UNIVERSIDAD CENTROCCIDENTAL
"LISANDRO ALVARADO"

Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

"ESTIMACION POR MAXIMA VEROSIMILITUD PARA LA DISTRIBUCION NORMAL MULTIVARIADA. "

Trabajo Especial de Grado presentado por

Br. José Gregorio Chirino R.

como requisito final
para obtener el título de Licenciado
en Ciencias Matemáticas
Área de Conocimiento: Probabilidad y Estadística.
Tutor: Ms.C. Luz Rodriguez.

Barquisimeto, Venezuela. Junio de 2009

Universidad Centroccidental
"Lisandro Alvarado"
Decanato de Ciencias y Tecnología
Licenciatura en Ciencias Matemáticas

ACTA
TRABAJO ESPECIAL DE GRADO

Los suscritos miembros del Jurado designado por el Jefe del Departamento de Matemáticas del Decanato de Ciencias y Tecnología de la Universidad Centroccidental "Lisandro Alvarado", para examinar y dictar el veredicto sobre el Trabajo Especial de Grado titulado:

"ESTIMACION POR MAXIMA VEROSIMILITUD PARA LA DISTRIBUCION NORMAL MULTIVARIADA. "

presentado por el ciudadano Br. José Gregorio Chirino R. titular de la Cédula de Identidad No. 16.749.418, con el propósito de cumplir con el requisito académico final para el otorgamiento del título de Licenciado en Ciencias Matemáticas.

Luego de realizada la Defensa y en los términos que imponen los Lineamientos para el Trabajo Especial de Grado de la Licenciatura en Ciencias Matemáticas, se procedió a discutirlo con el interesado habiéndose emitido el veredicto que a continuación se expresa:

[1] _____

Con una calificación de _____ puntos.

En fe de lo expuesto firmamos la presente Acta en la Ciudad de Barquisimeto a los _____ días del mes de _____ de _____.

_____	_____
TUTOR	FIRMA
_____	_____
PRINCIPAL	FIRMA
_____	_____
PRINCIPAL	FIRMA

OBSERVACIONES:

_____

_____

_____

[1] Aprobado ó Reprobado

A mis padres y madres: José Francisco (Cheo), Nerio de Jesús, Magaly y Dulce (mi Tiamamá).
Sin ustedes, nada....!

AGRADECIMIENTOS

Primeramente doy GRACIAS A DIOS por colocarme en este mundo, en esta familia, con estos amigos, en esta casa de estudios y en esta carrera que carga con ella un grupo de profesores con gran pedagogia y un sin fin de conocimientos para impartir.

En forma especial, gracias a mis padres Jose F. (Cheo), Magaly (Maga), Nerio (Mi viejo) y Dulce (mi Tiamamá) por guiarme por el mejor camino y por siempre brindarme su apoyo y cuidados, dejandome así enseñanzas en cada experiencia compartida.

A mi morena bella, Yarkin Corona, mi gran amor, gracias por todo el amor tan bello, sutíl y sublime que me has entregado, que a pesar de adversidades nos ha mantenido juntos, por eso y más gracias... Te Amo!

Para mis hermanos Jessica y Jonathan, su fidelidad, apoyo, amor y preocupación fueron muy bien invertidos en mi, les debo mucho y cuando digo mucho es MUCHO... Fueron motivo de inspiración para éste logro, al igual que mi sobrina bella, Valeska (mi Tatu), para quien espero el mejor futuro y que revase todas mis metas y logros. Junto a ellos quiero agradecer a mi cuñado Junior su apoyo y amistad incondicional.

A mi familia (Chirino y Ramirez) son una combinación, un kit, un combo fenomenal... A ustedes gracias por su apoyo y por su lucha por mantener vivo el principio familiar, la unidad. A todos los quiero mucho.

A mi tutora, mi profe Luz Rodriguez, GRACIAS por su guia, regaños que nunca fueron regaños sino consejos y por toda la paciencia que puso en mi tutoria, pero sobretodo GRACIAS por su amistad, pués sé que cuento con ella. Dios me la bendiga a usted y a su familia que está en proceso de expansión.

Y por último y no por eso menos importante a mis amigos y compañeros Marisela (Chetes), Gaby, Gladimar, Nestor (Memo), Marco, Jessica, Luis F, Mario, David, Adrian, Manuel, Elifer y Eliezer gracias por tantos momentos gratos, conversaciones inolvidables y por su amistad incondicional, en mi estas semillas siempre tendran buenos frutos.

A todos Gracias...!

INTRODUCCIÓN

El análisis multivariado es una rama de la estadística dedicada al estudio de variables aleatorias las cuales están correlacionadas entre sí.

La esencia de la aplicación del análisis multivariado envuelve la motivación de resolver problemas y llegar a respuestas númericas, o generar grandes opiniones acerca de un fenómeno natural, así como también proveer resultados que pueden ser usados como base para tomar decisiones.

Por otra parte, uno de los métodos más usados, para la estimación de parámetros es el método de Máxima Verosimilitud, las ventajas de usar dicho método serán discutidas y se muestra que éste se deriva asumiendo una distribución Gaussiana Multivariada para los datos.

Es bien conocido que la estimación por Máxima Verosimilitud es un atractivo método de inferencia paramétrica en estadística, tanto para el caso univariado como el multivariado. Por esa razón, en este trabajo se establece formalmente la teoría correspondiente a los Métodos de Estimación Paramétrica por Máxima Verosimilitud en el caso Multivariado.

Específicamente se estudia la distribución normal multivariada y algunos ejemplos son presentados para ilustrar la teoría. Este trabajo servirá como material de apoyo para trabajos futuros relacionados con el área de probabilidad y estadística, particularmente con el análisis multivariado.

ÍNDICE

Agradecimientos **i**

1. Antecedentes. **1**

1.1. Inferencia basada en máxima verosimilitud. 2

1.2. Máxima verosimilitud. 2

1.3. Naturaleza del problema de estimación. 3

2. Conceptos Fundamentales. **5**

2.1. Traza de una matriz. 5

2.1.1. Algunas propiedades de la traza. 5

2.2. Derivada de una función escalar de una matriz. 6

2.2.1. Propiedades de la derivada de una función. 6

2.2.2. Derivada de un vector respecto al vector Hessiano. 7

2.3. Variables aleatorias. 8

2.4. Función de distribución acumulada (Fda). 9

2.4.1. Propiedades de la Fda. 9

2.5. Densidad. 11

2.6. Distribución Marginal. 12

2.7. Distribución condicional. 12

2.8. Independencia. 13

2.9. Esperanza. 14

2.10. Momento de segundo orden. 14

2.11. La distribución Normal Mutivariada. 15

2.11.1. Densidad general. 16

2.11.2. Media y Covarianza muestral. 16

2.11.3. Distribución Normal Bivariante. 16

2.11.4. Independencia. 18

2.11.5. Estandarización. 18

2.12. Distribucion Wishart. 19
2.12.1. Propiedades de la distribución Wishart. 20
2.13. Criterio de Suficiencia. 21

3. Método de máxima verosimilitud para la distribución normal multivariada. 25

Referencias 39

Capítulo 1

Antecedentes.

El método de máxima verosimilitud no fué formulado o usado como una técnica general de estimación hasta que R.A. Fisher en 1912 introdujo una forma generalizada y rigurosa de éste. Fisher, más que introducir el concepto simple, publicó una serie de artículos en los cuales extendió estos conceptos a un comprensivo y unificado sistema de estadística matemática al igual que una filosofía de inferencia estadística la cual ha tenido un profundo y ancho desenvolvimiento.

Los problemas teóricos inherentes a los estimadores de máxima verosimilitud son principalmente aquellos concernientes a las propiedades de varianza de los estimadores, particularmente cuando el tamaño de la muestra es pequeña. Para tamaños de muestras grandes y cuando las observaciones son independientes, la teoría ha sido bien desarrollada, y existen una variedad de resultados para las propiedades asintóticas de los estimadores, en particular para la varianza de los mismos. A pesar de algunas desventajas, particularmente cuando los tamaños muestrales son pequeños, el método de máxima verosimilitud es atractivo, casi como una técnica práctica universal para formular ecuaciones de estimación.

Mardia y Marshall (1984) consideraron las propiedades asintóticas de los estimadores de máxima verosimilitud, y mostraron que las propiedades asintóticas usuales de consistencia y normalidad asintótica se satisfacen bajo la condición de dominio asintótico creciente.

La estimación por máxima verosimilitud (ML) es uno de los métodos más importantes para la estimación de parámetros en campos aleatorios Gaussianos. Ello es así por la versatilidad y buenas propiedades estadísticas que en general poseen los métodos inferenciales basados en la verosimilitud.

En la práctica, algunos investigadores prefieren hacer inferencia por ML (Mardia y Marshall, 1984; Vecchia, 1988; Jones y Vecchia, 1993), miéntras que otros prefieren usar otros métodos (Cressie, 1993; Zimmerman y Harville, 1991). Estos últimos argumentan que los estimadores ML de parámetros de covarianza son sesgados, sobre todo cuando se tienen muestras pequeñas y/o cuando el modelo incluye varios parámetros de regresión. Sin embargo, el sesgo es sólo un aspecto de un estimador, y para establecer conclusiones más sólidas, la varianza del estimador debe ser considerada, o mejor aún, el error cuadrático medio del estimador.

1.1. Inferencia basada en máxima verosimilitud.

Los métodos de inferencia basados de alguna forma en verosimilitud, entre estos el método de máxima verosimilitud, son tal vez los métodos más versátiles para ajustar datos de modelos estadísticos. En aplicaciones típicas, la meta es usar un modelo paramétrico para describir un conjunto de datos o un proceso que genere un conjunto de datos. Aparte de su fuerte motivación intuitiva, el mayor atractivo de los métodos estadísticos basados en alguna forma de verosimilitud consiste en que éstos pueden ser aplicados a una gran variedad de modelos y clases de datos (continuos, discretos, categóricos, censurados, truncados, etc.), donde otros métodos populares, tales como mínimos cuadrados, no proveen en general un método satisfactorio para hacer inferencia estadística.

Kitanidis (1983) fue (aparentemente) el primero en proponer el uso de métodos basados en la verosimilitud para estimación de parámetros en campos aleatorios Gaussianos. El uso del método de máxima verosimilitud para la estimación de parámetros en campos aleatorios Gaussianos fue desarrollado y estudiado por Kitanidis y Lane (1985), Mardia y Marshall (1984), Mardia y Watkins (1989), Jones y Vecchia (1993) y Vecchia (1988), entre otros.

