Modelos de Localización y Escala
Algunas consideraciones teóricas y aplicaciones a pequeñas muestras

-

Gabriela L. Damilano Scarpinello

-

Director: Pere Puig Casado

Universidad Autónoma de Barcelona
08193 Bellaterra - Barcelona - España
2005

It is not Knowledge, but the act of learning,
not possession but the act of getting there,
which grants the greatest enjoyment.

Karl Friedrich Gauss (1808).

De alguna manera estas palabras describen lo que siento en este momento. Es verdad que las páginas siguientes registran los resultados de nuestra investigación. Sin embargo, estoy convencida que encierran, quizás de manera imperceptible para algunos, la "memoria de un proceso de aprendizaje" en un sentido más amplio.

En lo que a mi formación académica e introducción al mundo de la investigación se refiere, son el producto de 7 años que he transitado de la mano de quien, más que mi director, ha sido la persona que me ha acompañado con su compromiso, solidez académica y estímulo constante en esto de "hacer aprendiendo" y "aprender haciendo".

Cierto es también, que este proceso no sería posible si no estuviésemos inmersos en un contexto que nos facilita la tarea. Me refiero a la UAB y la UNRC como instituciones y, especialmente, a los docentes, compañeros y amigos que en estos ámbitos uno tiene la suerte de encontrar y compartir algo más que lo meramente académico.

No obstante, a la par de la elaboración de este trabajo, como cualquier otro, corre la vida. La de todos los días, conformada por nuestros afectos, amigos y familia, esos que no saben de qué va esta memoria pero saben del esfuerzo, con sus desavenencias y sus logros, y aunque no figuren en la bibliografía, saben que son de alguna manera parte, no de la "posesión" sino del "haber llegado".

A todos ustedes que, saben a quien me refiero aunque no los nombre, han contribuido y alentado este trabajo y el maravilloso "acto de aprender" cada día, en cualquiera de sus formas, deseo expresar mi más sincera y sentida gratitud y de manera muy especial

A mis Padres, por acompañarme siempre con su afecto y apoyo incondicional
permitiéndome crecer libremente.

A Jorge, por estar conmigo por elección, la de construir día a día una vida de a dos
con todo lo que eso significa.

Gabriela.
Abril 2005.

Tabla de Contenidos

Introducción

En 1934 Fisher resalta dos amplias clases de modelos estadísticos paramétricos que juegan un rol preponderante tanto en la Estadística Matemática, por sus propiedades inferenciales, como en la Estadística Aplicada, por sus estructuras: Las Familias Exponenciales y las Familias de Grupos de Transformaciones.

Fisher (1934, p. 296-303) se refirió a estas últimas utilizando las siguientes frases:

"...A second case, of somewhat wider practical application...", *"...A typical case of such a relationship occurs in parameter of location..."*, *"...In a very frequent class of cases not only the origen but the scale of the distribution is also represented by a parameter to be estimated from the observations..."*

En este sentido, nuestra investigación se centra principalmente en el estudio de las características y los procedimientos de inferencia de unos destacados representantes de las Familias de Grupos de Transformaciones: los *Modelos de Localización y Escala.* Si bien el trabajo se desarrolla básicamente dentro del ámbito de la Estadística Matemática, nuestra intención es encontrar resultados que puedan ser aplicados en diferentes ámbitos.

El propósito principal de este trabajo es aportar y analizar nuevos modelos estadísticos que puedan utilizarse en situaciones prácticas. Dentro de estos modelos paramétricos, buscaremos también desarrollar métodos de estimación alternativos

al de máxima verosimilitud que, preservando sus propiedades asintóticas, sean más sencillos de calcular y tengan un buen comportamiento cuando el tamaño muestral sea pequeño. Nuestra especial atención a las *pequeñas muestras* está motivada por el hecho que, la limitada disponibilidad de datos, es una situación particularmente conflictiva en inferencia estadística.

Generalmente, los procedimientos usados en inferencia estadística están basados en el conocimiento de la distribución de los estadísticos involucrados. No obstante, en muchos problemas reales, su distribución exacta es intratable o incluso desconocida por lo cual casi siempre se deben utilizar métodos aproximados. El desarrollo que los procedimientos asintóticos de inferencia basados en la verosimilitud han tenido en las últimas décadas, ilustran claramente la estrecha relación entre la teoría asintótica y la Estadística Matemática. Por ejemplo, es bien conocido que el estimador de máxima verosimilitud se distribuye asintóticamente como una normal o que es asintóticamente eficiente. Sin embargo, estas aproximaciones asintóticas llamadas de Primer-Orden son válidas cuando el número de observaciones tiende a infinito, lo que nos lleva naturalmente a preguntarnos, ¿qué sucede cuando el tamaño muestral es pequeño?

Justamente, el estudio de esta cuestión aplicada a unos determinados modelos paramétricos en concreto, es una de las motivaciones de nuestra investigación, orientada básicamente en dos sentidos:

- Proponer y comparar estimadores asintóticamente competitivos del estimador de máxima verosimilitud y apropiados para pequeños tamaños muestrales, con especial atención en la estimación del parámetro de localización, línea recurrente en los capítulos 2 al 4.

- Mejorar las aproximaciones asintóticas clásicas, usando metodologías asintóticas

de Orden-Superior (*aproximación Saddlepoint*) en presencia de censura, otro aspecto que crea especiales problemas en el análisis de datos particularmente dentro del análisis de supervivencia y fiabilidad, que abordamos en el último capítulo.

A continuación y de manera más específica, detallaremos como está estructurada la memoria:

1. El capítulo 1 es instrumental. En él hacemos una breve descripción del marco teórico sobre el que se sustentan los resultados producto de esta investigación.

2. Casi 200 años atrás, Laplace estudió cómo estimar la media teórica de una distribución simétrica. Para ello propuso utilizar una combinación lineal de la mediana y media muestrales, de tal manera que la varianza asintótica fuese minimizada. En el capítulo 2, analizamos este tipo de estimadores del parámetro de localización y además caracterizamos a todos los modelos de localización simétricos para los cuales una combinación lineal de la media y mediana es un estimador asintóticamente eficiente del parámetro de localización (Teorema 2.2.1). Como corolario de este resultado surge inmediatamente que *el único modelo de localización simétrico tal que la media muestral es un estimador asintóticamente eficiente del parámetro de localización es la distribución normal* y que *la distribución de Laplace es el único modelo de localización simétrico tal que la mediana muestral es un estimador asintóticamente eficiente del parámetro de localización.* Extendiendo este resultado al caso particular del estimador de Hodges-Lehmann, mediante el Teorema 2.3.1 caracterizamos a *la distribución logística como el único modelo de localización simétrico para el cual este estimador es asintóticamente eficiente.*

3. Fruto del Teorema 2.2.1, una nueva familia de localización y escala simétrica es introducida en el capítulo 3. Para este caso particular, además de analizar sus características, estudiamos en detalle la estimación puntual de los parámetros y también la construcción de intervalos de confianza, aproximados y exactos. Demostramos que, los estimadores alternativos que presentamos, poseen propiedades asintóticas deseables que los hacen buenos competidores de los estimadores de máxima verosimilitud y son más sencillos de calcular. También realizamos un estudio de simulación para comparar ambos estimadores cuando tenemos pequeñas muestras.

4. En el capítulo 4, estudiamos los modelos de localización simétricos resultantes del Teorema 2.2.1 como una familia de distribuciones a tres parámetros: la *distribución Normal Truncada "Simetrizada" o "Doblada"* (NTS). La distribución NTS es la simetrización de la distribución normal truncada simple definida sobre $\mathbb{R}^+$ y también, un modelo de localización y escala simétrico con un parámetro extra θ, cuyo valor está relacionado directamente con el coeficiente de curtosis. Mediante la Proposición 4.3.1 demostramos que, si θ es reemplazado por un estimador consiste, el estimador del parámetro de localización basado en la combinación lineal de la media y la mediana preserva todas las propiedades asintóticas deseadas. Este resultado nos permite obtener estimadores alternativos al de máxima verosimilitud con buenas propiedades asintóticas. Concretamente, presentamos dos métodos: uno basado en la curtosis empírica, que se destaca por su sencillez de cálculo, y un algoritmo iterativo que puede implementarse fácilmente usando paquetes estadísticos estándares que trabajen con la distribución normal truncada. Como en todos los casos, y dado nuestro interés

especial en pequeñas muestras, realizamos un estudio basado en simulaciones a fin de comparar el comportamiento de los distintos estimadores propuestos. La distribución NTS, puede utilizarse para analizar conjuntos de datos que presenten una leve desvío de la distribución normal reflejada por un incremento de su coeficiente de curtosis, como lo mostramos con un ejemplo sobre tasas diarias de cambio de las monedas Dólar US y Euro.

5. En el Capítulo 5 abordamos la estimación puntual y por intervalos para los casos de una y dos muestras provenientes de familias de localización y escala en presencia de censura de Tipo I. Además de demostrar una condición suficiente para la unicidad del estimador de máxima verosimilitud (Teorema 5.1.1), aplicamos el estadístico Z^* basado en la aproximación asintótica de orden-superior Saddlepoint y estudiamos su comportamiento, especialmente cuando el tamaño muestral es pequeño. Para el caso de una muestra, analizamos en particular las distribuciones Normal y Valor Extremo. En la extensión de estos procedimientos para el problema de dos muestras, consideramos la comparación de medias de dos poblaciones Normales cuando las muestras son emparejadas o independientes (problema de Behrens-Fisher). También estudiamos la comparación de dos distribuciones de Weibull, donde el interés radica en contrastar si los parámetros de forma, correspondientes a los de escala de las distribuciones del Valor Extremo, pueden asumirse como iguales.

Durante el proceso de esta investigación hemos tratado de divulgar nuestros resultados mediante comunicaciones en diversos congresos de Estadística y mediante artículos en revistas internacionales. Los frutos de esta tarea divulgativa han sido los siguientes:

Comunicaciones en congresos

- Damilano, G., Puig, P. *Location and Scale models with Type I Censored Data.* 15th International Workshop on Statistical Modelling. Bilbao, 2000.

- Damilano, G., Puig, P. *Inferencia en una distribución de Weibull para muestras con censura.* VIII Seminario de Estadística Aplicada del Instituto Interamericano de Estadística (IASI). Panamá, 2001.

- Damilano, G., Puig, P. *Un Modelo de Localización y Escala Simétrico.* V Coloquio Latinoamericano de Sociedades de Estadística, Buenos Aires, 2002.

- Damilano, G., Puig, P. *On the symmetrized truncated normal distribution.* IX Conferencia Española de Biometría. La Coruña, 2003.

- Damilano, G., Puig, P. *A new look at an old problem: the double truncated normal distribution.* 6th World Congress of the Bernulli Society and the 67 Annual Meeting of the Institute of Mathematical Statistics, Barcelona, 2004.

Publicaciones

- Damilano y Puig (2002). Small Sample Asymptotics for Type I Censored Data. *Biometrical Journal*, **44**,no. 7, p. 867-876.

- Damilano y Puig (2004). Efficiency of a Linear Combination of the Median and the Sample Mean: The Double Truncated Normal Distribution. *Scandinavian Journal of Statistics*, **31**, no. 4, p. 629-637(9).

Finalmente deseamos destacar que, si bien esta memoria recoge algunos resultados sobre inferencias en Modelos de Localización y Escala con énfasis en aplicaciones prácticas a pequeñas muestras, el trabajo no está ni mucho menos cerrado. Por el contrario, consideramos que esta primera aproximación es un buen punto de partida para futuras investigaciones.

Capítulo 1

Marco Teórico

En este capítulo realizaremos una breve descripción del marco teórico sobre el que se sustentan los distintos tópicos que desarrollaremos en el presente trabajo.

Debido a que nuestra investigación se centra en procedimientos de inferencia paramétrica sobre modelos de localización y escala basados en la verosimilitud, serán éstos los principales puntos a desarrollar. Además, presentaremos otros conceptos y resultados que juegan un papel importante en el contenido de esta memoria y que, por tanto, serán utilizados como futuras referencias.

Cabe aclarar que la exposición de este capítulo no es exhaustiva ni detallada. No obstante, en las diferentes referencias bibliográficas que se citan oportunamente, puede encontrarse un completo análisis de los temas tratados.

1.1 Modelos de Localización y Escala

Las *Familias de Transformaciones* constituyen, junto a las Familias Exponenciales, una clase general de modelos particularmente relevantes en la teoría y la práctica de la Estadística. Además de contener un gran número de modelos que son útiles en las aplicaciones, la suposición de que el proceso generador de los datos pertenezca

a una de estas familias hace posible desarrollar procedimientos generales, simples y precisos, para realizar inferencias (ver Pace y Salvan, 1997).

Concretamente, vamos a considerar familias de distribuciones generadas por la acción de un grupo de transformaciones. A continuación vamos a precisar estos conceptos.

Definición 1.1.1. Familias Generadas por Grupos.
Sea Y una variable aleatoria continua definida sobre todos los reales, con función de densidad f_0. Consideremos además $\mathcal{G}$, un grupo de transformaciones a valores reales, uno a uno y medibles. Entonces,

$$\mathcal{F}_{\mathcal{G}} = \{g.f_0;\, g \in \mathcal{G}\,\}$$

donde $g.f_0$ denota la densidad de $g(Y)$, es la *Familia generada por* Y *bajo la acción del grupo* $\mathcal{G}$.

Restringiendo $g(.)$ a clases apropiadas, es posible construir familias de densidades que pueden usarse como modelos estadísticos. Si además, f_0 es de la forma $f_0(y;\psi)$ con ψ un vector de parámetros que toma valores en un conjunto abierto y no vacío $\Psi \subseteq \mathbb{R}^d$, entonces $\mathcal{F}_{\mathcal{G}}$ será un *Modelo Paramétrico.* También podemos introducir parámetros en el modelo mediante el mismo grupo $\mathcal{G}$.

Observemos que f_0 es un elemento de la familia, dado que la transformación identidad está contenida en el grupo $\mathcal{G}$. Un ejemplo particularmente importante, es la Familia de Localización y Escala generada por f_0.

Definición 1.1.2. Sea Y una variable aleatoria continua real, con densidad $f_0(y)$ y distribución $F_0(y)$, y sea $\mathcal{G}$ el grupo de transformaciones a valores reales $g(y) =$

$\mu + \sigma\, y$, donde $\mu \in \mathbb{R}$ y $\sigma \in \mathbb{R}^+$. Entonces $\mathcal{F}_{LE} = \{g.f_0\,,\, g \in \mathcal{G}\}$ es la *Familia de Localización y Escala* generada por $f_0(.)$, con densidad dada por

$$f(y;\mu,\sigma) = \frac{1}{\sigma} f_0(\frac{y-\mu}{\sigma}), -\infty < y < \infty \tag{1.1.1}$$

y función de distribución $F(y;\mu,\sigma) = F_0((y-\mu)/\sigma)$, siendo $F_0(y) = \int_{-\infty}^{y} f_0(t)dt$.

Como cada transformación g está identificada en $\mathcal{G}$ por $\psi = (\mu,\sigma) \in \mathbb{R}\text{x}\mathbb{R}^+$, las familias $\mathcal{F}_{LE}$ obtenidas por su acción son familias a dos parámetros. En particular μ se denomina *Parámetro de Localización* y σ es el *Parámetro de Escala.*

Observemos que si Y tiene densidad $f(y;\mu,\sigma) \in \mathcal{F}_{LE}$, la variable aleatoria $g(Y) = \mu' + \sigma' Y$ tendrá densidad $f(y;\mu'+\sigma'\mu,\sigma'\sigma) \in \mathcal{F}_{LE}$. Es decir, la estructura de la densidad no cambia pero si los valores de los parámetros. En este caso se dice que $\mathcal{F}_{LE}$ es una *familia invariante* bajo la acción del grupo de transformaciones $\mathcal{G}$.

Observemos también que los parámetros μ y σ de la variable aleatoria $g(Y) = \mu + \sigma Y$ no coinciden en general con la media y desviación estándar de Y, excepto en el caso especial en que la variable aleatoria generadora tenga media cero y varianza unitaria. Esto se deduce a partir de las siguientes relaciones evidentes

$$E\{g(Y)\} = \mu + \sigma E\{Y\}$$

$$Var\{g(Y)\} = \sigma^2 Var\{Y\}.$$

Por lo general, en la práctica, trabajaremos con una muestra aleatoria simple. Por tanto, tendremos como punto de partida un vector aleatorio $(Z_1,...,Z_N)$ de componentes independientes e idénticamente distribuidas (iid) con densidad $f_0(z) \in \mathcal{F}_{LE}$. Si en este caso consideramos el grupo $\mathcal{G}$ de transformaciones $g(z_1,...,z_N) = (\mu +$

$\sigma z_1, ..., \mu + \sigma z_N)$, entonces la familia generada por f_0 bajo la acción del grupo $\mathcal{G}$ es la *Familia de densidades de una muestra aleatoria simple de tamaño N de una* $\mathcal{F}_{LE}$.

Otra perspectiva interesante de las familias de localización y escala, que no analizaremos en este trabajo, es su extensión a los modelos de regresión. Por ejemplo, si $(Z_1, ..., Z_N) \sim N_N(0, I_N)$, las transformaciones $Y_i = \mu_i + \sigma Z_i$, $i = 1, ..., N$ con $\sigma > 0$ y vector de localización $(\mu_1, ..., \mu_N) = \beta_1 x_1 + ... + \beta_k x_k$ con $\beta_1, ..., \beta_k$ escalares arbitrarios y $x_1, ..., x_k$ vectores fijos linealmente independientes de $\mathbb{R}^N$, determinan el *Modelo de Regresión Lineal Clásico con Errores Normales.* Otra extensión, incluye a los *Modelos de Regresión No Lineales* como $Y = \eta(\beta, x) + \sigma Z$, donde $\eta(\beta, x)$ es una función no lineal de β.

Para concluir este punto, en el Cuadro I presentamos las funciones de densidad de algunos de los miembros más destacados de estas familias.

Cuadro I: Algunas Familias de Localización y Escala.

Distribución	Densidad
Normal: $N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{1}{2}(\frac{y-\mu}{\sigma})^2]$
Valor Extremo: $EV(\mu, \sigma)$	$\frac{1}{\sigma}\exp[(\frac{y-\mu}{\sigma}) - \exp(\frac{y-\mu}{\sigma})]$
Logística: $L(\mu, \sigma)$	$\frac{1}{\sigma}\frac{\exp[-(\frac{y-\mu}{\sigma})]}{\{1+\exp[-(\frac{y-\mu}{\sigma})]\}^2}$
Exponencial: $Exp(\mu, \sigma)$	$\frac{1}{\sigma}\ exp[-(\frac{y-\mu}{\sigma})]\mathbf{1}_{[\mu;\infty)}(y)$
Laplace: $La(\mu, \sigma)$	$\frac{1}{2\sigma}\exp[-\frac{\lvert y-\mu\rvert}{\sigma})]$

1.2 Verosimilitud y Cantidades Asociadas

El concepto de verosimilitud es crucial para la inferencia estadística en el paradigma de Fisher, quien fue el primero en delinearlo de manera específica y unívoca en 1912. Fisher se dedicó a estudiar sus propiedades y conceptos asociados, como "información" o "suficiencia", aumentando con ello la producción en Estadística Matemática durante varias décadas. Por ello, cabe aclarar que en esta sección sólo presentamos un breve resumen de algunos conceptos y resultados que serán posteriormente referenciados a lo largo del presente trabajo.

Esencialmente, la *verosimilitud* es un indicador de la coherencia entre los modelos estadísticos elegidos para analizar los datos y las observaciones. Da una idea de cuán *verosímiles* o plausibles son las observaciones obtenidas, es decir, mide la "probabilidad" de re-observar la muestra en una hipotética repetición del experimento.

Sea $Y = (Y_1, ...Y_n)$ un vector aleatorio, con densidad conjunta dada por $f(y; \psi)$, siendo ψ un vector de parámetros que toma valores en un conjunto abierto y no vacío $\Psi \subseteq \mathbb{R}^d$. La *Función de Verosimilitud* es una función de ψ para cada valor muestral $y = (y_1, ..., y_n)$,

$$L = L(\psi) = L(y; \psi) = f(y; \psi).$$

La función de verosimilitud brinda la información resumida acerca de ψ, basada en el modelo estadístico y en los datos observados. Usualmente es conveniente trabajar con el logaritmo natural de $L(\psi)$, la denominada *Función de Log-Verosimilitud*,

$$l = l(\psi) = l(y; \psi) = \log[L(y; \psi)].$$

En particular, si el vector de observaciones se obtiene a través de un muestreo

aleatorio simple de una densidad univariada $f(y;\psi)$, entonces:

$$L(y;\psi) = \prod_{i=1}^{n} f(y_i;\psi) \quad ; \quad l(y;\psi) = \sum_{i=1}^{n} log\left[f(y_i;\psi)\right].$$

A partir de ahora asumiremos que la función de log-verosimilitud es convenientemente regular, en el sentido que tiene derivadas parciales respeto de ψ hasta el orden requerido y que sus momentos son finitos. Denotaremos a las derivadas de $l(\psi)$ respecto de una componente r del vector de parámetros ψ mediante $l_r(\psi;y)$; por ejemplo, $l_{rr}(\psi;y)$ denota la derivada segunda respecto de ψ_r. Las dos primeras derivadas de $l(\psi)$ constituyen los componentes de dos conceptos básicos en la inferencia basada en el análisis de verosimilitud, que definimos a continuación.

Definición 1.2.1. El *Vector de Score*, es el vector de derivadas parciales de $l(\psi)$ con respecto a ψ,

$$u(\psi) = \frac{\partial l(\psi)}{\partial \psi} = (l_1, ..., l_d).$$

El vector de score cumple la propiedad $E_\psi\{u(\psi)\} = 0$, es decir, su primer momento es cero.

Definición 1.2.2. La *Matriz de Información de Fisher* también denominada *Matriz de Información Esperada*, es la varianza del vector de score y, como tal, es una matriz definida no-negativa:

$$i(\psi) = Var_\psi\{u(\psi)\} = E_\psi\{u(\psi)^2 - E_\psi\{u(\psi)\}\} = E_\psi\{(\frac{\partial l(\psi)}{\partial \psi})^2\}.$$

Bajo ciertas condiciones de regularidad que nosotros asumimos, también puede ser expresada alternativamente como $i(\psi) = E_\psi\{j(\psi)\}$. Es decir, el valor esperado de

la *Matriz de Información Observada* también denominada *Matriz Hessiana*, definida por

$$j(\psi) = -\frac{\partial^2 l(\psi)}{\partial \psi^2} = \begin{pmatrix} -l_{11} & \dots & -l_{1d} \\ . & . & . \\ -l_{d1} & \dots & -l_{dd} \end{pmatrix},$$

y por consiguiente, la matriz compuesta por las segundas derivadas de la función de log-verosimilitud respecto de ψ cambiadas de signo.

La notación i_{rs} o j_{rs} hará referencia al elemento (r, s) de la matriz correspondiente. En tanto que, cuando consideremos las inversas de estas matrices, utilizaremos la notación i^{rs} o j^{rs} respectivamente para indicar uno de sus elementos.

Definición 1.2.3. El *Estimador de Máxima Verosimilitud (EMV)* de ψ, usualmente denotado por $\widehat{\psi}$, se define como el valor que maximiza la función de verosimilitud $L(\psi)$ o equivalentemente la de log-verosimilitud $l(\psi)$. Es decir, $\widehat{\psi}$ es un valor en el espacio de parámetros $\Psi \subseteq \mathbb{R}^d$, para el cual $\hat{L} = L(\widehat{\psi}) = Sup_{\psi \in \Psi} L(\psi)$.

