destacar que en este ejemplo, y en los anteriores, el test se basa en un estadístico cuya distribución es conocida cuando H es cierta. Conocer esa distribución hace posible definir la región de rechazo que tendrá probabilidad α prefijada bajo H. El valor elegido como cota o punto de corte, para tomar la decisión, se llama *valor crítico* y por lo tanto, separa la región de aceptación de la región de rechazo.

Volvamos al problema general de test de hipótesis planteado al comienzo de esta sección. Sea H : $\boldsymbol{\theta} \in \Theta_1$ y K : $\boldsymbol{\theta} \in \Theta_2$; sea $\varphi(\mathbf{X})$ un test para estas dos hipótesis. Entonces

Definición 4. Se llama *función de potencia del test* $\varphi(\mathbf{X})$ a la función

$$\beta_\varphi(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\text{rechazar H}),$$

donde $P_{\boldsymbol{\theta}}$ indica la probabilidad cuando $\boldsymbol{\theta}$ es el valor verdadero.

En el caso que φ es un test no aleatorizado se tiene

$$\beta_\varphi(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\varphi(\mathbf{X}) = 1) = E_{\boldsymbol{\theta}}(\varphi(\mathbf{X})) \ .$$

Si φ es aleatorizado, $\varphi(\mathbf{X})$ puede interpretarse como la probabilidad de rechazar H, condicional a observar $\mathbf{X}$; luego se tiene

$$\varphi(\mathbf{X}) = P(\text{rechazar H}|\mathbf{X})$$

y resulta

$$\beta_\varphi(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(\text{ rechazar H}) = E_{\boldsymbol{\theta}}(P(\text{ rechazar H}|\mathbf{X})) = E_{\boldsymbol{\theta}}(\varphi(\mathbf{X})) \ .$$

Por lo tanto, en todos los casos se tiene

$$\beta_\varphi(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}(\varphi(\mathbf{X})) \ .$$

Expresemos ahora las probabilidades de los errores de un test en términos de $\beta_\varphi(\boldsymbol{\theta})$

- La probabilidad que ocurra un error de tipo 1 será $\beta_\varphi(\boldsymbol{\theta})$ para $\boldsymbol{\theta} \in \Theta_1$.
- La probabilidad que ocurra un error de tipo 2 será $(1 - \beta_\varphi(\boldsymbol{\theta}))$ para $\boldsymbol{\theta} \in \Theta_2$.

Un buen test deberá tener errores de tipo 1 y 2 pequeños, y por lo tanto debe tener una función de potencia $\beta_\varphi(\boldsymbol{\theta})$ que tome valores cercanos a 0 para $\boldsymbol{\theta} \in \Theta_1$ y valores cercanos a 1 para $\boldsymbol{\theta} \in \Theta_2$.

En realidad, no podemos hacer ambos errores pequeños simultáneamente. Más aún, para un tamaño de muestra dado para que decrezca la probabilidad de que ocurra un error de tipo 1, debemos aumentar la probabilidad de que ocurra un error de tipo 2 (o sea disminuir la potencia). Si queremos que ambos sean pequeños debemos aumentar la cantidad de observaciones.

Por ejemplo, en el Ejemplo 1, el test φ^* cumplía $\beta_{\varphi^*}(1/2) = 0.10$. Por otra parte, se verifica que $\beta_{\varphi^*}(\theta) = (1-\theta)^6 + 5.4\,\theta(1-\theta)^5$, con lo cual

tenemos la tabla siguiente que da la función de potencia del test φ^* para algunos valores de $\theta \in [0, 1/2]$

θ	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
$\beta_{\varphi^*}(\theta)$	1	0.944	0.85	0.736	0.616	0.498	0.389	0.295	0.215	0.149	0.1

Como vemos, la función de potencia de φ^* es una función decreciente de θ en el intervalo $[0, 1/2]$ que tiende a 1 cuando $\theta \to 0$ y tiende a 0.1 cuando $\theta \to 1/2$. Es decir, la probabilidad de error 2 tiende a 0 cuando $\theta \to 0$ y por lo tanto, se logran detectar bien alternativas lejanas a la hipótesis H.
Para los procedimientos que daremos $1-P(\text{error de tipo 1}) \geq P(\text{error de tipo 2})$. El objetivo será encontrar procedimientos con la menor probabilidad de tipo 2, fijada la probabilidad de tipo 1, es decir, buscaremos procedimientos con potencia grande para $\boldsymbol{\theta} \in \Theta_2$.

6.2.1 Nivel de significación de un test

La teoría clásica de test de hipótesis considera que el error de tipo 1 es mucho más grave que el error de tipo 2. Es decir, la situación de las hipótesis H y K no es simétrica; es mucho más grave rechazar H cuando es cierta que aceptarla cuando es falsa. Esto significa que se debe tener mucha evidencia sobre que H es falsa antes de rechazarla. Se observa en el Ejemplo 2 de la sección 1, que esta simetría corresponde a una situación real, puesto que antes de cambiar de droga, es decir rechazar H, habría que tener un grado de certeza muy alto respecto de que la nueva droga es mejor que la primera. Desde ahora en adelante H se denominará *hipótesis nula* y K *hipótesis alternativa*.

Veamos un ejemplo que servirá para fijar ideas y clarificar la mecánica de elección de H

Ejemplo 1. Supongamos que se quiere decidir si un paciente tiene o no tuberculosis, para proceder, en caso afirmativo, a suministrarle un tratamiento adecuado. Tendremos entonces dos hipótesis:

(A) El señor W está tuberculoso;

(B) El señor W no está tuberculoso.

Es claro que el médico responsable de la decisión considerará mucho más grave rechazar (A) cuando es cierta, que rechazar (B) cuando es cierta (esto es lo mismo que aceptar H cuando es falsa), puesto que en el primer caso

se expone al paciente a una agudización grave de su enfermedad, mientras que en el segundo se le aplicará un tratamiento que no necesita y cuyas consecuencias nunca serán comparables con el daño de no tratarlo estando enfermo.

Luego la hipótesis nula será H : "El señor W está tuberculoso"; y la alternativa K : "El señor W no está tuberculoso".

Como dijimos más arriba, supondremos que el error de tipo 1 (rechazar H cuando es cierta), es el más grave. Por lo tanto se va requerir que el error de tipo 1 del test a utilizar no sea mayor que un número $0 < \alpha < 0.5$ prefijado. Este número α es generalmente pequeño (entre 0.01 y 0.10) y se lo determina de acuerdo a la importancia del error de tipo 1. La siguiente definición formaliza este concepto.

Definición 5. El *nivel de significación* de un test φ está definido por

$$\alpha = \sup_{\boldsymbol{\theta} \in \Theta_1} \beta_{\boldsymbol{\theta}}(\varphi)$$

Luego, α es el supremo de la probabilidad de cometer un error de tipo 1.

Por lo tanto, fijado α, se buscará un test que tenga nivel de significación menor o igual que α. Un test con está propiedad asegurará que la probabilidad de rechazar la hipótesis nula H, cuando esta es cierta, no sea mayor que α.

Como existen muchos tests que tienen nivel de significación menor o igual que α para un problema determinado, debemos dar un criterio para elegir uno entre todos ellos. Resulta natural elegir entre todos los tests con la restricción de que su nivel de significación sea menor o igual que α aquel que tenga menor probabilidad de error de tipo 2. Esto motiva la siguiente definición.

Definición 6. Consideremos un problema general de test de hipótesis donde se observa un vector $\mathbf{X}$ con distribución $F(x, \boldsymbol{\theta})$,con $\boldsymbol{\theta} \in \Theta$, y se tiene que decidir entre las hipótesis H: $\boldsymbol{\theta} \in \Theta_1$ y K: $\boldsymbol{\theta} \in \Theta_2$. Diremos que un test φ es el test *más potente* de nivel menor o igual que α para una alternativa fija $\boldsymbol{\theta}_2 \in \Theta_2$ si

(a) $\sup_{\boldsymbol{\theta} \in \Theta_1} \beta_\varphi(\boldsymbol{\theta}) \leq \alpha$, es decir si φ tiene nivel de significación menor o igual que α

(b) Dado otro test φ^* de nivel menor o igual que α entonces se tiene

$$\beta_{\varphi^*}(\boldsymbol{\theta}_2) \leq \beta_\varphi(\boldsymbol{\theta}_2)$$

Es decir, la probabilidad de error cuando $\boldsymbol{\theta}_2$ es el verdadero valor es menor para el test φ que para cualquier otro φ^* de nivel menor o igual que α (o sea, $(1-\beta_{\varphi}(\boldsymbol{\theta}_2)) \leq (1-\beta_{\varphi^*}(\boldsymbol{\theta}_2))$).

Es claro que si cambiamos la alternativa $\boldsymbol{\theta}_2 \in \Theta_2$ por otro $\boldsymbol{\theta}_2' \in \Theta_2$, el test más potente para esta $\boldsymbol{\theta}_2'$ no tiene porque coincidir con el correspondiente a $\boldsymbol{\theta}_2$. Por ejemplo, si se quiere testear H : $\mu = \mu_0$ contra K : $\mu \neq \mu_0$, para una distribución $N(\mu, \sigma_0^2)$ con σ_0^2 conocida, resultará

$$\Theta_1 = \{\mu_0\} \quad ; \quad \Theta_2 = \{\mu \in I\!R : \mu \neq \mu_0\}.$$

Si se toma una alternativa fija $\mu_1 < \mu_0$, el test más potente de nivel α para esta alternativa no coincide con el test mas potente para una alternativa $\mu_2 > \mu_0$, como veremos más adelante.

Definición 7. Diremos que un φ es un test *uniformemente más potente*, UMP, de nivel menor o igual que α para H : $\boldsymbol{\theta} \in \Theta_1$ contra K : $\boldsymbol{\theta} \in \Theta_2$, si φ es el más potente de nivel menor o igual que α para todo $\boldsymbol{\theta}_2 \in \Theta_2$, es decir, si el mismo test es óptimo *cualquiera* sea la alternativa fija $\boldsymbol{\theta}_2 \in \Theta_2$ considerada.

Lo ideal sería encontrar (cuando existan) tests uniformemente más potentes de nivel menor o igual que α. Estudiaremos casos donde estos tests existen y otros donde no. En estos últimos habrá que elegir otros criterios para seleccionar el test a usar.

Definición 8. El *nivel crítico o p-valor* es el menor valor de significación para el que rechazamos la hipótesis H para una observación dada $\mathbf{x}$.

En el Ejemplo 1 de la sección 2, por ejemplo si observamos $X = 2$ el p-valor del test $\{\varphi_k\}$ que rechaza para valores pequeños de T, será $p = 7/64$.

Prefijado el nivel de significación α, y evaluado el p- valor, p, del test utilizado, rechazaremos H si $p < \alpha$.

A esta altura, la lógica de los tests puede parecer más clara. Es un argumento por contradicción destinado a mostrar que la hipótesis nula lleva a conclusiones absurdas y que por lo tanto, debe ser rechazada.

Supongamos que para un conjunto de datos dado, se evalúa el estadístico del test y se obtiene un p–valor de 0.001. Para interpretarlo, *debemos pensar que la hipótesis nula es cierta* e imaginamos a otros investigadores repitiendo la experiencia en idénticas condiciones. El valor 0.001 dice que sólo un investigador de cada 1000 puede obtener un valor del estadístico tan extremo como el obtenido. Por lo tanto, la diferencia entre los datos y lo que se espera de ellos bajo H no puede atribuirse meramente a variación aleatoria. Este

hecho lleva a una contradicción y por lo tanto, a abandonar nuestra hipótesis de que H era cierta.

Es tentador pensar que el p–valor da la probabilidad de que H sea cierta, pero no es así. No importa cuántas veces se repita el experimento, H será siempre cierta o siempre falsa. Es decir, el nivel crítico da la probabilidad de obtener evidencia en contra de la hipótesis nula suponiendo que ésta sea cierta. Por lo tanto, cuanto menor sea el p-valor más evidencia en contra de H tenemos, suponiendo que H es cierta.

6.3 Tests óptimos para el caso de hipótesis simple contra hipótesis simple

El caso más simple de problema de test de hipótesis es la situación donde Θ_1 y Θ_2 contengan cada uno un elemento. En este caso, se dice, H y K son *hipótesis simples*.

Si Θ_1 tuviera más de un elemento, H se llamará *hipótesis compuesta*, y lo mismo vale para K en relación a Θ_2.

En el caso en que H y K sean simples, un problema de test de hipótesis será de la forma

$$\mathrm{H} : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \qquad \text{contra} \qquad \mathrm{K} : \boldsymbol{\theta} = \boldsymbol{\theta}_2$$

Supongamos que $\mathbf{X}$ sea un vector discreto (o continuo) bajo $\boldsymbol{\theta}_1$ y $\boldsymbol{\theta}_2$ y que las funciones de densidad correspondientes sean $p(\mathbf{x}, \boldsymbol{\theta}_1)$ y $p(\mathbf{x}, \boldsymbol{\theta}_2)$. Luego, intuitivamente, parece razonable rechazar H si la "probabilidad" de obtener el valor observado $\mathbf{x}$ bajo $\boldsymbol{\theta}_2$ es grande comparada con la "probabilidad" de obtener $\mathbf{x}$ bajo $\boldsymbol{\theta}_1$, es decir, cuando

$$L_{21} = \frac{p(\mathbf{x}, \boldsymbol{\theta}_2)}{p(\mathbf{x}, \boldsymbol{\theta}_1)} \geq k_\alpha$$

donde k_α es una constante que depende del nivel α. Por lo tanto, se podría pensar en construir un test de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } L_{21} > k_\alpha \\ \gamma_\alpha & \text{si } L_{21} = k_\alpha \\ 0 & \text{si } L_{21} < k_\alpha \end{cases}$$

o equivalentemente,

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } p(\mathbf{x}, \boldsymbol{\theta}_2) > k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1) \\ \gamma_\alpha & \text{si } p(\mathbf{x}, \boldsymbol{\theta}_2) = k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1) \\ 0 & \text{si } p(\mathbf{x}, \boldsymbol{\theta}_2) < k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1) \end{cases} \tag{6.1}$$

donde $0 \leq \gamma_\alpha \leq 1$, correspondiendo el caso $k_\alpha = +\infty$ al test

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } p(\mathbf{x}, \boldsymbol{\theta}_1) = 0 \\ 0 & \text{si } p(\mathbf{x}, \boldsymbol{\theta}_1) > 0 \end{cases} \tag{6.2}$$

que tiene nivel 0.

Si queremos que el test (6.1) tenga nivel α debemos elegir k_α y β_α tales que se cumpla

$$E_{\boldsymbol{\theta}_1}(\varphi(\mathbf{X})) = \alpha \, . \tag{6.3}$$

Notemos que entonces, en este tipo de test k_α es una función decreciente de α.

Un test de la forma (6.1) se llama *test del cociente de verosimilitud*. El siguiente teorema establece que se pueden elegir k_α y γ_α de manera que se cumpla (6.3) y que usando estos valores en (6.1) se obtiene un test más potente de nivel menor o igual que α. Sin embargo, los tests de la forma (6.1) no garantizan la unicidad y es por ello, que para obtenerla le permitiremos a γ_α depender de $\mathbf{x}$.

Teorema 1 (de Neyman–Pearson)

(i) *Dado $0 \leq \alpha \leq 1$ se pueden elegir k_α y γ_α, $0 \leq \gamma_\alpha \leq 1$, tales que el test de la forma (6.1) satisfaga (6.3).*

(ii) *Sea un test de la forma (6.1) que satisface (6.3) para $\alpha > 0$ y de la forma (6.2) para $\alpha = 0$. Luego ese test es el más potente de nivel menor o igual que α para*

$$H : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \qquad \textit{contra} \qquad K : \boldsymbol{\theta} = \boldsymbol{\theta}_2 \, .$$

(iii) Si φ^* es un test uniformemente más potente de nivel $\alpha > 0$ para $H : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ versus $K : \boldsymbol{\theta} = \boldsymbol{\theta}_2$ entonces φ^* es de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \textit{si } p(\mathbf{x}, \boldsymbol{\theta}_2) > k_\alpha \, p(\mathbf{x}, \boldsymbol{\theta}_1) \\ \gamma_\alpha(\mathbf{x}) & \textit{si } p(\mathbf{x}, \boldsymbol{\theta}_2) = k_\alpha \, p(\mathbf{x}, \boldsymbol{\theta}_1) \\ 0 & \textit{si } p(\mathbf{x}, \boldsymbol{\theta}_2) < k_\alpha \, p(\mathbf{x}, \boldsymbol{\theta}_1) \end{cases} \tag{6.4}$$

excepto quizás en un conjunto $\mathcal{N}$ tal que $P_{\boldsymbol{\theta}_1}(\mathcal{N}) = P_{\boldsymbol{\theta}_2}(\mathcal{N}) = 0$.

Si φ^* es un test uniformemente más potente de nivel 0 para $H : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ versus $K : \boldsymbol{\theta} = \boldsymbol{\theta}_2$ entonces φ^* es de la forma (6.2) excepto quizás en un conjunto $\mathcal{N}$ tal que $P_{\boldsymbol{\theta}_1}(\mathcal{N}) = P_{\boldsymbol{\theta}_2}(\mathcal{N}) = 0$.

DEMOSTRACIÓN: (i) Si $\alpha = 0$ el test (6.2) tiene nivel 0. Sea entonces, $0 < \alpha \leq 1$.

Extendamos la definición de la variable aleatoria L_{21} al caso en que el denominador es 0,

$$L_{21} = \begin{cases} \frac{p(\mathbf{x},\boldsymbol{\theta}_2)}{p(\mathbf{x},\boldsymbol{\theta}_1)} & \text{si } p(\mathbf{x},\boldsymbol{\theta}_1) > 0 \\ 1 & \text{si } p(\mathbf{x},\boldsymbol{\theta}_1) = 0 \end{cases} .$$

Luego,

$$\begin{aligned} E_{\boldsymbol{\theta}_1}(\varphi(\mathbf{X})) &= P_{\boldsymbol{\theta}_1}(L_{21} > k_\alpha) + \gamma\, P_{\boldsymbol{\theta}_1}(L_{21} = k_\alpha) \\ &= 1 - P_{\boldsymbol{\theta}_1}(L_{21} \leq k_\alpha) + \gamma_\alpha\, P_{\boldsymbol{\theta}_1}(L_{21} = k_\alpha) . \end{aligned}$$

Si existe una constante k_0 tal que $P_{\boldsymbol{\theta}_1}(L_{21} \leq k_0) = \alpha$ tomamos $k_\alpha = k_0$ y $\gamma_\alpha = 0$. En caso contrario, siempre existe k_0 tal que

$$P_{\boldsymbol{\theta}_1}(L_{21} < k_0) \leq 1 - \alpha < P_{\boldsymbol{\theta}_1}(L_{21} \leq k_0) \tag{6.5}$$

y se cumple, $P_{\boldsymbol{\theta}_1}(L_{21} = k_0) > 0$. Definamos $k_\alpha = k_0$ y

$$\gamma_\alpha = \frac{P_{\boldsymbol{\theta}_1}(L_{21} \leq k_0) - (1-\alpha)}{P_{\boldsymbol{\theta}_1}(L_{21} = k_0)} .$$

Luego, por (6.5) $0 < \gamma_\alpha \leq 1$ y además $E_{\boldsymbol{\theta}_1}(\varphi(\mathbf{X})) = \alpha$.

Demostraremos (ii) en el caso continuo, el caso discreto es análogo reemplazando las integrales por sumatorias. Supongamos que φ sea de la forma (6.1) y satisfaga (6.3). Luego, por satisfacer (6.3) su nivel es igual a α.

Para mostrar que φ es el test más potente de nivel menor o igual que α, sólo falta mostrar que dado otro test φ^* de nivel menor o igual que α se tiene

$$\beta_\varphi(\boldsymbol{\theta}_2) \geq \beta_{\varphi^*}(\boldsymbol{\theta}_2) \tag{6.6}$$

(a) Supongamos primero $\alpha > 0$ con lo cual $k_\alpha < \infty$ en (6.1). Sea φ^* de nivel menor o igual que α. Consideremos la expresión

$$U(\mathbf{x}) = [\varphi(\mathbf{x}) - \varphi^*(\mathbf{x})]\,[p(\mathbf{x},\boldsymbol{\theta}_2) - k_\alpha\, p(\mathbf{x},\boldsymbol{\theta}_1)] . \tag{6.7}$$

Mostraremos que $U(\mathbf{x}) \geq 0$.

Supongamos primero que

$$p(\mathbf{x},\boldsymbol{\theta}_2) > k_\alpha\, p(\mathbf{x},\boldsymbol{\theta}_1) .$$

Luego, de acuerdo con (6.1), se tendrá $\varphi(\mathbf{x}) = 1$ y por lo tanto $\varphi(\mathbf{x}) \geq \varphi^*(\mathbf{x})$, de donde, $U(\mathbf{x}) \geq 0$.

Si $p(\mathbf{x}, \boldsymbol{\theta}_2) = k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1)$, es claro que $U(\mathbf{x}) = 0$.

Finalmente, si

$$p(\mathbf{x}, \boldsymbol{\theta}_2) < k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1)\ ,$$

entonces usando nuevamente (6.1), $\varphi(\mathbf{x}) = 0$, con lo cual $\varphi(\mathbf{x}) \leq \varphi^*(\mathbf{x})$ y por lo tanto $U(\mathbf{x}) \geq 0$.

Resulta entonces que

$$\int [\varphi(\mathbf{x}) - \varphi^*(\mathbf{x})]\, [p(\mathbf{x}, \boldsymbol{\theta}_2) - k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1)]\, d\mathbf{x} = \int U(\mathbf{x}) d\mathbf{x} \geq 0\ .$$

Por lo tanto,

$$\int (\varphi(\mathbf{x}) - \varphi^*(\mathbf{x})) p(\mathbf{x}, \boldsymbol{\theta}_2) d\mathbf{x} \geq k_\alpha \int (\varphi(\mathbf{x}) - \varphi^*(\mathbf{x})) p(\mathbf{x}, \boldsymbol{\theta}_1) d\mathbf{x}$$

o equivalentemente,

$$\beta_\varphi(\boldsymbol{\theta}_2) - \beta_{\varphi^*}(\boldsymbol{\theta}_2) \geq k_\alpha(\beta_\varphi(\boldsymbol{\theta}_1) - \beta_{\varphi^*}(\boldsymbol{\theta}_1))\ .$$

Por (6.3) se tiene $\beta_\varphi(\boldsymbol{\theta}_1) = \alpha$, como φ^* es un test de nivel de significación menor o igual que α, $\beta_{\varphi^*}(\boldsymbol{\theta}_1) \leq \alpha$, y entonces resulta

$$\beta_\varphi(\boldsymbol{\theta}_1) - \beta_{\varphi^*}(\boldsymbol{\theta}_1) \geq 0$$

con lo cual,

$$\beta_\varphi(\boldsymbol{\theta}_2) \geq \beta_{\varphi^*}(\boldsymbol{\theta}_2)\ .$$

Esto demuestra que φ es el test más potente de nivel de significación menor o igual que α si su nivel no es cero.