1.2. Máxima verosimilitud.

Existen métodos de estimación que están basados en distribuciones iniciales y funciones de pérdida, pero es útil poder aplicar un método relativamente sencillo

para construir un estimador sin tener que especificar una función de pérdida y una distribución inicial. A continuación se describe un método con éstas características que se denomina Método de Máxima Verosimilitud que se puede aplicar a la mayoría de los problemas, tiene un fuerte atractivo intuitivo y usualmente proporciona una estimación razonable de θ.

Además, si la muestra es grande, el método de máxima verosimilitud es quizas el método de estimación más ampliamente utilizado en estadística.

La función de verosimilitud provee, en general, una poderosa herramienta para cuantificar la información que los datos observados poseen acerca de los parámetros desconocidos. En el marco del modelo y los datos descritos antes, la función de log-verosimilitud del vector de parámetros

$$\eta = (\beta, \sigma^2, \vartheta) \in \Omega = \mathbb{R}^p \times (0, \infty) \times \Theta,$$

basada en los datos observados $z = (z_1, \ldots, z_n)'$, $z_i = z(s_i)$, viene dada, salvo por una constante aditiva, por

$$l(\eta; z) = -\tfrac{n}{2}\, \log(\sigma^2) - \tfrac{1}{2}\, \log(|\Sigma_\vartheta|) - \tfrac{1}{2\,\sigma^2}\, (\mathbf{z} - X\beta)' \Sigma_\vartheta^{-1} (z - X\beta).$$

El estimador de máxima verosimilitud de η, suponiendo que éste existe, es el vector $\hat{\eta}^{ml} = (\hat{\beta}, \hat{\sigma}^2, \hat{\vartheta}) \in \Omega$ que maximiza $l(\eta; z)$ como función de η, para z fijo, es decir, $\hat{\eta}^{ml} = \arg \max_{\eta \in \Omega} l(\eta; z)$.

1.3. Naturaleza del problema de estimación.

Supongamos que se va a seleccionar una muestra aleatoria $X_1, ..., X_n$ de una distribución cuya función de densidad de probabilidad es $f(x|\theta)$ donde el valor del parámetro θ es desconocido.
Supóngase además que el valor de θ debe pertenecer a un intervalo concreto Ω sobre la recta real, y que el valor de θ se debe estimar a partir de los valores observados de la muestra.

Un estimador del parámetro θ, basado en variables aleatorias $X_1, ..., X_n$ es una función $\delta(X_1, ..., X_n)$ que especifica el valor esperado de θ para cada conjunto de valores posibles de $X_1, ..., X_n$. En otras palabras, si los valores observados de $X_1, ..., X_n$ son $x_1, ..., x_n$ entonces el valor del estimador de θ es $\delta(x_1, ..., x_n)$ puesto que el valor de θ debe pertenecer al intervalo Ω.

Es conveniente distinguir los términos estimador y estimación.

Un estimador $\delta(X_1, ..., X_n)$ es una función de los variables aleatorias $X_1, ..., X_n$ y el es una variable aleatoria cuya distribución de probabilidad se puede obtener a partir de la distribución conjunta de $X_1, ..., X_n$.

Por otro lado, una estimación es un valor específico $\delta(x_1, ..., x_n)$ del estimador que se determina utilizando valores observados específicos $x_1, ..., x_n$ de la muestra aletoria $(X_1, ..., X_n)$.

Capítulo 2

Conceptos Fundamentales.

Este capítulo tiene como objetivo dar una breve introducción a la teoría de probailidad y estadística, haciendo énfasis en el método de máxima verosimilitud, estableciendo definiciones, propiedades y notaciones, que son necesarias para el buen desarrollo del trabajo.

2.1. Traza de una matriz.

Definición 2.1. La suma de los elementos de la diagonal de una matriz cuadrada es llamada la traza, es decir, si $A = (a_{ij})$ con $i, j = 1, \ldots, p$

$$trA = \sum_{j=1}^{p} a_{jj}.$$

Esta definición será de gran utilidad, por ejemplo, para definir algunas distribuciones de probabilidad para caso multivariado.

2.1.1. Algunas propiedades de la traza.

1. Supongamos que $A : p \times x$, $B : n \times p$, entonces

$$tr(AB) = tr(BA).$$

Por ejemplo, si x es un vector $p \times 1$

$$tr(xx') = tr(x'x) = x'x.$$

Este resultado se obtendrá porque el elemento ij de AB es $\sum_{\alpha=1}^{n} a_{i\alpha}b_{\alpha j}$ así,

$$tr(AB) = \sum_{i=1}^{p}\sum_{\alpha=1}^{n} a_{i\alpha}b_{\alpha j}.$$

Además, el elemento ij de BA es $\sum_{\alpha=1}^{n} b_{i\alpha}a_{\alpha j}$, así

$$tr(BA) = \sum_{i=1}^{n}\sum_{\alpha=1} b_{i\alpha}a_{\alpha j}.$$

2. Si $A : p \times n$, $B : n \times p$, entonces

$$tr(A + B) = tr(A) + tr(B).$$

3. Si α es un escalar y $A : p \times p$,

$$tr(\alpha A) = \alpha tr(A).$$

4. Si A es un escalar, $tr(A) = A$.

Por ejemplo, si $x : p \times 1$, $A : p \times p$, $x'Ax$ es un escalar. Así

$$tr(Axx') = x'Ax.$$

2.2. Derivada de una función escalar de una matriz.

Definición 2.2. La derivada de una función escalar f de una matriz $X = (x_{ij})$, con $i = 1, \ldots, p$ y $j = 1, \ldots, n$, esta definida como:

$$\frac{\partial f(X)}{\partial X} = \left(\frac{\partial f(X)}{\partial x_{ij}}\right), \qquad \begin{matrix} i = 1, \ldots, p \\ j = 1, \ldots, n \end{matrix}$$

2.2.1. Propiedades de la derivada de una función.

1. Sea $X : p \times p$, $|X| \neq 0$

Para $X \neq X'$,
$\frac{\partial}{\partial X}|X| = |X|(X^{-1})'$, para $X : p \times p, |X| \neq 0$.

Para $X = X'$,
$\frac{\partial}{\partial X}|X| = 2|X|X^{-1} - diag(|X|X^{-1})$.

2. Para $A' : p \times q$, $X : q \times p$
$$\frac{\partial}{\partial X} tr(A'X) = A.$$

3. Para $A' : p \times q$, $X : q \times p$
$$\frac{\partial}{\partial X} tr(X'A) = A.$$

4. Si $x : p \times 1$, $A : p \times p$
$$\frac{\partial}{\partial x}(x'Ax) = 2Ax.$$

5. Si $X : p \times p$ y $X = X'$,
$$\frac{\partial}{\partial X} trX^2 = 2X.$$

La prueba es inmediata de la definición.

2.2.2. Derivada de un vector respecto al vector Hessiano.

Definición 2.3. Sea $f(x)$ una función escalar del vector $x : p \times 1$. Entonces

$$\frac{\partial}{\partial x \partial x'} f(x) = \frac{\partial}{\partial x} \frac{\partial f(x)}{\partial x'}$$

Ejemplo:

Suponga que $f(x) = x'x$, donde $x = (x_i)$, $i = 1, ..., p$.

$$\frac{\partial^2}{\partial x \partial x'} f(x) = \frac{\partial}{\partial x} \left[\frac{\partial (x'x)}{\partial x'} \right] = \frac{\partial}{\partial x}(2x').$$

Pero,

$$\begin{aligned} \frac{\partial}{\partial x}(x') &= \left[\frac{\partial}{\partial x}(x_1), ..., \frac{\partial}{\partial x}(x_p) \right] \\ &= \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \\ &= I_p. \end{aligned}$$

Para $x : p \times 1$, $\frac{\partial x}{\partial x'} = I_p$ y
$$\frac{\partial^2}{\partial x \partial x'}(x'x) = 2I.$$

Una generalización inmediata de este resultado es la siguiente. Si $A : p \times p$ y $x : p \times 1$,

$$\frac{\partial^2}{\partial x \partial x'}(x'Ax) = A.$$

La matriz Hessiana de una función escalar $f(x)$ de un vector x, es definida como una matriz simétrica

$$H = \frac{\partial^2}{\partial x \partial x'} f(x);$$

donde,

$$h_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x).$$

La matriz es útil para examinar si una función tiene un valor extremo.

Suponga $x : p \times 1$ y que para $x = x_0$, $\frac{\partial}{\partial x} f(x) = 0$. Así, x_0 es un punto estacionario de $f(x)$. Si además, $H > 0$ para todo x, x_0 corresponde a un mínimo global de $f(x)$. Alternativamente, si $H < 0$ para todo x, x_0 correponde a un máximo global de $f(x)$. Aunque, en general, H no es definido positivo ni definido no-negativo sobre el rango entero de x, en análisis multivariado aplicado, la función objetivo involucra a menudo Hessianos.

Observación:

Funciones $f(x)$ con $H > 0$ son llamadas convexas, si $H < 0$ son llamadas concavas. Más general, $f(x)$ es convexa si para todo par de puntos x_1 y x_2 y para cualquier λ, con $0 < \lambda < 1$,

$$f[\lambda x_1 + (1 - \lambda)x_2] \leq \lambda f(x_1) + (1 - \lambda) f(x_2).$$

Si la inecuación no se cumple, $f(x)$ es concava.

2.3. Variables aleatorias.

La relación entre los sucesos del espacio muestral y el valor numérico que se les asigna se establece a través de variables aleatorias.

Definición 2.4. Una variable aleatoria es una función que asigna un valor numérico a cada suceso elemental del espacio muestral. Es decir, una variable aleatoria es una variable cuyo valor numérico está determinado por el resultado del experimento aleatorio. La variable aleatoria la denotaremos con letras mayúsculas $X, Y, ...$ y con las letras minúsculas $x, y, ...$ sus valores.

La variable aleatoria puede tomar un número numerable o no numerable de valores, dando lugar a dos tipos de variables aleatorias discretas y continuas.

Definición 2.5. Se dice que una variable aleatoria X es discreta si puede tomar un número finito o infinito, pero numerable, de posibles valores.

Definición 2.6. Se dice que una variable aleatoria X es continua si puede tomar un número infinito no numerable de valores, o bien, si puede tomar un número infinito de valores correspondientes a los puntos de uno más intervalos de la recta real.