La notación $\widehat{(\ )}$, indicará que la función o cantidad específica se evalúa en el EMV $\hat{\psi}$. Un ejemplo relevante es $\widehat{j} = j(\hat{\psi})$, que se denomina *Precisión Observada.*

Si Ψ es un conjunto abierto, el modelo es suficientemente regular y el EMV $\hat{\psi}$ es finito, entonces puede obtenerse como solución de las *Ecuaciones de Verosimilitud*, $\frac{\partial l(\psi)}{\partial \psi} = 0$. Muchas veces no es posible resolver de manera analítica estas ecuaciones y por ello existen una amplia variedad de métodos numéricos (Newton-Raphson por ejemplo) para aproximar el EMV satisfactoriamente.

Tanto $L(\psi)$ como $l(\psi)$ no dependen de la parametrización escogida para el modelo estadístico considerado, que denotaremos como $\mathcal{F}$. Una reparametrización del modelo es una función, $\omega : \Psi \longrightarrow \Omega$, uno a uno e infinitamente diferenciable, que aplica a cada $\psi \in \Psi$ en $\omega(\psi) \in \Omega$. Como ψ y $\omega(\psi)$ identifican el mismo elemento de $\mathcal{F}$ se

cumple que $L^{\Omega}(\omega) = L^{\Psi}(\psi(\omega))$, y sólo diferirán por el determinante del jacobiano de la transformación. Como consecuencia de ello, el estimador de máxima verosimilitud se dice que es *equivariante bajo reparametrizaciones*, es decir, $\hat{\omega} = \omega(\hat{\psi})$ y $\hat{\psi} = \psi(\hat{\omega})$.

La función de verosimilitud ofrece un resumen de los datos bajo un modelo estadístico. Sin embargo, muchas veces para extraer información sobre el parámetro ψ a partir de la muestra esto se realiza a través de estadísticos o combinantes.

Consideramos *estadístico* a cualquier función $T = T(Y)$, usualmente seleccionada para reducir la dimensión de Y haciendo una partición del espacio muestral; un estadístico T se dice *suficiente* para ψ si la distribución condicional de Y dado T es independiente de ψ. En particular, si un estadístico suficiente T de dimensión k puede re-expresarse mediante una transformación diferenciable uno a uno como $(\hat{\psi}, a)$, siendo a un estadístico y $\hat{\psi}$ el EMV, diremos entonces que el estadístico a de dimensión $d - k$ es *auxiliar*; si además, su distribución no depende de ψ, diremos entonces que a es un *estadístico "ancillary" o complementario.*

Llamaremos *combinante* a cualquier función $q = q(Y, \psi)$ que depende de los datos y del modelo estadístico a través del parámetro ψ. En particular, si la distribución del combinante es independiente de ψ para todo $\psi \epsilon \Psi$, entonces se denomina *Pívot.*

Estos conceptos son importantes en los procesos de inferencia. Por ejemplo, como veremos después, los estadísticos complementarios son muy utilizados en los procesos de reducción de datos a través de la marginalización o condicionamiento de la función de verosimilitud. Por otro lado, si ψ es un escalar y tenemos un pívot que es una función monótona de ψ, entonces es posible calcular intervalos de confianza exactos para ψ, simplemente estudiando la distribución de $q(y, \psi)$.

En este sentido, un resultado muy útil y del cual puede encontrarse una demostración detallada en Antle y Bain (1969) es el siguiente:

Teorema 1.2.1. *Sea $y_1, ..., y_n$ una muestra aleatoria simple de un modelo de localización y escala con densidad (1.1.1), tal que $L(y; \mu, \sigma) = \frac{1}{\sigma} \prod_{i=1}^{n} f_0(\frac{y_i - \mu}{\sigma})$. Entonces los EMV $\hat{\mu}$ y $\hat{\sigma}$ tienen la propiedad de que los combinantes $\frac{\hat{\mu}-\mu}{\sigma}$, $\frac{\hat{\sigma}}{\sigma}$ y $\frac{\hat{\mu}-\mu}{\hat{\sigma}}$ son cada uno de ellos un pívot, es decir, su distribución no depende de los parámetros de localización μ y escala σ.*

Este resultado puede hacerse extensivo ante la presencia de un parámetro extra (ver Eastman y Bain, 1974) de la siguiente manera:

Teorema 1.2.2. *Sea $y_1, ..., y_n$ una muestra aleatoria simple de un modelo de localización y escala con un parámetro extra θ, con densidad dada por $f(y; \mu, \sigma, \theta) = \frac{1}{\sigma} f_0(\frac{y-\mu}{\sigma}; \theta)$. Entonces, las distribuciones de $\frac{\hat{\mu}-\mu}{\sigma}$, $\frac{\hat{\sigma}}{\sigma}$, $\frac{\hat{\mu}-\mu}{\hat{\sigma}}$ y $\hat{\theta}$ dependen sólo de n y θ.*

En este caso, estos combinantes no son realmente pívots pues sus distribuciones dependen del parámetro extra θ. Sin embargo, debido a que permanecen independientes de los parámetros de localización y escala, a partir de ahora los denominaremos *Pseudo-Pívots.*

Otro concepto menos estricto que pívot es el llamado *Pívot de Primer-Orden* (ver Pace y Salvan, 1997), que se da cuando sólo su esperanza es independiente del parámetro ψ. Ejemplos de pívots de primer-orden son el vector de score $u(\psi)$ o las "unbiased estimating functions" esto es, combinantes para los cuales $E_\psi\{q(y; \psi\} = 0$ para todo $\psi \in \Psi$.

Un aspecto a tener en cuenta en la estimación de parámetros es qué tipo de propiedades deseamos que tengan los estimadores que utilizaremos. A lo largo de este trabajo prestaremos especial atención a las siguientes:

- *Insesgado.* Un estimador ψ^* se dice insesgado si su valor esperado es igual al verdadero parámetro ψ, $E\{\psi^*\} = \psi$; en caso contrario, diremos que el estimador es sesgado con sesgo dado por $E\{\psi^* - \psi\}$.

- *Consistente.* Un estimador ψ^* es consistente si converge en probabilidad al verdadero parámetro ψ, $\lim_{n\to\infty} P[|\psi^* - \psi| \geq \varepsilon] = 0$, $\forall \varepsilon > 0$. Observemos que ésta es una propiedad asintótica.

- *Eficiente.* Un estimador es eficiente cuando su varianza iguala a la *cota de Cramer-Rao.*

Teorema 1.2.3. *Cota de Cramer-Rao o Desigualdad de Información. Sea ψ^* un estimador de $\psi \in \Psi \subseteq \mathbb{R}^d$ con $E\{\psi^{*2}\} < \infty$, para el cual la derivada de $E\{\psi^*\}$ respecto de cada componente ψ_r, $r = 1, ..., d$, existe y puede obtenerse diferenciando bajo el símbolo de la integral. Entonces,*

$$Var\{\psi_r^*\} \geq \alpha^T \, i^{-1}(\psi) \, \alpha \,, \quad donde \quad \alpha_r = \frac{\partial E\{\psi^*\}}{\partial \psi_r}.$$

En particular, si el estimador es insesgado la cota de Cramer-Rao se reduce simplemente a $Var\{\psi_r^*\} \geq i^{rr}(\psi) = [i^{-1}(\psi)]_{rr}$, es decir la varianza del r-ésimo estimador ψ_r^* no puede ser menor al correspondiente elemento de la diagonal de la inversa de la matriz de información de Fisher.

Nota 1.2.1. La aplicación de la cota de Cramer-Rao requiere que se satisfagan ciertas condiciones de regularidad que asumiremos. En términos generales, se trata de condiciones de tipo técnico impuestas sobre la función de verosimilitud, y que aseguran que

el teorema del límite de Lindberg-Levy se pueda aplicar sobre las muestras del vector aleatorio $\frac{\partial l(\psi)}{\partial \psi}$. Entre otras, se requiere que existan los momentos, al menos de tercer orden de la v.a. subyacente Y, y que su soporte no dependa de los parámetros. Para más detalles ver, por ejemplo, Ferguson (1996).

La cota de Cramer-Rao (CCR) permite definir la máxima precisión de cualquier procedimiento de estimación. Si existe un estimador eficiente, es decir un estimador tal que su varianza iguala a la CCR, ningún otro estimador tendrá menor varianza. Puede demostrarse (ver Cramér, 1946) que si existe un estimador eficiente, éste es el estimador de máxima verosimilitud. Sin embargo en muchos problemas prácticos no existe un estimador eficiente. En estos casos la CCR permanece como una cota inferior.

No obstante, la importancia del EMV no se debe sólo a esta estrecha relación con la CCR sino que también posee propiedades asintóticas deseables en el contexto de cualquier problema, como veremos en la próxima sección.

1.3 Teoría Asintótica

Los procedimientos usados en la inferencia estadística están basados en el conocimiento de la distribución de los estadísticos involucrados. Sin embargo, en la mayoría de los problemas reales, no se dispone de las distribuciones exactas en una forma simple para ser usadas directamente.

Por ello, desarrollos recientes en la teoría estadística han recaído en las aproximaciones asintóticas. Por ejemplo, algunos modelos pueden definirse en base a consideraciones asintóticas como la normalidad unida al Teorema Central del Límite, o el uso de métodos asintóticos para encontrar el procedimiento estadístico óptimo.

Lo cierto es que, cuando la variable o el parámetro toman valores grandes o están próximos a un valor particular, el problema se simplifica y los resultados "límites" pueden usarse como aproximación a la solución original.

Por lo general las aproximaciones utilizadas son resultado de una apropiada combinación entre las técnicas asintóticas del análisis, como el desarrollo de expansiones asintóticas, y la teoría de la probabilidad.

Las llamadas aproximaciones de *Primer-Orden* son las determinadas en su mayoría por la aplicación del Teorema Central del Límite (TCL). A menudo, estas aproximaciones pueden ser mejoradas incorporando términos de orden-superior en las expansiones asintóticas, lo que se conoce en la literatura como *High-Order Asymptotics.* A pesar de que las aproximaciones serán usualmente válidas cuando el tamaño muestral tienda a infinito, esto no implica que la teoría asintótica no sea relevante para muestras de reducido tamaño. De hecho, en muchos casos las aproximaciones obtenidas pueden ser muy precisas aún para pequeñas muestras.

En la teoría asintótica es importante poder expresar cuán cerca está la aproximación utilizada del resultado exacto. En este sentido, la notación estándar de Mann-Wald (1943) es la que más se utiliza para denotar el orden de la magnitud de la aproximación asintótica. Concretamente, se dice que una sucesión de variables aleatorias Y_n es *asintóticamente de orden* $\mathbf{o}_p(a_n)$, si $(Y_n/a_n) \longrightarrow 0$ en probabilidad cuando n tiende a infinito. Mientras que, si $|Y_n/a_n|$ está acotado en probabilidad cuando n tiende a infinito, es decir, dado $\varepsilon > 0$, $\exists\ k \neq 0$ y $\exists n_0$ tal que $\forall n > n_0$ $Pr\{|Y_n/a_n| < k\} > 1 - \varepsilon$, entonces diremos que Y_n es *asintóticamente de orden* $\mathcal{O}_p(a_n)$.

1.3.1 Verosimilitud y Teoría Asintótica de Primer-Orden

El análisis asintótico para inferencias basado en la verosimilitud es un área que ha tenido considerables desarrollos en los últimos 30 años. Por esta razón, cabe aclarar que en este punto sólo resumiremos algunos de los principales resultados. Para un tratamiento en mayor profundidad ver por ejemplo, Cox y Hinkley (1974) o Barndorff-Nielsen y Cox (1994).

Una de las razones para la popularidad de los métodos de máxima verosimilitud es la existencia de aproximaciones generales y simples de la distribución de los estadísticos asociados. Los tres estadísticos de mayor interés en el presente trabajo son, el vector de score, el estimador de máxima verosimilitud y el estadístico del cociente de verosimilitud, cuyos resultados asintóticos más relevantes presentaremos a continuación.

Sean $Y_1, Y_2, ..., Y_n$ variables aleatorias iid con densidad $f(y;\psi)$, siendo $\psi \in \Psi \subseteq \mathbb{R}^d$ el vector de parámetros, y asumamos que se satisfacen las condiciones usuales de regularidad sobre $f(y;\psi)$. Supongamos además, que estamos interesados en realizar inferencia sobre todas las componentes de ψ (la inferencia en presencia de parámetros "nuisance" se discutirá en el próximo punto).

• El *vector de score* $u(\psi)$ se distribuye asintóticamente como una normal con media cero y matriz de varianza-covarianza igual a la matriz de información de Fisher, esto es

$$\sqrt{n}u(\psi) \longrightarrow^{D} \quad N_d(\mathbf{0}, i(\psi)).$$

Este conocido resultado se obtiene aplicando el TCL al vector de score expresado como una suma $u(\psi) = \sum_{j=1}^{n} u_j$, donde cada uno de sus términos $u_j = \partial \ln f(y_j;\psi)/\partial\psi$ tiene un valor esperado nulo, $E\{u_j\} = 0$.

La expansión de la expresión $u(\hat{\psi}) = 0$, es el punto de partida usual para establecer las propiedades asintóticas de primer orden del EMV y otros estadísticos basados en la función de verosimilitud.

• El *Estimador de Máxima Verosimilitud* $\hat{\psi}$ tiene una distribución asintótica normal con media ψ y matriz de varianza-covarianza igual a la inversa de la matriz de información de Fisher

$$\sqrt{n}(\hat{\psi} - \psi) \longrightarrow^{D} \quad N_d(\mathbf{0}, i^{-1}(\psi)).$$

Cramer (1946), presentó la primera demostración rigurosa de este resultado para el caso en que ψ es un escalar (d=1). El consideró que para cualquier solución consistente $\hat{\psi}$ de la ecuación $u(\psi) = 0$, esto es $u(\hat{\psi}) = \frac{\partial l(\hat{\psi})}{\partial \psi} = 0$ y $\hat{\psi} - \psi = \mathbf{o}_p(1)$, se puede llevar a cabo el desarrollo de Taylor de segundo orden en torno de a ψ del tipo,

$$0 = \frac{\partial l(\hat{\psi})}{\partial \psi} = \frac{\partial l(\psi)}{\partial \psi} + (\hat{\psi} - \psi)\frac{\partial^2 l(\hat{\psi})}{\partial \psi^2} + \mathcal{E}(\hat{\psi}, \psi).$$

A partir de aquí se obtiene la siguiente expresión equivalente,

$$\sqrt{n}(\hat{\psi} - \psi) = \sqrt{n}u(\psi)j^{-1}(\psi) - \sqrt{n}j^{-1}(\psi)\mathcal{E}(\hat{\psi}, \psi),$$

donde $\mathcal{E}(\hat{\psi}, \psi)$ es una función que no afecta al comportamiento asintótico de la parte izquierda de la igualdad y por tanto puede despreciarse. Entonces, dado que $j^{-1}(\psi)$ converge en probabilidad a $i^{-1}(\psi)$ y $u(\psi)$ lo hace en distribución a una $N(0, i(\psi))$, resulta que $\sqrt{n}(\hat{\psi} - \psi)$ converge en distribución a una $N(0, i^{-1}(\psi))$.

Posteriormente, Chanda (1954) y Doss(1962,1963) generalizaron esta demostración para al caso en que ψ es un vector de parámetros con dimensión $d > 1$.

Este resultado nos asegura que $lim_{n\to\infty} E\{\hat{\psi}\} = \psi$, es decir que el EMV es al menos *asintóticamente insesgado.* Observemos que su matriz de varianza-covarianza

asintótica es la inversa de la matriz de información de Fisher, $Avar(\hat{\psi}) = i^{-1}(\psi)$. Con lo cual, la varianza asintótica de cada componente $\hat{\psi}_r$ con $r = 1, ..., d$ alcanza la cota de Cramer-Rao y, por tanto, diremos que el EMV es *asintóticamente eficiente.*

En este sentido, muchas veces será de utilidad comparar la eficiencia asintótica de cualquier otro estimador respecto del EMV. Para ello, utilizaremos la *Eficiencia Asintótica Relativa* definida como

$$ARE(\psi_r^*, \hat{\psi}_r) = Avar(\hat{\psi}_r)/Avar(\psi_r^*),$$

donde $\hat{\psi}_r$ es el EMV y ψ_r^* cualquier otro estimador del r-ésimo parámetro ψ_r con $r = 1, ..., d$. Observemos que sólo si la ARE es igual a 1, es decir $Avar(\hat{\psi}_r) = Avar(\psi_r^*)$, el estimador ψ_r^* será también asintóticamente eficiente.

Otra propiedad asintótica importante del EMV es que es un estimador consistente, esto es $\hat{\psi}$ converge en probabilidad a ψ.

• El *Estadístico del Test de la Razón de Verosimilitud*

$$W(\psi) = 2\{l(\hat{\psi}) - l(\psi)\} \tag{1.3.1}$$

se distribuye asintóticamente como una χ^2 con d (la dimensión de ψ) grados de libertad. Lo mismo sucede con sus versiones asintóticamente equivalentes,

Estadístico del test de Score $\quad W_u(\psi) = u(\psi)^T \ i^{-1}(\psi) \ u(\psi)$

Estadístico del test de Wald $\quad W_e(\psi) = (\hat{\psi} - \psi)^T \ i(\psi) \ (\hat{\psi} - \psi)$

Si ψ es un escalar ($d = 1$), suele ser de utilidad considerar las versiones unilaterales de W, W_u y W_e que presentan en cada caso una distribución asintótica $N(0, 1)$. Estas versiones unilaterales se definen respectivamente como,

$$Z(\psi) = sig(\hat{\psi} - \psi)\sqrt{W(\psi)} \tag{1.3.2}$$

$$Z_u(\psi) = u(\psi)/\sqrt{i(\psi)}$$

$$Z_e(\psi) = (\hat{\psi} - \psi)\sqrt{i(\psi)}.$$

Cualquiera de estos estadísticos puede utilizarse para contrastar la hipótesis nula $H_0 : \psi = \psi_0$, o construir un intervalo de confianza para ψ. Es importante destacar además que, las distribuciones asintóticas no se ven afectadas por la sustitución de $i(\psi)$ por $i(\hat{\psi})$ o por $j(\hat{\psi})$.

1.3.2 Pseudo-Verosimilitud

Existen situaciones en las que la aplicación de los procedimientos de verosimilitud presentan serias dificultades. Por ejemplo, pueden existir problemas para expresar la función de verosimilitud como es el caso de algunos modelos de series temporales o modelos espaciales, o bien para obtener el EMV en presencia de parámetros *nuisance*, es decir, parámetros que no tienen un interés principal.

Este tipo de dificultades ha llevado al desarrollo de la noción de la *Pseudo-Likelihood* que, como se desprende de su denominación, es una función de los datos que se comporta, en algunos aspectos, como si fuese una genuina función de verosimilitud y puede por tanto usarse para similares propósitos. Este concepto involucra una amplia gama de procedimientos con idéntico espíritu y que sólo se diferencian por el problema que deba solucionarse o las particularidades de construcción. Para más detalles ver por ejemplo, Cox y Hinkley (1974), Barndorff-Nielsen y Cox (1989), Hinkley et al. (1991) o Pace y Salvan (1997).

Casos particulares de pseudo-verosimilitud son, por mencionar algunas, las funciones de *Marginal-Likelihood*, *Conditional-Likelihood*, *Partial-Likelihood*, *Quasi-Likelihood*, o la *Profile-Likelihood* que utilizaremos en el presente trabajo y, siguiendo a Pace y

Salvan (1997), detallamos a continuación.

Sea $Y = (Y_1, ...Y_n)$ un vector aleatorio con densidad conjunta dada por $f(y; \psi)$ con ψ un vector de parámetros que toma valores en un conjunto abierto y no vacíoΨ $\subseteq \mathbb{R}^d$. Supongamos además que $\psi = (\eta, \lambda)$ se divide en dos subvectores, el parámetro de interés η con dimensión d_0 y el parámetro "nuisance" λ de dimensión $d - d_0$.

En este caso, podemos considerar las particiones del vector de score $u(\psi) = (l_\eta, l_\lambda)$, de la matriz de información observada $j(\psi) = -\begin{pmatrix} l_{\eta\eta} & l_{\eta\lambda} \\ l_{\lambda\eta} & l_{\lambda\lambda} \end{pmatrix}$ y de la matriz de información de Fisher $i(\psi) = \begin{pmatrix} i_{\eta\eta} & i_{\eta\lambda} \\ i_{\lambda\eta} & i_{\lambda\lambda} \end{pmatrix}$; aquí $l_\cdot$ y $l_{\cdot\cdot}$ indican las primeras y segundas derivadas de $l(\psi)$ respectivamente.

En este contexto, una idea general para definir una pseudo-verosimilitud para la componente η, consiste en substituir el parámetro nuisance por un estimador consistente en la función de verosimilitud original.

Definición 1.3.1. La función de *Profile Likelihood* para η es la función

$$L_p = L_p(\eta) = L(\eta, \hat{\lambda}_\eta),$$

donde $\hat{\lambda}_\eta$ es el EMV de λ del modelo restringido para un valor prefijado de η. Así pues, de manera natural definimos la *Profile Log-Likelihood* como

$$l_p = l_p(\eta) = \log L_p(\eta) = \log L(\eta, \hat{\lambda}_\eta).$$

La función L_p puede ser pensada y usada como si fuese una verdadera función de verosimilitud. Podemos utilizar L_p para realizar inferencias sobre η, especialmente cuando la parte complementaria del modelo y los datos contienen poca o ninguna información sobre η.

En particular, el *profile-EMV* de η es igual al EMV general $\hat{\eta}$ basado en $L(\eta, \lambda)$ esto es, $L(\hat{\eta}, \hat{\lambda}_\eta) = \sup_\eta L(\eta, \hat{\lambda}_\eta) = \sup_\eta L_p(\eta) = L_p(\hat{\eta})$. Además, $(\hat{\eta}, \hat{\lambda}_\eta)$ se distribuye asintóticamente como una normal con vector de medias (η, λ) y matriz de varianza-covarianza dada por la inversa de la matriz de información de Fisher, que de acuerdo con las reglas de inversión de matrices particionadas resulta

$$i^{-1}(\psi) = \begin{pmatrix} i^{\eta\eta} & i^{\eta\lambda} \\ i^{\lambda\eta} & i^{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} (i_{\eta\eta} - i_{\eta\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\eta})^{-1} & -i_{\eta\eta}^{-1} i_{\eta\lambda} (i_{\lambda\lambda} - i_{\lambda\eta} i_{\eta\eta}^{-1} i_{\eta\lambda})^{-1} \\ -i_{\lambda\lambda}^{-1} i_{\lambda\eta} (i_{\eta\eta} - i_{\eta\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\eta})^{-1} & (i_{\lambda\lambda} - i_{\lambda\eta} i_{\eta\eta}^{-1} i_{\eta\lambda})^{-1} \end{pmatrix}.$$

Esto implica que $(\hat{\eta}, \hat{\lambda}_\eta)$ preserva las propiedades asintóticas del EMV, es decir que será asintóticamente eficiente e insesgado.

Alternativamente, la varianza asintótica del profile-EMV de η puede expresarse en términos de la varianza asintótica de $\hat{\lambda}_\eta$ como

$$i^{\eta\eta} = i_{\eta\eta}^{-1} + i_{\eta\eta}^{-1} i_{\eta\lambda} i^{\lambda\lambda} i_{\lambda\eta} i_{\eta\eta}^{-1}. \tag{1.3.3}$$

En esta última expresión, $i_{\eta\eta}^{-1}$ representa la varianza asintótica de $\hat{\eta}$ que podría obtenerse si λ fuese conocido y, por tanto, el término de la derecha es la mínima varianza asintótica adicional en la estimación de η si λ es desconocido, sólo alcanzada cuando los EMV $\hat{\eta}$ y $\hat{\lambda}$ se calculan conjuntamente (ver Parke, 1986).