(b) Si $\alpha = 0$, como el test dado por (6.2) tiene nivel cero queremos ver que dado φ^* con nivel 0 se cumple (6.6). Como φ^* tiene nivel 0,

$$\int \varphi^*(\mathbf{x}) p(\mathbf{x}, \boldsymbol{\theta}_1) d\mathbf{x} = 0\ .$$

Por lo tanto, $\varphi^*(\mathbf{x}) = 0$ en el conjunto $\{\mathbf{x} : p(\mathbf{x}, \boldsymbol{\theta}_1) > 0\}$ excepto quizás en un conjunto de medida 0. Por lo tanto, como $\varphi(\mathbf{x}) = 0$ si $p(\mathbf{x}, \boldsymbol{\theta}_1) > 0$ y $\varphi(\mathbf{x}) = 1$ si $p(\mathbf{x}, \boldsymbol{\theta}_1) = 0$ se tiene

$$\begin{aligned}\beta_\varphi(\boldsymbol{\theta}_2) - \beta_{\varphi^*}(\boldsymbol{\theta}_2) &= E_{\boldsymbol{\theta}_2}(\varphi(\mathbf{X})) - E_{\boldsymbol{\theta}_2}(\varphi^*(\mathbf{X})) \\ &= \int_{\{\mathbf{x}:\, p(\mathbf{x},\boldsymbol{\theta}_1)=0\}} [\varphi(\mathbf{X}) - \varphi^*(\mathbf{X})]\, p(\mathbf{x},\boldsymbol{\theta}_2) d\mathbf{x} \\ &\quad + \int_{\{\mathbf{x}:\, p(\mathbf{x},\boldsymbol{\theta}_1)>0\}} [\varphi(\mathbf{X}) - \varphi^*(\mathbf{X})]\, p(\mathbf{x},\boldsymbol{\theta}_2) d\mathbf{x} \\ &= \int_{\{\mathbf{x}:\, p(\mathbf{x},\boldsymbol{\theta}_1)=0\}} [1 - \varphi^*(\mathbf{X})]\, p(\mathbf{x},\boldsymbol{\theta}_2) d\mathbf{x} \geq 0\ .\end{aligned}$$

(iii) Haremos primero el caso $\alpha = 0$. Sea φ el test de la forma (6.2) y φ^* un test de nivel 0. Hemos visto que entonces $\varphi^*(\mathbf{x}) = 0$ en el conjunto $\{\mathbf{x} : p(\mathbf{x},\boldsymbol{\theta}_1) > 0\}$ excepto quizás en un conjunto $\mathcal{N}_1$ de medida 0. Luego, $P_{\boldsymbol{\theta}_1}(\mathcal{N}_1) = P_{\boldsymbol{\theta}_2}(\mathcal{N}_1) = 0$ y $\varphi^*(\mathbf{x}) = \varphi(\mathbf{x})$ en $\{\mathbf{x} : p(\mathbf{x},\boldsymbol{\theta}_1) > 0\} - \mathcal{N}_1$.

Falta ver que $\varphi^*(\mathbf{x}) = \varphi(\mathbf{x}) = 1$ en $\{\mathbf{x} : p(\mathbf{x},\boldsymbol{\theta}_1) = 0\}$ excepto quizás un conjunto de medida 0. Como

$$E_{\boldsymbol{\theta}_2}(\varphi(\mathbf{X})) = E_{\boldsymbol{\theta}_2}(\varphi^*(\mathbf{X}))$$

se cumple

$$\begin{aligned}0 &= \int_{\{\mathbf{x}:\, p(\mathbf{x},\boldsymbol{\theta}_1)=0\}} [\varphi(\mathbf{X}) - \varphi^*(\mathbf{X})]\, p(\mathbf{x},\boldsymbol{\theta}_2) d\mathbf{x} \\ &\quad + \int_{\{\mathbf{x}:\, p(\mathbf{x},\boldsymbol{\theta}_1)>0\}} [\varphi(\mathbf{X}) - \varphi^*(\mathbf{X})]\, p(\mathbf{x},\boldsymbol{\theta}_2) d\mathbf{x} \\ &= \int_{\{\mathbf{x}:\, p(\mathbf{x},\boldsymbol{\theta}_1)=0\}} [1 - \varphi^*(\mathbf{X})]\, p(\mathbf{x},\boldsymbol{\theta}_2) d\mathbf{x}\ .\end{aligned}$$

Pero $\varphi^* \leq 1$ luego el integrando es no negativo y la integral es cero si y solo si $\varphi^* = 1$ en el conjunto $\{\mathbf{x} : p(\mathbf{x},\boldsymbol{\theta}_1) = 0\} \cap \{\mathbf{x} : p(\mathbf{x},\boldsymbol{\theta}_2) > 0\}$ excepto quizás en un conjunto $\mathcal{N}_2$ de medida 0. Luego si

$$\mathcal{N} = \mathcal{N}_1 \cup \mathcal{N}_2 \cup (\{\mathbf{x} : p(\mathbf{x},\boldsymbol{\theta}_1) = 0\} \cap \{\mathbf{x} : p(\mathbf{x},\boldsymbol{\theta}_2) = 0\})$$

se tiene $P_{\boldsymbol{\theta}_1}(\mathcal{N}) = P_{\boldsymbol{\theta}_2}(\mathcal{N}) = 0$ y $\varphi^*(\mathbf{x}) = \varphi(\mathbf{x})$ para $\mathbf{x} \notin \mathcal{N}$.

Supongamos ahora $\alpha > 0$. Sea φ^* un test de nivel α uniformemente más potente para H versus K y φ el test dado por (6.1) que también es uniformemente más potente para H versus K por lo visto en (ii). Luego se cumple

$$E_{\boldsymbol{\theta}_1}(\varphi(\mathbf{X})) = E_{\boldsymbol{\theta}_1}(\varphi^*(\mathbf{X})) \qquad \text{y} \qquad E_{\boldsymbol{\theta}_2}(\varphi(\mathbf{X})) = E_{\boldsymbol{\theta}_2}(\varphi^*(\mathbf{X})) \tag{6.8}$$

Por otra parte, la función $U(\mathbf{x})$ definida en (6.7) es no negativa y por (6.8) $\int U(\mathbf{x})d\mathbf{x} = 0$. Luego, $U(\mathbf{x})$ debe ser nula excepto en un conjunto $\mathcal{N}$ de medida 0. Es decir, $(\varphi(\mathbf{x}) - \varphi^*(\mathbf{x}))(p(\mathbf{x}, \boldsymbol{\theta}_2) - k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1)) = 0$ para $\mathbf{x} \notin \mathcal{N}$. Por lo tanto, $\varphi(\mathbf{x}) = \varphi^*(\mathbf{x})$ en el conjunto $\{\mathbf{x} : p(\mathbf{x}, \boldsymbol{\theta}_2) \neq k_\alpha\, p(\mathbf{x}, \boldsymbol{\theta}_1)\} \cap \mathcal{N}^c$ de donde el resultado.

Observación. Si L_{21} es una variable continua no hay que preocuparse por γ_α, ya que $P(L_{21} = k_\alpha) = 0$.

Ejemplo 1. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a $N(\mu, \sigma_0^2)$ donde σ_0^2 es conocido, y supongamos que se quiere decidir entre H : $\mu = \mu_1$ contra K : $\mu = \mu_2$. Supongamos primero que $\mu_2 > \mu_1$. En este caso, el test más potente rechaza H si

$$\frac{p(X_1, \ldots, X_n; \mu_2)}{p(X_1, \ldots, X_n; \mu_1)} \geq k_\alpha$$

donde $p(X_1, \ldots, X_n; \mu)$ indica la densidad conjunta de $\mathbf{X} = (X_1, \ldots, X_n)$ cuando X_i tiene distribución $N(\mu, \sigma_0^2)$. Luego, $\varphi(X_1, \ldots, X_n) = 1$ si

$$L_{21} = \frac{(2\,\pi\sigma_0)^{-n/2} e^{-\Sigma_{i=1}^n (X_i - \mu_2)^2/2\sigma_0^2}}{(2\,\pi\sigma_0)^{-n/2} e^{-\Sigma_{i=1}^n (X_i - \mu_1)^2/2\sigma_0^2}} \geq k_\alpha$$

o sea $\varphi(X_1, \ldots, X_n) = 1$ si

$$e^{-\Sigma_{i=1}^n (X_i - \mu_2)^2/2\sigma_0^2 + \Sigma_{i=1}^n (X_i - \mu_1)^2/2\sigma_0^2} \geq k_\alpha$$

o equivalentemente, $\varphi(X_1, \ldots, X_n) = 1$ si

$$-\sum_{i=1}^n (X_i - \mu_2)^2 + \sum_{i=1}^n (X_i - \mu_1)^2 \geq 2\sigma_0^2 \ln k_\alpha\,.$$

Desarrollando el primer miembro de esta desigualdad, se tiene que $\varphi(X_1, \ldots, X_n) = 1$ si

$$2(\mu_2 - \mu_1) \sum_{i=1}^n X_i \geq 2\sigma_0^2 \ln k_\alpha + n\mu_2^2 - n\mu_1^2\,.$$

Como $\mu_2 - \mu_1 > 0$, se tiene que $\varphi(X_1, \ldots, X_n) = 1$ si

$$\sum_{i=1}^n X_i \geq \frac{2\sigma_0^2 \ln k_\alpha + n\mu_2^2 - n\mu_1^2}{2(\mu_2 - \mu_1)}$$

pero el segundo miembro de esta desigualdad es una constante, llamémosla k'_α.

Luego, el test más potente es de la forma

$$\varphi(X_1, \ldots, X_n) = 1 \qquad \text{si} \qquad \sum_{i=1}^{n} X_i \geq k'_\alpha$$

(puesto que las regiones de rechazo planteadas inicialmente y esta última son equivalentes). La constante k'_α deberá elegirse de modo que

$$E_{\mu_1}(\varphi(X_1, \ldots, X_n) = \alpha \tag{6.9}$$

Para encontrar el k'_α que hace que (6.9) se satisfaga, necesitaríamos una tabla de la distribución $N(n\mu_1, n\sigma_0^2)$, pero para trabajar más cómodamente transformamos el estadístico $\sum_{i=1}^{n} X_i$ en otro cuya distribución sea $N(0,1)$. Para esto escribimos el test de la siguiente forma $\varphi(X_1, \ldots, X_n) = 1$ si

$$\sqrt{n}\frac{(\overline{X}_n - \mu_1)}{\sigma_0} \geq \sqrt{n}\frac{(k'_\alpha/n - \mu_1)}{\sigma_0}$$

donde $\overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Nuevamente $\sqrt{n}(k'_\alpha/n - \mu_1)/\sigma_0$ es una constante que llamaremos k''_α. Luego el test puede ser escrito de la forma

$$\varphi(X_1, \ldots, X_n) = 1 \quad \text{si } \sqrt{n}\frac{(\overline{X}_n - \mu_1)}{\sigma_0} \geq k''_\alpha.$$

Calculemos k''_α. De acuerdo con el Teorema de Neyman–Pearson, debería tenerse que

$$\begin{aligned} \alpha &= E_{\mu_1}(\varphi(X_1, \ldots, X_n)) \\ &= P_{\mu_1}(\varphi(X_1, \ldots, X_n) = 1) \\ &= P_{\mu_1}(\sqrt{n}\frac{(\overline{X}_n - \mu_1)}{\sigma_0} \geq k''_\alpha)\,. \end{aligned}$$

Pero cuando μ es μ_1, $\sqrt{n}(\overline{X}_n - \mu_1)/\sigma_0$ es $N(0,1)$. Luego, k''_α debe ser igual a z_α.

Finalmente, el test queda como

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \sqrt{n}\frac{(\overline{X}-\mu_1)}{\sigma_0} \geq z_\alpha \\ \\ 0 & \text{si } \sqrt{n}\frac{(\overline{X}-\mu_1)}{\sigma_0} < z_\alpha \end{cases} \tag{6.10}$$

En este caso, no debemos preocuparnos por el caso en que $L_{21} = k_\alpha$ ya que la variable L_{21} es continua.

Si se hubiera tenido que $\mu_2 < \mu_1$, el test más potente de nivel de significación α hubiese resultado

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \sqrt{n}\frac{(\overline{X}_n - \mu_1)}{\sigma_0} \leq -z_\alpha \\ \\ 0 & \text{si } \sqrt{n}\frac{(\overline{X}_n - \mu_1)}{\sigma_0} > -z_\alpha \end{cases} \tag{6.11}$$

De (6.10) resulta que el test más potente para H : $\mu = \mu_1$ contra K : $\mu = \mu_2$ no depende de μ_2, es decir es el mismo cualquiera sea $\mu_2 > \mu_1$. Por lo tanto, el test dado por (6.10) es el test *uniformemente más potente* de nivel menor o igual que α para H : $\mu = \mu_1$ contra K : $\mu > \mu_1$.

Análogamente el test dado por (6.11) es el test *uniformemente más potente* de nivel menor o igual que α para H : $\mu = \mu_1$ contra K :$\mu < \mu_1$.

Calculemos ahora la función de potencia del test φ dado por (6.10), el que se puede escribir, haciendo manipuleo algebraico, como

$$\varphi(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si } \sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma_0} \geq z_\alpha + \sqrt{n}\frac{(\mu_1 - \mu)}{\sigma_0} \\ \\ 0 & \text{si } \sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma_0} < z_\alpha + \sqrt{n}\frac{(\mu_1 - \mu)}{\sigma_0} \end{cases} \tag{6.12}$$

Luego, la función de potencia del test φ definido por (6.10) está dada por

$$\beta_\varphi(\mu) = E_\mu(\varphi(\mathbf{X})) = P_\mu(\sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma_0} \geq z_\alpha + \sqrt{n}\frac{(\mu_1 - \mu)}{\sigma_0})$$

Pero cuando el valor de la media es μ, $\sqrt{n}(\overline{X}_n - \mu)/\sigma_0$ tiene distribución $N(0,1)$. Luego si Φ es la función de distribución de una variable aleatoria $N(0,1)$, se tendrá

$$\beta_\varphi(\mu) = 1 - \Phi(z_\alpha + \sqrt{n}\frac{(\mu_1 - \mu)}{\sigma_0}) \; .$$

Estudiaremos algunas propiedades de $\beta_\varphi(\mu)$.

A. $\beta_\varphi(\mu)$ para n fijo es una función creciente de μ, ya que Φ es una función creciente.

B. $\beta_\varphi(\mu_1) = \alpha$.

C. $\lim_{\mu\to+\infty}\beta_\varphi(\mu) = 1 - \lim_{x\to-\infty}\Phi(x) = 1 - 0 = 1.$

D. $\lim_{\mu\to-\infty}\beta_\varphi(\mu) = 1 - \lim_{x\to+\infty}\Phi(x) = 1 - 1 = 0.$

E. Para μ_2 fijo, $\mu_2 > \mu_1$, se tiene

$$\lim_{n\to\infty}\beta_\varphi(\mu_2) = 1 - \lim_{n\to\infty}\Phi(x) = 1 - 0 = 1.$$

De aquí se deduce que tomando n grande, para un μ_2 fijo, la probabilidad de error de tipo 2 se puede hacer tan pequeño como se quiera.

De A y B resulta que

$$\sup_{\mu\le\mu_1}\beta_\varphi(\mu) = \alpha,$$

y luego φ resulta un test de nivel igual que α para H : $\mu \le \mu_1$ contra K : $\mu > \mu_1$.

Veamos ahora que φ es el test de nivel $\le \alpha$, uniformemente más potente para estas mismas hipótesis. Sea φ^* otro test de nivel menor o igual que α para H : $\mu \le \mu_1$; también tendrá este nivel para H : $\mu = \mu_1$, pero φ es el test uniformemente más potente para H : $\mu = \mu_1$ contra K : $\mu > \mu_1$. Entonces se tiene

$$\beta_\varphi(\mu) \ge \beta_{\varphi^*}(\mu) \quad \forall\, \mu > \mu_1$$

y φ resulta el test más potente de nivel menor o igual que α para H : $\mu \le \mu_1$ contra K : $\mu > \mu_1$.

Luego hemos demostrado el siguiente teorema

Teorema 2.

(i) El test φ dado por (6.10) es el uniformemente más potente de nivel menor o igual que α para
(a) H : $\mu = \mu_1$ contra K : $\mu > \mu_1$
y para
(b) H : $\mu \le \mu_1$ contra K : $\mu > \mu_1$.
Su función de potencia viene dada por

$$\beta_\varphi(\mu) = 1 - \Phi(z_\alpha + \sqrt{n}(\mu_1 - \mu)/\sigma_0).$$

(b) En forma similar el test φ dado por (6.11) es el uniformemente más potente de nivel menor o igual que α para
(a) H : $\mu = \mu_1$ contra K : $\mu < \mu_1$

y para
(b) $H : \mu \geq \mu_1$ *contra* $K : \mu < \mu_1$.
Su función de potencia viene dada por

$$\beta_\varphi(\mu) = \Phi(-z_\alpha + \sqrt{n}(\mu_1 - \mu)/\sigma_0)$$

Ejemplo 2. Supongamos que se mide el grado de impurezas de un producto químico. El método de medición está afectado por un error que se supone $N(0, \sigma_0^2)$, con σ_0^2 conocida igual a 0.01. Además los errores correspondientes a diferentes mediciones son independientes entre sí. Se sabe que el producto es aceptable si el grado de impurezas es menor o igual que 0.7. Se hacen 64 observaciones, $X_1, \ldots, X_{64}$, y se quiere decidir entre las hipótesis: $\mu < 0.7$ ó $\mu \geq 0.7$. Se quiere encontrar un test de modo que la probabilidad de aceptar el producto, cuando éste no satisfaga las condiciones, sea menor que 0.05. Sabemos que cada X_i puede escribirse

$$X_i = \mu + \varepsilon_i$$

donde μ es el grado de impureza y ε_i el error de medición para la observación $i-$esima. Como los ε_i se supusieron normales e independientes, las X_i serán una muestra aleatoria de la distribución $N(\mu, \sigma_0^2)$.

Lo primero que tenemos que determinar es cuál hipótesis es H y cuál K. Tomamos H : $\mu \geq 0.7$, ya que rechazar esta hipótesis equivale a aceptar el producto, y esto queremos hacerlo solamente si estamos muy seguros de su bondad. Luego, se tiene el problema:

$$\text{H} : \mu \geq 0.7 \qquad \text{contra} \qquad \text{K} : \mu < 0.7$$

y por lo tanto, el test más potente de nivel 0.05 está dado por $\varphi(\mathbf{X}) = 1$ si

$$\sqrt{64}\frac{(\overline{X} - 0.7)}{0.1} \leq -z_{0.05} \, .$$

En las tablas se encuentra que $-z_{0.05} = -1.65$. Así, el test rechaza H, es decir, acepta el producto si

$$\overline{X} \leq \frac{-1.65 \times 0.1}{8} + 0.7 = 0.68 \, .$$

Supongamos ahora que se quiere conocer la probabilidad de cometer error de tipo 2, o sea, de aceptar H cuando es falsa (rechazar el producto cuando

cumple la especificación). Tenemos que calcular la función de potencia del test. De acuerdo con lo que hemos visto, será

$$\beta_\varphi(\mu) = \Phi\left(-1.65 - 8\frac{(\mu - 0.7)}{0.1}\right) = \Phi(54.35 - 80\mu) \ .$$

Si queremos, por ejemplo, calcular $\beta_\varphi(0.65)$, esto será uno menos la probabilidad de rechazar el producto cuando $\mu = 0.65$, luego

$$\beta_\varphi(0.65) = \Phi(54.35 - 80 \times 0.65) = \Phi(2.35) = 0.99 \ .$$

Esto quiere decir que la probabilidad de rechazar la droga, cuando $\mu = 0.65$ es 0.01.

6.4 Tests uniformemente más potentes para hipótesis unilaterales

Hemos visto en el parágrafo anterior la forma de encontrar tests más potentes en el caso de hipótesis simples

$$\mathrm{H} : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \qquad \text{contra} \qquad \mathrm{K} : \boldsymbol{\theta} = \boldsymbol{\theta}_1 \ .$$

Esta situación es principalmente de interés teórico puesto que aún las situaciones más simples que se presentan en la práctica, cuando $\theta \in I\!R$, implican problemas de la forma

(a) $\mathrm{H} : \theta = \theta_0$ contra $\mathrm{K} : \theta > \theta_0$

(b) $\mathrm{H} : \theta = \theta_0$ contra $\mathrm{K} : \theta < \theta_0$

(c) $\mathrm{H} : \theta \leq \theta_0$ contra $\mathrm{K} : \theta > \theta_0$

(d) $\mathrm{H} : \theta \geq \theta_0$ contra $\mathrm{K} : \theta < \theta_0$

(e) $\mathrm{H} : \theta = \theta_0$ contra $\mathrm{K} : \theta \neq \theta_0$

Los problemas (a) a (d) se denominan *unilaterales* y al (e) *bilateral*. Hemos visto que para el caso $N(\mu, \sigma_0^2)$ con σ_0^2 conocido se puede extender el test de Neyman–Pearson a hipótesis compuestas de la forma

$\mathrm{H} : \mu = \mu_0$ contra $\mathrm{K} : \mu > \mu_0$
$\mathrm{H} : \mu \leq \mu_0$ contra $\mathrm{K} : \mu > \mu_0$
$\mathrm{H} : \mu = \mu_0$ contra $\mathrm{K} : \mu < \mu_0$

H : $\mu \geq \mu_0$ contra K : $\mu < \mu_0$
obteniéndose tests uniformemente más potentes para estos problemas.

La obtención de tests uniformemente más potentes para hipótesis unilaterales a partir de Neyman–Pearson es siempre posible para ciertas familias de distribuciones que tienen una propiedad llamada de *cociente de verosimilitud monótono.*

Definición 1. Una familia de distribuciones discretas o continuas con densidad (o función de probabilidad puntual) $p(\mathbf{x}, \theta)$, $\theta \in \Theta \subset I\!R$ se dice de *cociente de verosimilitud monótono* (CVM) *en el estadístico* $T = r(\mathbf{X})$ donde r toma valores reales, si para todo par $\theta_1 < \theta_2$ se tiene

(i) Las distribuciones correspondientes a $p(\mathbf{x}, \theta_1)$ y $p(\mathbf{x}, \theta_2)$ son distintas

(ii) $p(\mathbf{x}, \theta_2)/p(\mathbf{x}, \theta_1) = g_{\theta_1\theta_2}(r(\mathbf{x}))$, donde $g_{\theta_1\theta_2}(t)$ es una función no decreciente en el conjunto

$$\mathcal{S} = \{t : t = r(\mathbf{x}) \quad \text{con } p(\mathbf{x}, \theta_1) > 0 \text{ó } p(\mathbf{x}, \theta_2) > 0\}$$

Observación. A los efectos de la Definición 1 si $p(\mathbf{x}, \theta_1) = 0$ y $p(\mathbf{x}, \theta_2) > 0$, el cociente $p(\mathbf{x}, \theta_2)/p(\mathbf{x}, \theta_1)$ se considerará igual a ∞.

Es sencillo mostrar que las familias exponenciales a un parámetro con $c(\theta)$ estrictamente monótona son de CVM.