2.4. Función de distribución acumulada (Fda).

Definición 2.7. Sean X, Y dos variables definidas conjuntamente, es decir, X e Y tienen una distribución de probabilidad conjunta cuya funcion de distribución acumulada (Fda) conjunta está dada por:

$$F(x, y) = P\{X \leq x, Y \leq y\}.$$

De manera general, cuando $X' = (X_1, \ldots, X_p)$ es un vector de variables aleatorias que son distribuidas conjuntamente, la Fda ésta dada por:

$$F(x) = F(x_1, \ldots, x_p) = P\{X_1 \leq x_1, \ldots, X_p \leq x_p\}.$$

2.4.1. Propiedades de la Fda.

Toda Fda multivariada F satisface las siguientes propiedades:

1. F es monótona no decreciente en cada componente de X.

   Basta probar que $F(E) \geq 0$ con $E \subset X \subset \mathbb{R}$.
   Sabemos que $P\{a \leq X \leq b\} = F(b) - F(a) = F(E)$, definiendo $X = \Omega$ (el espacio muestral) tenemos que $E \subset X$ es un evento, luego por axioma de probabilidad

   $$P\{E\} \geq 0 \Rightarrow F(E) \geq 0$$

   Así, se cumple lo que se queria probar.

2. $0 \leq F(x) \leq 1$

   Sea $S \subset X = \Omega$ (evento).

   Por lo anterior $F(S) \geq 0$. Falta probar que $F(S) \leq 1$ ahora $\Omega = S \subset S^c$. Entonces:

   $$P\{\Omega\} = P\{S \cup S^c\} = P\{S\} + P\{S^c\} = 1$$

   $$\Rightarrow P\{S\} = 1 - P\{S^c\} \leq 1 \qquad \text{(ya que } P\{S^c\} \geq 0\text{)},$$

   luego, $0 \leq P\{S\} \leq 1$, y como S es un evento arbitrario, se cumple la propiedad.

3. $F(-\infty, x_2, \ldots, x_p) = F(x_1, -\infty, \ldots, x_p) = \ldots = F(x_1, x_2, \ldots, -\infty)$.

   Sabemos que,

   $$F(x_1, x_2, \ldots, x_p) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_p} f(x_1, x_2, \ldots, x_p) dx_p \ldots dx_1,$$

   donde f es la función de densidad de $X = (X_1, \ldots, X_p)$. Como f es continua

   $$\begin{aligned} F(-\infty, x_2, \ldots, x_p) &= \int_{-\infty}^{-\infty} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{x_p} f(x_1, x_2, \ldots, x_p) dx_p \ldots dx_2 dx_1 \\ &= \int_{-\infty}^{x_p} \int_{-\infty}^{x_2} \ldots \int_{-\infty}^{-\infty} f(x_1, x_2, \ldots, x_p) dx_1 dx_2 \ldots dx_p \\ &= 0 \end{aligned}$$

   Análogamente se prueba para $F(x_1, -\infty, \ldots, x_p)$.

   Por lo tanto,

   $$F(-\infty, x_2, \ldots, x_p) = F(x_1, -\infty, \ldots, x_p) = \ldots = F(x_1, x_2, \ldots, -\infty).$$

4. $F(\infty, \infty, \ldots, \infty) = 1$

   $$F(\infty, \infty, \ldots, \infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_p) dx_p \ldots dx_2 dx_1 = 1,$$

   ya que f es una función de densidad de X.

5. La probabilidad de un rectángulo de dimensión p es no negativo. Probemos para $p = 2$

$$\begin{aligned} P\{x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2\} &= F(x_2, y_2) - F(x_2, y_1) - F(x_1, y_2) + F(x_1, y_1) \\ &\geq 0 \end{aligned}$$

Todas las propiedades son análogas al caso univariado excepto la última propiedad, existen funciones que cumplen las 4 propiedades y la última no la cumple, así no son Fda.

EJEMPLO:

Supongamos que:

$$F(x_1, x_2) = \begin{cases} 0, & si\ x_1 \leq 0\ o\ x_1 \leq 0, x_1 + x_2 \leq 1 \\ 1, & \text{en otro caso} \end{cases} \tag{2.1}$$

esta función satisface las 4 primeras propiedades, lo cual es suficiente para una Fda en el caso univariado, pero dado que

$$F(1,1) - F(1, \frac{1}{2}) - F(\frac{1}{2}, 1) + F(\frac{1}{2}, \frac{1}{2}) = -1$$

no es una Fda ya que no cumple con la última propiedad, así $F(x_1, x_2)$ no puede ser una Fda bivariada.

Asumiremos que todas las funciones $F(x)$ serán continuas, en consecuencia ésta será expresada como la integral de una función $f(x)$ llamada densidad, es decir:

$$F(X) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \dots \int_{-\infty}^{x_p} f(x) dx$$

2.5. Densidad.

Definición 2.8. Supongamos que $F(X)$ es continua, entonces del ejemplo anterior, la función de densidad conjunta (Fdc) de X es:

$$f(x) = f(x_1, \dots, x_p) = \frac{\partial^p}{\partial x_1 \dots \partial x_p} F(x).$$

Hay conjuntos donde los valores de x en la forma anterior no existen. Análogo al caso univariado, esto es una relación para la probabilidad de un evento (o conjunto de valores en el espacio de dimensión p) en términos de la densidad conjunta para $X : p \times 1$

$$P\{X \subseteq R\} = \int_R \dots \int f(x)dx. \tag{2.2}$$

para una región R.

2.6. Distribución Marginal.

En el análisis de datos multivariados, es típico comenzar con un vector con muchas componentes y, luego, encontrar posteriormente un subvector de interés. En tal caso, la distribución marginal de los subvectores es importante para la inferencia proporcional.

Definición 2.9. Sea $X' = (Y', Z')$, donde Y y Z son subvectores de $X : p \times 1$ [por ejemplo, $Y' \equiv (X_1, X_2)$, $Z' \equiv (X_3, \dots, Xp)$], entonces si $g(y)$, $h(z)$ denotan las densidades de Y, Z respectivamente, y si $f(x) = f(y, z)$ denota la densidad de X se tiene que:

$$g(y) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y, z)dz,$$

$$h(z) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(y, z)dy,$$

donde todas las integrales son tomadas sobre $(-\infty, \infty)$, $g(y)$ y $h(z)$ son llamadas las densidades marginales de Y y Z.

2.7. Distribución condicional.

La distribución condicional es de interés, y ocurre cuando un grupo de variables aleatorias están siendo estudiadas miéntras un segundo grupo se mantiene fijo.

Definición 2.10. Sean A y B dos eventos que pueden ocurrir en un espacio de 2-dimensiones, entonces por definición, la probabilidad condicional de B dado A esta dada por:

$$P(B|A) = \frac{P(AB)}{P(A)},$$

si $P(A) \neq 0$.

Si A es un evento donde la variable aleatoria X está en el intervalo $a \leq X \leq b$, y B es un evento donde la variable aleatoria Y esta en el intervalo $c \leq Y \leq d$, entonces

$$P\{c \leq Y \leq d | a \leq X \leq b\} = \frac{P\{a \leq X \leq b, c \leq Y \leq d\}}{P\{a \leq X \leq b\}}$$

y por (2.2)

$$P\{c \leq Y \leq d | a \leq X \leq b\} = \frac{\int_c^d \int_a^b f(x,y)dxdy}{\int_a^b g(x)dx}$$

donde $f(x,y)$ es la densidad conjunda de X, Y y $g(x)$ es la densidad marginal de X. La densidad condicional de Y dado $X = x$ esta definida como:

$$h(y|x) = \frac{f(x,y)}{g(x)}.$$

Así,

$$P\{c \leq Y \leq d | X = x\} = \int_c^d h(y|x)dy.$$

Generalizando a una dimensión p, sean $X' = (X_1, ..., X_p)$, $Y' = (X_1, ..., X_k)$ y $Z' = (X_{k+1}, ..., X_p)$ los vectores aleatorios y con letra minúscula denotaremos los valores observados. La densidad condicional de Y dado Z está dado por:

$$g(Y|Z) = \frac{f(y,z)}{h(z)} = \frac{f(x)}{h(z)}$$

donde $f(x)$ denota la densidad del vector aleatorio X, y $h(z)$ denota la densidad marginal del vector Z.

2.8. Independencia.

Dos vectores aleatorios Y y Z se dicen que son independientes si se cumple alguna de las siguientes condiciones:

$$f(y,z) = g(y)h(z),$$

o

$$F(y,z) = G(y)H(z),$$

o

$$P(y|z) = g(y),$$

donde $f(y,z)$, g(y) y h(z) son las densidades de $X = (Y,Z)$, Y y Z respectivamente; F, G y H son las respectivas Fda, y $P(y|z)$ es la densidad condicional de $Y|Z$.

2.9. Esperanza.

Sea $X : p \times 1$ un vector columna con X_i, con $i = 1, ..., p$, componentes aleatorias, donde $f(X) = f(x_1, ..., x_p)$ es la función de densidad conjunta.
Cuando ésta existe, la esperanza de un vector X esta definido como:

$$E(X) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix}$$

Análogamente, si $V : p \times n$, $E(V) = (E(V_{ij}))$, donde $V = (V_{ij})$.

2.10. Momento de segundo orden.

La confianza entre dos variables aleatorias Y y Z con momento de segundo orden finito, está definido como:

$$cov(Y,Z) = E[(Y - E(Y))(Z - E(Z))]$$

ésto cuantitativamente puede ser positivo, negativo o cero, la covarianza matricial de un vector X está dada como sigue:

$$\Sigma = (\sigma_{ij}) = E[(X - E(X))(X - E(X))']$$

para $i, j = 1, ..., p$. Un elemento típico de Σ es $\sigma_{ij} = E[(X_i - E(X_i))(X_j - E(X_j))']$, $i, j = 1, ..., p$ cuando $j = i$ los elementos están ubicados a lo largo de la diagonal de Σ y es llamada la varianza X.
Recordemos que

$$Var(X_i) = E(X_i - E(X_i))^2$$

si $i \neq j$, σ_{ij} es la covarianza de X_i y X_j. El coeficiente de correlación entre dos variables aleatorias escalares Y y Z con momento de segundo orden finito está definido como:

$$\rho = corr(Y, Z) = \frac{Cov(Y,Z)}{[(VarY)(VarZ)]^{\frac{1}{2}}}.$$

Esta es una medida de causa y efecto asociada con Y y Z. En general, $-1 \leq \rho \leq 1$, aunque en algunos casos, ρ es restringido a un intervalo más pequeño.