Otro tanto ocurre con la mayoría de la cantidades asociadas a la verosimilitud por ejemplo, dentro de las que consideramos más relevantes, tenemos las siguientes:

- *Profile Vector de Score:*

$$u_p(\eta) = \frac{\partial}{\partial \eta} l_p(\eta) = \frac{\partial}{\partial \eta} l(\eta, \hat{\lambda}_\eta) = l_\eta(\eta, \hat{\lambda}_\eta) + l_\lambda(\eta, \hat{\lambda}_\eta) \frac{\partial}{\partial \eta} \hat{\lambda}_\eta = l_\eta(\eta, \hat{\lambda}_\eta),$$

  donde la última igualdad se debe a que $l_\lambda(\eta, \hat{\lambda}_\eta) = 0$. Lamentablemente, su valor esperado no es nulo, como sucede con el vector de score genuino; en algunos casos regulares $E\{u_p(\eta)\}$ es asintóticamente nula con error de orden $\mathcal{O}_p(1)$.

- *Profile Matriz de Información Observada*

$$j_p(\psi) = -\frac{\partial^2}{\partial \eta^2} l_p(\eta) = -\frac{\partial^2}{\partial \eta^2} l(\eta, \hat{\lambda}_\eta) = -(l_{\eta\eta} - l_{\eta\lambda} l_{\lambda\lambda}^{-1} l_{\lambda\eta})$$

donde todas las derivadas son evaluadas en $(\eta, \hat{\lambda}_\eta)$. La última expresión se deduce a partir del hecho que la matriz de las segundas derivadas parciales puede expresarse como

$$\frac{\partial^2}{\partial \eta^2} l_p(\eta) = l_{\eta\eta}(\eta, \hat{\lambda}_\eta) + l_{\eta\lambda}(\eta, \hat{\lambda}_\eta) \frac{\partial}{\partial \eta} \hat{\lambda}_\eta$$

con $\frac{\partial \hat{\lambda}_\eta}{\partial \eta} = -(l_{\lambda\lambda}(\eta, \hat{\lambda}_\eta))^{-1} l_{\lambda\eta}(\eta, \hat{\lambda}_\eta)$, que se obtiene derivando la ecuación $l_\lambda(\eta, \hat{\lambda}_\eta) = 0$ respecto de η.

Observemos entonces, que $(j_p(\eta))^{-1} = -(l_{\eta\eta} - l_{\eta\lambda} l_{\lambda\lambda}^{-1} l_{\lambda\eta})^{-1} = j^{\eta\eta}$, donde esta última representa el bloque superior izquierdo de la partición de $j^{-1}(\psi)$, la inversa de la matriz de información observada.

- El *Profile Estadístico de la Razón de Verosimilitud*

$$W_p = 2\{l_p(\hat{\eta}) - l_p(\eta)\} = 2\{l(\hat{\eta}, \hat{\lambda}_\eta) - l(\eta, \hat{\lambda}_\eta)\}.$$

Coincide con W para la hipótesis sobre ψ especificada por un valor de η fijo. Por tanto, bajo las condiciones de regularidad usuales, se distribuirá asintóticamente como una $\chi^2_{d_0}$.

Además, si η es un escalar, la versión unilateral de W_p resulta

$$Z_p(\eta) = Sig(\hat{\eta} - \eta)\sqrt{2(l_p(\hat{\eta}) - l_p(\eta))},$$

con una distribución asintótica normal estándar.

La precisión inferencial puede no ser enteramente satisfactoria, especialmente cuando el número de parámetros nuisance es grande. Es por esto que en los últimos años se han propuesto varias modificaciones con el objeto de mejorar las propiedades asintóticas de la profile-verosimilitud. En Barndorff-Nielsen (1994), por ejemplo, pueden encontrarse muchas de estas propuestas basadas en aproximaciones asintóticas de orden superior.

Propiedades similares a las de la profile-likelihood se conservan en las situaciones en las que el parámetro nuisance es reemplazado por otro estimador λ^*, que no coincide con el EMV $\hat{\lambda}_\eta$ obtenido del modelo restringido para un valor prefijado de η.

Este es el caso del *Pseudo-EMV* presentado por Gong y Samaniego (1981) del cual se espera que tenga buenas propiedades asintóticas si las posee el estimador λ^*. De hecho, bajo ciertas condiciones de regularidad bastante estándares y fáciles de chequear, los autores demuestran que:

a) si λ^* es consistente entonces el pseudo-EMV $\hat{\eta}$ también lo es;

b) si el estimador λ^* es $\sqrt{n}$-consistente y asintóticamente normal entonces la distribución asintótica del pseudo-EMV es normal;

c) si λ^* es asintóticamente eficiente entonces el pseudo-EMV será también un estimador asintóticamente eficiente.

Sin embargo, estos autores se restringen al problema de dos parámetros es decir, consideran a η y λ como escalares. Parke (1986), con una notación más sencilla, extiende este resultado al caso multiparamétrico y bajo las mismas condiciones de regularidad dadas por Gong y Samaniego, demuestra que la distribución asintótica del pseudo-EMV $(\hat{\eta}, \lambda^*)$ es normal con vector de medias (η, λ) y matriz de varianza-covarianza dada por $\begin{pmatrix} i_{\eta\eta}^{-1} + i_{\eta\eta}^{-1} i_{\eta\lambda} \Sigma_{\lambda\lambda} i_{\lambda\eta} i_{\eta\eta}^{-1} & -i_{\eta\eta}^{-1} i_{\eta\lambda} \Sigma_{\lambda\lambda} \\ -\Sigma_{\lambda\lambda} i_{\lambda\eta} i_{\eta\eta}^{-1} & \Sigma_{\lambda\lambda} \end{pmatrix}$, donde $\Sigma_{\lambda\lambda}$ es la varianza

asintótica del estimador λ^*.

Observemos que ahora, en la varianza asintótica del psedo-EMV de η,

$$\Sigma_{\eta\eta} = i_{\eta\eta}^{-1} + i_{\eta\eta}^{-1} i_{\eta\lambda} \Sigma_{\lambda\lambda} i_{\lambda\eta} i_{\eta\eta}^{-1} \tag{1.3.4}$$

el segundo término representa la varianza asintótica atribuida al hecho de usar el estimador λ^*.

Teniendo presente la Cota de Cramer-Rao, podemos considerar a $\Sigma_{\eta\eta} - i^{\eta\eta} = i_{\eta\eta}^{-1} i_{\eta\lambda} (\Sigma_{\lambda\lambda} - i^{\lambda\lambda}) i_{\lambda\eta} i_{\eta\eta}^{-1}$ como una medida de la "ineficiencia asintótica" del pseudo-EMV de η basado en λ^*. A partir de aquí, puede deducirse que el pseudo-EMV $\hat{\eta}$ resultará asintóticamente eficiente sólo si se satisface que $\Sigma_{\eta\eta} - i^{\eta\eta} = \mathbf{0}$.

Dicho de otro modo, debido a que la matriz de información es definida positiva, sólo si los EMV de η y λ son asintóticamente no correlacionados ($i_{\eta\lambda} = 0$) o bien λ^* es tan eficiente como el EMV de λ ($\Sigma_{\lambda\lambda} = i^{\lambda\lambda}$), entonces el pseudo-EMV de η basado en λ^* será también asintóticamente eficiente.

1.3.3 Teoría Asintótica de Orden Superior (Higher-Order)

Varias técnicas se han desarrollado para mejorar la aproximación de la distribución exacta de los estadísticos. Un método muy conocido es utilizar los primeros términos de la expansión de *Edgeworth*, que es un desarrollo de orden $\mathcal{O}_p(n^{-1/2})$ donde el término principal es la densidad Normal Estándar. Sin embargo, esta aproximación tiende a ser pobre en las colas de la distribución (ver Field y Ronchetti, 1990).

En 1954, H. Daniels introdujo un nueva idea en estadística aplicando las denominadas *Técnicas del Saddlepoint (Punto de Silla)* para obtener una aproximación de la distribución de la media aritmética de una muestra. Es una expansión asintótica de orden $\mathcal{O}_p(n^{-1})$, muy precisa incluso en las colas de la densidad, que funciona

bien aún en el caso de muestras pequeñas. El término principal del desarrollo no es una densidad, por lo que se suele utilizar una versión renormalizada que tiene error relativo $\mathcal{O}_p(n^{-3/2})$.

Un resultado muy importante derivado de la aproximación Saddlepoint, es la fórmula p^* de Barndorff-Nielsen (1980), que aporta una aproximación para la densidad condicional del estimador de máxima verosimilitud $\hat{\psi}$ de la forma,

$$p(\hat{\psi}|a;\psi) = p^*(\hat{\psi};\psi|a)\{1+\mathcal{O}_p(n^{-3/2})\},$$

siendo $p^*(\hat{\psi};\psi|a) = c(a,\psi)|j(\hat{\psi};\hat{\psi},a))|^{1/2}\exp\{l(\psi;\hat{\psi},a)-l(\hat{\psi};\hat{\psi},a)\}$, donde $c(a,\psi)$ es la constante normalizadora, y $a = a(y)$ un estadístico "ancillary" (exacto o aproximado) es decir, tal que su distribución marginal no depende de ψ.

Para el caso particular de las familias de localización y escala, la formula p^* brinda la distribución exacta de $\hat{\psi} = (\hat{\mu},\hat{\sigma})$ condicionada a $a = (a_1,...,a_n)$, donde a es el estadístico de configuración $a_i = \frac{y_i-\hat{\mu}}{\hat{\sigma}}$ para $i = 1,...,n$. Esta coincidencia de p^* con la densidad exacta del EMV condicionada a un estadístico ancillary a, también se da para otros modelos que satisfacen ciertas condiciones de regularidad.

Otra característica importante de la fórmula p^* es que se transforma de manera regular bajo reparametrizaciones, es decir, si $\omega = \omega(\psi)$ es una reparametrización entonces $p^*(\hat{\omega};\omega,a) = p^*(\psi(\hat{\omega});\psi(\omega)|a)|\partial\psi(\hat{\omega})/\partial\hat{\omega}|$.

Sin embargo, las aplicaciones de la fórmula p^* no son usualmente directas. Más bien, es una herramienta general para obtener "refinamientos" específicos de los resultados asintóticos de primer-orden. Estos refinamientos conciernen particularmente a obtener mejores aproximaciones de las distribuciones asintóticas de los estadísticos W y Z. Es decir, se intenta de alguna manera acelerar la convergencia del estadístico

hacia su distribución asintótica. En este sentido, uno de los resultados más destacados es el estadístico Z^* obtenido por Barndorff-Nielsen (1991) mediante la aplicación directa de la fórmula p^*, que a continuación describiremos en qué consiste.

Supongamos que nuestro vector de parámetros $\psi = (\eta, \lambda)$ se divide en el parámetro de interés η un escalar y el vector de parámetros "nuisance" λ de dimensión $d-1$. Supongamos además que estamos interesados en contrastar la hipótesis nula $H_0 : \eta = \eta_0$ contra la alternativa $H_0 : \eta \neq \eta_0$.

Para ello es posible utilizar el estadístico de la razón de verosimilitud W con distribución asintótica χ^2_1 o su equivalente Z con distribución asintótica $N(0,1)$. No obstante, y a fin de mejorar estas aproximaciones asintóticas de primer orden, utilizaremos una transformación simple de Z cuya distribución asintótica está más cercana a la Normal Estándar. Se trata del estadístico Z^* definido como

$$Z^* = Z + \frac{1}{Z}\log(\frac{CU_p}{Z}), \tag{1.3.5}$$

donde C y Up deben calcularse para cada modelo en concreto de la siguiente manera:

$$C = \frac{|\tilde{l}_{\lambda\hat{\lambda}}|}{\{|\tilde{j}_{\lambda\lambda}||\hat{j}_{\lambda\lambda}|\}^{1/2}} \quad ; \quad U_p = -j_p(\hat{\eta})^{-1/2}\frac{\partial}{\partial\hat{\eta}}\{l_p(\eta) - l_p(\hat{\eta})\} \tag{1.3.6}$$

Aquí l indica como siempre la log-verosimilitud, pero expresada como función de los parámetros y de los EMV, esto es $l(\eta, \lambda, \hat{\eta}, \hat{\lambda})$. El tilde ($\backsim$) en la notación, indica que se está evaluando en $(\eta, \tilde{\lambda})$, donde $\tilde{\lambda} = \hat{\lambda}_\eta$ denota el estimador de máxima verosimilitud restringido de λ para un valor prefijado de η; el signo circunflejo ($\wedge$) indica, como siempre que lo utilizamos, la evaluación en el EMV $(\hat{\eta}, \hat{\lambda})$.

En la expresión de C intervienen los determinantes de matrices, todas de orden

$(d-1)$x$(d-1)$, cuyos elementos son las derivadas segundas de la función de log-verosimilitud. Concretamente, los elementos de la matriz del numerador son de la forma $(\tilde{l}_{\lambda\hat{\lambda}})_{rs} = \frac{\partial^2 l}{\partial\lambda_r \partial\hat{\lambda}_s}(\eta, \tilde{\lambda}, \hat{\eta}, \hat{\lambda})$ con $r, s = 1, 2, ..., d-1$. El término $j_{\lambda\lambda}$ del denominador es el bloque inferior derecho de la partición de la matriz de información observada con elemento genérico $-\partial^2 l(\eta, \lambda, \hat{\eta}, \hat{\lambda})/\partial\lambda_r \partial\lambda_s$ con $r, s = 1, 2, ..., d-1$.

La expresión de U_p está íntimamente relacionada con la profile-likelihood y sus cantidades asociadas descritas en el punto anterior. En este caso, $j_p(\hat{\eta})^{-1} = j^{\eta\eta}$ resulta un escalar y representa la estimación de la varianza asintótica de $\hat{\eta}$. La derivada parcial que aparece dentro de la expresión de U_p, debe entenderse como la derivada respecto de $\hat{\eta}$ de la profile log-likelihood l_p, es decir,

$$\frac{\partial}{\partial\hat{\eta}}\{l_p(\eta) - l_p(\hat{\eta})\} = \tilde{l}_{\hat{\eta}} + \frac{\partial\hat{\lambda}}{\partial\hat{\eta}}\tilde{l}_{\hat{\lambda}} - \hat{l}_{\hat{\eta}} - \frac{\partial\hat{\lambda}}{\partial\hat{\eta}}\hat{l}_{\hat{\lambda}},$$

donde $\frac{\partial\hat{\lambda}}{\partial\hat{\eta}} = -\tilde{l}_{\lambda\hat{\eta}}(\tilde{l}_{\lambda\hat{\lambda}})^{-1}$ se obtiene derivando la ecuación $l_\lambda(\eta, \hat{\lambda}_\eta, \hat{\eta}, \hat{\lambda}) = 0$ respecto de $\hat{\eta}$. Por tanto, toda la expresión puede escribirse equivalentemente como

$$\frac{\partial}{\partial\hat{\eta}}\{l_p(\eta) - l_p(\hat{\eta})\} = \{\tilde{l}_{\hat{\eta}} - \hat{l}_{\hat{\eta}} - \tilde{l}_{\lambda\hat{\eta}}(\tilde{l}_{\lambda\hat{\lambda}})^{-1}(\tilde{l}_{\hat{\lambda}} - \hat{l}_{\hat{\lambda}})\}.$$

Aplicaciones particulares de este estadístico Z^* que se distribuye asintóticamente como una $N(0,1)$ con error de orden $\mathcal{O}_p(n^{-3/2})$, serán presentadas en el capítulo 5. Para mayores detalles teóricos ver por ejemplo, Pace y Salvan (1997), Jensen (1993, 1994) o Reid (1996). Una versión equivalente de Z^*, es el método de Fraser y Reid (1995) aplicado también por Wong y Wu (2000) para familias de localización y escala con censura.

1.4 Modelos de Duración y Variables Limitadas

En muchas aplicaciones prácticas puede aparecer restricciones en una proporción de las observaciones de la muestra, ya sea por características intrínsecas de la distribución de la variable a lo que denominamos **Truncamiento**, o bien porque existen observaciones cuyo valor cae en áreas restringidas del espacio muestral, fenómeno conocido como **Censura**.

En esta sección describiremos estos patrones que crean especiales problemas en el análisis de datos. Así mismo, debido a que los datos censurados y/o truncados surgen en diversas situaciones experimentales en los denominados **Modelos de Duración**, presentaremos a continuación algunos conceptos básicos y resultados a los que haremos referencia en otros lugares de esta memoria.

1.4.1 Modelos de Duración

El análisis estadístico de lo que muchos autores se refieren como *tiempo de supervivencia*, *tiempo de vida* o *tiempo de fallo*, ha pasado a constituir un tópico importante para distintas áreas, especialmente en las ciencias biomédicas e ingenierías. El objetivo general es la modelización, a través de variables aleatorias no-negativas, de la duración o el tiempo que transcurre hasta que un determinado evento ocurre (ver Lawless, 1982).

Según sea la naturaleza de los datos tratados, estos modelos básicamente dan lugar a dos disciplinas: el *Análisis de Supervivencia*, que se centra en el estudio de los tiempos de vida de seres humanos y animales; la *Fiabilidad o Control de Calidad*, que lo hace sobre el tiempo de vida útil de aparatos o sistemas mecánicos. A pesar de que cada rama tiene lenguajes conceptuales y notaciones propias, comparten algunos

métodos y técnicas estadísticas que resumiremos a continuación.

Sea Y una variable aleatoria continua y no-negativa, que representa el tiempo de vida de los individuos de cierta población, con densidad f y distribución F.

Definición 1.4.1. La probabilidad de que un individuo sobreviva como mínimo hasta el tiempo t viene dada por la *Función de Supervivencia*

$$\overline{F}(t) = 1 - F(t) = Pr(Y \geq t) = \int_t^{\infty} f(y)dy,$$

también denominada *Función de Fiabilidad (Reliability Function).*

Definición 1.4.2. La *Razón de Fallo (Failure Rate Function)* se define como

$$r(t) = \frac{f(t)}{\overline{F}(t)} = \lim_{\Delta t \to 0} \frac{Pr(t < Y < t + \Delta t | Y > t)}{\Delta t}.$$

Esta función también se conoce como, *Función de Riesgo o Peligro (Hazard Function)*, *Tasa de Riesgo Instantánea* o *Fuerza de Mortalidad.*

Por su definición, la función de riesgo puede interpretarse como la probabilidad de muerte en el intervalo $[t; t + \Delta t]$ (sin importar cuan pequeño sea Δt), suponiendo que el individuo ha sobrevivido hasta el comienzo del intervalo dividida por Δt. El tiempo de vida en el intervalo $[t; t + \Delta t]$, dado que el individuo sobrevivió hasta el tiempo t, se denomina *Tiempo de Vida Residual* en t. Asociada a la función de riesgo $r(t)$, tenemos también la razón de fallo acumulado,

Definición 1.4.3. *Razón de Fallo Acumulado*

$$\Lambda(t) = \int_0^t r(y)dy.$$

Esta función, también conocida como *Función de Riesgo Acumulado* se relaciona con la función de supervivencia de la siguiente forma,

$$\exp\left[-\Lambda(t)\right] = \exp\left[-\int_0^t r(y)dy\right] = \exp\left[-\int_0^t \frac{f(y)}{\overline{F}(y)}dy\right] = \exp\left[\log(\overline{F}(t))\right] = \overline{F}(t),$$

con lo cual se tiene además que $f(t) = r(t)\exp[-\Lambda(t)]$.

Cuando la función de riesgo es monótona, podemos diferenciar dos clases de distribuciones, según $r(t)$ sea creciente o decreciente,

Definición 1.4.4. *Clases IFR y DFR.* Una variable aleatoria Y es de la *Clase IFR (Increasing Failure Rate)* si su razón de fallo $r(t)$ es creciente para todo $t \geq 0$. Por el contrario, si la función de riesgo $r(t)$ es decreciente para todo $t \geq 0$, se dice que Y es de la *Clase DFR (Decreasing Failure Rate).*

También, puede suceder que la función de riesgo sea constante, es decir el paso del tiempo no influye de ninguna manera en la supervivencia. La distribución Exponencial es la única de esta clase, que es a la vez IFR y DFR.

Intuitivamente, una variable es de la clase IFR, si el "riesgo" que tiene el individuo de morir aumenta a medida que pasa el tiempo. Considerando que normalmente esto es lo que sucede, tanto con los seres vivos (envejecimiento) como con los aparatos (deterioro), los modelos IFR son los más frecuentes en las aplicaciones.

Un resultado que permite determinar si una variable es de la clase IFR o no, en términos de su densidad, es el siguiente:

Teorema 1.4.1. *Si Y es una variable aleatoria continua con densidad $f(t)$ y la función $q(t) = \log(f(t))$ es cóncava, entonces Y pertenece a la clase IFR.*

La demostración de este resultado puede encontrarse en Barlow y Proschan (1981). Otro resultado interesante y que nos será de utilidad es,

Lema 1.4.2. *Una variable aleatoria continua Y, con función de densidad $f(t) \in \mathcal{C}^1$ y distribución $F(t)$, es de la clase IFR si y sólo si la función $\Lambda(t) = -log[\overline{F}(t)]$ es convexa para todo $t \geq 0$.*

Demostración.

Y es IFR $\Leftrightarrow r'(t) > 0 \ \forall t \Leftrightarrow \Lambda''(t) > 0 \ \forall t \Leftrightarrow \Lambda(t)$ es convexa para todo t. □

1.4.2 Censura

En las primeras referencias sobre este tema, las muestras censuradas se describían como truncadas, donde el número de observaciones no medidas era conocido. De acuerdo con Cohen (1991), fue J. E. Kerrich el primero que sugirió la denominación *censuradas* para este tipo de muestras. En estos casos, existen observaciones cuyo valor cae en áreas restringidas del espacio muestral, es decir, pueden ser identificadas y contadas pero no medidas.

Situaciones comunes que producen datos censurados son: cuando se realizan mediciones con aparatos que sólo aprecian cantidades inferiores o superiores a un valor fijo; cuando el experimento se realiza hasta un tiempo establecido de antemano; cuando se decide hacer un control de calidad hasta que hayan fallado una cantidad fija de unidades. En cualquier caso, por una razón u otra, no se cuenta con el valor de determinadas observaciones, pero se sabe que estarán por encima o por debajo de cierto límite, al que denominamos *punto de censura.*

Formalmente una observación y_i se dice que está *Censurada por la Derecha (izquierda)* en c_i, si el valor de la observación no es conocido pero se sabe que es mayor (menor) o igual que c_i. Si bien la censura por la derecha es lo más habitual en el análisis de supervivencia, también es concebible que la censura ocurra a la izquierda. Por

ejemplo, en los análisis clínicos el nivel de algunos constituyentes de la sangre puede estar por debajo del que detecta el aparato, aunque se sepa que su valor debe ser mayor que cero (ver Altman, 1991).

Además, puede suceder que $c_i = y_0 \; \forall \; i = 1, ..., N$ es decir, que tengamos un único punto de censura para todas las observaciones. A este caso particular, se le denomina *Censura Simple*. Un ejemplo claro, es el uso de un aparato que mide hasta una cierta cantidad y_0, todas las observaciones que superen ese punto estarán censuradas en y_0.

Por el contrario, si la censura ocurre en distintos puntos, se denomina *Censura Múltiple*. En este caso, cada observación y_i tiene su punto de censura asociado c_i y, en principio, todos los puntos de censura pueden ser diferentes. Un ejemplo, común en investigaciones biomédicas, es cuando los pacientes son dados de alta en distintos momentos del tratamiento, y el investigador conoce que el paciente sobrevivió un tiempo y_i hasta los diferentes tiempos de alta c_i.

También pueden presentarse otros tipos de censura más especiales como, Censura Progresiva, Censura por Intervalos, Censura Aleatoria, etc., que no comentaremos en este trabajo (ver Lawless, 1982).

Sin embargo, en lo que a problemas de inferencia se refiere, la distinción más importante se da en el hecho que la censura sea de Tipo I o de Tipo II. Esto lo detallaremos a continuación.