Teorema 1. *Sea la familia exponencial a un parámetro con función de densidad o probabilidad* $p(\mathbf{x}, \theta) = A(\theta)e^{c(\theta)r(\mathbf{x})}h(\mathbf{x})$ *con* $\theta \in \Theta \subset I\!R$. *Luego,*

(i) Si $c(\theta)$ *es estrictamente creciente la familia dada es de CVM en* $r(\mathbf{X})$

(ii) *Si* $c(\theta)$ *es estrictamente decreciente la familia dada es de CVM en* $-r(\mathbf{X})$

Demostración. Sólo demostraremos (i). La parte (ii) se demuestra idénticamente. En este caso se tiene si $\theta_1 < \theta_2$

$$\frac{p(\mathbf{x}, \theta_2)}{p(\mathbf{x}, \theta_1)} = \frac{A(\theta_2)}{A(\theta_1)}\, e^{(c(\theta_2)-c(\theta_1))r(\mathbf{x})} = g_{\theta_1\theta_2}(r(\mathbf{x}))$$

donde

$$g_{\theta_1\theta_2}(t) = \frac{A(\theta_2)}{A(\theta_1)}\, e^{(c(\theta_2)-c(\theta_1))t}$$

es una función creciente.

Por otro lado, por ser c estrictamente monótona, $\theta_1 \neq \theta_2$ implica $c(\theta_1) \neq c(\theta_2)$ y luego $p(\mathbf{x}, \theta_1)$ y $p(\mathbf{x}, \theta_2)$ corresponden a distribuciones diferentes. Luego, la familia dada es de cociente de verosimilitud monótono en $T = r(\mathbf{X})$.

Vamos a mostrar ahora que existen familias de CVM que no son exponenciales. Para ello consideramos el siguiente ejemplo

Ejemplo 1. Consideremos una muestra aleatoria $(X_1, \ldots, X_n)$ de una distribución $U[0, \theta]$ con $\theta \in I\!R^+$.

Luego, la familia de distribuciones conjuntas de $\mathbf{X} = (X_1, \ldots, X_n)$ se puede escribir

$$p(\mathbf{x}, \theta) = \frac{1}{\theta^n} I_{[0,\theta]}(\max_{1 \leq i \leq n} x_i)\, I_{[0,\infty]}(\min_{1 \leq i \leq n} x_i)\ . \tag{6.13}$$

Mostraremos que esta familia es de CVM en $r(\mathbf{X}) = \max_{1 \leq i \leq n} X_i$. Sea $\theta_2 > \theta_1$, luego, el conjunto $\mathcal{S} = \{t : r(\mathbf{x}) \text{ con } p(\mathbf{x}, \theta_1) > 0 \text{ o } p(\mathbf{x}, \theta_2) > 0\}$ resulta igual al intervalo $[0, \theta_2]$. Definiendo

$$g_{\theta_1 \theta_2}(t) = \frac{\theta_1^n}{\theta_2^n} \frac{I_{[0,\theta_2]}(t)}{I_{[0,\theta_1]}(t)}\ ,$$

se tiene que

$$\frac{p(\mathbf{x}, \theta_2)}{p(\mathbf{x}, \theta_1)} = g_{\theta_1 \theta_2}(r(\mathbf{x})).$$

Po lo tanto, bastará mostrar que $g_{\theta_1 \theta_2}(t)$ es monótona en $\mathcal{S}$. Pero

$$g_{\theta_1 \theta_2}(t) = \begin{cases} (\theta_1/\theta_2)^n & \text{si } 0 \leq t \leq \theta_1 \\ \infty & \text{si } \theta_1 \leq t \leq \theta_2\ . \end{cases}$$

Con lo cual, $g_{\theta_1 \theta_2}(t)$ es monótona y la familia dada por (6.13) es de CVM en $r(\mathbf{X})$. Por otro lado, la familia dada por (6.13) no es exponencial de acuerdo a lo visto en el ejercicio 2 del Capítulo 3.

Ejemplo 2. Consideremos una variable aleatoria X con distribución $\mathcal{C}(\theta, 1)$, $\theta \in I\!R$, o sea, su densidad viene dada por

$$p(x, \theta) = \frac{1}{\pi\,[1 + (x - \theta)^2]}\ .$$

Veremos que esta familia no es de cociente de verosimiltud monótono en $r(X) = X$.

Sea $\theta_2 > \theta_1$, luego, se tiene que

$$\frac{p(x,\theta_2)}{p(x,\theta_1)} = \frac{[1+(x-\theta_1)^2]}{[1+(x-\theta_2)^2]} = g_{\theta_1\theta_2}(x).$$

Sin embargo, la función $g_{\theta_1\theta_2}(x)$ no es monótona en x ya que $\lim_{x\to-\infty} g_{\theta_1\theta_2}(x) = \lim_{x\to+\infty} g_{\theta_1\theta_2}(x) = 1$.

El siguiente teorema nos permite encontrar tests UMP para familia con la propiedad de CVM.

Teorema 1. *Sea* $\mathbf{X}$ *un vector aleatorio con función de probabilidad o densidad perteneciente a la familia* $p(\mathbf{x},\theta)$ *con* $\theta \in \Theta \subset I\!R$, *que tiene la propiedad de ser de CVM en* $T = r(\mathbf{X})$. *Luego*

(i) *Existen* k_α *y* γ_α *tales que si definimos*

$$\varphi(\mathbf{X}) = \begin{cases} 1 & si\ T > k_\alpha \\ \gamma_\alpha & si\ T = k_\alpha \\ 0 & si\ T < k_\alpha \end{cases} \tag{6.14}$$

se satisface

$$E_{\theta_1}(\varphi(\mathbf{X})) = \alpha\ . \tag{6.15}$$

(ii) *Sea* φ *es un test de la forma (6.14) que satisface (6.15). Luego* φ *es el test uniformemente más potente UMP de nivel menor o igual que* α *para*

$$H : \theta = \theta_1 \text{ contra } K : \theta > \theta_1\ .$$

(iii) $\beta_\varphi(\theta)$ *es monótona no decreciente para todo* θ *y estrictamente creciente para todo* θ *tal que* $0 < \beta_\varphi(\theta) < 1$.

(iv) *Sea* φ *un test de la forma (6.14) que satisface (6.15). Luego,* φ *es el test uniformemente más potente de nivel menor o igual que* α *para*

$$H : \theta \leq \theta_1 \text{ contra } K : \theta > \theta_1\ .$$

DEMOSTRACIÓN: La demostración de (i) es idéntica a la dada en el Teorema de Neyman-Pearson.

Demostraremos (ii) suponiendo que si $\theta_2 > \theta_1$

$$\frac{p(\mathbf{x},\theta_2)}{p(\mathbf{x},\theta_1)} = g_{\theta_1\theta_2}(r(\mathbf{x}))$$

con $g_{\theta_1\theta_2}(t)$ estrictamente creciente. (Esta hipótesis no es necesaria, basta con que sea no decreciente.) En este caso, dado $\theta_2 > \theta_1$ el test dado por (6.14) se puede escribir como

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad g_{\theta_1\theta_2}(r(\mathbf{X})) > g_{\theta_1\theta_2}(k_\alpha) \\ \gamma_\alpha & \text{si} \quad g_{\theta_1\theta_2}(r(\mathbf{X})) = g_{\theta_1\theta_2}(k_\alpha) \\ 0 & \text{si} \quad g_{\theta_1\theta_2}(r(\mathbf{X})) < g_{\theta_1\theta_2}(k_\alpha) \end{cases}$$

y si llamamos $k'_\alpha = g_{\theta_1\theta_2}(k_\alpha)$ resulta

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \dfrac{p(\mathbf{X},\theta_2)}{p(\mathbf{X},\theta_1)} > k'_\alpha \\ \gamma_\alpha & \text{si} \quad \dfrac{p(\mathbf{X},\theta_2)}{p(\mathbf{X},\theta_1)} = k'_\alpha \\ 0 & \text{si} \quad \dfrac{p(\mathbf{X},\theta_2)}{p(\mathbf{X},\theta_1)} < k'_\alpha \ . \end{cases}$$

Como $\varphi(\mathbf{X})$ satisface (6.15), usando el Teorema 1 de 6.3 resulta que $\varphi(\mathbf{X})$ es el test más potente de nivel menor o igual que α para H : $\theta = \theta_1$ contra K : $\theta = \theta_2$. Como φ no depende de θ_2, este resultado vale para todo $\theta_2 > \theta_1$, luego φ es el test UMP de nivel menor o igual que α para H : $\theta = \theta_1$ contra K : $\theta > \theta_2$.

(iii) Sólo demostraremos que $\beta_\varphi(\theta)$ es monótona no decreciente.

Sean θ^* y θ^{**} cualesquiera, tales que $\theta^* < \theta^{**}$. Si llamamos $\alpha^* = E_{\theta^*}(\varphi(\mathbf{X}))$, resulta por (ii) que $\varphi(\mathbf{X})$ es el test más potente a nivel menor o igual que α^* para las hipótesis simples

$$\text{H} : \theta = \theta^* \text{ contra K} : \theta = \theta^{**} \ .$$

Consideremos ahora el test

$$\varphi^*(\mathbf{X}) = \alpha^* \ .$$

φ^* es un test de nivel α^*, luego φ^* es menos potente que φ en θ^{**}, es decir,

$$E_{\theta^{**}}(\varphi^*(\mathbf{X})) \leq E_{\theta^{**}}(\varphi(\mathbf{X}))$$

pero,

$$E_{\theta^{**}}(\varphi^*(\mathbf{X})) = \alpha^* = E_{\theta^*}(\varphi(\mathbf{X})) = \beta_\varphi(\theta^*)$$

y además

$$E_{\theta^{**}}(\varphi(\mathbf{X})) = \beta_\varphi(\theta^{**})$$

por lo tanto,

$$\beta_\varphi(\theta^*) \leq \beta_\varphi(\theta^{**}) \ ,$$

con lo que queda demostrado que $\beta_\varphi(\theta)$ es monótona no decreciente.

Para demostrar (iv), primero mostraremos que $\varphi(\mathbf{X})$ es un test de nivel menor o igual que α para

$$\mathrm{H} : \theta \leq \theta_1 \text{ contra } \mathrm{K} : \theta > \theta_1$$

o sea que

$$\sup_{\theta \leq \theta_1} \beta_\varphi(\theta) \leq \alpha \ .$$

Como $\beta_\varphi(\theta)$ es monótona creciente se tiene:

$$\sup_{\theta \leq \theta_1} \beta_\varphi(\theta) = \beta_\varphi(\theta_1) = \alpha$$

por (6.15).

Consideremos ahora otro test $\varphi^*(\mathbf{X})$ de nivel menor o igual que α para $\mathrm{H} : \theta \leq \theta_1$ contra $\mathrm{K} : \theta > \theta_1$, luego $\varphi^*(\mathbf{X})$ es de nivel menor o igual que α para $\mathrm{H} : \theta = \theta_1$ contra $\mathrm{K} : \theta > \theta_1$, pero por (ii) $\varphi(\mathbf{X})$ es el test uniformemente más potente para este problema, por lo tanto

$$\beta_\varphi(\theta) \geq \beta_{\varphi^*}(\theta) \quad \forall\, \theta > \theta_1 \ .$$

Análogamente se demuestra el siguiente teorema

Teorema 2. *Sea* $\mathbf{X}$ *un vector aleatorio con función de densidad perteneciente a la familia* $p(\mathbf{x}, \theta)$ *con* $\theta \in \Theta \subset I\!R$. *Supongamos que esta familia es CMV en* $r(\mathbf{X})$. *Luego*

(i) *Existen* k_α *y* γ_α *tales que si definimos*

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } r(\mathbf{X}) < k_\alpha \\ \gamma_\alpha & \text{si } r(\mathbf{X}) = k_\alpha \\ 0 & \text{si } r(\mathbf{X}) > k_\alpha \end{cases} \tag{6.16}$$

se satisface

$$E_{\theta_1}(\varphi(\mathbf{X})) = \alpha \tag{6.17}$$

(ii) *Sea $\varphi(\mathbf{X})$ es un test de la forma (6.16) que satisface (6.17). Luego φ es el test uniformemente más potente a nivel menor o igual que α para $H : \theta = \theta_1$ contra $K : \theta < \theta_1$.*

(iii) *$\beta_\varphi(\theta)$ es monótona no creciente para todo θ y estrictamente decreciente para todo θ tal que $0 < \beta_\varphi(\theta) < 1$.*

(iv) *Sea φ un test de la forma (6.16) que satisface (6.17). Luego φ es el test uniformemente más potente de nivel menor o igual que α para $H : \theta \geq \theta_1$ contra $K : \theta < \theta_1$.*

Para una versión más completa de este Teorema, ver Teorema 2 de 3.3 en Lehmann [2].

Ejemplo 3. Consideremos una muestra aleatoria $X_1, \ldots, X_n$ de una distribución perteneciente a la familia $N(\mu, \sigma_0^2)$ con σ_0^2 conocido. Luego, es fácil demostrar que la familia de distribuciones de la muestra es exponencial con $r(\mathbf{X}) = \sum_{i=1}^n X_i$ y $c(\mu) = \mu/\sigma_0^2$. Como $c(\mu)$ es creciente de acuerdo al Teorema 1, esta familia es de CMV en $r(\mathbf{X})$. Entonces para testear H : $\mu \leq \mu_1$ contra K : $\mu > \mu_1$, el test UMP de nivel menor o igual que α, es de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } \displaystyle\sum_{i=1}^n X_i \geq k_\alpha \\ \\ 0 & \text{si } \displaystyle\sum_{i=1}^n X_i < k_\alpha \end{cases}$$

con $E_{\mu_1}(\varphi(\mathbf{X}) = \alpha$.

En la Sección 6.3 ya habíamos demostrado este resultado y hallado el valor de k_α.

Ejemplo 4. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $Bi(\theta, 1)$.

En este caso la familia de distribuciones de $X_1, \ldots, X_n$ es exponencial con $T = r(\mathbf{X}) = \sum_{i=1}^n X_i$ y $c(\theta) = \ln(\theta/(1-\theta))$; como $c(\theta)$ es creciente, esta familia es de CMV en $r(\mathbf{X})$.

Luego, el test UMP de nivel menor o igual que α para H : $\theta \leq \theta_1$ contra K : $\theta > \theta_1$ será de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } T > k_\alpha \\ \gamma_\alpha & \text{si } T = k_\alpha \\ 0 & \text{si } T < k_\alpha \end{cases}$$

k_α y γ_α deberán ser elegidos de modo que

$$E_{\theta_1}(\varphi(\mathbf{X})) = P_{\theta_1}(T > k_\alpha) + \gamma_\alpha P_{\theta_1}(T = k_\alpha) = \alpha \ . \tag{6.18}$$

Como T tiene distribución $Bi(\theta, n)$ que es discreta, puede suceder que exista o no k tal que

$$P_{\theta_1}(T > k) = \alpha \tag{6.19}$$

Si existe k satisfaciendo (6.19), tomaremos ese valor como k_α y $\gamma_\alpha = 0$.
Si no existe k que verifique (6.19), siempre existirá k tal que

$$P_{\theta_1}(T > k) < \alpha < P_{\theta_1}(T \geq k) \ . \tag{6.20}$$

Este valor k será el k_α que eligiremos y reemplazándolo en (6.18) obtendremos

$$\gamma_\alpha = \frac{\alpha - P_{\theta_1}(T > k_\alpha)}{P_{\theta_1}(T = k_\alpha)} = \frac{\alpha - P_{\theta_1}(T > k_\alpha)}{P_{\theta_1}(T \geq k_\alpha) - P_{\theta_1}(T > k_\alpha)} \ .$$

Por (6.20) resulta que $0 < \gamma_\alpha < 1$.
Para encontrar el k_α que verifica (6.19) o (6.20) deberán usarse tablas binomiales.

Recordemos finalmente que

$$P_{\theta_1}(T \geq k_\alpha) = \sum_{k_\alpha \leq i \leq n} \binom{n}{i} \theta_1^i (1 - \theta_1)^{n-i} \ .$$

Supongamos que se tiene una muestra aleatoria X_1, X_2, X_3 de una distribución $Bi(\theta, 1)$ y se quiere testear H : $\theta \leq 1/3$ contra K : $\theta > 1/3$ con nivel de significación menor o igual que 0.1.

Cuando $\theta = 1/3$, la distribución de $T = \sum_{i=1}^{3} X_i$ está dada por

t	0	1	2	3
$p_T(t)$	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$

y por lo tanto, tenemos

k	-1	0	1	2	3
$P_{\frac{1}{3}}(T > k)$	1	$\frac{19}{27}$	$\frac{7}{27}$	$\frac{1}{27}$	0

Por lo tanto, no existe k_α que verifique (6.19) y el valor $k_\alpha = 2$ verifica (6.20), pues

$$P_{\frac{1}{3}}(T > 2) = \frac{1}{27} < 0.1 < P_{\frac{1}{3}}(T \geq 2) = P_{\frac{1}{3}}(T > 1) = \frac{7}{27}$$

y γ_α será entonces

$$\gamma_\alpha = \frac{0.1 - \frac{1}{27}}{\frac{6}{27}} = 0.27 \quad .$$

Como ejercicio se sugiere graficar la función de potencia de este test, y siendo el test aleatorizado, sugerir un mecanismo para decidir en caso en que $T = 2$.

Ejemplo 5. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $U[0, \theta]$.

El test uniformemente más potente para H : $\theta \leq \theta_1$ contra K : $\theta > \theta_1$, será de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } \max\limits_{1 \leq i \leq n} X_i \geq k_\alpha \\ \\ 0 & \text{si } \max\limits_{1 \leq i \leq n} X_i < k_\alpha \end{cases}$$

donde k_α verifica

$$E_{\theta_1}(\varphi(\mathbf{X})) = \alpha \; . \tag{6.21}$$

Teniendo en cuenta que la función de distribución de $T = \max_{1 \leq i \leq n} X_i$ es

$$F_T(t) = \begin{cases} 0 & \text{si} \quad t < 0 \\ (t/\theta)^n & \text{con} \quad 0 \leq t \leq \theta \\ 1 & \text{si} \quad t > 1 \end{cases}$$

y que debe cumplirse (6.21), se tiene $0 \leq k_\alpha \leq \theta_1$ y

$$P_{\theta_1}\left(\max_{1 \leq i \leq n} X_i \geq k_\alpha\right) = 1 - (k_\alpha/\theta_1)^n = \alpha \; ,$$

de donde resulta

$$k_\alpha = \theta_1 \sqrt[n]{1 - \alpha} \; .$$

6.5 Tests insesgados

En la mayoría de los casos en que la hipótesis alternativa es una hipótesis compuesta, no existe un test uniformemente más potente.

Ejemplo 1. Supongamos que se tiene una muestra aleatoria $X_1, \ldots, X_n$ de una distribución $N(\mu, \sigma_0^2)$ con σ_0 conocido y se desea testear $\mathrm{H} : \mu = \mu_0$ contra $\mathrm{K} : \mu \neq \mu_0$. Es fácil demostrar que no existe un test uniformemente más potente a nivel menor o igual que α.

Supongamos que tal test existiera y llamémoslo φ; entonces será el test más potente a nivel menor o igual que α para

$$\mathrm{H}_1 : \mu = \mu_0 \text{ contra } \mathrm{K}_1 : \mu = \mu_1 \quad (\mu_1 > \mu_0)$$

y para

$$\mathrm{H}_2 : \mu = \mu_0 \text{ contra } \mathrm{K}_2 : \mu = \mu_2 \quad (\mu_2 < \mu_0) \, .$$

Pero, por el Teorema 3 de la Sección 6.3 el test más potente para H_1 contra K_1 está dado por

$$\varphi_1(\mathbf{X}) = \begin{cases} 1 & \text{si } \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{\sigma_0} \geq z_\alpha \\ \\ 0 & \text{si } \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{\sigma_0} < z_\alpha \end{cases}$$

y el test más potente para H_2 contra K_2 está dado por

$$\varphi_2(\mathbf{X}) = \begin{cases} 1 & \text{si } \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{\sigma_0} \leq -z_\alpha \\ \\ 0 & \text{si } \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{\sigma_0} > -z_\alpha \end{cases}$$

Entonces, por la unicidad dada en el Teorema de Neyman-Pearson, φ debería coincidir con φ_1 y con φ_2 lo cual es imposible.

Recordemos que en el caso de estimadores puntuales tampoco existe en general uno de menor error cuadrático medio. Una manera de poder definir un estimador óptimo que se propuso en el Capítulo 3 fue restringiendo los estimadores a la clase de los insesgados. En el caso de test se procede en forma similar, restringiremos la clase de tests considerados a los que

llamaremos insesgados y luego se buscará el test uniformemente más potente en esta clase.

Definición 1. Sea una familia de distribuciones $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$. Se dirá que un test φ para testear H : $\boldsymbol{\theta} \in \Theta_1$ contra K : $\boldsymbol{\theta} \in \Theta_2$ *es insesgado* si

$$\sup_{\boldsymbol{\theta} \in \Theta_1} \beta_\varphi(\boldsymbol{\theta}) \leq \inf_{\boldsymbol{\theta} \in \Theta_2} \beta_\varphi(\boldsymbol{\theta})$$

El sentido de esta desigualdad es que la probabilidad de rechazar H cuando $\boldsymbol{\theta} \in \Theta_2$, es decir cuando H es falsa, no pude ser menor que cuando $\boldsymbol{\theta} \in \Theta_1$, es decir cuando H es verdadera.

Por lo tanto, un test insesgado de nivel α tiene función de potencia menor o igual que α para $\boldsymbol{\theta} \in \Theta_1$ y mayor o igual que α para $\boldsymbol{\theta} \in \Theta_2$.

Observemos que un test UMP de nivel α es insesgado.

Observación. Si la función de potencia $\beta_\varphi(\boldsymbol{\theta})$ del test φ es una función continua de $\boldsymbol{\theta}$ y φ es un test insesgado de nivel α, entonces $\beta_\varphi(\boldsymbol{\theta})$ debe valer α en la frontera Θ_F entre Θ_1 y Θ_2.
En particular, si $\Theta \subset I\!R$, $\Theta_1 = \{\theta_1\}$ y $\Theta_2 = \Theta - \{\theta_1\}$, o sea, si estamos testeando H : $\theta = \theta_1$ contra K : $\theta \neq \theta_1$, y φ es un test insesgado de nivel α se tiene

$$\begin{aligned} &\beta_\varphi(\theta_1) = \alpha \\ &\beta_\varphi(\theta) \geq \alpha \qquad \forall \theta \neq \theta_1 \ . \end{aligned}$$

Por lo tanto, si la función de potencia $\beta_\varphi(\theta)$ es derivable respecto de θ, φ debe cumplir

$$\beta'_\varphi(\theta_1) = \frac{\partial}{\partial \theta} \beta_\varphi(\theta)|_{\theta=\theta_1} = 0 \ . \tag{6.22}$$

En el caso particular de las familias exponenciales, la función de potencia de cualquier test es derivable y por lo tanto, los tests insesgados cumplen (6.22).

Definición 2. Se dirá que un test φ para testear H : $\boldsymbol{\theta} \in \Theta_1$ contra K : $\boldsymbol{\theta} \in \Theta_2$ *es uniformemente más potente de nivel* α *entre los insesgados*, IUMP, si

(a) φ tiene nivel α, o sea,

$$\sup_{\boldsymbol{\theta} \in \Theta_1} \beta_\varphi(\boldsymbol{\theta}) = \alpha$$

(b) φ es insesgado, es decir,

$$\beta_\varphi(\boldsymbol{\theta}) \geq \alpha \qquad \forall \boldsymbol{\theta} \in \Theta_2$$

(c) Dado otro test φ^* insesgado y de nivel α se verifica

$$\beta_\varphi(\boldsymbol{\theta}) \geq \beta_{\varphi^*}(\boldsymbol{\theta}) \qquad \forall \boldsymbol{\theta} \in \Theta_2 \ .$$

En la próxima Sección daremos un procedimiento general para encontrar tests para un problema determinado. En muchos casos este procedimiento da como resultado el test insesgado uniformente más potente.