Una matriz de correlación $R = (\rho_{ij})$, $i, j = 1, ..., p$; es útil para estudiar todas las asociaciones entre las componentes de un vector de variables simultáneamente. La matriz de correlación es calculada en muchos modelos usados en análisis de datos multivariados ya que la matriz R provee frecuentemente un rápido entendimiento dentro de muchas relaciones insospechadas.

Los elementos de la diagonal, ρ_{ij}, de la matriz de correlación deberían ser todos uno, y los elementos fuera de la diagonal dados por:

$$\rho_{ij} = corr(x_i, x_j) = \frac{Cov(x_i,x_j)}{[(Var(x_i))(Var(x_j))]^{\frac{1}{2}}}, \text{ con } i \neq j$$

además, los ρ_{ij} deberían tambíen satisfacer siempre la inecuación $-1 \leq \rho_{ij} \leq 1$ para $i, j = 1, ..., p$.

2.11. La distribución Normal Mutivariada.

La distribución más importante y fundamental de análisis multivariado aplicado es la distribución normal multivariada. Su mayor papel se debe al hecho que estandariza sumas de vector de datos independientes que siguen distribuciones multivariables arbitrarias, en muestras grandes, para seguir la distribución normal multivariada.

Este resultado es una generalización del teorema del límite central univariado a las dimensiones más altas. El papel central jugado por la distribución también es atribuible y no de manera pequeña al hecho que pueden obtenerse a menudo los resultados para la distribución Normal y no para otras distribuciones que, por muchas razones, no son atractivas como la Normal.

2.11.1. Densidad general.

Sea $X : p \times 1$ un vector aleatorio con función de densidad $f(x)$. Se dirá que X sigue una distribución Normal (p-variante) multivariada no singular con vector de media $\theta : p \times 1$ y matriz de covarianza $\Sigma : p \times p$ si

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\theta)'\Sigma^{-1}(x-\theta)\},$$

para $\Sigma > 0$. Si $|\Sigma| = 0$, la distribución de x es llamada singular o Normal degenerada y la densidad no existe.

2.11.2. Media y Covarianza muestral.

Sean $x_1, ..., x_N$ vectores observados $p \times 1$ independientes con una distribución $N(\theta, \Sigma)$. Denotemos el vector de la media muestral por $\overline{x}$ y definamoslo como:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

y denotemos la matriz de covarianza muestral como V, la cual estará determinada por:

$$V = \sum_{i=1}^{N}[(x_i - \overline{x})(x_i - \overline{x})'].$$

2.11.3. Distribución Normal Bivariante.

Sea $X : 2 \times 1$ un vector aleatorio bivariante con $\pounds(X) = N(\theta, \Sigma)$ y $\Sigma > 0$. Sea $\theta = (\theta_i)$ y $\Sigma = (\sigma_{ij})$, $i, = 1, 2$. Para simplificar tomemos $\sigma_{11} = \sigma_1^2$, $\sigma_{12} = \sigma_2^2$ y $\sigma_{12} = \rho\sigma_1\sigma_2$, donde ρ es el coeficiente de correlación entre X_1 y X_2. Expandiendo la densidad general de la distribución normal multivariada para $p = 2$, se encuentra fácilmente que la densidad $f(x) \equiv f(x_1, x_2)$ es dada por la expresión

$$\frac{1}{2\pi\sigma_1\sigma_2}\sqrt{1-\rho^2}\exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x_1-\theta_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\theta_1}{\sigma_1}\right)\left(\frac{x_2-\theta_2}{\sigma_2}\right) + \left(\frac{x_2-\theta_2}{\sigma_2}\right)^2\right]\right\}.$$

Aquí,

$$\Sigma = \begin{pmatrix} \theta_1^2 & \rho\theta_1\theta_2 \\ \rho\theta_1\theta_2 & \theta_2^2 \end{pmatrix}, \qquad \Sigma^{-1} = \begin{pmatrix} \frac{1}{\theta_1^2(1-\rho^2)} & \frac{1}{\theta_1\theta_2(1-\rho^2)} \\ \frac{1}{\theta_1\theta_2(1-\rho^2)} & \frac{1}{\theta_2^2(1-\rho^2)} \end{pmatrix}.$$

La expresión entre llaves de la función de densidad de la normal multivariada controla la variación de $f(x)$. Esto es, si la expresión dentro de las llaves es constante, $f(x)$ es constante, y recíprocamente.

Definición 2.11. Supongase que Z_1 y Z_2 son variables aleatorias independientes cada una de las cuales tiene una distribución normal tipificada. Entonces la f.d.p. conjunta $g(z_1, z_2)$ de Z_1 y Z_2 para cualesquiera valores de z_1 y z_2 está dada por la ecuación

$$g(z_1, z_2) = \frac{1}{2\pi} \exp[-\frac{1}{2}(z_1^2 z_2^2)].$$

Para cualesquiera constantes μ_1, μ_2, σ_1, σ_2 y ρ tales que $-\infty < \mu_i < +\infty$ $(i = 1, 2)$, $\sigma_i > 0$ $(i = 1, 2)$ y $-1 < \rho < 1$, se definen ahora dos nuevas variables aleatorias X_1 y X_2 como sigue:

$$X_1 = \sigma_1 Z_1 + \mu_1,$$

$$X_2 = \sigma_2[\rho Z_1 + (1 - \rho^2)^{\frac{1}{2}} Z_2] + \mu_2.$$

Se deducirá ahora la f.d.p. $f(x_1, x_2)$ de X_1 y X_2.
La transformación de Z_1 y Z_2 a X_1 y X_2 es una transformación lineal y se verificará que el determinante Δ de la matriz de coeficientes de X_1 y X_2 tiene el valor $\Delta = \theta_1\theta_2(1 - \rho^2)^{\frac{1}{2}}$. Por lo tanto, el jacobiano J de la transformación inversa de X_1 y X_2 a Z_1 y Z_2 es

$$J = \frac{1}{\Delta} = \frac{1}{\theta_1\theta_2(1 - \rho^2)^{\frac{1}{2}}}.$$

Puesto que $J > 0$, el valor de $|J|$ es igual al valor de J. Si se resuelven el par de ecuaciones anteriores para Z_1 y Z_2 en función de X_1 y X_2, entonces la f.d.p. conjunta $f(x_1, x_2)$ se puede obtener reemplazando z_1 y z_2 de la primera ecuación por sus expresiones en función de x_1 y x_2 y multiplicando luego por $|J|$. Se puede demostrar que el resultado de $f(x_1, x_2)$ para $-\infty < x_1 < +\infty$ y $-\infty < x_2 < +\infty$, viene dado por:

$$\frac{1}{2\pi\sigma_1\sigma_2}\sqrt{1 - \rho^2} \exp\left\{\frac{-1}{2(1 - \rho^2)}\left[\left(\frac{x_1 - \theta_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1 - \theta_1}{\sigma_1}\right)\left(\frac{x_2 - \theta_2}{\sigma_2}\right) + \left(\frac{x_2 - \theta_2}{\sigma_2}\right)^2\right]\right\}. \tag{2.3}$$

Cuando la f.d.p. conjunta de dos variables aleatorias X_1 y X_2 es de la forma de la ecuación anterior se dice que X_1 y X_2 tienen una distribución normal bivariante. Las

medias y las varianzas de la distribución normal bivariante, especificada por (2.3), se pueden deducir fácilmente de las ecuaciones de X_1 y X_2. Puesto que Z_1 y Z_2 son independientes y cada una tiene media 0 y varianza 1, resulta que $E(X_1) = \mu_1$, $E(X_2) = \mu_2$, $Var(X_1) = \sigma_1^2$ y $Var(X_2) = \sigma_2^2$. Además, se puede demostrar de las ecuaciones de X_1 y X_2 que la $Cov(X_1, X_2) = \rho\sigma_1\sigma_2$. Por lo tanto, la correlación de X_1 y X_2 es simplemente ρ.

En resumen, si X_1 y X_2 tienen una distribución normal bivariante cuya f.d.p. esta dada por (2.3), entonces

$$E(X_i) = \mu_i,\ Var(X_i) = \sigma_i^2, \text{ para } i = 1, 2.$$

Además,

$$\rho(X_1, X_2) = \rho.$$

Ha resultado conveniente introducir la distribución normal bivariante como la distribución conjunta de ciertas combinaciones lineales de variables aleatorias independientes que tienen distribución normal tipificada. Debe subrayarse, sin embargo, que la distribución normal bivariante aparece directa y naturalmente en muchos problemas prácticos.

2.11.4. Independencia.

Sea $\pounds(X)$ y $X : 2 \times 1$. Entonces, si $\rho = 0$ en Σ y Σ^{-1}, anteriormente nombrados, X_1 y X_2 no solo no están correlacionados, ellos también son independientes. Esto se ve fácilmente sustituyendo $\rho = 0$ en la función de densidad y observando que $f(x_1, x_2)$ reduce al producto de una función de x_1 y una función de x_2. Claro la converción también es verdad, esto es, si X_1 y X_2 son independientes, ellos también son no correlacionados; en esta dirección, los resultados se sostienen para toda la distribución bivariante (considerando que en la otra dirección, carencia de correlación generalmente no implica la independencia, aunque lo hace para la distribución Normal).

2.11.5. Estandarización.

Si $\pounds(X) = N(\theta, \Sigma)$, la distribución puede ser estandarizada por la transformación $Y = \Sigma^{\frac{-1}{2}}(X - \theta)$; esto es $\pounds(Y) = N(0, I)$.

Puesto que $J(X \to Y) = |\Sigma|^{\frac{1}{2}}$, la densidad de Y es

$$g(y) = \frac{1}{(2\pi)^{\frac{p}{2}}} \exp[-\frac{1}{2}y'y].$$

2.12. Distribucion Wishart.

Sea $V : p \times p$ simétrica y definida positiva. La matriz aleatoria V se dice que sigue la distribución no singular Wishart p-dimensional con matriz de escala Σ y n grados de libertad, $p \leq n$, si la distribución conjunta de los distintos elementos de V es continua con función de densidad dada por:

$$p(V) = \begin{cases} \frac{c|V|^{\frac{(n-p-1)}{2}}}{|\Sigma|^{\frac{n}{2}}} \exp\{-\frac{1}{2}tr(\Sigma^{-1}V)\}, & V > 0, \Sigma > 0 \\ 0, & \text{en otros casos} \end{cases} \tag{2.4}$$

donde c es un número constante definido como:

$$c = [2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma(\frac{n+1-j}{2})]^{-1}.$$

Si $n < p$, la distribución es singular y no es una densidad. Así, si $V \equiv (v_{ij})$ y $\Sigma^{-1} = (\sigma^{ij})$, para $V > 0$, entonces:

$$p(V) \ \alpha \ \frac{|V|^{\frac{(n-p-1)}{2}}}{|\Sigma|^{\frac{n}{2}}} \exp\{-\frac{1}{2}(\sum_{i=1}^{p} \sum_{j=1}^{p} v_{ij}\sigma^{ij})\}.$$

Esta relación será expresada como:

$$\pounds(V) = W(\Sigma, p, n).$$

La distribución Wishart se utiliza para representar la distribución muestral de las matrices de covarianza en muestras de variables normales multivariantes. En el caso escalar, la distribución que representa esta incertidumbre es la ji-cuadrado de Pearson, χ^2, y la distribución de Wishart estándar puede considerarse como una generalización multivariante de esta distribución.