Censura de Tipo I y II

Algunos experimentos se realizan por un período de tiempo determinado, de tal forma que sólo se tendrá un valor observado si éste es menor que un cierto valor prefijado. En este caso, se dice que los datos presentan *Censura de Tipo I*, o que son observaciones

a *Tiempo Censurado (Time Censored).*

De manera más precisa, una muestra con censura de tipo I surge si $1, 2, ..., N$ unidades experimentales se someten a valores limitados $c_1, c_2, ..., c_N$, de tal forma que la variable aleatoria Y_i sólo será realmente observada si $y_i \leq c_i$.

Una forma conveniente de expresar los datos en este caso es como pares de N variables aleatorias (X_i, ε_i), donde $X_i = min(Y_i, c_i)$ y ε_i es la función indicadora que determina si la observación está censurada $\varepsilon_i = 0$ $(y_i \geq c_i)$, o no $\varepsilon_i = 1$ $(y_i \leq c_i)$.

En la censura de tipo I, los puntos de censura c_i son constantes conocidas, y el número de datos censurados es una variable aleatoria. Por ejemplo, supongamos que deseamos analizar el tiempo hasta el fallo de N unidades durante un período pre-establecido t_0. Al culminar el experimento el número de las unidades que no fallaron (observaciones censuradas) será el valor observado de una variable aleatoria.

La Censura de Tipo II por el contrario, se da cuando el experimento se continúa hasta que hayan fallado una cantidad n, fijada de antemano, de las N unidades consideradas. En este caso se conoce a priori el número de datos censurados, $N - n$.

Formalmente, es el caso en que sólo los n menores valores de una muestra aleatoria de tamaño N son observados $(1 \leq n \leq N)$, los datos obtenidos son $y_{(1)}, y_{(2)}, ..., y_{(n)}$ que corresponden a los estadísticos de orden de la muestra.

1.4.3 Truncamiento

Cuando los datos de la muestra se extraen de un subconjunto de una población de interés, estamos frente al efecto de truncamiento. En este caso, una observación es excluida de la muestra si la variable aleatoria subyacente cae por debajo o por encima de un cierto valor.

Por ejemplo, muchos países tienen un nivel de ingreso mínimo debajo del cual no se deben pagar determinados impuestos. Por lo tanto, el registro de los impuestos presenta una distribución truncada por debajo o izquierda.

La función de densidad de una variable aleatoria Y con distribución $F(y)$, con truncamiento por la izquierda en y_0, viene dada por

$$f(y|y > y_0) = \frac{f(y)}{P(Y > y_0)} = \frac{f(y)}{1 - F(y_0)}.$$

Es decir, se obtiene reescalando la densidad original $f(y)$, ya que debido al truncamiento sólo se tienen en cuenta los valores mayores que y_0.

También pude darse el truncamiento por la derecha o arriba. En este caso los valores que no se consideran son los mayores al umbral y_0 y el tratamiento es análogo al truncamiento por la izquierda. En algunos casos puede ser necesario considerar también un doble truncamiento, es decir, se limitan las observaciones tanto por izquierda ($y_i > y_a$) como por derecha ($y_i < y_b$).

Es importante resaltar algunas propiedades de los dos primeros momentos centrales de las variables truncadas. Por ejemplo, la media de una variable aleatoria con truncamiento por la izquierda en y_0, $E(Y|Y > y_0) = \int_{y_0}^{\infty} y f(y|y > y_0)dy$ es mayor que la media de la variable aleatoria original; resulta a la inversa si el truncamiento es por la derecha. En tanto que, la varianza de la variable truncada, ya sea por debajo o por arriba, es siempre menor que la varianza de la variable original (ver Green, 1999).

La *Distribución Normal Truncada* (NT) es utilizada con frecuencia, especialmente en el ámbito de la Economía. El estudio de las propiedades de una variable aleatoria distribuida normalmente cuando ciertas observaciones están restringidas, ha sido un campo muy fértil con aplicaciones en diversas áreas. En el capítulo 4 volveremos a tratar con más detalle el caso de la distribución Normal Truncada.

1.5 Generación de Muestras Aleatorias

En diversos capítulos del presente trabajo será necesario simular muestras aleatorias provenientes de modelos específicos.

La generación de observaciones de una variable aleatoria X con distribución $F(x)$, no uniforme, se realiza usualmente a través de transformaciones a variables aleatorias $U(0,1)$. Existen diversas técnicas disponibles para simular variables aleatorias. La rapidez unida a la sencillez del método escogido serán criterios importantes para su implementación.

El procedimiento más generalizado es el *método de inversión.* Utiliza el hecho que la variable aleatoria $U = F(X)$ tiene una distribución $U(0,1)$. Asumiendo que $F(x)$ es estrictamente creciente, obtenemos que $X = F^{-1}(U)$.

Sin embargo, su implementación puede llevar a altos costos computacionales si, por ejemplo, el cálculo de $F^{-1}(x)$ debe realizarse a través de métodos numéricos. Por ello, muchas veces utilizaremos el método de *Acceptance-Rejection* (AR), quizás el más fiable para simular observaciones de una v.a. continua X con densidad $f(x)$, que describiremos a continuación.

Asumamos que existe una v.a. Y con una densidad conocida $g(y)$, tal que:

a) sabemos como generar observaciones de manera eficiente; por ejemplo Y se distribuye como una exponencial, normal, etc.

b) existe una constante $M > 0$ tal que $f(x) \leq Mg(x)$, $\forall x$; la densidad $g(y)$ con esta propiedad se denomina *densidad mayorante* (*majorizing density*).

Bajo estos supuestos, el algoritmo para el método Acceptance-Rejection se resume como sigue:

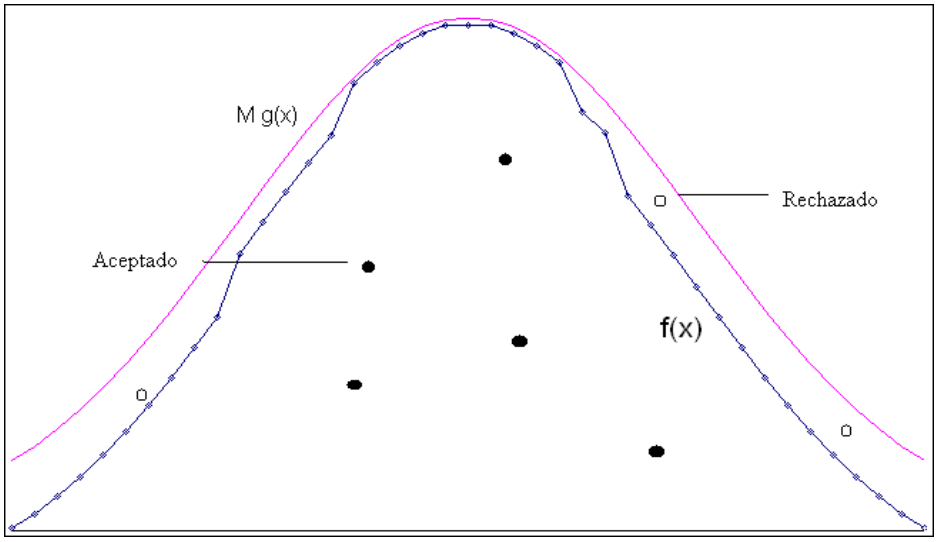

Figura 1.1: Método AR para la simulación de una muestra aleatoria con densidad $f(x)$

1. Generamos una observación y de la v.a. Y con densidad $g(y)$.

2. Generamos una observación u de una v.a. $U(0,1)$ e independiente de Y.

3. Si $u \leq \frac{f(y)}{Mg(y)}$ aceptamos a y como una observación de la v.a. que queremos simular X; de lo contrario volvemos al paso 1.

Es fácil demostrar que la v.a. X generada por este algoritmo tiene la densidad requerida $f(x)$ usando el hecho que $P(X \leq x) = P(Y \leq x | U \leq f(y)/(Mg(y))$ con Y y U independientes.

Más aún, la probabilidad de que un valor de Y sea aceptado es $1/M$. Por tanto, el número N de observaciones y_i necesarias para generar una observación de X sigue una distribución geométrica con media M. Esta constante es, por consiguiente, una medida de la eficiencia del algoritmo.

Entonces, debido a que deseamos que la probabilidad de rechazo sea pequeña, es conveniente escoger M lo más próximo a 1 posible, que es lo mismo que seleccionar $g(y)$ lo más cercana a la densidad requerida $f(x)$. Para más detalles de éste y otros métodos de simulación ver, por ejemplo, Rubistein (1981) o Law y Kelton (2000).

Ejemplo 1.5.1. *Supongamos que deseamos simular el valor absoluto $X = |Z|$, de una variable aleatoria Z con distribución normal estándar. La densidad de X está dada por $f(x) = 2\exp(-x^2/2)/\sqrt{2\pi}$, $x > 0$.*

Consideremos primeramente Y una variable aleatoria con distribución exponencial de parámetro 1, es decir $g(y) = \exp(-y)$, $y > 0$. Entonces

$$\frac{f(y)}{g(y)} = \sqrt{\frac{2}{\pi}} exp(-(y^2 - 2y + 1)/2)e^{1/2} = \sqrt{\frac{2e}{\pi}} exp(-(y-1)^2/2)$$

y tomando $M = \sqrt{2e/\pi}$, el algoritmo resulta:

1. *Generamos un valor $Exp(1)$, y.*
2. *Generamos un valor uniforme en $(0, 1)$, u.*
3. *Si $u \leq \exp(-(y-1)^2/2)$, aceptamos a y como un dato perteneciente a nuestra distribución. Si no, volvemos al paso 1.*

Capítulo 2

Familias de Localización Simétricas

En la mayoría de los modelos de medición se asume que las observaciones satisfacen la relación $x_i = \mu + \epsilon_i$ para $i = 1, 2, ..., n$, donde ϵ_i son errores independientes e idénticamente distribuidos. Usualmente, este error aleatorio se asume como simétrico en cero y, consecuentemente, las observaciones pueden describirse por modelos de localización simétricos.

El parámetro de localización μ puede ser estimado usando los bien conocidos estimadores muestrales robustos como la media, mediana, rango medio, etc., o combinaciones lineales de éstos. Ya en 1818, Laplace estudió como estimar la media poblacional μ de una distribución simétrica a través de una combinación lineal de la media y la mediana muestrales.

En este capítulo, analizamos este tipo de estimadores de μ y caracterizamos a todos los modelos de localización simétricos para los cuales una combinación lineal de la media y mediana muestrales es un estimador asintóticamente eficiente del parámetro de localización. Asimismo, mostramos cómo este resultado puede extenderse fácilmente a otros estimadores, tratando el caso particular del estimador de Hodges-Lehmann.

2.1 Combinación Lineal de la Media y Mediana Muestrales

Debido a que la media muestral $\bar{x}$ y la mediana muestral $\tilde{x}$ son estimadores insesgados de μ y esto es válido para cualquier combinación lineal de la forma $w\bar{x} + (1-w)\tilde{x}$ siendo w una constante, Laplace propuso usar este tipo de estimadores seleccionando w de manera que la varianza asintótica fuese minimizada (ver Stigler, 1973).

Sean $x_1, x_2, ..., x_n$, n observaciones independientes e idénticamente distribuidas de una variable aleatoria X con densidad $f(x-\mu)$ definida sobre $\mathbb{R}$, tal que $f(t)$ es simétrica en cero, siendo $\mu \in \mathbb{R}$ el parámetro de localización. Asumiremos además, que $f(t)$ es continua y positiva en cero.

La distribución conjunta asintótica de $(\bar{x}, \tilde{x})$ es una normal bivariada con media (μ, μ) y matriz de varianza-covarianza dada por

$$\frac{1}{n}\begin{pmatrix} v^2 & \tau/(2f(0)) \\ \tau/(2f(0)) & 1/(4f(0)^2) \end{pmatrix}, \tag{2.1.1}$$

donde $v^2 = E[(X-\mu)^2] = \int_{-\infty}^{\infty} t^2 f(t)dt$ y $\tau = E[|X-\mu|] = \int_{-\infty}^{\infty} |t| f(t)dt$. Este resultado fue obtenido originalmente por Laplace. Una demostración más general involucrando los percentiles muestrales puede encontrarse en Lin et al. (1980). Consecuentemente, el estimador de localización propuesto $\tilde{\mu} = w\bar{x} + (1-w)\tilde{x}$ tiene una distribución asintótica normal con media μ y varianza asintótica dada por

$$Avar(\tilde{\mu}) = n^{-1}(w^2 Var(\bar{x}) + (1-w)^2 Var(\tilde{x}) + 2Cov(\bar{x}, \tilde{x})),$$

que utilizando (2.1.1) puede expresarse como

$$Avar(\tilde{\mu}) = w^2 \frac{v^2}{n} + \frac{w(1-w)\tau}{f(0)n} + \frac{(1-w)^2}{4f(0)^2 n}. \tag{2.1.2}$$

Entonces, para minimizar la varianza asintótica basta con derivar (2.1.2) respecto de w e igualar a cero, o equivalentemente resolver para w la ecuación

$$2wv^2 + \frac{(1-w)\tau}{f(0)} - \frac{w\tau}{f(0)} - 2\frac{(1-w)}{4f(0)^2} = 0.$$

Así obtenemos que el valor de w que minimiza (2.1.2), tal como fue calculado por Laplace, es

$$w = \frac{1 - 2f(0)\tau}{4f(0)^2v^2 - 4f(0)\tau + 1}. \tag{2.1.3}$$

Observemos que si $w = 1$ resulta $\tilde{\mu} = \bar{x}$, indicando que la menor varianza para estimar μ se obtendría utilizando sólo la media. Análogamente, si $w = 0$ entonces $\tilde{\mu} = \tilde{x}$ y por tanto sólo consideraríamos la mediana.

Nota 2.1.1. En los modelos de localización y escala simétricos $f(x;\mu,\sigma) = \frac{1}{\sigma}g(\frac{x-\mu}{\sigma})$, el valor óptimo de w no depende del parámetro de escala σ. De hecho, si tenemos en cuenta que estos modelos son invariantes por cambios de escala y consideramos $v_g^2 = \int_{-\infty}^{\infty} t^2 g(t)dt$ y $\tau_g = \int_{-\infty}^{\infty} |t|g(t)dt$, es inmediato que $v^2 = \sigma^2 v_g^2$, $\tau = \sigma\tau_g$ y $f(0) = g(0)/\sigma$, con lo cual (2.1.1) resulta $\begin{pmatrix} \sigma^2 v_g^2 & \sigma^2\tau_g/(2g(0)) \\ \sigma^2\tau_g/(2g(0)) & \sigma^2/(4g(0)^2) \end{pmatrix}$. Por lo tanto el valor de w que minimiza la $Avar(\tilde{\mu})$ tiene la misma expresión dada en (2.1.3), reemplazando $f()$ por $g()$, y por consiguiente no depende de σ.

2.1.1 Ejemplos

- Distribución Normal: $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sigma}g(\frac{x-\mu}{\sigma})$, con $g(t) = \frac{e^{-t^2/2}}{\sqrt{2\pi}}$.

  Para este caso tenemos $f(0) = \frac{1}{\sigma}g(0) = \frac{1}{\sigma\sqrt{2\pi}}$; $\tau = \sigma\int_{-\infty}^{\infty}|t|\frac{e^{-t^2/2}}{\sqrt{2\pi}}dt = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}$ y $v^2 = \sigma^2\int_{-\infty}^{\infty} t^2\frac{e^{-t^2/2}}{\sqrt{2\pi}}dt = \sigma^2$, con lo cual (2.1.1) puede expresarse como $\begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2\pi/2 \end{pmatrix}$,

y el valor de w que minimiza la varianza es $w = 1$. Por tanto la menor varianza asintótica se obtiene al considerar sólo la media muestral como estimador. Para esta distribución $\bar{x}$ es además el estimador de máxima verosimilitud (EMV) de μ, con lo cual es asintóticamente eficiente. Por lo tanto es evidente que en este caso no podemos obtener una reducción de la varianza asintótica del estimador mediante una combinación lineal con la mediana.

- Distribución de Laplace: $f(x;\mu,\sigma) = \frac{1}{2\sigma}e^{-|\frac{x-\mu}{\sigma}|} = \frac{1}{\sigma}g(\frac{x-\mu}{\sigma})$ con $g(t) = \frac{1}{2}e^{-|t|}$.

  Aquí es $f(0) = \frac{1}{2\sigma}$, $\tau = \sigma\int_{-\infty}^{\infty}|t|\frac{e^{-|t|}}{2}dt = \sigma$ y $v^2 = \sigma^2\int_{-\infty}^{\infty}t^2\frac{e^{-|t|}}{2}dt = 2\sigma^2$ y (2.1.1) puede expresarse como $\begin{pmatrix} 2\sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{pmatrix}$. Con lo cual la varianza mínima se obtiene para $w = 0$. Esto es razonable ya que para la distribución de Laplace la mediana muestral es el EMV de μ y, por tanto, la aportación de la media muestral no contribuye a reducir la varianza del estimador.

A diferencia de los casos anteriores, en los que $w = 1$ y $w = 0$, los siguientes son ejemplos de algunas distribuciones simétricas en las cuales una combinación lineal de la media y la mediana muestrales aporta una reducción en la varianza asintótica del estimador estudiado.

- Distribución Logística: $f(x;\mu,\sigma) = \frac{e^{-(\frac{x-\mu}{\sigma})}}{\sigma(1+e^{-(\frac{x-\mu}{\sigma})})^2} = \frac{1}{\sigma}g(\frac{x-\mu}{\sigma})$, con $g(t) = \frac{e^{-t}}{(1+e^{-t})^2}$.

  Para este caso $f(0) = \frac{1}{4\sigma}$, $\tau = \sigma\int_{-\infty}^{\infty}|t|\frac{e^{-t}}{(1+e^{-t})^2}dt = \sigma\ln 4$ y $v^2 = \sigma^2\int_{-\infty}^{\infty}t^2\frac{e^{-t}}{(1+e^{-t})^2}dt = \frac{\sigma^2\pi^2}{3}$. Así (2.1.1) puede expresarse como $\begin{pmatrix} \sigma^2\pi^2/3 & 2\sigma^2\ln(4) \\ 2\sigma^2\ln(4) & 4\sigma^2 \end{pmatrix}$ y el valor de w que minimiza la varianza es $w = 0.703512$. Observemos que a pesar que la media tiene menor varianza asintótica que la mediana, el valor de $w \neq 1$ nos

indica que podemos obtener una mejora considerando también la contribución de la mediana muestral.

- Distribución Uniforme: $f(x;\mu,\sigma) = \frac{\mathbf{1}(x)_{[\mu-\sigma;\mu+\sigma]}}{2\sigma} = \frac{1}{\sigma}g(\frac{x-\mu}{\sigma})$, con $g(t) = \frac{\mathbf{1}(t)_{[-1;1]}}{2}$.

  Aquí es $f(0) = g(0)/\sigma = 1/2\sigma; \tau = \sigma\int_{-1}^{1}\frac{|t|}{2}dt = \sigma/2$ y $v^2 = \sigma^2\int_{-1}^{1}\frac{t^2}{2}dt = \sigma^2/3$ con lo cual, (2.1.1) resulta $\begin{pmatrix}\sigma^2/3 & \sigma^2/2\\ \sigma^2/2 & \sigma^2\end{pmatrix}$ y para $w = 3/2 = 1.5$ se obtiene la varianza mínima. Notemos que el valor de $w > 1$ indica que el aporte de la mediana al estimador $\tilde{\mu}$ viene dado por un término de valor negativo en la combinación lineal.

- Mixtura de Normales al 50% con $X_1 \sim N(\mu,\sigma_1)$ y $X_2 \sim N(\mu,\sigma_2)$, es decir que $f(x;\mu,\sigma_1,\sigma_2) = \frac{1}{2}(f_{X_1} + f_{X_2}) = \frac{1}{2}(\frac{1}{\sigma_1\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma_1^2}} + \frac{1}{\sigma_2\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma_2^2}})$.

  Por tanto $f(0) = \frac{1}{2}(\frac{1}{\sigma_1\sqrt{2\pi}} + \frac{1}{\sigma_2\sqrt{2\pi}}) = \frac{(\sigma_1+\sigma_2)^2}{2\sigma_1\sigma_2\sqrt{2\pi}}$. Además, es fácil comprobar que $\tau = \frac{1}{2}(\tau_{f_{X_1}} + \tau_{f_{X_2}}) = \frac{\sigma_1+\sigma_2}{\sqrt{2\pi}}$ y $v^2 = \frac{1}{2}(v^2_{f_{X_1}} + v^2_{f_{X_2}}) = \frac{\sigma_1^2+\sigma_2^2}{2}$, con lo cual (2.1.1) puede expresarse como $\begin{pmatrix}\frac{\sigma_1^2+\sigma_2^2}{2} & \sigma_1\sigma_2\\ \sigma_1\sigma_2 & \frac{2\pi\sigma_1^2\sigma_2^2}{(\sigma_1+\sigma_2)^2}\end{pmatrix}$.

  Anteriormente hemos visto para variables aleatorias normales, que el valor para el cual se minimiza la varianza asintótica de nuestro estimador es $w = 1$. Sin embargo, observemos que ahora el valor óptimo de w depende de los valores de σ_1 y σ_2; concretamente, definiendo $\theta = \sigma_1/\sigma_2$ resulta, $w = w(\theta) = \frac{a}{(\sigma_1+\sigma_2)^2(\sigma_1-\sigma_2)^2+a}$ donde $a = 2\theta(2\pi\theta - (\theta+1)^2)$. Es claro que $w(\theta) = w(\theta^{-1})$ y solamente resulta igual a 1 cuando $\theta = 1$, es decir si $\sigma_1 = \sigma_2$, situación en que la mixtura desaparece y sólo nos queda una única distribución normal. La figura (2.1) muestra los valores de $w(\theta)$ para $\theta \in (0,1]$. En esta figura podemos observar

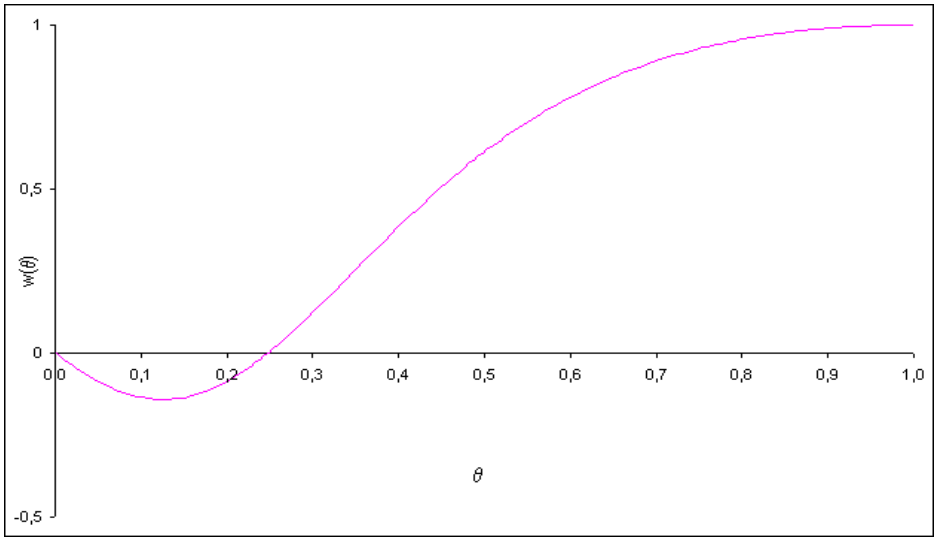

Figura 2.1: Valores de $w(\theta)$ para $\theta = \sigma_1/\sigma_2 \in (0, 1]$

que $w(\theta) = 0$, cuando $\theta = 0.25$ o $\theta = 4$ es decir, $w(\theta) = 0$ si una desviación estándar es cuatro veces la otra.

