FRANCISCO MONTES SUAY

ESTADÍSTICA MULTIVARIANTE (EXPLICADA)

UNIVERSITAT DE VALÈNCIA

Índice general

Capítulo 1

Álgebra matricial

ALGEBRA MATRICIAL

Se pretende en este primer capítulo recordar algunos conceptos referentes al cálculo matricial y materias con él relacionadas, por ser éste útil matemático de gran importancia para facilitar los cálculos que el manejo de muestras multivariantes implican.

Evidentemente resultaría ocioso insistir en algunos conceptos de sobra conocidos, a saber, definiciones referente a los distintos tipos de matrices, operaciones elementales con vectores y matrices, como suma, producto, producto por un escalar, etc. así como la correspondencia biunívoca existente entre las matrices y las formas lineales, el producto interior vectorial, la trasposición de matrices, la definición y obtención del determinante de una matriz cuadrada, etc.—

Ocupémonos ahora de otros conceptos, existen de algunos de ellos han sido olvidadas propiedades de interés.

1. INVERSA DE UNA MATRIZ

Solo existe en para las matrices cuadradas no singulares. Viene definida por $A.A^{-1} = A^{-1}.A = I$.

Se impulsa que su existencia esta ligada al valor del determinante de A, por cuanto si este es 0, aquella no existe. Propiedades de interés.

a) La inversa de una matriz simetrica tambien lo es

b) La inversa de la traspuesta es la traspuesta de la inversa.

c) La inversa del producto verifica

$$(ABC)^{-1} = C^{-1}.B^{-1}.A^{-1}$$

d) La inversa de una matriz diagonal es una matriz diagonal cuyos elementos son los inversos de los elementos primitivos.

2. RANGO DE UNA MATRIZ

Sea A una matriz $k \times p$, con $k \leq p$. El rango de A es el número de vectores fila linealmente independientes en la matriz. Si $p \leq k$, la definición es analoga sustituyendo fila por columna. Se demuestra, que en cualquier caso el rango de una matriz es unico, independientemente de su obtenido considerando los vectores filas o columnas.

Atendiendo al valor de los determinante de los menores complementarios en una matriz A, podemos afirmar que:

La matriz A es del rango r si contiene al menos un menor de $r \times r$ distinto de cero, siendo nulos todos los menores con dimensión mayor que r.

Algunas propiedades de interés:

a) El rango de A' y el de A son iguales.

b) El rango de A'A es igual al de A e igual al de AA'.

c) El rango de A no se altera si se pre o post-multiplica A por una matriz no singular.

Como consecuencia de c) podemos afirmar que el rango de una matriz A no cambia si

c.1) Se intercambian dos filas (columnas) cualesquiera

c.2) multiplicando cada fila (columna) por un escalar

c.3) añadiendo a una fila (columna) los elementos de otra previamente multiplicados por un escalar.

Todo ello por que las operaciones descritas en estas tres propiedades pueden llevarse a cabo mediante premultiplicación (caso de filas) o postmultiplicación (caso de columnas) de la matriz A por una matriz no singular.

$$A = \begin{bmatrix} 5 & 1 \\ -2 & 3 \\ 3 & 2 \end{bmatrix} \qquad E = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad EA = \begin{bmatrix} -2 & 3 \\ 5 & 1 \\ 3 & 2 \end{bmatrix}$$

Operaciones sobre las filas y columnas nos permiten transformar una matriz A de rango r, en su correspondiente canónica F, que tiene la forma

$$F = \begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 & \cdots & 0 \\ 0 & 0 & \cdots & \cdots & & 0 \end{bmatrix}$$

es decir, 1's en las r primeras posiciones de la diagonal principal y ceros en el resto. Si designamos por $R_s - R_p$ y $C_s - C_q$ estos cambios en filas y columnas podemos escribir

$$R_p - R_s \, A \, C_s - C_q = PAQ = F$$

Si A es cuadrada y de rango n (completo), entonces

$$R_q - R_1 \, A = I$$

y de aquí

$$\underline{A^{-1} = R_p - R_s \, I}$$

Este hecho es aprovechado para invertir matrices de forma rápida y sencilla, útil en los casos de matrices de elevada dimensión.

3. INVERSA GENERALIZADA

La definición dada anteriormente es solo aplicable a matrices cuadradas no singulares. Se trata de introducir un concepto de inversa que generalice el anterior.

La G-inversa de una matriz A de cualquier dimensión viene dada por:

$$(1) \quad AGA = A$$

Si la dimensión de A es $p \times q$ y su rango r, la dimensión de G es $q \times p$ y su rango es r.

Existen otras definiciones, por ejemplo la debida a Penrose que exige además

$$GAG = G \qquad (GA)' = GA \qquad (AG)' = AG$$

La obtención de la G definida en (1) puede hacerse a partir de la reducción canónica de A, como sigue:

$$PAQ = F = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$$

donde D es una matriz diagonal $r \times r$ (no necesariamente I) y el resto matrices nulas de dimensión adecuada. P y Q las anteriormente definidas a partir de operaciones en filas y columnas. Definimos ahora

$$F^- = \begin{bmatrix} D^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

entonces $G = QF^-P$, es una inversa generalizada de A, que desde luego no es única. Si A es cuadrada y no singular, entonces $G = A^{-1}$ y es única.

4. SISTEMAS DE ECUACIONES LINEALES

Recordemos que un sistema de la forma

$$a_{11}x_1 + \cdots + a_{1n}x_n = c_1$$

$$\vdots$$

$$a_{m1}x_1 + \cdots + a_{mn}x_n = c_m$$

admite una representación matricial mediante

$$Ax = c$$

con

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad c = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}$$

Si $c = [0]$, el sistema se llama homogéneo

(1) El sistema posee una solución si y sólo si la matriz ampliada $[A\ c]$ de dimensión $m \times (n+1)$ tiene el mismo rango que la matriz A. (Se dice entonces que el sistema es consistente)

Estudiaremos los tipos de sistemas y sus soluciones en función de los rangos de A y $[Ac]$.

a) Sistemas no homogeneos: A cuadrada y no singular

La solución es única y viene dada por

$$x = A^{-1}c$$

b) Sistemas no homogeneos: A $m \times n$ y de rango r

Si se satisface (1) existe una solución que puede obtenerse hallando r ecuaciones linealmente independientes y resolviendo para r incognitas en función de las constantes y de las restantes $n-r$ incognitas.

c) Sistemas homogeneos

Evidentemente el rango de $[A\ 0]$ es siempre el mismo que el de $[A]$ por tanto siempre tienen solución.

c.1) Si rango$(A) = n$ = número de incognitas, solución única dada por $x = [0]$

c.2) En cualquier otro caso solución del tipo expresado en b).

En cualquier caso la solución puede expresarse en función de la inversa generalizada mediante la siguiente expresión debida a Rao (1962), siendo el sistema consistente.

$$X^* = Gc + (GA - I)z$$

donde z es un vector $n \times 1$ de constantes arbitrarias. Obsérvese que si existe A^{-1}, entonces $X^* = A^{-1}C$, como ya sabíamos

5. MATRICES ORTOGONALES

Un vector se dice que es ortogonal respecto de otro si su producto interior es nulo.

Una matriz T se dice ortogonal si está compuesta por vectores fila ortogonales, siendo cuadrada. Más concretamente se exige que los vectores sean ortonormales, es decir de norma unidad. De esta definición se desprende que:

$$TT' = T'T = I$$

Algunas propiedades de interés:

a) Las columnas de una matriz ortogonal son ortogonales.

b) El determinante de una matriz ortogonal es siempre 1 o -1.

En efecto $|TT'| = |I| = 1$ y $|T| = |T'| = \pm 1$

3) El producto de matrices ortogonales de la misma dimensión es también una matriz ortogonal

5. FORMAS CUADRÁTICAS

Una forma cuadrática en las variables $x_1 \dots x_n$ es una expresión del tipo

$$f(x_1 \dots x_n) = a_{11}x_1^2 + \dots + a_{nn}x_n^2 + 2a_{12}x_1x_2 + \dots + 2a_{1n}x_1x_n + \dots + 2a_{n-1,n}x_{n-1}x_n =$$

$$= \sum\sum a_{ij}x_ix_j \quad \text{con } a_{ij} = a_{ji} \text{ pudiendo ser cero algún } a_{ij}$$

De inmediato se observa que una forma cuadrática admite notación matricial, como sigue:

$$f(x_1 \dots x_n) = x'Ax$$

donde A es una matriz simétrica $n \times n$. Las formas cuadráticas juegan un importante papel en estadística, tanto univariante como multivariante. Por ejemplo, la suma de los cuadrados de las desviaciones respecto de la media muestral

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum x_i\right)^2$$

es una forma cuadrática cuya matriz de coeficientes viene dada por:

$$A = \begin{bmatrix} \frac{N-1}{N} & \frac{-1}{N} & \dots & \frac{-1}{N} \\ \frac{-1}{N} & \frac{N-1}{N} & \dots & \frac{-1}{N} \\ \hline \frac{-1}{N} & \frac{-1}{N} & \dots & \frac{N-1}{N} \end{bmatrix}$$

Carácter de una forma cuadrática

a) Definida positiva, semidefinida positiva.

Una matriz simétrica A y su forma cuadrática asociada se dice que es definida positiva si

$$x'Ax > 0 \quad , \forall x \neq 0 .$$

Si $x'Ax \geq 0 \quad , \forall x \neq 0$ entonces se la denomina semidefinida positiva

b) Definida negativa, semidefinida negativa.

Id con los signos cambiados.

c) Indefinida

Si puede tomar cualquier valor positivo, negativo o nulo.

Algunas propiedades de las formas cuadráticas definidas y semidefinidas positivas

1) Las formas cuadráticas definidas positivas tienen matrices de rango completo. Es posible mediante completamientos de cuadrados sucesivos reducir dichas formas a expresiones del tipo

$$d_1y_1^2 + \dots + d_ny_n^2 \qquad d_i > 0 , \forall i$$

Algo similar ocurre con las semidefinidas positivas, pero ahora la expresión es

$$d_1 y_1^2 + \cdots + d_r y_r^2 \qquad d_i > 0 \qquad r \le n \text{ es el rango de la matriz.}$$

2) No obstante para determinar el carácter de una forma vamos a dar ahora una condición necesaria y suficiente. Formaremos en primer lugar la sucesión de los determinantes de los menores principales, a saber

$$p_0 = 1 \qquad p_1 = a_{11} \qquad p_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \quad \cdots \quad p_i = \begin{vmatrix} a_{11} & \cdots & a_{1i} \\ & & \\ a_{i1} & \cdots & a_{ii} \end{vmatrix} \quad \cdots \quad p_n = |A|$$

Si el rango de A es r, se dice que A es regular si $p_r \neq 0$ y no hay dos p_i consecutivos en la sucesión iguales a cero. Siempre es posible, intercambiando filas y columnas obtener una matriz simétrica en una forma regular. Entonces, si A es una matriz dispuesta regularmente, tenemos:

2.a) Una condición nec. y suficiente para la definición positiva es que $p_i > 0$, $i = 1, \ldots, n$

2.b) Id. para la semidefinición positiva es que $p_1 > 0, \ldots p_r > 0$ y las restantes $n-r$ p_i son nulas, donde $r \le n$.

Se puede demostrar aplicando esto que la forma cuadrática de la suma de los cuadrados de las desviaciones respecto de la media es semidefinida positiva con rango N-1 para cualquier matriz. Los distintos p_i tienen un valor

$$p_i = \frac{1}{N}\left(N - 1 - (i-1)\right).$$

6. RAICES Y VECTORES CARACTERISTICOS DE UNA MATRIZ

Las raíces características de una matriz A, $p \times p$, son las soluciones de la ecuación determinante

$$|A - \lambda I| = 0$$

El determinante es un polinomio de grado p en A y tiene por tanto p raíces. Aplicando el desarrollo de Laplace del determinante característico, podemos escribir

$$|A - \lambda I| = (-\lambda)^p + S_1 (-\lambda)^{p-1} + \cdots + S_{p-1}(-\lambda) + |A|$$

donde S_i es la suma de todos los determinantes de los menores principales de dimensión $i \times i$. S_1 será en particular la suma de los elementos de la diagonal principal de A, o sea $tr(A)$. De la teoría de las soluciones de una ecuación polinomial, tenemos de inmediato que:

1) El producto de las raíces características de A es igual a $|A|$

2) La suma de las raíces características de A es igual a $tr(A)$

Algunas propiedades de interés y de posterior uso de las raíces características son:

a) Las raíces características de una matriz simétrica con elementos reales son todas reales.

b) Las raíces características de una matriz definida positiva son todas positivas.

c) Para una matriz $n \times n$ semidefinida positiva de rango $r \le n$, hay exactamente r raíces características positivas y $n-r$ nulas.

d) Las raíces características no nulas del producto AB son iguales a las raíces no nulas de BA. Por tanto $tr(AB) = tr(BA)$

e) Las raíces características de una matriz diagonal son los elementos de la diagonal.

A cada raíz característica podemos asociarle un vector característico, cuyos elementos satisfacen la ecuación homogénea

$$[A - \lambda_i I] x_i = 0$$

como el determinante $|A - \lambda_i I| = 0$ existe siempre una solución no trivial para X_i que quedará determinado a excepción de un factor de escala. Dada la importancia de los vectores y raíces características obtenidas a partir de matrices simétricas, damos a continuación algunas propiedades interesantes.

a) Si $\lambda_i \neq \lambda_j$ son raíces características de una matriz simétrica A, entonces X_i y X_j, sus vectores asociados son ortogonales. En efecto:

$$AX_i = \lambda_i X_i \qquad AX_j = \lambda_j X_j$$

premultiplicando por X_j' y X_i' respectivamente

$$X_j' A X_i = \lambda_i X_j' X_i \quad y \quad X_i' A X_j = \lambda_j X_i' X_j$$

y dada la simetría de A, tenemos

$$\lambda_i X_j' X_i = \lambda_j X_i' X_j$$

y siendo $\lambda_i \neq \lambda_j$ esto puede ocurrir que $X_j' X_i = 0$.

b) Para cualquier matriz simétrica real A existe una matriz ortogonal P tal que

$$P'AP = D$$

donde D es una matriz diagonal cuyos elementos son las raíces características de A. Las columnas de P ~~son~~ son los vectores característicos normalizados de A

Estas propiedades para las matrices simétricas tienen una aplicación inmediata en las formas cuadráticas. En efecto, si aplicamos ~~a~~ la transformación ortogonal

$$X = PY$$

a las variables de la forma cuadrática $X'AX$, tendremos:

$$X'AX = Y'P'APY = Y'DY = \lambda_1 Y_1^2 + \cdots + \lambda_r Y_r^2$$

7. MATRICES PARTICIONADAS

Se trata de presentar una matriz ~~como~~ función de las submatrices que la componen y que vienen determinadas por agrupamiento de determinadas filas y columnas siguiendo criterios de homogeneidad, etc. Puede entonces una matriz aparecer como

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \cdots & \cdots & \cdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix}$$

donde A_{ij} contiene r_i filas y c_j columnas y donde todas las submatrices de una misma fila tienen el mismo número de filas y todas las de una misma columna el mismo número de columnas. Con estas matrices pueden llevarse a cabo las mismas operaciones que en el caso no agrupado, teniendo en cuenta, en todo caso, la naturaleza no escalar de los elementos. Así por ejemplo si A y B tienen submatrices de igual dimensión podemos escribir

$$A+B = \begin{bmatrix} A_{11}+B_{11} & \cdots & A_{1n}+B_{1n} \\ \cdots & \cdots & \cdots \\ A_{m1}+B_{m1} & \cdots & A_{mn}+B_{mn} \end{bmatrix}$$

o bien

$$AB = \begin{bmatrix} \sum_{i=1}^{n} A_{1i}B_{i1} & \cdots & \sum_{i=1}^{n} A_{1i}B_{ip} \\ \cdots & \cdots & \cdots \\ \sum_{i=1}^{n} A_{mi}B_{i1} & \cdots & \sum_{i=1}^{n} A_{mi}B_{ip} \end{bmatrix}$$

Existen expresiones para la matriz inversa de una matriz particionada en función de las submatrices que la conforman, así como para el determinante (caso de tratarse de una matriz cuadrada no singular). Por ejemplo, para una matriz de la forma

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$ con A_{11} y A_{22} cuadradas y ~~no singulares~~ (dada su naturaleza de menor principal)

$$A^{-1} = \begin{bmatrix} (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & -(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} & A_{22}^{-1} + A_{22}^{-1}A_{21}(A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}A_{12}A_{22}^{-1} \end{bmatrix}$$

para el determinante

$$|A| = |A_{11}| \cdot |A_{22} - A_{21}A_{11}^{-1}A_{12}| \quad \text{si} \quad A_{11} \text{ no singular, o bien}$$

$$|A| = |A_{22}| \cdot |A_{11} - A_{12}A_{22}^{-1}A_{21}| \quad \text{si} \quad A_{22} \text{ no singular.}$$

8. DIFERENCIACIÓN CON VECTORES Y MATRICES.

Sea $f(x)$ una función continua de los elementos del vector $x' = [x_1 \; \cdots \; x_p]$ cuyas primeras y segundas derivadas parciales existen para todos los puntos x de una determinada región del espacio euclídeo p-dimensional. El vector operador derivada parcial lo definimos como

$$\frac{\partial}{\partial x} = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_p} \end{bmatrix}$$

aplicándolo a $f(x)$ nos lleva a $\dfrac{\partial f(x)}{\partial x}' = \left[\dfrac{\partial f(x)}{\partial x_1} \; \cdots \; \dfrac{\partial f(x)}{\partial x_p} \right]$.

Algunos casos de funciones y sus derivadas especialmente importantes son:

1) La función constante

2) $f(x) = a'x$, $\dfrac{\partial f(x)}{\partial x}' = a'$

3) $f(x) = x'Ax$ una forma cuadrática. Escribamos la forma de la manera siguiente:

$$x'Ax = \sum \sum a_{ij}x_ix_j = a_{ii}x_i^2 + 2x_i \sum_{\substack{j=1 \\ i \neq j}} a_{ij}x_j + \sum \sum_{\substack{h \neq i \\ j \neq i}} a_{hj}x_hx_j$$

derivando parcialmente respecto de x_i

$$\frac{\partial f(x)}{\partial x_i} = 2a_{ii}x_i + 2\sum_{\substack{j=1 \\ i \neq j}} a_{ij}x_j = 2\sum_{j=1}^{p} a_{ij}x_j = 2a_i'x$$

si por a_i' designamos la i-ésima fila de la matriz simétrica A. Entonces el vector de las derivadas parciales vendrá dado por

$$\frac{\partial f(x)}{\partial x} = 2Ax$$

4) Si se trata de una función cuadrática más general $h(x) = (a - Cx)'K(a - Cx)$ con K simétrica $N \times N$, $C = [C_1, \ldots, C_p]$ matriz de constantes $N \times p$, a un vector $N \times 1$.

haciendo $u = a - Cx$, las derivadas de $h(x)$ pueden calcularse aplicando la regla de la cadena:

$$\frac{\partial h(x)}{\partial x_i} = \sum_{j=1}^{N} \frac{\partial (u'Ku)}{\partial u_j} \cdot \frac{\partial u_j}{\partial x_i} = -2 \sum_{j=1}^{N} k_j u \cdot c_{ji} = -2 u' K C_i \qquad i = 1, \dots, p$$

y de aquí

$$\frac{\partial h(x)}{\partial x} = -2\, c' k (a - Cx).$$

5) La matriz de las derivadas de segundo orden (parciales) se la denomina hessiano. En el caso de la forma cuadrática $f(x) = x'Ax$ dicha matriz es $2A$. El hessiano es necesariamente simétrico si nuestras condiciones originales de continuidad y existencia de las derivadas parciales de primero y segundo orden son satisfechas por $f(x)$.

Determinación de máximos y mínimos

Ya conocemos las condiciones necesarias para la existencia de maximos o mínimos. En cualquier lo más interesante para nosotros es las aplicaciones. Si queremos obtener los extremos estacionarios de $h(x)$ del apartado 4) anteriormente, tendremos que el valor x_0 donde se encuentra el extremo debe satisfacer:

$$c'Kcx = c'Ka$$

y si $c'Kc$ es de rango completo p, tendremos una solución única dada por $x = (c'Kc)^{-1} c'Ka$. Si k es definida positiva y C tiene rango p, el hessiano $2c'Kc$ es definido positivo y estaremos de un mínimo.

Maximos y mínimos restringidos

Se utiliza entonces el metodo de los multiplicadores de lagrange. Consiste esencialmente en lo siguiente

$$h(x,\lambda) = f(x) - \lambda [g(x) - c]$$

donde $f(x)$ es la función a maximizar o minimizar y $g(x) = c$ es la restricción. Para un valor estacionario restringido tendremos

$$\frac{\partial h(x,\lambda)}{\partial x} = \frac{\partial f(x)}{\partial x} - \lambda \frac{\partial g(x)}{\partial x} = 0$$

$$\frac{\partial h(x,\lambda)}{\partial \lambda} = -g(x) + c = 0$$

en la practica se obtiene x despues de haber eliminado λ en el sistema de ecuaciones.

Veamos un ejemplo para el caso de una forma cuadratica $f(x) = x'Ax$, con A definida positiva, sujeta a la restricción $x'x = 1$. entonces

$$h(x,\lambda) = x'Ax - \lambda(x'x - 1)$$

entonces

$$\frac{\partial h(x,\lambda)}{\partial x} = 2Ax - 2\lambda x = 0 \longrightarrow [A - \lambda I]x = 0$$

que es la ecuación característica de A. Premultiplicando por x' y haciendo de la restricción tenemos

$$x'[A - \lambda I]x = 0 \longrightarrow x'Ax - \lambda = 0 \longrightarrow \lambda = x'Ax$$

entonces si $f(x)$ debe ser un máximo, λ deberá ser la mayor de todas las raices características, y x el vector característico asociado. la restricción aún de que dicho vector debe tener módulo 1. Resulta igualmente para el mínimo de $f(x)$.

Derivada de una matriz y de un determinante

La derivada del determinante de A respecto de un elemento a_{ij} podemos encontrarla a partir del desarrollo de $|A|$ en cofactores de la i fila y la j columna. En efecto:

$$\frac{\partial |A|}{\partial a_{ij}} = \frac{\partial}{\partial a_{ij}} \left(a_{i1} A_{i1} + \cdots + a_{ij} A_{ij} + \cdots + a_{in} A_{in} \right) = A_{ij}$$

Para derivar una matriz respecto de los elementos que la constituyen tendremos: Sea $\mathbf{X}$ una matriz $m \times n$ con elemento genérico x_{ij} y su derivada respecto a este elemento será

$$\frac{\partial \mathbf{X}}{\partial x_{ij}} = J_{ij}$$

donde J_{ij} es una matriz $m \times n$ con un 1 en la posición ij-ésima y ceros en el resto. Si $\mathbf{X}$ es simétrica

$$\frac{\partial \mathbf{X}}{\partial x_{ij}} = J_{ij} + J_{ji} \qquad i \neq j$$

La regla para derivar un producto de matrices es similar a la utilizada para escalar. Supongamos entonces que $\mathbf{X}$ e $\mathbf{Y}$ son dos matrices conformes y que ambas son funciones ambas de z a través de $x_{ij}(z)$ e $y_{ij}(z)$ entonces

$$\frac{\partial \mathbf{XY}}{\partial z} = \frac{\partial \mathbf{X}}{\partial z} \mathbf{Y} + \mathbf{X} \frac{\partial \mathbf{Y}}{\partial z}$$

Esta fórmula nos va a permitir obtener la derivada de la inversa de una matriz cuadrada no singular. Sea

$$\mathbf{I} = \mathbf{X} \mathbf{X}^{-1}$$

$$\frac{\partial \mathbf{I}}{\partial x_{ij}} = \frac{\partial \mathbf{X}}{\partial x_{ij}} \cdot \mathbf{X}^{-1} + \mathbf{X} \cdot \frac{\partial \mathbf{X}^{-1}}{\partial x_{ij}} = 0$$

$$J_{ij} \mathbf{X}^{-1} + \mathbf{X} \frac{\partial \mathbf{X}^{-1}}{\partial x_{ij}} = 0 \longrightarrow \frac{\partial \mathbf{X}^{-1}}{\partial x_{ij}} = - \mathbf{X}^{-1} J_{ij} \mathbf{X}^{-1}$$

si $\mathbf{X}$ es simétrica

$$\frac{\partial \mathbf{X}^{-1}}{\partial x_{ij}} = \begin{cases} - \mathbf{X}^{-1} J_{ii} \mathbf{X}^{-1} & i = j \\ - \mathbf{X}^{-1} (J_{ij} + J_{ji}) \mathbf{X}^{-1} & i \neq j \end{cases}$$

Capítulo 2

Muestras de una normal multivariante

1. VARIABLES ALEATORIAS MULTIDIMENSIONALES

Un vector aleatorio multidimensional $\mathbf{X}$, es un vector del tipo

$$\mathbf{X}' = [X_1, \dots, X_p]$$

cuyos elementos son variables aleatorias continuas cuyas densidades vienen dadas por $f_i(x_i)$, y cuyas funciones de distribución son $F_i(x_i)$ $i = 1, \dots, p$. Recordemos que la función de distribución conjunta viene dada por

$$F(x_1, \dots, x_p) = Pr(X_1 \le x_1, \dots, X_p \le x_p).$$

Si la función es absolutamente continua, existirá $f(x_1, \dots, x_p)$ tal que

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

Si las variables $X_1, \dots, X_p$ son independientes la obtención de $f(x_1, \dots, x_p)$, y de $F(x_1, \dots, x_p)$ se simplifican enormemente aplicando el teorema de factorización, ambas

$$f(x_1, \dots, x_p) = f_1(x_1) \cdots f_p(x_p)$$

$$F(x_1, \dots, x_p) = F_1(x_1) \cdots F_p(x_p)$$

No siempre puede llevarse a cabo esta supuesta de independencia, depende de las condiciones del problema. En cualquier caso, en análisis multivariante no siempre se verificará la independencia entre las variables observadas. La densidad conjunta de cualquier subconjunto de variables del conjunto original, se puede obtener mediante:

$$g(x_1, \dots, x_p) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_{p+q})\, dx_{p+1} \cdots dx_{p+q}$$

para la función de distribución

$$G(x_1, \dots, x_p) = F(x_1, \dots, x_p, \infty, \dots, \infty)$$

Es importante que los límites de integración estén bien definidos, aunque frecuentemente siempre deberán ser definidos entre $-\infty$ y $+\infty$. Por ejemplo, para X_1, X_2 variables aleatorias i.i.d. con f.d.p. $g(y)$ y f.d. $G(y)$ si queremos obtener las f.d.p y f.d. de las nuevas variables

$$X_2 = \max(Y_1, Y_2)$$
$$X_1 = \min(Y_1, Y_2)$$

se obtiene fácilmente

$$f(x_1, x_2) = 2 g(x_1) g(x_2) \quad , \quad -\infty < x_1 \le x_2 < +\infty \quad , \ 0 \text{ en el resto}$$

Las marginales vienen dadas por

$$f_1(x_1) = \int_{-\infty}^{+\infty} 2\, g(x_1) g(x_2)\, dx_2 = \cancel{\int_{-\infty}^{x_1} 0} + \int_{x_1}^{+\infty} 2\, g(x_1) g(x_2)\, dx_2 = 2 g(x_1)[1 - G(x_1)]$$

análogamente

$$f_2(x_2) = \int_{-\infty}^{+\infty} 2\, g(x_1) g(x_2)\, dx_1 = \int_{-\infty}^{x_2} 2\, g(x_1) g(x_2)\, dx_1 + \cancel{\int_{x_2}^{+\infty} 0} = 2\, G(x_2) g(x_2)$$

Observese además que $f(x_1, x_2) \neq f_1(x_1) f_2(x_2)$ no hay pues independencia.

Condicionales

Se trata de obtener la densidad de $X_1 - X_p$ dado que las otras variables han tomado un valor determinado $X_i = x_i$ $i = p+1, \dots, p+q$. Recordemos que dicha densidad viene dada por

$$h(x_1 \dots x_p / x_{p+1} \dots x_{p+q}) = \frac{f(x_1 \dots x_{p+q})}{g(x_{p+1} \dots x_{p+q})}$$

donde $f(x_1 - x_{p+q})$ es la densidad conjunta y $g(x_{p+1} - x_{p+q})$ es la marginal de las variables que condicionan. En el caso de independencia la condicionada actuará/será directamente en la marginal correspondiente. La función de distribución se obtiene mediante integración adecuada. Por ejemplo, retomando con las variables X_1 y X_2 antes definidas

$$f(x_1/x_2) = \frac{g(x_1)}{g(x_2)} \qquad -\infty < x_1 \leq x_2 < \infty$$

$$0 \text{ en el resto}$$

la distribución correspondiente será

$$F(x_1/x_2) = \begin{cases} \dfrac{g(x_1)}{g(x_2)} & -\infty < x_1 \leq x_2 < \infty \\ 1 & x_2 \leq x_1 < \infty \end{cases}$$

Momentos

El vector media viene dado por

$$E(X') = [E(X_1), \dots, E(X_p)]$$

La extensión del concepto de varianza y covarianza al caso multidimensional viene dado por

$$E\{[X - E(X)][X - E(X)]'\} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \dots \sigma_{1p} \\ \sigma_{12} & \sigma_{22} \dots \sigma_{2p} \\ \sigma_{1p} & \sigma_{2p} \dots \sigma_{pp} \end{bmatrix} = \Sigma$$

Matriz que se la denomina matriz de covarianzas de X

Las expresiones matriciales para las varianzas y covarianzas de determinados cambios de variables vienen dadas por:

a) $Y = a_1 X_1 + \dots + a_p X_p = a' X$ $\quad var(Y) = var(a'X) = \sum\sum a_i a_j \sigma_{ij} = a' \Sigma a$

b) $Y = a' X$, $Z = b' X$ $\quad cov(Y,Z) = \{\sum a_i b_j \sigma_{ij} = a' \Sigma b$

c) Expresando de forma más general el cambio ; A matriz $r \times p$ B matriz $s \times p$

$Y = AX$ $\quad Z = BX$ tenemos $\quad cov(Y,Y) = A \Sigma A'$

$cov(Z,Z) = B \Sigma B'$

$cov(Y,Z) = A \Sigma B'$

Finalmente el coeficiente de correlación viene definido mediante

$$\rho_{ij} = \frac{cov(X_i, X_j)}{\sqrt{var(X_i) \cdot var(X_j)}}$$

La matriz de correlaciones viene dada por

$$P = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}$$

La relación existente entre las matrices de covarianza y de correlación viene dada por:

$$P = D\left(\frac{1}{\sigma_i}\right) \Sigma \, D\left(\frac{1}{\sigma_i}\right)$$

$$\Sigma = D(\sigma_i) \, P \, D(\sigma_i)$$

donde $D(\sigma_i)$ es la matriz diagonal de las desviaciones típicas de las variables.

2. LA DISTRIBUCION NORMAL MULTIVARIANTE

En adelante las muestras a las que hagamos referencia provendrán de poblaciones normales multivariantes lo que pudiera considerarse como una restricción puede justificarse en base a los siguientes hechos:

a) Cualquier vector aleatorio que surja como la suma de un gran número de vectores aleatorios distribuidos idéntica e independientemente, tiene una distribución normal multivariante a medida que el número de estos vectores crece sin límite. Esto no es más que aplicar el teorema central del límite en su versión multivariante. Este tipo de modelos suma parece ser bastante realista a la hora de explicar gran parte de los fenómenos que surgen en biología y en ciencias del comportamiento.

b) Diferentes modelos para los vectores considerados podrían conducir a muy diferentes distribuciones conjuntas de los elementos ~~considerados~~ cuya complejidad matemática impediría el desarrollo de las distribuciones muestrales de los test estadísticos y de los estimadores usuales. Tales distribuciones deberían ser conocidas para cada una de las poblaciones fundamentales del modelo. Sin embargo, parece probable que con la excepción de algunos casos patológicos, el teorema central del límite, en su versión multivariante garantice que las distribuciones para grandes muestras de los tests estadísticos conducirán a conclusiones similares acerca del estado de la naturaleza.

Hecha esta salvedad, justificación del posterior uso y abuso de la normal multivariante vamos con ella.

Recordemos que la distribución normal para el caso univariante tenía la expresión, en forma de función de densidad,

$$\phi(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad -\infty < x < \infty$$

Para el caso de n variables, $x_1 \cdots x_n$, todas normales e independientes tenemos:

$$\phi(x_1 \cdots x_n) = \frac{1}{(2\pi)^{n/2}\sigma_1 \cdots \sigma_n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right]$$

y si sustituimos

$$x' = [x_1 \cdots x_n] \qquad \mu' = [\mu_1 \cdots \mu_n] \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

la densidad conjunta anterior adquiere la expresión

$$\phi(x) = \frac{1}{(2\pi)^{n/2}\,|\Sigma|^{1/2}} \exp\left[-\frac{1}{2} z' \Sigma^{-1} z\right] \qquad \text{con} \quad z' = (x-\mu)' \;, \quad z = (x-\mu)$$

~~Sirve esta expresión recapitulativa~~

Esta expresión es una expresión particular de la densidad de una normal multivariante, porque, en efecto, no siempre el vector x considerado estará constituido por variable aleatorias independientes. La generalización de la anterior expresión se obtiene haciendo que Σ sea una matriz simetrica definida positiva, entonces la función $\phi(x)$ es positiva para cualquier valor de x y se cumple además que

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \phi(x_1 \cdots x_n)\, dx_1 \cdots dx_n = 1$$

Es decir, $\phi(x)$ es una función de densidad. En este caso tenemos que:

- el i-ésimo elemento del vector μ es la esperanza de x_i
- el i-ésimo elemento de la diagonal principal de Σ es la varianza de x_i
- el ij-ésimo elemento, σ_{ij}, de la matriz Σ es la covarianza de x_i y x_j. Evidentemente si las $P(P-1)/2$ covarianzas son nulas, entonces las x_i son independientes.

Consideremos con algún detalle el caso $n=2$, de gran importancia en estadística teórica.

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_{22} \end{bmatrix}$$

La densidad conjunta será

$$\phi(x_1 x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2} \frac{1}{1-\rho^2} \left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho \frac{x_1-\mu_1}{\sigma_1} \cdot \frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2 \right] \right\}$$

si tipificamos las variables mediante la transformación

$z_i = \dfrac{x_i - \mu_i}{\sigma_i}$, tendremos

$$\phi(z_1, z_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2} \frac{1}{1-\rho^2} \left[z_1^2 - 2\rho z_1 z_2 + z_2^2 \right] \right\}$$

Posteriormente volveremos sobre esta densidad y veremos cuál es el papel de ρ.

Ejes principales en la densidad multinormal

El exponente

$$(x-\mu)' \Sigma^{-1} (x-\mu)$$

es una forma cuadrática y por tanto especifica la ecuación de un elipsoide en el espacio n-dimensional cuando lo igualamos a alguna constante positiva. La familia de elipsoides que se genera cuando igualamos a distintos valores c_i de esta constante característica propia todos ellos están centrados en μ. El primero de los ejes principales del elipsoide es aquella línea que pasa a través de su mayor dimensión. Si representamos a cualquiera de las líneas que pasan por μ, mediante su punto de corte en la superficie del elipsoide, el eje principal tendrá más coordenadas que maximizarán la longitud del segmento que μ y x determinan. El cuadrado de dicha longitud viene dado por

$$(x-\mu)'(x-\mu)$$

El cálculo de este primer eje vendría pues dado mediante la solución de un problema de máximos condicionados. A saber, maximizar

$$(x-\mu)'(x-\mu)$$

Sujeto a la restricción

$$(x-\mu)' \Sigma^{-1} (x-\mu) = c$$

Aquí debemos usar los multiplicadores de Lagrange, tendremos

$$f(x) = (x-\mu)'(x-\mu) - \lambda\left[(x-\mu)'\Sigma^{-1}(x-\mu) - c\right]$$

y derivando

$$\frac{\partial f(x)}{\partial x} = 2(x-\mu) - 2\lambda\Sigma^{-1}(x-\mu) = 2\left[I - \lambda\Sigma^{-1}\right](x-\mu) \#$$

e igualando a cero

$$\left[I - \lambda\Sigma^{-1}\right](x-\mu) = 0$$

Es decir, que las coordenadas del eje mayor deben satisfacer la anterior ecuación, que como Σ^{-1} sea singular, admite también la forma:

$$\left[\Sigma - \lambda I\right](x-\mu) = 0$$

lo que da a entender que las coordenadas del 1er eje principal son proporcionales a los unos de los vectores característicos de Σ. Pero veamos a cual. Si premultiplicamos por $4(x-\mu)'$, tendremos:

$$4(x-\mu)'\Sigma(x-\mu) - 4\lambda(x-\mu)'\Sigma^{-1}(x-\mu) = 0$$

$$4(x-\mu)'(x-\mu) = 4\lambda c$$

Por tanto para c fijo la longitud del 1er eje principal se maximizará tomando λ igual a la mayor de las raíces características de Σ. Por tanto podemos afirmar:

"La posición del primer eje principal del elipsoide de concentración viene especificada por los cosenos directores del vector característico normalizado α_1 asociado a la mayor raíz característica λ_1 de Σ".

Observese que la longitud del eje viene entonces dada por $l = 2\sqrt{\lambda_1 c}$

El segundo eje principal vendrá dado por λ_2, cuando dicho valor sea la segunda mayor raíz característica de Σ. El proceso continúa hasta obtener los n ejes principales del elipsoide.

Si las raíces características son distintas todas ellas

$$\lambda_1 > \lambda_2 > \cdots > \lambda_{n-1} > \lambda_n$$

entonces, recordando una propiedad que asegura que los vectores asociados son ortogonales

$$\alpha_i'\alpha_j = 0 \quad , \lambda_i \neq \lambda_j$$

tendremos que los n ejes están univocamente determinados y además son todos ellos perpendiculares entre sí. Si hubiera dos raíces consecutivas iguales, quiere ello decir, que el elipsoide tiene una sección circular a través del plano que las direcciones de los vectores asociados correspondientes determinan. En este, aunque pueden encontrarse ejes perpendiculares entre sí, su posición no es de ningún modo única. En general si λ_i es una raíz característica cuyo orden de multiplicidad es r_i, cada uno de los r_i ejes a determinar puede ser elegido perpendicular con los $r_i - 1$ restantes y con los $n - r_i$, aunque dichos ejes pueden ocupar una infinidad de direcciones principales. En tal caso, el elipsoide tiene forma hiperesferica en el subespacio que determinan las r_i raíces características y recibe que tiene variación isotropica en este subespacio. Posteriormente nos ocuparemos nuevamente de este tema cuando lo relacionemos con una determinada técnica estadística.

Consideremos ahora la nueva variable $Y' = [Y_1 \dots Y_n]$ cuyos elementos tienen valores en los ejs principales del elipsoide de concentración. Esta variable está relacionada con el vector original $\mathbf{x}$, mediante la transformación

$$Y = A'(\mathbf{x} - \mu)$$

donde la i-ésima columna de la matriz A es el vector característico normalizado α_i. La ortogonalidad de A implica que la transformación consiste en una rotación rígida de los ejs originales hasta coincidir con los ejs principales del elipsoide regida de una traslación del antiguo centro al centro de del elipsoide. La matriz de covarianzas del nuevo vector Y vendrá dada por

$$\Sigma_Y = A'\Sigma A$$

que como sabemos es una matriz diagonal cuyos valores (en la diagonal principal son) λ_i. Por tanto la varianza de la variable que se mueve en el i-ésimo eje principal vendrá dada por

$$\text{var}(Y_i) = \alpha_i' \Sigma \alpha_i = \lambda_i$$

por tanto

$$\text{cov}(Y_i Y_j) = \alpha_i' \Sigma \alpha_j = 0 \quad i \neq j$$

a condición de que todas las λ_i sean distintas o bien, en caso contrario, los α_i se hayan elegido ortogonales.

Así pues la transformación de los ejs principales da lugar a variables incorreladas cuyas varianzas son proporcionales a la longitud de los ejs (a ser medidos más exactamente).

Apliquemos este resultado a la normal bivariante con las componentes tipificadas, por aquello de la simplicidad.

Definamos la familia de elipsoides (elipses, claro) mediante

$$h = \phi(z_1 z_2)$$

Su ecuación vendrá dada por

$$(1-\rho^2)c = z_1^2 - 2\rho z_1 z_2 + z_2^2$$

con $c = -2\log\left(2\pi h\sqrt{1-\rho^2}\right)$. Las raíces características de $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ son $\lambda_1 = 1+\rho$, $\lambda_2 = 1-\rho$ y los vectores asociados $\alpha_1' = \left[\frac{1}{2}\sqrt{2}, \frac{1}{2}\sqrt{2}\right]$, $\alpha_2' = \left[\frac{1}{2}\sqrt{2}, -\frac{1}{2}\sqrt{2}\right]$.

Si ρ es positivo el eje principal mayor es la línea $z_1 = z_2$ y el menor la línea $z_1 = -z_2$. En caso contrario se intercambian los ejs. Cuando $\rho = 0$, la elipse es un círculo y existen una infinidad de ejs principales, independientes de los anteriores y de las variables originales. Señalemos por último que en este caso particular en que las varianzas son iguales (σ iguales a 1) los ejs tienen la misma posición, independientemente del valor que tome ρ.

3. DISTRIBUCIONES MARGINALES Y CONDICIONALES DE LAS VARIABLES MULTINORMALES

Comenzaremos demostrando una proposición de utilidad posterior para diferenciar la distribución marginal de cualquier subconjunto de variables del vector $\mathbf{X} = (X_1 \dots X_n)$.

PROPOSICIÓN. Sea $\mathbf{X}$ un vector aleatorio con n componentes con una distribución normal multivariante. Entonces

$$Y = C\mathbf{X}$$

se distribuye como $N(C\mu, C\Sigma C')$ para C no singular.

Demostración.- Recordemos que la densidad de Y se obtiene a partir de la densidad de X, reemplazando x por

$$x = C^{-1}y$$

y multiplicando por el jacobiano de la transformación. Sabíamos que en el caso de una transformación lineal viene dado por $\text{mod}\,|C^{-1}|$. Tendremos que

$$\text{mod}\,|C^{-1}| = \frac{1}{\text{mod}\,|C|} = \sqrt{\frac{1}{|C|^2}} = \sqrt{\frac{|\Sigma|}{|C|\cdot|\Sigma|\cdot|C'|}} = \frac{|\Sigma|^{1/2}}{|C\Sigma C'|^{1/2}}$$

La forma cuadrática del exponente $Q=(x-\mu)'\Sigma^{-1}(x-\mu)$ se transformará en

$$Q = (C^{-1}y-\mu)'\Sigma^{-1}(C^{-1}y-\mu) = (C^{-1}y - C^{-1}C\mu)'\Sigma^{-1}(C^{-1}y - C^{-1}C\mu) =$$

$$= [C^{-1}(y-C\mu)]'\Sigma^{-1}[C^{-1}(y-C\mu)] = (y-C\mu)'(C^{-1})'\Sigma^{-1}C^{-1}(y-C\mu) =$$

$$= (y-C\mu)'(C\Sigma C')^{-1}(y-C\mu)$$

La densidad de Y vendrá entonces dada por

$$f(y) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{1/2}}\cdot\frac{\Sigma^{1/2}}{|C\Sigma C'|^{1/2}}\cdot\exp\left\{-\frac{1}{2}(y-C\mu)'(C\Sigma C')^{-1}(y-C\mu)\right\} =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}|C\Sigma C'|^{1/2}}\cdot\exp\left\{-\frac{1}{2}(y-C\mu)'(C\Sigma C')^{-1}(y-C\mu)\right\}.$$

Lo que demuestra la proposición.

Veamos ahora de estudiar las marginales. Iamos para ello a dividir el conjunto de variables en dos subconjuntos $X_1 \dots X_q$ y $X_{q+1} \dots X_n$ y formaremos los vectores

$$X^{(1)} = \begin{bmatrix} X_1 \\ \vdots \\ X_q \end{bmatrix}, \quad X^{(2)} = \begin{bmatrix} X_{q+1} \\ \vdots \\ X_n \end{bmatrix}$$

de manera que

$$(1) \quad X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

supongamos además que las variables que representan el vector X tienen una distribución conjunta normal multivariante con medias

$$E(X^{(1)}) = \mu^{(1)}, \quad E(X^{(2)}) = \mu^{(2)}$$

y covarianzas

$$E\left((X^{(1)}-\mu^{(1)})(X^{(1)}-\mu^{(1)})'\right) = \Sigma_{11}$$
$$E\left((X^{(2)}-\mu^{(2)})(X^{(2)}-\mu^{(2)})'\right) = \Sigma_{22}$$
$$E\left((X^{(1)}-\mu^{(1)})(X^{(2)}-\mu^{(2)})'\right) = \Sigma_{12} = \Sigma_{21}' = 0.$$

Decimos entonces que el vector X ha sido fraccionado en (1) en dos subvectores, que el vector media ha sido fraccionado análogamente en subvectores y que la matriz de covarianzas

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

ha sido fraccionada, análogamente, en submatrices.

Vamos a demostrar ahora, que en las condiciones anteriormente impuestas, a las submatrices de Σ, las variables $\bar{X}^{(1)}$ y $\bar{X}^{(2)}$ se distribuyen independientes y normalmente.

La inversa de Σ vendrá dada por

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix}$$

en cuanto a la forma cuadrática del exponente, tenemos

$$Q = (x-\mu)' \Sigma^{-1} (x-\mu) = \left[(x^{(1)}-\mu^{(1)})', (x^{(2)}-\mu^{(2)})' \right] \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} (x^{(1)}-\mu^{(1)}) \\ (x^{(2)}-\mu^{(1)}) \end{bmatrix} =$$

$$= \left[(x^{(1)}-\mu^{(1)})' \Sigma_{11}^{-1}, (x^{(2)}-\mu^{(2)})' \Sigma_{22}^{-1} \right] \begin{bmatrix} x^{(1)}-\mu^{(1)} \\ x^{(2)}-\mu^{(2)} \end{bmatrix} =$$

$$= (x^{(1)}-\mu^{(1)})' \Sigma_{11}^{-1} (x^{(1)}-\mu^{(1)}) + (x^{(2)}-\mu^{(2)})' \Sigma_{22}^{-1} (x^{(2)}-\mu^{(2)}) = Q_1 + Q_2$$

por otra parte $|\Sigma| = |\Sigma_{11}| \cdot |\Sigma_{22}|$. Entonces la densidad de $\bar{X}$ podemos escribirla como

$$f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \cdot e^{-\frac{1}{2}Q} = \frac{1}{(2\pi)^{\frac{q}{2}} |\Sigma_{11}|^{1/2}} \cdot e^{-\frac{1}{2}Q_1} \cdot \frac{1}{(2\pi)^{\frac{n-q}{2}} |\Sigma_{22}|^{1/2}} \cdot e^{-\frac{1}{2}Q_2} = f(x^{(1)}) \cdot f(x^{(2)})$$

que es el producto de dos normales multivariantes, a saber, $N(\mu^{(1)}, \Sigma_{11})$, $N(\mu^{(1)}, \Sigma_{22})$.

La marginal de $\bar{X}^{(1)}$ vendrá dada por la integral

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x)\, dx_{q+1} \cdots dx_n = f(x^{(1)}) \cdot \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x^{(2)})\, dx_{q+1} \cdots dx_n = f(x^{(1)})$$

Es decir, la marginal de $\bar{X}^{(1)}$ es una $N(\mu^{(1)}, \Sigma_{11})$. Análogamente obtendríamos $N(\mu^{(2)}, \Sigma_{22})$ para la marginal de $\bar{X}^{(2)}$ y como además $N(\mu, \Sigma)$ puede factorizarse como producto de ambas marginales podemos afirmar que ambas variables son independientes, pudiendo enunciar el siguiente teorema:

TEOREMA - Si $X_1 \ldots X_n$ tienen una densidad conjunta normal multivariante, una condición necesaria y suficiente para que un subconjunto de esta variable y el subconjunto complementario sean independientes es que cada componente de una variable de un conjunto y una variable del otro sea nula.

La necesidad se obtiene de inmediato teniendo en cuenta que

$$\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_i - \mu_i)(x_j - \mu_j) f(x^{(1)}) f(x^{(2)})\, dx^{(1)} \cdot dx^{(2)} =$$

$$= \left\{ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_i - \mu_i) f(x^{(1)})\, dx^{(1)} \right\} \left\{ \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} (x_j - \mu_j) f(x^{(2)})\, dx^{(2)} \right\} = 0.$$

Por tanto $\rho_{ij} = 0$ y vemos que incorrelación e independencia son equivalentes en el caso de normalidad.

Consideremos el caso especial en que se trata de una normal ~~multivariante~~ bivariante. Es decir

$X^{(1)} = X_1$, $X^{(2)} = X_2$, $\mu^{(1)} = \mu_1$, $\mu^{(2)} = \mu_2$, $\Sigma_{11} = \sigma_{11} = \sigma_1^2$, $\Sigma_{22} = \sigma_{22} = \sigma_2^2$, $\Sigma_{12} = \Sigma_{21} = \sigma_{12} = \sigma_1 \sigma_2 \rho_{12}$.

Así, si X_1 y X_2 tienen distribución normal bivariante serán independientes si y sólo si son incorrelados. En este caso la marginal de X_i será normal con media μ_i y varianza σ_i^2. Podemos concluir también el siguiente

COROLARIO.- Si X es un vector normal multivariante $N(\mu, \Sigma)$ y si un conjunto de variables de X es incorrelado con otro conjunto (el complementario), las distribuciones marginales del conjunto son una normal multivariante con medias, varianzas y covarianzas obtenidas tomando las correspondientes componentes de μ y Σ respectivamente.

Pero vamos a ir más allá en nuestro resultado. Vamos a comprobar que este resultado es válido aún cuando los subvectores obtenidos no sean independientes. Vamos para ello la siguiente transformación lineal no singular de los subvectores

$$Y^{(1)} = X^{(1)} + M X^{(2)}$$

$$Y^{(2)} = X^{(2)}$$

eligiendo M de tal manera que las componentes de $Y^{(1)}$ sean incorreladas con las componente de $Y^{(2)} = X^{(2)}$. La matriz M debe satisfacer, para ello, la siguiente ecuación

$$0 = E\left(Y^{(1)} - E(Y^{(1)})\right)\left(Y^{(2)} - E(Y^{(2)})\right)' = E\left(\left(X^{(1)} + M X^{(2)} - E(X^{(1)}) - M E(X^{(2)})\right)\left(X^{(2)} - E(X^{(2)})\right)'\right) =$$

$$= E\left(\left(X^{(1)} - E(X^{(1)})\right) + \left(M X^{(2)} - M E(X^{(2)})\right)\right)\left(X^{(2)} - E(X^{(2)})\right)' = \Sigma_{12} + M \Sigma_{22}$$

y de aquí

$$M = -\Sigma_{12} \Sigma_{22}^{-1}$$

y por tanto $$Y^{(1)} = X^{(1)} - \Sigma_{12} \Sigma_{22}^{-1} X^{(2)} .$$

Es decir

$$Y = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \end{bmatrix} = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix} = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} X$$

Si una transformación no singular de X, tendrá por tanto, de acuerdo con la primera proposición, una distribución normal dada por

$$E\begin{bmatrix} Y^{(1)} \\ Y^{(2)} \end{bmatrix} = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix} = \begin{bmatrix} \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)} \\ \mu^{(2)} \end{bmatrix} = \begin{bmatrix} \nu^{(1)} \\ \nu^{(2)} \end{bmatrix} = \nu$$

y matriz de covarianzas

$$C(Y, Y') = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{12}\Sigma_{22}^{-1} & I \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

Resulta entonces que de acuerdo con el teorema anterior $Y^{(1)}$ e $Y^{(2)}$ son independientes y aplicando el corolario $X^{(2)} = Y^{(2)}$ tiene como marginal una normal $N(\mu^{(2)}, \Sigma_{22})$. Podemos entonces enunciar el siguiente teorema

TEOREMA.- Si X tiene una distribución normal multivariante $N(\mu, \Sigma)$, la marginal de cualquier subconjunto de componentes de X es normal multivariante con medias, varianzas y covarianzas obtenidas tomando las correspondientes componentes de μ y Σ respectivamente.

Consideremos ahora la transformación

$$Z = D\Sigma$$

donde Z tiene q componentes y D es una matriz real $q \times n$. El valor esperado de Z viene dado por

$$E(Z) = D\mu.$$

y su matriz de covarianzas es

$$E\left(Z - D\mu\right)\left(Z - D\mu\right)' = D\Sigma D'$$

El caso $q = n$ y D no singular ya lo estudiamos al comienzo de este párrafo. Si $q \leq n$ y D es de rango q, podemos encontrar una matriz E, $(n-q) \times n$ de tal forma que

$$\begin{pmatrix} Z \\ W \end{pmatrix} = \begin{pmatrix} D \\ E \end{pmatrix} \Sigma$$

sea una transformación no singular. En este caso Z y W tienen distribución conjunta normal multivariante y Z de acuerdo con el teorema anterior tendrá una marginal normal cuyos parámetros ya conocemos. Tenemos en consecuencia el siguiente teorema:

TEOREMA.- Si Σ es un vector normal multivariante $N(\mu, \Sigma)$, entonces $Z = D\Sigma$ se distribuye $N(D\mu, D\Sigma D')$, donde D es una matriz real $q \times n$ de rango $q \leq n$.

Notación es posible llevar a cabo una generalización de este teorema para aquellos casos en que D sea una matriz cualquiera, no necesariamente de rango completo. En cualquier caso no vayamos de ello como un ejercicio.

Distribuciones condicionales

Vamos a tratar de encontrar las distribuciones condicionales derivadas de una normal multivariante.

Sea Σ un vector normal multivariante $N(\mu, \Sigma)$ con Σ no singular. Si llevamos a cabo la partición en dos subvectores

$$\Sigma = \begin{bmatrix} \Sigma^{(1)} \\ \Sigma^{(2)} \end{bmatrix}$$

de q y $n-q$ componentes respectivamente. Recordemos que en el párrafo anterior obteníamos la distribución conjunta de las variables $Y^{(1)} = \Sigma^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma^{(2)}$ e $Y^{(2)} = \Sigma^{(2)}$, cuya expresión era un producto de normales, a saber

$$n\left(y^{(1)} \mid \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)},\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) \cdot n\left(y^{(2)} \mid \mu^{(2)}, \Sigma_{22}\right)$$

La densidad conjunta de $\Sigma^{(1)}$ y $\Sigma^{(2)}$ podemos obtenerla a partir de la anterior expresión, sustituyendo $y^{(1)}$ por $x^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)}$ y $y^{(2)}$ por $x^{(2)}$ teniendo en cuenta que el Jacobiano de la transformación es 1. El resultado de esta sustitución será:

$$f(x^{(1)}, x^{(2)}) = \frac{1}{(2\pi)^{q/2}\left|\Sigma_{11.2}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left[(x^{(1)} - \mu^{(1)}) - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})\right]' \Sigma_{11.2}^{-1}\left[(x^{(1)} - \mu^{(1)}) - \Sigma_{12}\Sigma_{21}^{-1}(x^{(2)} - \mu^{(2)})\right]\right\}$$

$$\cdot \frac{1}{(2\pi)^{\frac{n-q}{2}}\left|\Sigma_{22}\right|^{1/2}} \exp\left\{-\frac{1}{2}(x^{(2)} - \mu^{(2)})'\Sigma_{22}^{-1}(x^{(2)} - \mu^{(2)})\right\}$$

donde $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Expresión ésta que desde luego, dado está, nada de la densidad de Σ, a saber $n(x \mid \mu, \Sigma)$. El interés de la factorización anterior quedará de manifiesto a continuación. En efecto, la densidad condicional de $\Sigma^{(1)}$ dado que $\Sigma^{(2)} = x^{(2)}$ será el cociente de la anterior expresión y la densidad marginal de $\Sigma^{(2)}$, que es precisamente el segundo factor.

Así pues

$$f(x^{(1)}/x^{(2)}) = \frac{1}{(2\pi)^{q/2}\,|\Sigma_{11.2}|^{1/2}} \exp\left\{-\frac{1}{2}\left[(x^{(1)}-\mu^{(1)}) - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})\right]' \Sigma_{11.2}^{-1}\left[(x^{(1)}-\mu^{(1)}) - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})\right]\right\}.$$

claramente, la densidad $f(x^{(1)}/x^{(2)})$ es una densidad normal q-variante con vector media

$$E(X^{(1)}/x^{(2)}) = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)}) = v(x^{(2)})$$

y matriz de covarianzas

$$E\left[\left[X^{(1)}-v(x^{(2)})\right]\left[X^{(1)}-v(x^{(2)})\right]'\right] = \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Hay que destacar en estas expresiones que la media de $X^{(1)}$ dado $x^{(2)}$ es simplemente una función lineal de $x^{(2)}$ mientras que la matriz de covarianzas correspondiente no depende para nada de $x^{(2)}$.

DEFINICION.- La matriz $\Sigma_{12}\Sigma_{22}^{-1}$ es la matriz de los coeficientes de regresión de $X^{(1)}$ sobre $x^{(2)}$.

El i,j-ésimo elemento de $\Sigma_{12}\Sigma_{22}^{-1}$ viene a menudo representado por

$$\beta_{ij\cdot q+1,\dots,n} \qquad \begin{matrix} i=1,\dots,q \\ j=q+1,\dots,n \end{matrix}$$

El vector $\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})$ se le denomina función de regresión.

Si por $\sigma_{ij\cdot q+1,\dots,n}$ designamos el elemento i,j-ésimo de $\Sigma_{11.2}$, llamaremos a esta cantidad el covariante parcial.

DEFINICION.-

$$\rho_{ij\cdot q+1,\dots,n} = \frac{\sigma_{ij\cdot q+1,\dots,n}}{\sqrt{\sigma_{ii\cdot q+1,\dots,n}}\cdot\sqrt{\sigma_{jj\cdot q+1,\dots,n}}} \qquad i,j = 1,\dots,q$$

es el coeficiente de correlación parcial entre X_i y X_j cuando $X_{q+1},\dots,X_n$ permanecen fijas.

DEFINICION.- Al vector aleatorio

$$X_{1.2} = X^{(1)} - \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})$$

se le llama el conjunto de las variables residuales y representa las discrepancias de los elementos de $X^{(1)}$ y de la predicción de sus valores hecha a partir de la relación lineal del vector media de la distribución condicional con las variables de $X^{(2)}$.