La teoría de los tests insesgados uniformemente más potentes escapa a las posibilidades de este curso y puede verse en Lehmann [3] o en Ferguson [2].

6.6 Test del cociente de máxima verosimilitud

Supongamos que se observa un vector $\mathbf{X}$, cuya distribución tiene función de densidad $p(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ y se quiere testear H : $\boldsymbol{\theta} \in \Theta_1$ contra K : $\boldsymbol{\theta} \in \Theta_2$ ($\Theta_1 \cup \Theta_2 = \Theta$).

Un procedimiento intuitivamente razonable y que da buenos resultados en una gran variedad de situaciones es el siguiente.

Tomemos estimadores de máxima verosimilitud de $\boldsymbol{\theta}$, suponiendo $\boldsymbol{\theta} \in \Theta_1$, llamémoslo $\widehat{\boldsymbol{\theta}}_1$ y análogamente suponiendo $\boldsymbol{\theta} \in \Theta_2$, $\widehat{\boldsymbol{\theta}}_2$; luego

$$p(\mathbf{X}, \widehat{\boldsymbol{\theta}}_1) = \max_{\boldsymbol{\theta} \in \Theta_1} p(\mathbf{X}, \boldsymbol{\theta})$$

y

$$p(\mathbf{X}, \widehat{\boldsymbol{\theta}}_2) = \max_{\boldsymbol{\theta} \in \Theta_2} p(\mathbf{X}, \boldsymbol{\theta}) \ .$$

Si $\widehat{\boldsymbol{\theta}}_1$ y $\widehat{\boldsymbol{\theta}}_2$ no dependieran de la muestra, podríamos considerar el test más potente para testear H* : $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_1$ contra K* : $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_2$, el cual es de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } L < k_\alpha \\ \gamma_\alpha & \text{si } L = k_\alpha \\ 0 & \text{si } L > k_\alpha \end{cases}$$

donde

$$L = \frac{1}{L_{21}} = \frac{p(\mathbf{X}, \widehat{\boldsymbol{\theta}}_1)}{p(\mathbf{X}, \widehat{\boldsymbol{\theta}}_2)}$$

y k_α se elige de manera que el test resulte de nivel α.

En algunos casos $\widehat{\boldsymbol{\theta}}_1$ y $\widehat{\boldsymbol{\theta}}_2$ pueden no existir, pero siempre tiene sentido hablar de L definido por

$$L = \frac{\sup_{\boldsymbol{\theta}\in\Theta_1} p(\mathbf{X},\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta_2} p(\mathbf{X},\boldsymbol{\theta})}$$

Intuitivamente, este test puede interpretarse como rechazando $\mathrm{H} : \boldsymbol{\theta} \in \Theta_1$ cuando "el valor más probable de Θ_2" tiene probabilidad considerablemente más grande que "el valor más probable de Θ_1".

En muchos casos, como por ejemplo cuando la dimensión de Θ_1 es menor que la dimensión de $\Theta = \Theta_1 \cup \Theta_2$, y $p(\mathbf{x},\boldsymbol{\theta})$ es continua, resulta que

$$\sup_{\boldsymbol{\theta}\in\Theta_2} p(\mathbf{X},\boldsymbol{\theta}) = \sup_{\boldsymbol{\theta}\in\Theta} p(\mathbf{X},\boldsymbol{\theta}) \tag{6.23}$$

En este caso, el test del cociente de máxima verosimilitud resulta equivalente a

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } L^* < k_\alpha \\ \gamma_\alpha & \text{si } L^* = k_\alpha \\ 0 & \text{si } L^* > k_\alpha \end{cases}$$

donde

$$L^* = \frac{\sup_{\boldsymbol{\theta}\in\Theta_1} p(\mathbf{X},\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta} p(\mathbf{X},\boldsymbol{\theta})} .$$

En general, es más fácil aplicar la forma del test basada en L^* cuando es posible, es decir, cuando (6.23) se cumple.

Ejemplo 1. Se tiene una muestra aleatoria $X_1, \ldots, X_n$ de una distribución $N(\mu, \sigma_0^2)$ con σ_0 conocido y se quiere testear $\mathrm{H} : \mu = \mu_0$ contra $\mathrm{K} : \mu \neq \mu_0$

Como en este caso $\Theta_1 = \{\mu_0\}$ tiene dimensión cero (se reduce a un punto) y $\Theta = \{\mu : -\infty < \mu < +\infty\}$ tiene dimensión uno, podemos usar el test basado en L^*.

Es claro que

$$\sup_{\mu\in\Theta_1} p(\mathbf{X},\mu) = (2\pi\sigma_0^2)^{-\frac{n}{2}} e^{-\frac{1}{2\,\sigma_0^2}\Sigma_{i=1}^n (X_i-\mu_0)^2}$$

y que

$$\sup_{\mu\in\Theta} p(\mathbf{X},\mu) = (2\pi\sigma_0^2)^{-\frac{n}{2}} e^{-\frac{1}{2\,\sigma_0^2}\Sigma_{i=1}^n (X_i-\overline{X})^2} .$$

Luego,

$$L^* = e^{-\frac{1}{2\sigma_0^2}\left(\Sigma_{i=1}^n (X_i-\mu_0)^2 - \Sigma_{i=1}^n (X_i-\overline{X})^2\right)}$$

y como

$$\sum_{i=1}^n (X_i-\mu_0)^2 - \sum_{i=1}^n (X_i-\overline{X})^2 = n(\overline{X}-\mu_0)^2$$

resulta

$$L^* = e^{-\frac{n}{2\sigma_0^2}(\overline{X}-\mu_0)^2} .$$

Sea $T = \sqrt{n}|\overline{X}-\mu_0|/\sigma_0$. Luego, $L^* = g(T)$ con g decreciente. Luego $\varphi(\mathbf{X})$ es equivalente a

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X}-\mu_0|}{\sigma_0} \geq k'_\alpha \\ \\ 0 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X}-\mu_0|}{\sigma_0} < k'_\alpha \, . \end{cases}$$

Obsérvese que este test resulta muy razonable intuitivamente, ya que se rechaza la hipótesis de que $\mu = \mu_0$ si $\overline{X}$ difiere sensiblemente de μ_0.

k'_α debe elegirse de modo tal que φ resulte de nivel α, es decir que

$$P_{\mu_0}(\sqrt{n}\frac{|\overline{X}-\mu_0|}{\sigma_0} \geq k_\alpha) = \alpha \, .$$

Pero como, cuando $\mu = \mu_0$ se tiene que$\sqrt{n}(\overline{X}-\mu_0)/\sigma_0$ tiene distribución $N(0,1)$, resulta que $k'_\alpha = z_{\alpha/2}$.

Ejemplo 2. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$ con varianza σ^2 desconocida y se desea testear H : $\mu = \mu_0$ contra K : $\mu \neq \mu_0$. En este caso,

$$\Theta_1 = \{(\mu_0, \sigma^2) : 0 < \sigma^2 < \infty\}$$

resulta de dimensión uno, y

$$\Theta = \{(\mu_1, \sigma^2) : -\infty < \mu < \infty, \; 0 < \sigma^2 < \infty\}$$

es de dimensión dos.

Por lo tanto utilizaremos el test basado en L^*. El estimador de máxima verosimilitud de (μ, σ^2) restringido a Θ_1 es $(\mu_0, \sum_{i=1}^n (X_i-\mu_0)^2/n)$ y el estimador de máxima verosimilitud de (μ, σ^2) sin restricciones es $(\overline{X}, \sum_{i=1}^n (X_i-\overline{X})^2/n)$.

Luego, se tiene

$$\sup_{(\mu,\sigma^2)\in\Theta_1} p(\mathbf{X},\mu,\sigma^2) = \frac{1}{e^{\frac{n}{2}}(2\pi)^{\frac{n}{2}}\left(\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{n}\right)^{\frac{n}{2}}}$$

y

$$\sup_{(\mu,\sigma^2)\in\Theta} p(\mathbf{X},\mu,\sigma^2) = \frac{1}{e^{\frac{n}{2}}(2\pi)^{\frac{n}{2}}\left(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n}\right)^{\frac{n}{2}}} .$$

Por lo tanto, L^* está dado por

$$L^* = \left[\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sum_{i=1}^n (X_i-\mu_0)^2}\right]^{\frac{n}{2}} .$$

Como

$$\sum_{i=1}^n (X_i-\mu_0)^2 = \sum_{i=1}^n (X_i-\overline{X})^2 + n(\overline{X}-\mu_0)^2$$

se tiene que

$$L^* = \left[1+\frac{n(\overline{X}-\mu_0)^2}{\sum_{i=1}^n (X_i-\overline{X})^2}\right]^{-\frac{n}{2}} .$$

Sea ahora

$$T = \sqrt{n}\frac{(\overline{X}-\mu_0)}{s}$$

donde $s^2 = \sum_{i=1}^n (X_i-\overline{X})^2/(n-1)$. Luego,

$$L^* = \left[\frac{1}{1+\frac{T^2}{n-1}}\right]^{\frac{n}{2}}$$

Como la función $1/(1+t^2/(n-1))$ es monótona decreciente de $|t|$, el test del cociente de máxima verosimilitud resulta equivalente a

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad |T| \geq k_\alpha \\ 0 & \text{si} \quad |T| < k_\alpha \end{cases}$$

y k_α deberá ser elegido de manera que el test resulte con nivel de significación α, es decir, de manera que

$$P_{\mu_0}(|T| \geq k_\alpha) = \alpha .$$

Como T tiene distribución student con $n-1$ grados de libertad, resulta

$$k_\alpha = t_{n-1,\frac{\alpha}{2}} \ .$$

Obsérvese que este test es completamente análogo al del Ejemplo 1, con la diferencia que se reemplaza σ por s y $z_{\alpha/2}$ por $t_{n-1,\frac{\alpha}{2}}$.

Ejemplo 3. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$ con media y varianza desconocidas. Supongamos que se quiere testear H : $\mu \leq \mu_0$ contra K : $\mu > \mu_0$. En este caso,

$$\Theta_1 = \{(\mu, \sigma^2) : \mu \leq \mu_0,\ \sigma^2 > 0\}$$

y

$$\Theta_2 = \{(\mu, \sigma^2) : \mu > \mu_0,\ \sigma^2 > 0\} \ .$$

Luego, la dimensión de Θ_1 es igual a la de Θ_2, y el test del cociente de máxima verosimilitud deberá hacerse con L y no con L^*. Como

$$p(\mathbf{X}, \mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2} \tag{6.24}$$

resulta

$$\ln p(\mathbf{X}, \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 \ . \tag{6.25}$$

Teniendo en cuenta que

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \overline{X})^2 + n(\overline{X} - \mu)^2$$

se obtiene que el estimador de máxima verosimilitud de μ en Θ_1, es igual a

$$\widehat{\mu}_1 = \begin{cases} \overline{X} & \text{si } \overline{X} \leq \mu_0 \\ \mu_0 & \text{si } \overline{X} > \mu_0 \end{cases} \tag{6.26}$$

y que el estimador de máxima verosimilitud de μ en Θ_2, es igual a

$$\widehat{\mu}_2 = \begin{cases} \overline{X} & \text{si } \overline{X} > \mu_0 \\ \mu_0 & \text{si } \overline{X} \leq \mu_0 \end{cases} \ . \tag{6.27}$$

El estimador de máxima verosimilitud de σ^2, para $\boldsymbol{\theta} \in \Theta_1$ es

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_1)^2$$

y para $\boldsymbol{\theta} \in \Theta_2$ es

$$\widehat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_2)^2 \, .$$

Luego, reemplazando en (6.24) se obtiene

$$\max_{(\mu,\sigma^2)\in\Theta_j} p(\mathbf{X}, \mu, \sigma^2) = [2\, e\, \pi \sum_{i=1}^{n}(X_i - \widehat{\mu}_j)^2/n]^{-\frac{n}{2}}$$

para $j = 1, 2$, de donde

$$L = \left[\frac{\sum_{i=1}^n (X_i - \widehat{\mu}_2)^2}{\sum_{i=1}^n (X_i - \widehat{\mu}_1)^2}\right]^{\frac{n}{2}} = \left[\frac{\sum_{i=1}^n (X_i - \overline{X})^2 + n(\overline{X} - \widehat{\mu}_2)^2}{\sum_{i=1}^n (X_i - \overline{X})^2 + (\overline{X} - \widehat{\mu}_1)^2}\right]^{\frac{n}{2}}$$

y usando (6.26) y (6.27) se deduce

$$L = \begin{cases} \left[\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]^{\frac{n}{2}} & \text{si} \quad \overline{X} \leq \mu_0 \\ \left[\dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2}\right]^{\frac{n}{2}} & \text{si} \quad \overline{X} > \mu_0 \end{cases} .$$

Si llamamos

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sqrt{\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}}}$$

se tiene

$$L = \begin{cases} (1 + \frac{T^2}{n-1})^{\frac{n}{2}} & \text{si} \quad \overline{X} \leq \mu_0 \\ (1 + \frac{T^2}{n-1})^{-\frac{n}{2}} & \text{si} \quad \overline{X} > \mu_0 \end{cases} .$$

Luego, el test del cociente de máxima verosimilitud es de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 \quad \text{si} \begin{cases} 1 + \dfrac{T^2}{n-1} \leq k_\alpha \quad \text{y} \quad \overline{X} \leq \mu_0 \quad \text{(A)} \\ \dfrac{1}{1+\frac{T^2}{n-1}} \leq k_\alpha \quad \text{y} \quad \overline{X} > \mu_0 \quad \text{(B)} \end{cases} \\ 0 \quad \text{si} \begin{cases} 1 + \dfrac{T^2}{n-1} > k_\alpha \quad \text{y} \quad \overline{X} \leq \mu_0 \quad \text{(C)} \\ \dfrac{1}{1+\frac{T^2}{n-1}} > k_\alpha \quad \text{y} \quad \overline{X} > \mu_0 \quad \text{(D)} \end{cases} \end{cases} .$$

Tomemos ahora $k_\alpha < 1$ (con $k_\alpha \geq 1$ se llega al mismo resultado), en este caso la primera desigualdad de (A) no puede ocurrir y la primera desigualdad de (C) ocurre siempre, luego $\varphi(\mathbf{X})$ se transforma en

$$\varphi(\mathbf{X}) = \begin{cases} 1 \quad \text{si} \quad \dfrac{1}{1+\frac{T^2}{n-1}} \leq k_\alpha \quad \text{y} \quad \overline{X} > \mu_0 \\ 0 \quad \text{si} \begin{cases} \overline{X} \leq \mu_0 \\ \dfrac{1}{1+\frac{T^2}{n-1}} > k_\alpha \text{ y} \quad \overline{X} > \mu_0 \end{cases} \end{cases} .$$

Esto es equivalente a

$$\varphi(\mathbf{X}) = \begin{cases} 1 \quad \text{si } |T| \geq {k'}_\alpha \quad \text{y} \quad T > 0 \\ 0 \quad \text{si} \begin{cases} |T| < {k'}_\alpha \quad \text{y} \quad T > 0 \\ T < 0 \quad , \end{cases} \end{cases}$$

de donde, se deduce que

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } T \geq {k'}_\alpha \\ 0 & \text{si } T < {k'}_\alpha \, . \end{cases}$$

Debemos ver ahora que se puede elegir ${k'}_\alpha$ de modo que el test resulte de nivel igual α. Esto significa que

$$\sup_{\{\mu \leq \mu_0, \sigma^2 > 0\}} P_{\mu,\sigma^2}(T \geq {k'}_\alpha) = \alpha \, .$$

Se puede pensar que el caso más desfavorable, en el cual hay mayor probabilidad de rechazar H, es en el caso límite $\mu = \mu_0$; por lo tanto parece razonable elegir ${k'}_\alpha$ de manera que

$$P_{\mu_0,\sigma^2}(T \geq {k'}_\alpha) = \alpha \ .$$

Pero cuando $\mu = \mu_0$, T tiene distribución de Student con $n-1$ grados de libertad, y por lo tanto debemos tomar

$${k'}_\alpha = t_{n-1,\alpha} \ .$$

El test φ resulta entonces

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad T \geq t_{n-1,\alpha} \\ 0 & \text{si} \quad T < t_{n-1,\alpha} \ . \end{cases}$$

Debemos probar ahora que este test tiene realmente nivel α, es decir que,

$$P_{\mu,\sigma^2}(T \geq t_{n-1,\alpha}) \leq \alpha \qquad \forall\, \mu \leq \mu_0 \ .$$

Para ello necesitaremos la siguiente definición.

Definición 1. Llamaremos distribución de Student no central con n grados de libertad y parámetro de no centralidad Δ, $-\infty < \Delta < \infty$, que simbolizaremos por $\mathcal{T}_n(\Delta)$ a la distribución de

$$\frac{U + \Delta}{\sqrt{V/n}}$$

donde U tiene distribución $N(0,1)$ donde V tiene distribución χ^2_n siendo U y V independientes.

Teorema 1. *Sea X_Δ una variable aleatoria con distribución de Student no central $\mathcal{T}_n(\Delta)$, definamos $c_{n,k}(\Delta)$ por*

$$c_{n,k}(\Delta) = P(X \geq k),$$

luego, $c_{n,k}(\Delta)$ es una función monótona creciente de Δ.

Demostración. Como X_Δ tiene distribución $\mathcal{T}_n(\Delta)$; se puede escribir

$$X_\Delta = \frac{U + \Delta}{\sqrt{V/n}}$$

donde U es una variable aleatoria $N(0,1)$ y V tiene distribución χ^2_n, independientes. Luego,

$$c_{n,k}(\Delta) = P(X_\Delta \geq k) = E\left[P(X_\Delta \geq k|V)\right],$$

pero

$$P(X_\Delta \geq k|V = v) = P\left(\frac{U+\Delta}{\sqrt{v/n}} \geq k\,|V = v\right) = 1 - \Phi(k\sqrt{\frac{v}{n}} - \Delta)\,.$$

Luego esta última probabilidad, para k, n y v fijos, es una función creciente de Δ. Por lo tanto, si $\Delta_1 < \Delta_2$ se tiene

$$P(X_{\Delta_1} \geq k|V = v) < P(X_{\Delta_2} \geq k|V = v)$$

con lo cual, tomando esperanza se obtiene

$$E\left(P(X_{\Delta_1} \geq k)|V\right) < E\left(P(X_{\Delta_2} \geq k)|V\right)$$

o sea

$$P(X_{\Delta_1} \geq k) < P(X_{\Delta_2} \geq k),$$

y por lo tanto $c_{n,k}(\Delta)$ es creciente en Δ.

Volvamos ahora al Ejemplo 3. Vamos a mostrar que el test φ dado por

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad T \geq t_{n-1,\alpha} \\ 0 & \text{si} \quad T < t_{n-1,\alpha} \end{cases}$$

tiene nivel de significación α. Como

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2}} = \frac{\sqrt{n}\frac{(\overline{X}-\mu_0)}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2}}}$$

resulta

$$T = \frac{\sqrt{n}\frac{(\overline{X}-\mu)}{\sigma} + \sqrt{n}\frac{(\mu-\mu_0)}{\sigma}}{\sqrt{\frac{1}{n-1}\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2}}}\,.$$

Llamando $U = \sqrt{n}(\overline{X} - \mu)/\sigma$ y $V = \sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2$ se tiene que U y V son independientes, y cuando los valores de los parámetros son μ y

σ^2, U tiene distribución $N(0,1)$ y V tiene distribución χ^2_{n-1}. Luego T tiene distribución $\mathcal{T}_{n-1}(\Delta)$ donde $\Delta = \sqrt{n}(\mu - \mu_0)/\sigma$. Además,

$$\beta_\varphi(\mu, \sigma^2) = P_{\mu,\sigma^2}(T \geq t_{n-1,\alpha}) = c_{n-1,t_{n-1,\alpha}}(\Delta) \ .$$

Resulta, por el Teorema 1, que $\beta_\varphi(\mu, \sigma^2)$ es una función creciente de μ para cada σ^2 fijo. Como, por otra parte, $\beta_\varphi(\mu_0, \sigma^2) = \alpha$, para todo σ^2, se tiene

$$\beta_\varphi(\mu, \sigma^2) < \alpha \qquad \forall\, \mu < \mu_0$$

y el test φ tiene nivel de significación α. También, a partir de la expresión de $\beta_\varphi(\mu, \sigma^2)$ se obtiene que el test φ es insesgado.

Análogamente, en el caso de testear H : $\mu \geq \mu_0$ contra K : $\mu < \mu_0$, el test del cociente de máxima verosimilitud vendrá dado por

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad T \leq t_{n-1,\alpha} \\ 0 & \text{si} \quad T > t_{n-1,\alpha} \ . \end{cases}$$

Para calcular la potencia de estos tests se pueden utilizar las tablas construídas por Owen [4].

Ejemplo 4. Supongamos nuevamente que tenemos una muestra aleatoria $X_1, \ldots, X_n$ de una distribución $N(\mu, \sigma^2)$ con μ y σ^2 desconocidos. Se desea testear H : $\sigma^2 \leq \sigma_0^2$ contra K : $\sigma^2 > \sigma_0^2$.

Se deduce haciendo un razonamiento análogo al ejemplo anterior que el test del cociente de máxima verosimilitud es de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } \sum_{i=1}^n (X_i - \overline{X})^2 \geq k_\alpha \\ 0 & \text{si } \sum_{i=1}^n (X_i - \overline{X})^2 < k_\alpha \ . \end{cases}$$

La constante k_α se debe elegir de manera que

$$\sup_{\sigma^2 \leq \sigma_0^2} P_{\sigma^2}\left(\sum_{i=1}^n (X_i - \overline{X})^2 \geq k_\alpha\right) = \alpha \ .$$

Determinemos k_α por el valor de σ^2 más desfavorable, o sea, σ_0^2. Luego, debemos elegir k_α tal que

$$P_{\sigma_0^2}\left(\sum_{i=1}^n (X_i - \overline{X})^2 \geq k_\alpha\right) = \alpha$$

o equivalentemente

$$P_{\sigma_0^2}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2} \geq \frac{k_\alpha}{\sigma_0^2}\right) = \alpha \ .$$

Como $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma_0^2$ tiene distribución χ^2_{n-1} cuando $\sigma^2 = \sigma_0^2$, se tiene que

$$k_\alpha = \sigma_0^2 \chi^2_{n-1,\alpha} \quad .$$

Para mostrar que el test tiene realmente nivel de significación α, bastará mostrar que la función de potencia es una función creciente y esto se deduce como sigue. Sea $D_n(k) = P(Y \geq k)$, donde Y es una variable aleatoria con distribución χ^2_n. Luego

$$\begin{aligned}
\beta_\varphi(\sigma^2) &= P_{\sigma^2}\left(\sum_{i=1}^n (X_i - \overline{X})^2 \geq \sigma_0^2 \chi^2_{n-1,\alpha}\right) \\
&= P_{\sigma^2}\left(\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \geq \frac{\sigma_0^2 \chi^2_{n-1,\alpha}}{\sigma^2}\right) \\
&= D_{n-1}\left(\frac{\sigma_0^2 \chi^2_{n-1,\alpha}}{\sigma^2}\right) \ ,
\end{aligned}$$

ya que cuando la varianza de cada X_i es σ^2 resulta que $\sum_{i=1}^n (X_i - \overline{X})^2/\sigma^2$ tiene distribución χ^2_{n-1}.