Otras aplicaciones de este tipo de estimadores pueden encontrarse en Samuel-Cahn (1994). Chan y He (1994) estiman el valor de w de manera no paramétrica.

Para las distribuciones normal y de Laplace la menor varianza asintótica de $\tilde{\mu}$ se obtiene en los estimadores de máxima verosimilitud, media y mediana respectivamente, que son asintóticamente eficientes. Pero que sucede en los ejemplos restantes?

Reemplazando en (2.1.2) la expresión de w dada en (2.1.3), la varianza asintótica del estimador $\tilde{\mu}$, puede escribirse como $Avar(\tilde{\mu}) = n^{-1}\frac{-\tau^2+v^2}{1-4f(0)\tau+4f(0)^2v^2}$. Esto nos permite determinar su eficiencia relativa asintótica (ARE), respecto del estimador de máxima verosimilitud. A continuación presentamos estos cálculos para las distribuciones consideradas anteriormente:

- Logística

  Para esta distribución $Avar(\tilde{\mu}) = \frac{3.137\sigma^2}{n}$ en tanto que, como señalan Antle et al. (1970) el estimador de máxima verosimilitud $\hat{\mu}$ tiene una varianza asintótica dada por $Avar(\hat{\mu}) = \frac{3\sigma^2}{n}$. Por tanto, la eficiencia relativa asintótica es $ARE[\tilde{\mu}, \hat{\mu}] = 3/3.137 = 0.96$. Esto quiere decir que es ligeramente más eficiente el EMV $\hat{\mu}$. De hecho, para muestras grandes, le basta con el 96% del tamaño de la muestra para estimar μ con igual eficiencia que $\tilde{\mu}$.

- Uniforme

  Para este caso tenemos que $Avar(\tilde{\mu}) = \frac{0.25\sigma^2}{n} = \frac{\sigma^2}{4n}$. Además es sabido que el EMV es el "midrange" $\hat{\mu} = \frac{x_{(n)}+x_{(1)}}{2}$, con $Avar(\hat{\mu}) = \frac{2\sigma^2}{(n+1)(n+2)}$. Así, la eficiencia relativa asintótica resulta $ARE[\tilde{\mu}, \hat{\mu}] = lim_{n\to\infty}\frac{8n}{(n+1)(n+2)} = 0$. Para esta distribución observamos pues un pésimo comportamiento del estimador propuesto en lo que a su eficiencia asintótica se refiere.

- Mixtura

  Para este ejemplo, al igual como ocurre para w, la varianza asintótica de ambos estimadores $\hat{\mu}$ y $\tilde{\mu}$ y por ende la $ARE[\tilde{\mu}, \hat{\mu}]$, dependen de θ. Concretamente,

$$Avar(\tilde{\mu}) = \frac{\sigma_2^2 2\theta^2(\pi(\theta^2+1) - (\theta+1)^2)}{n((\theta+1)^2(\theta-1)^2 + 2\theta(2\pi\theta - (\theta+1)^2))}$$

  y

$$Avar(\hat{\mu}) = \frac{\sigma_2^2 2\theta\sqrt{2\pi}}{n}[\int_{-\infty}^{\infty} \frac{t^2(e^{-t^2/2\theta^2}/\theta^2 + \theta e^{-t^2/2})^2}{(e^{-t^2/2\theta^2} + \theta e^{-t^2/2})} dt]^{-1}$$

  Esto nos lleva a que la eficiencia relativa asintótica de $\tilde{\mu}$ respecto del EMV $\hat{\mu}$ no sea tratable analíticamente y tengamos que calcularla numéricamente. En la figura 2.2 podemos observar el comportamiento de $ARE[\tilde{\mu}, \hat{\mu}]$ para distintos

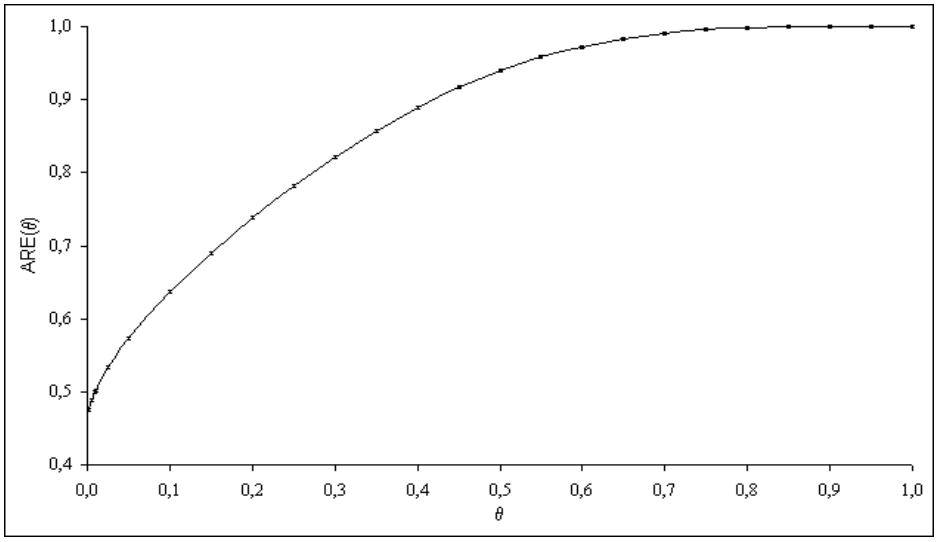

Figura 2.2: Eficiencia Asintótica Relativa $ARE[\tilde{\mu}, \hat{\mu}]$ para distintos valores de θ

valores de θ. Notemos en primer lugar que $ARE(\theta) = 1$ solamente si $\theta = 1$, es decir si $\sigma_1 = \sigma_2$ y por tanto la mixtura se reduce al caso de la distribución normal. Por otro lado, la eficiencia relativa asintótica de nuestro estimador es muy alta ($> 90\%$) para valores $\theta > 0.5$, es decir cuando la desviación estándar de una de las normales de la mixtura no sea superior al doble de la otra. Observemos también que la ARE es siempre superior al 47% ($ARE(0) \simeq 1/(\pi - 1)$).

En algunos de los ejemplos anteriores hemos constatado que el estimador $\tilde{\mu}$ presenta una eficiencia asintótica bastante alta y por tanto es un competidor razonable del EMV del parámetro de localización. Una pregunta inmediata que nos hacemos es, ¿para qué modelos este tipo de estimador es también asintóticamente eficiente?

En la próxima sección caracterizamos todos los modelos de localización simétricos que tienen esta propiedad.

2.2 Caracterización

Vamos a considerar los modelos estadísticos definidos sobre $\mathbb{R}$ con función de densidad $f(x-\mu;\psi)$, donde $\mu \in \mathbb{R}$ es el parámetro de localización, $\psi \in \Psi \subset \mathbb{R}^p$ es un vector p-dimensional de parámetros y f es simétrica en cero, es decir $f(t;\psi) = f(-t;\psi)$ para todo $\psi \in \Psi$. Varios modelos admiten esta representación, en particular todos los modelos de localización y escala simétricos. Para estos modelos el parámetro de localización μ representa tanto la esperanza como la media teórica.

Por razones técnicas, supondremos que $f(t;\psi)$ satisface las condiciones estándares de la desigualdad de Cramér-Rao (ver sección 1.2 del Capítulo 1). Más aún, para evitar que la distribución asintótica de $\tilde{x}$ sea degenerada, asumiremos que $f(t;\psi)$ es continua y positiva en cero.

En este caso, dada una muestra $x = (x_1, ..., x_n)$ la función de log-verosimilitud viene dada por

$$l(x;\mu,\psi) = \sum_{i=1}^{n} log f(x_i - \mu;\psi), \tag{2.2.1}$$

y la matriz de información de Fisher puede expresarse como

$$i = \begin{pmatrix} i_{\mu\mu} & i_{\mu\psi} \\ i_{\psi\mu} & i_{\psi\psi} \end{pmatrix}, \tag{2.2.2}$$

donde $i_{\psi\psi}$ es la matriz de información de Fisher de orden $p \times p$ del modelo restringido con densidad $f(t;\psi)$ y

$$i_{\mu\mu} = nE\{(\frac{\partial l}{\partial \mu})^2\} = nE\{\frac{f'(t;\psi)^2}{f(t;\psi)^2}\} = n\int_{-\infty}^{\infty} \frac{f'(t;\psi)^2}{f(t;\psi)} dt = 2n\int_{0}^{\infty} \frac{f'(t;\psi)^2}{f(t;\psi)} dt,$$

esto último por la simetría de $f(t;\psi)$. Por otro lado, $i_{\mu\psi}$ es un vector de orden $1 \times p$ con componente j-ésima dada por

$$i_{\mu\psi_j} = nE\{(\frac{\partial l}{\partial \mu}\frac{\partial l}{\partial \psi_j})\} = nE\{\frac{f'(t;\psi)}{f(t;\psi)}\frac{\partial l}{\partial \psi_j}(t;\psi)\} = n\int_{-\infty}^{\infty} f'(t;\psi)\frac{\partial l}{\partial \psi_j}(t;\psi)dt$$

que debido a la simetría de $\frac{\partial l}{\partial \psi_j}$ y la antisimetría de $f'(t;\psi)$ resulta,

$$= -n\int_{0}^{\infty} f'(t;\psi)\frac{\partial l}{\partial \psi_j}(t;\psi)dt + n\int_{0}^{\infty} f'(t;\psi)\frac{\partial l}{\partial \psi_j}(t;\psi)dt = 0,\ j = 1, ..., p$$

Por tanto, la matriz de varianza-covarianza asintótica del estimador de máxima verosimilitud (EMV) de (μ, ψ) tiene la simple expresión

$$\Sigma = \begin{pmatrix} 1/i_{\mu\mu} & 0 \\ 0 & i^{\psi\psi} \end{pmatrix}. \tag{2.2.3}$$

Es de destacar que el estimador de μ es ortogonal a todos los restantes. De aquí, la varianza asintótica del EMV de μ puede expresarse como

$$Avar(\hat{\mu}) = \frac{1}{2n\int_0^{\infty}\frac{f'(t;\psi)^2}{f(t;\psi)}dt}. \tag{2.2.4}$$

Nota 2.2.1. Para el caso de familias de localización y escala simétricas la matriz $i_{\psi\psi}$ también se reduce a un escalar, $i_{\sigma\sigma} = nE([\frac{\partial l}{\partial \sigma}]^2) = \frac{n}{\sigma^2}(\int_{-\infty}^{\infty}\frac{f'(t)^2}{f(t)}t^2dt - 1)$, donde ahora $t = \frac{(x-\mu)}{\sigma}$. Por lo tanto, la varianza asintótica del EMV $\hat{\sigma}$ es $Avar(\hat{\sigma}) = \frac{\sigma^2}{n}\{\int_{-\infty}^{\infty}\frac{f'(t)^2}{f(t)}t^2dt - 1\}^{-1}$.

A continuación estudiamos cuándo los estimadores de μ de la forma considerada, $\tilde{\mu} = w\bar{x} + (1-w)\widetilde{x}$, son también asintóticamente eficientes y, consecuentemente, buenos competidores del EMV de μ. El siguiente resultado caracteriza a todos los modelos de localización simétricos que tienen esta propiedad.

Teorema 2.2.1. *Los modelos de localización simétricos que tienen un estimador asintóticamente eficiente del parámetro de localización de la forma* $\tilde{\mu} = w\bar{x} + (1-w)\widetilde{x}$ *tienen la siguiente función de densidad*

$$f_\theta(x;\mu,\sigma) = \frac{c(\theta)}{\sigma}\exp(-\theta\frac{|x-\mu|}{\sigma} - \frac{(x-\mu)^2}{2\sigma^2}), \tag{2.2.5}$$

donde $\mu \in \mathbb{R}$ *y* $\sigma \in \mathbb{R}^+$ *son los parámetros de localización y escala respectivamente y* θ *es un valor fijo. Aquí* $c(\theta)^{-1} = \sqrt{2\pi}e^{\theta^2/2}(1 - Erf(\theta/\sqrt{2}))$, *y* $Erf(\cdot)$ *denota la función de error.*

Nota 2.2.2. La constante $c(\theta)$ puede expresarse en términos del cociente de Mills inverso $\lambda(\theta) = \frac{\varphi(\theta)}{1-\Phi(\theta)}$, también denominada función de razón de fallo de la distribución normal estándar. Concretamente, $c(\theta)^{-1} = \lambda(\theta)/2$ con lo cual, (2.2.5) puede expresarse equivalentemente como

$$f_\theta(x;\mu,\sigma) = \frac{\varphi(\theta)}{2(1-\Phi(\theta))\sigma}\exp(-\theta\frac{|x-\mu|}{\sigma} - \frac{(x-\mu)^2}{2\sigma^2}).$$

Además, $\lambda(\theta)$ interviene en la expresión de la esperanza de una variable aleatoria con distribución normal truncada lo que nos anticipa una relación entre esta distribución y (2.2.5) como veremos en el capítulo 4.

Demostración

Observemos que la varianza asintótica del estimador $\tilde{\mu} = w\bar{x} + (1-w)\tilde{x}$ puede calcularse, a partir de (2.1.2), como

$$\frac{w^2}{n}\int_{-\infty}^{\infty} t^2 f(t;\psi)dt + \int_{-\infty}^{\infty}\frac{w(1-w)}{f(0;\psi)n}|t|f(t;\psi)dt + \frac{(1-w)^2}{4f(0;\psi)^2 n}\int_{-\infty}^{\infty} f(t;\psi)dt,$$

que, por la simetría de $f(t;\psi)$ respecto de cero, resulta

$$\frac{2}{n}\int_0^{\infty} w^2t^2 f(t;\psi)dt + \frac{2}{n}\int_0^{\infty}\frac{w(1-w)}{f(0;\psi)}tf(t;\psi)dt + \frac{2}{n}\int_0^{\infty}\frac{(1-w)^2}{4f(0;\psi)^2}f(t;\psi)dt$$
$$= \frac{2}{n}\int_0^{\infty}((wt)^2 + 2\frac{w(1-w)}{2f(0;\psi)}t + \frac{(1-w)^2}{4f(0;\psi)^2})f(t;\psi)dt.$$

Finalmente, la podemos escribir en la forma

$$Avar(\tilde{\mu}) = \frac{2}{n}\int_0^\infty [wt + \frac{(1-w)}{2f(0;\psi)}]^2 f(t;\psi)dt. \tag{2.2.6}$$

Por hipótesis nuestro estimador $\tilde{\mu}$ es asintóticamente eficiente, luego la eficiencia asintótica relativa (ARE) con respecto al estimador de MLE debe ser igual a 1. Esto es, a partir de (2.2.4), (2.2.6) y denotando $d\nu(t) = f(t;\psi)dt$, tenemos

$$ARE(\hat{\mu},\tilde{\mu}) = \frac{Avar(\hat{\mu})}{Avar(\tilde{\mu})} = 4\int_0^\infty [wt + \frac{(1-w)}{2f(0;\psi)}]^2 d\nu(t) \int_0^\infty [\frac{f'(t;\psi)}{f(t;\psi)}]^2 d\nu(t) = 1$$

o, lo que es lo mismo,

$$\int_0^\infty [wt + \frac{(1-w)}{2f(0;\psi)}]^2 d\nu(t) \int_0^\infty [\frac{f'(t;\psi)}{f(t;\psi)}]^2 d\nu(t) = 1/4. \tag{2.2.7}$$

Por otra parte, la desigualdad de Cauchy-Schwarz para integrales establece que la parte izquierda de (2.2.7) es siempre mayor o igual que

$$[\int_0^\infty [wt + \frac{(1-w)}{2f(0;\psi)}][\frac{f'(t;\psi)}{f(t;\psi)}]d\nu(t)]^2 = [\int_0^\infty wtf'(t;\psi)dt + \int_0^\infty \frac{(1-w)}{2f(0;\psi)}f'(t;\psi)dt]^2.$$

Teniendo en cuenta que $\int_0^\infty f'(t;\psi)dt = -f(0;\psi)$ y que, integrando por partes, $\int_0^\infty tf'(t;\psi)dt = -1/2$, la cota inferior de la desigualdad resulta $[-w/2-1/2+w/2]^2 = (-1/2)^2 = 1/4$. Por lo tanto, la desigualdad de Cauchy-Schwarz para integrales establecida se torna en igualdad y la condición necesaria y suficiente para que esto ocurra en este caso, se traduce en la siguiente ecuación diferencial,

$$\frac{f'(t;\psi)}{f(t;\psi)} = k(wt + \frac{(1-w)}{2f(0;\psi)}),$$

cuyas soluciones son de la forma

$$f(t;\psi) = f(0;\psi)e^{k(w\frac{t^2}{2}+\frac{(1-w)}{2f(0;\psi)}t)}. \tag{2.2.8}$$

A partir de aquí, ψ permanece absolutamente identificado. Si ahora realizamos la extensión simétrica de $f(t;\psi)$ a todo $\mathbb{R}$ considerando la densidad $f(|t|;\psi)/2$, entonces (2.2.8) puede expresarse como

$$f(t;\psi) = \frac{f(0;\psi)}{2}\exp\{k(w\frac{t^2}{2} + \frac{(1-w)}{2f(0;\psi)}|t|)\}.$$

Sólo nos resta considerar el cambio de variable $t = x - \mu$ y la reparametrización $k = -(1 + 2\theta c(\theta))/\sigma^2$, $w = 1/(1 + 2\theta c(\theta))$ y $f(0;\psi) = 2c(\theta)/\sigma$ para obtener la densidad (2.2.5), como queríamos demostrar. □

Nota 2.2.3. La expresión $w = w(\theta) = \frac{1}{1+2\theta c(\theta)}$, que aparece en la demostración, también puede verse como una consecuencia directa de (2.1.3). Además, puede escribirse equivalentemente en términos del cociente inverso de Mills como $w = w(\theta) = \frac{1}{1+\theta\lambda(\theta)} = \frac{1-\Phi(\theta)}{(1-\Phi(\theta)+\theta\varphi(\theta))}$. Por otra parte, de (2.2.4) se deduce inmediatamente que la varianza asintótica del estimador de parámetro de localización $\tilde{\mu}$, es $Avar(\tilde{\mu}) = \frac{\sigma^2}{n(1+2\theta c(\theta))} = \sigma^2 w(\theta)/n$.

Nota 2.2.4. Observemos que en particular $w(0) = 1$. En este caso el Teorema 2.2.1 establece que *el único modelo de localización simétrico tal que la media muestral es un estimador asintóticamente eficiente del parámetro de localización es la distribución normal.* Análogamente, se puede mostrar que *la distribución de Laplace es el único modelo de localización simétrico tal que la mediana muestral es un estimador asintóticamente eficiente del parámetro de localización.*

Es sabido que la distribución normal es el único modelo de localización (bajo condiciones muy generales) tal que la media muestral es el EMV del parámetro de localización. Este es un antiguo resultado dado por Gauss, generalizado y modernizado por Teicher (1961). Similarmente, el único modelo de localización simétrico tal que la mediana muestral es el EMV del parámetro de localización es la distribución de Laplace (ver Rao y Ghosh, 1971).

Hasta donde sabemos, creemos que las caracterizaciones indicadas en la nota 2.2.4 son nuevas. De hecho, para un valor fijo de θ, el estimador $\tilde{\mu} = w\bar{x} + (1-w)\tilde{x}$ es insesgado y asintóticamente eficiente pero en general no coincide con el estimador de máxima verosimilitud de μ. En realidad, $\tilde{\mu}$ es un estimador más sencillo y fácil de calcular que el EMV como veremos en el próximo capítulo.

2.3 Extensión al Estimador de Hodges-Lehmann

Sean $x_1, x_2, ..., x_n$, observaciones independientes e idénticamente distribuidas de una variable aleatoria X con densidad $f(x-\mu)$ definida sobre $\mathbb{R}$, tal que $f(t)$ es simétrica en cero, siendo $\mu \in \mathbb{R}$ el parámetro de localización.

El estimador de Hodges-Lehmann (EHL) para el parámetro de localización μ, $\hat{\mu}_{hl}$, puede expresarse como la mediana de todos los pares de medias de las observaciones:

$$\hat{\mu}_{hl} = mediana\{\frac{x_i + x_j}{2}\} \;\; 1 \leq i \leq j \leq n.$$

Es decir, puede expresarse como la mediana de los denominados "promedios de Walsh".

Hodges y Lehmann (1963) propusieron este estimador puntual definiéndolo en términos del estadístico del test de los signos de Wilcoxon. Como puede observarse,

el EHL está basado en rangos lo que lo hace más resistente o robusto que la media muestral. Además, por ser una función simétrica de los estadísticos de orden, se sigue de manera inmediata que es un estimador insesgado del parámetro de localización, es decir, del centro de simetría.

La varianza asintótica de este estimador, altamente eficiente para distribuciones simétricas, está dada por

$$Avar(\hat{\mu}_{hl}) = [12\sigma^2(\int_{-\infty}^{-\infty} f(t)^2 dt)^2]^{-1}. \tag{2.3.1}$$

Puede demostrarse que su ARE respecto de la media muestral es por lo menos del 86.4%, es decir, que la eficiencia de la media muestral puede ser arbitrariamente peor que la del EHL, pero no puede ser mejor que el 15% aproximadamente (Hodges y Lehmann, 1956).

La eficiencia asintótica relativa del EHL respecto del estimador de máxima verosimilitud viene dada por,

$$\frac{Avar(\hat{\mu}_{hl})}{Avar(\hat{\mu})} = \frac{\int_{-\infty}^{-\infty} f'(t)^2/f(t)dt}{12(\int_{-\infty}^{-\infty} f(t)^2 dt)^2}$$

y sabemos, por la cota inferior de Cramer-Rao, que siempre se cumplirá

$$(\int_{-\infty}^{\infty} f(t)^2 dt)^2 \leq \frac{1}{12}\int_{-\infty}^{\infty} \frac{f'(t)^2}{f(t)} dt. \tag{2.3.2}$$

Gastwirth y Cohen (1970) calculan la eficiencia asintótica del EHL para varias distribuciones. Por ejemplo, señalan que para la distribución normal es alrededor del 95%, 75% para la distribución de Laplace y que es asintóticamente eficiente (100%) para la distribución logística.

La pregunta inmediata que nos surge entonces es, ¿para qué otros modelos el EHL es asintóticamente eficiente y por tanto buen competidor del EMV? El siguiente

resultado caracteriza a todos los modelos de localización simétricos que satisfacen dicha propiedad.

Teorema 2.3.1. *Los modelos de localización simétricos para los cuales el estimador de Hodges-Lehmann del parámetro de localización es asintóticamente eficiente, tienen la siguiente función de densidad*

$$f(x;\mu,\sigma) = \frac{\exp(-\frac{x-\mu}{\sigma})}{\sigma[1+\exp(-\frac{x-\mu}{\sigma})]^2}, \tag{2.3.3}$$

es decir, la correspondiente a la distribución logística de parámetros μ y σ.

Demostración

Para que el EHL sea asintóticamente eficiente debe darse la igualdad en (2.3.2). En primer lugar, veamos que esta desigualdad puede expresarse equivalentemente como

$$(\int_{-\infty}^{\infty} \frac{f'(t)}{f(t)}[F(t)-1/2]d\nu)^2 \leq \int_{-\infty}^{\infty}[F(t)-1/2]^2 d\nu \int_{-\infty}^{\infty}(\frac{f'(t)}{f(t)})^2 d\nu, \tag{2.3.4}$$

donde $F(t)$ es la función de distribución y $d\nu = f(t)dt$.