Como ya vimos anteriormente

$$\mathrm{Cov}(Y^{(1)}, Y^{(2)}) = 0 \longrightarrow \mathrm{cov}(X_{1.2}, X^{(2)}) = 0$$

hicimos ver que

$$\mathrm{Cov}(Y^{(1)}, Y^{(1)}) = \Sigma_{11.2} \longrightarrow \mathrm{cov}(X_{1.2}, X^{(1)}) = \Sigma_{11.2}$$

Como la numeración de las componentes de X es arbitraria y q también lo es, todo lo anterior sirve para definir la distribución condicional de cualquier conjunto de q componentes de X, fijando las $n-q$ restantes. Análogamente, dado que la marginal de cualquier r componentes de X es normal, podemos también definir la distribución condicional de cualesquiera q componentes ($q \le r$) dado que las restantes $r-q$ permanecen fijas. Podemos enunciar el siguiente teorema.

TEOREMA.- Sea X un vector aleatorio n-dimensional dividido en dos grupos que dan lugar a los subvectores $X^{(1)}$ y $X^{(2)}$. Supongamos que dividimos a la media μ de forma similar $\mu^{(1)}$ y $\mu^{(2)}$ y que la matriz de covarianzas Σ aparece dividida en las submatrices $\Sigma_{11}, \Sigma_{12}, \Sigma_{22}$ que son las matrices de covarianza de $X^{(1)}$, de $X^{(1)}$, $X^{(2)}$ respectivamente. Entonces, si la distribución de X es normal, la distribución condicional de $X^{(1)}$ dado $X^{(2)} = x^{(2)}$ es también normal con media $\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)}-\mu^{(2)})$ y matriz de covarianzas $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}$.

Consideremos, como un ejemplo de todo lo expuesto el caso de una normal bivariante y vamos a encontrar la condicional de X_1 dado $X_2 = x_2$. En este caso $\mu^{(1)} = \mu_1$, $\mu^{(2)} = \mu_2$, $\Sigma_{11} = \sigma_1^2$, $\Sigma_{12} = \rho\sigma_1\sigma_2$, $\Sigma_{22} = \sigma_2^2$. La matriz de los coeficientes de regresión vendrá dada por $\Sigma_{12}\Sigma_{22}^{-1} = \rho\cdot\sigma_1/\sigma_2$ y la matriz de covarianzas parciales será

$$\Sigma_{11\cdot 2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \sigma_1^2 - \rho\sigma_1\sigma_2\cdot\frac{1}{\sigma_2^2}\cdot\rho\sigma_1\sigma^2 = \sigma_1^2 - \rho^2\sigma_1^2 = \sigma_1^2(1-\rho^2)$$

entonces la densidad de X_1, dado $X_2 = x_2$ será

$$f(x_1/x_2) = \frac{1}{\left[2\pi\sigma_1^2(1-\rho^2)\right]^{1/2}}\exp\left\{-\frac{1}{2}\left(x_1 - \mu_1 - \rho\sigma_1/\sigma_2(x_2-\mu_2)\right)^2/\sigma_1^2(1-\rho^2)\right\}$$

Hay que hacer notar que para valores de ρ positivos, la media condicional aumenta a medida que lo hace x_2 y disminuye si ρ es negativo.

Algunas consideraciones geométricas

Una interpretación geométrica de la teoría expuesta puede ayudar a comprenderla mejor. La densidad $f(x_1,x_2)$ puede ser considerada como una superficie $z = f(x_1,x_2)$ sobre el plano x_1x_2. Si intersectamos esta superficie mediante el plano $x_2 = c$ obtendremos una curva $z = f(x_1,c)$ sobre la línea $x_2 = c$ en el plano x_1x_2. Las ordenadas de esta curva son proporcionales a la densidad condicional de X_1, dado $X_2 = x_2$; es decir, son proporcionales a las ordenadas de la curva de una distribución normal univariante. En el caso más general es conveniente considerar elipsoides de densidad constante en el espacio n-dimensional. Las superficies de densidad constante de $f(x_1 - x_q/c_{q+1},\ldots,c_n)$ son la intersección de las superficies de densidad constante de $f(x_1,\ldots,x_n)$ y los hiperplanos $x_{q+1} = c_{q+1},\ldots,x_n = c_n$. Estas son también elipsoides.

Para una mayor claridad de las ideas podemos, todavía, hacer algunas consideraciones acerca de una población real idealizada mediante una distribución normal. Consideremos por ejemplo una población de parejas padre-hijo. Si la población es razonablemente homogénea, las alturas de los padres y las alturas de los hijos tendrán aproximadamente una distribución normal. Una distribución condicional puede obtenerse considerando a los hijos de aquellos padres cuya altura sea, por ejemplo, 175 cms. Las alturas de estos hijos tendrán una distribución aproximadamente normal univariante. La media de esta distribución difiere de la media de las alturas de los hijos cuyos padres tienen una altura de 1.65 ms, por ejemplo, pero las varianzas serán las mismas.

Consideremos también las triplets constituidas por las alturas de un padre y sus dos hijos mayores. La colección de alturas de ambos hijos para padres de 175 cms. de altura es la distribución condicional de dos variables; la correlación entre las alturas de los dos hijos mayores es un coeficiente de correlación parcial. El hecho de mantener constante la altura de los padres elimina el efecto hereditario debido a los mismos. No obstante, cabría esperar una correlación positiva entre ambas alturas puesto que los efectos hereditarios de la madre y los ambientales tenderían a causar en las alturas de los hermanos variaciones similares (de la misma índole).

Como hemos señalado anteriormente, cualq distribución condicional obtenida a partir de una distribución normal es normal con media una función lineal de las variables fijadas y matriz de covarianzas constante. En el caso de distribuciones no normales la distribución condicional de un conjunto de variables sobre otras no tiene generalmente esta propiedad. No obstante, pueden construirse distribuciones no normales tales que algunas distribuciones condicionadas tengan la citada propiedad. Esto puede hacerse tomando como densidad para X el producto $n(x/\mu^{(1)} + \beta(x^{(2)} - \mu^{(2)}), \Sigma_{11\cdot 2})\, f(x^{(2)})$ donde $f(x^{(2)})$ es una densidad cualquiera.

El coficiente de correlación multiple

Consideremos nuevamente X fraccionado en $X^{(1)}$, $X^{(2)}$. Vamos a estudiar algunas propiedades de $\Sigma_{12}\Sigma_{22}^{-1}X^{(2)}$. Como solo estamos interesados, ahora, en funciones de las covarianzas y estas son invariantes por cambio de origen o de escala vamos a suponer $\mu=0$. Elijamos X_i una componente de $X^{(1)}$. Entonces

$$E\left(X_i / x^{(2)}\right) = \beta x^{(2)}$$

donde $\beta = \sigma_{(i)}\Sigma_{22}^{-1}$

siendo $\sigma_{(i)}$ la i-ésima fila de Σ_{12} definida mediante

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Consideramos ahora la función lineal de la, $x^{(2)}$, $\beta x^{(2)}$. Vimos anteriormente que la covariancia entre $X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}X^{(2)}$ y $X^{(2)}$ era nula, entonces las variables $X_i - \beta X^{(2)}$ y $X^{(2)}$ serán independientes.

Trataremos de buscar una función lineal $\alpha X^{(2)}$ para la que $(X_i - \alpha X^{(2)})$ tenga mínima varianza.

Puesto que $E(z^2) = E(zz')$ cuando z es un escalar, la varianza será

$$E\left(X_i - \alpha X^{(2)}\right)^2 = E\left[\left(X_i - \beta X^{(2)}\right) + (\beta - \alpha)X^{(2)}\right]^2 = E\left[\left(X_i - \beta X^{(2)}\right) + (\beta-\alpha)X^{(2)}\right]\left[\left(X_i' - X^{(2)'}\beta'\right) +$$

$$+ X^{(2)'}(\beta-\alpha)'\Big] = E\left[\left(X_i - \beta X^{(2)}\right)\left(X_i' - X^{(2)'}\beta'\right) + (\beta-\alpha)X^{(2)}X^{(2)'}(\beta-\alpha)'\right] =$$

$$= E\left[\left(X_i - \beta X^{(2)}\right)\left(X_i' - X^{(2)'}\beta'\right)\right] + (\beta-\alpha)E\left(X^{(2)}X^{(2)'}\right)(\beta-\alpha)' =$$

$$= E\left(X_i X_i'\right) - \beta E\left(X^{(2)}X_i'\right) - E\left(X_i X^{(2)'}\right)\beta' + \beta E\left(X^{(2)}X^{(2)'}\right)\beta' + (\beta-\alpha)E\left(X^{(2)}X^{(2)'}\right)(\beta-\alpha)' =$$

$$= \sigma_{ii} - \beta\sigma_{(i)}' - \sigma_{(i)}\beta' + \beta\Sigma_{22}\beta' + (\beta-\alpha)\Sigma_{22}(\beta-\alpha)' =$$

$$= \sigma_{ii} - \sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}' - \sigma_i\Sigma_{22}^{-1}\sigma_i' + \sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}' + (\beta-\alpha)\Sigma_{22}(\beta-\alpha)' =$$

$$= \left(\sigma_{ii} - \sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}'\right) + (\beta-\alpha)\Sigma_{22}(\beta-\alpha)'$$

Como Σ_{22} es definida positiva el segundo término de la última línea es no negativo y alcanzará su mínimo cuando sea nulo, es decir $\alpha=\beta$. Así, la función de regresión es la función de $X^{(2)}$ tal que $(X_i - \alpha X^{(2)})$ tiene mínima varianza.

Vamos ahora a demostrar que la máxima correlación entre X_i y $\alpha X^{(2)}$ se obtiene para $\alpha=\beta$. Sabemos que

$$E\left(X_i - \beta X^{(2)}\right)^2 \le E\left(X_i - c\alpha X^{(2)}\right)^2 \quad \forall c, \forall \alpha.$$

Por tanto

$$\sigma_{ii} + E\left(\beta X^{(2)}\right)^2 - 2E\left(X_i \beta X^{(2)}\right) \le \sigma_{ii} + c^2 E\left(\alpha X^{(2)}\right)^2 - 2cE\left(X_i \alpha X^{(2)}\right)$$

De aquí

$$2\frac{E\left(X_i \beta X^{(2)}\right)}{\sqrt{\sigma_{ii}}\sqrt{E\left(\beta X^{(2)}\right)^2}} - \frac{E\left(\beta X^{(2)}\right)^2}{\sqrt{\sigma_{ii}}\sqrt{E\left(\beta X^{(2)}\right)^2}} \ge 2c\frac{E\left(X_i \alpha X^{(2)}\right)}{\sqrt{\sigma_{ii}}\sqrt{E\left(\beta X^{(2)}\right)^2}} - c^2\frac{E\left(\alpha X^{(2)}\right)^2}{\sqrt{\sigma_{ii}}\sqrt{E\left(\beta X^{(2)}\right)^2}}$$

eligiendo $c^2 = \dfrac{E\left(\beta X^{(2)}\right)^2}{E\left(\alpha X^{(2)}\right)^2}$

tendremos

$$\frac{E\left(X_i \beta X^{(2)}\right)}{\sqrt{\sigma_{ii}}\sqrt{E\left(\beta X^{(2)}\right)^2}} \ge \frac{E\left(X_i \alpha X^{(2)}\right)}{\sqrt{\sigma_{ii}}\sqrt{E\left(\alpha X^{(2)}\right)^2}}$$

Podemos ahora enunciar el siguiente teorema:

TEOREMA - Sea X un vector aleatorio normal multivariante $N(\mu, \Sigma)$. Sean $X' = [X^{(1)'}, X^{(2)'}]$, $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ y sea β la i-ésima fila de $\Sigma_{12}\Sigma_{22}^{-1}$, $i = 1, \dots, q$. De todas las combinaciones lineales $\alpha X^{(2)}$, la combinación que minimiza la varianza de $X_i - \alpha X^{(2)}$ y que maximiza la correlación entre X_i y $\alpha X^{(2)}$ es la combinación lineal $\beta X^{(2)}$.

DEFINICIÓN - La máxima correlación entre X_i y la combinación lineal $\alpha X^{(2)}$ se denomina coeficiente de correlación múltiple entre X_i y $X^{(2)}$.

Se sigue que dicho coeficiente es

$$\bar{R}_{i.q+1,\dots,n} = \frac{E(\beta X^{(2)} X_i)}{\sqrt{\sigma_{ii}}\sqrt{E(\beta X^{(2)} X^{(2)'} \beta')}} = \frac{\sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}'}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}'}} = \frac{\sqrt{\sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}'}}{\sqrt{\sigma_{ii}}}$$

Una fórmula útil es la siguiente

$$1 - \bar{R}^2_{i.q+1,\dots,n} = \frac{\sigma_{ii} - \sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}'}{\sigma_{ii}} = \frac{|\Sigma^*|}{\sigma_{ii}|\Sigma_{22}|}$$

donde $\Sigma^* = \begin{bmatrix} \sigma_{ii} & \sigma_{(i)} \\ \sigma_{(i)}' & \Sigma_{22} \end{bmatrix}$ (recordemos que el determinante puede obtenerse en función de los determinantes de las submatrices mediante la expresión $|\Sigma^*| = |\Sigma_{22}| \cdot |\sigma_{ii} - \sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}'|$)

Puesto que

$$\sigma_{ii.q+1,\dots,n} = \sigma_{ii} - \sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}' \quad \text{(elemento de la diagonal de } \Sigma_{11.2})$$

se sigue que

$$1 - \bar{R}^2_{i.q+1,\dots,n} = \frac{\sigma_{ii.q+1,\dots,n}}{\sigma_{ii}} \longrightarrow \sigma_{ii.q+1,\dots,n} = \sigma_{ii}\left(1 - \bar{R}^2_{i,q+1,\dots,n}\right)$$

lo que demuestra que cualquier varianza condicional de un componente de X no puede ser mayor que la varianza. En efecto, cuanto mayor es $\bar{R}_{i.q+1,\dots,n}$, mayor es la reducción en varianza para la distribución condicional

~~Resultado relativo a que mide el coeficiente de correlación para variable y el coeficiente~~

Terminaremos lo referente al coeficiente de correlación múltiple demostrando la siguiente propiedad:

PROPIEDAD - El coeficiente de correlación múltiple es invariante bajo transformaciones no singulares de las variables originales.

Supongamos

$$\begin{aligned} Y_i &= a X_i + b \\ Y^{(2)} &= C X^{(2)} + d \end{aligned} \qquad (1)$$

donde a y b son escalares, C es una matriz no singular y d es un vector de constantes. Entonces el cuadrado del coeficiente de correlación ~~múltiple~~ múltiple de Y_i respecto de $Y^{(2)}$ es

$$\bar{R}^2_{i.q+1,\dots,n} = \frac{\{E(\beta^* Y^{(2)} Y_i)\}^2}{\sigma^*_{ii} \cdot E(\beta^* Y^{(2)} Y^{(2)'} \beta^{*'})} = \frac{\{E(\beta^* C X^{(2)} a X_i)\}^2}{a^2 \sigma_{ii} \cdot E(\beta^* C X^{(2)} (C X^{(2)})' \beta^{*'})} =$$

$$= \frac{\{a\,\sigma_{(i)} C' (C \Sigma_{22} C')^{-1} C \sigma_{(i)}'\}^2}{a^2 \sigma_{ii} \cdot (\sigma_{(i)} C' (C \Sigma_{22} C')^{-1} C \sigma_{(i)}')} = \frac{\sigma_i C' (C \Sigma_{22} C')^{-1} C \sigma_{(i)}'}{\sigma_{ii}} = \frac{\sigma_{(i)}\Sigma_{22}^{-1}\sigma_{(i)}'}{\sigma_{ii}}$$

Esta propiedad implica que la misma correlación puede obtenerse indistintamente partiendo de la matriz de correlaciones o de la matriz de covarianzas. Los coeficientes de regresión de las variables transformadas (1) son

$$\gamma = a\left[\sigma_{(i)} C'\right]\left[C \Sigma_{22} C'\right]^{-1} = a\,\sigma_{(i)} C' C'^{-1} \Sigma_{22}^{-1} C^{-1} = a\left(\sigma_{(i)}\Sigma_{22}^{-1}\right) C^{-1} = a\,\beta\, C^{-1}$$

Si C es una matriz diagonal de factores de escala, por ejemplo los inversos de las desviaciones Standard de las variables en $X^{(2)}$, el efecto de este cambio de escala consiste en el producto de cada elemento de β por a/δ_i. Por ejemplo, la expresión bivariante transformada nos coeficiente en la cantidad adimensional C haciendo $a = 1/\delta_1$, $c = 1/\delta_2$.

Finalmente, si Σ y Σ_{22} tienen igual rango, q, esto significa que X_1 puede ser expresada exactamente como una combinación lineal de las q variables que componen $X^{(2)}$. En este caso la correlación multiple es exactamente la unidad y β es el vector de los coeficientes de la combinación lineal. Si Σ_{22} tiene un rango menor que q, la expresión del coeficiente de correlación multiple es indeterminada y la correlación multiple puede ser redefinida como la máxima correlación de X_1 con una combinación lineal de un subconjunto de $X^{(2)}$ cuyo rango coincida con el de Σ_{22}.

4. FUNCION CARACTERISTICA. MOMENTOS

Definición.- La función característica de un vector aleatorio X es

$$\phi(t) = E\left(e^{it'X}\right) \quad \text{para cualquier vector real } t.$$

para darle un significado a esta definición definiremos la esperanza de una función vectorial aleatoria a valores complejos, mediante

Definición.- Sea $g(x)$ una función del vector aleatorio X con valores complejos, que podemos escribir de la forma $g(x) = g_1(x) + i g_2(x)$, donde $g_1(x)$ y $g_2(x)$ son funciones reales. Entonces la esperanza de $g(X)$ viene dada por

$$E(g(X)) = E(g_1(X)) + i E(g_2(X))$$

En particular

$$E\left(e^{it'X}\right) = E(\cos t'X) + i E(\operatorname{sen} t'X).$$

No es nuestro objetivo entrar en detalles acerca de las propiedades de la función característica, pero sin embargo algunos teoremas de gran importancia y utilidad.

Recordemos en primer lugar que el teorema de factorización, cuando se trata de variables independientes, también es verificado por las funciones características, de manera que si $X' = \left(X^{(1)'}, X^{(2)'}\right)$ siendo $X^{(1)}$, $X^{(2)}$ independientes entonces

$$\phi(t) = E\left(e^{it'X}\right) = E\left(e^{it^{(1)'}X^{(1)}}\right) \cdot E\left(e^{it^{(2)'}X^{(2)}}\right) = \phi_1(t^{(1)}) \cdot \phi_2(t^{(2)}) \quad \text{con } t' = \left(t^{(1)'}, t^{(2)'}\right).$$

Así, si X está constituido por n variables independientes tendremos

$$\phi(t) = E\left(e^{it'X}\right) = \prod_{j=1}^{n} E\left(e^{it_j X_j}\right) = \prod_{j=1}^{n} \phi_j(t_j).$$

Esta propiedad nos va a permitir obtener la función característica para un vector X normal multivariante.

TEOREMA.- Sea X un vector aleatorio $N(\mu, \Sigma)$, entonces su función característica viene dada por

$$\phi(t) = E\left(e^{it'X}\right) = e^{it'\mu - \frac{1}{2} t'\Sigma t}, \quad \forall t, \text{ vector real.}$$

Demostración.- Sabemos de la existencia de una matriz C no singular, tal que

$$C'\Sigma^{-1}C = I$$

y de aquí

$$\Sigma^{-1} = C'^{-1} \cdot C^{-1} = (CC')^{-1}$$

Hagamos

$$X - \mu = CY$$

entonces Y se distribuirá $N(0, I)$, como ya sabemos. La función característica de Y, viene dada por

$$\psi(u) = E\left(e^{iu'Y}\right) = \prod_{j=1}^{n} E\left(e^{iu_j Y_j}\right)$$

y como Y_j es $N(0,1)$, tendremos

$$\psi(u) = \prod_{j=1}^{n} E\left(e^{iu_j Y_j}\right) = \prod_{j=1}^{n} e^{-\frac{1}{2}u_j^2} = e^{-\frac{1}{2}u'u}.$$

A partir de aquí

$$\phi(t) = E\left(e^{it'X}\right) = E\left(e^{it'(CY+\mu)}\right) = e^{it'\mu} E\left(e^{it'CY}\right) = e^{it'\mu} e^{-\frac{1}{2}(t'C)(t'C)'}$$

para $t'C = u'$. Esta cadena de igualdades se obtiene fácilmente escribiendo las esperanzas como integrales y teniendo en cuenta las reglas que rigen los cambios de variable en una integral. ~~(cadena de igualdades que se obtiene fácilmente escribiendo las esperanzas como integrales y teniendo en cuenta las reglas que rigen los cambios de variable en una integral)~~. Tendremos finalmente

$$\phi(t) = e^{it'\mu - \frac{1}{2}t'CC't} = e^{it'\mu - \frac{1}{2}t'\Sigma t}.$$

Con el auxilio de los dos teoremas que enunciamos a continuación y del anterior resultado podrían haberse obtenido fácilmente más rápidamente algunos de los anteriores resultados.

TEOREMA (de Levy).- Si la variable X tiene una densidad $f(x)$ y una función característica $\phi(t)$, entonces

$$f(x) = \frac{1}{(2\pi)^n} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} e^{-it'x} \phi(t)\, dt_1 \cdots dt_n.$$

TEOREMA (de la continuidad de Levy-Cramer).- Sea $\{F_j(x)\}$ una sucesión de funciones de distribución y sea $\{\phi_j(t)\}$ la correspondiente sucesión de funciones características. Una condición necesaria y suficiente para que $F_j(x)$ converja a una distribución $F(x)$ es que, para cualquier t, $\phi_j(t)$ converja a un límite $\phi(t)$, continua en $t=0$. Entonces el límite $\phi(t)$ es además la función característica de la distribución límite $F(x)$.

Obsérvese que el primero de estos teoremas lo que afirma es que las funciones de densidad y características de un vector aleatorio están en correspondencia biunívoca. Entonces:

Si $Z = DX$, con X $N(\mu, \Sigma)$, la función característica de Z tendrá por expresión

$$\phi(t) = E\left(e^{it'Z}\right) = E\left(e^{it'DX}\right) = E\left(e^{i(D't)'X}\right) = e^{i(D't)'\mu - \frac{1}{2}(D't)'\Sigma(D't)} =$$

$$= e^{it'(D\mu) - \frac{1}{2}t'(D\Sigma D')t}$$

aplicando el teorema de Levy llegamos a la conclusión de que Z es normal con media $D\mu$ y matriz de covarianzas $D\Sigma D'$.

Análogamente el uso de las funciones características y sus propiedades nos permite demostrar el siguiente

TEOREMA.- Si cualquier combinación lineal de las componentes de un vector Y se distribuye normalmente, entonces Y se distribuye normalmente.

Demostración.- Supongamos la combinación lineal $Z = u'Y$ que se distribuye normalmente, tendremos que su función característica vendrá dada por

$$E\left(e^{itZ}\right) = E\left(e^{itu'Y}\right) = e^{itu'\mu - \frac{1}{2}t^2 u'\Sigma u}$$

siendo μ y Σ la media y matriz de covarianzas de Y. Si hacemos ahora $t=1$, tendremos que

$$E\left(e^{iu'Y}\right) = e^{iu'\mu - \frac{1}{2}u'\Sigma u}, \text{ por tanto } Y \text{ se distribuirá normalmente.}$$

Es muy importante insistir en el hecho de que hemos de ser todas las combinaciones lineales de las componentes de $\underline{X}$. Puede verse un contraejemplo en Anderson pags. 37 y 38.

Momentos de un vector aleatorio normal

Los momentos de $X_1, \dots, X_n$ con distribución conjunta normal pueden obtenerse a partir de la función característica. A saber:

$$E(X_j) = \frac{1}{i} \frac{\partial \phi}{\partial t_j}\Big|_{t=0} = \frac{1}{i}\left(-\sum_k \sigma_{jk} t_k + i\mu_j\right)\phi(t)\Big|_{t=0} = \mu_j$$

$$E(X_h X_j) = \frac{1}{i^2}\frac{\partial^2 \phi}{\partial t_h \partial t_j}\Big|_{t=0} = \frac{1}{i^2}\left\{\left(-\sum_k \sigma_{hk} t_k + i\mu_h\right)\left(-\sum_k \sigma_{kj} t_k + i\mu_j\right) - \sigma_{hj}\right\}\phi(t)\Big|_{t=0} =$$

$$= \sigma_{hj} - \mu_h \mu_j$$

y así inmediatamente.

5. MUESTRAS DE UNA NORMAL MULTIVARIANTE

En todo cuanto precede hemos estudiado las propiedades de una distribución normal multivariante como si los valores de sus parámetros fueran conocidos y estuvieran bajo control. Esta situación rara vez se cumple. Desgraciadamente en biología y ciencias del comportamiento no ocurre así generalmente y lo que pretendemos en este apartado es considerar métodos que permitan estimar los diferentes parámetros de la normal multivariante a partir de muestras relativamente pequeñas y estudiar, después, las propiedades muestrales de estas estimaciones.

Antes de entrar en detalles acerca de la obtención de las estimaciones insistir en la imperiosa necesidad de que la aleatorización de la muestra considerada se cumpla si queremos que nuestras observaciones tengan alguna validez. Es decir, se trata de que las unidades muestrales hayan sido tomadas independientemente unas de otras de una población homogénea. Estas unidades muestrales no pueden tener características comunes o rasgos que pudieran inducir alguna independencia entre ellas. Por ejemplo, una investigación de los niveles medios de cuatro componentes bioquímicos en el cerebro de cierta raza de ratas no puede estar basada en muestras extraídas de unas pocas camadas de ratas por grandes que éstas sean, sino en muestras extraídas aleatoriamente de ratas de la citada especie. De la misma manera, obtendremos estimaciones sesgadas y poco fiables del efecto general de una droga tranquilizante si la muestra estuviera basada en las respuestas diarias de un pequeño número de pacientes psiquiátricos estudiados durante varias semanas. De la misma forma que elementos de una misma camada de ratas tienen probablemente rasgos biológicos y genéticos comunes, parece razonable pensar que cierta condición afectiva del paciente tiene rasgos variando tan lentamente que harán las contribuciones de días sucesivos estadísticamente dependientes. Además, variaciones importantes en el comportamiento de una persona en la sala del hospital donde se encuentran los enfermos examinados inducirán probablemente cambios en los otros sujetos del estudio.

Estimaciones máximo-verosímiles del vector media y de la matriz de covarianzas

Trataremos de obtener los estimadores máximo-verosímiles de μ y Σ a partir de una muestra de N observaciones procedentes de una población normal p-variante cuyos parámetros son precisamente μ y Σ.

La elección, precisamente, de este tipo de estimadores sometidos a que sean solos o algunas simples transformaciones de los mismos poseen, normalmente, algunas propiedades óptimas de los estimadores. En el caso particular que ahora nos ocupa, los estimadores son asintóticamente eficientes.

Supongamos que nuestra muestra de tamaño N está constituida por los vectores p-variante $X_1, \dots, X_N$, $N > p$. La función de verosimilitud de la muestra viene dada por

$$L = \frac{1}{(2\pi)^{\frac{1}{2}pN}\,|\Sigma|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha-\mu)'\Sigma^{-1}(x_\alpha-\mu)\right]$$

Como el exponente aparece en términos de Σ^{-1} obtendremos primero los estimadores de μ y $\Sigma^{-1}=\Psi$. Recordemos que en la función L, las variables son μ y Σ, mientras que x_α son datos muestrales y por tanto perfectamente conocidos.

Para maximizar L, lo haremos a través de su log, ya que el log es una función creciente y el máximo coincidirá. Así pues

$$\log L = -\frac{1}{2}pN\log(2\pi) + \frac{1}{2}\log|\Psi| - \frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha-\mu)'\Psi(x_\alpha-\mu).$$

Definamos la media muestral como

$$\bar{x} = \frac{1}{N}\sum_{\alpha=1}^{N}x_\alpha = \begin{bmatrix} \frac{1}{N}\sum_{\alpha=1}^{N}x_{1\alpha} \\ | \\ \frac{1}{N}\sum_{\alpha=1}^{N}x_{p\alpha} \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ | \\ \bar{x}_p \end{bmatrix}$$

y la matriz de las sumas de cuadrados y productos cruzados de las desviaciones respecto a la media

$$A = \sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})' = \left[\sum_{\alpha=1}^{N}(x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j)\right] \quad i,j=1,\dots,p$$

Iremos a tratar de dar otra expresión a $\log L$, para ello necesitaremos algunas propiedades que exponemos a continuación.

<u>Lema</u>.- Sean $x_1,\dots,x_N$ N vectores y sea $\bar{x}$ la media antes definida. Entonces para cualquier vector b

$$\sum_{\alpha=1}^{N}(x_\alpha-b)(x_\alpha-b)' = \sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})' + N(\bar{x}-b)(\bar{x}-b)'$$

Demostración.-

$$\begin{aligned}\sum_{\alpha=1}^{N}(x_\alpha-b)(x_\alpha-b)' &= \sum_{\alpha=1}^{N}\left[(x_\alpha-\bar{x})+(\bar{x}-b)\right]\left[(x_\alpha-\bar{x})+(\bar{x}-b)\right]' = \\ &= \sum_{\alpha=1}^{N}\left[(x_\alpha-\bar{x})(x_\alpha-\bar{x})' + (x_\alpha-\bar{x})(\bar{x}-b)' + (\bar{x}-b)(x_\alpha-\bar{x})' + (\bar{x}-b)(\bar{x}-b)'\right] = \\ &= \sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})' + \left[\sum_{\alpha=1}^{N}(x_\alpha-\bar{x})\right](\bar{x}-b)' + (\bar{x}-b)\left[\sum_{\alpha=1}^{N}(x_\alpha-\bar{x})'\right] + N(\bar{x}-b)(\bar{x}-b)' = \\ &= \sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})' + N(\bar{x}-b)(\bar{x}-b)'\end{aligned}$$

pues $\sum_{\alpha=1}^{N}(x_\alpha-\bar{x}) = N\bar{x}-N\bar{x} = 0$.

Si hacemos $b=\mu$, tendremos

$$\sum_{\alpha=1}^{N}(x_\alpha-\mu)(x_\alpha-\mu)' = \sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x}) + N(\bar{x}-\mu)(\bar{x}-\mu)' = A + N(\bar{x}-\mu)(\bar{x}-\mu)'.$$

Por otra parte haciendo uso de la propiedad de la traza de una matriz que dice: $\mathrm{tr}(CD) = \mathrm{tr}(DC) = \sum c_{ij}d_{ji}$, tenemos

$$\begin{aligned}\sum_{\alpha=1}^{N}(x_\alpha-\mu)'\Psi(x_\alpha-\mu) &= \mathrm{tr}\sum_{\alpha=1}^{N}(x_\alpha-\mu)'\Psi(x_\alpha-\mu) = \mathrm{tr}\sum_{\alpha=1}^{N}\Psi(x_\alpha-\mu)(x_\alpha-\mu)' = \mathrm{tr}\,\Psi\left[A+N(\bar{x}-\mu)(\bar{x}-\mu)'\right] = \\ &= \mathrm{tr}\,\Psi A + \mathrm{tr}\,\Psi N(\bar{x}-\mu)(\bar{x}-\mu)' = \mathrm{tr}\,\Psi A + N(\bar{x}-\mu)'\Psi(\bar{x}-\mu)\end{aligned}$$

Podemos entonces escribir el $\log L$ de la siguiente manera

$$\log L = -\frac{1}{2}pN\log(2\pi) + \frac{1}{2}N\log|\Psi| - \frac{1}{2}\operatorname{tr}\Psi A - \frac{1}{2}N(\bar{x}-\mu)'\Psi(\bar{x}-\mu)$$

la matriz Ψ es definida positiva, por tanto $N(\bar{x}-\mu)'\Psi(\bar{x}-\mu) \geq 0$, $\forall(\bar{x}-\mu) \neq 0$, es decir que será nula cuando $\bar{x}-\mu=0$, o sea $\mu=\bar{x}$. Para maximizar el segundo y tercer término de la expresión anterior tendremos presente el siguiente resultado

LEMA.- Sea

$$f(C) = \frac{1}{2}N\log|C| - \frac{1}{2}\sum_{ij=1}^{p} c_{ij}d_{ji}$$

donde $C=[c_{ij}]$ es definida positiva y donde $D=[d_{ij}]$ es definida positiva. Entonces el maximo de $f(C)$ se alcanza para $C=ND^{-1}$ y este máximo es

$$f(ND^{-1}) = \frac{1}{2}pN\log N - \frac{1}{2}N\log|D| - \frac{1}{2}pN$$

Demostración.- Obsérvese que $f(C)$ tiende a $-\infty$ cuando C se aproxima a una matriz singular. Igualmente se demuestra (véase Anderson pag. 47) que ocurre lo mismo cuando algún o algunos elementos de C se aproximan a ∞ y/o $-\infty$. Los máximos de $f(C)$ vendrán pues dados igualando a cero las derivadas respecto a los elementos de C. Tenemos

$$\frac{\partial f}{\partial c_{kk}} = \frac{1}{2}\frac{N}{|C|}\frac{\partial|C|}{\partial c_{kk}} - \frac{1}{2}d_{kk} = \frac{1}{2}N\frac{C_{kk}}{|C|} - \frac{1}{2}d_{kk}$$

donde C_{kk} es el cofactor de c_{kk} en C. Para $k\neq l$

$$\frac{\partial f}{\partial c_{kl}} = \frac{1}{2}\frac{N}{|C|}\frac{\partial|C|}{\partial c_{kl}} - \frac{1}{2}(d_{kl}+d_{lk})$$

y teniendo en cuenta la simetría de C y de D, tenemos

$$\frac{\partial f}{\partial c_{kl}} = N\frac{C_{kl}}{|C|} - d_{kl}$$

donde C_{kl} es el cofactor de c_{kl} en C. Igualando a cero estas parciales y teniendo en cuenta que $C_{kl}/|C|$ es el kl-ésimo elemento de C^{-1}, tenemos $NC^{-1}=D$, y de aquí $C=ND^{-1}$. El correspondiente valor máximo para $f(ND^{-1})$ será

$$f(ND^{-1}) = \frac{1}{2}N\log|ND^{-1}| - \frac{1}{2}\operatorname{tr}ND^{-1}D = \frac{1}{2}N\log N^p|D^{-1}| - \frac{1}{2}\operatorname{tr}NI =$$

$$= \frac{1}{2}pN\log N - \frac{1}{2}N\log|D| - \frac{1}{2}Np$$

Aplicando este lema a la función $\log L$ podremos maximizarla para $\Psi = NA^{-1} = \left(\frac{1}{N}A\right)^{-1}$. Supondremos que A no es singular para que tenga sentido lo escrito (más tarde veremos que esto ocurre con probabilidad 1). En resumen los estimadores máximo verosímiles de μ y Ψ son $\hat{\mu}=\bar{x}$ y $\hat{\Psi}=NA^{-1}$. Para encontrar ahora el estimador de Σ necesitamos el siguiente lema y su corolario.

LEMA.- Sea $f(\theta)$ una función real definida en cierto conjunto S y sea ϕ una función real a valores simples, con una inversa de iguales características, de S a algún otro conjunto S^*; es decir, para cada $\theta\in S$ $\exists$ un único $\theta^*\in S^*$ e inversamente para cada $\theta^*\in S^*$. Sea

$$g(\theta^*) = f[\phi^{-1}(\theta^*)].$$

Entonces si $f(\theta)$ alcanza un máximo en $\theta=\theta_0$, $g(\theta^*)$ alcanza un máximo en $\theta^*=\theta_0^*=\phi(\theta_0)$. Si el maximo de $f(\theta)$ es único en θ_0, igual ocurre con el máximo de $g(\theta^*)$ en θ_0^*.

Demostración :- Por hipótesis

$$f(\theta_0) \geq f(\theta) \quad , \forall \theta \in S$$

Entonces para $\forall \theta^* \in S^*$, tendremos

$$g(\theta^*) = f[\phi^{-1}(\theta^*)] = f(\theta) \leq f(\theta_0) = g[\phi(\theta_0)] = g(\theta_0^*)$$

Resulta que $g(\theta^*)$ alcanza un máximo en θ_0^*. Si el máximo de $f(\theta)$ en θ_0 es único, la desigualdad anterior es estricta para $\theta \neq \theta_0$ y por tanto el máximo para $g(\theta^*)$ es único.

Tenemos el siguiente corolario.

COROLARIO .- Si sobre la base de una muestra dada $\hat{\theta}_1, \dots, \hat{\theta}_m$ son los estimadores máximo verosímiles de los parámetros $\theta_1, \dots, \theta_m$ de una distribución, entonces $\phi_1(\hat{\theta}_1, \dots, \hat{\theta}_m), \dots, \phi_m(\hat{\theta}_1, \dots, \hat{\theta}_m)$ son los estimadores máximo verosímiles de $\phi_1(\theta_1, \dots, \theta_m), \dots, \phi_m(\theta_1, \dots, \theta_m)$ si la transformación de $\theta_1, \dots, \theta_m$ a $\phi_1, \dots, \phi_m$ es uno a uno. Si los estimadores de $\theta_1, \dots, \theta_m$ son únicos, lo son también los de $\phi_1, \dots, \phi_m$.

Aplicando el corolario a nuestra situación podemos afirmar que el estimador máximo verosímil de Σ viene dado por $\hat{\Sigma} = \phi^{-1} = (1/N) A$. Resumiendo:

TEOREMA .- Si $X_1, \dots, X_N$ constituye una muestra de una $N(\mu, \Sigma)$ $(p < N)$, los estimadores máximo verosímiles de μ y Σ son $\hat{\mu} = \bar{x} = (1/N) \sum_\alpha x_\alpha$ y $\hat{\Sigma} = (1/N) \sum_\alpha (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$.

COROLARIO .- Si $X_1, \dots, X_N$ constituye una muestra de una $N(\mu, \Sigma)$, donde $\sigma_{ij} = \sigma_i \sigma_j \rho_{ij}$ $(\rho_{ii} = 1)$, el estimador máximo verosímil de μ es $\bar{x} = (1/N) \sum_\alpha x_\alpha$, la estimación máximo-verosímil de $\sigma_i^2 = \hat{\sigma}_i^2 = (1/N) \sum_\alpha (x_{i\alpha} - \bar{x}_i)^2 =$
$= (1/N) \left(\sum_\alpha x_{i\alpha}^2 - N \bar{x}_i^2 \right)$, donde $x_{i\alpha}$ es la i-ésima componente de x_α y $\bar{x}_i$ es la i-ésima componente de $\bar{x}$,) el estimador máximo-verosímil de ρ_{ij} es

$$\hat{\rho}_{ij} = \frac{\sum_\alpha (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_\alpha (x_{i\alpha} - \bar{x}_i)^2 \cdot \sum_\alpha (x_{j\alpha} - \bar{x}_j)^2}}$$

Demostración .- El conjunto de parámetros $\mu_i = \mu_i$, $\sigma_i^2 = \sigma_{ii}$ y $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ es una transformación uno a uno del conjunto de parámetros μ_i, σ_{ij}. Aplicando el corolario anterior las estimaciones máximo-verosímiles serán las correspondientes transformadas.

Vamos a finalizar este apartado de los estimadores máximo verosímiles de μ y Σ dando una interpretación geométrica de algunos de los elementos que aparecen en estas estimaciones.

Una conveniente interpretación de la muestra desde un punto de vista geométrico, viene en términos de las filas de la matriz de datos $\mathbb{X} = (X_1, \dots, X_N)$, a saber

$$\mathbb{X} = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \cdots & \cdots & \cdots \\ x_{p1} & \cdots & x_{pN} \end{bmatrix} = \begin{bmatrix} y_1' \\ \vdots \\ y_p' \end{bmatrix}$$

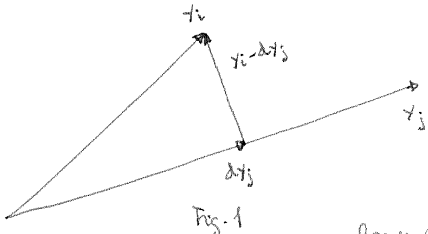


Fig. 1

El vector y_i puede ser considerado como un elemento en el espacio euclídeo N-dimensional. Más concretamente como el extremo de un vector en R^N, cuyo origen coincide con el origen de coordenadas.

Comprobemos en primer lugar que el coseno del ángulo entre y_i e y_j es $y_i y_j' / \sqrt{y_i y_i' y_j y_j'}$

En efecto, elijamos un escalar d de tal forma que dy_j sea ortogonal a $y_i - dy_j$. Entonces $dy_j (y_i - dy_j)' = 0$ y de aquí $d(y_j y_i' - dy_j y_j') = 0 \rightarrow d = y_j y_i' / y_j y_j'$. Si descomponemos ahora y_i en dy_j e $y_i - dy_j$ (ver fig. 1) tendremos que el valor absoluto del coseno del ángulo que forman y_i e y_j vendrá dado por el cociente entre la longitud de dy_j y la longitud de y_i, es decir

$$\sqrt{\frac{(dy_j)(dy_j)'}{y_i y_i'}} = \sqrt{\frac{d\, y_j y_j' d}{y_i y_i'}} = \sqrt{\frac{(y_i y_j')^2 (y_j y_j')(y_i y_i')}{(y_j y_j')(y_i y_i')(y_j y_j')}} = \frac{y_i y_j'}{\sqrt{(y_i y_i')(y_j y_j')}}$$

Podemos ahora dar una interpretación geométrica a a_{ii} y $a_{ij}/\sqrt{a_{ii}a_{jj}}$, donde a_{ij} son los elementos de la matriz A. Introducimos la recta equiangular que es una línea cuya dirección es paralela a la del vector $(1,1,\dots,1)$ y pasa por el origen. La proyección del vector y_i sobre la dirección que $\varepsilon=(1,1,\dots,1)$ determina viene dada por

$$\{(\varepsilon y_i')/(\varepsilon\varepsilon')\}\varepsilon = \left(\sum_\alpha x_{i\alpha}/\sum_\alpha 1\right) = \bar{x}_i\varepsilon = (\bar{x}_i,\dots,\bar{x}_i)$$

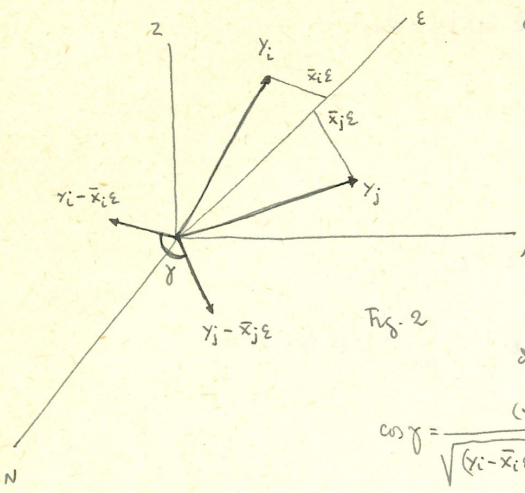


Fig. 2

Podemos ahora descomponer y_i en dos vectores, $\bar{x}_i\varepsilon$ en la dirección de ε y otro $y_i-\bar{x}_i\varepsilon$ proyección de y_i sobre el plano perpendicular a ε por el extremo de $\bar{x}_i\varepsilon$. El cuadrado de la longitud de $y_i-\bar{x}_i\varepsilon$ es $(y_i-\bar{x}_i\varepsilon)(y_i-\bar{x}_i\varepsilon)' = \sum_\alpha (x_{i\alpha}-\bar{x}_i)^2 = N\hat{\sigma}_{ii} = a_{ii}$. Si trasladamos ahora (fig. 2) los vectores $y_i-\bar{x}_i\varepsilon$ y $y_j-\bar{x}_j\varepsilon$ de manera que su origen coincida con el origen de coordenadas el ángulo entre ambos vendrá determinado por

$$\cos\gamma = \frac{(y_i-\bar{x}_i\varepsilon)(y_j-\bar{x}_j\varepsilon)'}{\sqrt{(y_i-\bar{x}_i\varepsilon)(y_i-\bar{x}_i\varepsilon)'\,(y_j-\bar{x}_j\varepsilon)(y_j-\bar{x}_j\varepsilon)'}} = \frac{\sum_\alpha (x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j)}{\sqrt{\sum_\alpha (x_{i\alpha}-\bar{x})^2 \cdot \sum_\alpha (x_{j\alpha}-\bar{x}_j)^2}} = r_{ij}$$

6.- DISTRIBUCION DEL VECTOR MEDIA MUESTRAL. INFERENCIAS ACERCA DE μ CUANDO Σ ES CONOCIDA

La distribución conjunta de μ y Σ.

Recordemos que en el caso univariante la media muestral y la varianza muestral se distribuyen independientemente una de otra y además, la media, en particular, se distribuye normalmente, cuando la población de origen es normal. El resultado similar puede demostrarse para el caso multivariante, previamente a ello hacemos al conjunto de las observaciones una transformación.

TEOREMA.- Supongamos que $X_1,\dots,X_N$ son independientes, con X_α distribuyéndose $N(\mu_\alpha,\Sigma)$. Sea $C=(C_{\alpha\beta})$ una matriz ortogonal. Entonces $Y_\alpha = \sum_{\beta=1}^N C_{\alpha\beta}X_\beta$ se distribuye según una $N(\nu_\alpha,\Sigma)$, con $\nu_\alpha = \sum_{\beta=1}^N C_{\alpha\beta}\mu_\beta$ y $Y_1,\dots,Y_N$ son independientes.

Demostración.- El conjunto de los vectores $\{Y_\alpha\}$ tiene una distribución conjunta normal puesto que el conjunto de todas sus componentes es un conjunto de combinaciones lineales de las componentes de $\{X_\alpha\}$ que tienen una distribución conjunta normal. En efecto, sabemos que

$$f(x_1,x_2,\dots,x_N) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}}\exp\left\{-\frac{1}{2}\sum_\alpha (x_\alpha-\mu_\alpha)'\Sigma^{-1}(x_\alpha-\mu_\alpha)\right\}$$

que puede escribirse de la forma

$$f(x_{11}x_{12}-x_{1p},x_{21}-x_{2p},\dots,x_{N1}-x_{Np}) = \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}}\exp\left\{-\frac{1}{2}\left((x_1-\mu_1)',\dots,(x_N-\mu_N)'\right)\Sigma^{*-1}\left((x_1-\mu_1),\dots,(x_N-\mu_N)\right)\right\}$$

donde Σ^* es una matriz de submatrices de la forma

$$\Sigma^* = \begin{bmatrix}\Sigma & 0 & \cdots & 0\\ 0 & \Sigma & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & \Sigma\end{bmatrix} \quad \text{de dimensión } Np\times Np$$

y cuyo determinante vale $|\Sigma^*| = |\Sigma|^N$.

El paso de los vectores X_α a los Y_α puede llevarse a cabo mediante una transformación del tipo

$$\begin{bmatrix}Y_{11}\\ Y_{21}\\ \vdots\\ Y_{N1}\\ \vdots\\ Y_{1p}\\ \vdots\\ Y_{Np}\end{bmatrix} = C^*\begin{bmatrix}X_{11}\\ X_{21}\\ \vdots\\ X_{N1}\\ \vdots\\ X_{1p}\\ \vdots\\ X_{Np}\end{bmatrix} \quad \text{con} \quad C^* = \begin{bmatrix}C & 0 & \cdots & 0\\ 0 & C & \cdots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \cdots & C\end{bmatrix}$$

La esperanza de Y_α es

$$E(Y_\alpha) = E\left(\sum_\beta c_{\alpha\beta} X_\beta\right) = \sum_\beta c_{\alpha\beta} E(X_\beta) = \sum_\beta c_{\alpha\beta} \mu_\beta = \nu_\alpha .$$

La matriz de covarianzas entre Y_α e Y_γ viene dada por

$$\begin{aligned} \mathrm{Cov}(Y_\alpha, Y_\gamma') &= E\left[(Y_\alpha - \nu_\alpha)(Y_\gamma - \nu_\gamma)'\right] = \\ &= E\left[\left(\sum_\beta c_{\alpha\beta}(X_\beta - \mu_\beta)\right)\left(\sum_\beta c_{\gamma\beta}(X_\xi - \mu_\xi)\right)'\right] = \\ &= \sum_{\beta,\xi} c_{\alpha\beta} c_{\gamma\xi} E\left[(X_\beta - \mu_\beta)(X_\xi - \mu_\xi)'\right] = \sum_{\beta,\xi} c_{\alpha\beta} c_{\gamma\xi} \delta_{\beta\xi} \Sigma = \\ &= \sum_\beta c_{\alpha\beta} c_{\gamma\beta} \Sigma = \delta_{\alpha\gamma} \Sigma \end{aligned}$$

donde $\delta_{\alpha\gamma}$ es la delta de Kronecker ($=1$ si $\alpha=\gamma$ e $=0$ si $\alpha\neq\gamma$). Este resultado muestra que Y_α es independiente de Y_γ y que Y_α tiene matriz de covarianzas Σ.

LEMA.- Si $C = [c_{\alpha\beta}]$ es ortogonal, entonces $\sum_{\alpha=1}^N X_\alpha X_\alpha' = \sum_{\alpha=1}^N Y_\alpha Y_\alpha'$, donde $Y_\alpha = \sum_\beta c_{\alpha\beta} X_\beta$.

Demostración.-

$$\sum_\alpha Y_\alpha Y_\alpha' = \sum_\alpha \left(\sum_\beta c_{\alpha\beta} X_\beta\right)\left(\sum_\gamma c_{\alpha\gamma} X_\gamma'\right) = \sum_{\beta,\gamma}\left(\sum_\alpha c_{\alpha\beta} c_{\alpha\gamma}\right) X_\beta X_\gamma' = \sum_{\beta,\gamma} \delta_{\beta\gamma} X_\beta X_\gamma' = \sum_\beta X_\beta X_\beta' .$$

Sean ahora $X_1, \ldots, X_N$ independientes, cada uno de ellos distribuido de acuerdo con una $N(\mu, \Sigma)$. Entonces existe una matriz ortogonal $B = [b_{\alpha\beta}]$ cuya última fila es

$$\left(1/\sqrt{N}, \ldots, 1/\sqrt{N}\right).$$

Esta transformación es una rotación en el espacio N dimensional en la que la línea equiangular descrita anteriormente se hace coincidir con el N-ésimo eje de coordenadas. Sea A la matriz de la suma de los cuadrados de las desviaciones respecto de la media muestral y los productos cruzados de las mismas, es decir $A = N\hat{\Sigma}$, y sea

$$Z_\alpha = \sum_\beta b_{\alpha\beta} X_\beta .$$

Entonces

$$Z_N = \sum_\beta b_{N\beta} X_\beta = \sum_\beta \frac{1}{\sqrt{N}} X_\beta = \sqrt{N}\,\bar{X}$$

Por el lema anterior tendremos:

$$A = \sum_{\alpha=1}^N X_\alpha X_\alpha' - N\bar{X}\bar{X}' = \sum_{\alpha=1}^N Z_\alpha Z_\alpha' - Z_N Z_N' = \sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$$

Puesto que Z_N es independiente de $Z_1, \ldots, Z_{N-1}$, $\bar{X}$ será independiente de A. Por otra parte

$$E[Z_N] = \sum_\beta b_{N\beta} E[X_\beta] = \sum_\beta \frac{1}{\sqrt{N}} \mu_\beta = \sum_\beta \frac{1}{\sqrt{N}} \mu = \sqrt{N}\,\mu ,$$

con lo que Z_N se distribuye $N(\sqrt{N}\mu, \Sigma)$ y por tanto $\bar{X} = \frac{1}{\sqrt{N}} Z_N$ será $N\left(\mu, \frac{1}{N}\Sigma\right)$.

Además

$$E(Z_\alpha) = \sum_\beta b_{\alpha\beta} E[X_\beta] = \sum_\beta b_{\alpha\beta} \mu = \sum_\beta b_{\alpha\beta} b_{N\beta} \sqrt{N}\,\mu = 0 , \quad \text{para } \alpha \neq N.$$

Podemos recoger todos estos resultados en el siguiente teorema

TEOREMA.- La media muestral de una muestra de tamaño N de una población $N(\mu,\Sigma)$ se distribuye de acuerdo con una $N\left(\mu,(1/N)\Sigma\right)$ e independientemente de $\hat{\Sigma}$, estimador máximo verosímil de Σ. $N\hat{\Sigma}$ se distribuye igual que $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$ donde Z_α se distribuye $N(0,\Sigma)$ e independientemente de Z_β $(\alpha\neq\beta)$.

Observemos primera parte que

$$E(\hat{\Sigma}) = \frac{1}{N}\sum_{\alpha=1}^{N-1} E\left(Z_\alpha Z_\alpha'\right) = \frac{N-1}{N}\Sigma$$

Resulta pues que $\hat{\Sigma}$ es un estimador sesgado de Σ. Lo podemos convertir en un estimador insesgado definiendo

$$S = \frac{1}{N-1}A = \frac{1}{N-1}\sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})'.$$

Matriz conocida como la matriz de covarianzas muestral. Sus elementos diagonal son los estimadores insesgados habituales de las varianzas de las componentes de Σ.

Diferencias acerca de μ cuando Σ es conocida

Un problema estadístico de considerable importancia es el de contrastar la hipótesis de que el vector media de una distribución normal es un vector dado, y un problema directamente relacionado con este el dar una región de confianza para dicho vector media. Vamos a estudiar estos problemas bajo el supuesto de conocer la matriz de covarianzas Σ. Más adelante consideraremos estos problemas para el caso en que sea desconocida Σ.

Veremos que en el caso univariante y bajo planteamientos análogos a los anteriores, los tests usuales de confianza para la media se basan en el hecho de que la diferencia entre la media muestral y la poblacional se distribuye normalmente con media 0 y varianza conocida. Un hecho similar utilizaremos en nuestro caso, también ahora esta diferencia se distribuye normalmente con vector media nulo y matriz de covarianzas conocida. Podríamos elegir límites para cada componente sobre la base de esta distribución, pero este procedimiento tiene la desventaja de que la elección de límite es arbitraria y en el caso de tests conducen a tests que pueden ser muy pobres frente a algunas alternativas, no sin dificultad de calcular los referidos límites por cuanto sobre no se dispone de tablas para la normal bivariante. El procedimiento que vamos a describir es sencillo en cuanto a la obtención de resultados y además pueden darse justificaciones teóricas e intuitivas generales.

El procedimiento se basa en el siguiente teorema:

TEOREMA.- Si las m-componentes de un vector Y se distribuyen de acuerdo con una $N(0,T)$ (no singulares), entonces $Y'T^{-1}Y$ se distribuye según una χ^2 con m grados de libertad.

Demostración.- Sea C una matriz no singular tal que $CTC'=I$ y definamos $Z=CY$. Entonces Z se distribuye normalmente con media 0 y matriz de covarianzas $E(ZZ')=E(CYY'C')=CTC'=I$. Entonces $Y'T^{-1}Y = Z'(C')^{-1}T^{-1}C^{-1}Z = Z'(CTC')^{-1}Z = Z'Z$ que es la suma de los cuadrados de las componentes de Z. Como cada componente es $N(0,1)$ e independiente de las restantes $Z'Z=Y'T^{-1}Y$ es una χ^2 con m grados de libertad.

Ahora bien, $\sqrt{N}(\bar{X}-\mu)$ se distribuye $N(0,\Sigma)$, aplicando el teorema

$$N(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu)$$

será una χ^2 con p grados de libertad. Este resultado es fundamental para determinar regiones de confianza para μ.

Sea $\chi_p^2(\alpha)$, el número tal que

$$Pr\left(\chi_p^2 \geq \chi_p^2(\alpha)\right) = \alpha.$$

Entonces

$$Pr\left(N(\bar{X}-\mu)'\Sigma^{-1}(\bar{X}-\mu) \geq \chi_p^2(\alpha)\right) = \alpha$$

Para contrastar la hipótesis $\mu=\mu_0$, utilizaremos como región crítica (o de rechazo)

$$N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right) \geq \chi^2_p(\alpha)$$

Intuitivamente podemos observar que la probabilidad de rechazar $\boldsymbol{\mu}_0$ será mayor que α si $\boldsymbol{\mu}$ está muy diferente de $\boldsymbol{\mu}_0$, puesto que en el espacio de las $\bar{\mathbf{x}}$, la expresión anterior representa una elipse centrada en $\boldsymbol{\mu}_0$ y cuando $\boldsymbol{\mu}$ está alejado de $\boldsymbol{\mu}_0$ la densidad de $\bar{\mathbf{x}}$ se encontrará en un punto fuera de la citada elipse, será pues difícil encontrar valores de $\bar{\mathbf{x}}$ cercanos a $\boldsymbol{\mu}_0$.

Una demostración análoga a la del teorema anterior nos llevaría a demostrar que $N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)$ se distribuye como una χ^2 no centrada con p g.l. y parámetro de no centralización $N\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)$ cuando $\bar{\mathbf{x}}$ es la media de una muestra de tamaño N de una $N\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$. La primera demostración del teorema anterior fue hecha por Pearson en 1900.

Consideremos ahora la siguiente afirmación hecha sobre la base de una muestra con media $\bar{\mathbf{x}}$: "La media de la distribución satisface

$$N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right) \leq \chi^2_p(\alpha)$$

considerada como una desigualdad sobre $\boldsymbol{\mu}^*$". De acuerdo con lo anterior la probabilidad de que extraigamos una muestra tal que sea cierta la afirmación es $1-\alpha$, puesto que el suceso $\left\{N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right) \geq \chi^2_p(\alpha)\right\}$ es equivalente a negar la afirmación. Así, el conjunto de valores $\boldsymbol{\mu}^*$ que satisfacen la última desigualdad suelen constituyen una región de confianza para $\boldsymbol{\mu}$ con un nivel de confianza de $1-\alpha$.