Consideremos un conjunto de m vectores aleatorios, $(X_1, ..., X_m)$, de dimensión p con la misma distribución $N_p(0, I)$. La estimación de su matriz de varianzas y covarianzas se obtendrá de $\sum_{i=1}^{m} X_i X_i'/m$, y el numerador de esta expresión

$$W = \sum_{i=1}^{m} X_i X_i' \tag{2.5}$$

que es una matriz cuadrada $p \times p$ simétrica y definida positiva, decimos que sigue una distribución Wishart con m grados de libertad. Esta información debe interpretarse en el sentido de que la distribución conjunta de los $\frac{1}{2}p(p+1)$ elementos distintos de W es

$$f(w_{11}, ..., w_{pp}) = c|V|^{\frac{(m-p-1)}{2}} \exp\{-\frac{1}{2}tr(W)\}$$

donde c es una constante para que la función integre uno.
Escribimos $W \sim W_p(m)$, donde p indica que se trata de la distribución de los elementos de una matriz cuadrada y simétrica de orden p, y m son los grados de libertad. Observemos que esta distribución depende únicamente de las dos medidas escalares del tamaño de la matriz: la traza y el determinante. Por lo tanto, todas las combinaciones de elementos de la matriz que conduzca a los mismos valores de estas medidas de tamaño tienen la misma probabilidad.

Consideremos, seguidamente, m vectores aleatorios $(X_1, ..., X_m)$ de distribución $N_p(0, \Sigma)$. La distribución de los elementos de la matriz

$$W = \sum_{i=1}^{m} x_i x_i' \tag{2.6}$$

es la distribución Wishart con m grados de libertad y matriz de parámetros Σ, dada por

$$f(w_{11}, ..., w_{pp}) = c|\Sigma|^{-\frac{m}{2}}|W|^{\frac{(m-p-1)}{2}} \exp\{-\frac{1}{2}tr\Sigma^{-1}W\}.$$

En general, si una matriz cuadrada y simétrica sigue la distribución Wishart, para una matriz simétrica Σ $(p \times p)$ no singular definida positiva de componentes constantes, diremos que dicha matriz sigue la distribución Wishart con m grados de libertad y matriz de parámetros Σ, la cual denotaremos por $W \sim W_p(m, \Sigma)$.

2.12.1. Propiedades de la distribución Wishart.

La distribución Wishart tiene las siguientes propiedades:

1. La esperanza de la distribución es:

$$E[W] = m\Sigma$$

lo que implica que W/m tiene esperanza Σ.

2. La suma de dos distribuciones χ^2 independientes es otra distribución χ^2 con grados de libertad igual a la suma de los grados de libertad de ambas. Análogamente, si $W_1 \sim W_p(m_1, \Sigma)$ y $W_2 \sim W_p(m_2, \Sigma)$ son independientes, entonces $W_1 + W_2 \sim W_p(m_1 + m_2, \Sigma)$.

3. Si A es una matriz $h \times p$ de constantes, y $W \sim W_p(m, \Sigma)$, la distribución de $AWA' \sim W_h(m, A^{-1}\Sigma A'^{-1})$. En efecto, como $W = \sum_{i=1}^{m} X_i X_i'$ la variable $AW'A$ será

$$A \sum_{i=1}^{m} x_i x_i' A' = \sum_{i=1}^{m} y_i y_i'$$

donde ahora y_i es $N(0, A\Sigma A')$, y aplicando la definición de la distribución Whishart se obtiene el resultado.

4. Si S es la matriz de varianzas y covarianzas muestral

$$S = \frac{1}{n} X'PX$$

donde $P = I - \frac{1}{n}11'$ es idempotente, entonces

$$nS \sim W_p(n-1, \Sigma).$$

Esta expresión indica que si tenemos el estimador

$$\widehat{S} = \frac{1}{(n-1)} X'PX = \frac{n}{(n-1)} S$$

su esperanza será σ, y $\widehat{S}$ será un estimador centrado para la matriz de varianzas. Podemos escribir que $(n-1)\widehat{s}^2$, donde s^2 es el estimador centrado de la varianza, sigue la distribución $\sigma^2 \chi^2_{n-1}$.

2.13. Criterio de Suficiencia.

En muchos problemas en los que se debe estimar un parámetro θ, es posible determinar un E.M.V. que sea apropiado. En algunos problemas, sin embargo, es posible que ninguno de estos estimadores sea apropiado. Podría no existir ningún E.M.V. o podria existir más de uno. Aún cuando el E.M.V. sea único, podría no ser el apropiado.

Supongamos que las variables aleatoria $X_1, ..., X_n$ constituyen una muestra aleatoria de una distribución discreta o continua con función de densidad $f(x|\theta)$. Supongamos además que el valor desconocido de θ debe pertenecer a un espacio paramétrico Ω.

Puesto que las variables aletorias $X_1, ..., X_n$ constituyen una muestra aletoria, se sabe que la f.d.p. conjunta $f_n(x|\theta)$ tiene la siguiente forma para una valor particular de $\theta \, \epsilon \, \Omega$

$$f_n(x|\theta) = f(x_1|\theta), ..., f(x_n|\theta).$$

El problema de estimar un valor de θ, se puede considerar como el problema de seleccionar por inferencia la distribución particular de esta familia que genera las observaciones $X_1, ..., X_n$.

Definición 2.12. Sea $X_1, ..., X_n$ una muestra aleatoria de una distribución discreta o continua con función de densidad $f(x|\theta)$. Cualquier función real $T = T(X_1, ..., X_n)$ de las observaciones de la muestra aleatoria se llama estadístico.

Definición 2.13. Sea X una cantidad aleatoria con función de probabilidad (densidad) $p(x|\theta)$. Entonces, el estadístico $T = T(X)$ es suficiente para el parámetro θ si

$$p(x|t, \theta) = p(x|t).$$

Teorema 2.1 (Criterio de Suficiencia)**.** *Si $T = T(X)$ es un estadístico suficiente para θ, entonces*

$$p(\theta|x) = p(\theta|t),$$

para toda densidad $p(\theta)$.

Demostración. Tenemos:

$$p(x|\theta) = \begin{cases} p(x, t|\theta) & \text{si } t = T(X) \\ 0 & \text{si } t \neq T(X). \end{cases} \tag{2.7}$$

Así,

$$\begin{aligned} p(x|\theta) &= p(x|t, \theta)p(t|\theta) \\ &= p(x|t)p(t|\theta), \qquad \text{(por la definición de suficiencia).} \end{aligned}$$

Pero, por el teorema de Bayes,

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &= p(x|t)p(t|\theta)p(\theta) \\ &\propto p(t|\theta)p(\theta), \qquad (p(x|t) \text{ no depende de } \theta) \\ &\propto p(\theta|t). \end{aligned}$$

Entonces $p(\theta|x) = kp(\theta|t)$, para algún $k > 0$.
Adicionalmente,

$$I = \int_\Theta p(\theta|x)d\theta = k\int_\Theta p(\theta|t)d\theta = k$$

y así, $p(\theta|x) = p(\theta|t)$.

□

Definición 2.14. El estadístico $T(X)$ es suficiente para θ si hay una función f tal que

$$p(\theta|x) \propto f(\theta, t).$$

Teorema 2.2 (Criterio de factorización de Neyman Fisher.)**.** *El estadístico T es suficiente para θ si y solo si*

$$p(x|\theta) = f(t, \theta)g(x)$$

donde f y g son funciones no negativas.

Demostración. ($\Rightarrow$)
Hemos visto que $p(x|\theta) = p(x|t)p(t|\theta)$. Entonces esto es suficiente para definir

$$g(x) = p(x|t) = p(x|T(X)) \text{ y } f(t, \theta) = f(t|\theta).$$

Así se completa la prueba.

($\Leftarrow$)
Tenemos que $p(x|\theta) = f(t, \theta)g(x)$.
Definiendo $A_t = \{x : T(x) = t\}$, la función de probabilidad de $T|\theta$ es

$$\begin{aligned} p(t|\theta) &= \int_{A_t} p(x|\theta) \\ &= f(t, \theta)\int_{A_t} g(x)dx \\ &= f(t, \theta)G(x), \qquad \text{para alguna función } G \end{aligned}$$

y así, $f(t,\theta) = \frac{p(t|\theta)}{G(x)}$.

Por otro lado, por la hipótesis de el teorema, $f(t,\theta) = \frac{p(x|\theta)}{g(x)}$.
De la equivalencia entre las ecuaciones tenemos

$$\frac{p(t|\theta)}{p(x|\theta)} = \frac{G(x)}{g(x)}.$$

Así, $p(x|t,\theta) = \frac{p(t|\theta)}{p(x|\theta)}$, entonces

$$p(x|t,\theta) = \frac{G(x)}{g(x)} = p(x|t),$$

donde este no depende de θ. Por lo tanto, T es suficiente para θ.

□

Teorema 2.3 (Caso multivariado.)**.** *Sea $p(x_1, x_2, \ldots, x_N|\phi)$ la distribución conjunta de N observaciones y $x_j : p\times 1$, con $j : 1, \ldots, N$, y una matriz de parámetros $\phi : q\times r$, entonces, $T(x_1, x_2, \ldots, x_N) \equiv T$ es suficiente para ϕ sí y solo sí existen funciones no negativas f y g tales que:*

$$p(x_1, x_2, \ldots, x_N|\phi) = f(T;\phi).g(x_1, x_2, \ldots, x_N). \tag{2.8}$$

Capítulo 3

Método de máxima verosimilitud para la distribución normal multivariada.

