UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMATICAS

Tesis
20

" ALGUNAS TECNICAS SOBRE DETECCION DE OUTLIERS "

JOAQUIN J. M. MUÑOZ GARCIA

Visado en Sevilla

Mayo de 1980

Fdo. Rafael Infante Macias

Fdo. Antonio Pascual Acosta

Tesis que presenta Joaquin J. M. Muñoz Garcia para optar al grado de Doctor en Ciencias Matematicas

Fdo. Joaquin J. M. Muñoz Garcia

UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMATICAS

" ALGUNAS TECNICAS SOBRE DETECCION DE OUTLIERS "

JOAQUIN J. M. MUÑOZ GARCIA

Memoria para optar al grado de Doctor en Ciencias Matematicas realizada bajo la direccion de los.

Prof. Dr. D. RAFAEL INFANTE MACIAS y
prof. Dr. D. ANTONIO PASCUAL ACOSTA.

Sevilla Mayo 1.980

ALGUNAS TECNICAS SOBRE

DETECCION DE OUTLIERS

JOAQUIN J. M. MUÑOZ GARCIA

Quiero expresar mi mas sincera gratitud al Pr. Dr. D. Rafael Infante Macias, director de esta memoria; por haber despertado en mi el interés por las Matematicas Aplicadas y por su constante estimulo y ayuda en la preparacion de esta.

Asimismo quiero expresar mi profundo agradecimiento al Pr. Dr. D. Antonio Pascual Acosta, a quien debo gran parte de mi formación sobre Calculo de Probabilidades y Estadistica Matemática y por contribuir en la realización y corrección de esta memoria.

Por ultimo quiero hacer llegar mi reconocimiento a todos aquellos que de un modo u otro, han contribuido a la realizacion de este trabajo , en especial a D. Luis Parras Guijosa.

Sevilla, Mayo de 1.980

A mi abuelo

CONTENIDO.

PROLOGO

Uno de los problemas que preocupan hoy dia fundamentalmente, a un estadistico cuando va a analizar un conjunto de datos , es la aparición de outliers o conjunto de observaciones que "parecen ser inconsistentes" con el resto del conjunto de datos. Lo que verdaderamente caracteriza a un outliers, es el "impacto " que produce en el estadistico cuando va a analizar la muestra.

Es evidente que la presencia de outliers en un conjunto de datos puede conducir a errores en nuestro intento de hacer inferencias acerca de la población de la que se han extraido, debido a que podrian falsear fuertemente las estimaciones o contrastes que hagamos sobre los parametros poblacionales.

La presencia de outliers plantea por tanto un problema fundamental en el análisis de los datos, por lo que es natural buscar medios para

detectar e interpretar la presencia de dichas observaciones , rechazandolas en algunas ocasiones para así restablecer las propiedades de los datos , o por lo menos teniendo en cuenta su presencia en todos los análisis estadisticos.

El objetivo de esta Memoria se centra en la obtención de nuevos criterios para la detección de outliers en poblaciones normales multivariantes.

En el primer capitulo despues de plantear de forma general el problema que conlleva la presencia de outliers en un conjunto de datos, se pasa a describir el esquema que ha de seguir cualquier regla o técnica de detección de outliers. En el último apartado del capitulo incluimos un breve resumen con comentarios históricos sobre estas técnicas primero para el caso de poblaciones univariantes y en el segundo epigrafe para poblaciones multivariantes.

El capitulo segundo de la Memoria se dedica a la obtención de criterios para la detección de outliers basados en el concepto de función influencia. Estas reglas estan dadas para situaciones en las que se requiere aplicar una técnica de analisis discriminante, y han de aplicarse una vez obtenida la muestra, antes del calculo de la función discriminante donde ya no podrán ser incluidas las observaciones consideradas outliers.

Utilizando como técnica el test de cociente de verosimilitudes, el capitulo tercero se dedica al estudio de un criterio para detectar outliers en muestras procedentes de poblaciones normales multivariantes. Se incluye al final un programa de ordenador para detección de outliers mediante este metodo y se aplica en el último epigrafe a una situación práctica

En el último capitulo de la Memoria y utilizando como estadistico el máximo del cuadrado de la distancia entre las matrices de sumas de cuadrado y sumas de productos de observaciones muestrales, damos un procedimiento para detectar la presencia de outliers en muestras procedentes de poblaciones normales bivariantes. Despues de dar un programa de ordenador para la resolución de esta técnica se aplica dicho procedimiento en una situación práctica.