En un espacio p-dimensional de $\bar{\mathbf{x}}$, $N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right) \geq \chi^2_p(\alpha)$ es la superficie y el exterior de un elipsoide centrado en $\boldsymbol{\mu}_0$, la forma del elipsoide depende de $\boldsymbol{\Sigma}^{-1}$ y el tamaño de $\left(1/N\right)\chi^2_p(\alpha)$ para $\boldsymbol{\Sigma}^{-1}$ dado. En el espacio p-dimensional de $\boldsymbol{\mu}^*$, $N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right) \leq \chi^2_p(\alpha)$ es la superficie y el interior de un elipsoide centrado en $\bar{\mathbf{x}}$. Si $\boldsymbol{\Sigma}^{-1}=\mathbf{I}$, la $\Pr\left\{N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right) \geq \chi^2_p(\alpha)\right\}=\alpha$, significa que la probabilidad de que la distancia entre $\bar{\mathbf{x}}$ y $\boldsymbol{\mu}$ sea mayor que $\left[\left(1/N\right)\chi^2_p(\alpha)\right]^{1/2}$ es igual a α.

TEOREMA.- Si $\bar{\mathbf{x}}$ es la media de una muestra de N elementos extraída de una $N\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ y si $\boldsymbol{\Sigma}$ es conocida, entonces $N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}_0\right) \geq \chi^2_p(\alpha)$ da una región crítica de tamaño α para contrastar la hipótesis $\boldsymbol{\mu}=\boldsymbol{\mu}_0$ y $N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}^*\right) \leq \chi^2_p(\alpha)$ da una región de confianza para $\boldsymbol{\mu}$ de nivel de confianza $1-\alpha$. $\chi^2_p(\alpha)$ es el percentil $1-\alpha$ de la distribución χ^2_p.

Caso de dos muestras

Una técnica similar puede ser utilizada para los problemas que surgen con dos muestras. Supongamos que tenemos una muestra $\left\{\mathbf{X}^{(1)}_\alpha\right\}$ $(\alpha=1,\dots,N_1)$ de la distribución $N\left(\boldsymbol{\mu}^{(1)},\boldsymbol{\Sigma}\right)$ y una muestra $\left\{\mathbf{x}^{(2)}_\alpha\right\}$ $(\alpha=1,\dots,N_2)$ de una segunda población normal $N\left(\boldsymbol{\mu}^{(2)},\boldsymbol{\Sigma}\right)$ ambas con la misma matriz de covarianzas. Entonces las dos medias muestrales

$$\bar{\mathbf{x}}^{(1)}=\frac{1}{N_1}\sum_{\alpha=1}^{N_1}\mathbf{x}^{(1)}_\alpha \qquad \bar{\mathbf{x}}^{(2)}=\frac{1}{N_2}\sum_{\alpha=1}^{N_2}\mathbf{x}^{(2)}_\alpha$$

se distribuirán independientemente $N\left(\boldsymbol{\mu}^{(1)},\left(1/N_1\right)\boldsymbol{\Sigma}\right)$ y $N\left(\boldsymbol{\mu}^{(2)},\left(1/N_2\right)\boldsymbol{\Sigma}\right)$ respectivamente. La diferencia entre ambas medias $\mathbf{y}=\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}$ se distribuye $N\left[\boldsymbol{\nu},\left(1/N_1+1/N_2\right)\boldsymbol{\Sigma}\right]$, donde $\boldsymbol{\nu}=\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}$. Así

$$\frac{N_1N_2}{N_1+N_2}\left(\mathbf{y}-\boldsymbol{\nu}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\nu}\right) \leq \chi^2_p(\alpha)$$

es una región de confianza para la diferencia $\boldsymbol{\nu}$ entre los dos vectores media y una región crítica para contrastar la hipótesis $\boldsymbol{\mu}^{(1)}=\boldsymbol{\mu}^{(2)}$ viene dada por

$$\frac{N_1N_2}{N_1+N_2}\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right)'\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\right) \geq \chi^2_p(\alpha).$$

Mahalanobis (1930) sugirió la utilización de $\left(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}\right)$ como una medida de la distancia entre las dos poblaciones.

Estadísticos suficientes para μ y Σ

Hemos demostrado que

$$\sum_{\alpha}(x_\alpha-\mu)(x_\alpha-\mu)' = A + N(\bar{x}-\mu)(\bar{x}-\mu)' \quad ,$$

$$\sum_{\alpha}(x_\alpha-\mu)'\Sigma^{-1}(x_\alpha-\mu) = tr(\Sigma^{-1}A) + N(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)$$

Así la densidad de $X_1 \cdots X_N$ podemos escribirla como

$$k\exp\left\{-\tfrac{1}{2}\left[N(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)+tr(\Sigma^{-1}A)\right]\right\} = k_1\exp\left\{-\tfrac{1}{2}N(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)\right\}k_2\exp\left\{-\tfrac{1}{2}tr(\Sigma^{-1}A)\right\}$$

Por tanto $\bar{x}$ y $(1/N)A$ forman un conjunto suficiente de estadísticos para μ y Σ. Si Σ es conocido, $\bar{x}$ es un estadístico suficiente para μ. Sin embargo, si μ es conocido, $(1/N)A$ no es un estadístico suficiente para Σ, pero sí lo es $(1/N)\sum_{\alpha=1}^{N}(x_\alpha-\mu)(x_\alpha-\mu)'$.

Recordemos que t es un estadístico suficiente para θ si

$$\prod_{\alpha=1}^{N} f(x_\alpha;\theta) = g(t;\theta).h(x_1,\ldots,x_N)$$

donde $f(x_\alpha;\theta)$ es la densidad de la α-ésima observación; $g(t;\theta)$ es una función de t y θ, y $h(x_1,\ldots,x_N)$ no depende de θ (Criterio de factorización de Fisher-Neymann).

Si un vector aleatorio q-dimensional Y tiene media ν y matriz de covarianzas Ψ, entonces

$$(y-\nu)'\Psi(y-\nu) = q+2$$

es denominado elipsoide de concentración de Y (vease Cramer, 344). La densidad definida mediante una distribución uniforme sobre el interior de este elipsoide tiene el mismo vector media y la misma matriz de covarianzas que Y.

Por otra parte, si θ es un vector de q parametros y t es un vector de estimadores insesgados de θ basados en N observaciones de una población con matriz de covarianzas Ψ, entonces el elipsoide

$$N(t-\theta)'E\left[\frac{\partial \log f}{\partial\theta}\right]\left[\frac{\partial \log f}{\partial\theta}\right]'(t-\theta) = q+2$$

está enteramente contenido en el elipsoide de concentración de t (vease Cramer p. 566). Si este último elipsoide coincide con el de concentración de t, entonces se dice que t es eficiente. En general el cuadrado de la razón del volumen de este último elipsoide y el del de concentración se define como la eficiencia relativa de t. En el caso de una distribución normal multivariante, si $\theta=\mu$, el vector de las medias muestrales $\bar{x}$ es eficiente. Si θ incluye a μ y a Σ, entonces $\bar{x}$ y S tienen una eficiencia relativa igual a $\left[(N-1)/N\right]^{p(p+1)/2}$.

En efecto

$$\log f = \log\left[\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\right] - \tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)$$

$$\frac{\partial \log f}{\partial \mu} = -\tfrac{1}{2}\left[-2\Sigma^{-1}(x-\mu)\right] \quad \text{y de aquí} \quad E\left[\frac{\partial\log f}{\partial\theta}\right]\left[\frac{\partial\log f}{\partial\theta}\right]' = E\left[(\Sigma^{-1}(x-\mu))(\Sigma^{-1}(x-\mu))'\right] =$$

$$= \Sigma^{-1}.E(x-\mu)(x-\mu)'.\Sigma^{-1} = \Sigma^{-1}$$

entonces el elipsoide correspondiente será

$$N(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) = q+2$$

pero $\bar{x}$ es $N\left(\mu, \frac{1}{N}\Sigma\right)$, entonces su elipsoide de concentración viene dado por

$$N(\bar{x}-\mu)'\frac{\Sigma^{-1}}{N^{-1}}(\bar{x}-\mu) = N(\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu) = q+2 \quad \text{. Ambos coinciden.}$$

Con frecuencia podemos encontrarnos con matrices de datos en las que al faltar algunas observaciones ponen en entredicho la utilización de las técnicas usuales para la obtención de las correspondientes estimaciones de μ y Σ. Situaciones de este tipo pueden presentarse por ejemplo, cuando algunos de los animales en estudio en una experimentación de laboratorio fallece por causas ajenas a la experiencia, cuando en grandes investigaciones interdisciplinarias surgen nuevas variables a medida que progresa el estudio. En cualquier caso, es esencial que las causas que originan las pérdidas de datos sean completamente independientes de la naturaleza o de los valores de las variables respuesta examinadas. Si el número de vectores de observaciones completas es pequeño y el coste de la obtención de datos es apreciable, preferiremos utilizar métodos de estimación y contrastes de hipótesis que optimicen el uso de la información disponible.

Las estimaciones máximo-verosímiles del vector media y de la matriz de covarianzas de una distribución multinormal mediante una matriz de datos incompleta con un modelo general aleatorio conduce a sistemas de ecuaciones no lineales cuya solución implica técnicas de análisis numérico. Pueden consultarse en este caso los trabajos de Wilks (1932), Elashoff y Afifi (1966), Timm (1970), Hartley an Hocking (1971) y Basickes (1977).

Un caso especial de pérdida de datos permitirá el cálculo directo de las estimaciones máximo-verosímiles de μ y Σ. Se trata del caso en que la matriz de datos obedece el modelo conocido con el nombre de "monótono" o "anidado" y que es de la forma

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1,K-1} & X_{1,K} \\ X_{21} & X_{22} & \cdots & X_{2,K-1} & - \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ X_{K-2,1} & X_{K-2,2} & \cdots & - & - \\ X_{K-1,1} & X_{K-1,2} & \cdots & - & - \\ X_{K,1} & \cdots & \cdots & - & - \end{bmatrix}$$

en la que los guiones indican bloques de observaciones perdidas. X_{ij} es una submatriz $N_i \times r_j$, donde $\sum_{i=1}^{k} N_i = N$, número total de unidades muestrales independientes y $\sum_{j=1}^{k} r_j = p$, el número de respuestas. Consecuentemente el vector media aparece fraccionado de la forma $\mu' = [\mu_1', \mu_2', \cdots, \mu_k']$ y la matriz de covarianzas aparece también fraccionada en matrices Σ_{ij} de dimensiones $r_i \times r_j$, $i,j = 1, \cdots, k$. Como consecuencia de la monotonía del modelo de la matriz de datos podemos escribir la verosimilitud de la muestra como

$$L(\mu, \Sigma) = f(X) = \left[\prod_{i=1}^{k} f(X_{i1})\right] \cdot \left[\prod_{i=1}^{k-1} f(X_{i2}/X_{i1})\right] \cdot \left[\prod_{i=1}^{k-2} f(X_{i3}/X_{i1}, X_{i2})\right] \cdots f(X_{1K}/X_{11} \cdots X_{1,K-1})$$

en la que la parte derecha de la igualdad aparece en términos de densidades condicionales a fin de evitar una notación excesivamente complicada. $f(X_{ij}/X_{i1}, \cdots, X_{i,j-1})$ es la verosimilitud condicional de la ij-ésima submatriz, cuya expresión puede obtenerse de la forma usual mediante el producto de N_i densidades multinormales r_i-dimensionales e independientes cuyos parámetros se obtienen aplicando las expresiones correspondientes a las densidades condicionadas. Se procede mediante la maximización sucesiva de los factores que aparecen en los corchetes: la primera de estas operaciones conduce a las estimaciones usuales $\hat{\mu}_1$ y $\hat{\Sigma}_{11}$ utilizando para ello las N unidades muestrales. El segundo factor proporciona estimaciones de $\mu_2 - \Sigma_{12}'\Sigma_{11}^{-1}\mu_1$, de los parámetros de regresión $\Sigma_{12}'\Sigma_{11}^{-1}$ y de la matriz de covarianzas condicional $\Sigma_{22} - \Sigma_{12}'\Sigma_{11}^{-1}\Sigma_{12}$. Con la ayuda de las estimaciones previas de μ_1 y Σ_{11} podemos obtener, a partir de los anteriores, las estimaciones máximo-verosímiles de μ_2, Σ_{12} y Σ_{22}. Continuamos de esta forma hasta que hayan sido estimados los parámetros incondicionales del k-ésimo conjunto. Este método de aproximación fue desarrollado por Anderson (1957) y fue extendido a más de dos conjuntos de variables por Bhargava (1962).

A continuación desarrollamos con detalle el método para el caso más sencillo, aquel en el que el modelo está formado tan sólo por dos conjuntos de variables. Sean $Y' \equiv (X_{11}, X_{21})$ y $\Sigma \equiv \Sigma_{12}$. Entonces

$$\hat{\mu}_1 = \frac{1}{N}\sum_{i=1}^{N} Y_i \quad , \quad \hat{\Sigma}_{11} = \frac{1}{N}\sum_{i=1}^{N} (Y_i - \bar{Y})(Y_i - \bar{Y})'$$

estimados en la forma usual a partir del conjunto completo de datos. Para las otras estimaciones tenemos

$$\bar{x} = \frac{1}{N_1}\sum_{i=1}^{N_1} x_i \qquad \bar{y}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} y_i \qquad \bar{y}_2 = \frac{1}{N_2}\sum_{i=N_1+1}^{N} y_i$$

$$A_{11}(N_1) = \sum_{i=1}^{N_1} (y_i - \bar{y}_1)(y_i - \bar{y}_1)' \qquad A_{11}(N) = \sum_{i=1}^{N} (y_i - \bar{y})(y_i - \bar{y})'$$

$$A_{22} = \sum_{i=1}^{N_1} (x_i - \bar{x})(x_i - \bar{x})' \qquad A_{12} = \sum_{i=1}^{N_1} (y_i - \bar{y}_1)(x_i - \bar{x})'$$

$$B = A_{11}^{-1}(N_1).A_{12}$$

Las estimaciones de los parametros estan son

$$\hat{\mu}_2 = \bar{x} - \left(\frac{N_2}{N}\right) B' (\bar{y}_1 - \bar{y}_2)$$

$$\hat{\Sigma}_{22} = \frac{1}{N_1}\left[A_{22} - A_{12}' A_{11}^{-1}(N_1) A_{12}\right] + \frac{1}{N} B' A_{11}(N) B$$

$$\hat{\Sigma}_{12} = \frac{1}{N} A_{11}(N) A_{11}^{-1}(N_1) A_{12}.$$

Las esperanzas y algunos momentos de segundo orden de estas estimaciones han sido calculadas por Morrison (1971). En particular $\hat{\mu}_2$ y $[N/(N-1)]\,\hat{\Sigma}_{12}$ son insesgados, mientras que es posible redefinir los divisores de los distintos términos de $\hat{\Sigma}_{22}$ para obtener el siguiente estimador insesgado:

$$\hat{\Sigma}_{22}(I) = \frac{1}{N_1 - r_1 - 1}\left[1 - \frac{r_1}{N_1 - r_1 - 2} + \frac{r_1(r_1+1)}{(N_1 - r_1 - 2)(N-1)}\right]\left[A_{22} - A_{12}' A_{11}^{-1}(N_1) A_{12}\right] + \frac{1}{N-1} B' A_{11}(N) B$$

Los valores de las varianzas de las estimaciones de μ_2, Σ_{22} y Σ_{12} para N_1 y N pequeños, indican que estas estimaciones son menos eficientes que las convencionales, obtenidas a partir, solamente, de las primeras N_1 observaciones completas cuando las correlaciones entre los dos conjuntos de variables son bajas. Normas para la elección del tipo de estimación más conveniente han sido indicadas, en la referencia anterior de Morrison (1971).

Ejemplo.- Una universidad utiliza una ecuación de regresión multiple para estimar los indices puntuales de acceso (IPA) de los solicitantes a partir de sus rangos de clase, tests de aptitud escolastica (TAS) y otros tipos de tests de interés general. Las estimaciones de los IPA se incluyen en los formularios resumen de los solicitantes para uso de los miembros del comité de admisión. Una muestra de $N = 34$ solicitudes evaluadas por un miembro de la facultad contiene cinco de ellas sin el correspondiente IPA, sino embargo, todas las solicitudes tienen TAS de matematicas y resultados de tests verbales entre otras medidas de tipo academico. Los datos así presentados son un ejemplo de modelo monotono de observaciones incompletas, con $r_1 = 2$ respuestas completas y $r_2 = 1$ incompletas, $N_1 = 29$ unidades muestrales completas y $N_2 = 5$ incompletas. Supondremos que las puntuaciones de los tests verbales (Y_1), del TAS de matematicas (Y_2) y del IPA (X) son una muestra trivariante de una distribución normal cuyos parámetros serán estimados mediante los métodos que acabamos de exponer.

Comenzaremos preparando una representación grafica informal de los valores del IPA frente a los resultados del TAS para los 29 vectores completos. Esta representación indica una ligera correlación positiva y parece que, en estas condiciones, las estimaciones basadas en observaciones incompletas hayan de ser más eficientes que las estimaciones basadas en los 29 resultados de los IPA. Tendremos pues

$$\bar{x} = 2.43 \qquad \bar{y}_1 = \begin{bmatrix} 56.38 \\ 61.31 \end{bmatrix} \qquad \bar{y}_2 = \begin{bmatrix} 55.80 \\ 55.60 \end{bmatrix} \qquad \bar{y} = \begin{bmatrix} 56.29 \\ 60.47 \end{bmatrix}$$

$$A_{11}(N_1) = \begin{bmatrix} 1292.83 & 206.59 \\ 206.59 & 1328.21 \end{bmatrix} \qquad A_{11}(N) = \begin{bmatrix} 1409.06 & 238.29 \\ 238.29 & 1668.47 \end{bmatrix} \qquad a_{22} = 2.9455$$

$$A_{11}^{-1}(N_1) = 10^{-4}\begin{bmatrix} 7.9321 & -1.2338 \\ -1.2338 & 7.7208 \end{bmatrix} \qquad B = 10^{-2}\begin{bmatrix} 1.6862 \\ 2.9142 \end{bmatrix}$$

y los estimadores máximo-verosímiles de los parámetros de IPA son

$$\hat{\mu}_x = 2.40 \qquad \hat{\sigma}_x^2 = 0.1034$$

La estimación de la media es ligeramente menor que la estimación usual basada en datos completos, $\bar{x}$. Para poder establecer comparaciones con $\hat{\sigma}_x^2$ a continuación presentamos distintas estimaciones alternativas de la varianza:

Estimación	Valor
MLE Univariante sesgada a_{22}/N	0.1016
MLE insesgada datos incompletos	0.1067
MLE univariante insesgada $a_{22}/(N-1)$	0.1052

La estimación basada en los datos incompletos, tanto en la sesgada como la insesgada, es ligeramente mayor que su correspondiente contrapartida univariante (basada en los datos completos).

Capítulo 3

Coeficientes de correlación muestrales

COEFICIENTES DE CORRELACIÓN MUESTRALES

1. INTRODUCCIÓN

Se pretende en este capítulo el estudio de los coeficientes de correlación muestrales que no serán más que las cantidades equivalentes a los coeficientes de correlación poblacionales definidos en el capítulo anterior. Calcularemos sus distribuciones de probabilidad y construiremos test de hipótesis y regiones de confianza para datas muestrales situaciones de interés.

En el caso de distribuciones normales conjuntas los coeficientes de correlación son la medida natural de la dependencia entre variables. Para las estimaciones de las correlaciones poblacionales utilizaremos los correspondientes muestrales y justificaremos el porqué.

Conviene recordar aquí la diferencia que existe entre los conceptos de regresión y correlación. En teoría de la regresión una variable se considera aleatoria o dependiente de otras fijas, o independiente. En teoría de la correlación consideramos todas las variables como aleatorias, y las tratamos todas ellas simétricamente. Si comenzamos con una distribución conjunta normal y mantenemos fijas todas las variables menos una, estudiaremos el modelo mínimo cuadrático puesto que la esperanza de la variable aleatoria en la distribución condicional es una función lineal de las variables que permanecieron fijas. Los coeficientes de regresión muestrales obtenidos mediante el método de los mínimos cuadrados son funciones de las varianzas muestrales y de las correlaciones.

Si contrastamos la independencia llegaremos a resultados análogos tanto si partimos de un caso como del otro (a saber: distribución conjunta normal o distribución condicional). La distribución de probabilidad bajo la hipótesis nula es la misma. La distribución de los criterios del test cuando la hipótesis nula no es cierta difiere en los dos casos. Si todas las variables pueden considerarse aleatorias utilizaremos la teoría de la correlación tal como aparece a continuación, si sólo una variable es aleatoria, utilizaremos la teoría de los mínimos cuadrados (posteriormente en otro capítulo).

2. COEFICIENTE DE CORRELACIÓN DE UNA MUESTRA BIVARIANTE

Hemos visto en el capítulo anterior que una estimación máximo-verosímil del coeficiente de correlación entre las componentes X_i y X_j del vector aleatorio $\mathbf{X}$, cuando utilizamos una muestra de tamaño N $(x_1, \dots x_N)$ es

$$r_{ij} = \frac{\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j)}{\sqrt{\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)^2} \sqrt{\sum_{\alpha=1}^{N} (x_{j\alpha} - \bar{x}_j)^2}}$$

Vamos a determinar la distribución de r_{ij} cuando la correlación poblacional entre X_i y X_j es cero y veremos como utilizar el coeficiente de correlación muestral para contrastar que el poblacional es cero. Desarrollaremos la teoría, por comodidad, para r_{12}. Como r_{12} depende sólo de las dos primeras componentes de cada x_α, bastará con considerar solamente la distribución conjunta de $(x_{11}, x_{21}), (x_{12}, x_{22}), \dots, (x_{1N}, x_{2N})$. Podemos entonces reformular el problema en términos de una distribución normal bivariante. Sean $x_1^* \dots x_N^*$ observaciones extraídas de

$$N\left[\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho\\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}\right]$$

Sea

$$r = r_{12} = \frac{a_{12}}{\sqrt{a_{11}}\sqrt{a_{22}}} \quad \text{con} \quad a_{ij} = \sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) \qquad i,j = 1,2$$

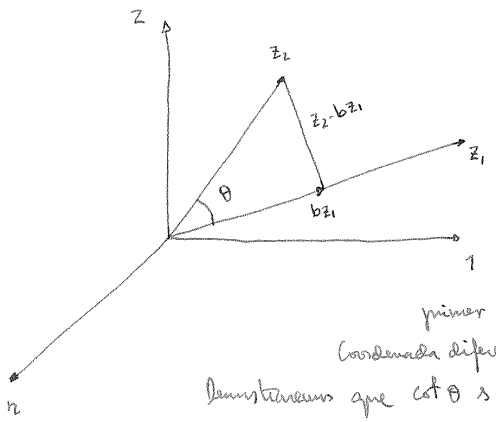
(siendo $x_{i\alpha}$ la i-ésima componente de x_α^*).

Sabemos que a_{ij} se distribuye como $\sum_{\alpha=1}^{n} z_{i\alpha} z_{j\alpha}$ $\quad i,j = 1,2$, donde $n = N-1$ y $(z_{1\alpha}, z_{2\alpha})$ se distribuye de acuerdo con una normal bivariante

$$N\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho\\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}\right]$$

e independientemente de $(z_{1\beta}, z_{2\beta})$, $\alpha \neq \beta$.

Sea $z_i' = (z_{i1}, \dots, z_{in})$, $i = 1, 2$. Estos dos vectores pueden representarse en un espacio n-dimensional.



El coeficiente de correlación es el coseno del ángulo θ entre z_1 y z_2. Para encontrar la distribución de $\cos\theta$ buscaremos primero la de la $\cot\theta$. Como $z_2 = (z_2 - bz_1) + bz_1$, podremos b como una función de z_1 y z_2 tal que $z_2 - bz_1$ sea ortogonal a bz_1. Entonces

$$\cot\theta = b\sqrt{z_1'z_1 / (z_2 - bz_1)'(z_2 - bz_1)}$$

Si z_1 permanece fijo podemos girar los ejes de manera que el primer eje de coordenadas coincida con z_1. Entonces bz_1 tiene todo la primera coordenada diferente de cero y $z_2 - bz_1$ tendrá la primera coordenada igual a cero.

Demostraremos que $\cot\theta$ es proporcional a una variable t cuando $\rho = 0$.

La distribución condicional de $Z_{2\alpha}$ dado $Z_{1\alpha} = z_{1\alpha}$ es $N(\beta z_{1\alpha}, \theta^2)$, donde $\beta = \rho\sigma_2/\sigma_1$ y $\theta^2 = \sigma_2^2(1-\rho^2)$ como vimos en su momento. La distribución conjunta de Z_2 dado $Z_1 = z_1$ es $N(\beta z_1, \theta^2 I)$ puesto que las $Z_{2\alpha}$ son independientes. Ahora bien, la densidad conjunta de Z_1 y Z_2 es

$$\prod_{\alpha=1}^{n} n\left[\begin{pmatrix} z_{1\alpha} \\ z_{2\alpha} \end{pmatrix} \middle| \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}\right]$$

y la marginal de Z_1 viene dada por

$$\prod_{\alpha=1}^{n} n\left(z_{1\alpha} \middle| 0, \sigma_1^2\right) = n\left(z_1 \middle| 0, \sigma_1^2 I\right)$$

por tanto la condicional de Z_2 dado $Z_1 = z_1$ será el cociente de la conjunta por la marginal de Z_1 en z_1, es decir

$$\prod_{\alpha=1}^{n} \left\{ n\left[\begin{pmatrix} z_{1\alpha} \\ z_{2\alpha} \end{pmatrix} \middle| \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}\right] \Big/ n\left(z_{1\alpha} \middle| 0, \sigma_1^2\right) \right\} = \prod_{\alpha=1}^{n} n\left(z_{2\alpha} \middle| \beta z_{1\alpha}, \theta^2\right).$$

Sea ahora $b = z_2'z_1 / z_1'z_1 \left(= a_{21}/a_{11}\right)$, entonces $bz_1'(z_2 - bz_1) = 0$, y sea $V = (z_2 - bz_1)'(z_2 - bz_1) = z_2'z_2 - b^2 z_1'z_1 \left(= a_{22} - a_{12}^2/a_{11}\right)$. Entonces $\cot\theta = b\sqrt{a_{11}/V}$. La situación de los ejes coordenados antes descrita supone elegir una matriz $(n \times n)$ ortogonal, C, cuya primera fila sea $(1/c)\, z_1'$ donde $c^2 = z_1'z_1$.

Podemos ahora aplicar un teorema del capítulo anterior que hacía referencia a transformación ortogonal, si hacemos $X\alpha = Z_{2\alpha}$ y $Y_\alpha = \sum_{\beta=1}^{n} C_{\alpha\beta} Z_{2\beta}$. Entonces las $\{Y_\alpha\}$ son independientes y se distribuyen normalmente con varianza σ^2 y medias

$$E(Y_1) = \sum_{\gamma=1}^{n} C_{1\gamma} E(Z_{2\gamma}) = \sum_{\gamma=1}^{n} C_{1\gamma} \beta z_{1\gamma} = \frac{\beta}{c} \sum_{\gamma=1}^{n} z_{1\gamma}^2 = \beta c$$

$$E(Y_\alpha) = \sum_{\gamma=1}^{n} C_{\alpha\gamma} E(Z_{2\gamma}) = \sum_{\gamma=1}^{n} C_{\alpha\gamma} \beta z_{1\gamma} = \beta c \sum_{\gamma=1}^{n} C_{\alpha\gamma} C_{1\gamma} = 0 \qquad , \alpha \neq 1.$$

Tenemos $b = \sum_\alpha Z_{2\alpha} z_{1\alpha} / \sum_\alpha z_{1\alpha}^2 = c \sum_\alpha Z_{2\alpha} C_{1\alpha} / \sum_\alpha z_{1\alpha}^2 = c Y_1 / c^2 = Y_1/c$ y del lema que presentabamos a continuación del teorema antes utilizado podemos deducir

$$V = \sum_\alpha Z_{2\alpha}^2 - b^2 \sum_\alpha z_{1\alpha}^2 = \sum_\alpha Y_\alpha^2 - Y_1^2 = \sum_{\alpha=2}^{n} Y_\alpha^2$$

que es independiente de b.

LEMA.— Si $(Z_{1\alpha}, Z_{2\alpha})$, $\alpha = 1, \dots, n$, son independientes, cada par con la distribución normal conjunta de media $\mu' = [0, 0]$ y $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}$, entonces la distribución condicional de $b = \sum_\alpha Z_{2\alpha} Z_{1\alpha} / \sum_\alpha Z_{1\alpha}^2$ y $V/\sigma^2 = \sum_\alpha (Z_{2\alpha} - bZ_{1\alpha})^2 / \sigma^2$ dado $Z_{1\alpha} = z_{1\alpha}$ $(\alpha = 1, \dots, n)$ es $N(\beta, \sigma^2/c^2)$, $c^2 = \sum_\alpha z_{1\alpha}^2$ y χ^2 con $n-1$ g.l. respectivamente y además b y V son independientes.

Si $\rho=0$, entonces $\beta=0$, por tanto b se distribuye incondicionalmente $N(0, \sigma^2/c^2)$ y

$$\frac{cb/\sigma}{\sqrt{\frac{V/\sigma^2}{n-1}}} = \frac{cb}{\sqrt{\frac{V}{n-1}}}$$

tiene una distribución condicional t con $n-1$ g.l. Pero bien, esta variable aleatoria es

$$\sqrt{n-1}\cdot\frac{(a_{11})^{1/2}a_{12}/a_{11}}{\sqrt{a_{22}-a_{12}^2/a_{11}}} = \sqrt{n-1}\,\frac{a_{12}/\sqrt{a_{11}a_{22}}}{\sqrt{1-[a_{12}^2/(a_{11}a_{22})]}} = \sqrt{n-1}\,\frac{r}{\sqrt{1-r^2}}$$

Así pues $\sqrt{n-1}\,r/\sqrt{1-r^2}$ tiene una distribución condicional t con $n-1$ g.l. La densidad de t es

$$\frac{\Gamma\left(\frac{1}{2}n\right)}{\sqrt{n-1}\,\Gamma\left[\frac{1}{2}(n-1)\right]\sqrt{\pi}}\left(1+\frac{t^2}{n-1}\right)^{-\frac{1}{2}n}$$

y la densidad de $r/\sqrt{1-r^2}$ viene dada por

$$\frac{\Gamma\left(\frac{1}{2}n\right)}{\Gamma\left(\frac{1}{2}(n-1)\right)\sqrt{\pi}}\left(1+w^2\right)^{-\frac{1}{2}n}$$

Puesto que $w = r(1-r^2)^{-\frac{1}{2}}$ y $dw/dr = (1-r^2)^{-\frac{3}{2}}$, la densidad de r vendrá dada por (haciendo $n=N-1$)

$$\frac{\Gamma\left(\frac{1}{2}(N-1)\right)}{\Gamma\left(\frac{1}{2}(N-2)\right)\sqrt{\pi}}\left(1-r^2\right)^{\frac{1}{2}(N-4)}.$$

No dijimos que esta es la densidad de r para z_1 fijo. Pero dado que no depende para nada de z_1 será también la marginal de r.

TEOREMA.- Sean $X_1, \ldots, X_n$ independientes, cada una con distribución $N(\mu, \Sigma)$. Si $\rho_{ij}=0$, la densidad de r_{ij} definido como antes, viene dada por la última expresión anterior.

Observemos que la densidad de r es simétrica respecto del origen. Además, para $N>4$, tiene una moda en $r=0$ y presenta saltos en los puntos ± 1 cuyo orden es $\frac{1}{2}(N-5)$ para N impar y $\frac{1}{2}N-3$ para N par. Puesto que la densidad es par los momentos impares son nulos, en particular la media es cero. Los momentos pares se pueden encontrar por integración (haciendo el cambio $x=r^2$ y utilizando la función β), se obtiene $E[r^{2m}]$ mediante la expresión $E[r^{2m}] = \Gamma\left(\frac{1}{2}(N-1)\right)\Gamma\left(\frac{1}{2}+m\right)\Big/\left\{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}(N-1)+m\right)\right\}$, en particular la varianza ($m=1$) es $1/(N-1)$.

El uso más importante del anterior teorema es buscar puntos de significación para contrastar la hipótesis de que un par de variables estén incorreladas. Consideremos la hipótesis

$$H_0: \rho_{ij}=0$$

para algún par (i,j). Parece razonable rechazar H_0 si r_{ij} difiere mucho de cero, pero cómo cuantificar este "mucho".

Supongamos que vamos a contrastar H_0 frente a $H_1: \rho_{ij}>0$. Rechazaremos H_0 si r_{ij} es mayor que una determinada cantidad r_0. La probabilidad de rechazar H_0 siendo cierta es

$$\int_{r_0}^{1} K_N(r)\,dr$$

donde $K_N(r)$ es la densidad del coeficiente de correlación muestral basado en N observaciones. Elegimos r_0 de manera que la anterior probabilidad coincida con el nivel de significación deseado. Si la alternativa es de la forma $\rho_{ij}<0$, entonces rechazaremos para $r_{ij}<-r_0$.

Para el caso de una alternativa bilateral, $\rho_{ij} \neq 0$. Entonces rechazaremos H_0 si $r_{ij} > r_1$ o $r_{ij} < -r_1$. La probabilidad de rechazo para H_0, siendo cierta, vendrá dada por

$$\int_{-1}^{-r_1} k_N(r)\,dr + \int_{r_1}^{1} k_N(r)\,dr.$$

Existen tablas (Fisher y Yates, 1942) para la distribución de r. De cualquier forma, teniendo en cuenta que $\sqrt{N-2}\, r/\sqrt{1-r^2}$ tiene una distribución t con $N-2$ g.l. podemos usar alternativamente esta distribución para la obtención de r_0 y r_1. Así, para alternativas del tipo $\rho_{ij} \neq 0$, rechazaremos H_0 si

$$\sqrt{N-2}\ \frac{|r_{ij}|}{\sqrt{1-r_{ij}^2}} > t_{N-2}(\alpha)$$

de forma análoga actuaríamos en los otros casos.

Por otra parte, en el transcurso de la deducción del test anterior quedó claro que el mismo estadístico puede ser utilizado para contrastar la hipótesis de que la regresión de Z_2 sobre Z_1 es cero. En términos de las observaciones originales, tenemos

$$\sqrt{N-2}\ \frac{r}{\sqrt{1-r^2}} = \frac{b\sqrt{\sum_\alpha (x_{1\alpha} - \bar{x}_1)^2}}{\sqrt{\sum_\alpha \left[x_{2\alpha} - \bar{x}_2 - b(x_{1\alpha} - \bar{x}_1)\right]^2/(N-2)}}$$

donde $b = \sum_\alpha (x_{2\alpha} - \bar{x}_2)(x_{1\alpha} - \bar{x}_1) \Big/ \sum_\alpha (x_{1\alpha} - \bar{x}_1)^2$ es el coeficiente de regresión mínimo cuadrático de $x_{2\alpha}$ sobre $x_{1\alpha}$. Ya hemos visto que el test $\rho_{12} = 0$ es equivalente al test de que la regresión de X_2 sobre x_1 es nula (es decir, que $\rho_{12}\sigma_2/\sigma_1 = 0$).

Distribución cuando $\rho \neq 0$. Test de hipótesis y regiones de confianza.

En este caso deberemos en primer lugar obtener la distribución conjunta de a_{11}, a_{12} y a_{22}. Ya vimos antes que las condicionales, para z_1 fijo, de $b = a_{12}/a_{11}$ y $v/\sigma^2 = (a_{22} - a_{12}^2/a_{11})/\sigma^2$ son independientes y más concretamente $N(\beta, \sigma^2/c^2)$ y χ^2 con $n-1$ g.l. respectivamente. Llamando la densidad de la χ^2_{n-1} mediante $g_{n-1}(v)$, la conjunta correspondiente de b y v será

$$n\left(b \mid \beta, \sigma^2/a_{11}\right) g_{n-1}(v/\sigma^2)/\sigma^2$$

La conjunta de z_1, b y v vendrá dada por

$$n\left(z_1 \mid 0, \sigma_1^2 I\right) \cdot n\left(b \mid \beta, \sigma^2/a_{11}\right) \cdot g_{n-1}(v/\sigma^2)/\sigma^2.$$

La densidad marginal de $z_1'z_1/\sigma_1^2 = a_{11}/\sigma_1^2$ se puede obtener de mediante

$$\int \cdots \int_{z_1'z_1 = a_{11}} n\left(z_1 \mid 0, \sigma_1^2 I\right) dW$$

y como se trata de la suma de cuadrados de n normales $(0,1)$, tendremos que ésta una χ^2 con n g.l., es decir

$$\frac{1}{\sigma_1^2}\, g_n\left(\frac{a_{11}}{\sigma_1^2}\right) = \int \cdots \int_{z_1'z_1 = a_{11}} n\left(z_1 \mid 0, \sigma_1^2 I\right) dW.$$

donde dW es el diferencial de volumen ordenado, más concretamente es un elemento de volumen sobre la esfera $z_1'z_1 = a_{11}$. La densidad conjunta de b, v y a_{11} es

$$\int \cdots \int_{z_1'z_1 = a_{11}} n\left(b \mid \beta, \sigma^2/a_{11}\right) g_{n-1}(v/\sigma^2)\, \frac{1}{\sigma^2}\, n\left(z_1 \mid 0, \sigma_1^2 I\right) dW =$$

$$= g_n\left(a_{11}/\sigma_1^2\right) \cdot n\left(b \mid \beta, \sigma^2/a_{11}\right) \cdot g_{n-1}\left(v/\sigma^2\right) \cdot \left(\sigma_1^2\sigma^2\right) =$$

$$= \frac{a_{11}^{\frac{1}{2}n-1}}{\left(2\sigma_1^2\right)^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)} \exp\left(-\frac{1}{2\sigma_1^2}a_{11}\right) \cdot \frac{\sqrt{a_{11}}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{a_{11}}{2\sigma^2}(b-\beta)^2\right] \cdot \frac{1}{\left(2\sigma^2\right)^{\frac{1}{2}(n-1)}\Gamma\left(\frac{1}{2}(n-1)\right)} v^{\frac{1}{2}(n-3)} \exp\left(-\frac{1}{2\sigma^2}v\right)$$

Hagamos ahora $b = a_{12}/a_{11}$, $v = a_{22} - a_{12}^2/a_{11}$. El Jacobiano de la transformación es

$$\frac{\partial(b,v)}{\partial(a_{11},a_{22})} = \begin{vmatrix} \frac{1}{a_{11}} & 0 \\ -2\frac{a_{12}}{a_{11}} & 1 \end{vmatrix} = \frac{1}{a_{11}} .$$

Así la densidad para a_{11}, a_{12}, a_{22} será

$$\frac{a_{11}^{\frac{1}{2}(n-3)}\left(\frac{a_{11}a_{22}-a_{12}^2}{a_{11}}\right)^{\frac{1}{2}(n-3)} \cdot e^{-\frac{1}{2}Q}}{2^n\sigma_1^n\left(\frac{\sigma_1^2\sigma_2^2-\rho^2\sigma_1^2\sigma_2^2}{\sigma_1^2}\right)\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n\right)\cdot\Gamma\left(\frac{1}{2}(n-1)\right)}$$

donde

$$Q = \frac{a_{11}}{\sigma_1^2} + \frac{a_{11}}{\sigma^2}\left(\frac{a_{12}^2}{a_{11}^2} - 2\rho\frac{\sigma_1\sigma_2}{\sigma_1^2}\cdot\frac{a_{12}}{a_{11}} + \rho^2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^4}\right) + \frac{1}{\sigma^2}\left(a_{22} - \frac{a_{12}^2}{a_{11}}\right) =$$

$$= a_{11}\left[\frac{1}{\sigma^2} + \frac{\rho^2\sigma_1^2\sigma_2^2}{\sigma_1^4\sigma_2^2(1-\rho^2)}\right] - 2a_{12}\frac{\rho\sigma_2}{\sigma_1\sigma_2^2(1-\rho^2)} + \frac{a_{22}}{\sigma_2^2(1-\rho^2)} =$$

$$= \frac{1}{1-\rho^2}\left(\frac{a_{11}}{\sigma_1^2} - 2\rho\frac{a_{12}}{\sigma_1\sigma_2} + \frac{a_{22}}{\sigma_2^2}\right) .$$

La densidad puede también escribirse

$$\frac{|A|^{\frac{1}{2}(n-3)}\, e^{-\frac{1}{2}Q}}{2^n|\Sigma|^{\frac{1}{2}n}\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n\right)\Gamma\left(\frac{1}{2}(n-1)\right)} .$$

que es un caso particular de la distribución de Wishart que obtendremos en el capítulo posterior.

Ahora la densidad de a_{11}, a_{22} y $r = a_{12}/\sqrt{a_{11}a_{22}}$ ($da_{12} = dr\sqrt{a_{11}a_{22}}$) vendrá dada por

$$\frac{a_{11}^{\frac{1}{2}n-1} \cdot a_{22}^{\frac{1}{2}n-1} \cdot (1-r^2)^{\frac{1}{2}(n-3)}\, e^{-\frac{1}{2}Q}}{2^n\left[\sigma_1^2\sigma_2^2(1-\rho^2)\right]^{\frac{1}{2}n}\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n\right)\cdot\Gamma\left(\frac{1}{2}(n-1)\right)}$$

donde

$$Q = \frac{1}{1-\rho^2}\left(\frac{a_{11}}{\sigma_1^2} - 2\rho r\frac{\sqrt{a_{11}}\sqrt{a_{22}}}{\sigma_1\sigma_2} + \frac{a_{22}}{\sigma_2^2}\right)$$

Para obtener la densidad de r deberemos integrar la anterior expresión respecto a a_{11} y a_{22} obteniendo la requerida real positiva. Podemos hacerlo de varias maneras, que dan lugar a diferentes expresiones para la densidad. Estudiaremos aquí el método más directo. Desarrollamos en primer lugar parte de la exponencial

$$\exp\left[\frac{\rho r\sqrt{a_{11}}\sqrt{a_{22}}}{(1-\rho^2)\sigma_1\sigma_2}\right] = \sum_{\alpha=0}^{\infty}\frac{\left(\rho r\sqrt{a_{11}}\sqrt{a_{22}}\right)^{\alpha}}{\alpha!\left[\sigma_2\sigma_1(1-\rho^2)\right]^{\alpha}} .$$

La nueva expresión para la densidad de a_{11}, a_{22} y r será

$$\frac{(1-\rho^2)^{\frac{1}{2}(n-3)}}{\sigma_1^n\sigma_2^n(1-\rho^2)^{\frac{1}{2}n}2^n\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n\right)\Gamma\left(\frac{1}{2}(n-1)\right)}\sum_{\alpha=0}^{\infty}\frac{(\rho r)^\alpha}{\alpha!(1-\rho^2)^\alpha\sigma_1^\alpha\sigma_2^\alpha}\cdot\left\{\exp\left[-\frac{a_{11}}{2(1-\rho^2)\sigma_1^2}\right]a_{11}^{\frac{1}{2}(n+\alpha)-1}\right\}\cdot\left\{\exp\left[-\frac{a_{22}}{2(1-\rho^2)\sigma_2^2}\right]\cdot a_{22}^{\frac{1}{2}(n+\alpha)-1}\right\}.$$

puesto que

$$\int_0^{+\infty}a_{11}^{\frac{1}{2}(n+\alpha)-1}\exp\left[-\frac{a_{11}}{2(1-\rho^2)\sigma_1^2}\right]da_{11}=\Gamma\left(\tfrac{1}{2}(n+\alpha)\right)\cdot\left[2\sigma_1^2(1-\rho^2)\right]^{\frac{1}{2}(n+\alpha)}$$

integrando la densidad (podemos integrar término a término) llegamos a

$$\frac{(1-\rho^2)^{\frac{1}{2}(n-3)}}{\sigma_1^n\sigma_2^n(1-\rho^2)^{\frac{1}{2}n}2^n\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n\right)\Gamma\left(\frac{1}{2}(n-1)\right)}\sum_{\alpha=0}^{\infty}\frac{(\rho r)^\alpha}{\alpha!(1-\rho^2)^\alpha\sigma_1^\alpha\sigma_2^\alpha}\cdot\Gamma^2\left[\tfrac{1}{2}(n+\alpha)\right]\cdot 2^{(n+\alpha)}\sigma_1^{(n+\alpha)}\cdot\sigma_2^{(n+\alpha)}(1-\rho^2)^{(n+\alpha)}=$$

$$=\frac{(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}n\right)\Gamma\left(\frac{1}{2}(n-1)\right)}\sum_{\alpha=0}^{\infty}\frac{(2\rho r)^\alpha}{\alpha!}\Gamma^2\left(\tfrac{1}{2}(n+\alpha)\right)$$

Utilizando la fórmula $\Gamma(z)\cdot\Gamma\left(z+\frac{1}{2}\right)=\sqrt{\pi}\,\Gamma(2z)/2^{2z-1}$ podemos modificar la constante. Enunciamos entonces el siguiente teorema:

Teorema.- El coeficiente de correlación en una muestra de tamaño N de una normal bivariante con correlación ρ se distribuye con densidad

$$\frac{2^{n-2}(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{(n-2)!\,\pi}\sum_{\alpha=0}^{\infty}\frac{(2\rho r)^\alpha}{\alpha!}\Gamma^2\left(\tfrac{1}{2}(n+\alpha)\right)\qquad\text{donde } n=N-1.$$

La distribución de r fue obtenida por primera vez por Fisher en 1915, quien dió para su densidad la siguiente expresión

$$\frac{(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}}{\pi(n-2)!}\left[\frac{d^{n-1}}{dx^{n-1}}\left\{\frac{\cos^{-1}(-x)}{\sqrt{1-x^2}}\right\}\Bigg|_{x=r\rho}\right]$$

que se obtiene de las expresiones iniciales haciendo los cambios $a_{11}=u\,e^{-v}$, $a_{22}=u\,e^{v}$.
Hotelling estudió exhaustivamente la distribución de r y recomendó el uso de la expresión

$$\frac{n-1}{\sqrt{2\pi}}\,\frac{\Gamma(n)}{\Gamma\left(n+\frac{1}{2}\right)}(1-\rho^2)^{\frac{1}{2}n}(1-r^2)^{\frac{1}{2}(n-3)}(1-\rho r)^{-n+\frac{1}{2}}\,\mathbf{F}\left(\tfrac{1}{2},\tfrac{1}{2};n+\tfrac{1}{2};\tfrac{1+\rho r}{2}\right)$$

donde

$$F(a,b;c;x)=\sum_{j=0}^{\infty}\frac{\Gamma(a+j)}{\Gamma(a)}\cdot\frac{\Gamma(b+j)}{\Gamma(b)}\cdot\frac{\Gamma(c)}{\Gamma(c+j)}\cdot\frac{x^j}{j!}$$ es una función hipergeométrica.

Esta expresión se obtiene también utilizando los cambios de variable utilizados por Fisher. La ventaja de utilizar esta fórmula para la densidad estriba en que la serie converge más rápidamente que en la expresión que dimos en el teorema anterior.

La distribución acumulada de r, $\Pr(r\le r^*)=F(r^*|N,\rho)$ ha sido tabulada por F. N. David (1938) para $\rho=0-.9$ con incrementos de .1, $N=3(1)25,30,100,200,400$, $r^*=-1(.05)1$. Dada la simetría de la densidad para r y ρ tenemos que $F(r^*|N,\rho)=1-F(-r^*|N,-\rho)$. Haremos el uso de las tablas en algunos procedimientos estadísticos, de los que nos ocuparemos a continuación.

a) Consideremos el problema de contrastar la hipótesis

$H_0 : \rho = \rho_0$

a partir de una muestra determinada.

Si la alternativa es del tipo $H_1: \rho > \rho_0$, rechazaremos H_0 si el coeficiente de correlación muestral es tal que supera a r_0, donde r_0 lo elegimos de manera que $1 - F(r_0 | N, \rho_0) = \alpha$, nivel de significación.

Si la alternativa es de la forma $H_1: \rho < \rho_0$, rechazaremos para $r_{ij} < r_0'$ un r_0' tal que $F(r_0' | N, \rho_0) = \alpha$. Para $H_1: \rho \neq \rho_0$, la zona de rechazo viene dada por $r_{ij} < r_1'$ y $r_{ij} > r_2$ donde r_2 y r_1' son tales que $F(r_1' | N, \rho_0) + (1 - F(r_2 | N, \rho_0)) = \alpha$. David sugiere que r_2 y r_1' sean elegidos de manera que ambas colas tengan la misma probabilidad e igual a $\frac{1}{2}\alpha$. Demostró en 1937 que para $N \geq 10$ y $|\rho| \leq 0.8$ la región crítica es similar a la región de un test insesgado de H_0, es decir, un test cuya función de potencia tiene un mínimo en ρ_0.

Debemos también señalar que cualquier test basado en r es invariante bajo transformaciones del tipo $x_{i\alpha}^* = c_i x_{i\alpha} + d_i$ $(\alpha = 1, \dots, N,\ i = 1,2,\ c_i > 0)$ y que r es el invariante único de los estadísticos suficientes (lo que significa que cualquier otro test invariante puede expresarse como función de r). El procedimiento anterior para contrastar la hipótesis $H_0: \rho = \rho_0$ frente a alternativas del tipo $\rho > \rho_0$ es uniformemente más potente frente a todos los test invariantes.

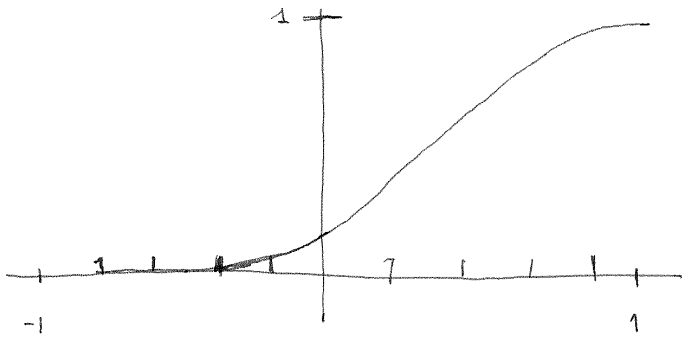
A modo de ejemplo, supongamos que deseamos contrastar la hipótesis $H_0: \rho = 0.5$ frente a la alternativa $\rho \neq 0.5$ a un nivel $\alpha = 5\%$ utilizando una correlación muestral obtenida a partir de una muestra de tamaño 15. En las tablas de David encontramos (por interpolación) $F(0.027 | 15, 0.5) = 0.025$ y $F(0.805 | 15, 0.5) = 0.975$. Por tanto rechazaríamos la hipótesis si r, en nuestra muestra, es menor que 0.027 o mayor que 0.805.

b) Podemos utilizar las tablas de David para calcular la función de potencia de un test de correlación. Si la región de rechazo de H_0 es $r > r_1$ y $r < r_1'$, la potencia del test es una función de la verdadera correlación ρ, $[1 - F(r_1 | N, \rho)] + [F(r_1' | N, \rho)]$; es decir la probabilidad de rechazar la hipótesis nula cuando la correlación poblacional es ρ.

Consideremos, por ejemplo la función de potencia del test para $\rho = 0$ que obtenemos en el ejemplo anterior. La región de rechazo (unilateral) es $r \geq 0.5494$ para $\alpha = 5\%$. Las probabilidades de rechazo son

ρ	-1.0	-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6	0.8	1
Prob	0.0000	0.0000	0.0004	0.0032	0.0147	0.0500	0.1376	0.3215	0.6235	0.9279	1.0000

que gráficamente da lugar a la siguiente representación



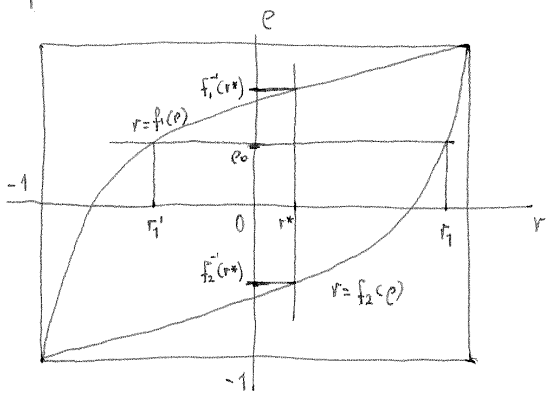
c) Los cálculos de David proporcionan también regiones de confianza para ρ. Para N dados, r_1' y r_1, los dos puntos de significación son funciones de ρ, a saber $f_1(\rho)$ y $f_2(\rho)$ respectivamente, de manera que

$$\Pr\{f_1(\rho) < r < f_2(\rho) \mid \rho\} = 1-\alpha$$

Las funciones $f_1(\rho)$ y $f_2(\rho)$ son funciones monótonas crecientes en ρ si r_1 y r_1' se eligen de manera que $1-F(r_1|N,\rho) = \frac{1}{2}\alpha = F(r_1'|N,\rho)$. Si $\rho = f_i^{-1}(r)$ es la inversa de $r = f_i(\rho)$, $i=1,2$ entonces la desigualdad $f_2(\rho) < r$ es equivalente a $\rho < f_1^{-1}(r)$ y $r < f_2(\rho)$ es equivalente a $f_2^{-1}(r) < \rho$. De esta forma la anterior probabilidad puede escribirse

$$\Pr\{f_2^{-1}(r) < \rho < f_1^{-1}(r) \mid \rho\} = 1-\alpha$$

Es decir la probabilidad de extraer una muestra y obtener un r de manera que el intervalo de extremos $[f_2^{-1}(r), f_1^{-1}(r)]$ cubra a ρ es precisamente $1-\alpha$. Se trata pues de un intervalo de confianza para ρ con nivel de confianza $1-\alpha$. Para valores de N y α dados



las curvas $r = f_1(\rho)$ y $r = f_2(\rho)$ tienen el aspecto de la figura. Cuando contrastamos $H_0: \rho = \rho_0$, la intersección de la línea $\rho = \rho_0$ con las dos curvas dadas determina r_1 y r_1' de manera que el nivel es el α fijado. Si queremos determinar una región de confianza para ρ sobre la base de r^*, coeficiente de correlación muestral, intersectaríamos las curvas anteriores en $r = r^*$ y determinaríamos los extremos del intervalo correspondiente. David proporciona estas curvas para $\alpha = 0.1, 0.05, 0.02$ y 0.01 y para varias N. Para el caso de test o regiones unilaterales utilizaríamos también una desigualdad de las anteriores.

También es posible el uso de $F(r|N,\rho)$ para determinar intervalos de confianza. Dado r^*, $f_1^{-1}(r^*)$ es el valor de ρ que hace que $F(r^*|N,\rho) = \frac{1}{2}\alpha$ y análogamente $f_2^{-1}(r^*)$ es aquel valor de ρ para el cual $1-F(r^*|N,\rho) = \frac{1}{2}\alpha$.

Por ejemplo, un intervalo de confianza con $1-\alpha = 0.95$ basado en un $r^* = 0.79\Omega$ obtenido a partir de una muestra de tamaño 10. Utilizando la tabla IV de David encontramos unos límites del intervalo 0.34 y 0.94

Criterio de la razón de verosimilitud

Es interesante encontrar el intervalo de razón de verosimilitud para contrastar $H_0: \rho = \rho_0$ a partir de una muestra $x_1 - x_N$ de una $N(\mu, \Sigma)$ con $\mu' = (\mu_1, \mu_2)$ y $\sigma_{11} = \sigma_1^2$, $\sigma_{12} = \sigma_{21} = \rho\sigma_1\sigma_2$ y $\sigma_{22} = \sigma_2^2$. La función de verosimilitud es

$$K\sigma_1^{-N}\sigma_2^{-N}(1-\rho^2)^{-N/2}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{a_{11}}{\sigma_1^2} - 2\rho\frac{a_{12}}{\sigma_1\sigma_2} + \frac{a_{22}}{\sigma_2^2} + N\frac{(\bar{x}_1-\mu_1)^2}{\sigma_1^2} - 2\rho N\frac{(\bar{x}_1-\mu_1)(\bar{x}_2-\mu_2)}{\sigma_1\sigma_2} + N\frac{(\bar{x}_2-\mu_2)^2}{\sigma_2^2}\right]\right\}$$

Si maximizamos la anterior expresión cuando los parámetros varían en Ω (con las restricciones $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, $\rho^2 < 1$) obtenemos $\hat{\sigma}_{1\Omega}^2 = a_{11}/N$, $\hat{\sigma}_{2\Omega}^2 = a_{22}/N$, $\hat{\rho}_\Omega = a_{12}/\sqrt{a_{11}a_{22}}$, $\hat{\mu}_{1\Omega} = \bar{x}_1$ y $\hat{\mu}_{2\Omega} = \bar{x}_2$.