Efectivamente, la parte izquierda de la desigualdad puede escribirse como

$$(\int_{-\infty}^{\infty} f'(t)F(t) - f'(t)/2dt)^2 = (\int_{-\infty}^{\infty} f'(t)F(t)dt - \int_{-\infty}^{\infty} f'(t)/2dt)^2 =$$
$$= (\int_{-\infty}^{\infty} f'(t)F(t)dt)^2 = ([f(t)F(t)]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(t)F'(t)dt)^2 = (\int_{-\infty}^{\infty} f(t)^2 dt)^2,$$

en tanto que para la parte derecha, el primer término resulta,

$$\int_{-\infty}^{\infty}[F(t)-1/2]^2 d\nu = \int_{-\infty}^{\infty}[F(t)-1/2]^2 f(t)dt = \int_0^1 [u-1/2]^2 du = 1/12,$$

por ser la varianza de una $U(0,1)$, y el segundo término es

$$\int_{-\infty}^{\infty}(\frac{f'(t)}{f(t)})^2 d\nu = \int_{-\infty}^{\infty}(\frac{f'(t)}{f(t)})^2 f(t)dt = \int_{-\infty}^{\infty}\frac{f'(t)^2}{f(t)}dt.$$

Luego, para que el EHL sea asintóticamente eficiente debe darse la igualdad en (2.3.4), que es la desigualdad de Cauchy-Schwarz para $G(t) = [F(t) - 1/2]$ y $H(t) = f'(t)/f(t)$. Esto ocurre sí y sólo sí $G = kH$, es decir sii $F(t) - 1/2 = kf'(t)/f(t)$ o equivalentemente sí y sólo sí

$$F(t) - k\frac{F''(t)}{F'(t)} = 1/2.$$

Las soluciones a esta ecuación diferencial de segundo orden, con las condiciones iniciales $F(0) = 1/2$ y $F(\infty) = 1$, son de la forma

$$F(t) = \frac{1}{1 + e^{-t/2k}} k > 0.$$

A partir de aquí, considerando que $t = x - \mu$ y haciendo $2k = \sigma$ obtenemos la distribución logística. Derivando para obtener la función de densidad, el teorema queda probado. □.

Nota 2.3.1. Al igual que Gastwirth y Cohen (1970), otros autores también se dan cuenta que el estimador Hodges-Lehmann es asintóticamente eficiente (fully efficient) para la distribución logística. Observemos sin embargo la diferencia con nuestro resultado, pues la caracterización del teorema 2.3.1 nos asegura que *la distribución logística es el único modelo de localización simétrico para el cual el EHL es asintóticamente eficiente.*

Park y Lindsay (1999) realizan un extensivo estudio mediante simulación para evaluar el comportamiento de varios estimadores robustos del parámetro de localización, entre ellos el EMV, el EHL y el estimador no paramétrico de Chan y He

comentado en la sección 2.1. A continuación, reproducimos parte de la información que los autores brindan respecto al sesgo y el error cuadrático medio (ECM) estimado, para muestras de tamaño $n = 25, 50, 75, 100$ de una distribución logística(0,1), en base a 10.000 simulaciones. Además, por nuestra parte incorporamos a estos resultados los correspondientes al estimador de la combinación lineal de la media y la mediana específico para la distribución logística, $\tilde{\mu} = 0.7035\bar{x} + 0.2965\tilde{x}$, considerado en este capítulo y simulados en idénticas condiciones.

Estimador	n=25		n=50		n=75		n=100	
	sesgo	ECM	sesgo	ECM	sesgo	ECM	sesgo	ECM
EMV	-0.0039	0.1219	0.0020	0.0617	0.0014	0.0405	0.0012	0.0306
$\hat{\mu}_{hl}$	-0.0044	0.1233	0.0020	0.0621	0.0012	0.0406	0.0011	0.0307
Chan y He	-0.0035	0.1287	0.0023	0.0653	0.0018	0.0428	0.0008	0.0324
$\tilde{\mu}$	0.0004	0.1261	0.0001	0.0631	0.0007	0.0417	-0.0004	0.0314

Como hemos comentado, $\tilde{\mu}$ y $\hat{\mu}_{hl}$ son estimadores insesgados del parámetro de localización, es decir, el valor teórico del sesgo es cero. Sin embargo, presentamos el sesgo estimado como una referencia experimental a fin de evaluar el sesgo del EMV y el del estimador de Chan y He que desconocemos. Desde este punto de vista, podemos observar que todos los estimadores presentan sesgos empíricos del mismo orden. Las fluctuaciones registradas puede apreciarse en la figura 2.3. Esto corrobora de algún modo, lo expresado por Antle *et al.* (1970) respecto a que, para la distribución logística, el sesgo del EMV es cero.

Los valores empíricos del ECM nos permiten analizar la eficiencia observada de los estimadores. En la figura 2.4 representamos estos valores relativizados por el tamaño muestral ($n * ECM$) para los cuatro estimadores considerados. Como se puede apreciar, $\hat{\mu}_{hl}$ y el EMV presentan un comportamiento muy similar con valores

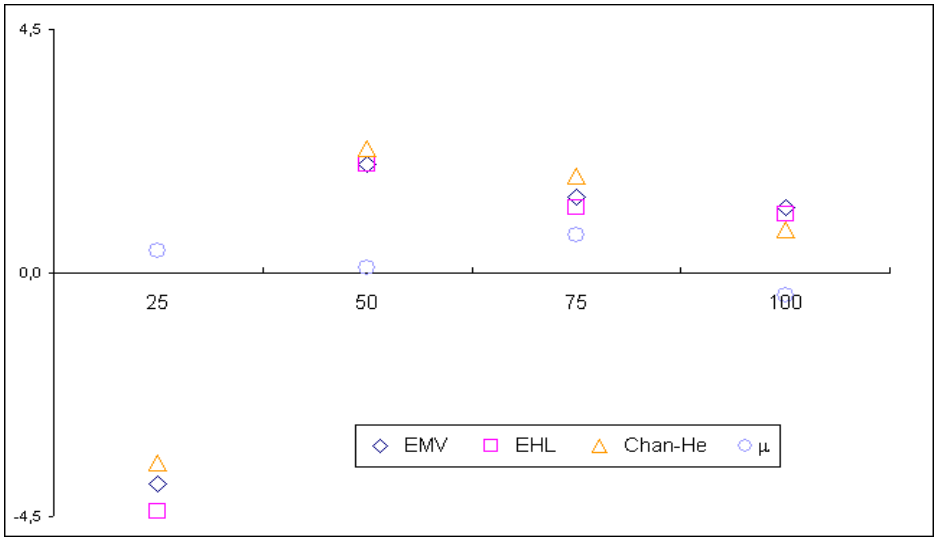

Figura 2.3: Sesgo observado de los estimadores $\tilde{\mu}$, de Chan y He, EHL y EMV para distintos tamaños muestrales en base a 10.000 simulaciones

muy cercanos a 3: recordemos que la cota inferior de Cramér-Rao del parámetro de localización para la distribución logística es $3\sigma^2/n$. Se observa un diferencia marcada respecto a los otros dos estimadores, resaltando particularmente el mal comportamiento del estimador de Chan y He.

Es interesante destacar que, aún siendo asintóticamente equivalentes, el EHL no resulta igual al EMV, como podemos apreciar en el siguiente ejemplo.

2.3.1 Ejemplo

Schmidt (2002) en su estudio ecológico sobre peces y características de la calidad del agua, señala que la salinidad es el mayor factor ambiental que afecta a la distribución de los peces en la bahía de Florida. Los datos que se presentan a continuación, corresponden a la salinidad media trimestral en 27 estaciones de dicha zona.

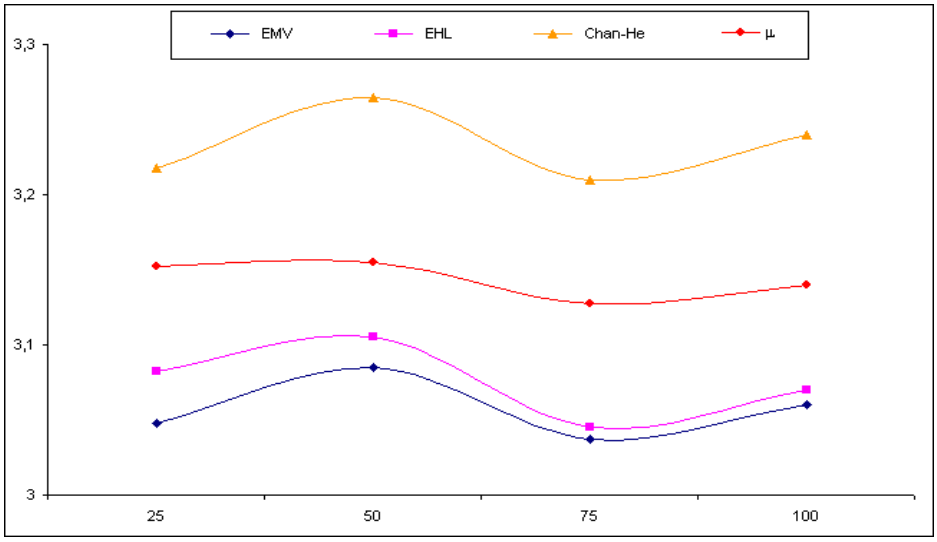

Figura 2.4: $n * ECM$ empíricos para los estimadores $\tilde{\mu}$, de Chan y He, EHL y EMV en base a 10.000 simulaciones

33.07	36.17	37.83	36.31	37.51	38.23	36.51	36.06	34.38
29.86	29.45	44.95	45.36	45.53	39.30	38.28	36.72	38.75
40.75	41.80	48.55	50.95	42.70	37.72	27.05	43.68	42.74

La figura 2.5, un "Q-Q plot" generado con el paquete SPSS, nos permite asumir que los datos provienen de un distribución logística. En este gráfico se representan los percentiles empíricos versus los esperados para una distribución logística de parámetros 38.897 y 3.112. Cabe aclarar que el SPSS estima los parámetros de la logística mediante el método de los momentos. Por tanto, estima el parámetro de localización mediante la media muestral y, debido a que la varianza de una variable aleatoria con distribución logística es $\pi^2\sigma^2/3$ (ver Antle *et al.*, 1970), el estimador del parámetro de escala resulta $\sqrt{3}s/\pi$, donde s es la desviación típica.

Para el parámetro de localización tenemos las siguientes estimaciones: la media muestral es de 38.897, la mediana es de 38.230 y los EMV son $\hat{\mu} = 38.851$ y $\hat{\sigma} = 3.1408$.

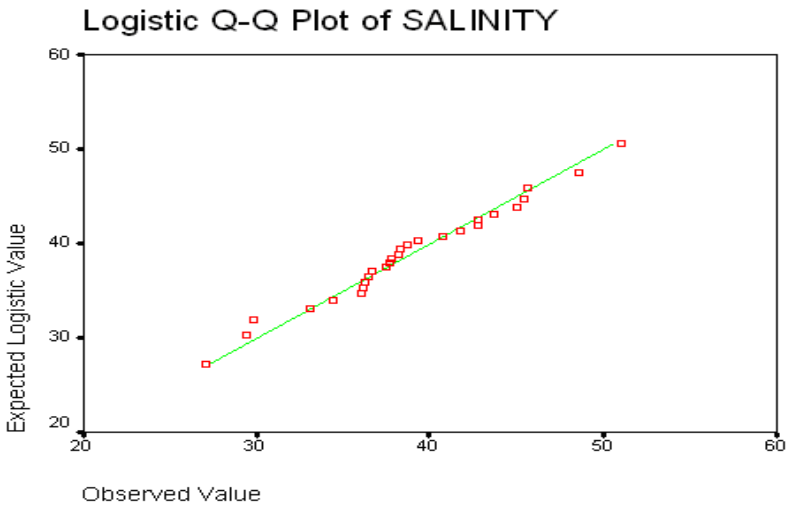

Figura 2.5: Percentiles empíricos versus los esperados para una distribución logística $L(38.896; 3.112)$

En tanto que, estimador de Hodges-Lehmann del parámetro de localización, es decir la mediana de los $n(n+1)/2 = 378$ promedios de Walsh, resulta $\hat{\mu}_{hl} = 38.780$.

El parámetro de escala puede estimarse sustituyendo μ por el valor correspondiente al EHL en la log-verosimilitud y maximizando respecto de σ. Para nuestro ejemplo, reemplazando μ por $\hat{\mu}_{hl}$ y maximizando, resulta $\hat{\sigma}_{hl} = 3.1405$. Observemos que $\hat{\sigma}_{hl}$ es un pseudo-EMV de σ (ver Capítulo 1) y resulta también asintóticamente eficiente por serlo $\hat{\mu}_{hl}$.

Capítulo 3

Una Familia de Localización y Escala Simétrica $(\theta = 1)$

En la caracterización del Teorema 2.2.1 se establece que θ puede ser un valor fijo arbitrario. En este capítulo estudiaremos la familia de localización y escala que se obtiene al considerar $\theta = 1$ en el modelo (2.2.5). Los resultados obtenidos son fácilmente generalizables al tomar otro valor cualquiera de θ.

Para este caso particular, analizaremos con detalle la estimación puntual de los parámetros y también la construcción de intervalos de confianza, aproximados y exactos, tanto para el parámetro de localización μ como para el de escala σ.

También presentaremos un ejemplo, cuyos datos se ajustan razonablemente bien mediante la distribución propuesta, para ilustrar los tópicos desarrollados.

3.1 La Familia $\theta = 1$

Fijando $\theta = 1$ en el modelo (2.2.5) obtenemos una nueva familia de localización y escala con densidad

$$f_1(x;\mu,\sigma) = \frac{c(1)}{\sigma}\exp(-\frac{|x-\mu|}{\sigma} - \frac{(x-\mu)^2}{2\sigma^2}), \tag{3.1.1}$$

donde $c(1) = 1/(\sqrt{2\pi}e^{1/2}(1 - Erf(1/\sqrt{2}))) = \varphi(1)/(2(1-\Phi(1))) = 0.762567$.

Obviamente, el modelo es simétrico respecto de la media poblacional ($E[X] = \mu$) y su desviación estándar se calcula multiplicando el parámetro de escala σ por una constante ($Var[X] = 0.4749\sigma^2$). Además, tiene la particularidad que su densidad se encuentra entre la Gaussiana ($\theta = 0$) y la de Laplace ($\theta \to \infty$).

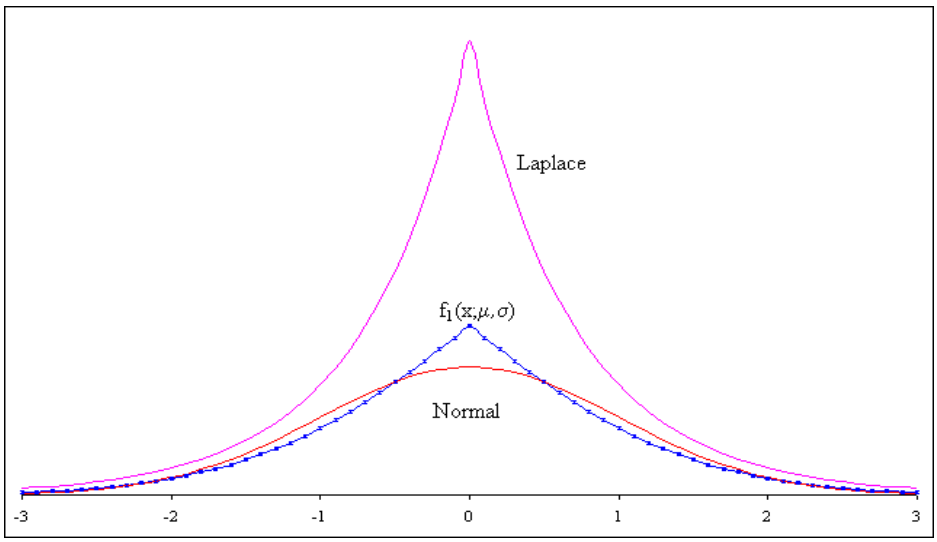

Figura 3.1: Densidades para las Distribuciones de Laplace, $f_1(x;\mu,\sigma)$ y Normal

Estas características pueden apreciarse en la figura 3.1, donde se representan las densidades para las tres distribuciones con media en cero y varianza igual a 1: es

decir, la nueva distribución con parámetros $\mu = 0$ y $\sigma = 1.4512$, la normal estándar $N(0,1)$, y la de Laplace con $\mu = 0$ y $\sigma = 0.7071$.

3.2 Generación de una Muestra Aleatoria

En esta sección, desarrollaremos procedimientos para obtener valores de una variable aleatoria X con densidad (3.1.1), mediante la simulación de valores pseudoaleatorios uniformemente distribuidos en el intervalo $[0,1]$. La primera idea es utilizar un método general, que sea rápido y eficaz, que sirva para simular datos de una distribución cualquiera. Por ello, aplicaremos el método AR, descrito en el capítulo 1.

Los procedimientos que desarrollaremos serán válidos para cualquier $\theta > 0$. Además, dado que es un modelo de localización y escala, bastará con simular valores para el caso estándar, es decir, $\mu = 0$ y $\sigma = 1$.

Por tanto, deseamos generar observaciones provenientes de una distribución con densidad dada por $f_\theta(x) = c(\theta)\exp(-\theta|x| - x^2/2)$. Para aplicar el método AR recordemos que es aconsejable escoger una densidad mayorante g cercana a f_θ, de manera que la variable aleatoria relacionada pueda generarse fácilmente. Observemos que el hecho de que la densidad f_θ se encuentra entre la normal y la de Laplace para $\theta \geq 0$, nos está indicando que podemos disponer de dos métodos alternativos para simular valores de nuestra distribución.

Método 1

Consideremos una v.a. Z con distribución $N(0,1)$, es decir $g(z) = \exp(-\frac{z^2}{2})/\sqrt{2\pi}$; por lo tanto, $M = c(\theta)\sqrt{2\pi}$ o equivalentemente $f/[Mg(z)] = \exp(-\theta|z|)$. Para este caso, el algoritmo resultante es el siguiente:

1. Generamos un valor normal estándar, z. Por ejemplo, utilizando el método de Box-Müller.

2. Generamos un valor uniforme en $(0, 1)$, u.

3. Si $u \leq \exp(-\theta|z|)$, aceptamos a z como un dato perteneciente a nuestra distribución. Si no, volvemos al paso 1.

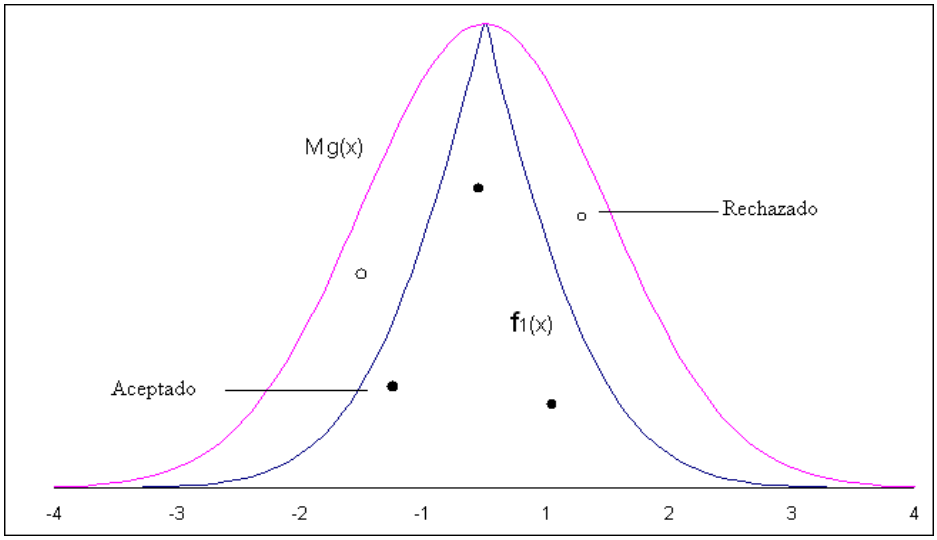

Figura 3.2: Densidades del Método AR para la simulación de una muestra aleatoria de $f_1(x; \mu, \sigma)$ considerando la normal estándar como mayorante.

Por otra parte, y como mencionamos en el capítulo 1, el hecho de que el número de observaciones necesarias para generar una observación de X tenga una distribución geométrica con media M, nos permite comparar la eficacia del método AR para distintos valores de θ. Concretamente, la probabilidad de aceptar un valor como proveniente del modelo (2.2.5) resulta $P(Aceptar) = 1/M = 2e^{\theta^2/2}(1 - \Phi(\theta))$, que

claramente es 1 para la distribución normal ($\theta = 0$) y tiende a cero a medida que crece θ, como podemos apreciar en la tabla 3.1.

θ	0	0.5	1	1.5	2	2.5	3	8	10
$P(Aceptar)$	1	0.70	0.52	0.41	0.34	0.28	0.24	0.09	0

Tabla 3.1: Probabilidad de aceptar que una observación proviene del modelo (2.2.5) para varios valores positivos de θ considerando la normal estándar como mayorante.

Observemos que para nuestro caso particular ($\theta = 1$), la probabilidad de rechazar una observación como proveniente de una distribución con densidad (3.1.1) es del 48%, en tanto que sería del 91% cuando $\theta = 8$.

Método 2

Ahora consideraremos como mayorante la distribución de Laplace$(0, 1/\theta)$, con densidad dada por $g(y) = \theta \exp(-\theta|y|)/2$. Tenemos que $M = 2c(\theta)/\theta$ y $f/[Mg(y)] = \exp(-y^2/2)$, con lo cual el algoritmo resulta:

1. Generamos un valor de la Laplace$(0, 1/\theta)$, y. En este caso, utilizamos el método de la función de distribución inversa (FDI).

2. Generamos un valor uniforme en $(0, 1)$, u.

3. Si $u \leq \exp(-y^2/2)$, aceptamos a y como un dato perteneciente a nuestra distribución. Si no, volvemos al paso 1.

θ	0	0.5	1	1.5	2	2.5	3	8	10
$P(Aceptar)$	0	0.44	0.66	0.77	0.84	0.89	0.91	0.97	1

Tabla 3.2: Probabilidad de aceptar que una observación proviene del modelo (2.2.5) para varios valores positivos de θ considerando la Laplace$(0, 1/\theta)$ como mayorante.

La tabla 3.2 nos muestra que la probabilidad de aceptar una observación como proveniente de la densidad (2.2.5) tiene un comportamiento inverso al del método 1. Resulta 0 para la distribución normal ($\theta = 0$) y tiende a 1 cuando θ crece. Observemos además que para nuestro caso particular ($\theta = 1$), ahora tenemos una probabilidad de rechazo de tan sólo el 34%.

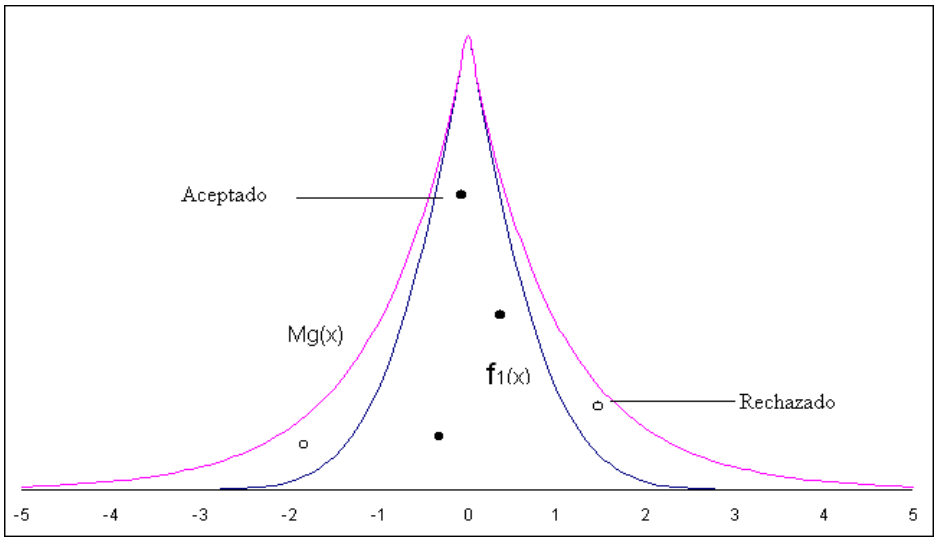

Figura 3.3: Densidades del Método AR para la simulación de una muestra aleatoria de $f_1(x; \mu, \sigma)$ considerando la Laplace$(0, 1/\theta)$ como mayorante.