Al considerar el problema de estimación recordemos que un Estadístico es una función que depende de una muestra aleatoria $X_1, X_2, \ldots, X_N$ tomada de una variable aleatoria X, correspondiente a una población, y que no contiene parámetros desconocidos.

En el caso univariado, un Estadístico Suficiente para determinado parámetro es buscado, en la práctica, haciendo uso del criterio de factorización de *Neyman Fisher*. Tal criterio es también aplicable a distribuciones multivariadas y es una generalización inmediata del caso univariado.

Supongamos que $x_1, x_2, \ldots, x_N$ es una muestra aleatoria de vectores $p \times 1$ independientes e idénticamente distribuidos como $N(\theta, \Sigma)$. Su densidad conjunta (Verosimilitud) esta dada por:

$$p(x_1, x_2, \ldots, x_N|\theta, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\{-\tfrac{1}{2}\textstyle\sum_{j=1}^{N}(x_j - \theta)'\Sigma^{-1}(x_j - \theta)\}.$$

El siguiente teorema determina que estimadores son suficientes para los parámetros θ y Σ de la distribución normal multivariada, su prueba estará basada en el Criterio de Suficiencia de Neyman Fisher para distribuciones multivariables, el cual enuncia que:

"Para $p(x_1, x_2, \ldots, x_N|\phi)$, la distribución conjunta de N observaciones y matriz de parámetros $\phi : q \times r$, entonces, $T(x_1, x_2, \ldots, x_N) \equiv T$ es suficiente para ϕ sí y solo

sí existen funciones no negativas f y g tales que:

$$p(x_1, x_2, \ldots, x_N|\phi) = f(T;\phi).g(x_1, x_2, \ldots, x_N)". \tag{3.1}$$

Teorema 3.1. *Si $x_1, x_2, \ldots, x_N$ son vectores observados $p \times 1$ independientes con distribución $N(\theta, \Sigma)$, entonces $(\overline{x}, V)$ es un estimador suficiente para (θ, Σ).*

Demostración. Supongamos que $x_1, x_2, \ldots, x_N$ es una muestra aleatoria de vectores $p \times 1$ mutuamente independientes e idénticamente distribuidos como una $N(\theta, \Sigma)$. Así, su probabilidad conjunta esta dada por:

$$p(x_1, x_2, \ldots, x_N|\theta, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\{-\frac{1}{2}\sum_{j=1}^{N}(x_j - \theta)'\Sigma^{-1}(x_j - \theta)\}. \tag{3.2}$$

Ahora probemos que (3.2) se puede escribir como (3.1) con $T = (\overline{x}, V)$ y $\phi = (\theta, \Sigma)$.
Notemos primero que en la función de densidad

$$p(x_1, x_2, \ldots, x_N|\theta, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\{-\frac{1}{2}\sum_{j=1}^{N}(x_j - \theta)'\Sigma^{-1}(x_j - \theta)\},$$

podemos reescribir la expresión $\sum_{j=1}^{N}(x_j - \theta)'\Sigma^{-1}(x_j - \theta)$ como:

$$\begin{aligned}
\sum_{j=1}^{N}(x_j - \theta)'\Sigma^{-1}(x_j - \theta) &= (x_1 - \theta)'\Sigma^{-1}(x_1 - \theta) + \cdots + (x_N - \theta)'\Sigma^{-1}(x_N - \theta) \\
&= tr\left[\Sigma^{-1}(x_1 - \theta)(x_1 - \theta)'\right] + \cdots + tr[\Sigma^{-1}(x_N - \theta)(x_N - \theta)'\big] \\
&= tr\left[\Sigma^{-1}(x_1 - \theta)(x_1 - \theta)' + \cdots + \Sigma^{-1}(x_N - \theta)(x_N - \theta)'\right] \\
&= tr\left[\Sigma^{-1}\sum_{j=1}^{N}(x_j - \theta)(x_j - \theta)'\right],
\end{aligned}$$

esto por propiedades de la traza de una matriz, y usando la simetría de Σ.
Así,

$$p(x_1, x_2, \ldots, x_N|\theta, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\{-\frac{1}{2}tr(\Sigma^{-1}A)\},$$

donde $A = \sum_{j=1}^{N}(x_j - \theta)(x_j - \theta)'$.

Más aún,

$$\begin{aligned}
A &= \sum_{j=1}^{N}(x_j - \theta)(x_j - \theta)' \\
&= \sum_{j=1}^{N}(x_j - \bar{x} + \bar{x} - \theta)(x_j - \bar{x} + \bar{x} - \theta)' \\
&= \sum_{j=1}^{N}[(x_j - \bar{x}) + (\bar{x} - \theta)][(x_j - \bar{x}) + (\bar{x} - \theta)]' \\
&= \sum_{j=1}^{N}[(x_j - \bar{x})(x_j - \bar{x})' + (x_j - \bar{x})(\bar{x} - \theta)' + (\bar{x} - \theta)(x_j - \bar{x})' + (\bar{x} - \theta)(\bar{x} - \theta)'] \\
&= \sum_{j=1}^{N}[(x_j - \bar{x})(x_j - \bar{x})'] + \sum_{j=1}^{N}[(x_j - \bar{x})(\bar{x} - \theta)' + (\bar{x} - \theta)(x_j - \bar{x})'] + \sum_{j=1}^{N}[(\bar{x} - \theta)(\bar{x} - \theta)'],
\end{aligned}$$

pero,

$\sum_{j=1}^{N}[(x_j - \bar{x})(\bar{x} - \theta)' + (\bar{x} - \theta)(x_j - \bar{x})'] = \sum_{j=1}^{N}[x_j\bar{x}' - x_j\theta' - \bar{x}\bar{x}' + \bar{x}\theta' + \bar{x}x_j' - \bar{x}\bar{x}' - \theta x_j' + \bar{x}'\theta]$

$= \sum_{j=1}^{N} x_j\bar{x}' - \sum_{j=1}^{N} x_j\theta' - \sum_{j=1}^{N} \bar{x}\bar{x}' + \sum_{j=1}^{N} \bar{x}\theta' + \sum_{j=1}^{N} \bar{x}x_j' - \sum_{j=1}^{N} \bar{x}\bar{x}' - \sum_{j=1}^{N} \theta x_j' + \sum_{j=1}^{N} \bar{x}'\theta$

$= \bar{x}' \sum_{j=1}^{N} x_j - \theta' \sum_{j=1}^{N} x_j - N\bar{x}\bar{x}' + N\bar{x}\theta' + \bar{x} \sum_{j=1}^{N} x_j' - N\bar{x}\bar{x}' - \theta \sum_{j=1}^{N} x_j' + N\bar{x}'\theta$

$= \bar{x}'N\bar{x} - \theta'N\bar{x} - N\bar{x}\bar{x}' + N\bar{x}\theta' + \bar{x}N\bar{x}' - N\bar{x}\bar{x}' - \theta N\bar{x}' + N\bar{x}'\theta$

$= 0$

entonces,

$$\begin{aligned} A &= \sum_{j=1}^{N}[(x_j-\bar{x})(x_j-\bar{x})'] + \sum_{j=1}^{N}[(\bar{x}-\theta)(\bar{x}-\theta)'] \\ &= \sum_{j=1}^{N}[(x_j-\bar{x})(x_j-\bar{x})'] + N[(\bar{x}-\theta)(\bar{x}-\theta)'] \\ &= V + N[(\bar{x}-\theta)(\bar{x}-\theta)'], \end{aligned}$$

donde $V = \sum_{j=1}^{N}[(x_j-\bar{x})(x_j-\bar{x})']$. Por lo tanto,

$$\begin{aligned} p(x_1,x_2,\ldots,x_N|\theta,\Sigma) &= \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}}\exp\{-\frac{1}{2}tr(\Sigma^{-1}A)\} \\ &= \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}}\exp\{-\frac{1}{2}tr[\Sigma^{-1}(V+N(\bar{x}-\theta)(\bar{x}-\theta)')]\}. \end{aligned}$$

Haciendo

$$g(x_1,x_2,\ldots,x_N) = 1 \text{ y}$$
$$f(T,\phi) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}}\exp\{-\frac{1}{2}tr[\Sigma^{-1}(V+N(\bar{x}-\theta)(\bar{x}-\theta)')]\},$$

tenemos que $p(x_1,x_2,\ldots,x_N|\theta,\Sigma)$ se puede escribir como:

$$f(T;\phi)g(x_1,x_2,\ldots,x_N).$$

Por lo tanto, de acuerdo al Criterio de Neyman Fisher, $\overline{x}$ y V son estimadores suficientes para θ y Σ, respectivamente.

□

A continuación se enunciará y demostrará un teorema que determina la forma explícita de los estimadores sufientes de los parámetros θ y Σ nombrados en el teorema anterior.

Teorema 3.2. *Sean $x_1, x_2, \ldots, x_N$ vectores $p\times 1$ independientes con una distribución $N(\theta,\Sigma)$, los estimadores de θ y Σ por el método de máxima verosimitud están dados por:*

$$\widehat{\theta} = \overline{x} \qquad y \qquad \widehat{\Sigma} = \frac{1}{N}\sum_{j=1}^{N}(x_j - \overline{x})(x_j - \overline{x})'$$

Demostración. Definamos $\Lambda = \Sigma^{-1}$ y $L(\theta, \Lambda) = \log p(x_1, x_2, \ldots, x_N|\theta, \Lambda)$. Recordemos que:

$$p(x_1, x_2, \ldots, x_N|\theta, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\{-\tfrac{1}{2}tr[\Sigma^{-1}(V + N(\bar{x} - \theta)(\bar{x} - \theta)')]\},$$

entonces debemos hallar los valores de θ y Σ que maximizan a $p(x_1, x_2, \ldots, x_N|\theta, \Sigma)$. Para facilitar los cálculos maximizamos $L(\theta, \Lambda)$, pues los valores que las maximizan coinciden por ser la función logaritmo una función creciente.