Como $D_n(k)$ es una función decreciente de k, $\beta_\varphi(\sigma^2)$ es una función creciente de σ^2.

Ejemplo 5. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$ con μ y σ^2 desconocidos y supongamos que se quiere testear H : $\sigma^2 = \sigma_0^2$ contra K : $\sigma^2 \neq \sigma_0^2$.

En este caso, el test del cociente de máxima verosimilitud es de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 \quad \text{si} \quad \dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2} \geq {k'}_\alpha \\ \\ 1 \quad \text{si} \quad \dfrac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_0^2} < {k''}_\alpha \\ \\ 0 \quad \text{en cualquier otro caso,} \end{cases}$$

Para que φ tenga nivel de significación α, se debe cumplir que

$$\begin{aligned}\beta_\varphi(\sigma_0^2) &= P_{\sigma_0^2}\left(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma_0^2} \geq {k'}_\alpha\right) \\ &\quad + P_{\sigma_0^2}\left(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma_0^2} < {k''}_\alpha\right) = \alpha\ .\end{aligned}$$

Luego, se debe tener que

$$k'_\alpha = \chi^2_{n-1,\beta} \quad \text{y} \quad {k''}_\alpha = \chi^2_{n-1,1-\gamma} \tag{6.28}$$

con $\beta + \gamma = \alpha$.

Si queremos que el test resulte insesgado, la derivada de la función de potencia debe ser cero en σ_0. Pero,

$$\begin{aligned}\beta_\varphi(\sigma^2) &= P_{\sigma^2}\left(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2} \geq \frac{{k'}_\alpha \sigma_0^2}{\sigma^2}\right) \\ &\quad + P_{\sigma^2}\left(\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2} < \frac{{k''}_\alpha \sigma_0^2}{\sigma^2}\right)\ ,\end{aligned}$$

con lo cual si llamamos Y a una variable con distribución χ^2_{n-1} obtenemos

$$\begin{aligned}\beta_\varphi(\sigma^2) &= P_{\sigma^2}\left(Y \geq \frac{{k'}_\alpha \sigma_0^2}{\sigma^2}\right) + P_{\sigma^2}\left(Y < \frac{{k''}_\alpha \sigma_0^2}{\sigma^2}\right) \\ &= 1 - P_{\sigma^2}\left(Y < \frac{{k'}_\alpha \sigma_0^2}{\sigma^2}\right) + P_{\sigma^2}\left(Y < \frac{{k''}_\alpha \sigma_0^2}{\sigma^2}\right)\ .\end{aligned}$$

Por lo tanto, si $f_Y(y)$ indica a la densidad de Y, la condición ${\beta'}_\varphi(\sigma_0^2) = 0$ es equivalente a

$$f_Y({k'}_\alpha)\, {k'}_\alpha = f_Y({k''}_\alpha)\, {k''}_\alpha$$

de donde se obtiene que ${k'}_\alpha$ y ${k''}_\alpha$ deberán ser elegidos de forma que

$$e^{-{k'}_\alpha/2}({k'}_\alpha)^{\frac{n-1}{2}} = e^{-{k''}_\alpha/2}({k''}_\alpha)^{\frac{n-1}{2}} \tag{6.29}$$

En la práctica se eligen $\gamma = \alpha/2$ y $\beta = \alpha/2$, aunque no satisfaga (6.29). Se puede mostrar que para $n \to \infty$ los β y γ que satisfacen (6.28) y hacen que se satisfaga (6.29) se aproximan a los valores elegidos. En realidad, la

aproximación es buena con tal que n no sea muy pequeño. Luego, el test que se usa viene dado por

$$\varphi(\mathbf{X}) = \begin{cases} 1 \quad \text{si} \begin{cases} \displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 \geq \sigma_0^2 \chi^2_{n-1,\frac{\alpha}{2}} \\ \\ \displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 \leq \sigma_0^2 \chi^2_{n-1,1-\frac{\alpha}{2}} \end{cases} \\ \\ 0 \quad \text{si} \quad \sigma_0^2 \chi^2_{n-1,1-\frac{\alpha}{2}} \leq \displaystyle\sum_{i=1}^{n}(X_i - \overline{X})^2 \leq \sigma_0^2 \chi^2_{n-1,\frac{\alpha}{2}} \end{cases}$$

Se puede mostrar que los tests obtenidos en los Ejemplos 1 a 5 son IUMP. Para estos resultados pueden consultarse el Capítulo 5 de Lehmann [3] o el Capítulo 5 de Ferguson [2].

6.7 Test con nivel de significación asintótico

La mayoría de los test de hipótesis, por ejemplo, los del cociente de verosimilitud, son de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si } T > k_\alpha \\ \gamma_\alpha & \text{si } T = k_\alpha \\ 0 & \text{si } T < k_\alpha \end{cases}$$

donde T es un estadístico basado en la muestra. Para encontrar k_α se requiere conocer la distribución de T para $\boldsymbol{\theta} \in \Theta_1$. Como en muchos casos esta distribución es muy compleja se puede reemplazar esta distribución por una asintótica. En este caso el test tendra un nivel de significación aproximado al deseado para muestras grandes. Esto motiva la siguiente definición.

Definición 1. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$ y supongamos que se quiere testear la hipótesis H : $\boldsymbol{\theta} \in \Theta_1$ contra K : $\boldsymbol{\theta} \in \Theta_2$. Se dirá que una sucesión de test $\varphi_n(X_1, \ldots, X_n)$ tiene nivel de significación asintótico α si

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\Theta_1} \beta_{\varphi_n}(\boldsymbol{\theta}) = \alpha$$

Es decir, que el nivel del test $\varphi_n(X_1, \ldots, X_n)$ se acerca a α cuando el tamaño de la muestra tiende a infinito.

Ejemplo 1. Supongamos que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución desconocida con media μ y varianza σ^2.

Supongamos que se quiere testear H : $\mu \leq \mu_0$ contra K : $\mu > \mu_0$. Llamemos

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \qquad \text{y} \qquad s^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} \, .$$

Ya hemos demostrado que

$$\sqrt{n}\frac{(\overline{X} - \mu_0)}{s}$$

converge en distribución a la $N(0,1)$ cuando la esperanza de las variables X_i es μ_0. Luego, si definimos

$$\varphi_n(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si} \quad \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{s} \geq z_\alpha \\ \\ 0 & \text{si} \quad \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{s} < z_\alpha \end{cases}$$

este test tiene nivel de significación asintótico α.

Del mismo modo, si se quiere testear H : $\mu = \mu_0$ contra K : $\mu \neq \mu_0$, un test de nivel de significación asintótico α será

$$\varphi_n(X_1, \ldots, X_n) = \begin{cases} 1 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X} - \mu_0|}{s} \geq z_\alpha \\ \\ 0 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X} - \mu_0|}{s} < z_\alpha \end{cases}$$

6.7.1 Distribución asintótica del test del cociente de máxima verosimilitud

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una distribución de densidad o probabilidad dada por $p(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \Theta$, donde Θ es un conjunto de $\mathbb{R}^p$ que contiene una esfera.

Supongamos que Θ_1 es un conjunto de dimensión menor que p, digamos de dimensión $p - j$, donde $1 \leq j \leq p$. Θ_1 puede venir expresado de varias formas diferentes. Por ejemplo, puede venir dado por j relaciones funcionales entre los parámetros $\theta_1, \ldots, \theta_p$, es decir,

$$\Theta_1 = \{\boldsymbol{\theta} \in \Theta : g_1(\boldsymbol{\theta}) = 0; \; g_2(\boldsymbol{\theta}) = 0, \ldots, g_j(\boldsymbol{\theta}) = 0\}$$

o bien, en forma paramétrica

$$\Theta_1 = \{\boldsymbol{\theta} = (\theta_1, ...\theta_p) : \theta_1 = h_1(\boldsymbol{\lambda}), \ldots, \theta_p = h_p(\boldsymbol{\lambda}),\ \boldsymbol{\lambda} \in \Lambda\},$$

donde $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{p-j})$ y $\Lambda \subset I\!R^{p-j}$ de dimensión $p - j$.

Supongamos que se está interesado en el siguiente problema de test de hipótesis:

$$\mathrm{H} : \boldsymbol{\theta} \in \Theta_1 \quad \text{contra} \quad \mathrm{K} : \boldsymbol{\theta} \in \Theta_2$$

con $\Theta = \Theta_1 \cup \Theta_2$. Luego, el test del cociente de máxima verosimilitud es de la forma

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad L^*(\mathbf{X}) \leq k_\alpha \\ 0 & \text{si} \quad L^*(\mathbf{X}) > k_\alpha \end{cases}$$

donde

$$L^*(\mathbf{X}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_1} p(\mathbf{X}, \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} p(\mathbf{X}, \boldsymbol{\theta})} \ .$$

Para determinar k_α es necesario conocer la distribución de $L^*(\mathbf{X})$ bajo H. Muchas veces esta es muy complicada y puede depender del valor particular $\boldsymbol{\theta} \in \Theta_1$ que se considere. Sin embargo, se puede mostrar que, bajo condiciones de regularidad muy generales en $p(\mathbf{x}, \boldsymbol{\theta})$, la distribución asintótica de $Z = -2 \ln L^*$ cuando $\boldsymbol{\theta} \in \Theta_1$ es χ^2_j. Luego un test de nivel de significación asintótico α está dado por

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad Z \geq \chi^2_{j,\alpha} \\ \\ 0 & \text{si} \quad Z < \chi^2_{j,\alpha} \end{cases}$$

Para ver la teoría asintótica del test del cociente de verosimilitud se puede ver Wald [5] y Chernoff [1]. Nosotros sólo daremos la distribución en el caso particular $\Theta \subset I\!R$ y $\mathrm{H} : \theta = \theta_0$ contra $\mathrm{K} : \theta \neq \theta_0$.

Teorema 1. *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución discreta o continua con densidad perteneciente a la familia $p(x, \theta)$ con $\theta \in \Theta$ y Θ un abierto en $I\!R$. Indiquemos por $p(\mathbf{x}, \theta)$ la densidad conjunta del vector $\mathbf{X} = (X_1, \ldots, X_n)$.*

Supongamos que se cumplen las siguientes condiciones (en lo que sigue suponemos que $\mathbf{X}$ es continuo, para el caso discreto habrá que reemplazar todos los signos $\int$ por $\sum$):

(A) *El conjunto $\mathcal{S} = \{x : p(x, \theta) > 0\}$ es independiente de θ.*

(B) *Para todo* x, $p(x,\theta)$ *tiene derivada tercera respecto de* θ *continua y tal que*

$$\Big|\frac{\partial^3 \ln p(x,\theta)}{\partial\theta^3}\Big| = \Big|\frac{\partial^2 \psi(x,\theta)}{\partial\theta^2}\Big| \leq K$$

para todo $x \in \mathcal{S}$ *y para todo* $\theta \in \Theta$, *donde*

$$\psi(x,\theta) = \frac{\partial \ln p(x,\theta)}{\partial\theta}\ .$$

(C) *Si* $h(\mathbf{X})$ *es un estadístico tal que* $E_\theta[|h(\mathbf{X})|] < \infty$ *para todo* $\theta \in \Theta$ *entonces se tiene*

$$\frac{\partial}{\partial\theta}\left[\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} h(\mathbf{x})p(\mathbf{x},\theta)d\mathbf{x}\right] = \int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} h(\mathbf{x})\frac{\partial p(\mathbf{x},\theta)}{\partial\theta}d\mathbf{x}$$

donde $d\mathbf{x} = (dx_1,\dots,dx_n)$.

(D)

$$0 < I_1(\theta) = E_\theta\left[\left(\frac{\partial \ln p(X_1,\theta)}{\partial\theta}\right)^2\right] < \infty\ .$$

Sea $\widehat{\theta}_n$ *un estimador de máxima verosimilitud de* θ *consistente, entonces si*

$$L^*(\mathbf{X}) = \frac{p(\mathbf{X},\theta_0)}{\sup_{\theta\in\Theta} p(\mathbf{X},\theta)} = \frac{p(\mathbf{X},\theta_0)}{p(\mathbf{X},\widehat{\theta}_n)}\ .$$

se tiene que $Z = -2\ln\left(L^*(\mathbf{X})\right)$ *tiene distribución asintótica* χ^2_1 *con lo cual el test*

$$\varphi(\mathbf{X}) = \begin{cases} 1 & si \quad Z \geq \chi^2_{1,\alpha} \\ \\ 0 & si \quad Z < \chi^2_{1,\alpha} \end{cases}$$

tiene nivel de significación asintótico α.

Demostración. Sea

$$\ell(\theta) = \ln p(\mathbf{X},\theta) = \sum_{i=1}^{n} \ln\left(p(X_i,\theta)\right)\ .$$

Indiquemos además por ℓ', ℓ'' y ℓ''' las derivadas hasta el orden tres respecto de θ de la función ℓ y por

$$\psi'(x,\theta) = \frac{\partial\psi(x,\theta)}{\partial\theta} \quad \text{y} \quad \psi''(x,\theta) = \frac{\partial^2\psi(x,\theta)}{\partial\theta^2}.$$

Luego, $\widehat{\theta}_n$ verifica

$$\ell'(\widehat{\theta}_n) = \sum_{i=1}^{n} \psi(X_i, \widehat{\theta}_n) = 0 \ .$$

Con lo cual, desarrollando en serie de Taylor alrededor de $\widehat{\theta}_n$ se obtiene:

$$\begin{aligned} \ell(\theta_0) - \ell(\widehat{\theta}_n) &= \ell'(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{1}{2}\ell''(\xi_n^1)(\theta_0 - \widehat{\theta}_n)^2 \\ &= \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^2 \left(\sum_{i=1}^{n} \psi'(X_i, \xi_n^1) \right) \\ &= \frac{1}{2} \left(n(\theta_0 - \widehat{\theta}_n)^2 \right) \frac{1}{n} \sum_{i=1}^{n} \psi'(X_i, \theta_0) + R_n \ , \end{aligned}$$

donde ξ_n^1 es un punto intermedio entre $\widehat{\theta}_n$ y θ_0 y

$$R_n = \frac{1}{2} \left(n(\theta_0 - \widehat{\theta}_n)^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} \psi'(X_i, \xi_n^1) - \frac{1}{n} \sum_{i=1}^{n} \psi'(X_i, \theta_0) \right) \ .$$

Aplicando el Teorema del valor medio se obtiene

$$R_n = \frac{1}{2} \left(n(\theta_0 - \widehat{\theta}_n)^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} \psi''(X_i, \xi_n^2)(\xi_n^1 - \theta_0) \right) \tag{6.30}$$

donde ξ_n^2 es un punto intermedio entre ξ_n^1 y θ_0. Observemos que por ser $\widehat{\theta}_n$ consistente, se obtiene entonces que $\xi_n^j \to \theta_0$ en probabilidad para $j = 1, 2$.

Reemplazando, obtenemos que

$$Z = 2 \left(\ell(\widehat{\theta}_n) - \ell(\theta_0) \right) = \left(n(\widehat{\theta}_n - \theta_0)^2 \right) A_n - R_n \tag{6.31}$$

donde $A_n = -\frac{1}{n} \sum_{i=1}^{n} \psi'(X_i, \theta_0)$.

Hemos visto en el Teorema 1 de 3.17 que cuando $\theta = \theta_0$

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to \mathrm{N}(0, \frac{1}{I_1(\theta_0)}) \quad \text{en distribución,}$$

con lo cual

$$I_1(\theta_0)\, n\, (\widehat{\theta}_n - \theta_0)^2 \to \chi_1^2 \quad \text{en distribución.} \tag{6.32}$$

Por otra parte, la ley de los grandes números implica que

$$\frac{1}{n} \sum_{i=1}^{n} \psi'(X_i, \theta_0) \to E(\psi'(X_1, \theta_0)) \quad \text{en probabilidad.} \tag{6.33}$$

Pero,

$$E_{\theta_0}(\psi'(X_1,\theta_0)) = -I_1(\theta_0)\ ,$$

luego, usando (6.32) y (6.33) se obtiene que

$$\left(n\,(\widehat{\theta}_n-\theta_0)^2\right)\,A_n \to \chi^2_1 \quad \text{en distribución.} \tag{6.34}$$

Por lo tanto, a partir de (6.31) y (6.34) deducimos que bastará probar que

$$R_n \to 0 \qquad \text{en probabilidad.} \tag{6.35}$$

Como $|\psi''(X_i,\theta)| \le K$ para todo θ, se tiene que

$$\left|\frac{1}{2n}\sum_{i=1}^{n}\psi''(X_i,\xi_n^2)(\xi_n^1-\theta_0)\right| \le \frac{K}{2}|(\xi_n^1-\theta_0)|$$

y luego como $\xi_n^1 \to \theta_0$ en probabilidad se deduce que:

$$\frac{1}{n}\sum_{i=1}^{n}\psi''(X_i,\xi_n^2)(\xi_n^1-\theta_0) \to 0 \quad \text{en probabilidad.} \tag{6.36}$$

Pero, (6.32) implica que $n(\widehat{\theta}_n-\theta_0)^2$ está acotado en probabilidad, luego (6.35)se obtiene de (6.30) y (6.36).

Ejemplo 1. Sea $X_1,\ldots,X_n$ una muestra de una distribución perteneciente a la familia $Bi(\theta,1)$, $0<\theta<1$, y supongamos que se quiere testear $\mathrm{H}:\theta=\theta_0$ contra $\mathrm{K}:\theta\neq\theta_0$. Luego el test del cociente de máxima verosimilitud es

$$L^* = \frac{p(\mathbf{x},\theta_0)}{\sup_{\theta\in\Theta}p(x,\theta)} = \frac{\theta_0^T(1-\theta_0)^{n-T}}{\overline{X}^T(1-\overline{X})^{n-T}}\ ,$$

donde $T=\sum_{i=1}^n X_i$. Luego,

$$Z = -2\ln L^* = 2T\ln\frac{\overline{X}}{\theta_0} + 2(n-T)\ln\frac{(1-\overline{X})}{1-\theta_0}$$

tiene una distribución asintótica χ^2_1 bajo H y un test de nivel asintótico estará dado por

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad Z > \chi^2_{1,\alpha} \\ \\ 0 & \text{si} \quad Z < \chi^2_{1,\alpha}\ . \end{cases}$$

6.8 Relación entre regiones de confianza y test

En esta sección se estudiará la relación que existe entre tests y regiones de confianza.

Supongamos que se tiene un vector aleatorio $\mathbf{X}$ con distribución perteneciente a la familia $F(\mathbf{x}, \boldsymbol{\theta})$ con $\boldsymbol{\theta} \in \Theta$ y supongamos que para cada $\boldsymbol{\theta}_0$ se tiene un test no aleatorizado de nivel α, $\varphi_{\boldsymbol{\theta}_0}$, para H : $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ contra K : $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

Se puede construir una región de confianza de nivel $(1 - \alpha)$ para $\boldsymbol{\theta}$ definiendo

$$S(\mathbf{X}) = \{\boldsymbol{\theta} : \varphi_{\boldsymbol{\theta}}(\mathbf{X}) = 0\}$$

Es decir, $S(\mathbf{X})$ es el conjunto de todos los $\boldsymbol{\theta} \in \Theta$ tales que la hipótesis de que el valor verdadero es $\boldsymbol{\theta}$, es aceptada cuando se observa $\mathbf{X}$.

Demostraremos que $S(\mathbf{X})$ así definida, es una región de confianza de nivel $1 - \alpha$ para $\boldsymbol{\theta}$

$$P_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in S(\mathbf{X})) = P_{\boldsymbol{\theta}}(\varphi_{\boldsymbol{\theta}}(\mathbf{X}) = 0) = 1 - P_{\boldsymbol{\theta}}(\varphi_{\boldsymbol{\theta}}(\mathbf{X}) = 1) = 1 - \alpha \ .$$

Recíprocamente, si se tiene una región de confianza $S(\mathbf{X})$ de nivel $1 - \alpha$ para $\boldsymbol{\theta}$, se puede construir un test de nivel α, $\varphi_{\boldsymbol{\theta}_0}$, para H : $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ contra K : $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

Definamos

$$\varphi_{\boldsymbol{\theta}_0}(\mathbf{X}) = \begin{cases} 1 & \text{si } \boldsymbol{\theta}_0 \notin S(\mathbf{X}) \\ 0 & \text{si } \boldsymbol{\theta}_0 \in S(\mathbf{X}) \ . \end{cases}$$

Mostraremos que este test tiene realmente nivel de significación α. Efectivamente,

$$P_{\boldsymbol{\theta}_0}(\varphi_{\boldsymbol{\theta}_0}(\mathbf{X}) = 1) = P_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0 \notin S(\mathbf{X})) = 1 - P_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0 \in S(\mathbf{X})) = 1 - (1 - \alpha) = \alpha \ .$$

Ejemplo 1. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma^2)$.

En el capítulo anterior hemos demostrado que un intervalo de confianza a nivel $(1 - \alpha)$ para μ viene dado por

$$S(\mathbf{X}) = [\overline{X} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \ , \ \overline{X} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$$

Construyamos el test correspondiente de nivel α para

$$\text{H} : \mu = \mu_0 \quad \text{contra} \quad \text{K: } \mu \neq \mu_0$$

$$\varphi_{\mu_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \mu_0 \notin S(\mathbf{X}) \\ 0 & \text{si} \quad \mu_0 \in S(\mathbf{X}) \end{cases}$$

pero $\mu_0 \in S(\mathbf{X})$ si y sólo si $|\mu_0 - \overline{X}| \leq t_{n-1,\frac{\alpha}{2}}(s/\sqrt{n})$, luego

$$\varphi_{\mu_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X} - \mu_0|}{s} > t_{n-1,\frac{\alpha}{2}} \\ \\ 0 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X} - \mu_0|}{s} \leq t_{n-1,\frac{\alpha}{2}} \ . \end{cases}$$

Por lo tanto, este test coincide con el obtenido en el Ejemplo 2 de 6.6, cuando obtuvimos el test del CMV para este problema. Recíprocamente, a partir de esta familia de tests si se usara el procedimiento indicado anteriormente para obtener intervalos de confianza, se llegará al intervalo inicial.