Si maximizamos en $w \in \Omega$ (Ω restringido para $\rho = \rho_0$) $\hat{\mu}_{1w} = \bar{x}_1$ y $\hat{\mu}_{2w} = \bar{x}_2$ pues la forma cuadrática de $\bar{x}_1 - \mu_1$ y $\bar{x}_2 - \mu_2$ en el exponente es definida negativa (y alcanza por tanto su máximo cuando vale 0). Para maximizar el resto tomamos logaritmos y tenemos

$$\log k - N \log \sigma_1 - N \log \sigma_2 - \frac{1}{2} N \log (1-\rho_0^2) - \frac{1}{2(1-\rho_0^2)} \left(\frac{a_{11}}{\sigma_1^2} - 2\rho_0 \frac{a_{11}}{\sigma_1 \sigma_2} + \frac{a_{22}}{\sigma_2^2} \right)$$

Derivando respecto de σ_i, $i=1,2$ e igualando a 0, obtenemos

$$-\frac{N}{\hat{\sigma}_{iw}} - \frac{1}{2(1-\rho_0^2)} \left(-2 \frac{a_{ii}}{\hat{\sigma}_{iw}^3} + 2\rho_0 \frac{a_{ij}}{\hat{\sigma}_{iw}^2 \hat{\sigma}_{jw}} \right) = 0 \qquad , i \neq j \ , i,j = 1,2$$

Esto es

$$\frac{a_{ii}}{\hat{\sigma}_{iw}^2} - \rho_0 \frac{a_{ij}}{\hat{\sigma}_{iw} \hat{\sigma}_{jw}} = N(1-\rho_0^2)$$

Sumando para $i=1, j=2$ y $j=1, i=2$, llegamos

$$\frac{a_{11}}{\hat{\sigma}_{1w}^2} - 2\rho_0 \frac{a_{12}}{\hat{\sigma}_{1w} \hat{\sigma}_{2w}} + \frac{a_{22}}{\hat{\sigma}_{2w}^2} = 2N(1-\rho_0^2).$$

Restando para $i=1, j=2$ e $i=2, j=1$, tenemos

$$\frac{a_{11}}{\hat{\sigma}_{1w}^2} = \frac{a_{22}}{\hat{\sigma}_{2w}^2} = \frac{a^2}{\hat{\sigma}^2} \ . \text{ Entonces}$$

$$\frac{a^2}{\hat{\sigma}^2} - \frac{\rho_0 r a^2}{\hat{\sigma}^2} = N(1-\rho_0^2)$$

Así pues

$$\frac{N(1-\rho_0^2)}{1-\rho_0 r} = \frac{a^2}{\hat{\sigma}^2} = \frac{\sqrt{a_{11} a_{22}}}{\hat{\sigma}_{1w} \hat{\sigma}_{2w}}$$

y

$$\left| \hat{\Sigma}_w \right| = (1-\rho_0^2) \hat{\sigma}_{1w}^2 \hat{\sigma}_{2w}^2 = \frac{(1-\rho_0 r)^2}{1-\rho_0^2} \cdot \frac{a_{11}}{N} \cdot \frac{a_{22}}{N} .$$

Los correspondientes máximos para w y Ω, son

$$\max_w L = \frac{(1-\rho_0^2)^{\frac{1}{2}N} N^N}{(2\pi)^N (1-\rho_0 r)^N a_{11}^{\frac{1}{2}N} a_{22}^{\frac{1}{2}N}} e^{-N}$$

$$\max_{\Omega} L = \frac{N^N}{(2\pi)^N (1-r^2)^{\frac{1}{2}N} a_{11}^{\frac{1}{2}N} a_{22}^{\frac{1}{2}N}} e^{-N}$$

El criterio de la razón de verosimilitud es por tanto

$$\frac{\max_w L}{\max_{\Omega} L} = \frac{(1-\rho_0^2)^{\frac{1}{2}N} (1-r^2)^{\frac{1}{2}N}}{(1-\rho_0 r)^N} = \left[\frac{(1-\rho_0^2)(1-r^2)}{(1-\rho_0 r)^2} \right]^{\frac{1}{2}N}$$

El test correspondiente es $(1-\rho_0^2)(1-r^2)(1-\rho r)^{-2} < c$, donde c se elige de manera que la probabilidad de la desigualdad, cuando las muestras se extraen de poblaciones normales bivariantes con $\rho=\rho_0$, sea el nivel de significación elegido. La región crítica puede también presentarse

$$(\rho_0^2 c - \rho_0^2 + 1) r^2 - 2\rho_0 c r + c - 1 + \rho_0^2 > 0,$$

o bien

$$r > \frac{\rho_0 c + (1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c + 1 - \rho_0^2} \quad , \quad r < \frac{\rho_0 c - (1-\rho_0^2)\sqrt{1-c}}{\rho_0^2 c + 1 - \rho_0^2}$$

Así, el test de la razón de verosimilitud de $H_0: \rho=\rho_0$ frente a la alternativa $H_1: \rho \neq \rho_0$ tiene una región crítica de la forma $r > r_1$ y $r < r_1'$, donde r_1 y r_1' no están elegidos de manera que la probabilidad de cada desigualdad $\leq \frac{1}{2}\alpha$ cuando H_0 es cierta, sino de la forma descrita en las dos últimas desigualdades y con c teniendo el significado antes citado (la probabilidad de ambas desigualdades igual a α).

La distribución asintótica de un coeficiente de correlación muestral y de la Z de Fisher

Se trata de demostrar en este apartado que a medida que aumenta el tamaño de la muestra la distribución del coeficiente de correlación muestral tiende a la normal. La distribución de una función particular de la correlación muestral, conocida como la Z de Fisher (obtenida por Fisher en 1921), y cuya varianza es aproximadamente independiente de la correlación poblacional, tiende más rápidamente a la normal. En primer lugar demostraremos un teorema central del límite para variables n-dimensionales.

TEOREMA.- Sean $Y_1, Y_2, \ldots$ vectores n-dimensionales independientes e idénticamente distribuidos con medias $E(Y_\alpha) = \nu$ y covarianzas $E[(Y_\alpha - \nu)(Y_\alpha - \nu)'] = T$. Entonces la distribución límite de $(1/\sqrt{n}) \sum_{\alpha=1}^{n} (Y_\alpha - \nu)$ cuando $n \to \infty$ es $N(0, T)$.

Demostración.- Sea

$$\phi_n(t,u) = E\left(\exp\left[iut' \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (Y_\alpha - \nu)\right]\right),$$

donde u es un escalar y t un vector de n componentes. Para valores de t fijos, $\phi_n(t,u)$ puede ser considerada como la función característica de $(1/\sqrt{n}) \sum_{\alpha=1}^{n} (t'Y_\alpha - E(t'Y_\alpha))$. Por el teorema central del límite (para variables unidimensionales), la distribución límite correspondiente será $N(0, t'Tt)$. Por tanto

$$\lim_{n\to\infty} \phi_n(t,u) = e^{-\frac{1}{2}u^2 t'Tt} \quad , \forall u, t$$

Entonces haciendo $u=1$, tenemos

$$\lim_{n\to\infty} E\left[\exp\left(it' \frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (Y_\alpha - \nu)\right)\right] = e^{-\frac{1}{2}t'Tt} \quad , \forall t$$

Como $e^{-\frac{1}{2}t'Tt}$ es continua para $t=0$, la convergencia es uniforme en algún entorno de cero. El teorema se cumple.

Vamos a demostrar a continuación que la distribución asintótica de la matriz de covarianzas muestral es normal.

TEOREMA.- Sea $A(n) = \sum_{\alpha=1}^{N} (X_\alpha - \bar{X})(X_\alpha - \bar{X})'$, donde $X_1, \ldots$ son todas independientes y $N(\mu, \Sigma)$ y $n = N-1$. Entonces la distribución asintótica de $B(n) = (1/\sqrt{n})[A(n) - n\Sigma]$ es normal con media 0 y covarianzas

$$E\left(b_{ij}(n) b_{kl}(n)\right) = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}.$$

Demostración.- Como demostramos anteriormente $A(n)$ se distribuye como $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, donde $Z_1, Z_2, \ldots$ se distribuyen $N(0, \Sigma)$ e independientes. Podemos ordenar los elementos de $Z_\alpha Z_\alpha'$ de forma vectorial como sigue

$$Y_\alpha' = [Z_{1\alpha}^2, Z_{1\alpha}Z_{2\alpha}, \ldots, Z_{2\alpha}^2, \ldots, Z_{p\alpha}^2]$$

los momentos Y_α pueden deducirse de los momentos de Z_α tal como se obtuvieron en su momento. Tenemos $E(z_{i\alpha} z_{j\alpha}) = \sigma_{ij}$, $E(z_{i\alpha} z_{j\alpha} z_{k\alpha} z_{l\alpha}) = \sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$, $E[(z_{i\alpha} z_{j\alpha} - \sigma_{ij})(z_{k\alpha} z_{l\alpha} - \sigma_{kl})] = \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}$ (se han obtenido mediante la función característica). Así, los vectores Y_α satisfacen las condiciones del teorema anterior siendo los elementos de ν los elementos de Σ ordenados de forma similar a como lo hemos hecho con Y_α y los elementos de T los que acabamos de obtener. Si los elementos de $A(n)$ los ordenamos también en forma vectorial, $W(n)$, entonces $W(n) - n\nu = \sum_{\alpha=1}^{n} (Y_\alpha - \nu)$. Aplicando ahora el teorema anterior $(1/\sqrt{n})[W(n) - n\nu]$ se distribuye asintóticamente normal con media 0 y la matriz de covarianzas la de Y_α.

Estamos particularmente interesados en la correlación muestral

$$r(n) = \frac{A_{ij}(n)}{\sqrt{A_{ii}(n)\, A_{jj}(n)}} \quad , \quad i \neq j$$

Esto puede también escribirse de la forma

$$r(n) = \frac{C_{ij}(n)}{\sqrt{C_{ii}(n) \cdot C_{jj}(n)}}$$

donde $c_{gh}(n) = A_{gh}(n) / \sqrt{\sigma_{gg}\sigma_{hh}}$. El conjunto $C_{ii}(n)$, $C_{jj}(n)$ y $C_{ij}(n)$ se distribuye como

$$\sum_{\alpha=1}^{n} \begin{pmatrix} z_{i\alpha}^* \\ z_{j\alpha}^* \end{pmatrix} (z_{i\alpha}^* \; z_{j\alpha}^*) = \sum_{\alpha=1}^{n} \begin{pmatrix} z_{i\alpha}/\sqrt{\sigma_{ii}} \\ z_{j\alpha}/\sqrt{\sigma_{jj}} \end{pmatrix} \left(z_{i\alpha}/\sqrt{\sigma_{ii}} \, , \, z_{j\alpha}/\sqrt{\sigma_{jj}} \right)$$

donde $(z_{i\alpha}^*, z_{j\alpha}^*)$ son independientes, cada una con distribución $N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right]$ y $\rho = \sigma_{ij}/\sqrt{\sigma_{ii}\sigma_{jj}}$. Sean

$$U(n) = \frac{1}{n} \begin{pmatrix} C_{ii}(n) \\ C_{jj}(n) \\ C_{ij}(n) \end{pmatrix} \qquad \text{y} \qquad b = \begin{pmatrix} 1 \\ 1 \\ \rho \end{pmatrix}$$

Entonces $\sqrt{n}\,(U(n) - b)$ es asintóticamente normal con media 0 y matriz de covarianzas

$$\begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1+\rho^2 \end{pmatrix}$$

Necesitamos ahora el siguiente teorema general

TEOREMA.- Sea $U(n)$ un vector con m componentes y b un vector fijo. Supongamos que $\sqrt{n}\,(U(n) - b)$ se distribuye asintóticamente $N(0, T)$. Sea $W = f(U)$ una función de un vector U cuyas primeras y segundas derivadas existen en un entorno de $U = b$. Sea $\left.\frac{\partial f(U)}{\partial u_i}\right|_{U=b}$ la i-ésima componente de ϕ_b. Entonces la distribución límite de $\sqrt{n}\,(f(U(n)) - f(b))$ es

$$N(0, \phi_b' T \phi_b).$$

La demostración del teorema puede consultarse en el texto de Cramer.

Desde luego el $U(n)$ antes definido, con b y T los correspondientes valores antes introducidos, satisface las condiciones del teorema. La función

$$r = \frac{u_3}{\sqrt{u_1 \cdot u_2}} = u_3 u_1^{-1/2} u_2^{-1/2}$$

satisface las condiciones de derivabilidad exigidas. Los elementos de ϕ_b son

$$\left.\frac{\partial r}{\partial u_1}\right|_{u=b} = \left. -\frac{1}{2} u_3 u_1^{-3/2} u_2^{-1/2} \right|_{u=b} = -\frac{1}{2}\rho \quad , \quad \left.\frac{\partial r}{\partial u_2}\right|_{u=b} = \left. -\frac{1}{2} u_3 u_1^{-1/2} u_2^{-3/2} \right|_{u=b} = -\frac{1}{2}\rho \quad , \quad \left.\frac{\partial r}{\partial u_3} = u_1^{-1/2} u_2^{-1/2} \right|_{u=b} = 1$$

y $f(\rho)=\rho$. La varianza asintótica de $\sqrt{n}(r(n)-\rho)$ es

$$\left(-\tfrac{1}{2}\rho,\ -\tfrac{1}{2}\rho,\ 1\right)\begin{pmatrix} 2 & 2\rho^2 & 2\rho \\ 2\rho^2 & 2 & 2\rho \\ 2\rho & 2\rho & 1-\rho^2 \end{pmatrix}\begin{pmatrix} -\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1 \end{pmatrix} = (\rho-\rho^3,\ \rho-\rho^3,\ 1-\rho^2)\begin{pmatrix} -\frac{1}{2}\rho \\ -\frac{1}{2}\rho \\ 1 \end{pmatrix} = 1-2\rho^2+\rho^4 = (1-\rho^2)^2.$$

Hemos obtenido el siguiente teorema:

TEOREMA.- Si $r(n)$ es el coeficiente de correlación muestral de una muestra de tamaño $N=(n+1)$ de una distribución normal con correlación ρ, entonces $\sqrt{n}\,(r(n)-\rho)/(1-\rho^2)$ [o $\sqrt{N}\,(r(n)-\rho)/(1-\rho^2)$] se distribuye asintóticamente $N(0,1)$.

Apoyandonos en el teorema anterior, si $f(x)$ es una función con primera y segunda derivadas en $x=\rho$, entonces $\sqrt{n}\,(f(r)-f(\rho))$ también es asintóticamente normal con media 0 y varianza $\left(\frac{df}{dx}\Big|_{x=\rho}\right)^2(1-\rho^2)^2$.

Es usual considerar f una función cuya varianza asintótica sea constante (la unidad) independiente del parámetro ρ. Una función de este tipo satisface la ecuación

$$f'(\rho)=\frac{1}{1-\rho^2}=\frac{1}{2}\left(\frac{1}{1+\rho}+\frac{1}{1-\rho}\right)$$

Podemos tomar $f(\rho)=\frac{1}{2}[\log(1+\rho)-\log(1-\rho)]=\frac{1}{2}\log\left[\frac{1+\rho}{1-\rho}\right]$. La llamada z de Fisher es

$$z=\frac{1}{2}\log\frac{1+r}{1-r} \qquad \text{y sea} \qquad \zeta=\frac{1}{2}\log\frac{1+\rho}{1-\rho}.$$

TEOREMA.- Sea z definida como antes, donde r es el coeficiente de correlación muestral de una muestra de tamaño $N=n+1$, de una normal bivariante con correlación ρ, sea ζ el valor antes definido. Entonces $\sqrt{n}(z-\zeta)$ es asintóticamente normal con media 0 y varianza 1.

Se puede demostrar que aproximadamente

$$E(z)\sim\zeta+\frac{\rho}{2n} \qquad E(z-\zeta)^2\sim\frac{1}{n-2}\sim E\left(z-\zeta-\frac{\rho}{2n}\right)^2$$

Esta última aproximación se sigue de la expresión

$$E(z-\zeta)^2=\frac{1}{n}+\frac{8-\rho^2}{4n^2}+\cdots$$

y es muy buena para valores de ρ^2/n^2 pequeños. Hotelling (1953) obtuvo los momentos de z hasta el orden n^{-3}. Una importante propiedad de la z de Fisher es que la aproximación a la normal es mucho más rápida que para r. David (1938) estableció comparaciones entre las probabilidades tabuladas y las probabilidades obtenidas suponiendo que z se distribuía normalmente. Recomendó que para $N>25$ se tome z como distribuido normalmente con media y varianzas los dadas anteriormente.

Antes de finalizar con los coeficientes de correlación daremos algunas pautas para la utilización de la z de Fisher.

a) Supongamos que deseamos contrastar la hipótesis $\rho=\rho_0$ sobre la base de una muestra de tamaño N frente a la alternativa $\rho\neq\rho_0$. Calculemos r y después z. Sea

$$\zeta_0=\frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}$$

La región de rechazo al 5% vendrá dada por

$$\sqrt{N-3}\,|z-\zeta_0|>1.96$$

Una región mejor es

$$\sqrt{N-3}\,\left|z-\zeta_0-\tfrac{1}{2}\rho_0/(N-1)\right|>1.96$$

b) Supongamos que tenemos una muestra de tamaño N_1 de una población y una muestra N_2 de otra población. Como contrastar la hipótesis de que los dos coeficientes de correlación son iguales, $\rho_1=\rho_2$? Del teorema de la z de Fisher sabemos que si la hipótesis nula es cierta z_1-z_2 es asintóticamente normal con media 0 y varianza $1/(N_1-3)+1/(N_2-3)$. Como región crítica al 5% utilizaremos

$$\frac{|z_1-z_2|}{\sqrt{1/(N_1-3)+1/(N_2-3)}} > 1.96$$

c) Bajo las condiciones del apartado anterior supongamos que $\rho_1=\rho_2=\rho$. Como utilizar los resultados de ambas muestras para estimar conjuntamente ρ? Puesto que z_1 y z_2 tienen varianzas $1/(N_1-3)$ y $1/(N_2-3)$ respectivamente, podemos estimar ζ mediante

$$\frac{(N_1-3)z_1+(N_2-3)z_2}{N_1+N_2-6}$$

y convertir esto en un estimador de ρ mediante la transformación inversa de la Z.

d) Sea r el coeficiente de correlación muestral de N observaciones. Como obtener un intervalo de confianza para ρ? Sabemos que aproximadamente

$$Pr\{-1.96 < \sqrt{N-3}\,(z-\zeta) < 1.96\} = 0.95$$

De aquí deducimos que $[-1.96/\sqrt{N-3}+z,\ 1.96/\sqrt{N-3}+z]$ es un intervalo de confianza para ζ. A partir de aquí obtenemos la región para ρ utilizando la transformación inversa $\rho=\tanh\zeta = (e^{\zeta}-e^{-\zeta})/(e^{\zeta}+e^{-\zeta})$ que es una transformación monotónica. La región es

$$\tanh\left(z-1.96/\sqrt{N-3}\right) < \rho < \tanh\left(z+1.96/\sqrt{N-3}\right).$$

3. COEFICIENTES DE CORRELACION PARCIAL

Recordemos que estos coeficientes son coeficientes de correlación en distribuciones condicionales. Como vimos anteriormente si X se distribuye $N(\mu,\Sigma)$ la distribución condicional del subvector $X^{(1)}$ dado $X^{(2)}=x^{(2)}$, viene dada por $N\left(\mu^{(1)}+\beta(x^{(2)}-\mu^{(2)}),\ \Sigma_{11.2}\right)$, donde

$$\beta = \Sigma_{12}\Sigma_{22}^{-1}$$
$$\Sigma_{11.2} = \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Las correlaciones parciales de $X^{(1)}$ dado $x^{(2)}$ son las correlaciones calculadas de la forma ya establecida a partir de $\Sigma_{11.2}$. Veamos como llevar a cabo la estimación de estos valores y como, a partir de ellos, podremos efectuar contraste de hipótesis.

En primer lugar nos vamos a ocupar del problema de la estimación. Supongamos que tenemos una muestra de tamaño N de una normal $N(\mu,\Sigma)$. Cuales serán las estimaciones máximo-verosímiles de las correlaciones parciales de $X^{(1)}$, $\rho_{ij\cdot q+1,\dots,p}$? Sabemos que el estimador máximo-verosímil de Σ es

$$\hat{\Sigma} = \frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})' \quad \text{con} \quad \bar{x}=\frac{1}{N}\sum_{\alpha=1}^{N}x_\alpha .$$

La independencia entre Σ y $\Sigma_{11.2}$ y β y Σ_{22} es uno a uno en virtud de las relaciones (definiciones) anteriores y de

$$\Sigma_{12} = \beta\Sigma_{22}$$
$$\Sigma_{11} = \Sigma_{11.2}+\beta\Sigma_{22}\beta'$$

por tanto, teniendo en cuenta una propiedad de los estimadores máximo-verosímiles podemos afirmar que los correspondientes estimadores de $\Sigma_{11.2}$ y β y Σ_{22} vendrán dados por $\hat{\Sigma}_{11.2}=\hat{\Sigma}_{11}-\hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}$, $\hat{\beta}=\hat{\Sigma}_{12}\cdot\hat{\Sigma}_{22}^{-1}$ y $\hat{\Sigma}_{22}$. Además los estimadores de los coeficientes de correlación parcial serán:

$$\hat{r}_{ij\cdot q+1,\dots,p} = \frac{\hat{\sigma}_{ij\cdot q+1,\dots,p}}{\sqrt{\hat{\sigma}_{ii\cdot q+1,\dots,p}\cdot\hat{\sigma}_{jj\cdot q+1,\dots,p}}} \qquad i,j = 1,\dots,q$$

donde $\hat{\sigma}_{ij\cdot q+1,\dots,p}$ es el ij-ésimo elemento de $\hat{\Sigma}_{11\cdot 2}$.

TEOREMA.- Sea $x_1,\dots,x_N$ una muestra de tamaño N de una $N(\mu,\Sigma)$. Los estimadores máximo-verosímiles de $\rho_{ij\cdot q+1,\dots,p}$, correlaciones parciales de las primeras q componentes condicionadas a las $p-q$ restantes, vienen dadas por

$$\hat{r}_{ij\cdot q+1,\dots,p} = \frac{a_{ij\,q+1,\dots,p}}{\sqrt{a_{ii\cdot q+1,\dots,p}\cdot a_{jj\cdot q+1,\dots,p}}}$$

donde

$$\left(a_{ij\cdot q+1,\dots,p}\right) = A_{11} - A_{12}A_{22}^{-1}A_{21} \quad \text{y}$$

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'.$$

La estimación $\hat{\rho}_{ij\cdot q+1,\dots,p}$, que denotaremos por $r_{ij\cdot q+1,\dots,p}$ es conocida como el coeficiente de correlación muestral entre X_i y X_j cuando $X_{q+1},\dots,X_p$ permanecen constante.

Dos interpretaciones geométricas pueden darse para estos resultados. A saber:

a) En un espacio p-dimensional consideramos los N puntos $x_1,\dots,x_N$, datos muestrales. La función de regresión

$$x^{(1)} = \bar{x}^{(1)} + \hat{\beta}\,(x^{(2)} - \bar{x}^{(2)})$$

es un hiperplano $p-q$ dimensional que resulta de la intersección de los q hiperplanos $p-1$ dimensionales

$$x_i = \bar{x}_i + \sum_{j=q+1}^{p} \hat{\beta}_{ij}(x_j - \bar{x}_j) \qquad i = 1,\dots,q$$

Aquí $\hat{\beta}_{ij}$ es un elemento de $\hat{\beta} = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1} = A_{12}A_{22}^{-1}$. La i-ésima fila de $\hat{\beta}$ es $[\hat{\beta}_{i\,q+1},\dots,\hat{\beta}_{ip}]$. La parte derecha de la anterior igualdad es la función de regresión mínimo-cuadrática de x_i sobre $x_{q+1},\dots,x_p$; sobre, si proyectamos los puntos $x_1,\dots,x_N$ en el hiperplano de las coordenadas $x_i, x_{q+1},\dots,x_p$, entonces la expresión anterior es el plano de regresión. El punto de coordenadas

$$x_i = \bar{x}_i + \sum_{j=q+1}^{N} \hat{\beta}_{ij}(x_{j\alpha} - \bar{x}_j) \qquad i = 1,\dots,q$$

$$x_j = x_{j\alpha}, \qquad j = q+1,\dots,p$$

está en el hiperplano anterior. La diferencia entre la i-ésima coordenada de x_α y el punto anterior es $y_{i\alpha} = x_{i\alpha} - \left[\bar{x}_i + \sum_{j=q+1}^{N} \hat{\beta}_{ij}(x_{j\alpha} - \bar{x}_j)\right]$ para $i = 1,\dots,q$ y 0 para las otras coordenadas. Pongamos

$$Y_\alpha = \begin{bmatrix} y_{1\alpha} \\ \vdots \\ y_{q\alpha} \end{bmatrix}.$$

Estos puntos pueden representarse como N puntos en un espacio q-dimensional. Entonces $A_{11\cdot 2} = \sum_{\alpha=1}^{N} Y_\alpha Y_\alpha'$.

b) Podemos, por contra, interpretar la muestra como p puntos en un espacio N-dimensional. Sea $Z_j = (x_{j1},\dots,x_{jN})$ el j-ésimo punto (las j-ésimas coordenadas de la muestra), y hagamos $Z_{p+1} = (1,\dots,1)$ para el $p+1$-ésimo punto. El punto de coordenadas $\bar{x}_1,\dots,\bar{x}_i$ es $\bar{x}_i Z_{p+1}$. La proyección de Z_i sobre el hiperplano definido mediante $Z_{q+1},\dots,Z_{p+1}$ es

$$Z_i^* = \bar{x}_i Z_{p+1} + \sum_{j=q+1}^{p} \hat{\beta}_{ij}(Z_j - \bar{x}_j Z_{p+1})$$

y sea el punto del hiperplano cuya distancia a z_i sea mínima. Sea $\tilde{z}_i$ el vector de z_i^* a z_i, es decir $z_i - z_i^*$, o mejor un equivalente trasladado de manera que su origen coincida con el origen de coordenadas. El conjunto de vectores $\tilde{z}_1, \dots, \tilde{z}_q$ son las proyecciones de $z_1, \dots, z_q$ en el hiperplano ortogonal a $z_{q+1}, \dots, z_p$. Entonces $\tilde{z}_i' \tilde{z}_i = a_{ii \cdot q+1, \dots, p}$ es el cuadrado de la longitud de $\tilde{z}_i$, o lo que es lo mismo, el cuadrado de la distancia de z_i^* a z_i. Así $\tilde{z}_i \tilde{z}_j / \sqrt{\tilde{z}_i' \tilde{z}_i \, \tilde{z}_j' \tilde{z}_j} = r_{ij \cdot q+1, \dots, p}$ es el coseno del ángulo entre $\tilde{z}_i$ y $\tilde{z}_j$.

Un ejemplo de coeficiente de correlación parcial

Consideremos algunos datos [(Hooker, 1907) J. Roy. Stat. Soc. Vol 70. pp. 1-42] acerca de la cosecha de heno (x_1) en cientos de libras de por acre, cantidad de lluvia caída durante la primavera (x_2) y temperaturas acumuladas superiores a 42°F en la primavera (x_3) para una determinada zona de Inglaterra durante un período de 20 años. Las estimaciones de $\mu_i, \sigma_i \, (= \sqrt{\sigma_{ii}})$ y ρ_{ij} son

$$\hat{\mu} = \bar{x} = \begin{bmatrix} 28.02 \\ 4.91 \\ 594 \end{bmatrix}, \quad \begin{bmatrix} \hat{\sigma}_1 \\ \hat{\sigma}_2 \\ \hat{\sigma}_3 \end{bmatrix} = \begin{bmatrix} 4.42 \\ 1.10 \\ 85 \end{bmatrix}, \quad \begin{bmatrix} 1 & \hat{\rho}_{12} & \hat{\rho}_{13} \\ \hat{\rho}_{21} & 1 & \hat{\rho}_{23} \\ \hat{\rho}_{13} & \hat{\rho}_{23} & 1 \end{bmatrix} = \begin{bmatrix} 1.00 & 0.80 & -0.40 \\ 0.80 & 1.00 & -0.56 \\ -0.40 & -0.56 & 1.00 \end{bmatrix}$$

A partir de las correlaciones observamos que la cantidad de lluvia caída (x_2) y la cosecha (x_1) están relacionadas positivamente, cosecha y temperatura lo están negativamente al igual que lluvia y temperatura. ¿Qué interpretación puede darse a la relación negativa aparente que existe entre la cosecha y la temperatura? ¿Tienden las altas temperaturas a causar bajas cosechas o las altas temperaturas están asociadas con bajas cantidades de lluvia y por tanto con bajas cosechas? Para responder a estas cuestiones podemos considerar la correlación entre cosecha y temperatura cuando la cantidad de lluvia es constante; es decir, se trata de utilizar los datos anteriores para estimar el coeficiente de correlación parcial entre x_1 y x_3. Su valor es

$$\frac{\hat{\sigma}_{13 \cdot 2}}{\sqrt{\hat{\sigma}_{11 \cdot 2} \cdot \hat{\sigma}_{33 \cdot 2}}} = 0.097$$

Así, al efecto de la lluvia eliminado, cosecha y temperatura vuelven relacionadas positivamente. La conclusión es que antes, altas lluvias y altas temperaturas aumentan las cosechas de heno, pero en casi todos los años observados las grandes lluvias han venido con bajas temperaturas y viceversa.

La distribución del coeficiente de correlación parcial

Las correlaciones parciales se obtienen a partir de $A_{11 \cdot 2} = A_{11} - A_{12} A_{22}^{-1} A_{21}$ de la misma forma mediante la que obtuvimos las correlaciones de A. Para obtener la distribución de las correlaciones demostramos que A se distribuía como $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$ donde los Z_α eran independientes y cada una se ellas $N(0, \Sigma)$. Actuaremos de forma similar ahora. Para ello demostraremos a continuación un teorema que será también de utilidad posteriormente.

Teorema. - Supongamos que $Y_1, \dots, Y_m$ son independientes con Y_α distribuida como $N(\Gamma w_\alpha, \Phi)$, donde w_α es un vector de r componentes. Sea $G = \sum_\alpha Y_\alpha w_\alpha' H^{-1}$ donde $H = \sum_\alpha w_\alpha w_\alpha'$ y es no singular. Entonces $\sum_{\alpha=1}^{m} Y_\alpha Y_\alpha' - GHG'$ se distribuye como $\sum_{\alpha=1}^{m-r} U_\alpha U_\alpha'$, donde U_α son independientes y $N(0, \Phi)$, y además independientes de G.

Demostración. - Sea $W = (w_1, \dots, w_m)$ y sea F una matriz cuadrada tal que $FHF' = I$, entonces $F'^{-1} H^{-1} F^{-1} = I$. Sea $E_2 = FW$, entonces $W = E_2 F^{-1}$. Así

$$E_2 E_2' = FWW'F = F \sum_\alpha w_\alpha w_\alpha' F' = FHF' = I.$$

Resulta que las filas de E_2 son vectores ortonormales. Será posible encontrar una matriz $E_1, (m-r) \times m$ tal que $E' = [E_1' \; E_2']$ sea ortogonal. Sea ahora $Y = (Y_1, \dots, Y_m) = UE$, o $U = YE'$ (es decir, $U_\alpha = \sum_\beta e_{\alpha\beta} Y_\beta$). Aplicando un conocido teorema, las columnas de U, es decir U_α, son independientes y distribuidas normalmente, cada una de ellas con covarianza Φ. La media vendrá dada por

$$E(U) = E(YE') = \Gamma W E' = \Gamma F^{-1} E_2 (E_1' E_2') = \left(\Gamma F^{-1} E_2 E_1' \quad \Gamma F^{-1} E_2 E_2'\right) = (0 \quad \Gamma F^{-1})$$

Para completar la demostración necesitamos demostrar que

$$\sum Y_\alpha Y_\alpha' - GHG' = \sum_{\alpha=1}^{m-r} U_\alpha U_\alpha' .$$

Sabemos que $\sum_{\alpha=1}^{m} Y_\alpha Y_\alpha' = \sum_{\alpha=1}^{m} U_\alpha U_\alpha'$ por ser la transformación que pasa de unos a otros, ortogonal. También se verifica

$$GHG' = (YW'H^{-1}) H (H^{-1} W Y') = UEE_2' (F^{-1})' H^{-1} F^{-1} E_2 E' U' = U \begin{pmatrix} E_1 \\ E_2 \end{pmatrix} E_2' E_2 (E_1' E_2') U' =$$

$$= U \begin{pmatrix} 0 \\ I \end{pmatrix} (0 \; I) U' = \sum_{\alpha=m-r+1}^{m} U_\alpha U_\alpha' .$$

Así pues

$$\sum_{\alpha=1}^{m} Y_\alpha Y_\alpha' - GHG' = \sum_{\alpha=1}^{m} U_\alpha U_\alpha' - \sum_{\alpha=m-r+1}^{m} U_\alpha U_\alpha' = \sum_{\alpha=1}^{m-r} U_\alpha U_\alpha' .$$

De las anteriores consideraciones se sigue que si $\Gamma = 0$, $E(U) = 0$, y obtenemos el siguiente Corolario

COROLARIO.- Si $\Gamma = 0$, la matriz GHG' del teorema anterior se distribuye como $\sum_{\alpha=m-r+1}^{m} U_\alpha U_\alpha'$ donde los U_α son independientes y distribuidas $N(0, \phi)$.

Podemos ahora encontrar la distribución de $A_{11.2}$ de la misma forma. Hemos que A se distribuía como $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$, donde los Z_α son independientes, cada una $N(0, \Sigma)$. Si fraccionamos Z_α en dos subvectores de dimensiones q y $p-q$ respectivamente,

$$Z_\alpha = \begin{pmatrix} Z_\alpha^{(1)} \\ Z_\alpha^{(2)} \end{pmatrix} .$$

Entonces $A_{ij} = \sum_{\alpha=1}^{N-1} Z_\alpha^{(i)} Z_\alpha^{(j)'}$. La distribución condicional de $Z_1^{(1)} \ldots Z_{N-1}^{(1)}$ dado $Z_1^{(2)} = z_1^{(2)}, \ldots, Z_{N-1}^{(2)} = z_{N-1}^{(2)}$ es

$$\frac{\prod_{\alpha=1}^{N-1} n(z_\alpha | 0, \Sigma)}{\prod_{\alpha=1}^{N-1} n(z_\alpha^{(2)} | 0, \Sigma)} = \prod_{\alpha=1}^{N-1} \frac{n(z_\alpha | 0, \Sigma)}{n(z_\alpha^{(2)} | 0, \Sigma)} = \prod_{\alpha=1}^{N-1} n\left(z_\alpha^{(1)} \middle| \beta z_\alpha^{(2)}, \Sigma_{11.2}\right)$$

donde $\beta = \Sigma_{12} \Sigma_{22}^{-1}$ y $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$. Aplicando ahora el teorema con $Z_\alpha^{(1)} = Y_\alpha$, $Z_\alpha^{(2)} = W_\alpha$, $N-1 = m$, $p-q = r$, $\beta = \Gamma$, $\Sigma_{11.2} = \phi$, $A_{11} = \sum Y_\alpha Y_\alpha'$, $A_{12} A_{22}^{-1} = G$, $A_{22} = H$. Vemos que la distribución condicional de $A_{11} - A_{12} A_{22}^{-1} A_{21} = A_{11} - (A_{12} A_{22}^{-1}) A_{22} (A_{22}^{-1} A_{12}) = A_{11.2}$ dado $Z_\alpha^{(2)} = z_\alpha^{(2)}$ es la de $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$ donde los U_α son independientes, cada uno con distribución $N(0, \Sigma_{11.2})$. Como esta distribución no depende de $\{z_\alpha^{(2)}\}$ entonces obtenemos finalmente:

TEOREMA.- La matriz $A_{11.2} = A_{11} - A_{12} A_{22}^{-1} A_{21}$ se distribuye como $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$, donde los U_α se distribuyen independientes, cada uno con distribución $N(0, \Sigma_{11.2})$

Como corolario obtenemos

COROLARIO.- Si $\Sigma_{12} = 0$ $(\beta = 0)$ entonces $A_{11.2}$ se distribuye como $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$ y $A_{12} A_{22}^{-1} A_{21}$ se distribuye como $\sum_{\alpha=N-(p-q)}^{N-1} U_\alpha U_\alpha'$ donde los U_α son independientes y $N(0, \Sigma_{11.2})$.

Se sigue de estos resultados que la distribución de $r_{ij.q+1,\ldots,p}$ basado en N observaciones es la misma que la de un coeficiente de correlación ordinario basado en $N-(p-q)$ observaciones con el correspondiente coeficiente de correlación poblacional dado por $\rho_{ij.q+1,\ldots,p}$. Enunciamos por fin el teorema

TEOREMA.- Si la función de distribución de r_{ij} basado en una muestra de tamaño N de una $N(0,\Sigma)$ con coeficiente de correlación poblacional parcial ρ_{ij} la designamos por $F(r|N,\rho_{ij})$, entonces la función de distribución del coeficiente de correlación parcial muestral $r_{ij\cdot q+1,\dots p}$ basado en una muestra de tamaño N de la misma normal es $F(r|N-(p-q),\rho_{ij\cdot q+1,\dots p})$.

Esta función de distribución fue obtenida por Fisher en el año 1924.

Test de hipótesis y regiones de confianza para coeficiente de correlación parcial.

Puesto que la distribución de un coeficiente de correlación parcial $r_{ij\cdot q+1,\dots p}$ basado en una muestra de tamaño N de una distribución con correlación poblacional $\rho_{ij\cdot q+1,\dots p}$ igual a cierto valor ρ, es decir, que la distribución de un coeficiente de correlación ordinario r basado en una muestra de tamaño $N-(p-q)$ de una distribución con el correspondiente coeficiente de correlación poblacional ρ. Todos los procedimientos de inferencia estadística antes descritos pueden utilizarse ahora. Se actúa de forma semejante pero reemplazando N por $N-(p-q)$. Veamos dos ejemplos de esta actuación.

Ejemplo 1. Supongamos que sobre la base de una muestra de tamaño N deseamos obtener un intervalo de confianza para $\rho_{ij\cdot q+1,\dots p}$. El coeficiente de correlación parcial muestral es $r_{ij\cdot q+1,\dots p}$. El procedimiento será utilizar las tablas de David para $N-(p-q)$. Por ejemplo para la ilustración de la cosecha de heno antes dada, si queremos construir un intervalo para $\rho_{12.3}$ con coeficiente de confianza 0.95, como $r_{12.3}=0.79$, utilizando la tabla $N-(p-q)=20-(2-1)=19$, tenemos $0.52<\rho_{12.3}<0.92$.

Ejemplo 2. Análogamente, si queremos utilizar la z de Fisher para contrastar un contraste de hipótesis para $\rho_{ij\cdot q+1,\dots p}=\rho_0$, frente a una alternativa bilateral, tenemos

$$z=\frac{1}{2}\log\frac{1+r_{ij\cdot q+1,\dots p}}{1-r_{ij\cdot q+1,\dots p}}$$

$$\xi_0=\frac{1}{2}\log\frac{1+\rho_0}{1-\rho_0}$$

Entonces $\sqrt{N-(p-q)-3}\,(z-\xi_0)$ se comparan con los puntos de significación de una normal standarizada. En el ejemplo de la cosecha de heno, si quisiéramos contrastar la hipótesis $\rho_{13.2}=0$, al nivel $\alpha=0.05$, tendríamos $\xi_0=0$ y $\sqrt{20-1-3}\,(0.0973)=0.3892$. y como $|0.3892|<1.96$ aceptaríamos la hipótesis.

4. EL COEFICIENTE DE CORRELACION MULTIPLE

Estimación de la correlación multiple

El coeficiente de correlación multiple poblacional lo estudiamos en capítulos anteriores. Por razones de conveniencia estudiaremos la estimación para el caso X_1 por un lado y $X_2,\dots,X_p$ por otro, no pondremos ningún índice alguno a R. Esta actuación no supone pérdida de generalidad por cuanto las variables pueden enumerarse para hacer válido lo obtenido ahora en cualquier otra situación. Recordemos que la expresión para el coeficiente de correlación multiple poblacional venía dada por

$$R=\frac{\beta\Sigma_{22}\beta'}{\sqrt{\sigma_{11}\beta\Sigma_{22}\beta'}}=\sqrt{\frac{\beta\Sigma_{22}\beta'}{\sigma_{11}}}=\sqrt{\frac{\sigma_{(1)}\Sigma_{22}^{-1}\sigma_{(1)}'}{\sigma_{11}}}$$

donde β, $\sigma_{(1)}$ y Σ_{22} son

$$\Sigma=\begin{pmatrix}\sigma_{11} & \sigma_{(1)}\\ \sigma_{(1)}' & \Sigma_{22}\end{pmatrix}\qquad \beta=\sigma_{(1)}\Sigma_{22}^{-1}$$

Dada una muestra $x_1,\dots,x_N$ $(N>p)$ estimamos Σ mediante $S=\left(N/(N-1)\right)\hat{\Sigma}$ o

$$\hat{\Sigma}=\frac{1}{N}A=\frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha-\bar{x})(x_\alpha-\bar{x})'=\begin{pmatrix}\hat{\sigma}_{11} & \hat{\sigma}_{(1)}\\ \hat{\sigma}_{(1)}' & \hat{\Sigma}_{22}\end{pmatrix}$$

y estimamos β mediante $\hat{\beta}=\hat{\sigma}_{(1)}\hat{\Sigma}_{22}^{-1}=a_{(1)}A_{22}^{-1}$. Definimos entonces el coeficiente de correlación multiple muestral como

$$R = \sqrt{\frac{\hat\beta\,\hat\Sigma_{22}\,\hat\beta'}{\hat\sigma_{11}}} = \sqrt{\frac{\hat\sigma_{(1)}\,\hat\Sigma_{22}^{-1}\,\hat\sigma_{(1)}'}{\hat\sigma_{11}}} = \sqrt{\frac{a_{(1)}\,A_{22}^{-1}\,a_{(1)}'}{a_{11}}}$$

Que el resultado de la estimación máximo-verosímil de $\bar R$ puede justificarse aplicando un corolario de un teorema presentado en capítulos anteriores referente a las estimaciones de funciones de los parámetros, y teniendo en cuenta que $\bar R$ $\bar\sigma_{(1)}$ y Σ_{22} pueden definirse mediante una transformación uno a uno de Σ. Otra expresión para R es

$$1-R^2 = \frac{|\hat\Sigma|}{\hat\sigma_{11}\,|\hat\Sigma_{22}|} = \frac{|A|}{a_{11}\,|A_{22}|}$$

R y $\hat\beta$ tienen en la muestra las mismas propiedades que $\bar R$ y β poseían en la población. Por ejemplo, de todos los vectores d con $(p-1)$ componentes que definen una combinación lineal $d\,x_\alpha^{(2)}$ de las componentes de $x_\alpha^{(2)}$, el vector $d=\hat\beta$ es el que minimiza $\sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar x_1)-d\,(x_\alpha^{(2)}-\bar x^{(2)})\right]^2$. En efecto, en primer lugar observemos que

$$\sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar x_1)-\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})\right]\left[x_\alpha^{(2)}-\bar x^{(2)}\right]' = a_{(1)}-\hat\beta\,A_{22} = 0$$

pues $\hat\beta = a_{(1)}\,A_{22}^{-1}$. Así

$$\sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar x_1)-d\,(x_\alpha^{(2)}-\bar x^{(2)})\right]^2 = \sum_{\alpha=1}^{N}\left\{\left[(x_{1\alpha}-\bar x_1)-\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})\right]+(\hat\beta-d)\,(x_\alpha^{(2)}-\bar x^{(2)})\right\}^2 =$$

$$= \sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar x_1)-\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})\right]^2 + (\hat\beta-d)\,A_{22}\,(\hat\beta-d)'$$

y como A_{22} es definida positiva (excepto para muestras que ocurren con probabilidad cero), el mínimo ocurrirá cuando $\hat\beta-d=0$. Este mínimo es

$$\sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar x_1)-\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})\right]^2 = a_{11}-2\hat\beta\sum_{\alpha=1}^{N}(x_\alpha^{(2)}-\bar x^{(2)})(x_{1\alpha}-\bar x_1)' + \hat\beta\,A_{22}\,\hat\beta' =$$

$$= a_{11}-2\,\hat\beta\,a_{(1)}' + a_{(1)}A_{22}^{-1}A_{22}A_{22}^{-1}a_{(1)}' = a_{11}-2\,a_{(1)}A_{22}^{-1}a_{(1)}' + a_{(1)}A_{22}^{-1}a_{(1)}' =$$

$$= a_{11}-\hat\beta\,A_{22}^{-1}\hat\beta' = a_{11}-a_{(1)}A_{22}^{-1}a_{(1)}' = a_{11.2}$$ (varianza condicionada para $q=1$)

De este resultado puede darse una interesante interpretación geométrica. El vector de N componentes cuya α-ésima componente es $x_{1\alpha}-\bar x_1$ es, como en su momento vimos, la proyección del vector de las componentes i-ésimas de la muestra sobre el plano ortogonal a la línea equiangular. Tenemos pues estos vectores. $d\,(x_\alpha^{(2)}-\bar x^{(2)})$ es la α-ésima componente de un vector en el hiperplano determinado por los $p-1$ últimos vectores. Como $\sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar x_1)-d\,(x_\alpha^{(2)}-\bar x^{(2)})\right]^2$ es la distancia entre el primer vector y la combinación lineal de los $p-1$ últimos vectores, $\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})$ es una componente de un vector que minimiza esta distancia. La interpretación del hecho que $\sum_{\alpha=1}^{N}\left[(x_{1\alpha}-\bar x_1)-\beta(x_\alpha^{(2)}-\bar x^{(2)})\right]\left[(x_\alpha^{(2)}-\bar x^{(2)})\right]=0$ es que el vector con componentes $(x_{1\alpha}-\bar x_1)-\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})$ es perpendicular a cada uno de los $p-1$ últimos vectores. Así resulta que el vector $\{\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})\}_{\alpha=1}^{N}$ es la proyección del primer vector sobre el hiperplano citado. El cuadrado de la longitud del vector proyección es

$$\sum_{\alpha=1}^{N}\left[\hat\beta\,(x_\alpha^{(2)}-\bar x^{(2)})\right]^2 = \hat\beta\,A_{22}\,\hat\beta' = a_{(1)}A_{22}^{-1}a_{(1)}'$$

y la longitud del primer vector $\sum_{\alpha=1}^{N}(x_{1\alpha}-\bar x_1)^2 = a_{11}$. Así R es el coseno del ángulo formado por el primer vector con su proyección.

Vimos anteriormente que el coeficiente de correlación ordinario era el coseno del ángulo formado por los vectores correspondientes en el plano ortogonal a la línea equiangular. Otra propiedad de R es que representa la máxima correlación entre $x_{1\alpha}$ y la combinación lineal de las componentes de $x_\alpha^{(2)}$. Esto corresponde con la propiedad geométrica de que R es el coseno del ángulo más pequeño entre el vector con componentes $(x_{1\alpha}-\bar x_1)$ y un vector en el hiperplano determinado por los otros $p-1$ vectores. Este resultado puede comprobarse desde cualquier punto de vista, el geométrico citado o el analítico, apoyándose en el resultado precedente y de forma análoga a como lo hicimos con $\bar R$.

Las interpretaciones geométricas vienen dadas todas ellas enteriminos de los vectores en el hiperplano $N-1$ dimensional ortogonal a la línea equiangular. Vimos anteriormente que el vector $(x_{i1}-\bar{x}_i, \dots, x_{iN}-\bar{x}_i)$ en este hiperplano podía ser designado por $(z_{i1}, \dots, z_{i,N-1})$ donde las z_{ix} son las coordenadas referidas a un sistema de coordenadas $N-1$ dimensional en el hiperplano. Demostrabamos que las nuevas coordenadas se obtenían a partir de las antiguas mediante la transformación $x_{ix} = \sum_{\alpha=1}^{N} z_{i\beta} b_{\beta\alpha}$ donde B es una matriz ortogonal cuya ultima fila es $(1/\sqrt{N}, \dots, 1/\sqrt{N})$. Entonces

$$a_{ij} = \sum_{\alpha=1}^{N} (x_{i\alpha}-\bar{x}_i)(x_{j\alpha}-\bar{x}_j) = \sum_{\alpha=1}^{N-1} z_{i\alpha} z_{j\alpha}$$

En adelante cuando nos refiramos a R definido enterminos de z_{ix} lo llamaremos como el "coeficiente de correlación multiple sin restaber las medias".

El cálculo de R supone la extracción de la raiz cuadrada del cociente de $a_{(1)} A_{22}^{-1} a_{(1)}'$ por a_{11}. Puesto que A se obtiene directamente de las observaciones solo, probablemente, el cálculo de $a_{(1)} A_{22}^{-1} a_{(1)}'$ requerirá alguna técnica especial.

Distribución del coeficiente de correlación multiple muestral cuando el poblacional es cero.

Recordemos que

$$R^2 = \frac{a_{(1)} A_{22}^{-1} a_{(1)}'}{a_{11}}$$

entonces

$$1 - R^2 = 1 - \frac{a_{(1)} A_{22}^{-1} a_{(1)}'}{a_{11}} = \frac{a_{11} - a_{(1)} A_{22}^{-1} a_{(1)}'}{a_{11}} = \frac{a_{11.2}}{a_{11}}$$

y

$$\frac{R^2}{1-R^2} = \frac{a_{(1)} A_{22}^{-1} a_{(1)}'}{a_{11.2}}.$$

Para $q=1$ recordamos de la sección anterior (la referente al coeficiente de correlación parcial) establece que cuando $\beta=0$, es decir, cuando $\bar{R}=0$, $a_{11.2}$ se distribuye como $\sum_{\alpha=1}^{N-p} V_\alpha^2$ y $a_{(1)} A_{22}^{-1} a_{(1)}'$ se distribuye como $\sum_{\alpha=N-p+1}^{N-1} V_\alpha^2$, donde las V_α son independientes y cada una con distribución $N(0, \sigma_{11.2})$. Entonces $a_{11.2}/\sigma_{11.2}$ y $a_{(1)} A_{22}^{-1} a_{(1)}'/\sigma_{11.2}$ se distribuyen independientemente como variables χ^2 con $N-p$ y $p-1$ grados de libertad respectivamente. Entonces

$$\frac{R^2}{1-R^2} \cdot \frac{N-p}{p-1} = \frac{a_{(1)} A_{22}^{-1} a_{(1)}'/\sigma_{11.2}}{a_{11.2}/\sigma_{11.2}} \cdot \frac{N-p}{p-1} = \frac{\chi^2_{p-1}}{\chi^2_{N-p}} \cdot \frac{N-p}{p-1} = F_{p-1, N-p}$$

es decir, una F de Snedecor con $p-1$ y $N-1$ grados de libertad. La densidad de F viene dada por

$$\frac{\Gamma\left(\frac{1}{2}(N-1)\right)}{\Gamma\left(\frac{1}{2}(p-1)\right)\Gamma\left(\frac{1}{2}(N-p)\right)} \left(\frac{p-1}{N-p}\right)^{\frac{1}{2}(p-1)} \cdot x^{\frac{1}{2}(p-1)-1} \cdot \left(1 + \frac{p-1}{N-p} x\right)^{-\frac{1}{2}(N-1)}.$$

La densidad de R, definido como

$$R = \sqrt{\frac{\frac{p-1}{N-p} x}{1 + \frac{p-1}{N-p} x}}$$

, con x una $F_{p-1, N-p}$

es

$$2 \frac{\Gamma\left(\frac{1}{2}(N-1)\right)}{\Gamma\left(\frac{1}{2}(p-1)\right)\Gamma\left(\frac{1}{2}(N-p)\right)} R^{p-2} \left(1-R^2\right)^{\frac{1}{2}(N-p)-1}$$

Podemos escribir todos estos resultados en el siguiente teorema.

TEOREMA .- Sea R el coeficiente de correlación multiple muestral entre X_1 y $X^{(2)\prime} = (X_2, \ldots, X_p)$ basado en una muestra de tamaño N tomada de una $N(\mu, \Sigma)$. Si $\bar{R}=0$ (es decir, si $(\sigma_{12}, \ldots, \sigma_{1p}) = \sigma_{(1)} = 0 = \beta$), entonces $\left[R^2/(1-R^2)\right] \cdot \left[(N-p)/(p-1)\right]$ se distribuye como una F de Snedecor con $p-1$ y $N-p$ grados de libertad.

Hay que señalar que $p-1$ es el número de componentes de $X^{(2)}$ y $N-p = N-(p-1)-1$. Si el número de componentes de $X^{(2)}$ fuera q, las cantidades serían q y $N-q-1$ respectivamente.

Por otra parte la cantidad $R^2/(1-R^2)$ es la que se obtiene en teoría de la regresión para contrastar hipótesis acerca de los coeficientes de regresión, más concretamente, si la regresión de X_1 sobre $X_2 \ldots X_p$ es nula.

Si $\bar{R} \neq 0$ la distribución de R es mucho más difícil de obtener. No comparemos de ella más tarde.

Contraste de hipótesis acerca de $\bar{R}$. $(H_0: \bar{R}=0)$

Vamos ahora a considerar el problema estadístico de contrastar la hipótesis $\bar{R}=0$ sobre la base de una muestra aleatoria N extraída de una población $N(0, \Sigma)$. Puesto que $\bar{R} \geq 0$ las alternativas consideradas son $\bar{R}>0$.

Obtengamos en primer lugar el test de la razón de verosimilitud de esta hipótesis. La función de verosimilitud es

$$L(\mu^*, \Sigma^*) = \frac{1}{(2\pi)^{\frac{1}{2}Np} |\Sigma^*|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum_{\alpha} (x_\alpha - \mu^*)' \Sigma^{*-1} (x_\alpha - \mu^*)\right].$$

Sea ω la región del espacio paramétrico especificado por la hipótesis nula. El criterio de la razón de verosimilitud se basa en

$$\lambda = \frac{\max_{\mu^*, \Sigma^* \in \omega} L(\mu^*, \Sigma^*)}{\max_{\mu^*, \Sigma^* \in \Omega} L(\mu^*, \Sigma^*)}.$$

Aquí Ω es el espacio de μ^*, Σ^* definido positivo y ω es la región de este espacio donde $\bar{R} = \sqrt{\sigma_{(1)} \Sigma_{22}^{-1} \sigma_{(1)}} / \sqrt{\sigma_{11}} = 0$; es decir donde $\sigma_{(1)} \Sigma_{22}^{-1} \sigma_{(1)} = 0$. Como Σ_{22} es definida positiva esta condición equivale a $\sigma_{(1)} = 0$. El máximo de $L(\mu^*, \Sigma^*)$ sobre Ω se alcanza para $\mu^* = \hat{\mu} = \bar{x}$ y $\Sigma^* = \hat{\Sigma} = (1/N) A$ y viene dado por

$$\max_{\mu^*, \Sigma^* \in \Omega} L(\mu^*, \Sigma^*) = \frac{N^{\frac{1}{2}pN} e^{-\frac{1}{2}pN}}{(2\pi)^{\frac{1}{2}pN} |A|^{\frac{1}{2}N}}.$$

En ω, la función de verosimilitud es

$$L(\mu^*, \Sigma^* \mid \sigma_{(1)}^* = 0) = \frac{1}{(2\pi)^{\frac{1}{2}N} \sigma_{11}^{*\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum (x_{1\alpha} - \mu_1^*)^2 / \sigma_{11}^*\right] \cdot \frac{1}{(2\pi)^{\frac{1}{2}(p-1)N} |\Sigma_{22}^*|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum (x_\alpha^{(2)} - \mu^{(2)*})' \Sigma_{22}^{-1} (x_\alpha^{(2)} - \mu^{(2)*})\right]$$

El primer factor se maximiza para $\mu_1^* = \hat{\mu}_1 = \bar{x}_1$ y $\sigma_{11}^* = \hat{\sigma}_{11} = (1/N) a_{11}$, y el segundo factor se maximiza para $\mu^{(2)*} = \hat{\mu}^{(2)} = \bar{x}^{(2)}$ y $\Sigma_{22}^* = \hat{\Sigma}_{22} = (1/N) A_{22}$. El valor máximo de la función es

$$\max_{\mu^*, \Sigma^* \in \omega} L(\mu^*, \Sigma^*) = \frac{N^{\frac{1}{2}N} e^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}N} a_{11}^{\frac{1}{2}N}} \cdot \frac{N^{\frac{1}{2}(p-1)N} \cdot e^{-\frac{1}{2}(p-1)N}}{(2\pi)^{\frac{1}{2}(p-1)N} |A_{22}|^{\frac{1}{2}N}}.$$

Así, el criterio de la razón de verosimilitud será

$$\lambda = \frac{|A|^{\frac{1}{2}N}}{a_{11}^{\frac{1}{2}N} |A_{22}|^{\frac{1}{2}N}} = (1-R^2)^{\frac{1}{2}N}$$

La región crítica del test vendrá dada por $\lambda < \lambda_0$, donde λ_0 se elige de manera que la probabilidad de la desigualdad, cuando $\bar{R}=0$, sea precisamente, α, nivel de significación elegido. Un test equivalente será

$$1 - \lambda^{2/N} = R^2 > 1 - \lambda_0^{2/N}.$$

Podemos construir un test a partir de $[R^2/(1-R^2)]\cdot[(N-p)/(p-1)]$, que es una función monótona de R. Para ello, teniendo en cuenta que si $\bar{R}=0$ el cociente anterior se distribuye como $F_{p-1,N-p}$, determinaremos la región crítica mediante la desigualdad

$$\frac{R^2}{1-R^2}\cdot\frac{N-p}{p-1} > F_{p-1,N-p}(\alpha)$$

donde $F_{p-1,N-p}(\alpha)$ es el punto de significación correspondiente a un nivel α. Este test es equivalente al obtenido mediante la razón de verosimilitud precisamente por ser una función monótona de R.

TEOREMA.- Dada una muestra $x_1,\dots,x_N$ tomada de una $N(\mu,\Sigma)$, el test de la razón de verosimilitud a un nivel de significación α para la hipótesis $\bar{R}=0$, donde $\bar{R}$ es el coeficiente de correlación multiple entre X_1 y $(X_2,\dots,X_p)$, viene dado por

$$\frac{R^2}{1-R^2}\cdot\frac{N-p}{p-1} > F_{p-1,N-p}(\alpha)$$

donde R es el coeficiente de correlación multiple muestral definido como antes.

Del hecho de que la densidad de R es monótona creciente en $\bar{R}$, como veremos a continuación, podemos asegurar que el test definido en el teorema es uniformemente más potente para contrastar $\bar{R}=0$ en la clase de los test que dependen de R. Como además R es invariante bajo transformaciones del tipo $x^*_{i\alpha}=Cx_{i\alpha}+d$ y $x^{(2)*}_\alpha = Cx^{(1)}_\alpha + \mathbf{e}$ y es la única función del estadístico suficiente que es invariante, podemos concluir que el test es el test invariante uniformemente más potente.

Ejemplo.- Consideremos nuevamente el ejemplo de la cosecha de heno. El coeficiente de correlación multiple muestral será, para nuestros datos

$$1-R^2 = \frac{\begin{vmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{23} \\ r_{23} & 1 \end{vmatrix}} = 0{,}357$$

y de aquí $R=0.802$. Para contrastar la hipótesis a un nivel $\alpha=0.01$ de que la cosecha de heno es independiente de la cantidad de lluvia y de la temperatura, comparemos $[R^2/(1-R^2)]\,[(20-3)/(3-1)] = 15.3$ con $F_{2,17}(0.01)=6.11$ y concluiremos que sí hay dependencia, pues rechazaremos la hipótesis nula $H_0:\bar{R}=0$.

Digamos meramente que el test de independencia entre X_1 y $(X_2,\dots,X_p)=X^{(2)}$ es equivalente al test de si la regresión de X_1 sobre $x^{(2)}$ (es decir, el valor condicional esperado de X_1 dado $X^{(2)}=x^{(2)}$), que es $\mu_1+\beta(x^{(2)}-\mu^{(2)})$, es nula en el sentido de que el vector de regresión $\beta=0$. $\beta=a_{(1)}A_{22}^{-1}$ es la estimación usual mínimo cuadrática de β, cuyo valor esperado es β y matriz de covarianzas $\sigma_{11.2}A_{22}^{-1}$ (cuando las $X^{(1)}_\alpha$ son fijas) y $a_{11.2}/(N-p)$ es la estimación usual de $\sigma_{11.2}$. Entonces

$$\frac{R^2}{1-R^2}\cdot\frac{N-p}{p-1} = \frac{\hat{\beta}A_{22}\hat{\beta}'}{a_{11.2}}\cdot\frac{N-p}{p-1}$$

es el estadístico F usual para contrastar la hipótesis de nulidad para el vector de coeficientes de regresión de X_1 sobre $x_2,\dots,x_p$.