Por lo tanto, disponemos de dos métodos para generar observaciones provenientes del modelo (2.2.5): el método 1 considerando la N$(0, 1)$ con $P(Aceptar) = 1/c(\theta)\sqrt{2\pi}$, y el método 2 tomando como mayorante a la Laplace$(0, 1/\theta)$ con $P(Aceptar) = \theta/2c(\theta)$.

Como puede observarse, las probabilidades se igualan para $\theta = 2/\sqrt{2\pi} \simeq 0.8$. Por consiguiente, el método 1 es más eficiente que el método 2 para valores de $\theta \leq 2/\sqrt{2\pi}$, y el método 2 es superior al método 1 para valores de θ superiores a $2/\sqrt{2\pi}$. Por tanto,

queda claro que para nuestro caso particular $\theta = 1 > 0.8$ debemos utilizar el método 2 por ser más eficiente. Más aún, la simulación de valores de una Laplace a través del método de la FDI requiere de sólo un valor $U(0,1)$, contra dos que se requieren para simular un valor de la normal si empleamos el algoritmo de Box-Müller.

Para la simulación de muestras aleatorias provenientes del modelo (2.2.5) para un valor fijo de $\theta > 0$ y con un tamaño muestral n pre-seleccionado, hemos escrito un programa donde se implementan los dos métodos descritos y se utiliza el apropiado, según sea el valor de θ mayor o menor a $2/\sqrt{2\pi}$. Los valores aleatorios generados de esta manera seguirán una distribución cuyo parámetro θ será el elegido por nosotros, con $\mu = 0$ y $\sigma = 1$. A partir de estos valores aleatorios, digamos t_i, podemos obtener otros con diferentes elecciones de μ y σ simplemente realizando la transformación lineal $\mu + \sigma t_i$.

3.3 Estimación Puntual

3.3.1 Estimación de μ: Combinación Lineal de Media y Mediana

Para esta nueva distribución con densidad (3.1.1) ($\theta = 1$), a partir de lo expuesto en la nota 2.2.3 obtenemos que $w = w(1) = 1/(1 + 2c(1)) = 1/2.5251 = 0.39602$ y por tanto,

$$\tilde{\mu} = w(1)\bar{x} + (1 - w(1))\tilde{x} = 0.39602\bar{x} + 0.60398\tilde{x}.$$

Además, y debido a que (3.1.1) es un caso particular del modelo (2.2.5), el Teorema 2.2.1 nos asegura que $\tilde{\mu}$ es un estimador insesgado y asintóticamente eficiente con

varianza asintótica

$$Avar(\tilde{\mu}) = w(1)\sigma^2/n = 0.39602\sigma^2/n.$$

El hecho que la mínima varianza se alcance para $w(1) = 0.39602$ ($w \neq 0$ y $w \neq 1$) nos indica que obtenemos una mejora en la eficiencia asintótica al considerar como estimador de μ a la combinación lineal de la media y la mediana muestrales. Lógicamente ambos estimadores presentan, individualmente, mayor varianza asintótica: $Avar(\bar{x}) = 0.4749\sigma^2/n$ y $Avar(\tilde{x}) = 0.4299\sigma^2/n$.

3.3.2 Estimador de Máxima Verosimilitud

Consideremos una muestra $x_1, x_2, ..., x_n$ proveniente de (3.1.1). Teniendo en cuenta la identidad $\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{n} = s^2 + (\bar{x} - \mu)^2$, siendo s^2 la varianza muestral, la función de log-verosimilitud puede escribirse como,

$$l_1(x; \mu, \sigma) = n\{\log \frac{c(1)}{\sigma} - \frac{s^2 + (\bar{x} - \mu)^2}{2\sigma^2} - \frac{\triangle(\mu)}{\sigma}\}, \tag{3.3.1}$$

donde $\triangle(\mu) = \sum_{i=1}^{n} \frac{|x_i-\mu|}{n}$.

Como la función (3.3.1) es diferenciable respecto de σ en el dominio de los parámetros, obtenemos la siguiente ecuación de verosimilitud:

$$\frac{\partial l_1}{\partial \sigma} = n\{-1/\sigma + (s^2 + (\bar{x} - \mu)^2)/\sigma^3 + \triangle(\mu)/\sigma^2\} = 0.$$

Resolviendo la ecuación cuadrática en σ resultante, la única solución positiva es

$$\hat{\sigma_0} = \frac{\triangle(\mu) + \sqrt{\triangle(\mu)^2 + 4(\bar{x} - \mu)^2 + 4s^2}}{2}. \tag{3.3.2}$$

Vamos a ver ahora cómo queda la función de log-verosimilitud (3.3.1) evaluada en $\hat{\sigma_0}$. Para ello, observemos que,

$$\hat{\sigma_0}^2 = \frac{\triangle(\mu)^2 + 2\triangle(\mu)\sqrt{\triangle(\mu)^2 + 4(s^2 + (\bar{x} - \mu)^2)} + \triangle(\mu)^2 + 4(\bar{x} - \mu)^2 + 4s^2)}{4} =$$

$$= \triangle(\mu)\frac{(\triangle(\mu) + \sqrt{\triangle(\mu)^2 + 4(s^2 + (\bar{x} - \mu)^2)})}{2} + s^2 + (\bar{x} - \mu)^2,$$

es decir, se cumple la identidad, $\hat{\sigma_0}^2 = \triangle(\mu)\hat{\sigma_0} + s^2 + (\bar{x} - \mu)^2$ o bien $\hat{\sigma_0}^2 - \triangle(\mu)\hat{\sigma_0} = s^2 + (\bar{x} - \mu)^2$.

Ahora sustituyendo estas expresiones en (3.3.1), podemos expresar la función de log-verosimilitud como,

$$l_1^*(x;\mu) = l_1(x;\mu,\hat{\sigma_0}) = n\{\log\frac{c(1)}{\hat{\sigma_0}} - \frac{\hat{\sigma_0}^2 - \triangle(\mu)\hat{\sigma_0}}{2\hat{\sigma_0}^2} - \frac{\triangle(\mu)}{\hat{\sigma_0}}\}$$

o equivalentemente,

$$l_1^*(x;\mu) = -\frac{n}{2} + n\log 2c(1) - n\{\log h(\mu) + \frac{\triangle(\mu)}{h(\mu)}\},$$

siendo $h(\mu) = \triangle(\mu) + \sqrt{\triangle(\mu)^2 + 4(\bar{x} - \mu)^2 + 4s^2}$.

Luego si todos los parámetros son desconocidos, el EMV del parámetro de localización μ, $\hat{\mu}$, se puede calcular maximizando $l_1^*(x;\mu)$, o bien minimizando la función $\{\log h(\mu) + \frac{\triangle(\mu)}{h(\mu)}\}$, cosa que hay que hacer numéricamente ya que no es posible realizarlo de manera analítica.

El parámetro de escala puede estimarse sustituyendo el valor de μ en (3.3.2) por los correspondientes estimadores que hemos considerado anteriormente. Así el EMV de σ será,

$$\hat{\sigma} = \frac{\triangle(\hat{\mu}) + \sqrt{\triangle(\hat{\mu})^2 + 4(s^2 + (\bar{x} - \hat{\mu})^2)}}{2}$$

mientras que el estimador alternativo de σ, construido a partir de $\tilde{\mu}$, puede expresarse como,

$$\tilde{\sigma} = \frac{\triangle(\tilde{\mu}) + \sqrt{\triangle(\tilde{\mu})^2 + 4(s^2 + (\bar{x} - \tilde{\mu})^2)}}{2}.$$

Además, y dado que este modelo es una familia de localización y escala simétrico, por la nota 2.2.1, la matriz de varianza-covarianza asintótica del EMV de (μ, σ) resulta ser $\frac{\sigma^2}{n}\begin{pmatrix} \{2\int_0^\infty \frac{f'(t)^2}{f(t)}dt\}^{-1} & 0 \\ 0 & \{\int_{-\infty}^{\infty} \frac{f'(t)^2}{f(t)}t^2dt - 1\}^{-1} \end{pmatrix}$. Evaluando numéricamente las integrales que aparecen dentro de la matriz, obtenemos $Avar(\hat{\mu}) = Avar(\tilde{\mu}) = 0.3960\sigma^2/n$ y $Avar(\hat{\sigma}) = Avar(\tilde{\sigma}) = 0.6780\sigma^2/n$. Es de destacar que el estimador $\tilde{\sigma}$ es un pseudo-EMV del parámetro de escala (ver capítulo 1) que en este caso resulta también asintóticamente eficiente.

3.3.3 Comparaciones entre Estimadores

Por lo dicho hasta aquí, disponemos de estimadores $(\tilde{\mu}, \tilde{\sigma})$ alternativos a los EMV $(\hat{\mu}, \hat{\sigma})$ que conservan sus mismas propiedades asintóticas (insesgados y eficientes) y además, son mucho más fáciles de calcular. Pero qué sucede cuando el tamaño de la muestra es pequeño? En esta sección analizaremos el comportamiento de ambos estimadores para pequeñas muestras.

En primer lugar, observemos que generalmente los estimadores $(\tilde{\mu}, \tilde{\sigma})$ no coinciden con los de máxima verosimilitud. Por ejemplo, utilizando el método 2 de la sección anterior, generamos una muestra de tamaño $n = 7$ correspondiente a una densidad (3.1.1) con $\mu = 0$ y $\sigma = 1$. Los valores obtenidos han sido los siguientes: $t_1 = 1.251$, $t_2 = -0.446$, $t_3 = -0.297$, $t_4 = -0.953$, $t_5 = 1.377$, $t_6 = -0.158$,$t_7 = 0.579$. Para este ejemplo, usando el software *Mathematica*, obtenemos que los EMV son $\hat{\mu} = 0.01030$ y $\hat{\sigma} = 1.28093$. Sin embargo, los estimadores alternativos resultan $\tilde{\mu} = -0.01888$ y

$\tilde{\sigma} = 1.28429$.

Por el Teorema 1.2.1 sabemos que $(\hat{\mu} - \mu)/\hat{\sigma}$ es un pívot. Este resultado es importante porque nos indica cómo podemos construir intervalos de confianza exactos de μ basándonos en los EMV. La siguiente proposición demuestra que esto mismo sucede para el combinante $\frac{(\tilde{\mu}-\mu)}{\tilde{\sigma}}$.

Proposición 3.3.1. *Los combinantes* $\frac{\tilde{\mu}-\mu}{\sigma}$, $\frac{\tilde{\sigma}}{\sigma}$ *y* $\frac{\tilde{\mu}-\mu}{\tilde{\sigma}}$ *son cada uno de ellos un pívot, es decir, su distribución no depende de los parámetros* μ *y* σ.

Demostración

Sea X una v.a. con densidad dada por (2.2.5). Cada una de las observaciones de X pueden expresarse como $x_i = \mu + \sigma z_i$ donde z_i son observaciones estandarizadas $(\mu = 0, \sigma = 1)$. A partir de aquí podemos reescribir los estimadores $\tilde{\mu}$ y $\tilde{\sigma}$ en términos de las z_i que no dependen de los parámetros. Concretamente,

$$\tilde{\mu} = w(\theta)\bar{x} + (1 - w(\theta))\tilde{x} = \mu + \sigma(w(\theta)\bar{z} + (1 - w(\theta))\tilde{z}) = \mu + \sigma\widetilde{\mu}_z$$

y

$$\triangle(\tilde{\mu}) = \frac{\sum_{i=1}^{n} |x_i - \tilde{\mu}|}{n} = \sigma \frac{\sum_{i=1}^{n} |z_i - \widetilde{\mu}_z|}{n} = \sigma \triangle_z.$$

Con lo cual

$$\tilde{\sigma} = \frac{\triangle(\tilde{\mu}) + \sqrt{\triangle(\tilde{\mu})^2 + 4\sum_{i=1}^{n}(x_i - \tilde{\mu})^2}}{2} = \sigma \frac{\triangle_z + \sqrt{\triangle_z^2 + 4\sum_{i=1}^{n}(z_i - \widetilde{\mu}_z)^2}}{2}.$$

Finalmente tenemos que,

$$\frac{\tilde{\mu} - \mu}{\tilde{\sigma}} = \frac{2\widetilde{\mu}_z}{\triangle_z + \sqrt{\triangle_z^2 + 4\sum_{i=1}^{n}(z_i - \widetilde{\mu}_z)^2}}$$

y

$$\frac{\tilde{\sigma}}{\sigma} = \frac{\triangle_z + \sqrt{\triangle_z^2 + 4\sum_{i=1}^{n}(z_i - \widetilde{\mu}_z)^2}}{2}.$$

Es decir, que las distribuciones de $(\tilde{\mu} - \mu)/\tilde{\sigma}$ y de $\tilde{\sigma}/\sigma$ no dependen de los parámetros. De aquí resulta inmediato que $(\tilde{\mu} - \mu)/\sigma$ también es un pívot. □

La Proposición anterior es válida para el modelo (2.2.5) con cualquier valor de θ prefijado. Particularmente, se cumple para la densidad (3.3.1) que estamos considerando.

Por lo expuesto hasta aquí, y como consecuencia de la proposición anterior, podemos entonces expresar el sesgo y el error cuadrático medio (ECM) para cada estimador de la siguiente manera:

- $\hat{\mu}$: $sesgo = E(\hat{\mu} - \mu) = \sigma k_n$ y $ECM = E(\hat{\mu} - \mu)^2 = \sigma^2 H_n$
- $\tilde{\mu}$: $sesgo = E(\tilde{\mu} - \mu) = \sigma k_n^{'}$ y $ECM = E(\tilde{\mu} - \mu)^2 = \sigma^2 H_n^{'}$
- $\hat{\sigma}$: $sesgo = E(\hat{\sigma} - \sigma) = \sigma l_n$ y $ECM = E(\hat{\sigma} - \sigma)^2 = \sigma^2 L_n$
- $\tilde{\sigma}$: $sesgo = E(\tilde{\sigma} - \sigma) = \sigma l_n^{'}$ y $ECM = E(\tilde{\sigma} - \sigma)^2 = \sigma^2 L_n^{'}$

donde k_n, $k_n^{'}$, l_n, $l_n^{'}$, H_n, $H_n^{'}$, L_n y $L_n^{'}$ sólo dependen del tamaño muestral y por lo tanto podrán calcularse mediante simulaciones.
Cabe aclarar que, si bien sabemos que $\tilde{\mu}$ es un estimador insesgado, esto es, el valor teórico de $k_n^{'}$ es cero, igualmente realizaremos las simulaciones como referencia experimental a fin de evaluar el sesgo de $\hat{\mu}$ que desconocemos.

A continuación vamos a describir el comportamiento del sesgo y el ECM de los estimadores, para tamaños muestrales pequeños, a partir de simulaciones del pívot correspondiente ($\mu = 0, \sigma = 1$), utilizando un programa ad-hoc que hemos desarrollado.

n	k_n	k_n'	H_n	H_n'	l_n	l_n'	L_n	L_n'
5	-0.0017	-0.0031	0.0960	0.0964	-0.1435	-0.1433	0.1384	0.1383
6	0.0052	0.0010	0.0776	0.0767	-0.1220	-0.1166	0.1148	0.1162
7	0.0038	0.0023	0.0689	0.0689	-0.1057	-0.1082	0.1013	0.0995
8	-0.0009	-0.0038	0.0589	0.0571	-0.0885	-0.0847	0.0873	0.0868
9	-0.0015	-0.0010	0.0529	0.0517	-0.0770	-0.0744	0.0763	0.0747
10	0.0055	-0.0023	0.0464	0.0446	-0.0679	-0.0688	0.0684	0.0685
11	-0.0032	-0.0015	0.0418	0.0410	-0.0644	-0.0665	0.0620	0.0614
12	0.0011	-0.0002	0.0382	0.0369	-0.0633	-0.0625	0.0588	0.0569
13	-0.0011	0.0031	0.0355	0.0364	-0.0517	-0.0558	0.0535	0.0529
14	0.0025	-0.0012	0.0328	0.0311	-0.0520	-0.0512	0.0495	0.0486
15	0.0020	-0.0007	0.0304	0.0308	-0.0463	-0.0431	0.0461	0.0456
16	-0.0019	0.0013	0.0281	0.0288	-0.0453	-0.0455	0.0427	0.0431
17	0.0019	0.0026	0.0270	0.0264	-0.0453	-0.0403	0.0404	0.0401
18	0.0004	0.0004	0.0247	0.0239	-0.0364	-0.0420	0.0377	0.0375
19	0.0025	-0.0003	0.0230	0.0228	-0.0392	-0.0388	0.0357	0.0354
20	-0.0006	-0.0011	0.0227	0.0222	-0.0334	-0.0340	0.0340	0.0340

Tabla 3.3: Valores relacionados con el sesgo y el ECM de los EMV y estimadores alternativos para distintos tamaños muestrales en base a 10.000 simulaciones.

Para pequeñas muestras ($5 \leq n \leq 20$), y con el programa mencionado, realizamos 2 grupos de 10.000 simulaciones, uno para los EMV y uno para los estimadores alternativos, a fin de evitar la posible correlación (por ejemplo, entre k_n y k_n'). En la tabla 3.3 resumimos los resultados obtenidos.

Parámetro de localización

Analicemos en primer lugar estos datos para los estimadores del parámetro de localización. A partir de la información que nos da la figura 3.4, podemos decir que experimentalmente el sesgo observado de $\hat{\mu}$ es equivalente al de $\tilde{\mu}$. No obstante, cuando el tamaño muestral es pequeño ($n \leq 10$), los valores de k_n presentan mayor fluctuación que los de k_n' cuyo valor teórico, recordemos, es cero por ser $\tilde{\mu}$ un estimador insesgado.

Esto se corrobora si analizamos la figura 3.5, donde hemos representado los valores

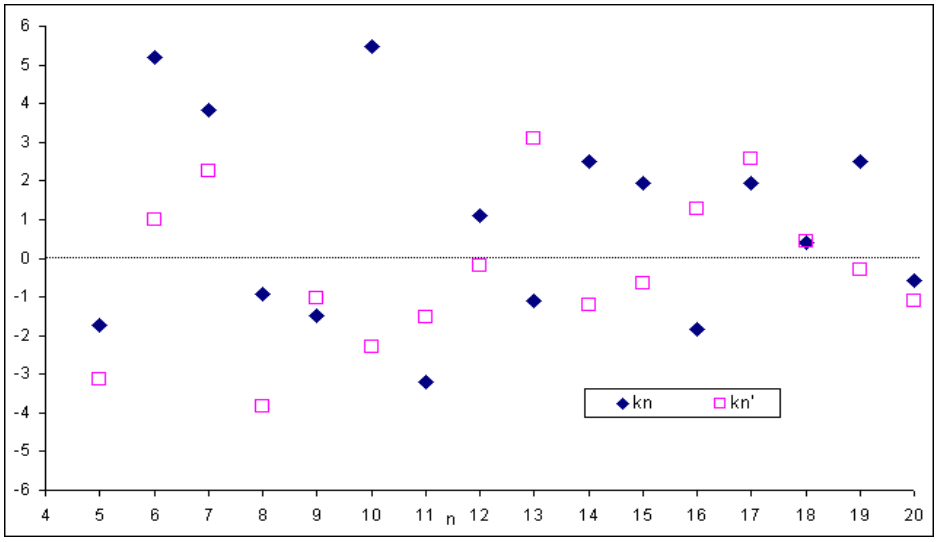

Figura 3.4: Valores de k_n y $k_n^{'}$ (multiplicados por 1000) para distintos tamaños muestrales en base a 10.000 simulaciones

observados de nH_n y $nH_n^{'}$, que nos permite apreciar experimentalmente la similitud del comportamiento del ECM de $\tilde{\mu}$ y $\hat{\mu}$. Además, a medida que el tamaño muestral crece, estos valores van disminuyendo sistemáticamente, tendiendo hacia la tercera serie representada, la constante 0.39762 ($nAvar(\hat{\mu})/\sigma^2 = nAvar(\tilde{\mu})/\sigma^2$). Esto indica lo que ya sabemos, que el ECM observado tiende a hacia la varianza asintótica común.

Parámetro de Escala

Ahora hagamos lo propio para los estimadores del parámetro de escala. Es de destacar el sesgo pronunciado que presentan ambos estimadores cuando el tamaño muestral es pequeño. Más aún, hemos observado que el EMV de σ subestima al parámetro de escala. Este es un resultado experimental interesante del que, lamentablemente, no tenemos ninguna constatación general teórica.

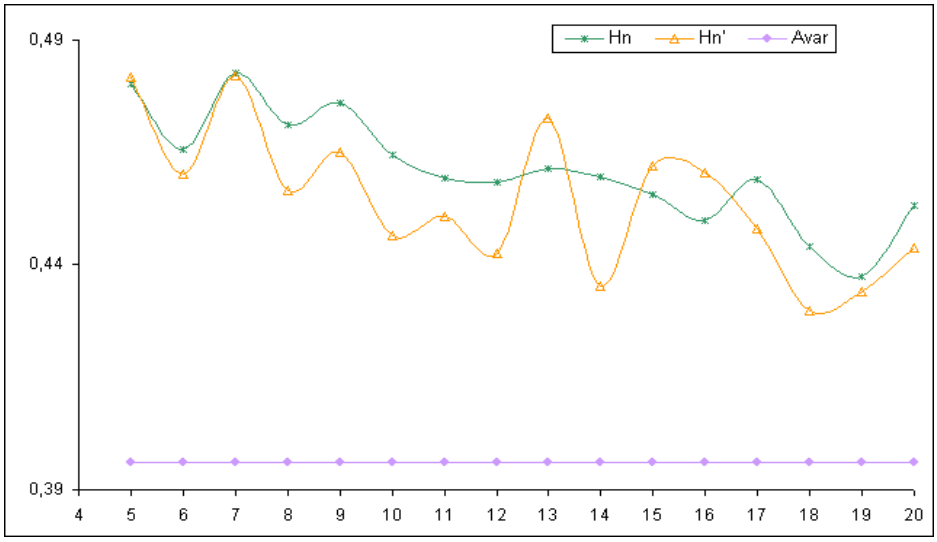

Figura 3.5: Valores de nH_n, $nH_n^{'}$ y $Avar$ para distintos tamaños muestrales en base a 10.000 simulaciones

De todas maneras, la figura 3.6 nos permite apreciar que el sesgo disminuye a medida que el tamaño muestral es mayor a pesar que, incluso para $n = 20$, los valores observados de l_n y $l_n^{'}$ son menores que -0.03. Además, en general no es posible notar diferencias marcadas en cuanto al comportamiento de $\tilde{\sigma}$ y $\hat{\sigma}$.

Respecto al ECM, a partir de la figura 3.7 observamos que, si bien nL_n y $nL_n^{'}$ presentan valores muy fluctuantes cuando n es pequeño, están más acentuados en el caso de $\hat{\sigma}$. Además, los valores observados tienden a estabilizarse rápidamente con el incremento del tamaño muestral hacia la constante 0.6780 ($nAvar(\hat{\sigma})/\sigma^2 = nAvar(\tilde{\sigma})/\sigma^2$).