Ejemplo 2. Sea $X_1, \ldots, X_{n_1}$ una muestra aleatoria de una distribución $N(\mu_1, \sigma^2)$ y sea $Y_1, \ldots, Y_{n_2}$ una muestra aleatoria de una distribución $N(\mu_2, \sigma^2)$ independiente de la primera. Se ha visto en el Capítulo 5 que

$$T = \sqrt{\frac{n_1\, n_2}{n_1 + n_2}}\frac{(\overline{X} - \overline{Y} - (\mu_1 + \mu_2))}{s}$$

donde

$$s^2 = \frac{1}{n_2 + n_1 - 2}\left(\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(Y_i - \overline{Y})^2\right)$$

tiene distribución de Student con $n_1 + n_2 - 2$ grados de libertad y que un intervalo de confianza para $\mu_1 - \mu_2$ está dado por

$$S(\mathbf{X}) = \left[\overline{X} - \overline{Y} - t_{n_1+n_2-2,\frac{\alpha}{2}}\ s\sqrt{\frac{n_1 + n_2}{n_1\, n_2}}\ , \overline{X} - \overline{Y} + t_{(n_1+n_2-2),\frac{\alpha}{2}}\ s\sqrt{\frac{n_1 + n_2}{n_1\, n_2}}\right]$$

Luego, si se quiere testear H : $\mu_1 - \mu_2 = \lambda_0$ contra K : $\mu_1 - \mu_2 \neq \lambda_0$, con nivel de significación α, se puede obtener un test haciendo

$$\varphi_{\lambda_0}(\mathbf{X}) = \begin{cases} 1 & \text{si } \lambda_0 \notin S(\mathbf{X}) \\ 0 & \text{si } \lambda_0 \in S(\mathbf{X}) \end{cases}$$

pero $\lambda_0 \in S(\mathbf{X})$ si y sólo si

$$\sqrt{\frac{n_1\, n_2}{n_1 + n_2}}\frac{|\overline{X} - \overline{Y} - \lambda_0|}{s} \leq t_{n_1+n_2-2,\frac{\alpha}{2}}.$$

Por lo tanto,

$$\varphi_{\lambda_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \sqrt{\dfrac{n_1\, n_2}{n_1+n_2}}\dfrac{|\overline{X}-\overline{Y}-\lambda_0|}{s} \geq t_{n_1+n_2-2,\frac{\alpha}{2}} \\ 0 & \text{si} \quad \sqrt{\dfrac{n_1\, n_2}{n_1+n_2}}\dfrac{|\overline{X}-\overline{Y}-\lambda_0|}{s} < t_{(n_1+n_2-2),\frac{\alpha}{2}} \end{cases}.$$

Hasta aquí hemos estudiado la relación entre regiones de confianza de nivel $1-\alpha$ para $\boldsymbol{\theta}$ y test de hipótesis para las hipótesis $\mathrm{H} : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ contra $\mathrm{K} : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Esta situación se puede generalizar al caso de

$$\mathrm{H} : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{contra} \quad \mathrm{K} : K(\boldsymbol{\theta}_0)$$

donde $K(\boldsymbol{\theta}_0)$ indica una alternativa cualquiera que no contiene a $\boldsymbol{\theta}_0$ si para cada $\boldsymbol{\theta}_0 \in \Theta$ se tiene un test de nivel α, $\varphi_{\boldsymbol{\theta}_0}$, resultará que

$$S(\mathbf{X}) = \{\boldsymbol{\theta} \in \Theta : \varphi_{\boldsymbol{\theta}}(\mathbf{X}) = 0\}$$

será una región con nivel de confianza $1-\alpha$. De la misma forma que antes $S(\mathbf{X})$ será el conjunto de todos los $\boldsymbol{\theta} \in \Theta$ tales que la hipótesis de que $\boldsymbol{\theta}$ es el verdadero valor es aceptada cuando se observa $\mathbf{X}$.

6.9 Cotas de confianza óptimas

Se verá ahora cómo la existencia de tests uniformemente más potentes para hipótesis unilaterales permite la construcción de intervalos de confianza unilaterales óptimas en el sentido definido en la sección 5.9.

Hemos demostrado en 6.4 que para *familias de cociente de verosimilitud monótono* existen tests UMP para las hipótesis:

$$\begin{aligned} &\mathrm{H}_1 : \theta = \theta_0 \quad \text{contra} \quad \mathrm{K}_1 : \theta > \theta_0 \\ &\mathrm{H}_2 : \theta = \theta_0 \quad \text{contra} \quad \mathrm{K}_2 : \theta < \theta_0 \end{aligned}$$

En estos casos vale el siguiente teorema

Teorema 1. *Sea φ_{θ_0} el test no aleatorizado (si existe) UMP para H_1 contra K_1, de nivel α. Dada $X_1, \ldots, X_n$ y siendo*

$$S(\mathbf{X}) = \{\theta \in \Theta : \varphi_{\theta}(\mathbf{X}) = 0\}$$

i) *$S(\mathbf{X})$ es una región de confianza de nivel $1-\alpha$ para θ.*

ii) *Si $\varphi^*_{\theta_0}$ es cualquier otro test no aleatorizado de nivel α para esas hipótesis y*

$$S^*(\mathbf{X}) = \{\theta \in \Theta : \varphi^*_\theta(\mathbf{X}) = 0\}$$

entonces $P_\theta\{\theta_0 \in S(\mathbf{X})\} \leq P_\theta\{\theta_0 \in S^*(\mathbf{X})\}$ para todo $\theta > \theta_0$.

Demostración. i) Por la definición de $S(\mathbf{X})$ sabemos que $\theta \in S(\mathbf{X})$ si y sólo si $\varphi_\theta(\mathbf{X}) = 0$, luego

$$P_\theta\{\theta \in S(\mathbf{X})\} = P_\theta\{\varphi_\theta(\mathbf{X}) = 0\} = 1 - \alpha$$

por ser φ_θ de nivel α.
ii) Igual que en i) $S^*(\mathbf{X})$ será una región de confianza de nivel $1-\alpha$. Por ser $\varphi_{\theta_0}(\mathbf{X})$ el test UMP para H_1 contra K_1 resulta que

$$\beta_{\varphi_{\theta_0}}(\theta) \geq \beta_{\varphi^*_{\theta_0}}(\theta) \qquad \forall\, \theta > \theta_0$$

o sea,

$$P_\theta\{\varphi_{\theta_0}(X) = 1\} \geq P_\theta\{\varphi^*_{\theta_0}(X) = 1\} \qquad \forall\, \theta > \theta_0\ .$$

Por lo tanto,

$$P_\theta\{\varphi_{\theta_0}(\mathbf{X}) = 0\} \leq P_\theta\{\varphi^*_{\theta_0}(\mathbf{X}) = 0\} \quad \forall\, \theta > \theta_0\ .$$

pero como $\theta_0 \in S(\mathbf{X})$ si y sólo si $\varphi_{\theta_0}(\mathbf{X}) = 0$ y $\theta_0 \in S^*(\mathbf{X})$ si y sólo si $\varphi^*_{\theta_0}(\mathbf{X}) = 0$, resulta

$$P_\theta\{\theta_0 \in S(\mathbf{X})\} \leq P_\theta\{\theta_0 \in S^*(\mathbf{X})\} \quad \forall\, \theta > \theta_0\ .$$

Un teorema similar puede demostrarse para H_2 contra K_2.

Veamos cómo son las regiones $S(\mathbf{X})$ en el caso del Teorema 1.

Teorema 2. *Sea $\mathbf{X}$ con distribución perteneciente a una familia $F(\mathbf{x}, \theta)$ de cociente de verosimilitud monótono en $T = r(\mathbf{X})$. Supongamos que la función de distribución $F_T(t, \theta)$ de T es continua para todo θ . Sea, para cada $\theta_0 \in \Theta$, $\varphi_{\theta_0}(\mathbf{X})$ el test UMP para $\mathrm{H}_1 : \theta = \theta_0$ contra $\mathrm{K}_1 : \theta > \theta_0$, o sea:*

$$\varphi_{\theta_0}(\mathbf{X}) = \begin{cases} 1 & si \quad T > k_\alpha(\theta_0) \\ 0 & si \quad T \leq k_\alpha(\theta_0) \end{cases}$$

Si además $F_T(t,\theta)$ es continua en θ para cada t fijo, la región de confianza

$$S(\mathbf{X}) = \{\theta \in \Theta : \varphi_\theta(\mathbf{X}) = 0\} = \{\theta \in \Theta : T = r(\mathbf{X}) \leq k_\alpha(\theta)\}$$

es el intervalo $I = [\underline{\theta}(\mathbf{X}), +\infty)$, donde

$$\underline{\theta}(\mathbf{X}) = \inf\{\theta \in \Theta : T \leq k_\alpha(\theta)\} .$$

DEMOSTRACIÓN. Ya hemos demostrado que si se tiene una familia de cociente de verosimilitud monótono en $T = r(\mathbf{X})$, el test UMP para $\mathrm{H}_1 : \theta = \theta_0$ contra K_1: $\theta > \theta_0$ es de la forma

$$\varphi_{\theta_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad T > k_\alpha(\theta_0) \\ \gamma_\alpha(\theta_0) & \text{si} \quad T = k_\alpha(\theta_0) \\ 0 & \text{si} \quad T < k_\alpha(\theta_0) \end{cases}$$

con $k_\alpha(\theta_0)$ y $\gamma_\alpha(\theta_0)$ tales que

$$E_{\theta_0}(\varphi_{\theta_0}(\mathbf{X})) = \alpha .$$

Como T tiene distribución continua, no es necesario aleatorizar y por lo tanto, el test UMP resulta

$$\varphi_{\theta_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad T > k_\alpha(\theta_0) \\ 0 & \text{si} \quad T \leq k_\alpha(\theta_0) . \end{cases}$$

Mostraremos que

(a) $k_\alpha(\theta)$ es una función no decreciente de θ.

(b) $k_\alpha(\theta)$ es una función continua a derecha.

(a) Sabemos que por ser φ_{θ_0} el test UMP de nivel α para H_1 contra K_1, la función de potencia de φ_{θ_0} es mayor o igual que el nivel para todo $\theta > \theta_0$. Luego, dado cualquier $\theta_1 > \theta_0$ se cumple

$$\begin{aligned} \alpha &= E_{\theta_0}(\varphi_{\theta_0}(\mathbf{X})) = P_{\theta_0}(T \geq k_\alpha(\theta_0)) \\ &\leq E_{\theta_1}(\varphi_{\theta_0}(\mathbf{X})) = P_{\theta_1}(T \geq k_\alpha(\theta_0)) . \end{aligned}$$

Como además

$$\alpha = E_{\theta_1}(\varphi_{\theta_1}(\mathbf{X})) = P_{\theta_1}(T \geq k_\alpha(\theta_1)) ,$$

tendremos

$$\alpha = P_{\theta_1}(T \geq k_\alpha(\theta_1)) \leq P_{\theta_1}(T \geq k_\alpha(\theta_0)) ,$$

y por lo tanto, es posible tomar $k_\alpha(\theta_1)$ tal que

$$k_\alpha(\theta_1) \geq k_\alpha(\theta_0) \ .$$

Con lo cual, $k_\alpha(\theta)$ es una función no decreciente de θ.
(b) Sea θ_n una sucesión decreciente que converge a θ, luego como $k_\alpha(.)$ es no decreciente se tiene

$$k_\alpha(\theta_n) \geq k_\alpha(\theta) \tag{6.37}$$

Sea $k = \lim_{n\to\infty} k_\alpha(\theta_n) = \inf_{n\geq 1} k_\alpha(\theta_n)$. Por (6.37) $k \geq k_\alpha(\theta)$, bastará mostrar que $k \leq k_\alpha(\theta)$.

Como $k \leq k_\alpha(\theta_n)$ se cumple

$$P_{\theta_n}(T \leq k) \leq P_{\theta_n}\left(T \leq k_\alpha(\theta_n)\right) = \alpha \ . \tag{6.38}$$

Pero además, como $F_T(k,\theta)$ es continua en θ se tiene

$$P_\theta(T \leq k) = \lim_{n\to\infty} P_{\theta_n}(T \leq k) \ . \tag{6.39}$$

Por lo tanto, (6.38) y (6.39) implican que

$$P_\theta(T \leq k) \leq \alpha = P_\theta\left(T \leq k_\alpha(\theta)\right)$$

luego, es posible tomar $k_\alpha(\theta)$ tal que $k \leq k_\alpha(\theta)$. Con lo cual, $k = k_\alpha(\theta)$ y $k_\alpha(\theta)$ es continua a derecha.

Veamos ahora que $\theta \in S(\mathbf{X})$ si y sólo si $\theta \geq \underline{\theta}(\mathbf{X})$.

Si $\theta \in S(\mathbf{X})$ entonces $T \leq k_\alpha(\theta)$ de donde $\theta \in \{\theta \in \Theta : T \leq k_\alpha(\theta)\}$ y $\theta \geq \underline{\theta}(\mathbf{X})$ que es el ínfimo de este conjunto.

Si $\theta > \underline{\theta}(\mathbf{X})$ entonces existe $\theta' \in \Theta$ tal que $T \leq k_\alpha(\theta')$ con $\underline{\theta}(X) < \theta' \leq \theta$. Pero como $k_\alpha(.)$ es creciente, resulta $T \leq k_\alpha(\theta)$ y por lo tanto, $\theta \in S(\mathbf{X})$.

Si $\theta = \underline{\theta}(\mathbf{X})$, existe una sucesión θ_n decreciente que converge a θ y tal que $\theta_n \in \{\theta \in \Theta : T \leq k_\alpha(\theta)\}$. Por lo tanto, $T \leq k_\alpha(\theta_n)$. Luego, la continuidad a derecha de $k_\alpha(\theta)$ implica que $T \leq k_\alpha(\theta)$ y por lo tanto, $\theta \in S(\mathbf{X})$.

Teorema 3. *Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una distribución perteneciente a una familia $F(\mathbf{x},\theta)$ de cociente de verosimilitud monótono en $T = r(\mathbf{X})$ y sea, para cada $\theta_0 \in \Theta$, $\varphi_{\theta_0}(\mathbf{X})$ el test UMP para* $\mathrm{H}_1 : \theta = \theta_0$ *contra* $\mathrm{K}_1 : \theta > \theta_0$, *o sea:*

$$\varphi_{\theta_0}(\mathbf{X}) = \begin{cases} 1 & si \quad T > k_\alpha(\theta_0) \\ 0 & si \quad T \leq k_\alpha(\theta_0) \end{cases}$$

suponiendo que la distribución, $F_T(t,\theta)$, de $T(\mathbf{X})$ es continua para todo θ. Supongamos además $F_T(t,\theta)$ es continua en θ para cada t fijo.

En estas condiciones

$$\underline{\theta}(\mathbf{X}) = \inf\{\theta \in \Theta : T \leq k_\alpha(\theta)\}$$

es una cota inferior para θ uniformemente óptima.

DEMOSTRACIÓN. De acuerdo a la definición de cota inferior con nivel de confianza $1-\alpha$ uniformemente óptima deberá demostrarse que

i) $P_\theta(\theta \geq \underline{\theta}(X)) = 1-\alpha$ para todo θ,

ii) si $\underline{\theta}^*$ es otra cota inferior a nivel α para θ

$$E_\theta(D(\theta,\underline{\theta})) \leq E_\theta(D(\theta,\underline{\theta}^*)) \quad \text{para todo } \theta \tag{6.40}$$

donde D es una medida de la subevaluación de $\underline{\theta}$ respecto de θ, definida por

$$D(\theta,\underline{\theta}) = \begin{cases} \theta - \underline{\theta} & \text{si} \quad \theta > \underline{\theta} \\ 0 & \text{si} \quad \theta \leq \underline{\theta}\,. \end{cases}$$

(i) se deduce del Teorema 1, ya que

$$S(\mathbf{X}) = \{\theta : \theta \geq \underline{\theta}(\mathbf{X})\}$$

es un intervalo de nivel de confianza $1-\alpha$.

(ii) Demostraremos que dada cualquier otra cota $\underline{\theta}^*$ a nivel $1-\alpha$

$$P_\theta\{\theta' \geq \underline{\theta}\} \leq P_\theta\{\theta' \geq \underline{\theta}^*\} \quad \text{para todo } \theta' \leq \theta\,. \tag{6.41}$$

Dado $\theta' \in \Theta$ definamos

$$\varphi^*_{\theta'}(X) = \begin{cases} 1 & \text{si } \theta' \leq \underline{\theta}^* \\ 0 & \text{si } \theta' > \underline{\theta}^*\,. \end{cases}$$

Luego $\varphi^*_{\theta'}(X)$ es un test de nivel α para $\mathrm{H} : \theta = \theta'$ contra $\mathrm{K} : \theta > \theta'$. Como $\varphi_{\theta'}(X)$ es el UMP para estas hipótesis, por Teorema 1, ii) sabemos que

$$P_\theta\{\theta' \geq \underline{\theta}(X)\} \leq P_\theta\{\theta' \geq \underline{\theta}^*(X)\} \quad \text{para todo } \theta \geq \theta'$$

y como esto se puede hacer para todo $\theta' \in \Theta$ resulta (6.41).

Se podría demostrar que si $\underline{\theta}$ cumple (6.41) entonces $\underline{\theta}$ cumple (6.40). Intuitivamente esto parece razonable, puesto que una cota inferior $\underline{\theta}$ de θ que cumple (6.41) es, en algún sentido, la "mayor" cota inferior y, en este caso, el defecto que presenta $\underline{\theta}$ respecto de θ debería ser lo más pequeño posible. Sin embargo la demostración de esta implicación está fuera de los alcances de este curso. (Para la demostración ver Lehmann [3], ejercicio 21, página 117.)

Ejemplo 1. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $U[0, \theta]$. Sabemos que el test UMP para H : $\theta = \theta_0$ contra K : $\theta > \theta_0$ es de la forma

$$\varphi_{\theta_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \max\limits_{1 \le i \le n} X_i > \theta_0 \sqrt[n]{1-\alpha} \\ 0 & \text{si} \quad \max\limits_{1 \le i \le n} X_i \le \theta_0 \sqrt[n]{1-\alpha} \end{cases}$$

En este caso, si $T = \max_{1 \le i \le n} X_i$ y $k_\alpha(\theta) = \theta \sqrt[n]{1-\alpha}$

$$S(\mathbf{X}) = \{\theta \in I\!R : \varphi_\theta(X) = 0\} = \{\theta \in I\!R : T \le k_\alpha(\theta)\}$$

resulta igual a

$$\begin{aligned} S(\mathbf{X}) &= \{\theta \in I\!R : \max_{1 \le i \le n} X_i \le \theta \sqrt[n]{1-\alpha}\} = \\ &= \{\theta \in I\!R : \theta \ge \frac{\max\limits_{1 \le i \le n} X_i}{\sqrt[n]{1-\alpha}}\} \end{aligned}$$

y $\underline{\theta}$ será

$$\underline{\theta}(\mathbf{X}) = \frac{\max\limits_{1 \le i \le n} X_i}{\sqrt[n]{1-\alpha}}$$

puesto que este es el menor valor que puede tomar θ que pertenece a $S(\mathbf{X})$.

Resulta entonces que

$$I = [\underline{\theta}(\mathbf{X}), +\infty) = [\frac{\max\limits_{1 \le i \le n} X_i}{\sqrt[n]{1-\alpha}}, +\infty)$$

es un intervalo de confianza unilateral para θ de nivel $1 - \alpha$ y que $\underline{\theta}$ es la mejor cota inferior para θ.

Ejemplo 2. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $N(\mu, \sigma_0^2)$ con σ_0^2 conocido. Sabemos que el test UMP para H : $\mu = \mu_0$ contra

K : $\mu > \mu_0$, es de la forma

$$\varphi_{\mu_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{\sigma_0} > z_\alpha \\ \\ 0 & \text{si} \quad \sqrt{n}\dfrac{(\overline{X} - \mu_0)}{\sigma_0} \leq z_\alpha \end{cases}$$

Procediendo en forma similar a la del Ejemplo 1, resulta

$$S(X) = \{\mu \in I\!R : \mu \geq \overline{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}}\} .$$

Luego,

$$\underline{\mu}(\mathbf{X}) = \overline{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}}$$

es la mejor cota inferior para μ y

$$I = [\underline{\mu}(X), +\infty) = [\overline{X} - z_\alpha \frac{\sigma_0}{\sqrt{n}}, +\infty)$$

es un intervalo unilateral de nivel $1 - \alpha$ para μ.

6.10 Relación entre intervalos de confianza con nivel asintótico $1-\alpha$ y test con nivel de significación asintótico α

Supongamos que $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución perteneciente a la familia $F(x, \boldsymbol{\theta})$ y que para cada $\boldsymbol{\theta}_0$ se tenga una sucesión de test $\varphi_{n\boldsymbol{\theta}_0}(X_1, \ldots, X_n)$ con nivel de significación asintótico $1 - \alpha$ para H : $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ contra K : $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Luego, puede construirse una sucesión de intervalos de confianza con nivel asintótico $1 - \alpha$ definiendo

$$S_n(X_1, \ldots, X_n) = \{\boldsymbol{\theta} : \varphi_{n\boldsymbol{\theta}}(\mathbf{X}) = 0\} .$$

Recíprocamente, dada una sucesión de intervalos de confianza $S_n(X_1, \ldots, X_n)$ de nivel asintótico $1 - \alpha$, si definimos

$$\varphi_{n\boldsymbol{\theta}_0}(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \boldsymbol{\theta}_0 \notin S(X_1, \ldots, X_n) \\ 0 & \text{si} \quad \boldsymbol{\theta}_0 \in S(X_1, \ldots, X_n) \end{cases}$$

se tiene que $\varphi_n{}_{\boldsymbol{\theta}_0}$ es una sucesión de test con nivel de significación asintótico α para H : $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ contra K : $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. (Se deja como ejercicio la demostración de estos enunciados.)

Ejemplo 1. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una distribución $Bi(\theta, 1)$. Ya se ha visto que

$$\sqrt{n}\frac{(\overline{X} - \theta_0)}{\sqrt{\theta_0(1-\theta_0)}}$$

converge en distribución a la $N(0,1)$ cuando $\theta = \theta_0$.

Un intervalo de confianza para θ, con nivel asintótico $1-\alpha$ viene dado por

$$S_n(\mathbf{X}) = \{\theta : \sqrt{n}\frac{|\overline{X} - \theta|}{\sqrt{\theta(1-\theta)}} < z_{\frac{\alpha}{2}}\}$$

Luego, un test de significación asintótico α para H : $\theta = \theta_0$ contra K : $\theta \neq \theta_0$, viene dado por

$$\varphi(\mathbf{X}) = \begin{cases} 1 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X} - \theta_0|}{\sqrt{\theta_0(1-\theta_0)}} \geq z_{\frac{\alpha}{2}} \\ \\ 0 & \text{si} \quad \sqrt{n}\dfrac{|\overline{X} - \theta_0|}{\sqrt{\theta_0(1-\theta_0)}} < z_{\frac{\alpha}{2}} \ . \end{cases}$$

REFERENCIAS

1. Chernoff, H. (1954). On the distribution of the likelihood ratio. Ann. Math. Statist. 25: 573-578.

2. Ferguson, T.S. (1967). Mathematical Statistics. A Decision Theoretic Approach. Academic Press.

3. Lehmann, E.L. (1994). Testing Statistical Hypothesis. Chapman and Hall.

4. Owen, D.B. (1965). The power of Student's t test. J. Amer. Statist. Assoc. 60: 320-333.

5. Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. 54: 426-483.

Chapter 7

Estimación Robusta

7.1 El problema de la robustez para el modelo de posición

Sea el modelo de posición y escala

$$x_i = \mu + \sigma u_i, 1 \leq i \leq n, \tag{7.1}$$

donde μ y σ son parámetros de posición y escala respectivamente, $u_1, ..., u_n$ son variables i.i.d. con distribución F. En este caso, $x_1, ..., x_n$ resulta una muestra aleatoria de $F_{\mu\sigma}$, donde $F_{\mu\sigma}(x) = F((x-\mu)/\sigma)$ Por ejemplo las x_i pueden ser distintas mediciones de una misma magnitud física μ medida con un error σu_i .