Es interesante resaltar, por último, que $\bar{R}$ es la única función de μ y Σ que es invariante bajo cambios de posición y escala de las variables y bajo cambios que impongan transformaciones lineales no singulares de $X^{(2)}$. Análogamente R es la única función de $\bar{x}$ y $\hat{\Sigma}$, estadísticos suficientes para μ y Σ, que es invariante bajo transformaciones similares.

Distribución del coeficiente de correlación multiple muestral cuando el poblacional es diferente de 0.

Vamos a encontrar la distribución de R cuando la hipótesis nula no se cumple. Encontraremos que esta distribución depende solo del coeficiente de correlación multiple poblacional, $\bar{R}$.

Consideremos en primer lugar la distribución condicional de $R^2/(1-R^2) = a_{(1)}A_{22}^{-1}a'_{(1)}/a_{11.2}$ dado $z^{(2)}_\alpha = z^{(2)}_\alpha$, $\alpha=1,2,\dots,n$. Bajo estas condiciones las $z_{1\alpha}$ se distribuyen independientemente como $N(\beta z^{(2)}_\alpha, \sigma_{11.2})$, donde $\beta=\sigma_{(1)}\Sigma_{22}^{-1}$ y $\sigma_{11.2}=\sigma_{11}-\sigma_{(1)}\Sigma_{22}^{-1}\sigma'_{(1)}$. Las condiciones son las de un teorema enunciado cuando estudiábamos las correlaciones parciales (véase párrafo 3 de este mismo capítulo), con $y_\alpha=z_{1\alpha}$, $\Gamma=\beta$, $w_\alpha=z^{(2)}_\alpha$ $(r=p-1)$, $\phi=\sigma_{11.2}$, $m=n$.

Entonces $a_{11.2} = a_{11} - a_{(1)} A_{22}^{-1} a_{(1)}$ se corresponde con $\sum_1^n Y_\alpha Y_\alpha' - GHG'$ y en consecuencia $a_{11.2}/\sigma_{11.2}$ tiene una distribución χ^2 con $n-(p-1)$ grados de libertad. Por otra parte $a_{(1)} A_{22}^{-1} a_{(1)}' = (a_{(1)} A_{22}^{-1}) A_{22} (A_{22}^{-1} a_{(1)}')$ se corresponde con GHG' y se distribuye como $\sum U_\alpha^2 \ (\alpha = n-(p-1)+1, \dots, n)$ donde $\text{var}(U_\alpha) = \sigma_{11.2}$ y

$$E(U_{n-p+2}, \dots, U_n) = \Gamma F^{-1},$$

donde, a su vez, $FHF' = I \ (H = F^{-1}(F')^{-1})$. Entonces $a_{(1)} A_{22}^{-1} a_{(1)}'/\sigma_{11.2}$ se distribuye como $\sum_\alpha \left(U_\alpha/\sqrt{\sigma_{11.2}}\right)^2$, donde $\text{var}\left(U_\alpha/\sqrt{\sigma_{11.2}}\right) = 1$ y

$$\sum_\alpha \left(\frac{E(U_\alpha)}{\sqrt{\sigma_{11.2}}}\right)^2 = \frac{1}{\sigma_{11.2}} \Gamma F^{-1} (\Gamma F^{-1})' = \frac{\Gamma H \Gamma'}{\sigma_{11.2}} = \frac{\beta A_{22} \beta'}{\sigma_{11.2}}$$

Así pues (condicionalmente) $a_{(1)} A_{22}^{-1} a_{(1)}'/\sigma_{11.2}$ tiene una distribución χ^2 no centrada con $p-1$ grados de libertad y parámetro de no centralización $\beta A_{22} \beta'/\sigma_{11.2}$. Podemos recoger todos estos resultados en el siguiente teorema

TEOREMA.- Sea R el coeficiente de relación múltiple muestral entre X_1 y $X^{(2)} = (X_2, \dots, X_p)$ basado en N observaciones $(x_{11}, x_1^{(2)}), \dots, (x_{1N}, x_N^{(2)})$. La distribución condicional de $[R^2/(1-R^2)][(N-p)/(p-1)]$ dado $x_\alpha^{(2)}$ fijo es una F de Snedecor no centrada con $p-1$ y $N-p$ grados de libertad y parámetro de no centralización $\beta A_{22} \beta'/\sigma_{11.2}$.

La densidad condicional de $F = [R^2/(1-R^2)][(N-p)/(p-1)]$ es de la forma

$$\frac{(p-1)\, e^{-\frac{1}{2}\beta A_{22}\beta'/\sigma_{11.2}}}{(N-p)\,\Gamma\left(\frac{1}{2}(N-p)\right)} \sum_{\alpha=0}^{\infty} \frac{\left(\frac{\beta A_{22}\beta'}{2\sigma_{11.2}}\right)^\alpha \left[\frac{(p-1)x}{N-p}\right]^{\frac{1}{2}(p-1)+\alpha-1} \Gamma\left(\frac{1}{2}(N-1)+\alpha\right)}{\alpha!\,\Gamma\left[\frac{1}{2}(p-1)+\alpha\right]\left[1+\frac{(p-1)x}{N-p}\right]^{\frac{1}{2}(N-1)+\alpha}}$$

y la densidad condicional de $W = R^2$ es $\left(dx = [(N-p)/(p-1)]\,(1-w)^{-2}\,dw\right)$

$$\frac{e^{-\frac{1}{2}\beta A_{22}\beta'/\sigma_{11.2}}}{\Gamma\left(\frac{1}{2}(N-p)\right)} (1-w)^{\frac{1}{2}(N-p)+1} \sum_{\alpha=0}^{\infty} \frac{\left(\frac{\beta A_{22}\beta'}{\sigma_{11.2}}\right)^\alpha w^{\frac{1}{2}(p-1)+\alpha-1} \Gamma\left(\frac{1}{2}(N-1)+\alpha\right)}{\alpha!\,\Gamma\left(\frac{1}{2}(p+1)+\alpha\right)}$$

Para obtener la densidad incondicional necesitamos multiplicar la expresión anterior por la marginal de $Z_1^{(2)} \dots Z_n^{(2)}$ obteniendo así la conjunta de W y $Z_\alpha^{(2)}$, y luego integrar respecto del último conjunto de variables, lo que nos dará, finalmente, la marginal de W. Tenemos

$$\frac{\beta A_{22}\beta'}{\sigma_{11.2}} = \frac{\beta \sum_{\alpha=1}^n Z_\alpha^{(2)} Z_\alpha^{(2)\prime} \beta'}{\sigma_{11.2}} = \sum_{\alpha=1}^n \left(\frac{\beta Z_\alpha^{(2)}}{\sqrt{\sigma_{11.2}}}\right)^2.$$

Puesto que la distribución de $Z_\alpha^{(2)}$ es $N(0, \Sigma_{22})$, la distribución de $\beta Z_\alpha^{(2)}/\sqrt{\sigma_{11.2}}$ es normal con media cero y varianza

$$E\left(\frac{\beta Z_\alpha^{(2)}}{\sqrt{\sigma_{11.2}}}\right)^2 = \frac{E\left(\beta Z_\alpha^{(2)} Z_\alpha^{(2)\prime} \beta'\right)}{\sigma_{11.2}} = \frac{\beta \Sigma_{22} \beta'}{\sigma_{11} - \beta \Sigma_{22} \beta'} = \frac{\beta \Sigma_{22} \beta'/\sigma_{11}}{1 - \beta \Sigma_{22} \beta'/\sigma_{11}} = \frac{\bar{R}^2}{1-\bar{R}^2}.$$

Así pues $\left(\beta A_{22}\beta'/\sigma_{11.2}\right)/\left[\bar{R}^2/(1-\bar{R}^2)\right]$ es una χ^2 con n grados de libertad. Pongamos $\bar{R}^2/(1-\bar{R}^2) = \phi$. Entonces $\beta A_{22}\beta'/\sigma_{11.2} = \phi \chi_n^2$. Calculemos

$$E\left(e^{-\frac{1}{2}\phi\chi_n^2} \cdot \left(\frac{\phi\chi_n^2}{2}\right)^\alpha\right) = \frac{\phi^\alpha}{2^\alpha} \int_0^\infty u^\alpha e^{-\frac{1}{2}\phi u} \frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)} u^{\frac{1}{2}n-1} e^{-\frac{1}{2}u}\, du = \frac{\phi^\alpha}{2^\alpha} \int_0^\infty \frac{1}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)} u^{\frac{1}{2}n+\alpha-1} \cdot e^{-\frac{1}{2}(u+\phi)}\, du =$$

$$= \frac{\phi^\alpha}{(1+\phi)^{\frac{1}{2}n+\alpha}} \frac{\Gamma\left(\frac{1}{2}n+\alpha\right)}{\Gamma\left(\frac{1}{2}n\right)} \int_0^\infty \frac{1}{2^{\frac{1}{2}n+\alpha}\Gamma\left(\frac{1}{2}n+\alpha\right)} v^{\frac{1}{2}n+\alpha-1} \cdot e^{-\frac{1}{2}v}\, dv = \frac{\phi^\alpha}{(1+\phi)^{\frac{1}{2}n+\alpha}} \cdot \frac{\Gamma\left(\frac{1}{2}n+\alpha\right)}{\Gamma\left(\frac{1}{2}n\right)}.$$

Aplicando este resultado para nuestro desarrollo obtenemos como densidad para R^2 (Fisher 1928)

$$\frac{(1-R^2)^{\frac{1}{2}(n-p+1)}\,(1-\bar{R}^2)^{\frac{1}{2}n}}{\Gamma\left(\frac{1}{2}(n-p+1)\right)\Gamma\left(\frac{1}{2}n\right)} \sum_{\mu=0}^{\infty} \frac{\left(\bar{R}^2\right)^\mu \left(R^2\right)^{\frac{1}{2}(p-1)+\mu-1} \Gamma^2\left(\frac{1}{2}n+\mu\right)}{\mu!\,\Gamma\left(\frac{1}{2}(p-1)+\mu\right)}.$$

EJEMPLOS

(1) Se sospecha que los niveles de secreción de dos compuestos bioquímicos durante una situación de stress están correlacionados. Los procesos biológicos que están implicados en este tipo de situaciones augurran que la correlación, de existir, debe ser positiva. Para contrastar la hipótesis nula $H_0: \rho = 0$ a un nivel $\alpha = 0.01$ frente a la alternativa $H_1: \rho > 0$ deseamos diseñar un experimento basado en un número suficiente, N, de ensayos independientes de manera que una correlación poblacional tan pequeña como $\rho = 0.20$ pueda ser detectada con una potencia 0.95. Los valores asimétricos de las probabilidades α y β reflejan las consecuencias relativas de tomar en cuenta una correlación falsa o pasar por alto una correlación real, aunque pequeña.

Utilizaremos como estadístico la z de Fisher. En estas condiciones la región crítica para la hipótesis nula viene dada por aquellos valores de z tales que

$$(z - \xi_0)\sqrt{N-3} \geq z_\alpha \longrightarrow z \geq z_\alpha/\sqrt{N-3} + \xi_0$$

donde

$$z = \tfrac{1}{2}\log\frac{1+r}{1-r} \quad \text{y} \quad \xi_0 = \tfrac{1}{2}\log\frac{1+\rho_0}{1-\rho_0} \quad \text{y} \quad z_\alpha \,/\, P(z \geq z_\alpha) = \alpha .$$

En nuestro caso $\rho_0 = 0$ y además podemos llevar a cabo una aproximación mediante la normal, con lo que z_α se calcula a partir de las tablas de la normal tipificada.

La potencia del test es la probabilidad de rechazar la hipótesis nula cuando es falsa, por lo tanto, en nuestro caso y recurriendo nuevamente a la aproximación mediante la normal, dicha potencia viene dada por

$$1 - \beta(\rho) = 1 - \Phi\left[z_\alpha + (\xi_0 - \xi)\sqrt{N-3}\right] \longrightarrow \Phi\left(\frac{z_\alpha/\sqrt{N-3} + \xi_0 - \xi}{1/\sqrt{N-3}}\right) = \Phi\left(z_\alpha + (\xi_0 - \xi)\sqrt{N-3}\right)$$

donde $\beta(\rho)$ es la probabilidad del error de tipo II, es decir, aceptar H_0 cuando es falsa y el verdadero valor es ρ. Así las cosas, $z_{0.01} = 2.33$ (obtenido de unas tablas de la $N(0,1)$), $\xi_0 = 0$, y $\xi = \tfrac{1}{2}\log\frac{1+\rho}{1-\rho}$ que para $\rho = 0.20$, valor a partir del cual queremos detectar correlaciones con dicha potencia (0.95), vale $\xi = 0.2027$. Así pues la potencia del test debe de satisfacer la condición

$$0.95 = 1 - \Phi\left(2.33 - 0.2027\sqrt{N-3}\right)$$

es decir,

$$\Phi\left(2.33 - 0.2027\sqrt{N-3}\right) = 0.05$$

y como Φ es la función de distribución de una $N(0,1)$, entonces

$$2.33 - 0.2027\sqrt{N-3} = -1.645$$

y de aquí $N = 389$. Es decir, al menos 389 pares de observaciones deben recogerse en el estudio para que se verifiquen las condiciones exigidas. El número es elevado, pero es el precio que debemos pagar para alcanzar los niveles α y β de probabilidad, tan rigurosos, deseados.

(2) Se ha aplicado determinado test de inteligencia a 933 individuos. Dicho test admite cuatro tipos de respuestas que se corresponden con otras tantas variables a las que designaremos por C, comportamiento, V, comprensión verbal, A, edad y E, educación. Para estas cuatro variables se ha obtenido la siguiente matriz de correlaciones

$$\begin{array}{c} \\ C \\ V \\ A \\ E \end{array}\!\!\begin{array}{c} \begin{array}{cccc} C & V & A & E \end{array} \\ \begin{bmatrix} 1 & 0.72 & -0.44 & 0.60 \\ & 1 & -0.13 & 0.68 \\ & & 1 & -0.29 \\ & & & 1 \end{bmatrix} \end{array} \xrightarrow{\text{matriz de varianzas}} \begin{bmatrix} 126.07 & 116.44 & -53.98 & 20.85 \\ & 267.47 & -20.46 & 30.31 \\ & & 119.38 & -9.80 \\ & & & 9.58 \end{bmatrix}$$

Las desviaciones standard para cada una de ellas fueron 11.228, 14.404, 10.926 y 3.098, respectivamente. Las correlaciones parciales cuando mantenemos constantes la edad y la educación vienen dadas por las siguientes matrices:

Edad Constante

$$\begin{array}{c} \\ G \\ V \\ E \end{array}\begin{bmatrix} 1 & 0.74 & 0.55 \\ & 1 & 0.68 \\ & & 1 \end{bmatrix}$$

Educación constante

$$\begin{array}{c} \\ G \\ V \\ A \end{array}\begin{bmatrix} 1 & 0.53 & -0.33 \\ & 1 & 0.10 \\ & & 1 \end{bmatrix}$$

Análogamente, la correlación parcial $r_{12.34}$ entre G y V cuando ambas, A y E permanecen constante vale

$$r_{12.34} = 0.62.$$

Si queremos llevar a cabo contraste de hipótesis del tipo $H_0: \rho = 0$ frente a $H_1: \rho \neq 0$, para $\alpha = 0.05$, podemos utilizar la t de student (ya sabemos que $\sqrt{N'-2}\, r/\sqrt{1-r^2}$ es una t de student con $N'-2$ grados de libertad). Ahora bien, como $N = 933$ supone una muestra de gran tamaño, podemos aproximar muy bien la t mediante la normal, haciendo caso omiso de la pérdida de grados de libertad (1 o 2 según los casos) ocasionada cuando mantenemos constantes 1 o 2 variables en el cálculo de las correlaciones parciales. En estas condiciones los valores críticos para r vienen dados por $r = \pm 0.064$, lo que supone que la hipótesis $H_0: \rho = 0$ sea rechazada cuando la aplicamos a cada uno de los coeficientes de correlación parcial obtenidos.

Calculemos ahora el coeficiente de correlación múltiple del comportamiento frente las otras tres variables y los coeficientes de regresión correspondientes, tenemos

$$\bar{R}^2_{1.234} = 0.644 \qquad \hat{\beta}' = [0.485, -0.346, 0.289]$$

los coeficientes de regresión vienen dados en las unidades de las variables originales.

Si queremos contrastar la hipótesis $H_0: \bar{R} = 0$ podemos hacer uso del hecho de que $\left(R^2/1-R^2\right)\cdot\left((N-P)/(P-1)\right)$, bajo la hipótesis nula se distribuye como una F con $p-1$ y $N-p$ grados de libertad. Ahora para $\alpha = 0.05$, tendremos

$$F_{3,929}(0.05) \simeq 2.60$$

como

$$\frac{R^2}{1-R^2}\cdot\frac{N-P}{P-1} = \frac{0.644}{0.356}\cdot\frac{929}{3} = 560.18 > 2.60$$

rechazamos la hipótesis H_0. Recordemos que este test es equivalente al efectuado para contrastar la hipótesis $\beta = 0$, Aceptar que $R \neq 0$ supone aceptar que $\beta \neq 0$, lo que sólo mismo, admitir una dependencia de tipo lineal entre G y las otras variables, siendo $\hat{\beta}'$ una estimación de los correspondientes coeficientes.

Capítulo 4

El estadístico T^2 generalizado

EL ESTADISTICO T^2 GENERALIZADO

1. INTRODUCCION

Uno de los grupos de problemas, en estadística univariante, más importante es aquel que se ocupa de las cuestiones concernientes a la media de una distribución dada cuando la varianza de la distribución es desconocida. Sobre la base de una muestra se puede buscar decidir si la media es igual a un número determinado de antemano, o bien dar un intervalo que contenga dicha media. El estadístico utilizado usualmente en estadística univariante es la diferencia entre la media muestral, $\bar{x}$, y la media hipotética poblacional dividida por la desviación típica muestral, s. Si la distribución muestreada es $N(\mu, \sigma)$, entonces

$$t = \sqrt{N}\,\frac{\bar{x}-\mu}{s}$$

tiene una distribución t de Student con $N-1$ g.l., donde N es el número de observaciones en la muestra. Sobre la base de este hecho podemos llevar a cabo test de hipótesis $\mu = \mu_0$, donde μ_0 viene especificado, o bien podemos construir un intervalo de confianza para el parámetro desconocido μ.

El estadístico multivariante equivalente al cuadrado de t viene dado por

$$T^2 = N(\bar{x}-\mu)' S^{-1} (\bar{x}-\mu),$$

donde $\bar{x}$ es el vector media muestral y S es la matriz de covarianza muestral. Nos ocuparemos de este estadístico y de su utilización para resolver los problemas equivalentes en estadística multivariante, así como de otras utilizaciones del mismo. El primero en introducir el estadístico T^2 fue Hotelling en 1931 quien además estudió su distribución teórica bajo la hipótesis nula que se había planteado.

2. OBTENCION Y DISTRIBUCION DEL ESTADISTICO T^2 GENERALIZADO

Derivación del estadístico T^2 como una función del criterio de la razón de verosimilitud

Aunque el estadístico T^2 tiene varios usos, comenzaremos demostrando que el test de la razón de verosimilitud de la hipótesis $\mu = \mu_0$ sobre la base de una muestra de una población $N(\mu, \Sigma)$ se apoya precisamente en dicho estadístico T^2. Supongamos que tenemos N observaciones $x_1 \dots x_N$ $(N > p)$. La función de verosimilitud es

$$L(\mu, \Sigma^{-1}) = \frac{|\Sigma^{-1}|^{\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}pN}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha-\mu)'\Sigma^{-1}(x_\alpha-\mu)\right]$$

El criterio de la razón de verosimilitud es

$$\lambda = \frac{\max\limits_{\Sigma^{-1}} L(\mu_0, \Sigma^{-1})}{\max\limits_{\mu, \Sigma^{-1}} L(\mu, \Sigma^{-1})}$$

Como ya sabemos el máximo para el denominador se alcanza sustituyendo μ y Σ^{-1} por sus correspondientes estimadores máximo-verosímiles

$$\hat{\mu}_\Omega = \bar{x} \quad , \quad \hat{\Sigma}_\Omega = \frac{1}{N}\sum_\alpha (x_\alpha - \bar{x})'(x_\alpha - \bar{x})'$$

Cuando $\mu = \mu_0$, la función de verosimilitud se maximiza en

$$\hat{\Sigma}_\omega = \frac{1}{N}\sum_\alpha (x_\alpha - \mu_0)'(x_\alpha - \mu_0)$$

como ya vimos en su momento. Aplicando un lema utilizado en el capítulo 2 referente a la obtención del máximo de la función de verosimilitud podemos afirmar que

$$\max_{\Sigma^{-1}, \mu} L(\mu, \Sigma^{-1}) = \frac{1}{(2\pi)^{\frac{1}{2}pN}\,|\hat{\Sigma}_\Omega|^{\frac{1}{2}N}}\, e^{-\frac{1}{2}pN} \quad \text{y} \quad \max_{\Sigma^{-1}} L(\mu_0, \Sigma^{-1}) = \frac{1}{(2\pi)^{\frac{1}{2}pN}\,|\hat{\Sigma}_\omega|^{\frac{1}{2}N}}\, e^{-\frac{1}{2}pN}$$

Así, el criterio de la razón de verosimilitud es

$$\lambda = \frac{|\hat{\Sigma}_\Omega|^{\frac{1}{2}N}}{|\hat{\Sigma}_\omega|^{\frac{1}{2}N}} = \frac{|\sum (x_\alpha - \bar{x})(x_\alpha - \bar{x})'|^{\frac{1}{2}N}}{|\sum (x_\alpha - \mu_0)(x_\alpha - \mu_0)'|^{\frac{1}{2}N}} = \frac{|A|^{\frac{1}{2}N}}{|A + N(\bar{x}-\mu_0)(\bar{x}-\mu_0)'|^{\frac{1}{2}N}}$$

con $A = \sum (x_\alpha - \bar{x})(x_\alpha - \bar{x})' = (N-1)S$.

Para $|B| \neq 0$, utilizando una propiedad de las matrices, tenemos

$$\begin{vmatrix} B & C \\ D & E \end{vmatrix} = \begin{vmatrix} B & C \\ D & E \end{vmatrix} \cdot \begin{vmatrix} I & -B^{-1}C \\ 0 & I \end{vmatrix} = \begin{vmatrix} B & 0 \\ D & E - DB^{-1}C \end{vmatrix} = |B| \cdot |E - DB^{-1}C|$$

Aplicando dos veces este resultado a $\lambda^{2/N}$, tenemos

$$\lambda^{2/N} = \frac{|A|}{|A + [\sqrt{N}(\bar{x}-\mu_0)][\sqrt{N}(\bar{x}-\mu_0)]'|} = \frac{|A|}{\begin{vmatrix} 1 & \sqrt{N}(\bar{x}-\mu_0)' \\ -\sqrt{N}(\bar{x}-\mu_0) & A \end{vmatrix}} = \frac{|A|}{\begin{vmatrix} A & \sqrt{N}(\bar{x}-\mu_0) \\ -\sqrt{N}(\bar{x}-\mu_0)' & 1 \end{vmatrix}} =$$

$$= \frac{|A|}{|A| \cdot |1 + N(\bar{x}-\mu_0)' A^{-1} (\bar{x}-\mu_0)|} = \frac{1}{|1 + N(\bar{x}-\mu_0)' A^{-1}(\bar{x}-\mu_0)|} = \frac{1}{1 + T^2/(N-1)}$$

donde $T^2 = N(\bar{x}-\mu_0)' S^{-1} (\bar{x}-\mu_0) = (N-1) N (\bar{x}-\mu_0)' A^{-1} (\bar{x}-\mu_0)$.

El test de la razón de verosimilitud se define mediante la región crítica (o región de rechazo)

$$\lambda \leq \lambda_0$$

donde λ_0 se elige de manera que la probabilidad de que $\lambda \leq \lambda_0$ cuando $\mu = \mu_0$ (hipótesis nula cierta) sea igual al nivel de significación. Si operamos adecuadamente la anterior expresión es equivalente a

$$T^2 \geq T_0^2,$$

donde $T_0^2 = (N-1)(\lambda_0^{-2/N} - 1)$.

TEOREMA.- El test de la razón de verosimilitud de la hipótesis $\mu = \mu_0$ para la distribución $N(\mu, \Sigma)$ viene dado por $T^2 \geq T_0^2$, donde T^2 se define como anteriormente y T_0^2 se elige de manera que $Pr(T^2 \geq T_0^2)$, bajo la hipótesis nula, sea igual al nivel de significación elegido.

El test basado en la t de Student, para el caso univariante, tiene la propiedad de que cuando se contrasta la hipótesis $\mu = 0$ es invariante con respecto a transformaciones de escala. Si la variable aleatoria escalar $\bar{X}$ se distribuye como $N(\mu, \sigma)$ entonces $\bar{X}^* = c\bar{X}$ se distribuye $N(c\mu, c\sigma)$ que está en la misma clase y la hipótesis $E(\bar{X}) = 0$ equivale a $E(\bar{X}^*) = cE(\bar{X}) = 0$. Si las observaciones x_α se transforman análogamente, $x_\alpha^* = cx_\alpha$, entonces, para $c > 0$, t^* calculada a partir de x_α^* es la misma que la t calculada a partir de las x_α. Así, cualquiera que sea la unidad de medida el estadístico resulta ser el mismo.

El test T^2 generalizado tiene la misma propiedad. Si el vector aleatorio $\bar{X}$ se distribuye $N(\mu, \Sigma)$, entonces $\bar{X}^* = C\bar{X}$ (para $|C| \neq 0$) se distribuye $N(C\mu, C\Sigma C')$ que está en la misma clase, la hipótesis $E(\bar{X}) = 0$ equivale a $E(\bar{X}^*) = CE(\bar{X}) = 0$. Si las observaciones x_α se transforman de la misma manera, $x_\alpha^* = Cx_\alpha$, entonces T^2 y T^{*2} son iguales.

Esta consecuencia de que $\bar{x}^* = C\bar{x}$, $A^* = CAC'$ y del siguiente lema:

LEMA.- Para cualquier matriz C, $p \times p$, no singular, H de iguales características y k un vector cualquiera

$$k'H^{-1}k = (Ck)'(CHC')^{-1}(Ck).$$

Demostración.- $(Ck)'(CHC')^{-1}(Ck) = k'C'(C')^{-1}H^{-1}C^{-1}Ck = k'H^{-1}k.$

Demostraremos más adelante que de todos los test invariantes bajo estas transformaciones, este es uniformemente más potente.

Podemos dar ahora una interpretación geométrica de la raíz $2/N$-ésima del criterio de la unión de verosimilitud

$$\lambda^{2/N} = \frac{|\sum (x_\alpha - \bar{x})(x_\alpha - \bar{x})'|}{|\sum (x_\alpha - \mu_0)(x_\alpha - \mu_0)'|}$$

en términos de paralelotopos. En una representación p-dimensional el numerador de $\lambda^{2/N}$ es la suma de los cuadrados de los volúmenes de todos los paralelotopos con ejes principales p vectores, cada uno con un extremo en el punto $\bar{x}$ y el otro en x_α. El denominador es la suma de cuadrados de los volúmenes de los paralelotopos cuyos ejes principales tienen un extremo en μ_0 y el otro en x_α. Si la suma de los cuadrados de los volúmenes que involucran vectores que emanan de $\bar{x}$, el centro de de los x_α, es mucho menor que la de los volúmenes que involucran vectores que emanan de μ_0, entonces rechazamos la hipótesis de que μ_0 es la media de la distribución.

Existe también una interpretación en el caso de una representación N-dimensional. Sea $y_i = (x_{i1}, \dots, x_{iN})$ el vector i-ésimo. Entonces

$$\sqrt{N}\,\bar{x}_i = \sum_{\alpha=1}^{N} \frac{1}{\sqrt{N}} x_{i\alpha}$$

es la distancia desde el origen de la proyección de y_i en la línea equiangular (con cosenos directores $1/\sqrt{N}, 1/\sqrt{N}, \dots, 1/\sqrt{N}$). Las coordenadas de la proyección son $(\bar{x}_i, \dots, \bar{x}_i)$. Entonces $(x_{i1} - \bar{x}_i, \dots, x_{iN} - \bar{x}_i)$ es la proyección de y_i sobre el plano, a través del origen, perpendicular a la línea equiangular. El numerador de $\lambda^{2/N}$ es el cuadrado del volumen p-dimensional del paralelotopo con ejes principales los vectores $(x_{i1} - \bar{x}_i, \dots, x_{iN} - \bar{x}_i)$. Un punto $(x_{i1} - \mu_{0i}, \dots, x_{iN} - \mu_{0i})$ se obtiene a partir de y_i una traslación paralela a la línea equiangular (a una distancia $\sqrt{N}\,\mu_{0i}$). El denominador es el cuadrado del volumen de los paralelotopos con ejes principales dichos vectores. Entonces $\lambda^{2/N}$ es la relación de los cuadrados de dichos volúmenes.

La distribución de T^2

Encontraremos en esta sección la distribución de T^2 bajo condiciones generales, incluido el caso en que la hipótesis nula no es cierta. Sea $T^2 = Y' S^{-1} Y$, donde Y se distribuye $N(\mu, \Sigma)$ y nS se distribuye independientemente como $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ con Z_α independientes entre sí, cada una con una distribución $N(0, \Sigma)$. El T^2 antes utilizado es un caso especial de esta definición con $Y = \sqrt{N}(\bar{x} - \mu_0)$ y $\nu = \sqrt{N}(\mu - \mu_0)$ y $n = N - 1$. Sea D una matriz no singular tal que $D \Sigma D' = I$, y definimos

$$Y^* = DY$$
$$S^* = DSD'$$
$$\nu^* = D\nu .$$

Entonces $T^2 = Y^{*\prime} S^{*-1} Y^*$, de acuerdo con el lema anterior, donde Y^* se distribuye $N(\nu^*, I)$ y nS^* se distribuye independientemente como $\sum_{\alpha=1}^{n} Z_\alpha^* Z_\alpha^{*\prime} = \sum_{\alpha=1}^{n} DZ_\alpha (DZ_\alpha)'$ con $Z_\alpha^* = DZ_\alpha$ independientes entre sí, cada una con distribución $N(0, I)$. Observemos que $\nu' \Sigma^{-1} \nu = \nu^{*\prime} (I)^{-1} \nu^* = \nu^{*\prime} \nu^*$.

Sea la primera fila de una matriz ortogonal Q, de dimensión $p \times p$, definida mediante

$$q_{1i} = \frac{y_i^*}{\sqrt{y^{*\prime} y^*}} \qquad i = 1, \dots, p$$

lo que es permisible por cuanto se ha de verificar $\sum_i q_{1i}^2 = 1$. Las otras $p-1$ filas pueden definirse de forma arbitraria de acuerdo con una propiedad de estas matrices. Como Q depende de los Y^* se tratará de una matriz aleatoria. Sean ahora

$$U = QY^*$$
$$B = Q n S^* Q' .$$

De la definición de Q se sigue:

$$u_1 = \sum q_{1i} y_i^* = \sqrt{y^{*\prime} y^*}$$
$$u_j = \sum q_{ji} y_i^* = \sqrt{y^{*\prime} y^*} \sum q_{ji} q_{1i} = 0 . \qquad j \neq 1 .$$

Entonces

$$\frac{T^2}{n} = u' B^{-1} u = (u_1\ 0 \dots 0) \begin{bmatrix} b^{11} & b^{12} & \dots & b^{1p} \\ \dots & \dots & \dots & \dots \\ b^{p1} & b^{p2} & \dots & b^{pp} \end{bmatrix} \begin{pmatrix} u_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = u_1^2 . b^{11}$$

donde $(b^{ij}) = B^{-1}$. Haciendo uso de una propiedad de las matrices particionadas tenemos,

$$1/b^{11} = b_{11} - b_{(1)} . B_{22}^{-1} b'_{(1)} = b_{11.2-p}$$

donde

$$B = \begin{pmatrix} b_{11} & b_{(1)} \\ b'_{(1)} & B_{22} \end{pmatrix},$$

y $T^2 = n u_1^2 / b_{11.2-p} = n\, y^{*\prime} y^* / b_{11.2-p}$. La distribución condicional de B dado Q es la de $\sum_{\alpha=1}^{n} V_\alpha V'_\alpha$, donde condicionalmente los $V_\alpha = Q z_\alpha^*$ son independientes entre sí, cada uno con una distribución $N(0, I)$. Por otra parte, aplicando las propiedades obtenidas en el capítulo anterior para la distribución de las matrices $A_{11.2}$, cuando estudiábamos el coeficiente de correlación parcial, sabemos que $b_{11.2-p}$ se distribuye condicionalmente como $\sum_{\alpha=1}^{n-(p-1)} W_\alpha^2$ donde condicionalmente los W_α son independientes entre sí, cada una $N(0,1)$; es decir, $b_{11.2-p}$ se distribuye condicionalmente como una χ^2 con $n-(p-1)$ gr. de libertad. Puesto que la distribución condicional de $b_{11.2-p}$ no depende de Q, ello significa que se distribuye incondicionalmente como χ^2. La cantidad $y^{*\prime} y^*$ tiene una distribución χ^2 no centrada con p grados de libertad y parámetro de no centralización $v^{*\prime} v^* = v' \Sigma^{-1} v$. En definitiva T^2/n se distribuye como la razón de una χ^2 no centrada y una χ^2 independientes.

TEOREMA.- Sea $T^2 = y' S^{-1} y$, donde y se distribuye $N(v, \Sigma)$ y nS independientemente como $\sum_{\alpha=1}^{n} z_\alpha z'_\alpha$ con z_α independientes entre sí, cada una $N(0, \Sigma)$. Entonces $(T^2/n)[(n-p+1)/p]$ se distribuye como una F no centrada con p y $n-p+1$ grados de libertad y parámetro de no centralización $v' \Sigma^{-1} v$. Si $v=0$ la distribución es una F centrada.

Llamaremos a esta distribución una distribución T^2 con n grados de libertad.

COROLARIO.- Sea $x_1 \dots x_N$ una muestra de una $N(\mu, \Sigma)$ y sea $T^2 = N(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0)$. La distribución de $[T^2/(N-1)][(N-p)/p]$ es una F no centrada con p y $N-p$ grados de libertad y parámetro de no centralización $N(\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0)$. Si $\mu = \mu_0$, entonces la distribución es una F centrada.

La densidad y tablas de la F no centrada se estudiarán en un párrafo posterior.

3. USOS DEL ESTADISTICO T^2

Ya hemos visto que el test de la razón de verosimilitud de la hipótesis $\mu = \mu_0$ sobre la base de una muestra de tamaño N de una $N(\mu, \Sigma)$ es equivalente a

$$T^2 \geq T_0^2 .$$

Si el nivel de significación es α, elegimos T_0 a partir de una tabla F centrada de manera que

$$T_0^2 = \frac{(N-1)p}{N-p} F_{p, N-p}(\alpha),$$

siendo $F_{p,N-p}(\alpha)$ el punto que deja a su derecha un área α. La elección del nivel de significación depende o puede hacerlo de la potencia del test, de ello nos ocuparemos más tarde.

El estadístico T^2 se calcula con facilidad a partir de la muestra. En efecto

$$A^{-1} (\bar{x} - \mu_0) = b$$

es la solución de la ecuación

$$Ab = (\bar{x} - \mu_0),$$

entonces

$$\frac{T^2}{N-1} = N (\bar{x} - \mu_0)' b .$$

Obsérvese que $T^2/(N-1)$ es la raíz no nula de

$$\left| N(\bar{x}-\mu_0)(\bar{x}-\mu_0)' - \lambda A \right| = 0 .$$

LEMA.- Si v es un vector de p componentes y si B es una matriz no singular de orden p, entonces $v'B^{-1}v$ es la raíz no nula de

$$|vv' - \lambda B| = 0 .$$

Demostración.- La raíz λ_1, no nula de la ecuación anterior está asociada con un vector característico β que satisface

$$vv'\beta = \lambda_1 B\beta .$$

Como $\lambda_1 \neq 0$, $v'\beta \neq 0$. Multiplicando por la izquierda por $v'B^{-1}$, tenemos

$$(v'B^{-1}v)(v'\beta) = \lambda_1 (v'\beta),$$

lo que demuestra el lema. En el caso anterior $v = \sqrt{N}(\bar{x}-\mu_0)$ y $B = A$.

Una región de confianza para el vector media

Si μ es la media de $N(\mu, \Sigma)$ sabemos que la probabilidad de extraer una muestra de tamaño N con media $\bar{x}$ y matriz de covarianzas S, tal que

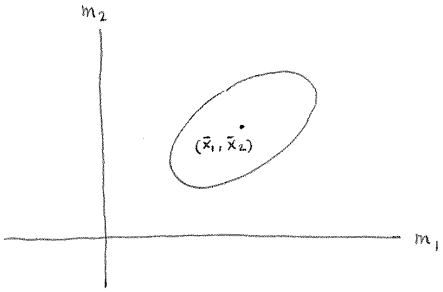
$$N(\bar{x}-\mu)' S^{-1} (\bar{x}-\mu) \leq T_0^2(\alpha)$$

es precisamente $1-\alpha$.

Así pues, si calculamos la expresión anterior para una muestra particular, tenemos una confianza $1-\alpha$ de que la citada expresión sea cierta en lo que a μ se refiere. La desigualdad

$$N(\bar{x}-m)' S^{-1} (\bar{x}-m) \leq T_0^2(\alpha)$$

es el interior y la frontera de un elipsoide p-dimensional con centro en $\bar{x}$ y cuyo tamaño y forma dependen de S^{-1} y α. (Véase la figura para el caso bidimensional). Afirmamos que μ está en el elipsoide. Lógicamente el elipsoide es aleatorio por cuanto depende de muestras aleatorias.



Problema de dos muestras

Otra situación en la cual se utiliza el estadístico T^2 es aquella en la que se pretende contrastar la hipótesis de la igualdad de medias para dos poblaciones normales multivariantes cuyas matrices de covarianza se desconocen pero se suponen iguales. Supongamos $y_1^{(i)} \cdots y_{N_i}^{(i)}$ es una muestra de $N(\mu^{(i)}, \Sigma)$, $i = 1,2$. Queremos contrastar la hipótesis $\mu^{(1)} = \mu^{(2)}$. $\bar{y}^{(i)}$ se distribuye $N(\mu^{(i)}, (1/N_i)\Sigma)$. Consecuentemente $\sqrt{N_1 N_2/(N_1+N_2)}\,(\bar{y}^{(1)} - \bar{y}^{(2)})$ se distribuye como $N(\mu^{(1)} - \mu^{(2)}, \Sigma)$ y si la hipótesis nula $\mu^{(1)} = \mu^{(2)}$ es cierta esta distribución será $N(0, \Sigma)$. Si hacemos

$$S = \frac{1}{N_1 + N_2 - 2} \left\{ \sum_{\alpha=1}^{N_1} (y_\alpha^{(1)} - \bar{y}^{(1)})(y_\alpha^{(1)} - \bar{y}^{(1)})' + \sum_{\alpha=1}^{N_2} (y_\alpha^{(2)} - \bar{y}^{(2)})(y_\alpha^{(2)} - \bar{y}^{(2)})' \right\}$$

entonces $(N_1 + N_2 - 2)S$ se distribuye como $\sum_{\alpha=1}^{N_1+N_2-2} Z_\alpha Z_\alpha'$ con Z_α $N(0, \Sigma)$. Así pues

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\bar{y}^{(1)} - \bar{y}^{(2)})' S^{-1} (\bar{y}^{(1)} - \bar{y}^{(2)})$$

se distribuye como T^2 con $N_1 + N_2 - 2$ grados de libertad. La región crítica viene dada por

$$T^2 \geq \frac{(N_1+N_2-2)p}{N_1+N_2-p-1} F_{p,N_1+N_2-p-1}(\alpha)$$ con nivel de significación α.

Un ejemplo de aplicación de esta teoría lo encontramos en Fisher (1936). Sea x_1 = longitud del sépalo, x_2 = anchura del sépalo, x_3 = longitud del pétalo y x_4 = anchura del pétalo. Se toman 50 observaciones de la población Iris versicolor (1) y otras 50 de la población Iris setosa (2). Los datos resumidos encontramos son los siguientes

$$\bar{x}^{(1)} = \begin{pmatrix} 5.936 \\ 2.770 \\ 4.260 \\ 1.326 \end{pmatrix}, \quad \bar{x}^{(2)} = \begin{pmatrix} 5.006 \\ 3.428 \\ 1.462 \\ 0.246 \end{pmatrix}, \quad 98S = \begin{pmatrix} 19.1434 & 9.0356 & 9.7634 & 3.2394 \\ & 11.8658 & 4.6232 & 2.4746 \\ & & 12.2978 & 3.8794 \\ & & & 2.4604 \end{pmatrix}$$

El valor de $T^2/98 > 26.334$ y $T^2/98 \times 95/4 = 625.5$. Este valor es altamente significativo (comparado con el punto F para 4 y 95 grados de libertad), en consecuencia rechazaríamos la hipótesis de igualdad de medias en ambas poblaciones.

Un problema de q-muestras

Después de considerar el anterior ejemplo, Fisher obtiene una tercera muestra de una población a la que supone con igual matriz de covarianzas. Lleva a cabo 50 medidas, las mismas, sobre individuos Iris borgnica. Existe una razón teórica para creer que la estructura genética de estas tres poblaciones es tal que sus vectores medias están relacionados como sigue:

$$3\mu^{(1)} = \mu^{(3)} + 2\mu^{(2)}$$

siendo $\mu^{(3)}$ el vector media de la tercera población.

Se trata de un caso particular del siguiente problema general. Sean $\{x_\alpha^{(i)}\}$ $\alpha = 1,\dots,N_i$, $i = 1,\dots,q$ muestras extraídas de $N(\mu^{(i)}, \Sigma)$ $i = 1,\dots,q$, respectivamente. Hagamos la siguiente hipótesis H:

$$\sum_{i=1}^{q} \beta_i \mu^{(i)} = \mu,$$

donde $\beta_1,\dots,\beta_q$ son escalares dados y μ un vector dado. El criterio a aplicar en esta situación es

$$T^2 = c \left(\sum_{i=1}^{q} \beta_i \bar{x}^{(i)} - \mu \right)' S^{-1} \left(\sum_{i=1}^{q} \beta_i \bar{x}^{(i)} - \mu \right),$$

donde $\bar{x}^{(i)} = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} x_\alpha^{(i)}$, $\left(\sum_{i=1}^{q} N_i - q\right) S = \sum_{i=1}^{q} \sum_{\alpha=1}^{N_i} (x_\alpha^{(i)} - \bar{x}^{(i)})(x_\alpha^{(i)} - \bar{x}^{(i)})'$, $1/c = \sum_{i=1}^{q} \frac{\beta_i^2}{N_i}$.

Este T^2 tiene una distribución T^2 de Hotelling con $\sum_{i=1}^{q} N_i - q$ grados de libertad. Fisher, en realidad, supuso que las matrices de covarianzas podían ser diferentes y resolvió el problema aplicando una técnica de la que no comparemos más bonde.

Un problema de simetría

Consideremos el siguiente contraste de hipótesis H: $\mu_1 = \mu_2 = \cdots = \mu_p$ sobre la base de una muestra $x_1,\dots,x_N$ extraída de una población $N(\mu, \Sigma)$ donde $\mu' = (\mu_1,\dots,\mu_p)$. Sea C cualquier matriz $(p-1)\times p$ de rango $p-1$ tal que

$$C\varepsilon = 0$$

donde $\varepsilon' = (1,\dots,1)$. Entonces

$$y_\alpha = C x_\alpha \quad , \alpha = 1,\dots,N$$

tiene media $C\mu$ y matriz de covarianzas $C\Sigma C'$. La hipótesis H supone ahora $C\mu = 0$, en efecto de $C\varepsilon = 0$ obtenemos

$$\sum_{j=1}^{p} c_{ij} = 0 \quad , i = 1,\dots,p-1$$

pero $$C\mu = \begin{bmatrix} \sum c_{1j}\mu_j \\ \vdots \\ \sum c_{p-1,j}\mu_j \end{bmatrix}$$ si H se cumple $\sum_{j=1}^{p} c_{ij}\mu_j = \mu \left(\sum_{j=1}^{p} c_{ij}\right) = 0$, donde $\mu = \mu_1 = \mu_2 = \cdots = \mu_p$.

Utilizaremos entonces el estadístico

$$T^2 = N\,\bar{y}'\,S^{-1}\,\bar{y}$$

con $\bar{y} = \frac{1}{N}\sum_{\alpha=1}^{N} y_\alpha = C\,\bar{x}$, $S = \frac{1}{N-1}\sum_{\alpha}(y_\alpha - \bar{y})(y_\alpha - \bar{y})' = \frac{1}{N-1}\,C\left[\sum_{\alpha}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'\right]C'$

Este estadístico tiene una distribución T^2 con $N-1$ grados de libertad para una distribución $(p-1)$-dimensional. Este estadístico T^2 es invariante bajo cualquier transformación lineal en las $p-1$ dimensiones que sea ortogonal a E. En consecuencia el estadístico es independiente de la elección que se haga para C.

Un ejemplo que ilustre este problema viene dado por Rao (1948). Sea N la cantidad de corcho extraída, de un árbol productor de corcho, en la cara Norte del mismo, sean E, S, W definidas de forma similar para las caras Este, Sur y Oeste respectivamente. El conjunto de estas cuatro cantidades para un mismo árbol se considera una observación de una normal 4-variante. La cuestión a dilucidar es si los árboles tienen la misma cantidad de corcho en cada una de sus caras. Hacemos la siguiente transformación,

$$y_1 = N - E - W + S$$
$$y_2 = S - W$$
$$y_3 = N - S.$$

El número de observaciones es de 28.
El vector de medias es

$$\bar{y} = \begin{bmatrix} 8.86 \\ 4.50 \\ 0.86 \end{bmatrix}$$

La matriz de covarianzas para y es

$$S = \begin{bmatrix} 128.72 & 61.41 & -21.02 \\ & 56.93 & -28.30 \\ & & 63.53 \end{bmatrix}$$

El valor de $T^2/(N-1)$ es 0.768. El estadístico $0.768 \times 25/3 = 6.402$ es una F con 3 y 25 grados de libertad. Para $\alpha = 0.01$ $F_{3,25} = 4.68$ y rechazaríamos la hipótesis.

4. LA DISTRIBUCION DE T^2 BAJO HIPOTESIS ALTERNATIVAS: LA FUNCION POTENCIA.

Hemos visto anteriormente que $(T^2/n)\left((N-p)/p\right)$ tiene una distribución F no centrada. Se trata ahora de estudiar las distribuciones χ^2 y F no centradas, la tabulación de estas últimas y su aplicación a procedimientos basados en T^2.

La distribución χ^2 es la distribución de la suma de cuadrados de variables aleatorias (escalares) independientes normales con media 0 y varianza 1; la χ^2 no centrada es la generalización de la anterior cuando las medias pueden ser distintas de cero. Sea Y un vector de p componentes distribuido $N(\nu, I)$. Sea Q una matriz ortogonal cuya primera fila viene dada por

$$q_{1i} = \frac{\nu_i}{\sqrt{\nu'\nu}}.$$

Entonces $Z = QY$ se distribuye $N(\lambda, I)$, donde $\lambda' = [\sqrt{\nu'\nu}, 0, \ldots, 0]$ y $Z = \sqrt{\nu'\nu}$, como ya vimos anteriormente (apartado 2 de este capítulo). Sea $V = Y'Y = Z'Z = \sum_{i=1}^{p} Z_i^2$, entonces $W = \sum_{i=2}^{p} Z_i^2$ tiene una distribución χ^2 con $p-1$ grados de libertad y Z_1 y W tienen una densidad conjunta

$$\frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}(z_1 - z)^2}\cdot\frac{1}{2^{\frac{1}{2}(p-1)}\,\Gamma\left(\frac{1}{2}(p-1)\right)}\,w^{\frac{1}{2}(p-1)-1}e^{-\frac{1}{2}w} = C\,e^{-\frac{1}{2}(z^2+z_1^2+w)}\,w^{\frac{1}{2}(p-3)}\,e^{zz_1} =$$

$$= C\,e^{-\frac{1}{2}(z^2+z_1^2+w)}\,w^{\frac{1}{2}(p-3)}\sum_{\alpha=0}^{\infty}\frac{z^\alpha z_1^\alpha}{\alpha!}$$

donde $C^{-1} = 2^{\frac{1}{2}p}\sqrt{\pi}\,\Gamma\left[\frac{1}{2}(p-1)\right]$. La densidad conjunta de $V = W + Z_1^2$ y Z_1 se obtiene sustituyendo $w = v - z_1^2$ (el jacobiano

de la transformación vienda la 1),

$$C\, e^{-\frac{1}{2}(z^2+v)}\,(v-z_1^2)^{\frac{1}{2}(p-3)} \sum_{\alpha=0}^{\infty} \frac{z^\alpha z_1^\alpha}{\alpha!}.$$

La densidad conjunta de v y $u = z_1/\sqrt{v}$ es ($dz_1 = \sqrt{v}\,du$)

$$C\, e^{-\frac{1}{2}(z^2+v)}\, v^{\frac{1}{2}(p-2)}\,(1-u^2)^{\frac{1}{2}(p-3)} \sum_{\alpha=0}^{\infty} \frac{z^\alpha u^\alpha v^{\frac{1}{2}\alpha}}{\alpha!}$$

siendo el campo de variación de z_1 dado v, $[-\sqrt{v}, \sqrt{v}]$ y el de u, $[-1,1]$. Cuando integramos en esta última expresión respecto de u, término a término, los términos para α impar son una integral nula pues para tales términos se trata de una función impar de u. En las otras integramos haciendo el cambio $u = \sqrt{s}$ ($du = \frac{1}{2} ds/\sqrt{s}$), obteniendo

$$\int_{-1}^{1} (1-u^2)^{\frac{1}{2}(p-3)}\, u^{2\beta}\, du = 2\int_0^1 (1-u^2)^{\frac{1}{2}(p-3)}\, u^{2\beta}\, du = \int_0^1 (1-s)^{\frac{1}{2}(p-3)}\, s^{\beta-\frac{1}{2}}\, ds = B\left[\tfrac{1}{2}(p-1), \beta+\tfrac{1}{2}\right] =$$

$$= \frac{\Gamma\left[\frac{1}{2}(p-1)\right]\cdot\Gamma\left(\beta+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}p+\beta\right)}$$ de acuerdo con la relación existente entre las funciones Beta y Gamma.

La densidad de v viene pues dada por:

$$\frac{1}{2^{\frac{1}{2}p}\sqrt{\pi}}\, e^{-\frac{1}{2}(z^2+v)}\, v^{\frac{1}{2}p-1} \sum_{\beta=0}^{\infty} \frac{(z^2)^\beta v^\beta}{(2\beta)!}\cdot\frac{\Gamma\left(\beta+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}p+\beta\right)}.$$

Esta es la densidad de una χ^2 no centrada con p grados de libertad y parámetro de no centralización z^2.

TEOREMA.- Si el vector Y de p componentes se distribuye $N(U, I)$, entonces $V = Y'Y$ tiene una densidad χ^2 no centrada con p grados de libertad, siendo $z^2 = U'U$ el parámetro de no centralización.

Sea ahora V una χ^2 no centrada con p grados de libertad y parámetro de no centralización z^2 y sea W, independiente de la anterior, una χ^2 con m grados de libertad. Vamos a encontrar la densidad de $F = (V/p)/(W/m)$, que es una F no centrada con parámetro z^2. La densidad conjunta de V y W es la densidad de V multiplicada por la de W, que es $2^{-\frac{1}{2}m}\Gamma^{-1}\left(\frac{1}{2}m\right) w^{\frac{1}{2}m-1} e^{-\frac{1}{2}w}$. La densidad conjunta de F y W ($dw = pw\,df/m$) es

$$\frac{e^{-\frac{1}{2}z^2}}{2^{\frac{1}{2}(p+m)}\sqrt{\pi}\,\Gamma\left(\frac{1}{2}m\right)}\, e^{-\frac{1}{2}w(1+pf/m)} \cdot \frac{p}{m}\sum_{\beta=0}^{\infty} \frac{(z^2)^\beta\,\Gamma\left(\beta+\frac{1}{2}\right)}{(2\beta)!\,\Gamma\left(\frac{1}{2}p+\beta\right)}\cdot\left(\frac{pf}{m}\right)^{\frac{1}{2}p+\beta-1} w^{\frac{1}{2}(p+m)+\beta-1}.$$

La densidad marginal se obtiene integrando para w entre 0 y $+\infty$,

$$\frac{p\, e^{-\frac{1}{2}z^2}}{m\sqrt{\pi}\,\Gamma\left(\frac{1}{2}m\right)} \sum_{\beta=0}^{\infty} \frac{(2z^2)^\beta\,\Gamma\left(\beta+\frac{1}{2}\right)}{(2\beta)!\,\Gamma\left(\frac{1}{2}p+\beta\right)}\left(\frac{pf}{m}\right)^{\frac{1}{2}p+\beta-1}\cdot\frac{\Gamma\left(\frac{1}{2}(p+m)+\beta\right)}{(1+pf/m)^{\frac{1}{2}(p+m)+\beta}}.$$

Si utilizamos la fórmula de la duplicación de la función gamma

$$(2\beta)! = \Gamma(2\beta+1) = \Gamma\left(\beta+\tfrac{1}{2}\right)\Gamma\left(\beta+\tfrac{1}{2}\right)\, 2^{2\beta}/\sqrt{\pi}$$

la densidad queda finalmente

$$\frac{p\, e^{-\frac{1}{2}z^2}}{m\,\Gamma\left(\frac{1}{2}m\right)} \sum_{\beta=0}^{\infty} \frac{(z^2/2)^\beta\cdot\left(pf/m\right)^{\frac{1}{2}p+\beta-1}\cdot\Gamma\left(\frac{1}{2}(p+m)+\beta\right)}{\beta!\,\Gamma\left(\frac{1}{2}p+\beta\right)\left(1+pf/m\right)^{\frac{1}{2}(p+m)+\beta}}$$

que es la densidad de una F no centrada con p y m grados de libertad respectivamente y parámetro de no centralización z^2.

Si $T^2 = N(\bar{x}-\mu_0)' S^{-1}(\bar{x}-\mu_0)$ se basa en una muestra de tamaño N de una $N(\mu,\Sigma)$, entonces $(T^2/n)\left(\frac{N-p}{p}\right)$ tiene una distribución F no centrada con p y $N-p$ grados de libertad y parámetro de no centralización $N(\mu-\mu_0)'\Sigma^{-1}(\mu-\mu_0) = \tau^2$. La expresión de la densidad de T^2 será

$$\frac{e^{-\frac{1}{2}\tau^2}}{(N-1)\Gamma\left(\frac{1}{2}(N-p)\right)} \sum_{\beta=0}^{\infty} \frac{(\tau^2/2)^\beta \left[t^2/(N-1)\right]^{\frac{1}{2}p+\beta-1} \Gamma\left(\frac{1}{2}N+\beta\right)}{\beta!\,\Gamma\left(\frac{1}{2}p+\beta\right)\left[1+t^2/(N-1)\right]^{\frac{1}{2}N+\beta}}.$$

Tang (1938) ha obtenido tablas de la probabilidad de aceptar la hipótesis nula (es decir, la probabilidad del error del tipo II) para varios valores del τ^2 y para niveles de significación 0.05 y 0.01. El número de grados de libertad f_1 que allí aparece es el p aquí considerado [1 (1) 8], y f_2 es $n-p+1$ [2, 4 (1) 30, 60, ∞], el parámetro de no centralización ϕ está relacionado con τ^2 mediante la expresión

$$\phi = \frac{\tau}{\sqrt{p+1}} \qquad [1\ (\tfrac{1}{2})\ 3\ (1)\ 8].$$

A modo de ejemplo, supongamos que $p=4$, $n-p+1=20$ y que queremos contrastar la hipótesis $\mu=0$ al nivel 1%. Deseamos conocer la probabilidad de aceptar la hipótesis nula cuando $\phi=2.5$ $(\tau^2=31.25)$. Dicha probabilidad es 0.227. Si pensamos que la desventaja de aceptar la hipótesis nula cuando N,μ,Σ cumplen que $\tau^2=31.25$ es menor que la desventaja de rechazarla cuando es cierta, entonces podemos encontrar razonable aceptar el test en estas condiciones. Sin embargo, si la desventaja del error del tipo I es aproximadamente igual que la del tipo II, parece razonable desear disminuir la probabilidad del error del tipo II. Para ello, adoptando un nivel $\alpha=5\%$, la probabilidad de un error del tipo II (para $\phi=2.5$) es solo 0.043.

Existen también tablas (Emma Lehmer, 1944) que proporcionan ϕ para α dado y para una probabilidad de error del tipo II dada. Estas tablas son útiles para ver qué valor de τ^2 es necesario para hacer la probabilidad de aceptar la hipótesis nula suficientemente baja cuando $\mu \neq 0$. Por ejemplo, si queremos ser capaces de rechazar la hipótesis $\mu=0$ sobre la base de una muestra para unos μ y Σ dados, podemos ser capaces de elegir N de manera que $N\mu'\Sigma^{-1}\mu = \tau^2$ sea suficientemente grande. Ahora bien, la dificultad con estas consideraciones es que normalmente no conocemos los valores de μ y Σ (y por tanto, τ^2) para los cuales queremos que la probabilidad del rechazo tenga un cierto valor.

7. ALGUNAS PROPIEDADES OPTIMAS DEL TEST T^2

Supongamos que queremos contrastar la hipótesis $\mu=0$ sobre la base de N observaciones $x_1, \dots, x_N$ de una $N(\mu,\Sigma)$. Consideremos en primer lugar la clase de los tests basados en los estadísticos $A = \sum(x_\alpha-\bar{x})(x_\alpha-\bar{x})'$ y $\bar{x}$ que son invariantes con respecto a transformaciones del tipo $A^* = CAC'$ y $\bar{x}^* = C\bar{x}$, donde C es no singular. La transformación $x_\alpha^* = Cx_\alpha$ deja al problema invariante; esto es, en términos de x_α^* contrastamos la hipótesis $E(x_\alpha^*)=0$ dado que $x_1^*, \dots, x_N^*$ son N observaciones procedentes de una población normal multivariante. Parece razonable buscar una solución que sea también invariante con respecto a estas transformaciones; es decir, buscaremos una región crítica que no se altere mediante una transformación lineal no singular (la definición de la región será la misma en diferentes sistemas de coordenadas).