Así,

$$\begin{aligned}
L(\theta, \Lambda) &= \log p(x_1, x_2, \ldots, x_N|\theta, \Sigma) \\
&= \log\left\{\frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\{-\frac{1}{2}tr\Sigma^{-1}(V + N(\bar{x} - \theta)(\bar{x} - \theta)')\}\right\} \\
&= \log(1) - \log[(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}] - \frac{1}{2}tr\Sigma^{-1}[V + N(\bar{x} - \theta)(\bar{x} - \theta)'] \\
&= -\log[(2\pi)^{\frac{Np}{2}}] + \frac{N}{2}\log|\Sigma|^{-1} - \frac{1}{2}tr\Sigma^{-1}[V + N(\bar{x} - \theta)(\bar{x} - \theta)'] \\
&= -\log[(2\pi)^{\frac{Np}{2}}] + \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}tr\Sigma^{-1}V + N\frac{1}{2}tr\Sigma^{-1}(\bar{x} - \theta)(\bar{x} - \theta)' \\
&= -\log[(2\pi)^{\frac{Np}{2}}] + \frac{N}{2}\log|\Lambda| - \frac{1}{2}tr\Lambda V + N\frac{1}{2}tr\Lambda(\bar{x} - \theta)(\bar{x} - \theta)'.
\end{aligned}$$

Derivando este último resultado respecto a θ, y utilizando las propiedades de la traza y de la derivada tenemos:

$$\begin{aligned}
\frac{\partial}{\partial\theta}L(\theta, \Lambda) &= -\frac{N}{2}\frac{\partial}{\partial\theta}tr\Lambda(\bar{x} - \theta)(\bar{x} - \theta)' \\
&= -\frac{N}{2}\frac{\partial}{\partial\theta}(\bar{x} - \theta)'\Lambda(\bar{x} - \theta) \\
&= -\frac{N}{2}[-2\Lambda(\bar{x} - \theta)] \\
&= -\frac{N}{2}[2\Lambda(\theta - \bar{x})].
\end{aligned}$$

Luego, igualando la derivada a cero y resolviendo tenemos:

$$\begin{aligned}
\frac{\partial}{\partial\theta}L(\theta, \Lambda) = 0 &\Rightarrow -\frac{N}{2}[2\Lambda(\theta - \bar{x})] = 0 \\
&\Rightarrow \Lambda(\theta - \bar{x}) = 0 \\
&\Rightarrow \theta - \bar{x} = 0.
\end{aligned}$$

Así,

$$\widehat{\theta} = \bar{x}. \tag{3.3}$$

Por otro lado,

$$\begin{aligned} \frac{\partial^2}{\partial\theta\partial\theta'}L(\theta,\Lambda) &= -\frac{N}{2}\frac{\partial}{\partial\theta'}[2\Lambda(\theta-\bar{x})] \\ &= -N\Lambda, \end{aligned}$$

la cual es claramente negativa, pues $N > 0$ y $\Sigma > 0$, así $\overline{x}$ corresponde a un máximo de $L(\theta, \Lambda)$.

Ahora, derivando $L(\theta, \Lambda)$ con respecto a Λ obtenemos:

$$\begin{aligned} \frac{\partial}{\partial\Lambda}L(\theta,\Lambda) &= \frac{N}{2}\frac{\partial}{\partial\Lambda}\log|\Lambda| - \frac{1}{2}\frac{\partial}{\partial\Lambda}tr\Lambda V - \frac{N}{2}\frac{\partial}{\partial\Lambda}tr\Lambda(\bar{x}-\theta)(\bar{x}-\theta)' \\ &= \frac{N}{2}\frac{\partial}{\partial\Lambda}\log|\Lambda| - \frac{1}{2}\frac{\partial}{\partial\Lambda}tr\Lambda[V + N(\bar{x}-\theta)(\bar{x}-\theta)'] \\ &= \frac{N}{2}\frac{\partial}{\partial\Lambda}\log|\Lambda| - \frac{1}{2}\frac{\partial}{\partial\Lambda}tr\Lambda A \\ &= \frac{N}{2}(2\Lambda^{-1} - diag\Lambda^{-1}) - \frac{1}{2}(2A - diag(A)), \end{aligned}$$

luego, haciendo la derivada igual a cero y resolviendo obtenemos:

$$\begin{aligned} \frac{N}{2}(2\Lambda^{-1} - diag\Lambda^{-1}) - \frac{1}{2}(2A - diag(A)) = 0 &\Rightarrow N(2\Lambda^{-1} - diag\Lambda^{-1}) - (2A - diag(A)) = 0 \\ &\Rightarrow 2N\Lambda^{-1} - 2A = 0 \\ &\Rightarrow \Lambda^{-1} = \frac{A}{N}. \end{aligned}$$

Por lo tanto, el estimador viene dado como:

$$\widehat{\Sigma} = \frac{\widehat{A}}{N} = \frac{1}{N}\sum_{j=1}^{N}(x_j - \widehat{\theta})(x_j - \widehat{\theta})' = \frac{1}{N}\sum_{j=1}^{N}(x_j - \overline{x})(x_j - \overline{x})', \tag{3.4}$$

pues $\widehat{\theta} = \bar{x}$.

La función $\log|\Lambda|$ es concava y puesto que $tr(\Lambda V)$ es lineal en Λ, $L(\theta, \Lambda)$ es concava en Λ, así $\widehat{\Sigma}$ debe corresponder a un máximo.

De esto, se aprecia claramente que ésta matriz es definida positiva.

Podemos concluir, de los resultados obtenidos en (3.3) y (3.4), que los estimadores suficientes para θ y Σ tienen (respectivamente) la siguiente forma:

$$\widehat{\theta} = \overline{x} \qquad \text{y} \qquad \widehat{\Sigma} = \tfrac{1}{N}\sum_{j=1}^{N}(x_j - \overline{x})(x_j - \overline{x})'$$

□

Para hacer estimación de los parámetros de una distribución normal multivariada por intervalos de confianza o para realizar test de hipótesis, es de interés conocer la distribución de los estadísticos suficientes. A continuación enunciamos los siguientes teoremas.

Teorema 3.3. *Si $x_1, x_2, \ldots, x_N$ son vectores observados $p \times 1$ independientes de una $N(\theta, \Sigma)$ y $\overline{x}$ es el vector de media muestral, entonces:*

$$\pounds(\overline{x}) = N(\theta, \tfrac{\Sigma}{N}).$$

Demostración. Sean $x_1, x_2, \ldots, x_N$ observaciones de vectores p-variantes independientes, distribuidos como $N(\theta, \Sigma)$.

Consideremos el vector de la media muestral $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

Luego, como $\pounds(x_i) = N(\theta, \Sigma)$ y los $x_i's$ son independientes, se tiene:

$$E(\overline{x}) = E(\frac{1}{N}\sum_{i=1}^{N} x_i) = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N} \theta = \frac{1}{N} N\theta = \theta. \tag{3.5}$$

Así de (3.5) tenemos,

$$E(\overline{x}) = \theta.$$

También tenemos que,

$$Var(\overline{x}) = Var(\frac{1}{N}\sum_{i=1}^{N} x_i) = \frac{1}{N^2}\sum_{i=1}^{N} Var(x_i) = \frac{1}{N^2}\sum_{i=1}^{N} \Sigma = \frac{1}{N^2} N\Sigma = \frac{1}{N}\Sigma. \tag{3.6}$$

Por lo tanto, de (3.6) obtenemos

$$Var(\overline{x}) = \frac{\Sigma}{N}.$$

De lo anterior concluimos que,

$$\pounds(\overline{x}) = N(\theta, \tfrac{\Sigma}{N}).$$

□

Teorema 3.4. *Sean* $X_\alpha : p \times 1$, *y* $X_1, X_2, \ldots, X_N$ *mutuamente independientes, con* $\pounds(X_\alpha) = N(\theta, \Sigma)$, $\Sigma > 0$, *para* $\alpha = 1, ..., N$; *denotemos el vector de media muestral por* $\overline{X} = \frac{1}{N}\sum_{\alpha=1}^{N} X_\alpha$. *Definamos* $V = \sum_{\alpha=1}^{N}(X_\alpha - \overline{X})(X_\alpha - \overline{X})' = \sum_{\alpha=1}^{N} X_\alpha X_\alpha' - NXX'$. *Entonces,* $p + 1 \leq N$, $V > 0$, $\pounds(V) = W(\Sigma, p, n)$, *con* $n = N - 1$.

Demostración. Recordemos de la definición de la distribución Wishart, vista en los conceptos fundamentales, que para un conjunto de m vectores aleatorios $(X_1, ..., X_m)$, de dimensión p con la misma distribución $N_p(0, I)$ se tiene que, si $W = \sum_{i=1}^{m} X_i X_i'$ es una matriz cuadrada $p \times p$ simétrica y definida positiva, entonces decimos que ésta sigue una distribución Wishart con m grados de libertad.

Así, para probar que $\pounds(V) = W(\Sigma, p, n)$, con $n = N - 1$, es suficiente reducir el estadístico V a uno que tenga la forma YY', donde $Y = (Y_1, ..., Y_n)$, $Y : p \times n$, $p \leq n$, y $\pounds(Y_\alpha) = N(0, \Sigma)$, $\Sigma > 0$, para $\alpha = 1, ..., N$.

Definamos entonces un conjunto de vectores $Y_1, Y_2, \ldots, Y_n$, $n = N - 1$ con las siguientes propiedades:

1. $Y_1, Y_2, \ldots, Y_n$ son mutuamente independientes.

2. $\pounds(Y_\alpha) = N(0, \Sigma)$, $\alpha = 1, ..., n$,

3. $V = \sum_{\alpha=1}^{n} Y_\alpha Y_\alpha'$.

Transformemos los vectores X_α usando cualquier matriz ortogonal cuyos elementos de la última fila sean todos iguales, una de tales matrices es la que tiene a $N^{-\frac{1}{2}}$ como elementos de su última fila.

Denotemos por $\Gamma = (\gamma_{ij})$, $i, j = 1, \ldots, N$, cualquier matriz ortogonal $N \times N$ la cual satisface que $\gamma_{Nj} = N^{-\frac{1}{2}}$, $j = 1, \ldots, N$ (los elementos de la última fila son todos iguales).
Definamos:

$$Y_i = \sum_{j=1}^{N} \gamma_{ij} X_j, \qquad i = 1, \ldots, N. \tag{3.7}$$

Puesto que $\pounds(X_j) = N(\theta, \Sigma)$, para todo j, entonces $\pounds(\gamma_{ij} X_j) = N(\gamma_{ij}\theta, \gamma_{ij}^2 \Sigma)$, ya que

$$E(\gamma_{ij} X_j) = \gamma_{ij} E(X_j) = \gamma_{ij}\theta.$$

y

$$Var(\gamma_{ij} X_j) = \gamma_{ij}^2 Var(X_j) = \gamma_{ij}^2 \Sigma.$$

Sumando sobre los j, es decir, $\sum_{j=1}^{N} \gamma_{ij} X_j$ se obtiene

$$\pounds(Y_i) = N(\phi_i, a\Sigma),$$

donde $\phi_i = (\sum_{j=1}^{N} \gamma_{ij})\theta$ y $a = \sum_{j=1}^{N} \gamma_{ij}^2$. Por la ortogonalidad de Γ, $a = 1$. Más aún, puesto que $\gamma_{Nj} = N^{-\frac{1}{2}}$, es decir, $\gamma_{Nj} N^{\frac{1}{2}} = 1$, con $j = 1, \ldots, N$, así tenemos que:

$$\phi_i = (\sum_{j=1}^{N} \gamma_{ij})(\gamma_{Nj} N^{\frac{1}{2}})\theta.$$

Nuevamente por ortogonalidad de Γ, para $i \neq N$, $\sum_{j=1}^{N} \gamma_{ij}\gamma_{Nj} = 0$. Así, $\phi_i = 0$, $i = 1, \ldots, n,$ lo cual establece la propiedad 2.