Si $F = \Phi$, la función de una distribución N(0,1), entonces las x_i tienen distribución N(μ, σ^2). Por lo tanto, un estimador óptimo de μ es $\bar{x} = \sum_{i=1}^{n} x_i/n$. Efectivamente este estimador es IMVU y minimax. Es importante señalar que para que $\bar{x}$ tenga estas propiedades, la distribución de los u_i debe ser exactamente N(0,1). Sin embargo, en la mayoría de las aplicaciones prácticas a lo sumo se puede asegurar los errores de medición tienen distribución *aproximadamente normal*. Por lo tanto, cabe preguntarse cual será el comportamiento de estimador $\bar{x}$ en este caso.

Una forma de determinar distribuciones aproximadamente normales es considerar entornos de contaminación de la función de distribución

Φ de la N(0,1). Un entorno de contaminación de tamaño ϵ de la distribución Φ se define por

$$\mathcal{V}_\epsilon = \{F : F = (1-\epsilon)\Phi + \epsilon H \text{ con } H \text{ arbitraria}\}. \tag{7.2}$$

La distribución $F = (1-\epsilon)\Phi + \epsilon H$ corresponde a que las observaciones con probabilidad $1-\epsilon$ provienen de la distribución Φ y con probabilidad ϵ de la distribución H.

En efecto supongamos que se tienen tres variables aleatoria independientes : Z con distribución Φ, V con distribución H, y W con distribución Bi$(1,\epsilon)$. Definamos entonces la variable aleatoria U de la siguiente manera

$$U = \begin{cases} Z & \text{si} \quad W = 0 \\ V & \text{si} \quad W = 1 \end{cases}.$$

Luego

$$\begin{aligned} F_U(u) &= P(U \le u) = P(U \le u,\ W = 0) + P(U \le u,\ W = 1) \\ &= P(U \le u |\ W = 0)P(W = 0) + P(U \le u |\ W = 1)P(W = 1) \\ &= (1-\epsilon)\Phi(u) + \epsilon H(u). \end{aligned}$$

Con lo cual, si ϵ es pequeño (por ejemplo .05 o .10) esto significará que la gran mayoría de las observaciones se obtendrán a partir de la distribución Φ, es decir serán normales. Por lo tanto, podemos afirmar que si ϵ es pequeño y $F \in \mathcal{V}_\epsilon$, entonces F está cerca de Φ. Supongamos que tenemos una muestra aleatoria $x_1, ..., x_n$ de $F \in V_\epsilon$. Por lo tanto una proporción $(1-\epsilon)$ de las observaciones estarán dadas por (7.1) con u_i proveniente de una distribución Φ, y una proporción ϵ tendrán el correspondiente u_i proveniente de la distribución H. Estas últimas observaciones serán denominadas puntos atípicos o *outliers,* y pueden ser debidas a realizaciones del experimento en circunstancias anormales u otros factores de error como, por ejemplo, una equivocación en la transcripción del dato.

Lo que vamos a mostrar a continuación es que aunque ϵ sea pequeño el comportamiento del estimador $\bar{x}$ puede ser muy ineficiente para distribuciones $F \in \mathcal{V}_\epsilon$.

Primero mostraremos que si

$$F = (1-\epsilon)\Phi + \epsilon H, \tag{7.3}$$

entonces

$$E_F(u) = (1-\epsilon)E_\Phi(u) + \epsilon E_H(u). \tag{7.4}$$

Además, si $E_H(u) = 0$, se tiene

$$\text{var}_F(u) = (1-\epsilon)\text{var}_\Phi(u) + \epsilon \text{var}_H(u). \tag{7.5}$$

Para mostrar (7.4) supongamos que la H tiene densidad h , y sea φ la densidad correspondiente a Φ. Luego la densidad de F es

$$f = (1-\epsilon)\varphi + \epsilon h,$$

y luego

$$E_F(u) = \int_{-\infty}^{\infty} uf(u)du = (1-\epsilon)\int_{-\infty}^{\infty} u\varphi(u)du + \epsilon\int_{-\infty}^{\infty} uh(u)du = (1-\epsilon)E_\Phi(u) + \epsilon E_H(u).$$

Para mostrar (7.5), observemos que

$$\begin{aligned}
\text{var}_F(u) &= \int_{-\infty}^{\infty} u^2 f(u)du \\
&= (1-\epsilon)\int_{-\infty}^{\infty} u^2\varphi(u)du + \epsilon\int_{-\infty}^{\infty} u^2 h(u)du = \\
&= (1-\epsilon) + \epsilon\ \text{var}_H(u).
\end{aligned}$$

Consideremos ahora al estimador $\hat{\mu} = \bar{x}$, donde la muestra $x_1, ..., x_n$ son generadas por (7.1) donde las u_i son independientes con distribución dada por (7.3) con $E_H(u) = 0$

Luego

$$\text{var}_F(\bar{x}) = \frac{\sigma^2\ \text{var}_F(u)}{n} = \frac{\sigma^2((1-\epsilon) + \epsilon\ \text{var}_H(u))}{n}.$$

Luego, si $\epsilon = 0$, entonces $\text{var}(\bar{x}) = \sigma^2/n$. En cambio una contaminación de tamaño ϵ puede producir un aumento de la varianza ilimitado, ya que $\text{var}_H(u)$ puede ser ilimitada, inclusive infinita.

Esta extrema sensibilidad de $\bar{x}$ a una contaminación con una proporción pequeña de outliers también puede verse de la siguiente forma. Supongamos que se tiene una muestra $x_1, ..., x_n$ y se agrega una observación $x_{n+1.}$ Si esta observación es un outlier, su influencia en $\bar{x}$ puede

ser ilimitada. En efecto sean $\bar{x}_n$ y $\bar{x}_{n+1}$ el promedio basado en n y $n+1$ observaciones respectivamente. Luego se tiene

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1} = \bar{x}_n + \frac{1}{n+1}(x_{n+1} - \bar{x}_n),$$

y por lo tanto e $\bar{x}_{n+1}$ puede tomar valores tan altos (o tan bajos) como se quiera con tal de tomar x_{n+1} suficientemente lejos de $\bar{x}_n$.

Supongamos que tenemos el modelo de posición dado por (7.1) donde la distribución F de los u_i es simétrica respecto de 0. Como en este caso μ es también la mediana de las observaciones, un estimador alternativo será $\tilde{\mu}$ =mediana$(x_1, ..., x_n)$. Ordenemos los datos $x_1, ..., x_n$ de menor a mayor obteniendo los valores $x_{(1)} \leq ... \leq x_{(n)}$. Luego la mediana estará dada por

$$\tilde{\mu} = \begin{cases} x_{(m+1)} & \text{si} \quad n = 2m+1 \\ x_{(m)} + x_{(m+1)} & si \quad n = 2m \end{cases} .$$

Veamos que este estimador es mucho más resistente a outliers que la media. En efecto, para que la mediana tome un valor ilimitado no es suficiente agregar un outlier, sino se requiere por lo menos $n/2$ outliers.

Un estimador como la mediana que es poco sensible a outliers se denomina **robusto**

La distribución de $\tilde{\mu}$ para muestras finitas es muy complicada aún en el caso de muestras normales. Sin embargo, podremos derivar su distribución asintótica. Para ello necesitamos una version del Teorema Central del Límite para arreglos triangulares que enunciaremos sin demostración.

Teorema Central del Límite. Sean para cada n natural, $v_{n1}, ...v_{nn}$, v variables aleatoria independientes igualmente disribuidas. Supongamos que existan constantes $M > 0$ y $m > 0$, tales que $|v_{ni}| \leq M$ y $\lim_{n\to\infty} \text{var}(v_{ni}) \geq m$. Luego se tiene que

$$\frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{(v_{ni} - E(v_{ni}))}{\text{var}(v_{ni})^{1/2}} \xrightarrow{D} \text{N}(0, 1).$$

El siguiente Teorema establece la distribución asintótica de la mediana.

Teorema 1. Sea $x_1, ..., x_n$ una muestra aleatoria de una distribución F con una única mediana μ y con una densidad f tal que $f(\mu) > 0$. Entonces si $\tilde{\mu}_n$ es la mediana de la muestra, se tiene que

$$n^{1/2}(\tilde{\mu}_n - \mu) \xrightarrow{D} N\left(0, \frac{1}{4f^2(\mu)}\right).$$

Demostración: Para facilitar la demostración consideraremos solo el caso que $n = 2m + 1$. Tenemos que demostrar

$$\lim_{n\to\infty} P(n^{1/2}(\tilde{\mu}_n - \mu) \leq y) = \Phi(2f(\mu)y), \tag{7.6}$$

donde Φ es la función de distribución correspondiente a N(0,1)

Es inmediato que

$$P(n^{1/2}(\tilde{\mu}_n - \mu) \leq y) = P\left(\tilde{\mu}_n \leq \mu + \frac{y}{n^{1/2}}\right). \tag{7.7}$$

Sea

$$v_{ni} = \begin{cases} 1 & \text{si} \quad x_i \leq \mu + \dfrac{y}{n^{1/2}} \\ 0 & \text{si} \quad x_i > \mu + \dfrac{y}{n^{1/2}} \end{cases}, \ 1 \leq i \leq n. \tag{7.8}$$

Como v_{ni} tiene distribución Bi($F(\mu + yn^{-1/2}, 1)$ se tiene

$$E(v_{ni}) = \nu_n = F(\mu + \frac{y}{n^{1/2}}),$$

y

$$\text{var}(v_{ni}) = \nu_n(1 - \nu_n).$$

De acuerdo a la definición de mediana se tiene que

$$P\left(\tilde{\mu}_n \leq \mu + \frac{y}{n^{1/2}}\right) = P\left(\sum_{i=1}^{n} v_{ni} \geq \frac{n}{2}\right)$$

$$= P\left(\frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{(v_{ni} - \nu_n)}{(\nu_n(1 - \nu_n))^{1/2}} \geq \frac{(n/2 - n\nu_n)}{(n\nu_n(1 - \nu_n))^{1/2}}\right). \tag{7.9}$$

Como $|v_{ni}| \leq 1$, y $\lim_{n\to\infty}\text{var}(v_{ni}) = 1/4$. se cumplen las hipótesis del Teorema Central del Límite. Luego

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{(v_{ni}-\nu_n)}{(\nu_n(1-\nu_n))^{1/2}} \xrightarrow{D} N(0,1). \tag{7.10}$$

Usando el hecho de que $F(\mu) = 1/2$, y el Teorema del Valor Medio tenemos

$$\frac{(n/2-n\nu_n)}{n^{1/2}} = n^{1/2}\left(F(\mu) - F(\mu+\frac{y}{n^{1/2}})\right) = -n^{1/2}f(\mu_n^*)\frac{y}{n^{1/2}} = -yf(\mu_n^*),$$

donde μ_n^* es un punto intermedio entre μ y ν_n. Luego usando el hecho que $\nu_n \to 1/2$ y $\mu_n^* \to \mu$, resulta

$$\frac{(n/2-n\nu_n)}{(n\nu_n(1-\nu_n))^{1/2}} \to -2yf(\mu). \tag{7.11}$$

Luego, usando (7.7), (7.9), (7.10) y (7.11) tenemos que

$$\begin{aligned}\lim_{n\to\infty} P(n^{1/2}(\tilde{\mu}_n-\mu)\leq y) &= P\left(\tilde{\mu}_n \leq \mu + \frac{y}{n^{1/2}}\right)\\ &= 1-\Phi(-2f(\mu)y) = \Phi(2f(\mu)y),\end{aligned}$$

y por lo tanto hemos probado (7.6). □

Observación 1. El Teorema 1 implica que $\tilde{\mu}_n \xrightarrow{p} \mu$. También puede probarse que $\tilde{\mu}_n \xrightarrow{a.s.} \mu$, pero no se dará la demostración.

Apliquemos ahora este resultado al modelo (7.1) y supongamos que la distribución F de las u_i sea simétrica respecto de 0 con densidad f. En este caso se tendrá que la mediana de la distribución $F_{\mu\sigma}$ es μ y

$$f_{\mu\sigma}(x) = \frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right),$$

y por lo tanto,

$$f_{\mu\sigma}(\mu) = \frac{1}{\sigma}f(0).$$

Luego, de acuerdo al Teorema 1, se tendrá

$$n^{1/2}(\tilde{\mu}_n-\mu) \xrightarrow{D} N\left(0, \frac{\sigma^2}{4f^2(0)}\right).$$

Si $F = \Phi$, entonces $f(0) = 1/\sqrt{2\pi}$ y entonces

$$n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{D} N\left(0, \frac{\pi}{2}\sigma^2\right).$$

Por otro lado, $n^{1/2}(\bar{x}_n - \mu)$ tiene distribución N(0,σ^2). Por lo tanto la varianza asintótica de $\hat{\mu}_n$ es aproximadamente 57% más alta que la varianza de $\bar{x}_n$. Esto significa que la propiedad que tiene la mediana de ser poco sensible a observaciones atípicas tiene como contrapartida negativa ser 57% menos eficiente que $\bar{x}_n$ en el caso de errores normales. De todas maneras esto es menos grave que el comportamiento de $\bar{x}_n$ bajo una contaminación con outliers. En efecto, recordemos que en este caso una fracción de outliers tan pequeña como se quisiera podía provocar que la varianza se hiciese infinita.

Sin embargo, lo ideal sería tener un estimador robusto, es decir poco sensible a outliers y que simultáneamente fuera altamente eficiente cuando los datos son normales. En las secciones siguientes vamos a tratar entonces de encontrar estimadores con estas propiedades.

7.2 M-estimadores de posición

7.2.1 Definición de M-estimadores

Consideremos el modelo (7.1) y supongamos que conozcamos la distribución F de las u_i. y el parámetro de escala σ. Estas hipótesis no son muy realistas y más adelante las eliminaremos. Sin embargo será conveniente suponerlas momentáneamente para simplificar el planteo del problema. Supongamos que F tiene una densidad que llamaremos $f = F'$. Luego, la densidad de cada x_i será

$$f_{\mu\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right),$$

y luego la función de verosimilitud correspondiente a la muestra $x_1, ..., x_n$ será

$$L(\mu) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f\left(\frac{x_i - \mu}{\sigma}\right).$$

Tomando logaritmos, como σ se supone conocida, se tendrá que el estimador de máxima verosimilitud de μ que llamaremos $\hat{\mu}_f$(la f como subscripto indica que corresponde a que las u_i tienen densidad f) estará dado por el valor que maximiza

$$\sum_{i=1}^{n} \log f\left(\frac{x_i - \mu}{\sigma}\right).$$

Equivalentemente, podemos decir que $\hat{\mu}_f$ minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho_f\left(\frac{x_i - \mu}{\sigma}\right), \tag{7.12}$$

donde

$$\rho_f(u) = -\log f(u) + \log f(0).$$

Por ejemplo, si f corresponde a la distribución N(0,1). Entonces $\rho_f(u) = u^2/2$, y entonces el estimador de máxima verosimilitud minimiza

$$S(\mu) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

o equivalentemente, el que minimiza

$$S(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2,$$

el cual es precisamente $\bar{x}_n$.

Si f corresponde a la distribución doble exponencial, entonces

$$f(u) = \frac{1}{2}e^{-|u|},\ -\infty < u < \infty,$$

y por lo tanto $\rho_f(u) = |u|$. Entonces en este caso el estimador de máxima verosimilitud corresponde a minimizar

$$S(\mu) = \sum_{i=1}^{n} |x_i - \mu|, \tag{7.13}$$

y el valor que minimiza (7.13) es precisamente la mediana de la muestra.

En el párrafo anterior hemos visto los inconvenientes de media y la mediana muestral. Si conociéramos exactamente f, podríamos utilizar el estimador de máxima verosimilitud, del cual conocemos que tiene varianza asintótica mínima y que está dado por (7.12). Como en general se tiene sólo un conocimiento aproximado de f, por ejemplo que corresponde a una distribución de $\mathcal{V}_\epsilon$, Huber (1964) definió los M-estimadores para el modelo de posición como el valor $\hat{\mu}$ valor que minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right), \tag{7.14}$$

donde la función ρ es elegida independientemente de f y de tal manera que tenga las propiedades deseadas:

1. El estimador es altamente eficiente cuando f corresponde a la distribución N(0,1)

2. El estimador es poco sensible a contaminación por outliers, en particular es altamente eficiente para toda f correspondiente a una distribución de $\mathcal{V}_\epsilon$.

A la función ρ que define al M-estimador se le pedirá las siguientes propiedades

A1 La función ρ es derivable. Denominaremos $\psi = \rho'$.

A2 La función ρ es par.

A3 La función $\rho(u)$ es monótona no decreciente en $|u|$.

A4 Se cumple que $\rho(0) = 0$.

Huber (1964) propuso una familia de funciones ρ intermedias entre las correspondientes a la distribución N(0,1) y a la doble exponencial. Esta funciones es cuadrática para valores de valor absoluto pequeños y lineal para valores absolutos grandes. Más precisamente, para cada $k \geq 0$ se define ρ_k^H por

$$\rho_k^H(u) = \begin{cases} -ku - k^2/2 & \text{si} \quad u < -k \\ u^2/2 & \text{si} \quad |u| \leq k \\ ku - k^2/2 & \text{si} \quad u > k \end{cases} .$$

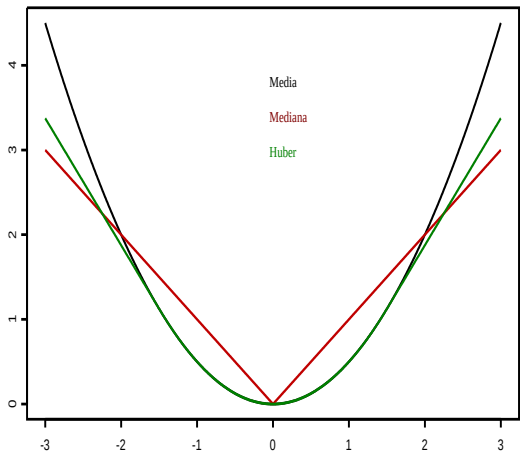

Figure 7.1: Funciones ρ correspondientes a la Media (en negro), la mediana (en rojo) y el M-estimador con función de Huber (en verde)

En la Figura 7.1 se grafican las funciones ρ correspondiente la media a la mediana y a la función de Huber. Obsérvese que las funciones ρ_k^H resultan derivables en todos los puntos, incluidos los puntos de cambio k y $-k$. Más adelante mostraremos que eligiendo k convenientemente los M-estimadores basadas en estas funciones gozan de las propiedades 1 y 2 enunciadas en esta sección.

Para encontrar el valor mínimo de $S(\mu)$ en (7.14) que define el M-estimador podemos encontrar sus punto críticos derivando. De esta manera obtenemos la siguiente ecuación

$$A(\mu) = \sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right) = 0. \tag{7.15}$$

El siguiente Teorema muestra que bajo ciertas condiciones la ecuación 7.15 tiene solución y corresponde a un mínimo de $S(\mu)$.

Teorema 2. Supongamos que ψ es continua impar, no decreciente y para algún a se tiene $\psi(a) > 0$. Entonces

(i) La ecuación (7.15) tiene al menos una raíz.

(ii) Toda raíz de (7.15) corresponde a un mínimo de $S(\mu)$.

(iii) Las raíces de (7.15) forman un intervalo.

(iv) Si ψ es estrictamente creciente hay una única raíz de (7.15).

Demostración. (i) Sea $M = \max_{1\leq i\leq n} x_i$ y $m = \min_{1\leq i\leq n} x_i$. Sea $\mu_1 = m-\sigma a$ y $\mu_2 = M+\sigma a$. Luego $(x_i-\mu_1)/\sigma \geq a$ para todo i y $(x_i-\mu_2)/\sigma \leq -a$ para todo i. Luego $\psi((x_i-\mu_1)/\sigma) \geq \psi(a) > 0$ para todo i y $\psi((x_i-\mu_2)/\sigma) \leq \psi(-a) = -\psi(a) < 0$ para todo i. Luego $A(\mu_1) > 0$ y $A(\mu_2) < 0$. Como $A(\mu)$ es continua, existe un punto μ_0 entre μ_2 y μ_1 tal que $A(\mu_0) = 0$.

(ii) Como $S'(\mu) = (-1/\sigma)A(\mu)$,es fácil ver que $S(\mu) - S(\mu_0) = (-1/\sigma)\int_{\mu_0}^{\mu} A(u)du$. Supongamos que μ_0 es una raíz de $A(\mu)$. Supongamos que $\mu_0 > 0$. Habrá que mostrar que

$$S(\mu_0) \leq S(\mu), \forall \mu. \tag{7.16}$$

Vamos a mostrar (7.16) solamente para $\mu > \mu_0$. El caso $\mu < \mu_0$ se demostrará similarmente. Tomemos $\mu > \mu_0$, luego

$$S(\mu) = \frac{1}{\sigma}\int_{\mu_0}^{\mu} A(u)du.$$

Como ψ es no decreciente resulta A no creciente. Luego como $A(\mu_0) = 0$, resulta $A(\mu) \leq 0$ para $\mu > \mu_0$. Por lo tanto resulta $\int_{\mu_0}^{\mu} A(u)du \leq 0$, y por lo tanto

$$S(\mu) \geq S(\mu_0).$$

En el caso $\mu < \mu_0$ se demuestra similarmente que también vale (7.16).

(iii) Supongamos que $\mu_1 < \mu_2$ sean raíces de A, y sea un valor μ tal que $\mu_1 < \mu < \mu_2$. Tenemos que mostrar que también $A(\mu) = 0$. Como A es no creciente se tendrá

$$0 = A(\mu_1) \geq A(\mu) \geq A(\mu_2) = 0.$$

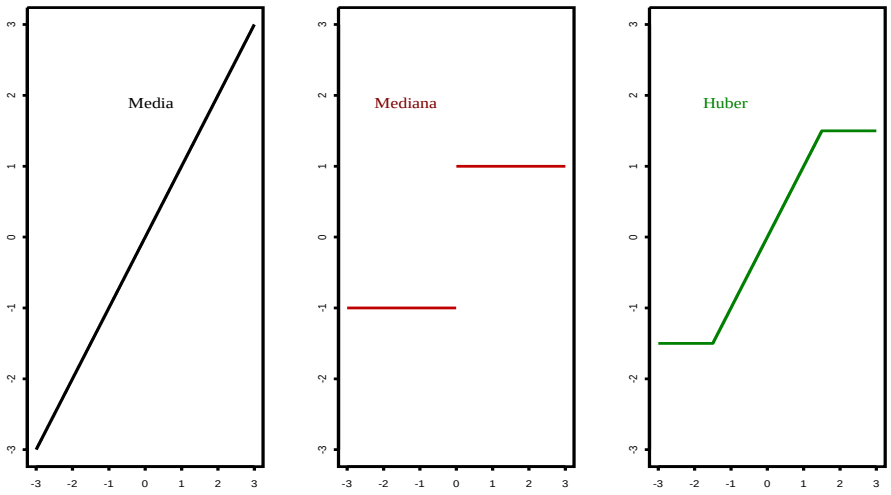

Figure 7.2: Funciones ψ correspondientes a la Media (en negro), la mediana (en rojo) y el M-estimador con función de Huber (en verde)

y luego $A(\mu) = 0$.