TEOREMA.- Dadas las observaciones $x_1, \dots, x_N$ procedentes de una $N(\mu,\Sigma)$, de todos los tests de $\mu=0$ basados en $\bar{x}$ y $A=\sum(x_\alpha-\bar{x})(x_\alpha-\bar{x})'$ que son invariantes con respecto a transformaciones del tipo $\bar{x}^* = C\bar{x}$, $A^* = CAC'$ (C no singular), el test T^2 es uniformemente más potente.

Demostración.- Como vimos anteriormente cualquier test basado en T^2 es invariante. Por otra parte, esta función es la única invariante para la que si $f(\bar{x},A)$ es invariante, entonces $f(\bar{x},A) = f(\bar{x}^*,I)$, donde la única coordenada de $\bar{x}^*$ diferente de cero es la primera y su valor es $\sqrt{\bar{x}'A^{-1}\bar{x}}$ (Existe una matriz C, no singular, tal que $C\bar{x}=\bar{x}^*$ y $CAC'=I$). Así $f(\bar{x},A)$ depende solo de $\bar{x}'A^{-1}\bar{x}$. Por tanto un test invariante debe de estar basado en $\bar{x}'A^{-1}\bar{x}$. Finalmente, podemos aplicar el lema fundamental de Neyman-Pearson a la distribución de T^2 para encontrar el test uniformemente más potente basado en T^2, frente a la alternativa simple $\tau^2 = N\mu'\Sigma^{-1}\mu$. El test uniformemente más potente de $\tau^2=0$ está basado en el cociente de las distribuciones de T^2 cuando $\tau^2 = N\mu'\Sigma^{-1}\mu$ y cuando $\tau^2=0$. Dicho test es

Si $T^2 = N(\bar{x}-\mu_0)' S^{-1}(\bar{x}-\mu_0)$ se basa en una muestra de tamaño N de una $N(\mu,\Sigma)$, entonces $(T^2/n)\left((N-p)/p\right)$ tiene una distribución F no centrada con p y $N-p$ grados de libertad y parámetro de no centralización $N(\mu-\mu_0)'\Sigma^{-1}(\mu-\mu_0) = z^2$. La expresión de la densidad de T^2 será

$$\frac{e^{-\frac{1}{2}z^2}}{(N-1)\,\Gamma\left(\frac{1}{2}(N-p)\right)} \sum_{\beta=0}^{\infty} \frac{(z^2/2)^\beta \left[t^2/(N-1)\right]^{\frac{1}{2}p+\beta-1} \Gamma\left(\frac{1}{2}N+\beta\right)}{\beta!\,\Gamma\left(\frac{1}{2}p+\beta\right)\left[1+t^2/(N-1)\right]^{\frac{1}{2}N+\beta}}.$$

Zang (1938) ha obtenido tablas de la probabilidad de aceptar la hipótesis nula (es decir, la probabilidad del error del tipo II) para varios valores de z^2 y para niveles de significación 0.05 y 0.01. El número de grados de libertad f_1 que allí aparece es el p aquí considerado [1 (1) 8], y f_2 es $n-p+1$ [2, 4 (1) 30, 60, ∞], el parámetro de no centralización ϕ está relacionado con z^2 mediante la expresión

$$\phi = \frac{z}{\sqrt{p+1}} \qquad [1\ (\tfrac{1}{2})\ 3\ (1)\ 8].$$

A modo de ejemplo, supongamos que $p=4$, $n-p+1=20$ y que queremos contrastar la hipótesis $\mu=0$ al nivel 1%. Deseamos conocer la probabilidad de aceptar la hipótesis nula cuando $\phi=2.5$ ($z^2=31.25$). Dicha probabilidad es 0.227. Si pensamos que la desventaja de aceptar la hipótesis nula cuando N, μ, Σ cumplen que $z^2=31.25$ es menor que la desventaja de rechazarla cuando es cierta, entonces podremos encontrar razonable aceptar el test en estas condiciones. Si embargo, si la desventaja del error del tipo I es aproximadamente igual que la del tipo II, parece razonable desear disminuir la probabilidad del error del tipo II. Para ello, adoptando un nivel $\alpha=5\%$, la probabilidad de un error del tipo II (para $\phi=2.5$) es solo 0.043.

Existen también tablas (Emma Lehmer, 1944) que proporcionan ϕ para α dado y para una probabilidad de error del tipo II dada. Estas tablas son útiles para ver qué valor de z^2 es necesario para hacer la probabilidad de aceptar la hipótesis nula suficientemente baja cuando $\mu \neq 0$. Por ejemplo, si queremos ser capaces de rechazar la hipótesis $\mu=0$ sobre la base de una muestra para unos μ y Σ dados, podemos ser capaces de elegir N de manera que $N\mu'\Sigma^{-1}\mu = z^2$ sea suficientemente grande. Ahora bien, la dificultad con estas consideraciones es que normalmente no conocemos los valores de μ y Σ (y por tanto, z^2) para los cuales queremos que la probabilidad del rechazo tenga un cierto valor.

7. ALGUNAS PROPIEDADES ÓPTIMAS DEL TEST T^2

Supongamos que queremos contrastar la hipótesis $\mu=0$ sobre la base de N observaciones $x_1, \ldots, x_N$ de una $N(\mu,\Sigma)$. Consideremos en primer lugar la clase de los tests basados en los estadísticos $A = \sum(x_\alpha-\bar{x})(x_\alpha-\bar{x})'$ y $\bar{x}$ que son invariantes con respecto a transformaciones del tipo $A^* = CAC'$ y $\bar{x}^* = C\bar{x}$, donde C es no singular. La transformación $x_\alpha^* = Cx_\alpha$ deja el problema invariante; esto es, en términos de x_α^* contrastamos la hipótesis $E(x_\alpha^*)=0$ dado que $x_1^*, \ldots, x_N^*$ son N observaciones procedentes de una población normal multivariante. Parece razonable buscar una solución que sea también invariante con respecto a estas transformaciones; es decir, buscaremos una región crítica que no se altere mediante una transformación lineal no singular (la definición de la región será la misma en diferentes sistemas de coordenadas).

TEOREMA.- Dadas las observaciones $x_1, \ldots, x_N$ procedentes de una $N(\mu,\Sigma)$, de todos los tests de $\mu=0$ basados en $\bar{x}$ y $A=\sum(x_\alpha-\bar{x})(x_\alpha-\bar{x})'$ que son invariantes con respecto a transformaciones del tipo $\bar{x}^* = C\bar{x}$, $A^* = CAC'$ (C no singular), el test T^2 es uniformemente más potente.

Demostración.- Como vimos anteriormente cualquier test basado en T^2 es invariante. Por otra parte, esta función es la única invariante para la que si $f(\bar{x},A)$ es invariante, entonces $f(\bar{x},A) = f(\bar{x}^*,I)$, donde la única coordenada de $\bar{x}^*$ diferente de cero es la primera y su valor es $\sqrt{\bar{x}'A^{-1}\bar{x}}$ (Existe una matriz C, no singular, tal que $C\bar{x}=\bar{x}^*$ y $CAC'=I$). Así $f(\bar{x},A)$ depende solo de $\bar{x}'A^{-1}\bar{x}$. Por tanto un test invariante debe de estar basado en $\bar{x}'A^{-1}\bar{x}$. Finalmente, podemos aplicar el lema fundamental de Neyman-Pearson a la distribución de T^2 para encontrar el test uniformemente más potente basado en T^2, frente a la alternativa simple $z^2 = N\mu'\Sigma^{-1}\mu$. El test uniformemente más potente de $z^2=0$ está basado en el cociente de las densidades de T^2 cuando $z^2 = N\mu'\Sigma^{-1}\mu$ y cuando $z^2=0$. Dicho test es

$$c < e^{-\frac{1}{2}c^2} \sum_{\alpha=0}^{\infty} \frac{(c^2/2)^\alpha \left(t^2/n\right)^{\frac{1}{2}p+\alpha-1} \left(1+t^2/n\right)^{-\frac{1}{2}(n+1)+\alpha} \Gamma\left(\frac{1}{2}(n+1)+\alpha\right)}{\alpha!\,\Gamma\left(\frac{1}{2}p+\alpha\right)} \Big/ \frac{\left(t^2/n\right)^{\frac{1}{2}p-1}\left(1+t^2/n\right)^{-\frac{1}{2}(n+1)}\Gamma\left(\frac{1}{2}(n+1)\right)}{\Gamma\left(\frac{1}{2}p\right)} =$$

$$= \frac{\Gamma\left(\frac{1}{2}p\right)}{\Gamma\left(\frac{1}{2}(n+1)\right)} e^{-\frac{1}{2}c^2} \sum_{\alpha=0}^{\infty} \frac{(c^2/2)^\alpha \Gamma\left(\frac{1}{2}(n+1)+\alpha\right)}{\alpha!\,\Gamma\left(\frac{1}{2}p+\alpha\right)} \left(\frac{t^2/n}{1+t^2/n}\right)^\alpha.$$

La parte derecha de la desigualdad es una función estrictamente creciente de $\frac{t^2/n}{1+t^2/n}$ y por tanto de t^2. Así pues la desigualdad se reduce a una de la forma $t^2 > k$ para k elegido adecuadamente. Como no depende de la alternativa Σ^2, el test es uniformemente más potente.

DEFINICIÓN.- Una función crítica $\psi(\bar{x}, A)$ es una función con valores entre 0 y 1 (ambos incluidos) tal que $E\left(\psi(\bar{x}, A)\right) = \varepsilon$, el nivel de significación, cuando $\mu = 0$.

Un test aleatorizado existe en rechazar la hipótesis con probabilidad [redacted] $\psi(x, B)$, cuando $\bar{x} = x$, $A = B$. Un test no aleatorizado se define cuando $\psi(\bar{x}, A)$ toma únicamente los valores 0 y 1. Utilizando de manera apropiada el lema de Neyman-Pearson para funciones críticas obtenemos el siguiente corolario.

COROLARIO.- Sobre la base de las observaciones $x_1, \dots, x_N$ de $N(\mu, \Sigma)$, de todos los test aleatorizados basados en $\bar{x}$ y A que son invariantes con respecto a transformaciones $\bar{x}^* = C\bar{x}$ y $A^* = CAC'$ (C no singular), el test T^2 es uniformemente más potente.

TEOREMA.- Sobre la base de N observaciones de $N(\mu, \Sigma)$, de todos los test de $\mu = 0$ que son invariantes con respecto a transformaciones $x_\alpha^* = Cx_\alpha$ (C no singular), el test T^2 es un test uniformemente más potente; es decir, el test T^2 es al menos tan potente como cualquier otro test invariante.

Demostración.- Sea $\psi(x_1, \dots, x_N)$ una función crítica de un test invariante. Entonces

$$E\left(\psi(x_1, \dots, x_N)\right) = E_{\bar{x}, A}\left\{E\left(\psi(x_1, \dots, x_N) \mid \bar{x}, A\right)\right\}$$

Puesto que $\bar{x}, A$ son estadísticas suficiente para μ y Σ, $E\left(\psi(x_1, \dots, x_N) \mid \bar{x}, A\right)$ depende solo de $\bar{x}$ y A. El invariante tiene la misma potencia que $\psi(x_1, \dots, x_N)$. Resulta así que cada test en la clase mayor puede ser reemplazado por un test en la clase menor (los que dependen solo de $\bar{x}$ y A) que tiene la misma potencia. Aplicando el Corolario anterior se completa la demostración.

TEOREMA.- Dadas las observaciones $x_1, \dots, x_N$ de una $N(\mu, \Sigma)$, de todos los test de $\mu = 0$ basados en $\bar{x}$ y A con potencia dependiendo solo de $\tau^2 = N\mu'\Sigma^{-1}\mu$, el test T^2 es uniformemente más potente.

Demostración.- Puede verse en Anderson pags. 117, 118.

6. EL PROBLEMA DE BEHRENS-FISHER EN EL CASO MULTIVARIANTE

Vamos a dar ahora una solución análoga a la solución de Scheffé (1943) para el problema de Behrens-Fisher. Sean $\{x_\alpha^{(i)}\}$, $\alpha = 1, \dots, N_i$, $i = 1, 2$ muestras procedentes de $N(\mu^{(i)}, \Sigma_i)$, $i = 1, 2$. Deseamos contrastar la hipótesis $\mu^{(1)} = \mu^{(2)}$. La media $\bar{x}^{(1)}$ de la primera muestra se distribuye normalmente con esperanza

$$E(\bar{x}^{(1)}) = \mu^{(1)}$$

y matriz de covarianzas

$$E\left[\left(\bar{x}^{(1)} - \mu^{(1)}\right)\left(\bar{x}^{(1)} - \mu^{(1)}\right)'\right] = \frac{1}{N_1}\Sigma_1.$$

Análogamente, para la media de la segunda muestra tenemos

$$E(\bar{x}^{(2)}) = \mu^{(2)}$$

y matriz de covarianzas

$$E\left[\left(\bar{x}^{(2)} - \mu^{(2)}\right)\left(\bar{x}^{(2)} - \mu^{(2)}\right)'\right] = \frac{1}{N_2}\Sigma_2.$$

Así $\bar{x}^{(1)} - \bar{x}^{(2)}$ tiene media $\mu^{(1)} - \mu^{(2)}$ y matriz de covarianzas $\left(1/N_1\right)\Sigma_1 + \left(1/N_2\right)\Sigma_2$. Ahora bien, no podemos utilizar la técnica desarrollada anteriormente puesto que

$$\sum_{\alpha=1}^{N_1} (x_\alpha^{(1)} - \bar{x}^{(1)})(x_\alpha^{(1)} - \bar{x}^{(1)})' + \sum_{\alpha=1}^{N_2} (x_\alpha^{(2)} - \bar{x}^{(2)})(x_\alpha^{(2)} - \bar{x}^{(2)})'$$

no tiene la distribución de Wishart con matriz de covarianzas múltiplo de $\left(1/N_1\right)\Sigma_1 + \left(1/N_2\right)\Sigma_2$.

Si $N_1 = N_2 = N$, podemos utilizar el test T^2 de una manera obvia. Veamos. Sea $y_\alpha = x_\alpha^{(1)} - x_\alpha^{(2)}$ (suponiendo que la numeración de las observaciones en las dos muestras es independiente de las mismas observaciones). Entonces y_α se distribuye normalmente con media $\mu^{(1)} - \mu^{(2)}$ y matriz de covarianzas $\Sigma_1 + \Sigma_2$, e independientemente de y_β $(\beta \neq \alpha)$. Sea $\bar{y} = \frac{1}{N}\sum_\alpha y_\alpha = \bar{x}^{(1)} - \bar{x}^{(2)}$ y definamos S mediante

$$(N-1)S = \sum_{\alpha=1}^{N} (y_\alpha - \bar{y})(y_\alpha - \bar{y})' = \sum_{\alpha=1}^{N} \left(x_\alpha^{(1)} - x_\alpha^{(2)} - \bar{x}^{(1)} + \bar{x}^{(2)}\right)\left(x_\alpha^{(1)} - x_\alpha^{(2)} - \bar{x}^{(1)} + \bar{x}^{(2)}\right)'.$$

Entonces

$$T^2 = N\,\bar{y}'\,S^{-1}\,\bar{y}$$

es adecuado para contrastar la hipótesis $\mu_y = \mu^{(1)} - \mu^{(2)} = 0$ y T^2 tiene la distribución T^2 con $N-1$ grados de libertad. Hay que señalar en este punto que si hubiéramos admitido que $\Sigma_1 = \Sigma_2$ habríamos utilizado un estadístico T^2 con $2N-2$ grados de libertad; hemos pues perdido $N-1$ grados de libertad en la construcción de un test que es independiente de las dos matrices de covarianzas.

Volvamos nuevamente al caso en que $N_1 \neq N_2$. Por conveniencia supongamos $N_1 < N_2$. Definimos entonces

$$y_\alpha = x_\alpha^{(1)} - \sqrt{\frac{N_1}{N_2}}\,x_\alpha^{(2)} + \frac{1}{\sqrt{N_1 N_2}} \sum_{\beta=1}^{N_1} x_\beta^{(2)} - \frac{1}{N_2}\sum_{\gamma=1}^{N_2} x_\gamma^{(2)} \qquad \alpha = 1, \ldots, N_1$$

El valor esperado de y_α será

$$E(y_\alpha) = \mu^{(1)} - \sqrt{\frac{N_1}{N_2}}\,\mu^{(2)} + \frac{N_1}{\sqrt{N_1 N_2}}\,\mu^{(2)} - \frac{N_2}{N_2}\,\mu^{(2)} = \mu^{(1)} - \mu^{(2)}$$

La matriz de covarianzas de y_α e y_β es

$$E\left(y_\alpha - E(y_\alpha)\right)\left(y_\beta - E(y_\beta)\right)' = E\left[\left(x_\alpha^{(1)} - \mu^{(1)}\right) - \sqrt{\frac{N_1}{N_2}}\left(x_\alpha^{(2)} - \mu^{(2)}\right) + \frac{1}{\sqrt{N_1 N_2}}\sum_{\gamma=1}^{N_1}\left(x_\gamma^{(2)} - \mu^{(2)}\right) - \frac{1}{N_2}\sum_{\gamma=1}^{N_2}\left(x_\gamma^{(2)} - \mu^{(2)}\right)\right] \cdot$$

$$\left[\left(x_\beta^{(1)} - \mu^{(1)}\right)' - \sqrt{\frac{N_1}{N_2}}\left(x_\beta^{(2)} - \mu^{(2)}\right)' + \frac{1}{\sqrt{N_1 N_2}}\sum_{\gamma=1}^{N_1}\left(x_\gamma^{(2)} - \mu^{(2)}\right)' - \frac{1}{N_2}\sum_{\gamma=1}^{N_2}\left(x_\gamma^{(2)} - \mu^{(2)}\right)'\right] =$$

$$= \delta_{\alpha\beta}\,\Sigma_1 + \frac{N_1}{N_2}\,\delta_{\alpha\beta}\,\Sigma_2 + \Sigma_2\left(-2\,\frac{1}{N_2} + \frac{2}{N_2}\sqrt{\frac{N_1}{N_2}} + \frac{N_1}{N_1 N_2} - 2\,\frac{N_1}{\sqrt{N_1 N_2}\,N_2} + \frac{N_2}{N_2^2}\right) =$$

$$= \delta_{\alpha\beta}\left(\Sigma_1 + \frac{N_1}{N_2}\,\Sigma_2\right)$$

Así un estadístico apropiado para contrastar la hipótesis $\mu^{(1)} - \mu^{(2)} = 0$, que tiene una distribución T^2 con $N_1 - 1$ grados de libertad, es

$$T^2 = N_1\,\bar{y}'\,S^{-1}\,\bar{y}$$

donde

$$\bar{y} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} y_\alpha = \bar{x}^{(1)} - \bar{x}^{(2)}$$

y

$$(N_1 - 1)S = \sum_{\alpha=1}^{N_1} (y_\alpha - \bar{y})(y_\alpha - \bar{y})' = \sum_{\alpha=1}^{N_1}\left(x_\alpha^{(1)} - \bar{x}^{(1)} - \sqrt{\frac{N_1}{N_2}}\left(x_\alpha^{(2)} - \frac{1}{N_1}\sum_{\beta=1}^{N_1} x_\beta^{(2)}\right)\right)\left(x_\alpha^{(1)} - \bar{x}^{(1)} - \sqrt{\frac{N_1}{N_2}}\left(x_\alpha^{(2)} - \frac{1}{N_1}\sum_{\beta=1}^{N_1} x_\beta^{(2)}\right)\right)'$$

lo que en términos de $u_\alpha = x_\alpha^{(1)} - \sqrt{N_1/N_2}\,x_\alpha^{(2)}$, $\alpha = 1, \ldots, N_1$ puede escribirse como

$$(N_1-1)S = \sum_{\alpha=1}^{N_1} (u_\alpha - \bar{u})(u_\alpha - \bar{u})'$$

donde $\bar{u} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} u_\alpha$.

Este procedimiento ha sido sugerido por Scheffé (1943), que usó de esta manera el caso univariante. Demostró Scheffé que en dicho caso, esta técnica daba los intervalos de confianza más pequeños utilizando la t de Student. La ventaja del método estriba en el hecho de utilizar $\bar{x}^{(1)} - \bar{x}^{(2)}$, que es el estadístico más importante para contrastar hipótesis acerca de $\mu^{(1)} - \mu^{(2)}$. El sacrificio de observaciones a la hora de estimar la matriz de covarianzas no es desdeñable importante. La extensión del método de Scheffé al caso multivariante se debe a Bennett (1951).

El método es generalizable a situaciones con más de dos poblaciones. Sean $\{x_\alpha^{(i)}\}$, $\alpha = 1, \dots, N_i$, $i = 1, \dots, q$ muestras provenientes de $N(\mu^{(i)}, \Sigma_i)$ $i = 1, \dots, q$, respectivamente. Consideramos la siguiente hipótesis

$$\sum_{i=1}^{q} \beta_i \mu^{(i)} = \mu.$$

donde los β_i son escalares dados, al igual que el vector μ. Si los N_i son distintos, sea N_1 la menor de ellos.

Definamos

$$y_\alpha = \beta_1 x_\alpha^{(1)} + \sum_{i=2}^{q} \beta_i \sqrt{\frac{N_1}{N_i}} \left(x_\alpha^{(i)} - \frac{1}{N_1}\sum_{\beta=1}^{N_1} x_\beta^{(i)} + \frac{1}{\sqrt{N_1 N_i}} \sum_{\gamma=1}^{N_i} x_\gamma^{(i)} \right).$$

Entonces

$$E(y_\alpha) = \beta_1 \mu^{(1)} + \sum_{i=2}^{q} \left(\beta_i \sqrt{\frac{N_1}{N_i}} \left(\mu^{(i)} - \frac{1}{N_1} N_1 \mu^{(i)} + \frac{N_i}{\sqrt{N_1 N_i}} \mu^{(i)} \right) \right) = \sum_{i=1}^{q} \beta_i \mu^{(i)}.$$

y

$$E\left(y_\alpha - E(y_\alpha)\right)\left(y_\beta - E(y_\beta)\right)' = \delta_{\alpha\beta} \left(\sum_{i=1}^{q} \frac{\beta_i^2 N_1}{N_i} \Sigma_i \right).$$

Sean $\bar{y}$ y S definidas mediante

$$\bar{y} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} y_\alpha = \sum_{i=1}^{q} \beta_i \bar{x}^{(i)}, \quad \text{con } \bar{x}^{(i)} = \frac{1}{N_i}\sum_{\beta=1}^{N_i} x_\beta^{(i)},$$

$$(N_1 - 1) S = \sum_{\alpha=1}^{N_1} (y_\alpha - \bar{y})(y_\alpha - \bar{y})'.$$

Entonces

$$T^2 = N_1 (\bar{y} - \mu)' S^{-1} (\bar{y} - \mu)$$

es un estadístico apropiado para contrastar la hipótesis propuesta, y cuando la hipótesis es cierta este estadístico tiene una distribución T^2 con $N_1 - 1$ grados de libertad. Mediante un cambio del tipo

$$u_\alpha = \sum_{i=1}^{q} \beta_i \sqrt{\frac{N_1}{N_i}}\, x_\alpha^{(i)} \qquad \alpha = 1, \dots, N_1$$

S puede expresarse de la forma

$$(N_1 - 1) S = \sum_{\alpha=1}^{N_1} (u_\alpha - \bar{u})(u_\alpha - \bar{u})'.$$

Para terminar un último problema que puede resolverse mediante este procedimiento es el de contrastar la hipótesis de que dos subvectores de un determinado vector normal multivariante tienen igual media. Sea

$$x = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}$$

distribuido normalmente con media

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}$$

y matriz de covarianzas

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Supongamos que $x^{(1)}$ y $x^{(2)}$ tienen cada uno q componentes. Entonces $x^{(1)} - x^{(2)}$ se distribuye normalmente con media $\mu^{(1)} - \mu^{(2)}$ y matriz de covarianzas

$$E\left((x^{(1)} - \mu^{(1)}) - (x^{(2)} - \mu^{(2)})\right)\left((x^{(1)} - \mu^{(1)}) - (x^{(2)} - \mu^{(2)})\right)' = \Sigma_{11} - \Sigma_{21} - \Sigma_{12} + \Sigma_{22}.$$

Para contrastar la hipótesis $\mu^{(1)} = \mu^{(2)}$ utilizaremos el estadístico T^2 definido mediante:

$$T^2 = N\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' \left(S_{11} - S_{21} - S_{12} + S_{22}\right)^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)$$

donde la media muestral y la matriz de covarianzas muestral han sido precisadas de forma similar a como lo han sido μ y Σ.

EJEMPLOS

① En una investigación de las respuestas al test de Wechsler para la inteligencia de adultos aplicado a hombres y mujeres de edad avanzada se obtuvieron las siguientes medias para las respuestas verbal y de comportamiento

$$\begin{bmatrix} \bar{x}_V \\ \bar{x}_C \end{bmatrix} = \begin{bmatrix} 55.24 \\ 34.97 \end{bmatrix}$$

siendo la muestra de un tamaño $N = 101$ y los sujetos de edades comprendidas entre los 60 y 64 años. La matriz de covarianzas muestral de las respuestas fue

$$S = \begin{bmatrix} 210.54 & 126.99 \\ 126.99 & 119.68 \end{bmatrix}$$

Deseamos contrastar, a un nivel $\alpha = 0.01$ la hipótesis nula de que las observaciones proviene de una población con vector media

$$\mu_0 = \begin{bmatrix} 60 \\ 50 \end{bmatrix}.$$

El test está basado en la aplicación del estadístico T^2, cuyo valor vendría dado por

$$T^2 = N(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) = 101 \cdot [55.24 - 60,\ 34.97 - 50] \begin{bmatrix} 0.01319 & -0.01400 \\ -0.01400 & 0.02321 \end{bmatrix} \begin{bmatrix} 55.24 - 60 \\ 34.97 - 50 \end{bmatrix} = 357.43$$

el valor $\left[T^2/(N-1)\right]\left[(N-p)/p\right]$ es una F de Snedecor con p y $N-p$ grados de libertad, por tanto

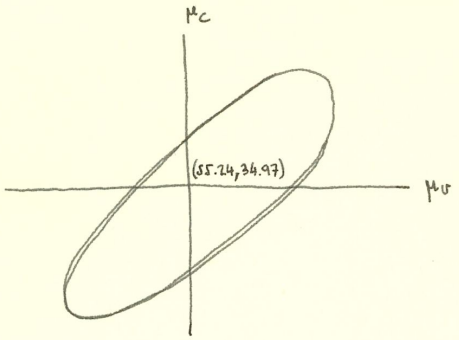
$$F = \frac{357.43}{100} \cdot \frac{99}{2} = 176.93$$

como $F_{0.01,2,99} \simeq 4.98$, y $F \gg F_{0.01,2,99}$ rechazaremos la hipótesis formulada. A la vista de los resultados obtenidos cabe pensar que la población origen de nuestros datos debe poseer un vector media menor que el postulado en H_0.

Podemos también hacer uso de T^2 para obtener una región al 99% de confianza para el vector media de la población. Necesitaremos el valor de $F_{0.01,2,99}$ que podemos obtener por interpolación entre los de $F_{0.01,2,60}$ y $F_{0.01,2,120}$, lo que da para $F_{0.01,2,99}$ un valor 4.83.

La elipse de confianza vendría dada por la ecuación

$$1.33\,(\mu_V - 55.24)^2 - 2.83\,(\mu_V - 55.24)(\mu_C - 34.97) + 2.34\,(\mu_C - 34.97)^2 \leq 9.76$$



(2) Cuarenta y nueve ancianos que participaron en un estudio interdisciplinario del envejecimiento humano fueron clasificados en dos categorías de diagnóstico, "factor renal presente" y "factor renal ausente", sobre la base de un intenso examen psiquiátrico. La escala de inteligencia de adultos de Wechsler fue administrada a todos los sujetos por un investigador independiente, y ciertos subtests mostraron grandes diferencias entre los dos grupos. A la vista de estos resultados queremos contrastar la hipótesis de que los grupos provienen de una misma población.

Los datos muestrales se recogen en la siguiente tabla (datos medios)

	Grupo	
Subtest	Factor renal ausente $N_1 = 37$	Factor renal $N_2 = 12$
Información	12.57	8.75
Similaridades	9.57	5.33
Aritmética	11.49	8.50
Figuras completas	7.97	4.75

La matriz de covarianzas compensada viene dada por

$$S = \begin{bmatrix} 11.2583 & 9.4042 & 7.1489 & 3.3830 \\ & 13.5318 & 7.3830 & 2.5532 \\ & & 11.5744 & 2.6170 \\ & & & 5.8085 \end{bmatrix}$$

El valor de T^2 para dos muestras, en este caso vale

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) = 22.05$$

y el correspondiente valor de F, s

$$F = \frac{N_1 + N_2 - p - 1}{(N_1 + N_2 - 2)\,p} T^2 = \frac{44}{47.4}\, 22.05 = 5.16.$$

Bajo la hipótesis de igualdad de medias la probabilidad de exceder semejante valor de F sería menor que 0.005, por tanto rechazaremos la hipótesis formulada para niveles de significación convencionales del 5% o del 1%.

(3) Se estudian los cambios de concentración en el plasma de ácidos grasos libres (FFA) y glucosa sanguínea (G) en 12 pacientes esquizofrénicos y 13 voluntarios normales, después de inyecciones intramusculares de insulina. Los cambios resumen en la siguiente tabla

	Cambio medio	
	Esquizofrénicos	Normales
G, mg, %	-25.6	-31.1
FFA, meqiv/litro	-0.06	-0.15
Matrices de suma de cuadrados y productos cruzados	$\begin{bmatrix} 3455 & 9.9492 \\ 9.9492 & 0.1105 \end{bmatrix}$	$\begin{bmatrix} 3509 & -3.2408 \\ -3.2408 & 0.0865 \end{bmatrix}$
Matriz de covarianzas "pooled"	$S = \begin{bmatrix} 302 & 0.292 \\ 0.292 & 0.00856 \end{bmatrix}$	

Bajo el supuesto de que los valores de FFA y G tienen una distribución normal bivariante con una matriz de covarianzas que no se afecta por diagnósticos psiquiátricos, deseamos contrastar la hipótesis de que los vectores medios en los sujetos normales y en los esquizofrénicos son equivalentes. Observemos que la hipótesis de la igualdad de la matriz de covarianzas aparece válida respecto de los elementos de la diagonal principal de la matriz de covarianzas muestral, aunque no tanto para los términos fuera de ésta, es decir la covariancia entre ambas variables.

El valor de T^2 para esta hipótesis

$$H_0 : \mu_1 - \mu_2 = 0$$

viene dado por $T^2 = 6.03$ y el valor de F asociado es $F = 2.88$. Observemos que este valor está comprendido entre los de $F_{0.1,2,22} = 2.56$ y $F_{0.05,2,22} = 3.44$. Ello significa que a un nivel $\alpha = 0.05$ no rechazaríamos la hipótesis nula, mientras que la aceptaríamos para un nivel superior $\alpha = 0.1$.

(4) Los niveles de FFA en sangre se midieron en 15 sujetos normales que se sometieron voluntariamente a sesiones de hipnosis. Durante estas sesiones se les pidió que experimentaran miedo, angustia y cólera y se les midió la cantidad (o mejor el cambio en la concentración) de FFA después de cada una de estas situaciones ficticias. Bajo el supuesto de que cada una de estas situaciones produce el mismo grado de cambio, los investigadores desean comprobar que no había cambios apreciables entre las diferentes sesiones de stress a las que se sometió a los individuos. Los cambios medios de FFA fueron

$$\bar{x}_1 = 2.699 \qquad \bar{x}_2 = 2.178 \qquad \bar{x}_3 = 2.558 .$$

El problema puede resolverse como un caso particular de los problemas de simetría antes presentados. Podemos efectuar el siguiente cambio

$$y_{i1} = x_{i1} - x_{i2} \qquad y_{i2} = x_{i1} - x_{i3} \qquad , \quad i = 1, \ldots, 15$$

es decir

$$y = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

La hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3$ es ahora equivalente a $H_0' : C\mu = 0$, donde $C\mu$ es la media de y y μ es la media de x. La matriz de covarianzas muestral para y es

$$S = \begin{bmatrix} 1.7343 & 1.1666 \\ 1.1666 & 2.7733 \end{bmatrix} \quad \text{y} \quad S^{-1} = \begin{bmatrix} 0.8041 & -0.3382 \\ -0.3382 & 0.5029 \end{bmatrix}$$

El valor de $T^2 = N \bar{y}' S^{-1} \bar{y}$ viene dado por $T^2 = 2.64$ y el correspondiente valor de $F = 1.24$ y como $F_{0.5,2,13} = 3.80$, aceptaremos la hipótesis formulada, es decir, que los cambios de FFA en los diferentes situaciones de stress son iguales.

Capítulo 5

Clasificación de observaciones

CLASIFICACION DE OBSERVACIONES

1. EL PROBLEMA DE LA CLASIFICACION

El problema de la clasificación surge cuando un investigador realiza varias medidas sobre un individuo y desea clasificarlo en una de varias categorías sobre la base de estas medidas. El investigador no puede identificar al individuo directamente con una categoría y debe por tanto recurrir a las medidas en cuestión. En muchos casos puede suponerse que hay un número finito de categorías o poblaciones a las que el individuo puede pertenecer y cada una de las poblaciones se caracteriza por una distribución de probabilidad de las medidas. Así, un individuo a considerado como una observación aleatoria de una población. La cuestión a: Dado un individuo con ciertas medidas, de qué población proviene?

El problema de la clasificación puede ser considerado como un problema de funciones estadísticas de decisión. Tenemos un cierto número de hipótesis: cada una de las hipótesis a que la distribución de la observación a sea dada. Debemos aceptar una de estas hipótesis y rechazar las otras. Si sólo existen dos poblaciones, nos encontramos ante un problema elemental de contrastar una hipótesis de una distribución específica frente a otra.

En algunos casos, las categorías están especificadas de antemano en el sentido de que la distribución de probabilidad de las medidas es completamente conocida. En otros casos, la forma de cada una de las distribuciones puede ser conocida, pero los parámetros de la distribución deben estimarse a partir de una muestra de la población.

Consideremos un ejemplo de clasificación. Los alumnos que desean ingresar en la universidad son sometidos a una batería de tests; el vector de los resultados es un conjunto de medidas x. El estudiante puede ser miembro de una población que consiste en aquellos estudiantes que están capacitados para sacar adelante con éxito los estudios, potencialmente claro está, o bien pertenecer a la población de los que no completarán con éxito dichos estudios. El problema es clasificar un estudiante sobre la base de sus resultados en el examen de admisión.

En este capítulo desarrollaremos la teoría de clasificación en términos generales para aplicarla posteriormente a aquellos casos en que se involucren poblaciones con distribución normal.

2. NORMAS PARA UNA BUENA CLASIFICACION

Consideraciones previas.- En la construcción de un proceso de clasificación se desea minimizar la probabilidad de mala clasificación, o más específicamente, se desea minimizar, por término medio, los malos efectos de clasificaciones erróneas. Iremos precisando más este concepto. Por conveniencia consideraremos el caso de dos categorías. Luego trataremos el caso general.

Supongamos que un individuo a una observación de una de estas dos poblaciones, π_1 o π_2. La clasificación de una observación depende del vector de medidas $x' = (x_1 \cdots x_p)$ sobre este individuo. Construiremos una regla de manera que si un individuo está caracterizado mediante ciertos conjuntos de valores de $x_1 \cdots x_p$ será clasificado como perteneciente a π_1; si tiene otros valores será clasificado como perteneciente a π_2.

Podemos considerar una observación como un punto en un espacio p-dimensional. Dividimos el espacio en dos regiones. Si la observación está en R_1, calificamos al individuo como de π_1, si la observación cae en R_2 lo clasificaremos como de π_2.

Siguiendo un procedimiento de este tipo, el investigador puede cometer dos tipos de error. Clasificar al individuo como de π_2 cuando realmente procede de π_1 y viceversa. Necesitamos conocer la indeseabilidad relativa de estas dos clases de malaclasificación. Designaremos por $C(2/1) > 0$ el coste del primero de los errores y por $C(1/2) > 0$ el otro. Estos costes pueden venir medidos en cualquier tipo de unidades. Cuanto más tarde veremos lo verdaderamente importante será la relación entre los dos costes.

En la tabla que presentamos a continuación aparecen descritas las anteriores situaciones. Obviamente, un buen procedimiento

		Decisión del estadístico	
		π_1	π_2
Población de origen	π_1	0	$C(2/1)$
	π_2	$C(1/2)$	0

es aquel que minimiza, en algún sentido, el coste de las clasificaciones erróneas.

Caso de dos poblaciones

Veremos a continuación dos formas de definir el coste mínimo. En un caso supondremos que tenemos probabilidades a priori de ambas poblaciones. Sean q_1 y q_2 las probabilidades de que una observación provenga de las poblaciones Π_1 y Π_2 respectivamente. Las propiedades probabilísticas de Π_1 serán especificadas mediante una función de distribución. Por comodidad consideraremos el caso en que la distribución de probabilidad es absolutamente continua y existe la función de densidad, aunque el caso discreto requiere un tratamiento muy similar. Sean $p_i(x)$, $i=1,2$ las densidades de probabilidad de las poblaciones Π_i, $i=1,2$. Si tenemos una región R_1 de clasificación en la que asignamos la observación a Π_1, la probabilidad de clasificar correctamente una observación que realmente proviene de Π_1 es

$$P(1|1,R) = \int_{R_1} p_1(x)\,dx ,$$

y la probabilidad de mala clasificación de una observación que procede de Π_1 es

$$P(2|1,R) = \int_{R_2} p_1(x)\,dx .$$

Análogamente tendríamos

$$P(2|2,R) = \int_{R_2} p_2(x)\,dx$$

y

$$P(1|2,R) = \int_{R_1} p_2(x)\,dx .$$

Puesto que la probabilidad de extraer una observación de Π_1 es q_1, la probabilidad de extraer una observación de Π_1 y clasificarla correctamente es $q_1 \cdot P(1|1,R)$. Análogamente obtendríamos las probabilidades de las tres restantes situaciones posibles.

¿Cuál sería la pérdida esperada o el coste medio de mala clasificación? Es la suma de los productos de los costes de cada una de las clasificaciones erróneas por la probabilidad de que ocurra; a saber

$$C(2|1)\cdot P(2|1,R)\cdot q_1 + C(1|2)\cdot P(1|2,R)\cdot q_2$$

Es este coste medio el que se desea minimizar. Es decir, se quiere dividir el espacio en dos regiones R_1 y R_2 de manera que la pérdida esperada sea lo menos posible. Un procedimiento que minimiza la anterior expresión para unas q_1 y q_2 dadas es conocido como un procedimiento Bayes.

El otro caso a considerar es aquel en el que no se conocen probabilidades a priori. En este caso la pérdida esperada si la observación proviene de Π_1 es

$$C(2|1)\cdot P(2|1,R) = r(1,R);$$

y para observaciones procedentes de Π_2,

$$C(1|2)\cdot P(1|2,R) = r(2,R).$$

No sabemos de donde proviene nuestra observación y no conocemos las probabilidades de que pertenezca a una u otra población.

Un procedimiento R es *al menos tan bueno* como un procedimiento R^* si $r(1,R) \le r(1,R^*)$ y $r(2,R) \le r(2,R^*)$. R es mejor que R^* si al menos una de estas desigualdades es estricta. Normalmente no hay procedimientos que sean mejor que cualquier otro o al menos tan bueno como cualquier otro. Un procedimiento R se dice que es *admisible* si no existe ningún otro que sea mejor; estamos interesados en la clase de los procedimientos admisibles. Demostraremos que bajo ciertas condiciones esta clase es la misma que la clase de los procedimientos Bayes. Una clase de procedimientos es *completa* si para cualquier procedimiento exterior existe uno en la clase que es mejor; una clase es *esencialmente completa* si para cualquier procedimiento exterior existe uno en la clase que al menos tan bueno como aquel. Una clase *mínima completa* (si existe) es una clase completa tal que ningún subconjunto propio es una clase completa; una definición similar puede darse para las clase *mínimas esencialmente completas*. Bajo ciertas condiciones demostraremos que la clase

admisible y esencialmente completa. Para simplificar la discusión consideraremos que dos procedimientos son 5 (2)
iguales si difieren solo en conjuntos de probabilidad nula. De hecho, a través de cuanto requirá haremos afirmaciones que son ciertas "excepción hecha de conjuntos de probabilidad nula" aunque no lo hagamos constar explícitamente.

Un principio que usualmente conduce a un procedimiento único es el principio del minimax. Un procedimiento es minmax si la máxima pérdida esperada, $r(i,R)$ es un mínimo. Desde un punto de vista conservador, este puede ser considerado como un procedimiento óptimo. Un estudio detallado de estos conceptos puede encontrarse en los textos de Wald (1950) y Blackwell y Girshick (1954)

3. PROCEDIMIENTOS DE CLASIFICACION EN UNA DE DOS POBLACIONES CUANDO SE CONOCEN LAS DISTRIBUCIONES DE PROBABILIDAD

Probabilidades a priori conocidas

Puesto que conocemos las probabilidades a priori podemos definir la distribución conjunta de la población y del conjunto de variables observadas. La probabilidad de que una observación provenga de Π_1 y de que cada una de sus variables sea menor que la correspondiente componente de y viene dada por

$$\int_{-\infty}^{y_p} \cdots \int_{-\infty}^{y_1} q_1 p_1(x)\, dx_1 \cdots dx_p$$

Podemos también definir la probabilidad condicional de que una observación provenga de una cierta población dados los valores de las variables observadas. Por ejemplo, la probabilidad condicional de venir de la población Π_1, dada una observación x, es

$$\frac{q_1 \cdot p_1(x)}{q_1 \cdot p_1(x) + q_2 \cdot p_2(x)}.$$

Supongamos por un momento que $C(1/2) = C(2/1) = 1$. Entonces la pérdida esperada es

$$q_1 \cdot \int_{R_2} p_1(x) \cdot dx + q_2 \int_{R_1} p_2(x) \cdot dx.$$

Esta es también la probabilidad de una mala clasificación; por tanto, deseamos minimizar la probabilidad de una clasificación errónea.

Para una observación dada X minimizamos la probabilidad de una mala clasificación asignando la población que tiene una probabilidad condicional mayor. Si

$$\frac{q_1 p_1(x)}{q_1 \cdot p_1(x) + q_2 \cdot p_2(x)} \geq \frac{q_2 \cdot p_2(x)}{q_1 \cdot p_1(x) + q_2 \cdot p_2(x)}$$

elegiremos la población Π_1. En cualquier otro caso elegimos Π_2. Puesto que minimizamos la probabilidad de mala clasificación en cada punto, la minimizamos sobre el espacio entero. Así, la regla es

$$\boxed{\begin{aligned} R_1 &: \quad q_1 \cdot p_1(x) \geq q_2 \cdot p_2(x) \\ R_2 &: \quad q_1 \cdot p_1(x) < q_2 \cdot p_2(x) \end{aligned}}$$

En caso de igualdad la observación puede ser asignada indistintamente a una u otra población. Si $q_1 \cdot p_1(x) + q_2 \cdot p_2(x) = 0$ para un x dado, también en este caso el punto puede asignarse a cualquiera de las regiones.

Probemos ahora formalmente que esta regla es el mejor procedimiento. Para cualquier procedimiento $R^* = (R_1^*, R_2^*)$ la probabilidad de mala clasificación es

$$q_2 \int_{R_2^*} p_1(x)\, dx + q_2 \int_{R_1^*} p_2(x)\, dx = \int_{R_2^*} [q_1 p_1(x) - q_2 p_2(x)]\, dx + q_2 \int p_2(x)\, dx.$$

El segundo sumando del segundo miembro es una constante y en cuanto al primer sumando se minimiza si R_2^* incluye los puntos x tales que $q_1 \cdot p_1(x) - q_2 \cdot p_2(x) < 0$ y excluye los puntos para los que $q_1 \cdot p_1(x) - q_2 p_2(x) > 0$. Si suponemos que

$$P\left\{ \frac{p_1(x)}{p_2(x)} = \frac{q_2}{q_1} \,\middle|\, \Pi_i \right\} = 0 \qquad i = 1, 2$$

entonces el procedimiento Bayes es único excepto para conjuntos de probabilidad nula.

Hemos con esto resuelto el problema cuando los costs son iguales a iguales a uno. Supongamos ahora que no ocurre así. Se trataría en este caso de minimizar

$$C(2/1)\cdot q_1 \int_{R_2} p_1(x)\,dx + C(1/2)\cdot q_2 \int_{R_1} p_2(x)\,dx$$

eligiremos entonces R_1 y R_2 de manera que

$$R_1: \; [C(2/1)\cdot q_1]\, p_1(x) \geq [C(1/2)\cdot q_2]\cdot p_2(x)$$

$$R_2: \; [C(2/1)\cdot q_1]\, p_1(x) < [C(1/2)\cdot q_2]\, p_2(x)$$

puesto que $C(2/1)$ y $C(1/2)$ son constantes no negativas. Otra forma de definir R_1 y R_2 sería

$$R_1: \; \frac{p_1(x)}{p_2(x)} \geq \frac{C(1/2)\cdot q_2}{C(2/1)\cdot q_1}$$

$$R_2: \; \frac{p_1(x)}{p_2(x)} < \frac{C(1/2)\cdot q_2}{C(2/1)\cdot q_1}$$

TEOREMA.- Si q_1 y q_2 son las probabilidades a priori de extraer una observación de la población Π_1 con probabilidad (densidad) $p_1(x)$ y de la población Π_2 con densidad de probabilidad $p_2(x)$, respectivamente, y si los costes de mala clasificación de una observación de Π_1 como procedente de Π_2 es $C(2/1)$ y de una observación de Π_2 como procedente de Π_1 es $C(1/2)$, entonces las regiones de clasificación R_1 y R_2, definidas mediante las anteriores expresiones, minimizan el coste esperado. Si además

$$P\left\{ \frac{p_1(x)}{p_2(x)} = \frac{q_2\, C(1/2)}{q_2\, C(2/1)} \Big/ \Pi_i \right\} = 0 \qquad , i = 1,2$$

entonces el procedimiento es único casi por todos puntos.

Cuando no se conocen probabilidades a priori

En muchos ejemplos de clasificación el estadístico no puede asignar probabilidades a priori a las dos poblaciones. En este caso trataremos de encontrar la clase de los procedimientos admisibles, es decir, el conjunto de procedimientos que no puede ser mejorado.

Probaremos primero que un procedimiento Bayes es admisible. Sea $R = (R_1, R_2)$ un procedimiento Bayes para q_1 y q_2 dados; existirá un procedimiento $R^* = (R_1^*, R_2^*)$ tal que $P(1/2, R^*) \leq P(1/2, R)$ y $P(2/1, R^*) \leq P(2/1, R)$ con al menos una desigualdad estricta. Puesto que R es Bayes

$$q_1 \cdot P(2/1, R) + q_2 \cdot P(1/2, R) \leq q_1 \cdot P(2/1, R^*) + q_2 \cdot P(1/2, R^*).$$

Esta desigualdad puede escribirse

$$q_1 \left[P(2/1, R) - P(2/1, R^*) \right] \leq q_2 \left[P(2/2, R^*) - P(1/2, R) \right].$$

Supongamos que $q_2 > 0$. Entonces si $P(1/2, R^*) \leq P(1/2, R)$, la parte derecha de la desigualdad será menor o igual a 0, por tanto $P(2/1, R) \leq P(2/1, R^*)$. Si $q_2 \neq 0$, razonando de forma análoga llegamos a que $P(1/2, R) \leq P(1/2, R^*)$. Así R^* no es mejor que R y R es admisible. Si $q_2 = 0$, entonces de la desigualdad obtenemos $0 \leq q_1 \left[P(1/2, R^*) - P(1/2, R) \right]$. Para un procedimiento Bayes esto supone que R_1 contiene aquellos x tales que $p_2(x) = 0$. Por tanto $P(1/2, R) = 0$ y si R^* ha de ser mejor que R, deberá ocurrir que $P(1/2, R^*) = 0$. En cuanto a $P(2/1, R)$ si $P\left(p_2(x) = 0 / \Pi_1 \right) = 0$, entonces $P(2/1, R) = P\left(p_2(x) > 0 / \Pi_1 \right) = 1$. Ahora bien si $P(1/2, R^*) = 0$, entonces R_1^* contiene solo puntos para los que $p_2(x) = 0$, pero $P(2/1, R^*) = P\left(R_1^* / \Pi_1 \right) = P\left(p_2(x) > 0 / \Pi_1 \right) = 1$, y por tanto R^* no es mejor que R.

TEOREMA.- Si $P\left(p_2(x) = 0 / \Pi_1 \right) = 0$ y $P\left(p_1(x) = 0 / \Pi_2 \right) = 0$ entonces todo procedimiento Bayes es admisible.

Probemos ahora el inverso, es decir, que todos procedimientos admisibles son Bayes. Supondremos

$$P\left\{\frac{p_1(x)}{p_2(x)} = k/\pi_i\right\} = 0 \quad , \quad i=1,2 \; ; \; 0 \le k \le \infty \quad \left[\frac{p_1(x)}{p_2(x)} = \infty \text{ significa que } p_2(x) = 0\right].$$

Entonces para cualquier q_2 el procedimiento Bayes es único. Además la función de distribución de $p_1(x)/p_2(x)$ para π_1 y π_2 es continua.

Sea R un procedimiento admisible. Entonces existe un k tal que

$$P(2/1, R) = P\left\{\frac{p_1(x)}{p_2(x)} \le k/\pi_1\right\} = P(2/1, R^*)$$

donde R^* es el procedimiento Bayes correspondiente a $q_2/q_1 = k$ [es decir, $q_2 = 1/(1+k)$]. Puesto que R es admisible, tendremos $P(1/2, R) \le P(1/2, R^*)$. Sin embargo el teorema anterior afirma que R^* es admisible y por tanto $P(1/2, R) \ge P(1/2, R^*)$ de aquí la igualdad de ambos. En definitiva R es un procedimiento Bayes y por la unicidad de estos procedimientos debe ser precisamente R^*.

TEOREMA.- Si $P\left\{p_1(x)/p_2(x) = k/\pi_i\right\} = 0$, $i=1,2$; $0 \le k \le \infty$, entonces cualquier procedimiento admisible es un procedimiento Bayes.

De acuerdo con lo visto en la demostración de este último teorema la clase de los procedimientos Bayes es completa. Evidentemente dicha característica no basta por sí sola para definir una clase mínimal completa puesto que coincide con la clase de los procedimientos admisibles.

Consideremos finalmente el procedimiento minimax. Sea $P(i/j, q_2) = P(i/j, R)$, donde R es el procedimiento Bayes correspondiente a q_2. $P(i/j, q_2)$ es una función continua de q_2. Así, $P(2/1, q_2)$ varía de 1 a 0 cuando q_2 lo hace de 0 a 1, mientras que $P(1/2, q_2)$ lo hace, al mismo tiempo, de 0 a 1. Existirá pues un valor de q_2, q_2^*, tal que $P(2/1, q_2^*) = P(1/2, q_2^*)$. Esta es la solución minimax y es óptima, puesto que si existiera otro procedimiento R^* tal que $\max\{P(2/1, R^*), P(1/2, R^*)\} \le P(2/1, q_2^*) = P(1/2, q_2^*)$ esto contradice el hecho de ser toda solución Bayes admisible.

4. CLASIFICACION EN UNA DE DOS POBLACIONES NORMALES MULTIVARIANTES CONOCIDAS

Utilizaremos el procedimiento general descrito anteriormente para el caso de dos poblaciones normales multivariantes con igual matriz de covarianza, a saber, $N(\mu^{(1)}, \Sigma)$ y $N(\mu^{(2)}, \Sigma)$, donde $\mu^{(i)\prime} = (\mu_1^{(i)}, \dots, \mu_p^{(i)})$ es el vector de medias de la i-ésima población $(i=1,2)$ y Σ la matriz de covarianzas común a ambas poblaciones. La densidad de la población i es

$$p_i(x) = \frac{1}{(2\pi)^{\frac{1}{2}p}\,|\Sigma|^{1/2}} \cdot \exp\left[-\tfrac{1}{2}(x-\mu^{(i)})'\Sigma^{-1}(x-\mu^{(i)})\right].$$

La razón de densidades viene dada por

$$\frac{p_1(x)}{p_2(x)} = \frac{\exp\left[-\frac{1}{2}(x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)})\right]}{\exp\left[-\frac{1}{2}(x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})\right]} = \exp\left[-\frac{1}{2}\{(x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)}) - (x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})\}\right].$$

La región de clasificación en π_1, R_1, es el conjunto de x's para los que este cociente es $\ge k$ (k elegido adecuadamente). Puesto que la función logarítmica es una función monótona creciente, la desigualdad puede ser escrita en términos de los logaritmos como

$$-\frac{1}{2}\{(x-\mu^{(1)})'\Sigma^{-1}(x-\mu^{(1)}) - (x-\mu^{(2)})'\Sigma^{-1}(x-\mu^{(2)})\} \ge \log k.$$

lo que desarrollado conduce a

$$-\frac{1}{2}\left[x'\Sigma^{-1}x - x'\Sigma^{-1}\mu^{(1)} - \mu^{(1)\prime}\Sigma^{-1}x + \mu^{(1)\prime}\Sigma^{-1}\mu^{(1)} - x'\Sigma^{-1}x + x'\Sigma^{-1}\mu^{(2)} + \mu^{(2)\prime}\Sigma^{-1}x - \mu^{(2)\prime}\Sigma^{-1}\mu^{(2)}\right] \ge \log k$$

y arreglado adecuadamente, llegamos a

$$x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \tfrac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \ge \log k.$$

El primer miembro de la izquierda es conocido como la función discriminante, y es una función lineal de las componentes del vector de las observaciones.

Aplicando los resultados antes obtenidos llegamos al siguiente teorema.

TEOREMA.- Si Π_i tiene una densidad normal multivariante $p_i(x)$, $i=1,2$, las mejores regiones de clasificación vienen dadas por

$$R_1: \; x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \geq \log k$$

$$R_2: \; x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(1)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) < \log k.$$

Si las probabilidades a priori q_1 y q_2 son conocidas, entonces k viene dado por

$$k = \frac{q_2 C(1/2)}{q_1 C(2/1)}.$$

El caso particular en que las dos poblaciones son igualmente probables y los costes son iguales, $k=1$ y $\log k = 0$. Entonces la región de clasificación en Π_1 es

$$R_1: \; x'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) \geq \frac{1}{2}(\mu^{(1)}+\mu^{(2)})\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}).$$

Si no sabemos nada acerca de q_1 y q_2 podemos seleccionar $\log k = C$, sobre la base de igualar las pérdidas esperadas debidas a malclasificación.

Sea X una observación aleatoria. Deseamos encontrar la distribución de

$$U = X'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}),$$

sobre el supuesto de que X se distribuye $N(\mu^{(1)}, \Sigma)$ y luego sobre el supuesto de que X lo hace $N(\mu^{(2)}, \Sigma)$. Cuando X es $N(\mu^{(1)}, \Sigma)$, U es normal con media

$$E_1(U) = \mu^{(1)'}\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) - \frac{1}{2}(\mu^{(1)}+\mu^{(2)})\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) = \frac{1}{2}(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}),$$

y varianza

$$\mathrm{var}_1(U) = E_1\left[(\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(X-\mu^{(1)})(X-\mu^{(1)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)})\right] = (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}).$$

Sea α la distancia entre las dos poblaciones, a saber

$$\alpha = (\mu^{(1)}-\mu^{(2)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}).$$

Luego U se distribuye $N(\frac{1}{2}\alpha, \alpha)$ cuando X es $N(\mu^{(1)}, \Sigma)$. Si X es $N(\mu^{(2)}, \Sigma)$ entonces

$$E_2(U) = \frac{1}{2}(\mu^{(2)}-\mu^{(1)})'\Sigma^{-1}(\mu^{(1)}-\mu^{(2)}) = -\frac{1}{2}\alpha$$

siendo la varianza la misma, puesto que esta depende de los momentos de segundo orden de X y estos son iguales en ambos casos. Así, U se distribuye $N(-\frac{1}{2}\alpha, \alpha)$ en este segundo caso.

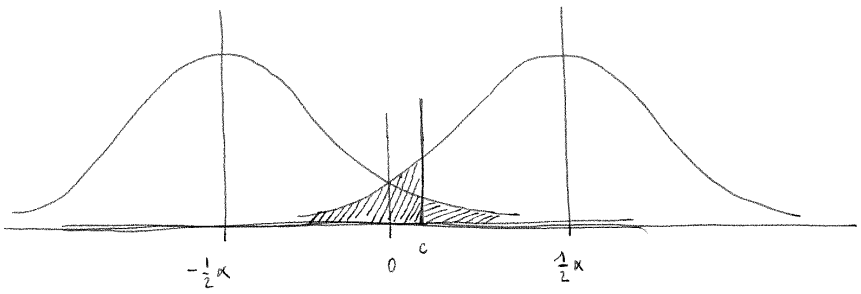
La probabilidad de malclasificación, si la observación procede de Π_1 es

$$P(2/1) = \int_{-\infty}^{C} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2}(z-\frac{1}{2}\alpha)^2/\alpha} dz = \int_{-\infty}^{(C-\frac{1}{2}\alpha)/\sqrt{\alpha}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$$

y la probabilidad de malclasificación cuando la observación procede de Π_2 es

$$P(1/2) = \int_{C}^{\infty} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2}(z+\frac{1}{2}\alpha)^2/\alpha} dz = \int_{(C+\frac{1}{2}\alpha)/\sqrt{\alpha}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy$$

ambas probabilidades aparecen representadas, en las figuras de la página siguiente, como las partes sombreadas de cada una de las colas adyacentes.