CAPITULO PRIMERO

INTRODUCCION A LAS TECNICAS DE DETECCION DE OUTLIERS

CONTENIDO.

1.1. PLANTEAMIENTO GENERAL DEL PROBLEMA.

Según ANSCOMBE (1.960) existen tres fuentes de variabilidad sobre los elementos de una muestra o de un conjunto de datos extraidos de una cierta población y que se pueden clasificar de la siguiente forma:

Variabilidad intrinseca.- Es la variación inherente a la población. Tal variación surge de una forma natural y no puede ser reducida sin modificar la población objeto del estudio.

Variabilidad debida al error en las medidas.- Es el error que lleva consigo la falta de precisión de los instrumentos de medida utilizados para realizar el experimento. Dentro de esta fuente de variabilidad podemos incluir el error de transcripción de las observaciones.

Variabilidad debida al error de ejecución.- Es debida a una recolección imperfecta de los datos. Puede ocurrir que tomemos una muestra sesgada o que incluyamos en la misma observaciones no representativas de la población que estamos estudiando.

Basandose en estas fuentes de variabilidad Anscombe distingue dos tipos de observaciones

- Outliers: que son aquellas observaciones que presentan una gran variabilidad de tipo intrinseco.
- Falsas observaciones: que son debidas a un gran error de ejecución y de medida.

Sin embargo al igual que BARNETT (1.978) discrepamos de esta terminologia debido al hecho que ante un conjunto de datos extraidos de una cierta población, generalmente no podemos determinar hasta donde llega una fuente de variabilidad y donde comienza la otra. Por ello seguimos en esta memoria el criterio de llamar outliers a aquella o aquellas observaciones que presentan una gran variabilidad de cualquiera de los tipos descritos por Anscombe, con respecto al conjunto principal de datos.

De entre las diversas definiciones del concepto "outliers" que aparecen a lo largo de las numerosas publicaciones estadisticas sobre el tema extraemos a continuación, las que consideramos mas interesantes.

Para GUMBEL (1.960) " los outliers son aquellos valores que parecen demasiado grandes o demasiado pequeñas comparadas con el resto de las observaciones".

En 1.961 FERGUSON da la siguiente idea de outliers " En una muestra extraida de una cierta población aparecen una o varias observaciones que sorprendentemente , se encuentran lejos del grupo principal de datos"

GRUBBS (1.969) afirma que un outliers " es una observación que se presenta fuertemente desviada de los otros miembros de la muestra".

Por ultimo Barnett (1.978) , considera un outliers " como una observación o conjunto de observaciones que parecen ser inconsistentes con el resto del conjunto de datos".

De estas definiciones se deduce que lo que caracteriza a una observacion outliers es el "impacto" que produce en el estadistico cuando va a analizar los datos.

Es evidente que la presencia de outliers en un conjunto de datos, puede conducirnos a errores en nuestro intento de hacer inferencias acerca de la población de la que se han extraido, debido a que falsearan fuertemente las estimaciones o contrastes que hagamos sobre los parametros poblacionales.

La presencia de outliers plantea por tanto un problema fundamental en el analisis de datos.

Podemos distinguir para la resolución de este problema dos grandes áreas: Reglas o técnicas de detección de outliers y estimación robusta. El objetivo de esta memoria se encuentra dentro de la primera area y por ello trataremos de una manera suscinta el problema de la estimación ro-

busta, en este primer capitulo.

Por técnica de detección de outliers entendemos un procedimiento que permite detectar de una manera objetiva la presencia de outliers en una muestra y determinar qué observaciones son outliers. En esta definición hemos de destacar la palabra "objetiva" , debido a que existen autores, que no consideran la existencia de outliers sino simplemente de falsas observaciones en el sentido dado por Anscombe y que piensan que éstas deben ser detectadas de forma "subjetiva" por la experiencia del observador.