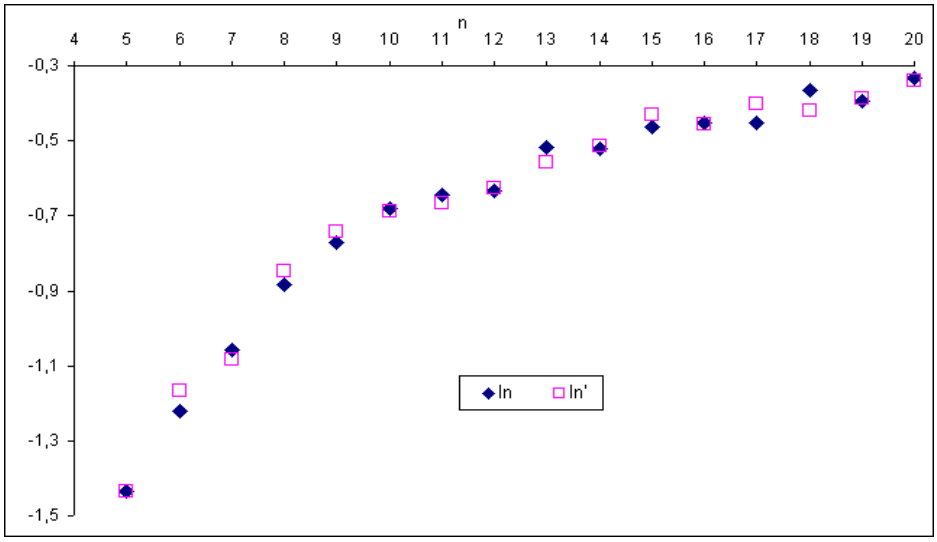

Figura 3.6: Valores de l_n y $l_n^{'}$ (multiplicados por 10) para distintos tamaños muestrales en base a 10.000 simulaciones

3.4 Estimación por Intervalos

En primer lugar, nos interesa construir un intervalo de confianza para el parámetro de localización μ. Debido a que $\frac{\tilde{\mu}-\mu}{\tilde{\sigma}}$ es un pívot, como se demostró en la proposición 3.3.1, y el estimador $\tilde{\mu}$ tienen una distribución asintótica normal con media μ y $Avar(\tilde{\mu}) = 0.39602\sigma^2/n$, se deduce que

$$\sqrt{n}\frac{\tilde{\mu}-\mu}{\tilde{\sigma}\sqrt{0.3960}}, \tag{3.4.1}$$

resulta ser también un pívot tal que su distribución asintótica es una normal estándar.

Esta aproximación a través de la distribución normal nos permite obtener, en el caso de muestras grandes, un **Intervalo de Confianza Aproximado** para μ de la forma habitual, es decir a partir de la expresión,

$$\mu = \tilde{\mu} \pm z_{\alpha/2}\frac{\tilde{\sigma}\sqrt{0.3960}}{\sqrt{n}}, \tag{3.4.2}$$

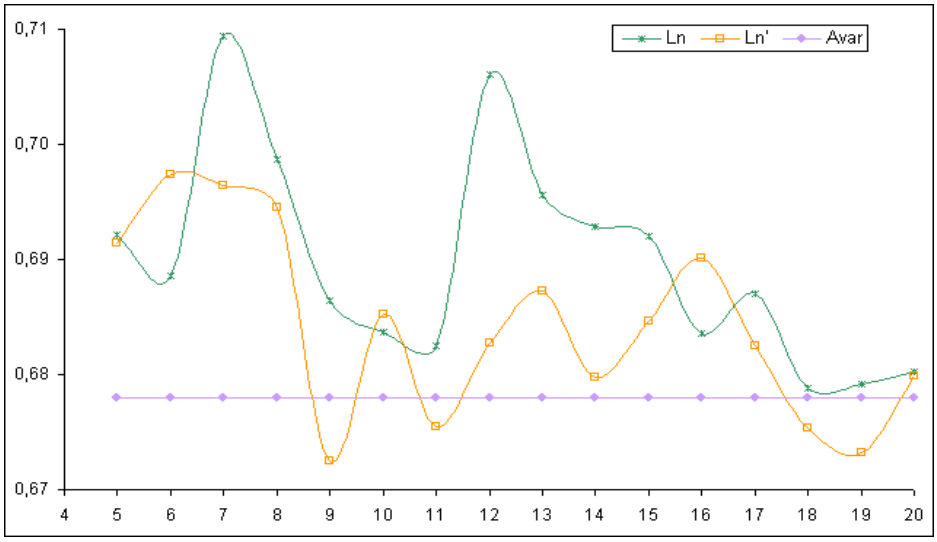

Figura 3.7: Valores de nL_n, $nL_n^{'}$ y $Avar$ para distintos tamaños muestrales en base a 10.000 simulaciones

siendo $z_{\alpha/2}$ el percentil de la normal estándar correspondiente a un nivel de confianza del $100(1-\alpha)\%$.

Sin embargo, no siempre es posible contar con muestras grandes. Muy por el contrario, a menudo debemos realizar inferencias con pequeñas muestras. En estas situaciones no es lícito utilizar la expresión (3.4.2) para construir intervalos de confianza. Para solucionar este problema, en lugar de utilizar la distribución asintótica del pívot, podemos calcular los percentiles de su distribución exacta mediante simulaciones y así obtener un **Intervalo de Confianza Exacto** para la media poblacional μ, presumiblemente de la forma,

$$\mu = \tilde{\mu} \pm y_{\alpha/2} \frac{\tilde{\sigma}\sqrt{0.3960}}{\sqrt{n}},$$

donde ahora $y_{\alpha/2}$ será el percentil de la distribución exacta del pívot correspondiente con un nivel del $100(1-\alpha)\%$.

Tabla 3.4: *Percentiles de la distribución exacta de* $\frac{\sqrt{n}(\tilde{\mu}-\mu)}{\tilde{\sigma}\sqrt{.39602}}$

	$\alpha/2$		
n	0.05	0.025	0.005
5	2.48	3.23	4.73
10	2.06	2.54	3.51
15	1.95	2.42	3.34
20	1.89	2.35	3.13
25	1.84	2.22	3.02
50	1.75	2.09	2.77
100	1.69	2.04	2.68
∞	1.65	1.96	2.57

En la tabla 3.4 presentamos los percentiles ($\alpha/2 = .05, .025, .005$) obtenidos a partir de 10.000 simulaciones para diferentes tamaños muestrales. Para ello, utilizamos la metodología descrita en la sección anterior para simular valores del pívot, con $\theta = 1$ y un tamaño de muestra deseado n, y posteriormente calculamos sus percentiles.

Nota 3.4.1. Observemos que al considerar $y_{1-\alpha/2} = -y_{\alpha/2}$ estamos asumiendo que la distribución exacta del estadístico $\sqrt{n}\frac{\tilde{\mu}-\mu}{\tilde{\sigma}\sqrt{0.3960}}$ es simétrica. Si bien no tenemos una constatación teórica, nos apoyamos en los resultados experimentales que así lo indican, aún cuando el tamaño muestral es pequeño como puede apreciarse en la figura 3.8 para $n = 5$.

Nota 3.4.2. El hecho de poder utilizar un pívot para calcular intervalos de confianza del parámetro de localización no es posible cuando los datos están censurados. Entonces pueden utilizarse otras técnicas como por ejemplo las aproximaciones "Saddlepoint" desarrolladas en el capítulo 5.

Con el mismo razonamiento podemos construir intervalos de confianza aproximados y exactos para el parámetro de escala σ. Como es sabido, el EMV $\hat{\sigma}$ y también el alternativo $\tilde{\sigma}$ tienen una distribución asintótica normal con media σ y varianza

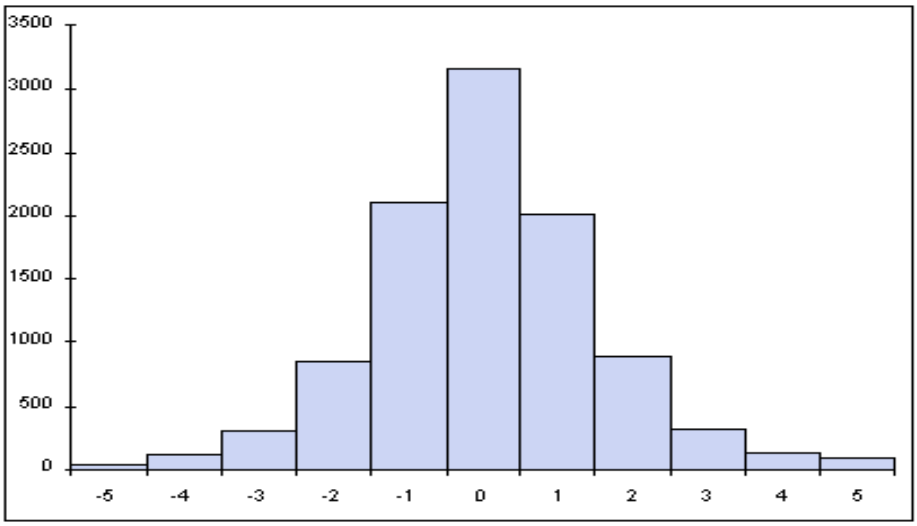

Figura 3.8: Histograma de la distribución exacta del pívot $\frac{\sqrt{n}(\tilde{\mu}-\mu)}{\tilde{\sigma}\sqrt{.39602}}$ para $n = 5$ en base a 10.000 simulaciones

asintótica $\sigma^2 0.6780/n$. Además, por la proposición 3.3.1, $\frac{\tilde{\sigma}}{\sigma}$ es un pívot. Por consiguiente, $\frac{\sqrt{n}}{\sqrt{0.6780}}\frac{(\tilde{\sigma}-\sigma)}{\sigma}$ es también un pívot cuya distribución asintótica es una $N(0,1)$. Esto nos permite construir, en el caso de muestras grandes, un **Intervalo de Confianza Aproximado** para σ de la forma habitual como

$$\sigma \in (\frac{\tilde{\sigma}\sqrt{n}}{\sqrt{n}+z_{\alpha/2}\sqrt{0.6780}}; \frac{\tilde{\sigma}\sqrt{n}}{\sqrt{n}-z_{\alpha/2}\sqrt{0.6780}}),$$

siendo $z_{\alpha/2}$ el percentil de la normal estándar correspondiente a un nivel de confianza del $100(1-\alpha)\%$.

Cuando las muestras son pequeñas, podemos nuevamente calcular los percentiles de la distribución exacta del pívot mediante simulaciones y así obtener un **Intervalo de Confianza Exacto** para σ de la forma

$$\sigma \in (\frac{\tilde{\sigma}\sqrt{n}}{\sqrt{n}+y_{1-\alpha/2}\sqrt{0.6780}}; \frac{\tilde{\sigma}\sqrt{n}}{\sqrt{n}+y_{\alpha/2}\sqrt{0.6780}}),$$

donde ahora y_p es el valor de la distribución exacta del pívot que deja a su derecha un área p.

Nota 3.4.3. Observemos que ahora consideramos ambos percentiles $y_{\alpha/2}$ y $y_{1-\alpha/2}$ debido a que la distribución exacta de $\frac{\sqrt{n}}{\sqrt{0.6780}}\frac{(\tilde{\sigma}-\sigma)}{\sigma}$ no es aparentemente simétrica como podemos apreciar en la figura 3.9 para $n = 5$.

Tabla 3.5: *Percentiles de la distribución exacta de* $\sqrt{n}\frac{(\tilde{\sigma}-\sigma)}{\sigma\sqrt{0.6780}}$

	$\alpha/2$			$1-\alpha/2$		
n	0.05	0.025	0.005	0.95	0.975	0.995
5	-1.75	-1.94	-2.18	1.25	1.61	2.33
10	-1.75	-1.96	-2.42	1.42	1.77	2.56
15	-1.74	-1.99	-2.40	1.46	1.83	2.51
20	-1.74	-1.98	-2.42	1.56	1.91	2.62
25	-1.73	-1.99	-2.44	1.48	1.86	2.50
50	-1.71	-2.04	-2.55	1.53	1.86	2.56
100	-1.69	-1.99	-2.57	1.64	1.98	2.71
∞	-1.64	-1.96	-2.57	1.65	1.96	2.57

En la tabla 3.5 presentamos los percentiles ($\alpha/2 = .05, .025, .005$ y $1-\alpha/2 = .95, .975, .995$) obtenidos a partir de 10.000 simulaciones para diferentes tamaños muestrales y considerando $\theta = 1$.

3.5 Un Ejemplo

A continuación ilustraremos los tópicos hasta aquí desarrollados a través de un ejemplo. Los datos que se presentan en la tabla 3.6 han sido tomados de un estudio de fiabilidad sobre el *graphite H590* (Margetson y Cooper, 1984) y corresponden a las fuerzas de rotura por stress de 41 ejemplares de vigas cortados a partir de un bloque de carbón simple.

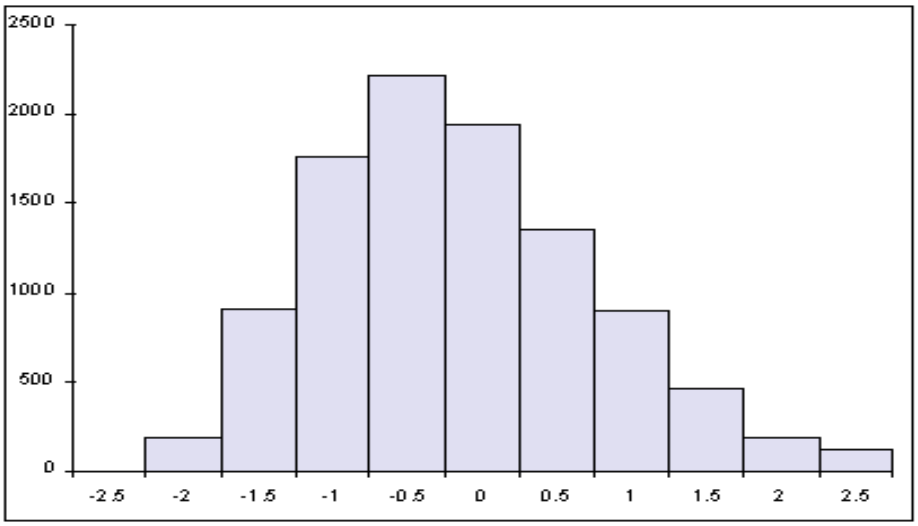

Figura 3.9: Histograma de la distribución exacta del pívot $\frac{\sqrt{n}(\tilde{\sigma}-\sigma)}{\sigma\sqrt{0.6780}}$ para $n = 5$ en base a 10.000 simulaciones

▷ **Ajuste**

Asumiremos que los datos provienen de una población con una distribución de la familia de localización y escala simétrica con $\theta = 1$. El mismo ejemplo es analizado por Cheng y Stephens (1989) quienes, aplicando un test de bondad de ajuste usando el estadístico de Moran con parámetros estimados, rechazan que los datos provengan de una población $N(\mu, \sigma)$; asimismo comentan que con algunos tests de bondad de ajuste bien conocidos, el modelo normal sería aceptado.

En nuestro caso, con un simple Q-Q plot representado en la figura 3.10, podemos observar que los datos se ajustan razonablemente bien mediante la distribución que proponemos. En la figura 3.11 también presentamos el histograma de los datos y, superpuestas, las curvas de los ajustes realizados con la densidad (3.1.1), la normal y la Laplace.

Tabla 3.6: Fuerza de rotura por stress de 41 vigas de H590 graphite (en MPa $\times 10^6$)

27.55	29.89	30.07	30.65	31.23	31.53	31.53
31.82	32.23	32.28	32.69	32.98	33.28	33.28
33.74	33.74	33.86	33.86	33.86	34.15	34.15
34.15	34.44	34.62	34.74	34.74	35.03	35.03
35.32	35.44	35.61	35.61	35.73	35.90	36.20
36.78	37.07	37.36	37.36	37.36	40.28	

▷ **Estimación Puntual**

Para este conjunto de $n = 41$ datos obtenemos los siguientes estadísticos de resumen:

$$\bar{x} = 34.08 \;\; ; \;\; \tilde{x} = 34.15 \;\; ; \;\; min = 27.55 \;\; ; \;\; max = 40.28 \;\; ; \;\; s = 2.39$$

Para obtener el estimador de máxima verosimilitud de los parámetros, debemos maximizar la función de log-verosimilitud numéricamente. En este caso, los EMV de (μ, σ) que obtenemos son,

$$\hat{\mu} = 34.15 \text{ y } \hat{\sigma} = 3.47.$$

En tanto que los estimadores alternativos, resultantes de considerar la combinación lineal de la media y la mediana muestrales, para estos datos son:

$$\tilde{\mu} = 34.12 \text{ y } \tilde{\sigma} = 3.47.$$

A fin de analizar el sesgo de los estimadores en este caso, también hemos realizado 2 grupos de 10.000 simulaciones para un tamaño muestral de $n = 41$ obteniendo lo siguiente: para $\hat{\mu}$ resultó $k_n = 0.0009$ apenas mayor que $k_n^{'} = 0.0002$ correspondiente al valor observado de $\tilde{\mu}$ que está muy próximo a cero, su valor teórico por ser insesgado.

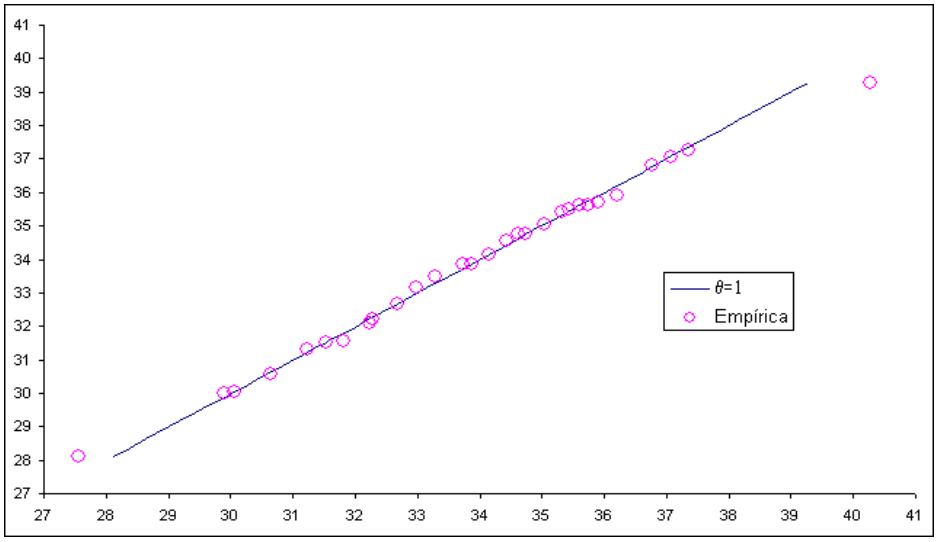

Figura 3.10: Percentiles empíricos versus los esperados para el ajuste mediante $f_1(x;\mu;\sigma)$.

En cambio, para $\tilde{\sigma}$ obtuvimos $l_n^{'} = -0.014$, que si bien es menor a $l_n = -0.018$ el valor correspondiente a $\hat{\sigma}$, sigue indicando un sesgo considerable para ambos estimadores del parámetro de escala. Por ejemplo, para un valor de $\sigma = 3$, cuyo orden de magnitud se corresponde con los datos experimentales, obtendríamos sesgos negativos con valores de 4.2% y 5.4% respectivamente.

▷ Intervalos de Confianza

Realicemos ahora estimaciones por intervalos para la media poblacional μ y el parámetro de escala σ. En primer lugar, utilizando el procedimiento habitual basado en la distribución asintótica normal estándar de ambos pívots, obtenemos para estos datos los siguientes intervalos de confianza aproximados,

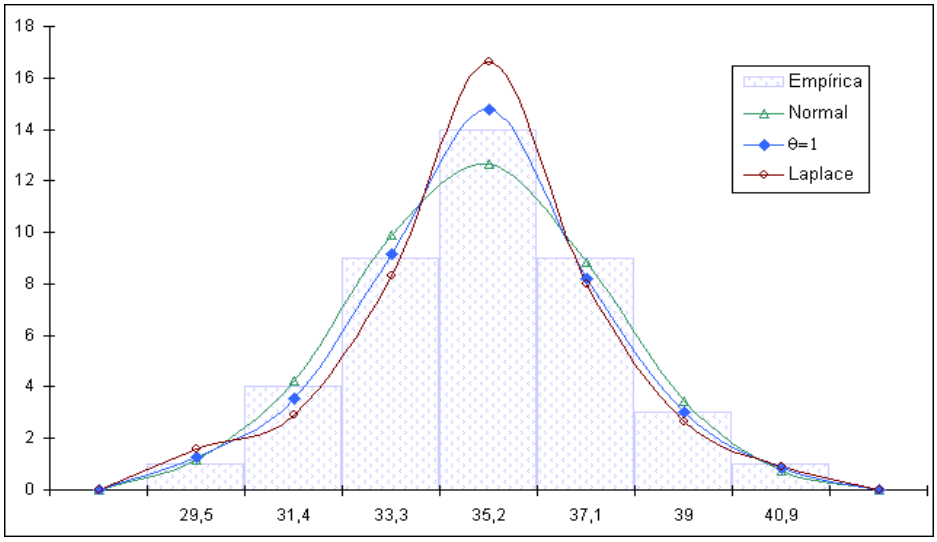

Figura 3.11: Histograma de los Datos y los ajustes por $f_1(x;\mu;\sigma)$, normal y Laplace

Nivel de confianza del 90% $(\alpha/2 = .05)$:

$$\mu \,\epsilon\, (33.56\,;\ 34.68) \text{ y } \ \sigma \,\epsilon\, (2.86\,;\ 4.41).$$

Nivel de confianza del 95% $(\alpha/2 = .025)$:

$$\mu \,\epsilon\, (33.45\,;\ 34.79) \text{ y } \ \sigma \,\epsilon\, (2.77\,;\ 4.64).$$

Para calcular un intervalo exacto para la media poblacional para este ejemplo $(n = 41)$ obtenemos, a partir de las correspondientes simulaciones, que los percentiles $(\alpha/2 = .05, .025)$ del pívot son 1.78 y 2.14 respectivamente. Para el parámetro de escala σ tenemos que, para un 90% de confianza, $y_{.05} = -1.67$ y $y_{.95} = 1.59$, en tanto que para una confianza del 95%, $y_{0.025} = -1.95$ y $y_{0.975} = 1.89$. Por tanto, los intervalos de confianza exactos resultan,

Nivel de confianza del 90%:

$$\mu \,\epsilon\, (33.51\,;\ 34.73) \text{ y } \ \sigma \,\epsilon\, (2.88\,;\ 4.42).$$

Nivel de confianza del 95%:

$$\mu \, \epsilon \, (33.39 \, ; \, 34.85) \text{ y } \; \sigma \, \epsilon \, (2.79 \, ; \, 4.63).$$

Observemos que no existe prácticamente diferencia entre los intervalos de igual nivel de confianza construidos para el parámetro de escala mediante la aproximación normal o la distribución exacta del pívot. En tanto que los correspondientes al parámetro de localización construidos con la distribución exacta presentan un pequeño incremento en su amplitud (< 0.1) respecto a los obtenidos a partir de la aproximación normal.

Por último, supongamos que los datos provienen de una distribución normal. Entonces, un intervalo de confianza para la media vendría dado por $\bar{x} \pm t_{\alpha/2} s/\sqrt{n}$, siendo $t_{\alpha/2}$ el percentil de la distribución t-student con $n-1$ grados de libertad para un nivel de confianza del $100(1-\alpha)\%$. Para nuestro ejemplo obtenemos los siguiente intervalos,

$$\text{Nivel de confianza del 90\%:} \quad \mu \, \epsilon \, (33.33 \, ; \, 34.83).$$

$$\text{Nivel de confianza del 95\%:} \quad \mu \, \epsilon \, (33.21 \, ; \, 34.95).$$

En este caso y para cada nivel de confianza, los intervalos obtenidos son de mayor amplitud que los construidos tanto con la aproximación normal como con la distribución exacta del pívot, asumiendo que los datos provienen de una distribución con densidad dada por (3.1.1). Observemos además, que el método de la t de student es considerado "robusto" y por tanto es posible utilizarlo incluso cuando no hay normalidad pero el tamaño muestral es relativamente grande.