Para la solución minimax elegiríamos c de manera que

$$C(1/2)\int_{(c+\frac{1}{2}\alpha)/\sqrt{\alpha}}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy = C(2/1)\int_{-\infty}^{(c-\frac{1}{2}\alpha)/\sqrt{\alpha}}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy .$$

TEOREMA.- Si las poblaciones π_i tienen densidades multinormales $p_i(x)$, $i=1,2$, la solución minimax viene dada por aquellas regiones que satisfacen las inecuaciones del anterior teorema con $\log k = C$ determinado para que verifique la igualdad anterior, en la que $C(i/j)$ representan los dos costes de mala clasificación.

Hemos de señalar que si los dos costes $C(i/j)$ son iguales, entonces $c=0$ y la probabilidad de mala clasificación es

$$\int_{\sqrt{\alpha}/2}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy .$$

En caso de costes desiguales, c deberá ser determinado mediante un método aproximado a partir de las tablas de la normal.

Observemos finalmente que los desplazamientos que aparecen en expresión que determina R_1 y R_2 constituyen al vector

$$\delta = \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}).$$

Es interesante señalar que $x'\delta$ es la función lineal que maximiza

$$\frac{[E_1(x'd) - E_2(x'd)]^2}{\text{var}(x'd)}$$

para todas las posibles elecciones de d. El numerador de este cociente es

$$(1)\quad [\mu^{(1)'}d - \mu^{(2)'}d]^2 = d'[(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})']d$$

siendo el denominador

$$(2)\quad d'E[(x - E(x))(x - E(x))']d = d'\Sigma d .$$

Deseamos maximizar (1) manteniendo (2) constante. Si λ es un multiplicador de lagrange, estamos buscando el máximo de

$$d'[(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})']d - \lambda(d'\Sigma d - 1).$$

Derivando respecto a las componentes de d e igualando a cero obtendremos

$$2[(\mu^{(1)} - \mu^{(2)})(\mu^{(1)} - \mu^{(2)})']d = 2\lambda\Sigma d .$$

Puesto que $(\mu^{(1)} - \mu^{(2)})'d$ es un escalar, k, podemos expresar la anterior igualdad como

$$\mu^{(1)} - \mu^{(2)} = \frac{\lambda}{k}\Sigma d$$

y la solución es proporcional a δ.

Señalemos finalmente que si tenemos una muestra de tamaño N de cada una de las poblaciones π_1 y π_2, podemos utilizar la media muestral y clasificarla como procedente de $N[\mu^{(1)}, \frac{1}{N}\Sigma]$ o $N[\mu^{(2)}, \frac{1}{N}\Sigma]$.

5. CLASIFICACION EN UNA DE DOS POBLACIONES NORMALES MULTIVARIANTES CUANDO LOS PARAMETROS SON ESTIMADOS

El criterio de clasificación.- Hasta ahora hemos supuesto que las poblaciones eran conocidas exactamente. En muchas de las aplicaciones de esta teoría las poblaciones no son conocidas, pero deben ser inferidas a partir de las muestras, una de cada población. Nos ocuparemos a continuación del caso en que tenemos una muestra de cada una de las poblaciones, ambas normales multivariantes, y deseamos utilizar esta información para clasificar otras observaciones como procedentes de una de las dos poblaciones.

Supongamos que tenemos una muestra $x_1^{(1)}$ — $x_{N_1}^{(1)}$ de $N(\mu^{(1)}, \Sigma)$ y una muestra $x_1^{(2)}$ — $x_{N_2}^{(2)}$ de $N(\mu^{(2)}, \Sigma)$. La mejor estimación de $\mu^{(1)}$ es $\bar{x}^{(1)} = \sum_1^{N_1} x_\alpha^{(1)}/N_1$ y de $\mu^{(2)}$ es $\bar{x}^{(2)} = \sum_1^{N_2} x_\alpha^{(2)}/N_2$, y la Σ es S definida mediante

$$(N_1+N_2-2)\,S = \sum_1^{N_1} (x_\alpha^{(1)} - \bar{x}^{(1)})(x_\alpha^{(1)} - \bar{x}^{(1)})' + \sum_1^{N_2} (x_\alpha^{(2)} - \bar{x}^{(2)})(x_\alpha^{(2)} - \bar{x}^{(2)})'.$$

Sustituimos estos valores por los correspondiente parámetros en la expresión obtenida anteriormente para obtener

$$x' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}) - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}).$$

El primer termino de esta suma es la función discriminante basada en dos muestras [sugerida por Fisher (1936)]. Es una función lineal que tiene la mayor "variancia entre muestras" relativa a la "variancia dentro de las muestras". La idea es utilizar la anterior expresión como criterio de clasificación de la misma forma a como fue utilizada en el caso de conocer los parametros de las poblaciones.

Cuando las poblaciones eran conocidas, razonamos que el criterio de clasificación es el mejor en el sentido de que minimizaba la pérdida esperada en el caso de probabilidades a priori conocidas y generaba la clase de procedimientos admisibles cuando las probabilidades a priori no son conocidas. Ahora no podemos justificarnos en el mismo sentido. Sin embargo, parece intuitivamente razonable que el criterio proporcione buenos resultados. Otro criterio será indicado en un punto posterior.

Supongamos que tenemos una muestra x_1 — x_N de una u otra población, π_1 o π_2, y deseamos llevar a cabo un clasificación como un todo y no elemento a elemento. Definimos entonces S como

$$(N_1+N_2+N-3)\,S = \sum_1^{N_1} (x_\alpha^{(1)} - \bar{x}^{(1)})(x_\alpha^{(1)} - \bar{x}^{(1)})' + \sum_1^{N_2} (x_\alpha^{(2)} - \bar{x}^{(2)})(x_\alpha^{(2)} - \bar{x}^{(2)})' + \sum_1^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$$

donde

$$\bar{x} = \frac{1}{N}\sum_1^N x_\alpha .$$

Entonces el criterio es

$$\left[\bar{x} - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})\right]' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}).$$

Puede demostrarse que a mayores valores de N, menores son las probabilidades de mala clasificación.

La distribución del criterio.- Sea

$$V = X' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}) - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}) = \left[X - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})\right]' S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)})$$

con X, $\bar{x}^{(1)}$, $\bar{x}^{(2)}$, S aleatorios.

La distribución de V es extremadamente complicada. Depende de los tamaños de las muestras y de los valores desconocidos de $\mu_\alpha^{(1)}$ y $\mu_\alpha^{(2)}$. Sean ahora

$$Z = X - \frac{1}{2}(\bar{x}^{(1)} + \bar{x}^{(2)})$$

$$Y = \bar{x}^{(1)} - \bar{x}^{(2)},$$

entonces

$$V = Z' S^{-1} Y .$$

La esperanza de Y es $\mu^{(1)} - \mu^{(2)}$ y la matriz de covariancias es $[1/N_1 + 1/N_2]\,\Sigma$. Z se distribuye normalmente con media

$$E_1(z) = \frac{1}{2}(\mu^{(1)} - \mu^{(2)})$$

si $\underline{x}$ procede de π_1, y

$$E_2(z) = \frac{1}{2}(\mu^{(2)} - \mu^{(1)})$$

si $\underline{x}$ procede de π_2. En ambos casos la matriz de covarianzas es $\left[1 + 1/4N_1 + 1/4N_2\right]\Sigma$. La matriz de covarianzas entre z e y es

$$-\left(\frac{1}{2N_1} + \frac{1}{2N_2}\right)\Sigma .$$

Si $N_1 = N_2$ esta matriz de covarianzas es nula. En este caso se ve fácilmente que la distribución de V para $\underline{x}$ procedente de π_1 es la misma que la de $-V$ para $\underline{x}$ procedente de π_2. Así, si $V \geq 0$ es la región de clasificación en π_1, la probabilidad de clasificar erróneamente a $\underline{x}$ cuando procede de π_1 es igual a la probabilidad de hacerlo, igualmente, mal cuando procede de π_2.

La distribución de V ha sido estudiada por Anderson, Stitt-Greaves y Wald.

La distribución asintótica del criterio.- En el caso de grandes muestras para ambas poblaciones podemos aplicar la teoría de las distribuciones límite. Puesto que $\bar{\underline{x}}^{(i)}$ es la media de la muestra con N_i variables independientes (distintas muestras) todas ellas con distribución $N(\mu^{(i)}, \Sigma)$, sabemos que

$$\lim_{N_1 \to \infty} \bar{\underline{x}}^{(1)} = \mu^{(1)} \quad (p).$$

análogamente

$$\lim_{N_2 \to \infty} \bar{\underline{x}}^{(2)} = \mu^{(2)} \quad (p)$$

y

$$\lim S = \Sigma \quad (p)$$

cuando N_1, N_2 o ambos tienden a ∞. Y de esta última igualdad se tiene

$$\lim S^{-1} = \Sigma^{-1} \quad (p)$$

límite en probabilidad

puesto que los límites en probabilidad de sumas, diferencias, productos y cocientes de variables aleatorias son las sumas, productos y cocientes de los límites en probabilidad siempre que el límite de cada denominador sea diferente de cero. Entonces, aplicando todo esto

$$\lim_{N_1, N_2 \to \infty} S^{-1}\left(\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)}\right) = \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \quad (p)$$

$$\lim_{N_1, N_2 \to \infty} \left(\bar{\underline{x}}^{(1)} + \bar{\underline{x}}^{(2)}\right)' S^{-1}\left(\bar{\underline{x}}^{(1)} - \bar{\underline{x}}^{(2)}\right) = (\mu^{(1)} + \mu^{(2)})' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \quad (p).$$

Se sigue que la distribución asintótica de V es la distribución de V estudiada en el párrafo anterior. Para muestras suficientemente grandes de π_1 y π_2 podemos utilizar el criterio como si conociéramos exactamente las poblaciones cometiendo sólo un pequeño error.

TEOREMA.- Sea V definido como antes. La distribución asintótica de V cuando $N_1 \to \infty$ y $N_2 \to \infty$ es $N\left(\frac{1}{2}\alpha, \alpha\right)$ si $\underline{x}$ procede de la población π_1 y es $N\left(-\frac{1}{2}\alpha, \alpha\right)$ si $\underline{x}$ procede de la población π_2. Donde $\alpha = (\mu^{(1)} - \mu^{(2)})' \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$.

Otra derivación del criterio.- Una obtención convenientemente nemotécnica del criterio es el uso de la regresión de una variable ficticia dada por Fisher en su trabajo de 1936. Sea

$$y_\alpha^{(1)} = \frac{N_2}{N_1 + N_2}, \quad \alpha = 1, \ldots, N_1 \qquad y_\alpha^{(2)} = \frac{-N_1}{N_1 + N_2} \quad \alpha = 1, \ldots, N_2 .$$

Se trata de encontrar la regresión sobre las variables $x_\alpha^{(i)}$ eligiendo b de manera que se minimice

$$\sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}\left[y_\alpha^{(i)} - b'(x_\alpha^{(i)} - \bar{x})\right]^2$$

donde

$$\bar{x} = \left[N_1 \bar{x}^{(1)} + N_2 \bar{x}^{(2)}\right] / (N_1 + N_2) .$$

Las "ecuaciones normales" son

$$\sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}(x_\alpha^{(i)} - \bar{x})(x_\alpha^{(i)} - \bar{x})' b = \sum_{i=1}^{2}\sum_{\alpha=1}^{N_i} y_\alpha^{(i)}(x_\alpha^{(i)} - \bar{x}) = \frac{N_1 N_2}{N_1+N_2}\left[(\bar{x}^{(1)} - \bar{x}) - (\bar{x}^{(2)} - \bar{x})\right] = \frac{N_1 N_2}{N_1+N_2}(\bar{x}^{(1)} - \bar{x}^{(2)}) .$$

La matriz que multiplica a b puede ser escrita como

$$\sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}(x_\alpha^{(i)} - \bar{x})(x_\alpha^{(i)} - \bar{x})' = \sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}(x_\alpha^{(i)} - \bar{x}^{(i)})(x_\alpha^{(i)} - \bar{x}^{(i)})' + N_1(\bar{x}^{(1)} - \bar{x})(\bar{x}^{(1)} - \bar{x})' + N_2(\bar{x}^{(2)} - \bar{x})(\bar{x}^{(2)} - \bar{x})' =$$

$$= \sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}(x_\alpha^{(i)} - \bar{x}^{(i)})(x_\alpha^{(i)} - \bar{x}^{(i)})' + \frac{N_1 N_2}{N_1+N_2}(\bar{x}^{(1)} - \bar{x}^{(2)})(\bar{x}^{(1)} - \bar{x}^{(2)})' .$$

En definitiva las anteriores ecuaciones normales pueden ser escritas como

$$Gb = (\bar{x}^{(1)} - \bar{x}^{(2)})\left[\frac{N_1 N_2}{N_1+N_2} - \frac{N_1 N_2}{N_1+N_2}(\bar{x}^{(1)} - \bar{x}^{(2)})' b\right]$$

donde

$$G = \sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}(x_\alpha^{(i)} - \bar{x}^{(i)})(x_\alpha^{(i)} - \bar{x}^{(i)})' .$$

Puesto que $(\bar{x}^{(1)} - \bar{x}^{(2)})' b$ es un escalar, vemos que la solución b de la anterior expresión es proporcional a

$$S^{-1}(\bar{x}^{(1)} - \bar{x}^{(2)}) .$$

El criterio de la razón de verosimilitud.- Otro criterio que puede ser utilizado en la clasificación es el de la razón de verosimilitud. Consideramos que contrastamos la hipótesis nula de que $x, x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ son extraídos de $N(\mu^{(1)}, \Sigma)$ y $x_1^{(2)}, \ldots, x_{N_2}^{(2)}$ lo son de $N(\mu^{(2)}, \Sigma)$, frente a la alternativa de que $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ pertenecen a $N(\mu^{(1)}, \Sigma)$, mientras que $x, x_1^{(2)}, \ldots, x_{N_2}^{(2)}$ lo hacen a $N(\mu^{(2)}, \Sigma)$, con $\mu^{(1)}, \mu^{(2)}$ y Σ no especificados. Bajo la primera hipótesis los estimadores máximos verosímiles de $\mu^{(1)}, \mu^{(2)}$ y Σ son

$$\hat{\mu}_1^{(1)} = (N_1 \bar{x}^{(1)} + x)/(N_1 + 1)$$

$$\hat{\mu}_1^{(2)} = \bar{x}^{(2)}$$

$$\hat{\Sigma}_1 = \frac{1}{N_1 + N_2 + 1}\left[\sum_{\alpha=1}^{N_1}(x_\alpha^{(1)} - \hat{\mu}_1^{(1)})(x_\alpha^{(1)} - \hat{\mu}_1^{(1)})' + (x - \hat{\mu}_1^{(1)})(x - \hat{\mu}_1^{(1)})' + \sum_{\alpha=1}^{N_2}(x_\alpha^{(2)} - \hat{\mu}_1^{(2)})(x_\alpha^{(2)} - \hat{\mu}_1^{(2)})'\right]$$

puesto que

$$\sum_{\alpha=1}^{N_1}(x_\alpha^{(1)} - \hat{\mu}_1^{(1)})(x_\alpha^{(1)} - \hat{\mu}_1^{(1)})' + (x - \hat{\mu}_1^{(1)})(x - \hat{\mu}_1^{(1)})' = \sum_{\alpha=1}^{N_1}(x_\alpha^{(1)} - \bar{x}^{(1)})(x_\alpha^{(1)} - \bar{x}^{(1)})' + N_1(\bar{x}^{(1)} - \hat{\mu}_1^{(1)})(\bar{x}^{(1)} - \hat{\mu}_1^{(1)})' + (x - \hat{\mu}_1^{(1)})(x - \hat{\mu}_1^{(1)})' =$$

$$= \sum_{\alpha=1}^{N_1}(x_\alpha^{(1)} - \bar{x}^{(1)})(x_\alpha^{(1)} - \bar{x}^{(1)})' + \frac{N_1}{N_1+1}(x - \bar{x}^{(1)})(x - \bar{x}^{(1)})'$$

podemos escribir $\hat{\Sigma}_1$ de la forma

$$\hat{\Sigma}_1 = \frac{1}{N_1 + N_2 + 1}\left[G + \frac{N_1}{N_1+1}(x - \bar{x}^{(1)})(x - \bar{x}^{(1)})'\right]$$

donde G es la del apartado anterior. Bajo los supuestos de la hipótesis alternativa encontramos (por consideraciones de simetría)

que los estimadores máximo verosímiles de los parámetros son

$$\hat{\mu}_2^{(1)} = \bar{x}^{(1)}$$

$$\hat{\mu}_2^{(2)} = (N_2\bar{x}^{(2)} + x)/(N_2+1)$$

$$\hat{\Sigma}_2 = \frac{1}{N_1+N_2+1}\left[C + \frac{N_2}{N_2+1}(x-\bar{x}^{(2)})(x-\bar{x}^{(2)})'\right].$$

El criterio de la razón de verosimilitud es, por tanto, la potencia $(N_1+N_2+1)/2$-ésima de

$$\frac{|\hat{\Sigma}_2|}{|\hat{\Sigma}_1|} = \frac{\left|C + \frac{N_2}{N_2+1}(x-\bar{x}^{(2)})(x-\bar{x}^{(2)})'\right|}{\left|C + \frac{N_1}{N_1+1}(x-\bar{x}^{(1)})(x-\bar{x}^{(1)})'\right|}$$

que puede también ser escrito de la forma

$$\frac{1 + \frac{N_2}{N_2+1}(x-\bar{x}^{(2)})'C^{-1}(x-\bar{x}^{(2)})}{1 + \frac{N_1}{N_1+1}(x-\bar{x}^{(1)})'C^{-1}(x-\bar{x}^{(1)})}$$

La región de clasificación en π_1 contendrá en aquellos puntos para los que el cociente anterior es mayor que una cantidad dada.

6. CLASIFICACIÓN EN UNA DE VARIAS POBLACIONES

Consideremos ahora el problema de clasificar una observación en una de varias poblaciones. Extenderemos las consideraciones hechas en los apartados anteriores al caso de más de dos poblaciones. Sean $\pi_1, \dots, \pi_m$, m poblaciones con funciones de densidad $p_1(x), \dots, p_m(x)$ respectivamente. Deseamos dividir el espacio de observaciones en m regiones exhaustivas y mutuamente excluyentes, $R_1, \dots, R_m$. Si una observación cae en R_i diremos que procede de π_i. Sean $C(j/i)$ los costes de mala clasificación de una observación que procede de la población π_i como procedente de π_j. La probabilidad de esta clasificación errónea es

$$P(j/i, R) = \int_{R_j} p_i(x)\,dx.$$

Supongamos que tenemos probabilidades a priori de las poblaciones, $q_1, \dots, q_m$. Entonces la pérdida esperada es

$$\sum_{i=1}^{m} q_i \left\{ \sum_{\substack{j=1 \\ i\neq j}}^{m} C(j/i)\, P(j/i, R) \right\}.$$

Deseamos elegir $R_1, \dots, R_m$ para hacer mínima esta expresión.

Puesto que tenemos probabilidades a priori para las poblaciones, podemos definir la probabilidad condicional de que una observación proceda de la población π_i dados unos valores de las componentes del vector x. Esta probabilidad viene dada por

$$\frac{q_i \cdot p_i(x)}{\sum_{k=1}^{m} q_k \cdot p_k(x)}.$$

Si clasificamos la observación como perteneciente a π_j, la pérdida esperada es

$$\sum_{\substack{i=1 \\ i\neq j}}^{m} \frac{q_i \cdot p_i(x)}{\sum_{k=1}^{m} q_k \cdot p_k(x)} \cdot C(j/i)$$

Minimizaremos la pérdida esperada en estas condiciones, si elegimos j de manera que minimice la anterior expresión; es decir, consideramos

$$\sum_{\substack{i=1 \\ i\neq j}}^{m} q_i \cdot p_i(x)\, C(j/i)$$

para todos los j y seleccionamos aquel j que proporciona el mínimo. (En el caso de igualdad la elección es irrelevante). Este procedimiento asigna el punto x a una de las R_j. Siguiendo este procedimiento para cada x, definiremos nuestras regiones $R_1, \ldots, R_m$. El procedimiento de clasificación entonces es clasificar una observación como procedente de π_j si cae en R_j.

TEOREMA.- Si q_i es la probabilidad a priori de extraer una observación de una población π_i (con densidad $p_i(x)$, $i=1,\ldots,m$) y el coste de mal clasificar una observación de π_i como procedente de π_j es $C(j/i)$, entonces las regiones de clasificación $R_1, \ldots, R_m$, que minimizan el coste esperado, se definen asignando x a R_k si

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i \cdot p_i(x)\, C(k/i) < \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i \cdot p_i(x)\, C(j/i) \quad , \quad j=1,\ldots,m \ , \ j\neq k.$$

Si esta desigualdad se verifica para todos los índices, excepto para h de ellos en los que se convierte en una igualdad, el punto en cuestión puede ser asignado a cualquiera de los $h+1$ posibles correspondientes.

Si la probabilidad de que se verifique la igualdad, en lugar de la desigualdad escrita, es cero para cada k y j bajo π_i (para cada i), entonces el procedimiento es único casi por todas partes.

Verifiquemos ahora este resultado. Sea

$$h_j(x) = \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i \cdot p_i(x)\, C(j/i) .$$

La pérdida esperada de un procedimiento R es

$$\sum_{j=1}^{m} \int_{R_j} h_j(x)\, dx = \int h(x)\, dx$$

donde $h(x) = h_j(x)$ para x en R_j. Para el procedimiento Bayes descrito en el teorema $h_i(x)$ es $h^*(x) = \min_i h_i(x)$. Así la diferencia entre la pérdida esperada para cualquier procedimiento R y la de R^* es

$$\int [h(x) - \min_i h_i(x)]\, dx = \sum_j \int_{R_j} [h_j(x) - \min_i h_i(x)]\, dx \geq 0 ,$$

verificándose la igualdad sólo si $h_j(x) = \min_i h_i(x)$ para x en R_j, excepción hecha de conjuntos de probabilidad cero.

Veamos como aplicar este método en aquella situación para la que $C(j/i) = 1$, $\forall i,j$ $i\neq j$. Dentro de R_k

$$\sum_{\substack{i=1 \\ i\neq k}}^{m} q_i \cdot p_i(x) < \sum_{\substack{i=1 \\ i\neq j}}^{m} q_i \cdot p_i(x) \qquad j\neq k .$$

Restando $\displaystyle\sum_{\substack{i=1 \\ i\neq j,k}}^{m} q_i \cdot p_i(x)$ de ambos miembros de la desigualdad, tendremos

$$q_j \cdot p_j(x) < q_k \cdot p_k(x) \qquad j\neq k .$$

En este caso el punto x está en R_k si k es el índice para el que $q_i \cdot p_i(x)$ es un máximo, es decir π_k es la población más probable.

Supongamos ahora que no tenemos probabilidades a priori, entonces no podemos definir una pérdida esperada minimal para un procedimiento de clasificación. Sin embargo, podemos definir una pérdida esperada sobre la condición de que las observaciones provienen de una población dada. La pérdida esperada si la observación proviene de π_i es

$$\sum_{\substack{j=1 \\ j\neq i}}^{m} C(j/i) \cdot P(j/i,R) = r(i,R)$$

Un procedimiento R es al menos tan bueno como R^* si $r(i,R) \leq r(i,R^*)$, $i=1,\ldots,m$; R es mejor si al menos

una desigualdad estricta. R es admisible si no hay ningún procedimiento R^* que mejor. Una clase de procedimientos es completa si para cualquier procedimiento R exterior a la clase existe un procedimiento R^* en la clase que es mejor.

Veamos ahora que un procedimiento Bayes es admisible. Sea R un procedimiento Bayes y R^* otro procedimiento cualquiera. Puesto que R es Bayes

$$\sum_{i=1}^{m} q_i r(i,R) \le \sum_{i=1}^{m} q_i r(i,R^*).$$

Supongamos que $r(i,R^*) \le r(i,R)$, $i=2,\dots,m$ y $q_i > 0$. Entonces

$$q_1 [r(1,R) - r(1,R^*)] \le \sum_{i=2}^{m} q_i [r(i,R^*) - r(i,R)] \le 0.$$

y $r(1,R^*) \not> r(1,R)$. Análogamente, si $q_i > 0$ y $r(j,R^*) \le r(j,R)$ para $j \neq i$, entonces $r(i,R) \le r(i,R^*)$. En definitiva, R^* no puede ser mejor que R y R es admisible.

TEOREMA.- Si $q_i > 0$, $i=1,\dots,m$, entonces R es un procedimiento admisible.

Supongamos ahora que $C(i/j) = 1$, $i \neq j$ y $P\{P_i(x) = 0/\pi_j\} = 0$. Esta última condición implica que los $p_i(x)$ son positivos sobre el mismo conjunto (excepción hecha de conjuntos de probabilidad nula). Supongamos ahora que $q_i = 0$ para $i=1,\dots,t$ y $q_i > 0$ para $i=t+1,\dots,m$. Entonces para la solución Bayes, R_i $(i=1,\dots,t)$ es vacía (excepto para conjuntos de probabilidad nula) como se comprueba fácilmente de la definición dada para R_i bajo esta situación particular, anteriormente (sólo significa $p_m(x) = 0$, $x \in R_i$). Se sigue entonces que $r(i,R) = \sum_{j \neq i} P(j/i,R) = 1 - P(i/i,R) = 1$ para $i=1,\dots,t$. Entonces $R_{t+1},\dots,R_m$ es una solución Bayes para el problema que subsumía a $p_{t+1}(x),\dots,p_m(x)$ y $q_{t+1},\dots,q_m$. Del teorema que acabamos de enunciar se sigue que no hay ningún procedimiento R^* para el que $P(i/i,R^*) = 0$, $i=1,\dots,t$, que sea mejor que el procedimiento Bayes. Consideremos ahora un procedimiento R^* tal que R_i^* incluya un conjunto de probabilidad distinta de cero tal que $P(1/1,R^*) > 0$. Para que R^* fuera mejor que R

$$P(i/i,R) = \int_{R_i} p_i(x)\,dx \le \int_{R_i^*} p_i(x)\,dx = P(i/i,R^*) \qquad , i = 1,\dots,m$$

En tal caso, un procedimiento R^{**} donde R_i^{**} es vacío para $i=1,\dots,t$, $R_i^{**} = R_i^*$, $i=t+1,\dots,m-1$ y $R_m^{**} = R_m^* \cup R_1^* \cup \dots \cup R_t^*$ daría riesgos tales como

$$P(i/i,R^{**}) = 0 \qquad i=1,\dots,t$$
$$P(i/i,R^{**}) = P(i/i,R^*) \ge P(i/i,R) \qquad i=t+1,\dots,m-1$$
$$P(m/m,R^{**}) > P(m/m,R^*) \ge P(m/m,R),$$

entonces $(R_{t+1}^{**},\dots,R_m^{**})$ sería mejor que $(R_{t+1},\dots,R_m)$ para el problema de decisión sobre $m-t$ poblaciones, lo que contradice la discusión precedente.

TEOREMA.- Si $C(i/j) = 1$, $i \neq j$ y $P\{P_i(x) = 0/\pi_j\} = 0$, entonces todo procedimiento Bayes es admisible.

Probemos ahora que los procedimientos admisibles son procedimientos Bayes. Restringiremos nuestra atención al caso $m=3$. Supondremos que

$$P\left\{\frac{P_i(x)}{P_j(x)} = K/\pi_k\right\} = 0 \qquad i \neq j \ , \ 0 \le K < \infty .$$

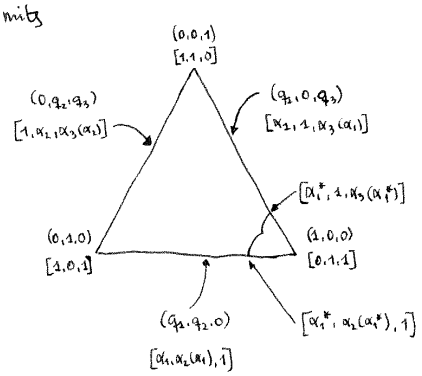
Esto implica que las distribuciones de probabilidad de $P_i(x)/P_j(x)$, para cualquier π_k, son continuas y que la conjunta de dos de sus cocientes es también continua.

Sea $\alpha_i(R) = 1 - P(i/i,R)$ la probabilidad de efectuar una decisión errónea cuando se utiliza el procedimiento R y se muestrea en π_i. Cuando R es un procedimiento Bayes, $\alpha_i(R)$ es una función de q_1, q_2, q_3, sea decir $\alpha(q_1,q_2,q_3)$. Es una función continua de q_1, q_2, q_3; por ejemplo

$$\alpha(q_1,q_2,q_3) = 1 - P\left\{\frac{P_2(x)}{P_1(x)} \le \frac{q_2}{q_1}, \frac{P_3(x)}{P_1(x)} \le \frac{q_3}{q_1} \,\middle|\, \pi_1\right\},$$

siendo continuas las distribuciones de $\frac{P_2(x)}{P_1(x)}$ y $\frac{P_3(x)}{P_1(x)}$. Es conveniente pensar en (q_1,q_2,q_3) como las coordenadas baricéntricas de

en punto. los límites



del espacio de tripletes y los valores de las funciones sobre dichos límites se indican en la figura.

Sea ahora R^* un procedimiento admisible, y sea $\alpha_i(R^*) = \alpha_i^*$. Demostraremos que R^* es un procedimiento Bayes. Consideremos la totalidad de los procedimientos Bayes para los que (q_1, q_2, q_3) dan lugar a $\alpha_1(q_1,q_2,q_3) = \alpha_1^*$. Cuando $q_3 = 0$, tenemos en efecto un problema de dos decisiones y entonces $\alpha_2 = \alpha_2(\alpha_1^*)$ que es el menor de los α_2 dado $\alpha_1 = \alpha_1^*$ (obtenido de los resultados para el caso de dos poblaciones); sea $\alpha_2(\alpha_1^*) \le \alpha_2^*$; y $\alpha_3 = 1$. Análogamente, cuando $q_2 = 0$, $\alpha_3 = \alpha_3(\alpha_1^*) \le \alpha_3^*$ y $\alpha_2 = 1$. El conjunto de puntos q_1, q_2, q_3 para los que $\alpha_1(q_1,q_2,q_3) = \alpha_1^*$ es una curva continua[1] desde el punto con $[\alpha_1^*, \alpha_2(\alpha_1^*), 1]$ hasta el punto con $[\alpha_1^*, 1, \alpha_3(\alpha_1^*)]$. Puesto que α_2 varía continuamente desde $\alpha_2(\alpha_1^*)$ hasta 1, existirá un punto donde $\alpha_2 = \alpha_2^*$. Existe un procedimiento Bayes $\bar{R}$ tal que $\alpha_1(\bar{R}) = \alpha_1^*$ y $\alpha_2(\bar{R}) = \alpha_2^*$. Puesto que $\bar{R}$ es admisible, por el teorema anterior, $\alpha_3(\bar{R}) \le \alpha_3^*$. Pero puesto que R^* es admisible $\alpha_3(\bar{R}) = \alpha_3^*$. Por la unicidad de las soluciones Bayes tenemos que $R^* = \bar{R}$.

TEOREMA.- Si $P\left\{\frac{p_i(x)}{p_j(x)} = k \,\middle|\, \pi_i\right\} = 0 \quad i \neq j$, $0 \le k < \infty$, entonces cualquier procedimiento admisible es un procedimiento Bayes.

La demostración del anterior teorema demuestra que la clase de los procedimientos Bayes es completa. Para cualquier procedimiento dado R^*, existe un procedimiento Bayes $\bar{R}$ que es al menos tan bueno (lo que supone la completitud sencial). Pero si $\bar{R}$ y R^* son igualmente buenos entonces son el mismo (excepto para conjuntos de probabilidad cero).

TEOREMA.- Si $P\left\{\frac{p_i(x)}{p_j(x)} = k \,\middle|\, \pi_i\right\} = 0$, la clase de los procedimientos Bayes es minimal completa.

Consideremos también ahora la solución minimax. Existe una solución Bayes para la cual $\alpha_1 = \alpha_2 = \alpha_3$. Puesto que el procedimiento es admisible, no existe otro que tenga una menor probabilidad máxima de error (es decir, cada uno de los riesgos es menor). Esto proporciona el procedimiento minimax.

Se pueden consultar los trabajos antes citados para completar esta teoría. Añadamos tan solo que Von Mises (1945) encontró la solución del problema minimax por un camino diferente.

7. CLASIFICACION EN UNA DE VARIAS POBLACIONES NORMALES MULTIVARIANTES

Apliquemos la teoría del anterior párrafo al caso en que cada una de las poblaciones tiene una distribución normal. Supondremos que las medias son diferentes y las matrices de covarianzas coinciden. Sea $N(\mu^{(i)}, \Sigma)$ la distribución de π_i. En un principio supondremos conocidos los parámetros. Para costes generales y probabilidades a priori conocidas podemos formar funciones como las del párrafo anterior y definir las regiones R_j como aquellas constituidas por puntos x tales que la j-ésima función es mínima.

En lo que sigue supondremos que los costes de mala clasificación son iguales. Entonces utilizaremos las funciones

$$u_{jk}(x) = \log \frac{p_j(x)}{p_k(x)} = \left[x - \tfrac{1}{2}\left(\mu^{(j)} + \mu^{(k)}\right)\right]' \Sigma^{-1} \left(\mu^{(j)} - \mu^{(k)}\right).$$

Si las probabilidades a priori son conocidas, las regiones R_j vienen definidas por aquellos x que satisfacen

$$R_j : u_{jk}(x) > \log \frac{q_k}{q_j}, \quad k = 1, \ldots, m; \; k \neq j.$$

---

(1) A lo largo de cada eje $q_3 = (1-k)(1-q_2)$ y $q_2 = k(1-q_3)$, $0 < k < 1$, α_1 decrece continua y monótonamente de 1 a 0. Sea $q_2 = q_2(k)$ el valor de q_2 tal que $\alpha_1 = \alpha_1^*$; entonces $q_2(k)$ es una función continua de k [por la continuidad de $\alpha_1(q_1,q_2,q_3)$ y la monotonía de α_1 como función de q_2 dado k].

TEOREMA – Si q_i es la probabilidad a priori de extraer una observación de $\pi_i = N(\mu^{(i)}, \Sigma)$ $(i=1,\dots,m)$ y si los costes de mala clasificación son iguales, entonces las regiones de clasificación $R_1,\dots,R_m$ que minimizan el coste esperado vienen definidas por la relación anterior, siendo $U_{jk}(x)$ definida como antes.

Hay que señalar que cada $U_{jk}(x)$ es la función de clasificación relacionada con las poblaciones j-ésima y k-ésima, y $U_{jk}(x) = -U_{kj}(x)$. Puesto que se trata de funciones lineales, la región R_i está acotada por hiperplanos. Si las medias definen un hiperplano $(m-1)$-dimensional (por ejemplo, si los vectores media $\mu^{(i)}$ son linealmente independientes y $p \geq m-1$), entonces R_i está acotada por $m-1$ hiperplanos.

En el caso en que no se conozcan probabilidades a priori, la región R_j viene definida mediante las desigualdades

$$U_{jk}(x) \geq c_j - c_k \quad , \quad k=1,\dots,m; \; k \neq j$$

Las constantes c_k pueden tomarse no negativas. Este conjunto de regiones forman la clase de procedimientos admisibles. Para el procedimiento minimax estas constantes se eligen de manera que $P(i/i,R)$ sean iguales.

Veamos ahora como obtener las probabilidades de una clasificación correcta. Si X es una observación aleatoria, consideremos las variables aleatorias

$$U_{ji} = \left[x - \tfrac{1}{2}(\mu^{(i)} + \mu^{(j)})\right]' \Sigma^{-1} (\mu^{(i)} - \mu^{(j)}).$$

Con $U_{ji} = -U_{ij}$. Utilizamos $m(m-1)/2$ funciones de clasificación si las medias definen un hiperplano $(m-1)$-dimensional. Si X procede de π_j, entonces U_{ji} se distribuye como $N(\tfrac{1}{2}\alpha_{jii}, \alpha_{jii})$ donde

$$\alpha_{jii} = (\mu^{(j)} - \mu^{(i)})' \Sigma^{-1} (\mu^{(j)} - \mu^{(i)}).$$

La covariancia entre U_{ji} y U_{jk} es

$$\alpha_{jki} = (\mu^{(j)} - \mu^{(k)})' \Sigma^{-1} (\mu^{(j)} - \mu^{(i)}).$$

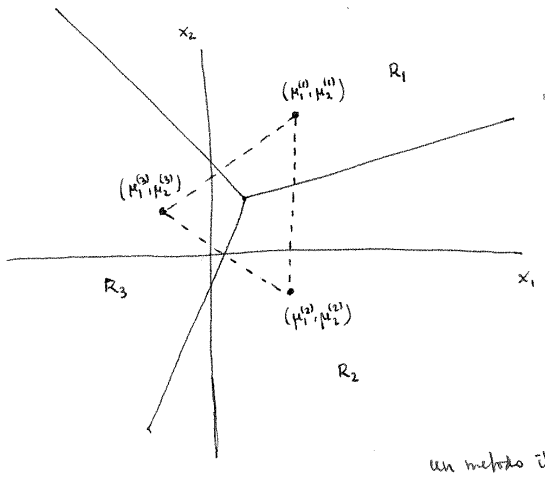
Para determinar las constantes c_j consideramos las integrales

$$P(j/j,R) = \int_{c_j - c_m}^{\infty} \cdots \int_{c_j - c_1}^{\infty} f_j \, dU_{j1} \cdots dU_{j,j+1} \, dU_{j,j-1} \cdots dU_{jm},$$

donde f_j es la densidad de U_{ji} $(i=1,2,\dots,m)$, $(i \neq j)$.

TEOREMA – Si π_i es $N(\mu^{(i)}, \Sigma)$ y los costes de mala clasificación son iguales, las regiones de clasificación $R_1,\dots,R_m$ que minimizan la máxima pérdida esperada condicional, están definidas mediante $U_{jk}(x) \geq c_j - c_k$ $k=1,\dots,m$; $k \neq j$, donde $U_{jk}(x)$ está definido como antes. Las constantes c_j se determinan de manera que las integrales expresadas den los valores de $P(j/j,R)$ sean iguales.

Como un ejemplo consideremos el caso $m=3$. No hay pérdida de generalidad si hacemos $p=2$, para densidades con mayor p podemos proyectar sobre el plano bidimensional determinado por las medias de las tres poblaciones si no son colineales (o dicho, podemos transformar el vector X en U_{12}, U_{13} y $p-2$ otras coordenadas, donde estas últimas $p-2$ componentes se distribuyen independientemente de U_{12} y U_{13} con medias cero). Las regiones R_j están determinadas mediante tres semirrectas, tal como se muestra en la figura. Si el procedimiento es minimax, no podemos desplazar la recta entre R_1 y R_2 más cerca de $(\mu_1^{(1)}, \mu_2^{(1)})$, la recta entre R_2 y R_3 más cerca de $(\mu_1^{(3)}, \mu_2^{(3)})$ y conservar todavía la igualdad $P(1/1,R) = P(2/2,R) = P(3/3,R)$, sin dejar un triángulo que no esté incluido en región alguna. Por tanto, como las regiones deben cubrir totalmente el espacio, las líneas deben encontrarse en un punto, y la igualdad de probabilidades determina $c_i - c_j$ univocamente.



Para realizar esto en un caso específico en el que tenemos valores numéricos para los vectores $\mu^{(1)}, \mu^{(2)}, \mu^{(3)}$ y la matriz Σ, consideraríamos las tres $(=p+1)$ distribuciones conjuntas, para los U_{ij} $(j \neq i)$. Probaríamos inicialmente con los valores $c_i = 0$, y utilizando las tablas del Teorema de la normal bivariante calcularíamos $P(i/i,R)$, mediante un método iterativo estudiaríamos los c_i que más se aproximan a la condición.

La teoría precedente parte del supuesto de conocer los parámetros de las poblaciones. Si no fueran conocidos y si tenemos disponible una muestra de cada población, sustituiríamos en $U_{ij}(x)$ cada parámetro por su correspondiente estimación. Sean $x_1^{(i)}, \dots, x_{N_i}^{(i)}$

las observaciones procedentes de $N(\mu^{(i)}, \Sigma)$, $i=1,\dots,m$. Estimamos $\mu^{(i)}$ mediante

$$\bar{x}^{(i)} = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} x_\alpha^{(i)}$$

y Σ mediante S definida por

$$\left(\sum_{i=1}^{m} N_i - m\right) S = \sum_{i=1}^{m} \sum_{\alpha=1}^{N_i} (x_\alpha^{(i)} - \bar{x}^{(i)})(x_\alpha^{(i)} - \bar{x}^{(i)})'.$$

Entonces, el análogo de $u_{ij}(x)$ es

$$v_{ij}(x) = [x - \tfrac{1}{2}(\bar{x}^{(i)} + \bar{x}^{(j)})]' S^{-1} (\bar{x}^{(i)} - \bar{x}^{(j)}).$$

Si las variables son absolutas, las distribuciones son diferentes de aquellas de las u_{ij}. No obstante, cuando $N_i \to \infty$, la distribución conjunta se aproxima a la de u_{ij}. Por tanto, para muestras enormemente grandes podemos utilizar la teoría anterior.

8. UN EJEMPLO DE CLASIFICACIÓN EN UNA DE VARIAS POBLACIONES NORMALES MULTIVARIANTES

Rao (1948) considera tres poblaciones consistentes en la casta de los Brahmin (π_1), la casta de los Artesanos (π_2), y la casta de los Korwa (π_3) de la India. Las medidas para cada uno de los individuos de una casta son, la estatura (x_1), altura sentado (x_2), anchura nasal (x_3) y altura nasal (x_4). Las medias de estas variables en las tres poblaciones aparecen en la tabla.

	Brahmin π_1	Artesanos π_2	Korwa π_3
Estatura (x_1)	164.51	160.53	158.17
Altura sentado (x_2)	86.43	81.47	81.16
Anchura nasal (x_3)	25.49	23.84	21.44
Altura nasal (x_4)	51.24	48.62	46.72

La matriz de correlaciones para todas las poblaciones es

1.0000	0.5849	0.1774	0.1974
	1.0000	0.2094	0.2120
		1.0000	0.2910
			1.0000

Las desviaciones típicas son $\sigma_1 = 5.74$, $\sigma_2 = 3.20$, $\sigma_3 = 1.75$, $\sigma_4 = 3.50$. Suponemos que cada población es normal. Nuestro problema es dividir el espacio de las cuatro variables x_1, x_2, x_3, x_4 en tres regiones de clasificación. Supondremos que los costes de malaclasificación son iguales. Encontraremos a) Un conjunto de regiones bajo el supuesto de que extraer una nueva observación de cada población es igualmente probable ($q_1 = q_2 = q_3 = 1/3$), y b) un conjunto de regiones tales que la mayor probabilidad de malaclasificación es minimizada (solución minimax).

Calculamos en primer lugar los coeficientes de $\Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$ y $\Sigma^{-1}(\mu^{(1)} - \mu^{(3)})$, ~~después~~ $\Sigma^{-1}(\mu^{(2)} - \mu^{(3)}) = \Sigma^{-1}(\mu^{(1)} - \mu^{(3)}) - \Sigma^{-1}(\mu^{(1)} - \mu^{(2)})$. Luego calculamos $\frac{1}{2}(\mu^{(i)} + \mu^{(j)})' \Sigma^{-1}(\mu^{(i)} - \mu^{(j)})$. Los resultados obtenidos son las siguientes funciones discriminantes

$$u_{12}(x) = -0.0708 x_1 + 0.4990 x_2 + 0.3373 x_3 + 0.0887 x_4 - 43.13$$
$$u_{13}(x) = 0.0003 x_1 + 0.3550 x_2 + 1.1063 x_3 + 0.1375 x_4 - 62.49$$
$$u_{23}(x) = 0.0711 x_1 - 0.1440 x_2 + 0.7690 x_3 + 0.0488 x_4 - 19.36.$$

Las otras tres funciones son $u_{21}(x) = -u_{12}(x)$, $u_{31}(x) = -u_{13}(x)$ y $u_{32}(x) = -u_{23}(x)$. Si existen probabilidades a priori y son iguales el mejor conjunto de ~~regiones~~ regiones de clasificación es

R_1: $u_{12}(x) \geq 0$, $u_{13}(x) \geq 0$

R_2: $u_{21}(x) \geq 0$, $u_{23}(x) \geq 0$

R_3: $u_{31}(x) \geq 0$, $u_{32}(x) \geq 0$.

Por ejemplo si obtenemos una observación individual x, con $u_{12}(x) \geq 0$ y $u_{13}(x) \geq 0$, la clasificaremos como Brahmin.

Para encontrar las probabilidades de malaclasificación cuando un individuo es extraído de la población π_i, necesitamos las

medias, varianzas y covarianzas de los pares correspondientes de las U's. Estos valores aparecen en la tabla siguiente

Población de X	U	Medias	Desviaciones típicas	Correlación
Π_1	U_{12}	1.491	1.727	0.8658
	U_{13}	3.487	2.641	
Π_2	U_{21}	1.491	1.727	-0.3894
	U_{23}	1.031	1.436	
Π_3	U_{31}	3.487	2.641	0.7983
	U_{32}	1.031	1.436	

Las probabilidades de mala clasificación se obtienen mediante el uso de tablas para la normal bivariante. Estas probabilidades son 0.21 para Π_1, 0.42 para Π_2 y 0.25 para Π_3. Por ejemplo, si las medidas son hechas sobre un Brahmin, la probabilidad de que sea clasificado como Anteano o Konva es 0.21.

La solución minimax se obtiene escribiendo las constantes c_1, c_2 y c_3 de manera que las probabilidades de una mala clasificación sean iguales. Las regiones de clasificación son:

$$R_1' : U_{12}(x) \geq 0.54 \quad , \quad U_{13}(x) \geq 0.29 \ ;$$

$$R_2' : U_{21}(x) \geq -0.54 \quad , \quad U_{23}(x) \geq -0.25 \ ;$$

$$R_3' : U_{31}(x) \geq -0.29 \quad , \quad U_{32}(x) \geq 0.25 \ .$$

La probabilidad de mala clasificación común es 0.30. Así, la máxima probabilidad de error ha sido reducida de 0.42 a 0.30.

EJEMPLOS

(1) En un ejemplo del capítulo anterior hablabamos de un grupo de 49 ancianos que provenían de dos poblaciones, la Π_1, factor senil ausente y la Π_2, factor senil presente. Recordemos que los datos obtenidos eran

	Población: Factor senil ausente	Factor senil presente
$\bar{x}_1^{(i)}$	12.57	8.75
$\bar{x}_2^{(i)}$	9.57	5.33
$\bar{x}_3^{(i)}$	11.49	8.50
$\bar{x}_4^{(i)}$	7.97	4.75
	$N_1 = 37$	$N_2 = 12$

$$S = \begin{bmatrix} 11.2553 & 9.4042 & 7.1489 & 3.3830 \\ & 13.3830 & 7.3830 & 2.5532 \\ & & 11.5744 & 2.6170 \\ & & & 5.8085 \end{bmatrix}$$

Iramos a calcular la función discriminante lineal para estos dos grupos poblacionales. El vector diferencia de medias es

$$(\bar{x}_1 - \bar{x}_2)' = [3.82, 4.24, 2.99, 3.22]$$

La función discriminante lineal viene dada por

$$y = 0.030 x_1 + 0.204 x_2 + 0.010 x_3 + 0.443 x_4 \ .$$

Vemos que los tests segundo y cuarto, similaridades y frases completas, respectivamente, dominan la función, mientras que los tests de información y aritmética tienen una ponderación casi despreciable en la función y. En investigaciones posteriores la atención del investigador podría centrarse en aquellos dos como indicadores de la calidad del factor senil.

El término independiente puede obtenerse bien a partir de la expresión

$$\frac{1}{2}(\bar{x}_1 + \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) = 4.76$$

o bien como el punto medio de los valores medios de la función discriminante en un grupo y otro. Así es $\bar{y}_1 = 5.97$, $\bar{y}_2 = 3.54$

$$\frac{\bar{y}_1 + \bar{y}_2}{2} = 4.76.$$

Si suponemos que nada sabemos acerca de las probabilidades a priori q_1 y q_2 de las poblaciones Π_1, Π_2, respectivamente y adoptamos trabajamos con la hipótesis de estas iguales, $c(1/2) = c(2/1)$, la solución minimax conduce a un punto de discriminación $c=0$, por tanto.

R_1: $\quad y - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) \geq 0$

R_2: $\quad$ ——————— < 0

o bien

R_1: $\quad y(x) \geq 4.76$

R_2: $\quad y(x) < 4.76$

Aplicando esta regla a los individuos de las dos muestras utilizadas obtendríamos la siguiente ~~clasificación~~ clasificación de los mismos.

		Diagnóstico Psiquiátrico		
		F.S.A.	F.S.P	Total
Clasificación mediante la función discriminante	F.S.A	29	4	33
	F.S.P	8	8	16
	Total	37	12	49

La tabla da una medida de la capacidad de la función discriminante lineal para reproducir el diagnóstico. Obsérvese que la estimación de las probabilidades teóricas de mala clasificación mediante las correspondientes proporciones muestrales no son muy adecuadas por cuanto en nuestro caso, $P(2/1)$ si concuerdan, pero no así en el otro $P(1/2)$. En efecto ~~los~~ los valores de α, necesario para conocer la distribución de $U = \left[\bar{x} - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})\right]' \Sigma^{-1} \left[\mu^{(1)} - \mu^{(2)}\right]$, puede estimarse mediante

$$\hat{\alpha} = (\bar{x}^{(1)} - \bar{x}^{(2)})' S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) = 2.42$$

U tiene una distribución $N\left(\frac{1}{2}\alpha, \alpha\right)$ si $\bar{x} \in \Pi_1$ y $N\left(-\frac{1}{2}\alpha, \alpha\right)$ si $\bar{x} \in \Pi_2$. Utilizando la estimación encontrada para α, tendremos

$U \simeq N\left(\frac{1}{2}\hat{\alpha}, \hat{\alpha}\right)$ si $\bar{x} \in \Pi_1$

$U \simeq N\left(-\frac{1}{2}\hat{\alpha}, \hat{\alpha}\right)$ si $\bar{x} \in \Pi_2$

entonces

$$\boxed{P(2/1) = P(U \leq 0) = P\left(\frac{U - \frac{1}{2}\hat{\alpha}}{\sqrt{\hat{\alpha}}} \leq \frac{-1.21}{\sqrt{2.42}}\right) = P(z \leq -0.78) = \Phi(-0.78) \simeq 0.2177}$$

análogamente

$$\boxed{P(1/2) \simeq 0.2177}$$

Las estimaciones, mediante la tabla de estas cantidades serían

$$\hat{P}(2/1) = \frac{8}{37} \simeq 0.2162$$

pero

$$\hat{P}(1/2) = \frac{4}{12} \simeq 0.3333 \quad \text{que supone una estimación por exceso.}$$

En el ejemplo que sigue damos una alternativa a este método, para tratar de mejorar estas probabilidades de mala clasificación tan elevadas.

(2) La información acerca de las frecuencias relativas con que las observaciones de las dos poblaciones son encontradas puede ser de utilidad en clasificación de estas observaciones mediante funciones discriminantes. Vamos a reclasificar los individuos del ejemplo anterior haciendo uso de una regla Bayes. Necesitamos para ello probabilidades a priori sobre cada población, por lo que llevaremos a cabo una estimación de las mismas mediante las correspondiente frecuencias relativas cuyos son: $\hat{q}_1 = 37/49$, $\hat{q}_2 = 12/49$. Los costos $C(1/2)$ y $C(2/1)$ se toma clasificación los supondremos iguales.

Así las cosas, la regla de clasificación viene dada por:

$$R_1: \quad y(x) - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2) \geq \log \frac{q_2}{q_1}$$

$$R_2: \quad y(x) - \frac{1}{2}(\bar{x}_1 + \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2) < \log \frac{q_2}{q_1}$$

$$\log \frac{q_2}{q_1} = \log \frac{12}{37} = -1.126$$

o sea

$$R_1: \quad y(x) \geq 3.634$$

$$R_2: \quad y(x) < 3.634$$

La aplicación de esta regla conduce al siguiente cuadro de clasificación:

		Diagnóstico Psiquiatrico		
		F.S.A	F.S.P	
clasificación mediante la función discriminante	F.S.A	37	5	42
	F.S.P.	0	7	7
		37	12	49

La introducción de las probabilidades a priori ha disminuido considerablemente, menos de la mitad, el número de individuos mal clasificados que se obtenía mediante la regla simplista del ejemplo anterior.

Capítulo 6

La distribución de la matriz de covarianzas muestral y de la varianza muestral generalizada

LA DISTRIBUCION DE LA MATRIZ DE COVARIANZAS MUESTRAL Y DE LA VARIANZA MUESTRAL GENERALIZADA

1. INTRODUCCION

La matriz de covarianzas muestral es $S = [1/(N-1)] \sum_{\alpha} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$ y es un estimador de la matriz de covarianzas poblacional Σ. Hemos estudiado anteriormente la distribución de probabilidad de $A = (N-1)S$ para el caso 2×2. Lo que haremos ahora es generalizar este resultado para una matriz A de cualquier orden. Cuando $\Sigma = I$, esta distribución es en un sentido una generalización de la distribución χ^2. La distribución de A (o de S), a menudo llamada distribución de Wishart, es fundamental en estadística multivariante. Nos ocuparemos de ella y estudiaremos algunas de sus propiedades.

La varianza muestral generalizada se define como $|S|$ y es una especie de medida de la dispersión de la muestra. Su distribución también es considerada en su momento.

2. LA DISTRIBUCION DE WISHART

En este párrafo obtendremos la distribución de $A = \sum_{\alpha=1}^{N} (X_\alpha - \bar{X})(X_\alpha - \bar{X})'$, donde los X_α son independientes, cada una con distribución $N(\mu, \Sigma)$. Como vimos en un capítulo anterior, A se distribuye como $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, donde $n = N-1$ y los Z_α son independientes, cada una con distribución $N(0, \Sigma)$. Demostraremos que la distribución (o mejor) la densidad de A, para A definida positiva es

$$\frac{|A|^{\frac{1}{2}(n-p-1)} \cdot \exp\left\{-\frac{1}{2}\operatorname{tr}\Sigma^{-1}A\right\}}{2^{\frac{1}{2}np}\,\pi^{p(p-1)/4} \cdot |\Sigma|^{\frac{1}{2}n} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n+1-i)\right)} \qquad (1)$$

Obtendremos en primer lugar (1) para $\Sigma = I$. Utilizaremos aquí repetidamente el siguiente caso especial de un teorema enunciado cuando estudiábamos la distribución del coeficiente de correlación parcial. Si los escalares U_α son independientes y U_α tiene la distribución $N(\Gamma w_\alpha, \phi)$, entonces $\sum_\alpha U_\alpha^2 - \sum_\alpha U_\alpha w_\alpha' \left(\sum_\alpha w_\alpha w_\alpha'\right)^{-1} \sum_\alpha U_\alpha w_\alpha$ se distribuye como $\sum_{\alpha=1}^{n-q} V_\alpha^2$, donde q es el número de componentes de w_α y los V_α son independientes, cada una con una distribución $N(0, \phi)$, e independientemente de $\sum U_\alpha w_\alpha' \left(\sum w_\alpha w_\alpha'\right)^{-1}$. En particular si $\phi = 1$, entonces $\sum_\alpha U_\alpha^2 - \sum_\alpha U_\alpha w_\alpha \left(\sum_\alpha w_\alpha w_\alpha'\right)^{-1} \sum_\alpha U_\alpha w_\alpha'$ tiene una distribución χ^2 con $n-q$ grados de libertad, cuya densidad es

$$\frac{1}{2^{\frac{1}{2}(n-q)}\,\Gamma\left(\frac{1}{2}(n-q)\right)} \cdot t^{\frac{1}{2}(n-q)-1} e^{-\frac{1}{2}t}.$$

Tenemos posteriormente que $\sum U_\alpha w_\alpha$ es normal; si $\Gamma = 0$, entonces $E\left(\sum_\alpha U_\alpha w_\alpha\right) = 0$ y la matriz de covarianzas es

$$E\left(\sum_\alpha U_\alpha w_\alpha \cdot \sum_\beta U_\beta w_\beta'\right) = \sum_{\alpha\beta} w_\alpha w_\beta' \delta_{\alpha\beta} = \sum_\alpha w_\alpha w_\alpha'.$$

Sea $a_{(i)} = (a_{i,i+1}, a_{i,i+2}, \ldots, a_{ip})$ y $A_{ii} = (a_{jk})$, $j,k = i \cdots p$.

Entonces

$$A_{ii} = \begin{pmatrix} a_{ii} & a_{(i)} \\ a_{(i)} & A_{i+1,i+1} \end{pmatrix}.$$

Sea $a_{ii\cdot i+1,\ldots,p} = a_{ii} - a_{(i)} A_{i+1,i+1}^{-1} a_{(i)}$. El conjunto $(Z_{i1}, \ldots, Z_{in})$ se distribuye independientemente de $(Z_{j1}, \ldots, Z_{jn})$, $j \neq i$ (a cargo de $\Sigma = I$), y por tanto, condicionando a $Z_{j\alpha} = z_{j\alpha}$ $(j \neq i, \alpha = 1, \ldots, n)$, los elementos de $(Z_{i1}, \ldots, Z_{in})$ se distribuyen independientemente, cada uno $N(0,1)$, que es de la forma $N(\Gamma w_\alpha, \phi)$ con $\Gamma = 0$ y $\phi = 1$. Sea $Z_\alpha^{(i)\prime} = (Z_{j\alpha}, Z_{j+1,\alpha}, \ldots, Z_{p\alpha})$. Entonces $a_{(i)} = \sum_\alpha Z_{i\alpha} Z_\alpha^{(i+1)\prime}$. Aplicando el caso especial del teorema citado encontramos que condicionado a $Z_\alpha^{(i+1)} = z_\alpha^{(i+1)}$, $a_{ii\cdot i+1,\ldots,p}$ tiene una distribución χ^2 con $n-(p-i)$ grados de libertad e independiente de a_{ij} que se distribuye condicionalmente como $N(0, A_{i+1,i+1})$.