(iv) Supongamos que $A(\mu) = 0$. Veremos que no puede haber otra raíz de A. Sea primero $\mu^* > \mu$, como en este caso A es estrictamente decreciente se tendrá $A(\mu^*) < 0$. Similarmente se demuestra que si $\mu^* < \mu$, entonces $A(\mu^*) > 0$. □

Como vamos a ver más adelante la función ψ cumple un papel muy importante en la teoría de M-estimadores. Para la función ρ correspondiente a la media, resulta $\psi(u) = u$, para la función ρ correspondiente mediana $\psi(u) = |u|$, y para la funciones ρ_k^H, las correspondientes derivadas ψ_k^H están dadas por

$$\psi_k^H(u) = \begin{cases} -k & \text{si} \quad u < -k \\ u & \text{si} \quad |u| \leq k \\ k & \text{si} \quad u > k \end{cases} .$$

la cual corresponde a una identidad truncada. En Fig. 7.2 se grafican estas tres funciones ψ.

Como consecuencia de la propiedad A2, la función ψ es impar .

Para que el M-estimador sea robusto como veremos más adelante se requerirá que la función ψ sea acotada.

7.2.2 Propiedades asintóticas de M-estimadores

La condición de consistencia de Fisher, requerida para que el M-estimador converja a μ está dada por

$$E_{F_{\mu\sigma}}\left(\psi\left(\frac{x-\mu}{\sigma}\right)\right)=0,$$

y de acuerdo a (7.1), esto es equivalente a

$$E_F(\psi(u))=0. \tag{7.17}$$

Esta condición se cumple automaticamente si F tiene una densidad simétrica respecto de 0 ya que en ese caso se tendrá

$$E_F(\psi(u))=\int_{\infty}^{\infty} uf(u)du=0,$$

ya que $uf(u)$ será una función impar.

Luego, se tendrá el siguiente Teorema que muestra la consistencia de los M-estimadores:

Teorema 3. Sean $x_1,...x_n$ variables aleatorias independientes que satisfacen el modelo (7.1). Consideremos un estimador $\hat{\mu}_n$ solución de (7.15), donde ψ y F satisfacen (7.17) . Luego $\hat{\mu}_n$ converge en casi todo punto a μ en cualquiera de los siguientes casos

1. La función ψ es estrictamente creciente.
2. La función ψ es no decreciente, $\psi(u)>\psi(0)$ y $F(u)>F(0)$ para todo $u>0$.

Demostración: Solamente mostraremos el Teorema para el caso 1. Consideremos $\epsilon>0$. Luego como ψ es estrictamente creciente tenemos que $\psi(u-\epsilon)<\psi(u)$, y luego

$$E_F\psi(u-\epsilon)<E_F\psi(u)=0.$$

Por lo tanto

$$E_{F_{\mu\sigma}}\psi\left(\frac{x-(\mu+\epsilon)}{\sigma}\right) = E_F\psi(u-\epsilon) < 0. \tag{7.18}$$

Similarmente se puede probar que

$$E_{F_{\mu\sigma}}\psi\left(\frac{x-(\mu-\epsilon)}{\sigma}\right) = E_F\psi(u+\epsilon) > 0. \tag{7.19}$$

Sea ahora

$$G_n(\mu^*) = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{x_i-\mu^*}{\sigma}\right),$$

luego el M-estimador $\hat{\mu}_n$ satisface

$$G_n(\hat{\mu}_n) = 0. \tag{7.20}$$

Por otro lado usando la ley de los grandes números y (7.18) y (7.19) se tiene que con probabilidad 1 existe un n_0 tal que para todo $n > n_0$ se tiene que

$$G_n(\mu+\epsilon) < 0,\ G_n(\mu-\epsilon) > 0,$$

y por lo tanto como G_n es monótona decreciente, se tiene que el valor $\hat{\mu}_n$ satisfaciendo (7.20) tendrá que satisfacer que

$$\mu - \epsilon < \hat{\mu}_n < \mu + \epsilon.$$

Esto prueba la consistencia de $\hat{\mu}_n$.

El siguiente teorema muestra la asintótica normalidad de los M-estimadores

Teorema 4. Sean $x_1, ...x_n$ variables aleatorias independientes que satisfacen el modelo (7.1). Consideremos un estimador $\hat{\mu}_n$ solución de (7.15), donde ψ y F satisfacen (7.17). Supongamos que $\hat{\mu}_n$ es consistente, y que además ψ tiene dos derivadas continuas y ψ'' es acotada. Luego se tiene que

$$n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{D} N(0, \sigma^2 V(\psi, F)),$$

donde

$$V(\psi, F) = \frac{E_F\psi^2(u)}{(E_F\psi'(u))^2}. \tag{7.21}$$

Demostración. El M-estimador $\hat{\mu}_n$ satisface

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}_n}{\sigma}\right) = 0,$$

y haciendo un desarrollo de Taylor en el punto μ se tiene

$$0 = \sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^{n} \psi'\left(\frac{x_i - \mu}{\sigma}\right)\frac{\hat{\mu}_n - \mu}{\sigma} + \frac{1}{2}\sum_{i=1}^{n} \psi''\left(\frac{x_i - \mu_n^*}{\sigma}\right)\frac{(\hat{\mu}_n - \mu)^2}{\sigma^2},$$

donde μ_n^* es un punto intermedio entre $\hat{\mu}_n$ y μ.

Luego, haciendo un despeje parcial de $(\hat{\mu}_n - \mu)$ se tiene

$$(\hat{\mu}_n - \mu) = \frac{\displaystyle\sum_{i=1}^{n} \psi\left((x_i - \mu)/\sigma\right)}{\displaystyle\frac{1}{\sigma}\sum_{i=1}^{n} \psi'\left((x_i - \mu)/\sigma\right) - \frac{1}{2}\frac{(\hat{\mu}_n - \mu)}{\sigma^2}\sum_{i=1}^{n} \psi''\left((x_i - \mu_n^*)/\sigma\right)},$$

y luego

$$n^{1/2}(\hat{\mu}_n - \mu) = \frac{\displaystyle\frac{1}{n^{1/2}}\sum_{i=1}^{n} \psi\left((x_i - \mu)/\sigma\right)}{\displaystyle\frac{1}{n\sigma}\sum_{i=1}^{n} \psi'\left((x_i - \mu)/\sigma\right) - \frac{1}{2\sigma^2}(\hat{\mu}_n - \mu)\frac{1}{n}\sum_{i=1}^{n} \psi''\left((x_i - \mu_n^*)/\sigma\right)}. \tag{7.22}$$

Sea

$$A_n = \frac{1}{n^{1/2}}\sum_{i=1}^{n} \psi\left((x_i - \mu)/\sigma\right) = \frac{1}{n^{1/2}}\sum_{i=1}^{n} \psi\left(u_i\right),$$

$$B_n = \frac{1}{n}\sum_{i=1}^{n} \psi'\left((x_i - \mu)/\sigma\right) = \frac{1}{n}\sum_{i=1}^{n} \psi'\left(u_i\right),$$

y

$$C_n = \frac{1}{2}(\hat{\mu}_n - \mu)\frac{1}{n}\sum_{i=1}^{n} \psi''\left((x_i - \mu_n^*)/\sigma\right).$$

Luego

$$n^{1/2}(\hat{\mu}_n - \mu) = \frac{A_n}{\sigma^{-1}\ B_n + \sigma^{-2} C_n}. \tag{7.23}$$

Por el Teorema Central del Límite se tiene

$$A_n \xrightarrow{D} N(0, E_F(\psi^2(u))). \tag{7.24}$$

Por la Ley Fuerte de los Grandes Números se tiene

$$B_n \xrightarrow{p} E_F(\psi'(u)). \tag{7.25}$$

Finalmente, por hipótesis existe una constante K tal que $|\psi''(u)| < K$. Luego $|C_n| < (K/2)(\hat{\mu}_n - \mu)$. Usando el hecho de que $\hat{\mu}_n \xrightarrow{p} \mu$, se tiene que

$$C_n \xrightarrow{p} 0. \tag{7.26}$$

Usando (7.23)-(7.26) se deduce el Teorema. □

7.2.3 M-estimador minimax para la varianza asintótica

El problema que vamos a desarrollar en esta sección es el de elegir la función ρ o equivalentemente la función ψ del M-estimador. En está sección vamos a utilizar como criterio minimizar la varianza asintótica del M-estimador dada en (7.21). Si conociéramos la distribución F de las u_i, utilizaríamos el M-estimador que tiene como función ψ la dada por

$$\psi(u) = \frac{d \log f(u)}{du},$$

es decir el estimador de máxima verosimilitud. Este estimador minimiza la varianza asintótica $V(\psi, F)$ dada en (7.21). Cuando existe la posibilidad de que hubieran outliers la distribución F no es conocida exactamente y por lo tanto no podemos usar este estimador.

La solución que propuso Huber (1964) es la siguiente. supongamos que F esté en el entorno de contaminación dado por (7.2), pero restringiendo H a distribuciones simétricas respecto de 0. Para esto definimos un nuevo entorno de distribuciones de Φ

$$\mathcal{V}_\epsilon^* = \{F : F = (1-\epsilon)\Phi + \epsilon H \text{ con } H \text{ simétrica}\}. \tag{7.27}$$

Luego, si usa el M-estimador basado en la función ψ. la mayor varianza posible en este entorno está dada por

$$V^*(\psi) = \sup_{F \in \mathcal{V}_\epsilon^*} V(\psi, F).$$

El criterio de Huber para elegir el M-estimador es utilizar la función ψ^* que minimice $V^*(\psi)$. Estos estimadores se denominarán minimax (minimizan la máxima varianza asintótica en el entorno de contaminación V_ε^*. En Huber (1964) se muestra que ψ^* está en la familia ψ_k^H, donde k depende de la cantidad de contaminación ϵ.

7.2.4 M-estimadores con escala desconocida

La definición de los M-estimadores dada en (7.14) supone que σ es conocida. Sin embargo, en la práctica σ es desconocida. En estos casos podemos reemplazar en esta ecuación σ por un estimador $\hat{\sigma}$, y el M-estimador se definirá por el valor $\hat{\mu}$ que minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\hat{\sigma}_n}\right). \tag{7.28}$$

Si queremos que el M-estimador resultante de μ sea robusto, será necesario que $\hat{\sigma}$ también lo sea. El estimador insesgado usual de σ dado por

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}(x_i - \bar{x})^2$$

no es robusto. En efecto, es fácil ver que una observación lo pueda llevar fuera de todo límite. Un estimador robusto de σ es el llamado MAD (median absolute deviation), que está definido por

$$\hat{\sigma}^2 = A \text{ mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \le i \le n\},$$

donde

$$\tilde{\mu}_n = \text{mediana}\{x_i : 1 \le i \le n\},$$

y donde A es una constante que hace que el estimador sea consistente a σ en el caso de que las observaciones sean una muestra aleatoria de una $\mathrm{N}(\mu, \sigma^2)$.

Vamos ahora a deducir cual debe ser el valor de A. Sean $x_1, ..., x_n$ una muestra de una distribución N(μ,σ^2). Entonces podemos escribir $x_i = \mu + \sigma u_i$, donde $u_1, ..., u_n$ es una muestra aleatoria de una distribución N(0,1). En este caso tenemos que

$$x_i - \tilde{\mu}_n = (\mu - \tilde{\mu}_n) + \sigma u_i$$

y

$$\text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\} = \text{mediana}\{|(\mu - \tilde{\mu}_n) + \sigma u_i|,\ 1 \leq i \leq n\}.$$

Como de acuerdo a lo visto en Observación 1, $\lim(\mu - \tilde{\mu}_n) = 0$ casi seguramente, se tendrá que

$$\lim_{n\to\infty} \text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\} = \lim_{n\to\infty} \text{mediana}\{|\sigma u_i|,\ 1 \leq i \leq n\}\ \}$$

$$= \sigma \lim_{n\to\infty} \text{mediana}\{|u_i|,\ 1 \leq i \leq n\},\ \text{c.s.}. \qquad (7.29)$$

Si u es N(0,1), entonces $|u|$ tiene distribución $2\Phi - 1$. Sea entonces $B = \text{mediana}(2\Phi - 1)$, luego por lo visto en Observación 1 se tiene

$$\lim_{n\to\infty} \text{mediana}\{|u_i|,\ 1 \leq i \leq n\} = B,\ \text{c.s.}$$

y usando (7.29)

$$\lim_{n\to\infty} \text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\} = \sigma B \text{ c.s.}$$

Luego $A = 1/B$. La constante B se calcula de la siguiente manera

$$2\Phi(B) - 1 = 0.5,$$

o sea

$$\Phi(B) = 0.75,\ B = \Phi^{-1}(0.75) = 0.675.$$

Luego se tendrá que el estimador MAD de σ viene dado por

$$\hat{\sigma}^2 = \frac{1}{0.6745}\text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\}.$$

Cuando el M-estimador se obtiene minimizando (7.28), la ecuación (7.15) se transforma en

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\hat{\sigma}}\right) = 0. \tag{7.30}$$

Las propiedades asintóticas del estimador $\hat{\mu}$ solución de (7.30) son similares a las del estimador correspondiente al caso de σ conocida. El siguiente Teorema se dará sin demostración.

Teorema 5. Sean $x_1, ...x_n$ variables aleatorias independientes que satisfacen el modelo (7.1). Consideremos un estimador $\hat{\mu}_n$ solución de (7.30), donde ψ es impar y F es simétrica respecto de 0. Supongamos que $\hat{\mu}_n$ es consistente a μ y $\hat{\sigma}_n$ es consistente a σ, y que además ψ tiene dos derivadas continuas y ψ'' es acotada. Luego se tiene que

$$n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{D} N(0, \sigma^2 V(\psi, F)),$$

donde V está dada por (7.21)

7.2.5 Algoritmos para calcular M-estimadores

A continuación vamos a describir tres algoritmos para computar el M-estimador definido como la solución de (7.30).

Algoritmo basado en medias ponderadas iteradas (MPI)

Llamemos $w(u) = \psi(u)/u$. Luego la ecuación (7.30).se puede escribir como

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) w\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0,$$

o sea

$$\sum_{i=1}^{n} x_i w\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = \hat{\mu} w\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right),$$

y haciendo un despeje "parcial"de $\hat{\mu}$ se tiene

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu})/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)}. \tag{7.31}$$

En realidad esta expresión no es un verdadero despeje, ya que el miembro derecho también aparece $\hat{\mu}$. Sin embargo esta fórmula nos va a sugerir un algoritmo iterativo para calcular $\hat{\mu}$.

En efecto, consideremos un estimador inicial $\hat{\mu}_0$ de μ, como por ejemplo la mediana.Luego podemos definir

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu}_0)/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu}_0)/\hat{\sigma}\right)},$$

y en general si ya tememos definido $\hat{\mu}_h$, podemos definir $\hat{\mu}_{h+1}$ por

$$\hat{\mu}_{h+1} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu}_h)/\hat{\sigma}\right)}. \tag{7.32}$$

Se puede mostrar que este si ψ es continua, entonces cuando este algoritmo iterativo converge, lo hace a una solución de (7.30). En efecto supongamos que $\lim_{h\to\infty} \hat{\mu}_h = \hat{\mu}$, luego tomando limite en ambos lados de (7.32), se tendrá

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu})/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)}. \tag{7.33}$$

Pero esta ecuación es precisamente (7.31) , que ya hemos visto es equivalente a (7.30).

La ecuación (7.33) muestra a $\hat{\mu}$ como promedio pesado de las x_i y pesos proporcionales a $w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)$. Como en general $w(u)$ es una función par monótona no creciente en $|u|$, (7.33) se puede interpretar como que el M-estimador da a cada observación un peso que penaliza las observaciones para las cuales $|x_i - \hat{\mu}|/\hat{\sigma}$ es grande. Para la media se tiene $w(u) = 1$, y para el estimador basado en la función ψ_k^H, la correspondiente función de peso está dada por

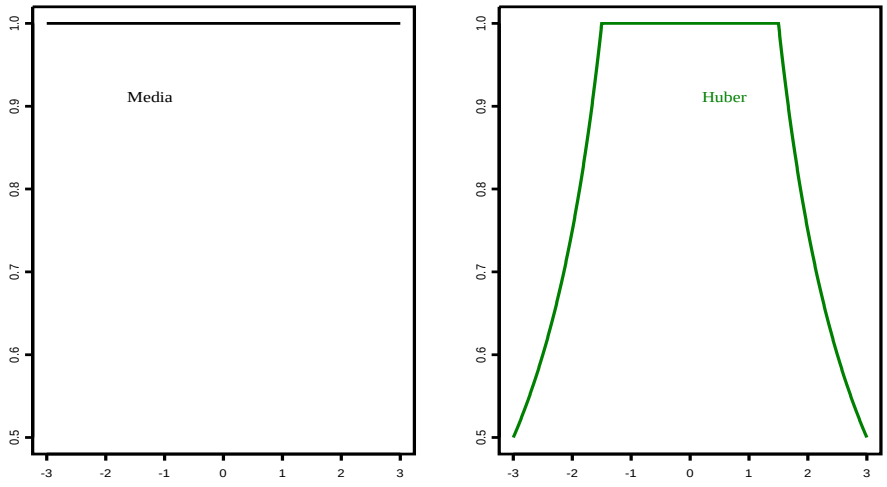

Figure 7.3: Funciones de peso w correspondientes a la Media (en negro) y al M-estimador con función de Huber (en verde)

$$w_k^H(u) = \begin{cases} 1 & \text{si} \quad |u| \leq k \\ \frac{k}{|u|} & \text{si} \quad |u| > k \end{cases} .$$

El gráfico de esta función se encuentra en la Figura 7.3.

Algoritmo basado en medias de pseudovalores iteradas (MPVI)

Definamos el pseudovalor $x_i^*(\mu)$ por

$$x_i^*(\mu) = \mu + \hat{\sigma}\psi\left((x_i - \hat{\mu})/\hat{\sigma}\right).$$

Luego se tiene

$$\psi\left((x_i - \hat{\mu})/\hat{\sigma}\right) = (x_i^*(\mu) - \hat{\mu})/\hat{\sigma},$$

y reemplazando en (7.30) se tiene la ecuación para el M-estimador es

$$\sum_{i=1}^{n}(x_i^*(\hat{\mu}) - \hat{\mu})/\hat{\sigma} = 0.$$

Haciendo un despeje parcial de $\hat{\mu}$ se tiene

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i^*(\hat{\mu}). \tag{7.34}$$

Es decir, se tiene expresado el M-estimador como promedio simple de los pseudo valores. Esta fórmula no permite calcular el M-estimador directamente, ya que el miembro derecho también depende de $\hat{\mu}$. Sin embargo, nos sugiere el siguiente algoritmo iterativo. Partiendo de un estimador inicial $\hat{\mu}_0$, consideramos la siguiente fórmula recursiva para $\hat{\mu}_h$

$$\hat{\mu}_{h+1} = \frac{1}{n}\sum_{i=1}^{n} x_i^*(\hat{\mu}_h). \tag{7.35}$$

Es interesante calcular los pseudovalores correspondientes a ψ_k^H, los cuales están dados por

$$x_i^*(\mu) = \begin{cases} \mu - k\hat{\sigma} & \text{si} \quad x_i < \mu - k\hat{\sigma} \\ x_i & \text{si} \quad |x_i - \mu| \leq k\hat{\sigma} \\ \mu + k\hat{\sigma}\ V & \text{si} \quad x_i > \mu + k\hat{\sigma} \end{cases}.$$

Es decir, si x_i pertenece al intervalo $[\mu - k\hat{\sigma},\ \mu + k\hat{\sigma}]$, el pseudovalor $x_i^*(\mu)$ es igual a la observación x_i. Si x_i está fuera de este intervalo el psudovalor se define como el extremo del intervalo más cercano.

Vamos a ver ahora que si $\lim_{h\to\infty} \hat{\mu}_h = \hat{\mu}$ y ψ es continua, entonces $\hat{\mu}$ es el M-estimador solución de (7.30). En efecto, tomando límite en ambos miembros de (7.35) se obtiene (7.34), que ya hemos visto es equivalente a (7.30).

Algoritmo de Newton Raphson (NR)

De acuerdo a lo visto anteriormente, el algoritmo de Newton Raphson para calcular la raíz de (7.30) tiene la siguiente fórmula recursiva

$$\hat{\mu}_{h+1} = \hat{\mu}_h + \hat{\sigma}\frac{\sum_{i=1}^{n} \psi\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}{\sum_{i=1}^{n} \psi'\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}. \tag{7.36}$$

Para el caso de que $\psi = \psi_k^H$, está formula toma una expresión particularmente interesante.

Para cada valor μ dividamos el conjunto de observaciones en tres conjuntos

$$\begin{aligned} D_- &= \{i : (x_i - \hat{\mu}_h)/\hat{\sigma} < -k\}, \\ D_0 &= \{i : |x_i - \hat{\mu}_h|/\hat{\sigma} \leq k\}, \\ D_+ &= \{i : (x_i - \hat{\mu}_h)/\hat{\sigma} > k\}. \end{aligned}$$

Es fácil ver que se tiene

$$\psi_k^H\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right) = \begin{cases} -k & \text{si} \quad i \in D_- \\ (x_i - \hat{\mu}_h)/\hat{\sigma} & \text{si} \quad i \in D_0 \\ k & \text{si} \quad i \in D_+ \end{cases},$$

y

$$\psi_k^{H\prime}\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right) = \begin{cases} 0 & \text{si} \quad i \in D_-(\hat{\mu}_h) \\ 1 & \text{si} \quad i \in D_0(\hat{\mu}_h) \\ 0 & \text{si} \quad i \in D_+(\hat{\mu}_h) \end{cases}.$$

Llamando n_-, n_0 y n_-, al número de elementos de D_-, D_0 y D_+ y reemplazando en (7.36), se tiene

$$\hat{\mu}_{h+1} = \hat{\mu}_h + \hat{\sigma}\frac{k(n_+ - n_-) + \sum\limits_{i \in D_0}(x_i - \hat{\mu}_h)/\hat{\sigma}}{n_0} = \frac{n_+ - n_-}{n_0}\hat{\sigma}k + \frac{1}{n_0}\sum_{i \in D_0} x_i.$$

Obsérvese que el miembro derecho de esta última fórmula solo depende de D_-, D_0 y D_+. Estos tres conjuntos forman una partición del conjunto $\{1, 2, ..., n\}$. Es claro que hay un número finito de estas particiones, y por lo tanto si $\hat{\mu}_h$ converge lo debe hacer en un número finito de pasos.

Convergencia de los algoritmos iterativos

Se puede demostrar que los 3 algoritmos iterativos que hemos estudiado MPI, MPVI, y NR convergen a la raíz de (7.30) cuando ψ es monótona no decreciente cuando ésta es única. Si (7.30) tiene más de una raíz, se puede demostrar que si $[\hat{\mu}_1,\ \hat{\mu}_2]$ es el intervalo de soluciones, entonces dado $\epsilon > 0$, existe h_0 tal que $\hat{\mu}_h \in [\hat{\mu}_1 - \epsilon,\ \hat{\mu}_2 + \epsilon]$ para todo $h > h_0$.