Por otro lado, para $i \neq k$, $i \neq N$ y $k \neq N$,

$$\begin{aligned} E(Y_i Y_k') &= E\left[(\sum_{j=1}^{N} \gamma_{ij} X_j)(\sum_{\alpha=1}^{N} \gamma_{k\alpha} X_\alpha')\right] \\ &= \sum_{j=1}^{N}\sum_{\alpha=1}^{N} \gamma_{ij}\gamma_{k\alpha} E(X_j X_\alpha'). \end{aligned}$$

Para esto último analicemos los siguientes casos:
Para $j = \alpha$, se tiene:

$$\sum_{j=1}^{N}\sum_{\alpha=1}^{N} \gamma_{ij}\gamma_{k\alpha} = \sum_{j=1}^{N} \gamma_{ij}\gamma_{kj} = 0.$$

Por lo tanto,

$$E(Y_i Y_k') = 0$$

Luego, para $j \neq \alpha$

$$\begin{aligned}
E(Y_i Y_k') &= \sum_{j=1}^{N}\sum_{\alpha=1}^{N} \gamma_{ij}\gamma_{k\alpha}\theta\theta' \\
&= \left[\sum_{j=1}^{N}\sum_{\alpha=1}^{N} \gamma_{ij}\gamma_{k\alpha} - \sum_{j=1}^{N}\gamma_{ij}\gamma_{kj}\right]\theta\theta' \\
&= \left[\sum_{j=1}^{N}\gamma_{ij}(\gamma_{Nj}N^{\frac{1}{2}})\sum_{\alpha=1}^{N}\gamma_{k\alpha}(\gamma_{N\alpha}N^{\frac{1}{2}}) - \sum_{j=1}^{N}\gamma_{ij}\gamma_{kj}\right]\theta\theta' \\
&= \left[(N^{\frac{1}{2}}\sum_{j=1}^{N}\gamma_{ij}\gamma_{Nj})(N^{\frac{1}{2}}\sum_{\alpha=1}^{N}\gamma_{k\alpha}\gamma_{N\alpha}) - \sum_{j=1}^{N}\gamma_{ij}\gamma_{kj}\right]\theta\theta' \\
&= \left[N(\sum_{j=1}^{N}\gamma_{ij}\gamma_{Nj})(\sum_{\alpha=1}^{N}\gamma_{k\alpha}\gamma_{N\alpha}) - (\sum_{j=1}^{N}\gamma_{ij}\gamma_{kj})\right]\theta\theta'
\end{aligned}$$

ya que cada término entre paréntesis desaparece por ortogonalidad en Γ. Entonces

$$E(Y_i Y_k') = 0. \tag{3.8}$$

Esto establece la propiedad 1.

Finalmente de (3.7) tenemos que:

$$\begin{aligned}
\sum_{i=1}^{N} Y_i Y_i' &= \sum_{i=1}^{N}(\sum_{j=1}^{N}\gamma_{ij}X_j)(\sum_{\alpha=1}^{N}\gamma_{i\alpha}X_\alpha') \\
&= \sum_{i=1}^{N}\sum_{\alpha=1}^{N}(\sum_{j=1}^{N}\gamma_{ij}\gamma_{i\alpha})X_j X_\alpha' \\
&= \sum_{j=1}^{N}(\sum_{\substack{i=1 \\ \alpha=j}}^{N}\gamma_{ij}^2)X_j X_j' + \sum_{j=1}^{N}\sum_{\substack{\alpha=1 \\ \alpha\neq j}}^{N}(\sum_{i=1}^{N}\gamma_{ij}\gamma_{i\alpha})X_j X_\alpha',
\end{aligned}$$

puesto que los términos entre paréntesis deberían ser unos y ceros, respectivamente, se sigue que

$$\sum_{i=1}^{N} Y_i Y_i' = \sum_{i=1}^{N} X_i X_i'.$$

Notemos que de (3.7), $Y_N = \sqrt{N}\,\overline{X}$. Así,

$$\begin{aligned} V &= \sum_{\alpha=1}^{N} X_\alpha X'_\alpha - N\overline{X}\,\overline{X}' \\ &= \sum_{\alpha=1}^{N} Y_\alpha Y'_\alpha - Y_N Y'_N \\ &= \sum_{\alpha=1}^{N-1} Y_\alpha Y'_\alpha \end{aligned}$$

Lo que establece la propiedad 3.

Por lo tanto, podemos concluir que para $Y_1, Y_2, \ldots, Y_n$ mutuamente independientes con $\mathcal{L}(Y_\alpha) = N(0, \Sigma)$, $\alpha = 1, ..., n$, y para $V = \sum_{\alpha=1}^{n} Y_\alpha Y'_\alpha$, se tiene que $\mathcal{L}(V) = W(\Sigma, p, n)$, con $n = N - 1$ grados de libertad.

□

Veamos, mediante el siguiente teorema, como influye la independencia de la muestra observada sobre los estimadores suficientes del teorema 1, cuya forma fué determinada y expresada en el teorema 2.

Teorema 3.5. *Si $x_1, x_2, \ldots, x_N$ son vectores $p \times 1$ independientes, observados de una $N(\theta, \Sigma)$, entonces $\overline{x}$ y V son independientes.*

Demostración. Del Teorema anterior se tiene que si $\Gamma = (\gamma_{ij})$, con $i, j = 1, \ldots, N$ es cualquier matriz ortogonal $N \times N$ con $\gamma_{Nj} = N^{-\frac{1}{2}}$, $j = 1, \ldots, N$, y si

$$Y_i = \sum_{j=1}^{N} \gamma_{ij} x_j,$$

con $i = 1, \ldots, N$, entonces los $Y'_i s$ son independientes con media cero para $i = 1, \ldots, N-1$ y $Y_N = \sqrt{N}\,\overline{x}$.

Sin embargo, probaremos ahora que la independencia también se cumple para $i = N$.

Como consecuencia, para $i \neq N$,

$$\begin{aligned} Cov(Y_N, Y_i) &= E[(Y_N - E(Y_N))(Y_i - E(Y_i))'] \\ &= E[(Y_N - \sqrt{N}\theta)(Y_i' - 0)] \\ &= E[(Y_N - \sqrt{N}\theta)Y_i'] \\ &= E(Y_N Y_i') - \sqrt{N}\theta E(Y_i') \\ &= E(Y_N Y_i') \\ &= 0, \end{aligned}$$

esto por (3.8). Así concluimos que

$$Cov(Y_N, Y_i) = 0, \qquad i = 1, \ldots, N-1.$$

Puesto que, $\overline{x}$ solo depende de Y_N, y como se mostro en el teorema 3.4, V solo depende de $Y_1, \ldots, Y_{N-1}$, obtenemos que, $\overline{x}$ y V son independientes.

□

En conclusión, para $x_1, x_2, \ldots, x_N$ una muestra aleatoria de vectores $p \times 1$ independientes e idénticamente distribuidos como $N(\theta, \Sigma)$ y densidad conjunta (Verosimilitud) dada por:

$$p(x_1, x_2, \ldots, x_N | \theta, \Sigma) = \frac{1}{(2\pi)^{\frac{Np}{2}} |\Sigma|^{\frac{N}{2}}} \exp\{-\frac{1}{2} \sum_{j=1}^{N} (x_j - \theta)' \Sigma^{-1} (x_j - \theta)\}$$

hemos conseguido en una primera fase los estimadores suficientes para los parámetros θ y Σ, dados por $\overline{x}$ y V respectivamente.

Luego, se determinaron los estimadores por máxima verosimilitud para tales parámetros, donde se apreció que estos coinciden con los estimadores suficientes encontrados en el teorema 3.1.

Seguidamente se estudió el tipo de distribución que seguian los estimadores de θ y Σ, obeniendo de esta manera que $\pounds(\overline{x}) = N(\theta, \frac{\Sigma}{N})$ y $\pounds(V) = W(\Sigma, p, n)$, con $n = N - 1$ grados de libertad. Además, en un último análisis se probó la independencia de los estimadores $\overline{x}$ y V.

De esta manera se finaliza el estudio de los estimadores de máxima verosimilitud para la distribución normal multivariada.

Referencias

[1] Anderson, T.W.(1958). An Introduction to Multivariate Statistical Analysis. New York: John Wiley and Sons.

[2] Anderson, Hair, Tatham y Black. (2005). Análisis Multivariante. 5ta. Edición. Pearson Prentice Hall.

[3] Berger J.O. (1985). Statistical Decision Theory and Bayesian Analysis. 2nd ed. New York: Springer.

[4] Bernardo, J.M. y Smith A.F.M. (1994). Bayesian Theory. New York: Wiley.

[5] Box, G.E.P y Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley Publishing Co.

[6] Bradley, P.C. y Thomas, A.L. (1996). Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall.

[7] Casella, G. y Berger, R.L. (1990). Statistical Inference. Thomson.

[8] Congdon Peter (2001). Statistical Modelling. New York: John Wiley and Sons.

[9] Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

[10] Johnson, R. y Wichern, D. (1998). Applied Multivariate Statistical Analysis. 4ta. Edition. Prentice Hall.

[11] Mendenhall, Scheaffer y Wackerly. (1986). Estadística Matemática con Aplicaciones. 3ra. Edición.

[12] Peña Daniel. (2002). Análisis de Datos Multivariantes. McGraw-Hill, Madrid.

[13] Press, James S. (1982). Applied Multivariate Analysis: using Bayesian and Frequentist Methods of Inference. Second Edition. Malabar, Florida.

[14] West, M. and Harrison, J. (1997). Bayesian Forecasting and Dynamic Models, 2nd.ed. Springer, New York.