Se observará que la distribución condicional depende de $z_\alpha^{(i+1)}$ solo a través de $A_{i+1,i+1}$; es decir, la densidad de $a_{11\cdot 2,\ldots,p}$, $a_{(1)}$, $a_{22\cdot 3\ldots p}$, $a_{(2)}$, $\ldots$, $a_{p-1,p-1\cdot p}$, $a_{(p-1)}$, a_{pp} es

$$f_1(a_{11\cdot 2,\ldots,p}\, a_{(1)} / A_{22}) \cdots f_{p-1}(a_{p-1,p-1\cdot p}\, a_{(p-1)} / a_{pp}) \cdot f(a_{pp}) =$$

$$= \frac{a_{pp}^{\frac{1}{2}n-1} \cdot e^{-\frac{1}{2}a_{pp}}}{2^{\frac{1}{2}n}\,\Gamma\left(\frac{1}{2}n\right)} \cdot \prod_{i=1}^{p-1} \left\{ \frac{a_{ii\cdot i+1,\ldots,p}^{\frac{1}{2}[n-(p-i)]-1} \cdot e^{-\frac{1}{2}a_{ii\cdot i+1,\ldots,p}}}{2^{\frac{1}{2}[n-(p-i)]}\,\Gamma\left[\frac{1}{2}[n-(p-i)]\right]} \cdot \frac{e^{-\frac{1}{2}a_{(i)} A_{i+1,i+1}^{-1} a_{(i)}'}}{(2\pi)^{\frac{1}{2}(p-i)} |A_{i+1,i+1}|^{\frac{1}{2}}} \right\}$$

Encontramos la densidad de $a_{ii}, a_{(i)}, \dots, a_{pp}$ sustituyendo en la expresión anterior $a_{ii\cdot i+1,\dots,p} = a_{ii} - a_{(i)} A_{i+1,i+1}^{-1} a'_{(i)}$, y multiplicando por el Jacobiano, que es 1, y simplificando posteriormente la expresión. El exponente de e en la anterior igualdad es

$$-\frac{1}{2}\left[a_{pp} + \sum_{i=1}^{p-1} a_{ii\cdot i+1,\dots,p} + \sum_{i=1}^{p-1} a_{(i)} A_{i+1,i+1}^{-1} a'_{(i)}\right] = -\frac{1}{2}\left[a_{pp} + \sum_{i=1}^{p-1}\left(a_{ii} - a_{(i)} A_{i+1,i+1}^{-1} a'_{(i)}\right) + \sum_{i=1}^{p-1} a_{(i)} A_{i+1,i+1}^{-1} a'_{(i)}\right] =$$

$$= -\frac{1}{2}\sum_{i=1}^{p} a_{ii} = -\frac{1}{2}\operatorname{tr} A .$$

Utilizando una propiedad de los puntuales, que asegura que

$$a_{ii\cdot i+1,\dots,p} = \frac{\begin{vmatrix} a_{ii} & a_{(i)} \\ a'_{(i)} & A_{i+1,i+1} \end{vmatrix}}{|A_{i+1,i+1}|} = \frac{|A_{ii}|}{|A_{i+1,i+1}|}$$

encontramos que $(a_{pp} = A_{pp})$

$$a_{pp}\prod_{i=1}^{p-1} a_{ii\cdot i+1,\dots,p} = a_{pp}\prod_{i=1}^{p-1}\frac{|A_{ii}|}{|A_{i+1,i+1}|} = |A_{11}| = |A| .$$

Entonces

$$a_{pp}^{-\frac{1}{2}n+1}\prod_{i=1}^{p-1}\frac{a_{ii\cdot i+1,\dots,p}^{\frac{1}{2}[n-(p-i)]-1}}{|A_{i+1,i+1}|^{\frac{1}{2}}} = |A|^{\frac{1}{2}(n-p-1)}\cdot a_{pp}^{\frac{1}{2}(p-1)}\cdot\prod_{i=1}^{p-1}\frac{|A_{ii}|^{\frac{1}{2}(i-1)}}{|A_{i+1,i+1}|^{\frac{1}{2}i}} = |A|^{\frac{1}{2}(n-p-1)} .$$

La potencia de π en el denominador es $\frac{1}{2}[(p-1)+(p-2)+\cdots+1] = \frac{1}{2}\left[p(p-1)/2\right]$. Puesto que $\pi^{\left(\frac{1}{2}n\right)\prod_{i=1}^{p-1}\pi\left\{\frac{1}{2}[n-(p-i)]\right\}} = \prod_{i=1}^{p}\pi^{\left[\frac{1}{2}(n-i+1)\right]}$, encontramos que la densidad de $a_{11}, a_{12}, \dots, a_{pp}$ es

$$\frac{|A|^{\frac{1}{2}(n-p-1)}\, e^{-\frac{1}{2}\operatorname{tr} A}}{2^{\frac{1}{2}np}\,\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]} .$$

que es la expresión de (1) cuando $\Sigma = I$.

Obtengamos la distribución para Σ arbitrario. Sea $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, donde los Z_α son independientes, cada una con distribución $N(0,\Sigma)$, y sea $A^* = \sum_{\alpha=1}^{n} Z_\alpha^* Z_\alpha^{*\prime}$, donde los Z_α^* son independientes, cada una con distribución $N(0,I)$. Entonces la densidad de A^* es la obtenida. Sea C una matriz triangular arbitraria $(c_{ij}=0,\ i>j)$ tal que $C\Sigma C' = I$ (siempre existe C por una propiedad de las matrices definidas positivas). La distribución de $CAC' = \sum_{\alpha=1}^{n}(CZ_\alpha)(CZ_\alpha)'$ es la misma que la de A^* puesto que la distribución de CZ_α es la de Z_α^*. Por tanto obtendremos la distribución de A sustituyendo $A^* = CAC'$ en la densidad obtenida y multiplicando por el Jacobiano.

LEMA.— Sea A^* una matriz simétrica que transformamos en otra matriz simétrica A mediante $A^* = CAC'$, donde C es una matriz triangular no singular. El jacobiano de la transformación es $\operatorname{mod}|C|^{p+1}$ (mod. indica módulo).

Demostración.— La transformación es

$$a_{ij}^* = \sum_{k\ell} c_{ik} a_{k\ell} c_{j\ell} .$$

Las derivadas parciales son

$$\frac{\partial a_{ij}^*}{\partial a_{kk}} = c_{ik} c_{jk}$$

$$\frac{\partial a_{ij}^*}{\partial a_{k\ell}} = c_{ik} c_{j\ell} + c_{i\ell} c_{jk} \qquad \ell \neq k$$

Ordenamos la matriz de las derivadas parciales en el orden $a_{11}, a_{12}, \dots, a_{1p}, a_{22}, \dots, a_{pp}$; la posición de la fila correspondiente a a_{ij}^* y la posición de la columna a a_{ij}. La matriz es

$$\begin{bmatrix} c_{11}^2 & 2c_{11}c_{12} & \cdots & 2c_{11}c_{1p} & c_{12}^2 & \cdots & c_{1p}^2 \\ 0 & c_{11}c_{22} & \cdots & c_{11}c_{2p} & c_{12}c_{22} & \cdots & c_{1p}c_{2p} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & c_{11}c_{pp} & c_{1p}c_{2p} & \cdots & c_{1p}c_{pp} \\ 0 & 0 & \cdots & 0 & c_{22}^2 & \cdots & c_{2p}^2 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & c_{pp}^2 \end{bmatrix}$$

que es una matriz triangular. El determinante es el producto de los elementos de la diagonal, que vale $\prod_{j=1}^{p} c_{jj}^{p+1} = |C|^{p+1}$. lo que demuestra el lema.

Si en la densidad de A^* reemplazamos A por CAC', de $C\Sigma C' = I$ tenemos $1 = |C\Sigma C'| = |C|.|\Sigma|.|C'| = |\Sigma|.|CC'|$; así $|CC'| = 1/|\Sigma|$ y $\mod|C| = 1/\sqrt{|\Sigma|}$. Además $\Sigma = C^{-1}(C')^{-1} = (C'C)^{-1}$. Así $\operatorname{tr} CAC' = \operatorname{tr} AC'C = \operatorname{tr} A\Sigma^{-1}$, $|CAC'| = |C|.|A|.|C'| = |A|.|CC'| = |A|/|\Sigma|$. Estas substituciones y el Jacobiano antes obtenido dan por resultado la expresión (1) con que hemos iniciado el parrafo.

TEOREMA.- Supongamos los vectores p-dimensionales $Z_1, \ldots, Z_n$ $(n \geq p)$, que son independientes, cada uno distribuido $N(0,\Sigma)$. Entonces la densidad de $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ es

$$\frac{|A|^{\frac{1}{2}(n-p-1)}\, e^{-\frac{1}{2}\operatorname{tr} A\Sigma^{-1}}}{2^{\frac{1}{2}np}\, \pi^{p(p-1)/4}\, |\Sigma|^{\frac{1}{2}n} \prod_{i=1}^{p} \Gamma\left[\tfrac{1}{2}(n+1-i)\right]}$$

para A definida positiva, y 0 en el resto.

COROLARIO.- Supongamos los vectores p-dimensionales $X_1, \ldots, X_N$ $(N > p)$, independientes, con distribución $N(\mu,\Sigma)$. Entonces la densidad de $A = \sum_{\alpha=1}^{N} (X_\alpha - \bar{X})(X_\alpha - \bar{X})'$ es la anterior con $n = N-1$.

Es costumbre derivar la densidad obtenida mediante $W(A/\Sigma, n)$ y la correspondiente distribución como $W(A/\Sigma, n)$ o bien $W(\Sigma, n)$. La primera obtención de esta distribución (Wishart, 1928) fue mediante un argumento geométrico, que estaba muy relacionado con la demostración que hemos presentado aquí. Sean $v_i' = (z_{i1}, \ldots, z_{in})$ un vector en el espacio R^n. Los elementos diagonales de A son los cuadrados de las longitudes de estos vectores, $a_{ii} = v_i' v_i$, y los elementos de fuera de la diagonal están relacionados con los longitudes y con los ángulos entre los vectores puesto que $r_{ij} = a_{ij}/\sqrt{a_{ii}a_{jj}}$ es el coseno del ángulo entre v_i y v_j. La matriz A describe las longitudes y configuración de estos vectores.

El elemento de probabilidad de $\sqrt{v_i'v_i}, v_i'v_{i+1}, \ldots, v_i'v_p$ dados $v_{i+1}, \ldots, v_p$ es aproximadamente la probabilidad de que v_i caiga en la región para la que $\sqrt{a_{ii}} < v_i'v_p < a_{ip} + da_{ip}$. El primer par de desigualdades define la región entre dos hiperplanos. En esta región la densidad $(2\pi)^{-\frac{1}{2}n} \exp(-\frac{1}{2} v_i'v_i)$ es aproximadamente constante. La intersección de las regiones es un casquete esférico en $n-(p-i)$ dimensiones con una sección transversal de $p-i$ dimensiones. El volumen de esta sección es aproximadamente $d\sqrt{a_{ii}}\, da_{i,i+1} \cdots da_{ip} / |A_{i+1,i+1}|^{1/2}$. El radio al cuadrado del casquete esférico es $a_{ii.i+1,\ldots,p} = a_{ii} - a_{(i)} A_{i+1,i+1}^{-1} a_{(i)}$. El área (o volumen) de esta superficie es la potencia $[n-(p-i)-1]$-ésima del radio por el área de la superficie de la esfera (en $n-(p-i)$ dimensiones) de radio unidad. El área de la superficie de la esfera unidad es $G(n-(p-i)) = 2\pi^{\frac{1}{2}[n-(p-i)]} / \Gamma\{\frac{1}{2}[n-(p-i)]\}$. Así la probabilidad de $\sqrt{v_i'v_i}, v_i'v_{i+1}, \ldots, v_i'v_p$ es

$$\frac{e^{-\frac{1}{2}v_i'v_i}}{(2\pi)^{\frac{1}{2}n}} \cdot \frac{2\pi^{\frac{1}{2}[n-(p-i)]}}{\Gamma\{\frac{1}{2}[n-(p-i)]\}} \cdot a_{ii.i+1,\ldots,p}^{\frac{1}{2}[n-(p-i)-1]} \cdot \frac{d\sqrt{a_{ii}}\, da_{i,i+1} \cdots da_{ip}}{|A_{i+1,i+1}|^{\frac{1}{2}}}$$

El elemento de probabilidad de $v_i'v_i, v_i'v_{i+1}, \ldots, v_i'v_p$ supone la sustitución de $d\sqrt{a_{ii}} = da_{ii}/(2\sqrt{a_{ii}})$. Esto conduce al i-ésimo término del producto antes obtenido y que representaba la densidad de $a_{11.2,\ldots,p}, a_{(1)}, a_{22.3,\ldots,p}, a_{(2)}, \ldots, a_{p-1,p-1.p}, a_{(p-1)}, a_{pp}$.

Existen otros muchos métodos para llegar a este resultado (véase Anderson pag. 158)

Para terminar con este párrafo daremos la distribución conjunta de las varianzas y covarianzas muestrales. Hemos demostrado que $N\hat{\Sigma}$, para muestras de tamaño N que provienen de una $N(\mu,\Sigma)$ se distribuye como $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ donde los Z_α son independientes, cada una con

distribución $N(0,\Sigma)$ y $n=N-1$, $E(N\hat{\Sigma})=n\Sigma$, y $E(S)=\Sigma$, donde $S=(N/n)\hat{\Sigma}$.

TEOREMA.- Supongamos que $\mathbf{X}_1,\dots,\mathbf{X}_N$ $(N>p+1)$ se distribuyen independientemente, cada una como $N(\mu,\Sigma)$. Entonces la distribución de $S=(1/n)\sum_\alpha(\mathbf{X}_\alpha-\bar{\mathbf{X}})(\mathbf{X}_\alpha-\bar{\mathbf{X}})'$ es $W[(1/n)\Sigma,n]$, donde $n=N-1$ y $W[(1/n)\Sigma,n]$ es la distribución de Wishart con matriz de covarianzas $(1/n)\Sigma$ y n grados de libertad.

Demostración.- Claramente, $S=(1/n)A=\sum_{\alpha=1}^{n}[(1/\sqrt{n})\mathbf{Z}_\alpha][(1/\sqrt{n})\mathbf{Z}_\alpha]'$, donde las $(1/\sqrt{n})\mathbf{Z}_\alpha$ son independientes, con distribución $N(0,(1/n)\Sigma)$. Entonces basta aplicar el teorema anterior a estos nuevos $\mathbf{Z}_\alpha^*=(1/\sqrt{n})\mathbf{Z}_\alpha$.

Si $n<p$, la matriz $A=\sum_{\alpha=1}^{n}\mathbf{Z}_\alpha\mathbf{Z}_\alpha'$ no tiene una densidad de probabilidad. No obstante, nos referiremos a la distribución correspondiente, como a la distribución de Wishart.

3. ALGUNAS PROPIEDADES DE LA DISTRIBUCIÓN DE WISHART

La Función Característica

La función característica de la distribución de Wishart puede obtenerse fácilmente a partir de la distribución de las observaciones. Supongamos que $\mathbf{Z}_1,\dots,\mathbf{Z}_n$ se distribuyen independientemente, cada una con densidad

$$\frac{1}{(2\pi)^{\frac{1}{2}p}|\Sigma|^{\frac{1}{2}}}\exp\left(-\tfrac{1}{2}\mathbf{z}'\Sigma^{-1}\mathbf{z}\right).$$

Sea

$$A=\sum_{\alpha=1}^{n}\mathbf{Z}_\alpha\mathbf{Z}_\alpha'.$$

Introducimos la matriz $\Theta=(\theta_{ij})$ con $\theta_{ij}=\theta_{ji}$. La función característica de $a_{11},a_{22},\dots,a_{pp},2a_{12},2a_{13},\dots,2a_{p-1,p}$ es

$$E\left[\exp\left(i\,\mathrm{tr}(A\Theta)\right)\right]=E\left[\exp\left(i\,\mathrm{tr}\sum_{\alpha=1}^{n}\mathbf{Z}_\alpha\mathbf{Z}_\alpha'\Theta\right)\right]=E\left[\exp\left(i\,\mathrm{tr}\sum_{\alpha=1}^{n}\mathbf{Z}_\alpha'\Theta\mathbf{Z}_\alpha\right)\right]=E\left[\exp\left(i\sum_{\alpha=1}^{n}\mathbf{Z}_\alpha'\Theta\mathbf{Z}_\alpha\right)\right]$$

en virtud del hecho de que $\mathrm{tr}\,EFG=\sum c_{ij}f_{jn}g_{ki}=\mathrm{tr}\,FGE$. Como las $\mathbf{Z}_\alpha$ son independientes, tendremos

$$E\left[\exp\left(i\sum_{\alpha=1}^{n}\mathbf{Z}_\alpha'\Theta\mathbf{Z}_\alpha\right)\right]=\prod_{\alpha=1}^{n}E\left[\exp\left(i\mathbf{Z}_\alpha'\Theta\mathbf{Z}_\alpha\right)\right]=\left\{E\left[\exp\left(i\mathbf{Z}_\alpha'\Theta\mathbf{Z}_\alpha\right)\right]\right\}^n.$$

Para Θ real existe una matriz, B, real no singular tal que

$$B'\Sigma^{-1}B=\mathbf{I}$$

$$B'\Theta B=D$$

donde D es una matriz diagonal. Si hacemos

$$\mathbf{z}=B\mathbf{y}$$

entonces

$$E\left[\exp\left(i\mathbf{z}'\Theta\mathbf{z}\right)\right]=E\left[\exp\left(i\mathbf{y}'D\mathbf{y}\right)\right]=E\left[\prod_{j=1}^{p}\exp\left(i d_{jj}y_j^2\right)\right]=\prod_{j=1}^{p}E\left[\exp\left(i d_{jj}y_j^2\right)\right].$$

El j-ésimo término en el producto es la $E[\exp(id_{jj}y_j^2)]$ donde y_j tiene una distribución $N(0,1)$; es pues la función característica de la distribución χ^2 con un grado de libertad, es decir $(1-2id_{jj})^{-1/2}$. Así

$$E\left[\exp\left(i\mathbf{z}'\Theta\mathbf{z}\right)\right]=\prod_{j=1}^{p}(1-2id_{jj})^{-\frac{1}{2}}=\left|I-2iD\right|^{-\frac{1}{2}}$$

puesto que $I-2iD$ es una matriz diagonal. Teniendo en cuenta lo que vale D y lo que vale I, tendremos

$$|I-2iD|=|B'\Sigma^{-1}B-2iB'\Theta B|=|B'(\Sigma^{-1}-2i\Theta)B|=|B'|\cdot|\Sigma^{-1}-2i\Theta|\cdot|B|=|B|^2\cdot|\Sigma^{-1}-2i\Theta|$$

$|B'|\cdot|\Sigma^{-1}|\cdot|B|=|I|=1$, y $|B|^2=1/|\Sigma^{-1}|$. Combinando estos resultados, obtenemos

$$E\left[\exp\left(i\,\mathrm{tr}(A\Theta)\right)\right]=\frac{|\Sigma^{-1}|^{\frac{1}{2}n}}{|\Sigma^{-1}-2i\Theta|^{\frac{1}{2}n}}.$$

Puede demostrarse que este resultado es válido con tal que $\mathrm{re}(\sigma^{jk} - 2i\theta_{jk})$ sea definida positiva. En particular es cierto para toda θ real.

TEOREMA.- Si $Z_1, \dots, Z_n$ son independientes, cada una con distribución $N(0,\Sigma)$, la función característica de $A_{11}, \dots, A_{pp}, 2A_{12}, \dots, 2A_{p-1,p}$, donde $(A_{ij}) = A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, es

$$f(\Theta) = \frac{|\Sigma^{-1}|^{\frac{1}{2}n}}{|\Sigma^{-1} - 2i\Theta|^{\frac{1}{2}n}}$$

Podemos obtener los momentos de A, bien a partir de la función característica, bien a partir de la distribución original. El valor esperado de A_{ij} es

$$E(A_{ij}) = E\left[\sum_{\alpha=1}^{n} Z_{i\alpha} Z_{j\alpha}\right] = \sum_{\alpha=1}^{n} \sigma_{ij} = n\sigma_{ij}.$$

Para obtener las covarianzas necesitamos

$$E(A_{ij}A_{k\ell}) = E\left[\sum_{\alpha,\beta=1}^{n} Z_{i\alpha} Z_{j\alpha} Z_{k\beta} Z_{\ell\beta}\right] = E\left[\sum_{\alpha=1}^{n} Z_{i\alpha} Z_{j\alpha} Z_{k\alpha} Z_{\ell\alpha}\right] + E\left[\sum_{\substack{\alpha,\beta=1 \\ \alpha\neq\beta}}^{n} Z_{i\alpha} Z_{j\alpha} Z_{k\beta} Z_{\ell\beta}\right] =$$

$$= n\left(\sigma_{ij}\sigma_{k\ell} + \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk}\right) + n(n-1)\sigma_{ij}\sigma_{k\ell} = n^2\sigma_{ij}\sigma_{k\ell} + n\sigma_{ik}\sigma_{j\ell} + n\sigma_{i\ell}\sigma_{jk},$$

siendo los momentos de cuarto orden los definidos en un momento cuando estudiamos la normal multivariante. Así, la covarianza entre A_{ij} y $A_{k\ell}$ es

$$E\left[(A_{ij} - n\sigma_{ij})(A_{k\ell} - n\sigma_{k\ell})\right] = n\sigma_{ik}\sigma_{j\ell} + n\sigma_{i\ell}\sigma_{jk}.$$

Para $i=k$ y $j=\ell$ obtenemos la varianza de A_{ij}

$$E\left[(A_{ij} - n\sigma_{ij})^2\right] = n\left(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}\right).$$

La suma de matrices de Wishart

Supongamos que A_i $(i=1,2)$ se distribuyen independientemente como $W(\Sigma, n_i)$ respectivamente. Entonces A_1 se distribuye como $\sum_{\alpha=1}^{n_1} Z_\alpha Z_\alpha'$ y A_2 como $\sum_{\alpha=n_1+1}^{n_1+n_2} Z_\alpha Z_\alpha'$, donde las Z_α son independientes, cada una $N(0,\Sigma)$. Entonces

$$A = A_1 + A_2$$

se distribuye como $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ con $n = n_1 + n_2$. Es decir, A se distribuye $W(\Sigma, n)$. Obviamente la suma de q matrices, distribuidas independientemente, cada una con distribución Wishart con covarianza Σ, tiene una distribución Wishart con matriz Σ y número de grados de libertad igual a la suma de los grados de libertad de las matrices componentes.

Una cierta transformación lineal

Haremos frecuentemente la transformación

$$A = CBC'$$

donde C es una matriz no singular $p \times p$. Si A se distribuye como $W(\Sigma, n)$, entonces B se distribuye como $W(\Phi, n)$ donde

$$\Phi = C^{-1}\Sigma C'^{-1}.$$

Esto se prueba mediante el siguiente argumento:

$$A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha',$$

donde Z_α se distribuyen independientemente, cada una como $N(0,\Sigma)$. Entonces

$$Y_\alpha = C^{-1} Z_\alpha$$

se distribuye como $N(0,\Phi)$. Por tanto,

$$B = \sum_{\alpha=1}^{n} Y_\alpha Y_\alpha' = C^{-1} \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha' C'^{-1} = C^{-1} A C'^{-1}$$

se distribuye como $W(\Phi, n)$. El Jacobiano de la transformación, $\left|\frac{\partial(A)}{\partial(B)}\right|$, viene dado por

$$\left|\frac{\partial(A)}{\partial(B)}\right| = \frac{W(B, \Phi, n)}{W(A, \Sigma, n)} = \frac{|B|^{\frac{1}{2}(n-p-1)} |\Sigma|^{\frac{1}{2}n}}{|A|^{\frac{1}{2}(n-p-1)} |\Phi|^{\frac{1}{2}n}} = \text{mod } |C|^{p+1}.$$

Distribuciones marginales

Si A se distribuye $W(\Sigma, n)$, la distribución marginal de cualquier subconjunto de elementos de A puede ser difícil de obtener. No obstante, la distribución marginal de algunos conjuntos de elementos puede obtenerse con facilidad. Tratamos algo de esto con los siguientes teoremas

TEOREMA.- Sean A y Σ fraccionadas en q y $p-q$ filas y columnas

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Si A se distribuye $W(\Sigma, n)$, entonces A_{11} lo hace $W(\Sigma_{11}, n)$.

Demostración.- A se distribuye como $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, donde los Z_α son independientes, cada uno $N(0, \Sigma)$. Fraccionamos Z_α en dos subvectores de q y $p-q$ componentes respectivamente,

$$Z_\alpha = \begin{pmatrix} Z_\alpha^{(1)} \\ Z_\alpha^{(2)} \end{pmatrix}.$$

Entonces $Z_\alpha^{(1)}$ son independientes, cada uno con distribución $N(0, \Sigma_{11})$ y A_{11} se distribuye como $\sum_{\alpha=1}^{n} Z_\alpha^{(1)} Z_\alpha^{(1)'}$, que tiene una distribución $W(\Sigma_{11}, n)$.

TEOREMA.- Sean A y Σ fraccionadas en $p_1, \dots, p_q$ filas y columnas ($p_1 + p_2 + \dots + p_q = p$)

$$A = \begin{pmatrix} A_{11} & \cdots & A_{1q} \\ \vdots & & \vdots \\ A_{q1} & \cdots & A_{qq} \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \cdots & \Sigma_{1q} \\ \vdots & & \vdots \\ \Sigma_{q1} & \cdots & \Sigma_{qq} \end{pmatrix}$$

Si $\Sigma_{ij} = 0$ para $i \neq j$ y si A se distribuye como $W(\Sigma, n)$, entonces $A_{11}, A_{22}, \dots, A_{qq}$ son independientes y se distribuyen $W(\Sigma_{jj}, n)$.

Demostración.- Fraccionamos Z_α de la forma

$$Z_\alpha = \begin{pmatrix} Z_\alpha^{(1)} \\ \vdots \\ Z_\alpha^{(q)} \end{pmatrix}$$

Puesto que $\Sigma_{ij} = 0$, $Z_\alpha^{(i)}$ y $Z_\beta^{(j)}$ son independientes, $\forall \alpha, \beta$. Entonces $A_{ii} = \sum_{\alpha=1}^{n} Z_\alpha^{(i)} Z_\alpha^{(i)'}$ es independiente de $A_{jj} = \sum_{\alpha=1}^{n} Z_\alpha^{(j)} Z_\alpha^{(j)'}$. El resto del teorema se deriva del anterior.

4. EL TEOREMA DE COCHRAN

El teorema de Cochran es útil para probar que ciertas "formas cuadráticas vectoriales" se distribuyen como sumas de "cuadrados vectoriales". Es una afirmación estadística de un teorema algebraico. Damos en primer lugar la proposición algebraica concerniente a variables escalares.

LEMA.- Si

$$q_i = \sum_{\alpha,\beta=1}^{N} a_{\alpha\beta}^{i} Y_\alpha Y_\beta, \qquad i = 1, \dots, m$$

es de rango r_i y

$$\sum_{i=1}^{m} q_i = \sum_{\alpha=1}^{N} Y_\alpha^2$$

Entonces una condición necesaria y suficiente para que exista una transformación ortogonal de $\{x_\alpha\}$ a $\{z_\alpha\}$ tal que

$$q_i = \sum_{\alpha = r_1 + \cdots + r_{i-1}+1}^{r_1 + \cdots + r_i} z_\alpha^2$$

es que

$$r_1 + \cdots + r_m = N$$

Demostración.- La necesidad de la condición es obvia porque la suma de los rangos no puede ser menor que N si se verifica que

$$\sum_{i=1}^{m} q_i = \sum_{\alpha=1}^{N} y_\alpha^2$$

y no puede ser mayor que N ni la transformación que pasa a $z_1, \ldots, z_{r_1+r_2+\cdots+r_m}$ ha de ser no singular. Probemos pues la suficiencia. De una propiedad matricial sabemos que existe una matriz no singular D, tal que

$$D' \begin{pmatrix} I & 0 & 0 \\ 0 & -I & 0 \\ 0 & 0 & 0 \end{pmatrix} D = A_1$$

donde $A_1 = (a^1_{\alpha\beta})$ y la suma de los órdenes de I y $-I$ es r_1. Sea $d_{\alpha\beta} = b^{(1)}_{\alpha\beta}$, $\alpha = 1, \ldots, r_1$, $\beta = 1, \ldots, N$. Entonces $z_\alpha = \sum_{\beta=1}^{N} b^{(1)}_{\alpha\beta} y_\beta$ $(\alpha = 1, \ldots, r_1)$ forman un conjunto de r_1 funciones lineales de $\{y_\beta\}$ tal que

$$q_1 = \sum_{\alpha=1}^{r_1} c_\alpha z_\alpha^2$$

donde $c_\alpha = 1$ o -1. En general existe un conjunto de r_i funciones lineales

$$z_\alpha = \sum_{\beta=1}^{N} b^{(i)}_{\alpha\beta} y_\beta \qquad \alpha = r_1 + \cdots + r_{i-1} + 1, \ldots, r_1 + \cdots + r_i.$$

tal que

$$q_i = \sum_{\alpha = r_1 + \cdots + r_{i-1}+1}^{r_1 + \cdots + r_i} c_\alpha z_\alpha^2 .$$

Así

$$\sum_{i=1}^{m} q_i = \sum_{\alpha=1}^{N} c_\alpha z_\alpha^2 ,$$

pero como la $\sum_{i=1}^{m} q_i$ es definida positiva los c_α han de ser todos positivos, $c_\alpha = 1$, $\forall \alpha$. Por tanto se verifica la igualdad

$$q_i = \sum_{\alpha = r_1 + \cdots + r_{i-1}+1}^{r_1 + \cdots + r_i} z_\alpha^2 .$$

Asimismo sustituyendo los z_α para todos los i, tendremos

$$z_\alpha = \sum_{\beta=1}^{N} b_{\alpha\beta} y_\beta \qquad , \qquad \alpha = 1, \ldots, N$$

pero

$$\sum_{i=1}^{m} q_i = \sum_{\alpha=1}^{N} z_\alpha^2 = \sum_{\alpha=1}^{N} y_\alpha^2 ,$$

lo que supone que la transformación que pasa de los $\{y_\alpha\}$ a los $\{z_\alpha\}$ es ortogonal $\left(\text{a saber, } I = B'IB = B'B\right)$.

Una expresión alternativa del lema es la siguiente: Sea r_i el rango de la matriz de orden N, A_i, matriz que es además simétrica $(i = 1, \ldots, m)$, y supongamos $\sum_{i=1}^{m} A_i = I$. Una condición necesaria y suficiente para que exista una matriz ortogonal

$$B = \begin{pmatrix} B_1 \\ \vdots \\ B_m \end{pmatrix}$$

tal que $A_i = B_i' B_i$ y que $\sum_{i=1}^{m} r_i = N$.

Estableceremos a continuación el teorema de Cochran.

TEOREMA DE COCHRAN.- Supongamos que Y_α se distribuye $N(\mathbf{0}, \Sigma)$ independientemente de Y_β $(\alpha \neq \beta)$. Supongamos que la matriz $A_i = (a^i_{\alpha\beta})$ utilizada para formar

$$Q_i = \sum_{\alpha,\beta=1}^{N} a^i_{\alpha\beta} Y_\alpha Y_\beta' \qquad i = 1, \dots, m$$

y de rango r_i y supongamos

$$\sum_{i=1}^{m} Q_i = \sum_{\alpha=1}^{N} Y_\alpha Y_\alpha'.$$

Entonces una condición necesaria y suficiente para que Q_i $(i = 1, \dots, m)$ se distribuya como

$$\sum_{r_1 + \cdots + r_{i-1} + 1}^{r_1 + \cdots + r_i} Z_\alpha Z_\alpha'$$

donde las Z_α son $N(0, \Sigma)$ e independientes y Q_i se distribuya independientemente de Q_j $(i \neq j)$ es que

$$r_1 + \cdots + r_m = N.$$

COROLARIO.- Si $r_i \geq p$, entonces Q_i se distribuye como $W(\Sigma, r_i)$

Demostración.- Si $r_1 + r_2 + \cdots + r_m = N$ se verifica, existe una matriz ortogonal B tal que $A_i = B_i' B_i$. Puesto que B es ortogonal

$$Z_\alpha = \sum_{\beta=1}^{N} b_{\alpha\beta} Y_\beta$$

son independientes, cada uno con distribución $N(0, \Sigma)$, por una propiedad ya estudiada de este tipo de transformaciones. Vemos que

$$Q_i = \sum_{\alpha,\beta=1}^{N} a^i_{\alpha\beta} Y_\alpha Y_\beta' = \sum_{\alpha\beta} \sum_{\gamma} b_{\gamma\alpha} b_{\gamma\beta} Y_\alpha Y_\beta' = \sum_{\gamma = r_1 + \cdots + r_{i-1} + 1}^{r_1 + \cdots + r_i} Z_\gamma Z_\gamma'.$$

La necesidad se obtiene mediante una argumentación análoga a la del lema anterior.

Este teorema es útil para generalizar resultados de análisis de la varianza unidimensional, de ello nos ocuparemos más tarde. Como un ejemplo de aplicación del teorema, probaremos que en una muestra de tamaño N la media de las matrices resultante al multiplicar una observación por su traspuesta, que es exactamente N veces el vector media muestral por su traspuesta y un múltiplo de la matriz de covarianzas muestral se distribuyen independientemente con distribución Wishart singular y no singular respectivamente. Sea $Y_1, \dots, Y_N$ vectores aleatorios independientes $N(0, \Sigma)$. Utilizaremos las matrices $(a^{(1)}_{\alpha\beta}) = (1/N)$ y $(a^{(2)}_{\alpha\beta}) = [\delta_{\alpha\beta} - (1/N)]$

Entonces

$$Q_1 = \sum_{\alpha,\beta=1}^{N} \frac{1}{N} Y_\alpha Y_\beta' = N \bar{Y} \bar{Y}'$$

$$Q_2 = \sum_{\alpha,\beta=1}^{N} \left(\delta_{\alpha\beta} - 1/N\right) Y_\alpha Y_\beta' = \sum_{\alpha=1}^{N} Y_\alpha Y_\alpha' - N(\bar{Y}\bar{Y}') = \sum_{\alpha=1}^{N} (Y_\alpha - \bar{Y})(Y_\alpha - \bar{Y})'.$$

Obviamente

$$Q_1 + Q_2 = \sum_{\alpha=1}^{N} Y_\alpha Y_\alpha'.$$

La primera matriz es de rango 1, la segunda matriz es de rango $(N-1)$ (puesto que el rango de la suma de dos matrices es menor o igual que la suma de los rangos y el rango de la segunda matriz es menor que N). Las condiciones del teorema se satisfacen; por tanto Q_1

se distribuye como ZZ', donde Z es $N(0,\Sigma)$ y Q_2 se distribuye independientemente como $W(\Sigma, N-1)$.

5. LA VARIANZA GENERALIZADA

El concepto multivariante equivalente a la varianza, σ^2, es una distribución multivariante y la matriz de covarianzas Σ. Otro concepto multivariante análogo es el escalar $|\Sigma|$, que se le llama varianza generalizada de la distribución multivariante. Análogamente, la varianza generalizada de la muestra de vectores $x_1, \dots, x_N$ es

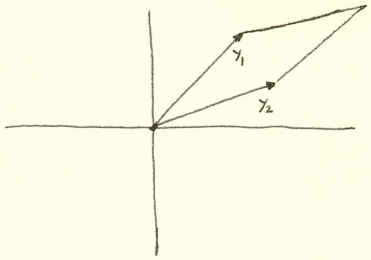
$$|S| = \left| \frac{1}{N-1} \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' \right|.$$

En algún sentido es una medida de extensión. Consideramos aquí este concepto porque recurriremos a él en muchos contrastes de hipótesis basados en la razón de verosimilitud.

Una interpretación geométrica de la varianza muestral generalizada puede darse en términos de p puntos en el espacio N-dimensional. Sea $z_\alpha = x_\alpha - \bar{x}$, y hagamos

$$\begin{pmatrix} y_1' \\ \vdots \\ y_p' \end{pmatrix} = (z_1 \cdots z_N).$$

Los N componentes del vector y_i son las i-ésimas componentes de $z_1, \dots, z_N$. Entonces $a_{ii} = y_i' y_i$ es el cuadrado de la longitud del i-ésimo vector y $a_{ij} = y_i' y_j$ es el producto del coseno del ángulo entre y_i e y_j por las longitudes de y_i e y_j. Consideremos ahora la figura basada en dos de estos vectores. Si $p=2$, tenemos un paralelogramo con y_1 e y_2 como ejes principales; si $p=3$ tenemos un paralelotopo con y_1, y_2 e y_3 como ejes principales. Para p cualquiera, la figura es un paralelotopo en el hiperplano p-dimensional definido mediante $y_1, \dots, y_p$. Está delimitado por pares de hiperplanos $(p-1)$-dimensionales, un hiperplano del par estando definido mediante $p-1$ vectores y el otro pasando por el extremo del vector que falta. El determinante $|A| = |(N-1)S| = (N-1)^p |S|$ es el cuadrado del volumen de este paralelotopo.



TEOREMA. Sea $Y = (y_1 \cdots y_p)$, donde los y_i son vectores N-dimensionales. Entonces el cuadrado del volumen p-dimensional del paralelotopo con $y_1, \dots, y_p$ como ejes principales es $|Y'Y| = |A| = (N-1)^p |S|$.

Demostración. El teorema es cierto si $p=1$, puesto que $|Y'Y| = y_1' y_1$, cuadrado de la longitud de y_1. Supongamos cierto para $p=k-1$ y probemos para $p=k$. Señalemos primero que si dos paralelotopos k-dimensionales tienen bases consistentes en paralelotopos $(k-1)$-dimensionales de igual volumen e igual altitud, sus volúmenes (k-dimensionales) serán iguales (pues cuanto dicho volumen es la integral del volumen $(k-1)$-dimensional). Puesto que es cierto para un paralelotopo rectangular, el volumen k-dimensional es el producto de la altura por el volumen $(k-1)$-dimensional de la base. Una combinación lineal de $y_1, \dots, y_{k-1}$, por ejemplo, $c_1 y_1 + \cdots + c_{k-1} y_{k-1}$ es un vector en la base del paralelotopo y la mínima longitud de $v = y_k - (c_1 y_1 + \cdots + c_{k-1} y_{k-1})$ es la altura del paralelotopo con $y_1, \dots, y_k$ como ejes principales. La longitud de v se minimiza eligiendo $c_1, \dots, c_{k-1}$ de manera que

$$0 = y_j' v = y_j' y_k - (c_1 y_j' y_1 + \cdots + c_{k-1} y_j' y_{k-1}), \quad j = 1, \dots, k-1.$$

Sea $Y_k = (y_1 \cdots y_k)$ e $Y_{k-1} = (y_1 \cdots y_{k-1})$,

$$C = \begin{pmatrix} 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{k-1} \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Entonces $|C| = 1$ y $(Y_{k-1} v) = Y_k C$. Tenemos

$$|Y_k' Y_k| = |C'| \cdot |Y_k' Y_k| \cdot |C| = |C' Y_k' Y_k C| = \left| \begin{pmatrix} Y_{k-1}' \\ v' \end{pmatrix} (Y_{k-1} v) \right| = \begin{vmatrix} Y_{k-1}' Y_{k-1} & 0 \\ 0 & v'v \end{vmatrix} = |Y_{k-1}' Y_{k-1}| \, v'v.$$

Puesto que hemos supuesto que $|Y_{k-1}' Y_{k-1}|$ era el cuadrado del volumen $(k-1)$-dimensional de la base del paralelotopo y $v'v$ es el cuadrado de la altura, el producto es el cuadrado del volumen del paralelotopo k-dimensional. Hemos probado el teorema por inducción.

Veremos más tarde que muchos conceptos de estadística multivariante tienen una interpretación en términos de estos volúmenes. Estos volúmenes son análogos a las distancias que surgen en casos especiales cuando $p=1$.

Consideremos ahora una interpretación geométrica de $|A|$ en términos de N puntos en un espacio p-dimensional. Sean $z_1 - z_N$ definidos como antes, representan N puntos en el espacio p-dimensional. Cuando $p=1$, $|A| = \sum_\alpha z_{1\alpha}^2$, que es la suma de los cuadrados de las distancias de cada punto al origen. En general $|A|$ es la suma de los volumenes de todos los paralelotopos formados tomando como ejes principales p vectores del conjunto $z_1 - z_N$.

Vemos que

$$|A| = \begin{vmatrix} \sum_\alpha z_{1\alpha}^2 & \cdots & \sum_\alpha z_{1\alpha} z_{p-1,\alpha} & \sum_\beta z_{1\beta} z_{p\beta} \\ \vdots & & \vdots & \vdots \\ \sum_\alpha z_{p-1,\alpha} z_{1\alpha} & \cdots & \sum_\alpha z_{p-1,\alpha}^2 & \sum_\beta z_{p-1,\beta} z_{p\beta} \\ \sum_\alpha z_{p\alpha} z_{1\alpha} & \cdots & \sum_\alpha z_{p\alpha} z_{p-1,\alpha} & \sum_\beta z_{p\beta}^2 \end{vmatrix} = \sum_\beta \begin{vmatrix} \sum_\alpha z_{1\alpha}^2 & \cdots & \sum_\alpha z_{1\alpha} z_{p-1,\alpha} & z_{1\beta} z_{p\beta} \\ & & & \\ \sum_\alpha z_{p-1,\alpha} z_{1\alpha} & \cdots & \sum_\alpha z_{p-1,\alpha}^2 & z_{p-1,\beta} z_{p\beta} \\ \sum_\alpha z_{p\alpha} z_{1\alpha} & \cdots & \sum_\alpha z_{p\alpha} z_{p-1,\alpha} & z_{p\beta}^2 \end{vmatrix}$$

por la regla del desarrollo de un determinante. Aplicando este procedimiento repetidas veces a las columnas, tendremos

$$|A| = \sum_{\alpha_1, \ldots \alpha_p = 1}^{N} |z_{i\alpha_j} \cdot z_{j\alpha_j}|.$$

Por el teorema anterior el cuadrado del volumen del paralelotopo con ejes principales $z_{\alpha_1} - z_{\alpha_p}$ es

$$V^2_{\alpha_1, \ldots \alpha_p} = \left| \sum_\beta z_{i\beta} z_{j\beta} \right|$$

donde la suma sobre β se extiende a todos los $(\alpha_1, \ldots \alpha_p)$. Si desarrollamos este determinante de la misma forma que lo hicimos con $|A|$ obtenemos

$$V^2_{\alpha_1, \ldots \alpha_p} = \sum |z_{i\beta_j} z_{j\beta_j}|$$

donde la suma es para cada β_j sobre el rango $(\alpha_1, \ldots \alpha_p)$. Sumando ahora sobre todos los posibles conjuntos $(\alpha_1 - \alpha_p)$ obtendremos precisamente $|A|$. Así pues $|A|$ es la suma de los cuadrados de los volumenes de todos los paralelotopos formados por conjuntos de p vectores de los z_α como ejes principales. Si reemplazamos z_α mediante $x_\alpha - \bar{x}$, podemos enunciar el siguiente teorema:

TEOREMA.- Sean $x_1, \ldots x_N$ una muestra de tamaño N y sea $|S|$ el determinante de la matriz de covarianzas muestral. Entonces $|S|$ es proporcional a la suma de los cuadrados de los volumenes de todos los paralelotopos formados utilizando como ejes principales p vectores de la forma $(x_i - \bar{x})$, elegidos los x_i entre los $x_1, \ldots, x_N$, siendo el factor de proporcionalidad $1/(N-1)^p$.

El equivalente poblacional de $|S|$ es $|\Sigma|$, del que también puede darse una interpretación. Sabemos que

$$P\{ X'\Sigma^{-1}X \leq \chi^2_p(\alpha) \} = 1 - \alpha$$

si X se distribuye $N(0, \Sigma)$, es decir, la probabilidad de que X caiga en el interior del elipsoide $x'\Sigma^{-1}x = \chi^2_p(\alpha)$, es $1-\alpha$. El volumen de este elipsoide viene dado por $C(p)\,|\Sigma|^{\frac{1}{2}}\,[\chi^2_p(\alpha)]^{\frac{1}{2}p}/p$, con $C(p) = \dfrac{2\pi^{\frac{1}{2}p}}{\Gamma(\frac{1}{2}p)}$

Distribución de la varianza muestral generalizada

La distribución de $|S|$ es la misma que la de $|A|/(N-1)^p$, donde

$$A = \sum_{\alpha=1}^{n} z_\alpha z_\alpha'$$

y z_α se distribuye, independientemente de $z_\beta\ (\alpha \neq \beta)$, como $N(0, \Sigma)$, y $n = N-1$. Hagamos

$$W_\alpha = C z_\alpha$$

donde C ha sido elegida de manera que $C\Sigma C' = I$. Entonces W_α se distribuye, independientemente de $W_\beta\ (\alpha \neq \beta)$, como $N(0, I)$ y

$$|B| = |A| / |\Sigma|$$

donde

$$B = \sum_{\alpha=1}^{n} W_\alpha W_\alpha' = \sum_{\alpha=1}^{n} C z_\alpha z_\alpha' C' = CAC'.$$

y $|C| \cdot |\Sigma| \cdot |C'| = 1$.

Como hicimos en un apartado anterior, sea $B_{ii} = (b_{jk})$, $j,k = i, \ldots, p$, $b_{(i)} = (b_{i,i+1}, b_{i,i+2}, \ldots b_{ip})$, y $b_{ii.i+1,\ldots,p} = b_{ii} - b_{(i)} B^{-1}_{i+1,i+1} b'_{(i)}$. Entonces

$$|B| = b_{11.2,\ldots,p}\, b_{22.3,\ldots,p} \cdots b_{pp}.$$

Como vimos en el apartado 2 de este capítulo, los $b_{11.2,\ldots,p}, b_{22.3,\ldots,p}, \ldots, b_{pp}$ son independientes, y $b_{ii.i+1,\ldots,p}$ tiene una distribución χ^2 con $n-(p-i)$ grados de libertad. Así $|B|$ se distribuye como $\chi^2_n \cdot \chi^2_{n-1} \cdots \chi^2_{n-p+1}$.

TEOREMA.- La distribución de la varianza generalizada $|S|$ de una muestra $X_1, \ldots, X_N$ de una $N(\mu, \Sigma)$ es la misma que la distribución de $|\Sigma|/(N-1)^p$ por el producto de p factores independientes, siendo la distribución del factor i-ésimo una χ^2 con $N-i$ grados de libertad.

Podemos dar meramente, una interpretación geométrica de este teorema. Sea $Y_i = (W_{i1}, \ldots, W_{in})$ un vector en un espacio n-dimensional. Los p vectores $Y_1, \ldots Y_p$ se distribuyen se distribuyen independientemente; cada componente de Y_i se distribuye como $N(0,1)$. $|B|$ es el cuadrado del volumen del paralelotopo definido mediante $Y_1, \ldots, Y_p$. Sea U_i el vector desde Y_i ortogonal a $Y_{i+1}, \ldots, Y_p$. Entonces el cuadrado del volumen del paralelotopo definido mediante $Y_{i+1}, \ldots Y_p, Y_i$ es el volumen al cuadrado del paralelotopo definido mediante $Y_{i+1}, \ldots, Y_p$ por $U_i' U_i = a_{ii.i+1,\ldots,p}$ (longitud al cuadrado de U_i). Se sigue que $|B| = U_1'U_1 \cdot U_2'U_2 \cdots U_p'U_p$. U_i es ortogonal a $U_{i+1}, \ldots, U_p$, por tanto la distribución condicional de U_i es normal en $(n-p+i)$ dimensiones, y las $n-(p-i)$ coordenadas se distribuyen $N(0,1)$. Así $U_i'U_i$ tiene una χ^2 con $n-(p-i)$ grados de libertad (independiente de las otras U_i's).

Para $p=1$ o 2, la distribución exacta de $|S|$ puede obtenerse con relativa facilidad, pero para valores de p más elevados las integrales involucradas en el proceso no son sencillas. No obstante, los momentos de $|S|$ pueden obtenerse fácilmente a partir de la igualdad

$$|S| = |A|/(N-1)^p$$

$|A|$ puede escribirse

$$|A| = |\Sigma| \cdot |B| = |\Sigma| \cdot \chi^2_{N-1} \cdot \chi^2_{N-2} \cdot \ldots \cdot \chi^2_{N-p}.$$

Puesto que el h-ésimo momento de una χ^2 con m grados de libertad es $2^h \Gamma\left(\frac{1}{2}m+h\right)/\Gamma\left(\frac{1}{2}m\right)$, y el momento de un producto de variables aleatorias independientes es el producto de los momentos, el h-ésimo momento de $|A|$ es

$$|\Sigma|^h \prod_{i=1}^{p} \left\{ 2^h \frac{\Gamma\left(\frac{1}{2}(N-i)+h\right)}{\Gamma\left(\frac{1}{2}(N-i)\right)} \right\} = 2^{hp} |\Sigma|^h \frac{\prod_{i=1}^{p} \Gamma\left[\frac{1}{2}(N-i)+h\right]}{\prod_{i=1}^{p} \Gamma\left[\frac{1}{2}(N-i)\right]}.$$

Así

$$E[|A|] = |\Sigma| \prod_{i=1}^{p} (N-i),$$

$$\mathrm{var}(|A|) = |\Sigma|^2 \prod_{i=1}^{p} (N-i) \left[\prod_{j=1}^{p} \Gamma(N-j+2) - \prod_{j=1}^{p} \Gamma(N-j) \right].$$

En el caso $p=1$ y $p=2$, podemos dar la distribución de $V = |A|/|S|$. Para $p=1$, V se distribuye como una χ^2 con $N-1$ grados de libertad. Para $p=2$, encontramos que el momento $k/2$ de V es

$$E\left[V^{k/2}\right] = 2^k \frac{\Gamma\left(\frac{1}{2}(N-1)+k/2\right) \Gamma\left(\frac{1}{2}(N-2)+k/2\right)}{\Gamma\left(\frac{1}{2}(N-1)\right) \cdot \Gamma\left(\frac{1}{2}(N-2)\right)}$$

Utilizando la fórmula de la duplicación para las funciones gamma,

$$2^{2\alpha-1} \Gamma(\alpha) \cdot \Gamma\left(\alpha+\tfrac{1}{2}\right) = \sqrt{\pi}\, \Gamma(2\alpha)$$

obtenemos como expresión para el momento $k/2$ de V

$$E\left[V^{k/2}\right] = \frac{\Gamma(N-2+k)}{\Gamma(N-2)}$$

que es el momento k-ésimo de $1/2$ de una variable con la distribución χ^2 con $2N-4$ grados de libertad. Resultamos el siguiente teorema:

TEOREMA DE CARLEMAN.- Si $\{\mu_i\}$ $(i=1,2,\dots)$ es una sucesión de números tal que

$$\sum_{k=1} \left(\frac{1}{\mu_{2k}}\right)^{1/2k}$$

es divergente, entonces al menos una distribución tiene la sucesión de momentos $\{\mu_i\}$.

La demostración puede encontrarse en Shohat y Tamarkin (1943) (The problem of the moments N.Y., A.M.S.). Utilizando el criterio de divergencia

$$\left(\frac{1}{\mu_{2k}}\right)^{1/2k} > \frac{1}{N-3+2k} > C\,\frac{1}{k}$$

para un C (>0) adecuado y para k suficientemente grande, podemos verificar que las condiciones del teorema anterior se satisfacen para los momentos de $\sqrt{v}$. Así $2\sqrt{v}$ se distribuye como una χ^2 con $2N-4$ grados de libertad.

Una aproximación a la densidad de $V^{1/p}$ viene dada por (Hoel, 1937)

$$\frac{c^{\frac{1}{2}p(N-p)}\, y^{\frac{1}{2}p(N-p)+1} e^{-cy}}{\Gamma\left(\frac{1}{2}p(N-p)\right)}$$

donde

$$c = \frac{p}{2}\left(1 - \frac{(p-1)(p-2)}{2N}\right)^{N/p}.$$

La distribución asintótica de la varianza generalizada

Sea $|B|/n^p = V_1(n).V_2(n)\dots V_p(n)$, donde las V's se distribuyen independientemente y $nV_i(n) = \chi^2_{n-p+i}$. Puesto que χ^2_{n-p+i} se distribuye como $\sum_{\alpha=1}^{n-p+i} W_\alpha^2$, con los W_α independientes y $N(0,1)$, el teorema central del límite (aplicado a W_α^2) establece que

$$\frac{nV_i(n) - (n-p+i)}{\sqrt{2(n-p+i)}} = \sqrt{n}\,\frac{V_i(n) - 1 + \frac{p-i}{n}}{\sqrt{2}\sqrt{1-\frac{p-i}{n}}}$$

se distribuye asintóticamente como $N(0,1)$. Entonces $\sqrt{n}\left[V_i(n)-1\right]$ se distribuye asintóticamente como $N(0,2)$. Sea ahora

$$u(n) = \begin{pmatrix} V_1(n) \\ \vdots \\ V_p(n) \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{y} \quad \sqrt{n}\left(u(n)-b\right)$$

que se distribuye asintóticamente como $N(0,T)$, siendo $T = 2I$. Definimos ahora $|B|/n^p = W = f(u_1,\dots,u_p) = u_1u_2\cdots u_p$, siendo $\partial f/\partial u_i\big|_{u=b} = 1$ y $\phi_b' T \phi_b = 2p$. Aplicando ahora un teorema límite ya utilizando cuando dimos a conocer la transformación z de Fisher para coeficientes de correlación, tendremos

$$\sqrt{n}\left(\frac{|B|}{n^p} - 1\right)$$

se distribuye asintóticamente como $N(0,2p)$.

TEOREMA.- Sea S una matriz de covarianzas muestral de orden p con n grados de libertad. Entonces $\sqrt{n}\left(|S|/|\Sigma| - 1\right)$ es asintóticamente normal con media 0 y varianza $2p$.

6. DISTRIBUCION DE UN CONJUNTO DE COEFICIENTES DE CORRELACION PARA UNA MATRIZ DE COVARIANZAS POBLACIONAL DIAGONAL.

En su momento obtuvimos la distribución para un coeficiente de correlación simple cuando la correspondiente correlación poblacional era cero. Aquí vamos a encontrar la densidad para el conjunto r_{ij} $i \neq j$ $ij = 1 \dots p$ cuando $\rho_{ij} = 0$, $i \neq j$.

Comenzaremos estudiando la distribución de A cuando Σ es diagonal, es decir $W((\sigma_{ii}\delta_{ij}), n)$. La densidad de A es

$$\frac{|a_{ij}|^{\frac{1}{2}(n-p-1)} \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{a_{ii}}{\sigma_{ii}}\right)}{2^{\frac{1}{2}np}\,\pi^{\frac{1}{4}p(p-1)}\prod_{i=1}^{p}\sigma_{ii}^{\frac{1}{2}n}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)}$$

puesto que

$$|\Sigma| = \begin{vmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_{pp} \end{vmatrix} = \prod_{i=1}^{p}\sigma_{ii}$$

Hacemos la transformación

$$(1)\quad a_{ij} = \sqrt{a_{ii}}\sqrt{a_{jj}}\, r_{ij} \qquad i \neq j$$

$$(2)\quad a_{ii} = a_{ii}$$

El Jacobiano de esta transformación viene dado por el producto del Jacobiano de (2) por el Jacobiano de (1) para a_{ii} fijo. El Jacobiano de (1) es el determinante de una matriz diagonal de orden $p(p-1)/2$ cuyos elementos en la diagonal principal son $\sqrt{a_{ii}}\sqrt{a_{jj}}$. Puesto que cada índice, por ejemplo k, aparece en el conjunto r_{ij}, $i<j$, $p-1$ veces, el Jacobiano es

$$J = \prod_{i=1}^{p} a_{ii}^{\frac{1}{2}(p-1)}.$$

Sustituyendo en la densidad de A obtendremos la densidad conjunta de $\{a_{ii}\}$ y $\{r_{ij}\}$

$$\frac{\left|\sqrt{a_{ii}}\sqrt{a_{jj}}\, r_{ij}\right|^{\frac{1}{2}(n-p-1)} \cdot \exp\left(-\frac{1}{2}\sum_{i=1}^{p}\frac{a_{ii}}{\sigma_{ii}}\right)}{2^{\frac{1}{2}np}\,\pi^{\frac{1}{4}p(p-1)}\prod_{i=1}^{p}\sigma_{ii}^{\frac{1}{2}n}\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)} \prod_{i=1}^{p} a_{ii}^{\frac{1}{2}(p-1)} = \frac{|r_{ij}|^{\frac{1}{2}(n-p-1)}}{\pi^{\frac{1}{4}p(p-1)}\prod_{i=1}^{p}\Gamma\left[\frac{1}{2}(n+1-i)\right]} \prod_{i=1}^{p}\left\{\frac{a_{ii}^{\frac{1}{2}n-1}\cdot\exp\left(-\frac{1}{2}\frac{a_{ii}}{\sigma_{ii}}\right)}{2^{\frac{1}{2}n}\sigma_{ii}^{\frac{1}{2}n}}\right\}$$

puesto que

$$\left|\sqrt{a_{ii}}\sqrt{a_{jj}}\, r_{ij}\right| = \left(\prod_{i=1}^{p} a_{ii}\right)|r_{ij}|$$

donde $r_{ii} = 1$. En el último término del producto del miembro derecho de la igualdad obtenido después de la sustitución, hagamos $a_{ii}/(2\sigma_{ii}) = u_i$; entonces la integral de este término es

$$\int_0^\infty \frac{a_{ii}^{\frac{1}{2}n-1}\cdot\exp\left(-\frac{1}{2}\frac{a_{ii}}{\sigma_{ii}}\right)}{2^{\frac{1}{2}n}\,\sigma_{ii}^{\frac{1}{2}n}}\, da_{ii} = \int_0^\infty u_i^{\frac{1}{2}n-1} e^{-u_i}\, du_i = \Gamma\left(\frac{1}{2}n\right)$$

por la definición de la función Γ (o bien por el hecho de que a_{ii}/σ_{ii} tiene una densidad χ^2 con n grados de libertad). Por tanto la densidad de r_{ij} es

$$\frac{\left[\Gamma\left(\frac{1}{2}n\right)\right]^p |r_{ij}|^{\frac{1}{2}(n-p-1)}}{\prod_{i=1}^{p}\Gamma\left(\frac{1}{2}(n+1-i)\right)\pi^{\frac{1}{4}p(p-1)}}$$

TEOREMA.- Si $X_1 \dots X_N$ son N vectores independientes, cada uno con distribución $N(\mu_i, (\sigma_{ii}\delta_{ij}))$, entonces la densidad de los coeficientes de correlación muestrales viene dada por la fórmula anterior, donde $n = N-1$.