Para dar cualquier técnica de detección de outliers, en general se ha de seguir el siguiente esquema:

Se determina un estadistico que debe incluir la observación u observaciones que suponemos que son outliers, a continuación se halla la distribución que sigue dicho estadistico bajo la hipótesis de que todas las observaciones constituyen una muestra procedente de una misma población.

Para un nivel de significación α prefijado, que según GRUBBS (1.969) debe variar entre el 1% y el 5%, y que puede interpretarse como el riesgo de considerar una buena observación como outliers, se determina el valor critico o percentil de la distribución teórica del estadistico correspondiente a esa probabilidad α . Esto nos permite determinar la región critica de esta regla de detección.

Si para la muestra dada, el valor del estadistico, cae en la región critica, diremos que existen observaciones outliers, para un nivel de significación α .Una vez detectada la presencia de outliers la técnica o regla utilizada nos permitirá en cada caso determinar qué observación u observaciones pueden ser consideradas outliers.

Una vez que se sabe que observaciones son outliers, el procedimiento a seguir podria ser uno de los siguientes.

Una primera solución seria despreciar dichas observaciones y tomar unas nuevas en su lugar , si esto es posible, y aplicar otra vez la tecnica de detección de outliers a este nuevo conjunto de datos. No obstante aconsejamos no despreciar las observaciones consideradas outliers y reservarlas para un analisis posterior de las mismas ya que este nos puede conducir a determinar las fuentes de variabilidad que han dado lugra a la presencia de outliers, pudiendolas corregir para un futuro experimento.

Otra solución y pensamos que esta se ha de realizar cuando detectamos un numero de outliers próximo a $[\frac{n}{2}]$ (parte entera de $\frac{n}{2}$), es tomar una nueva muestra y realizar de nuevo el estudio. Si no es posible tomar esta nueva muestra, entonces hemos de cambiar el modelo de distribución supuesto al principio por otro modelo que puede ser incluso una mixtura de distribuciones.

Como indicamos anteriormente otra forma posible de abordar el problema es basandose en la robustez. Puede ocurrir que no sea posible realizar una técnica para detectar outliers o aunque existan outliers, que queramos retenerlos en la muestra porque pensemos que los mismos contienen una información que nos interesa o por cualquier otra circunstancia. Estos problemas pueden resolverse mediante la estimación robusta de los parametros poblacionales sobre los que queremos realizar la inferencia.

1.2. COMENTARIOS HISTORICOS SOBRE TECNICAS DE DETECCION DE CUTLIERS.

Incluimos en este apartado un resumen histórico sobre las diferentes reglas o métodos para la detección de outliers. Conviene indicar que aunque los inicios de estas técnicas para poblaciones univariantes pueden remontarse hasta mediados del siglo XVIII no es hasta la decada de los cincuenta, cuando aparecen las primeras referencias sobre detección de outliers para poblaciones multivariantes, debido a los problemas que surgen en el caso multivariante y a los que haremos referencia mas adelante.

1.2.1. Outliers en poblaciones univariantes.

Los inicios de estas técnicas se pueden encontrar en el siglo XVIII cuando Boscovich en 1755, intenta determinar la elipticidad de la Tierra promediando las medidas de exceso de los grados polares alrededor del ecuador, descartando los valores extremos. Daniel Bernouilli en su trabajo " Dijudicatio maxime probabilis plurium observationum discrepantium atque verisimilima inductio inde formada " de 1777, aprueba el rechazo de observaciones extremas. Y Legendre en su obra " Nouvelles methodes pour la determination des orbites des cometes" en 1.805 recomienda tambien el rechazo de observaciones.

Sin embargo la escuela germana de astronomos en un trabajo publicado por Bessel hacia 1.838, afirma al efecto, que nunca una observación será rechazada por su gran residuo pues el conjunto de todas las observaciones contribuiran al resultado y afirman al mismo tiempo. " Creemos que unicamente mediante una observación estricta de esta regla, se elimina la arbitrariedad de los resultados".

Vemos como en estos comienzos el rechazo de observaciones se realiza unicamente a criterio del observador. Así Saunder en su trabajo de 1.90

llega a escribir que la experiencia práctica del individuo ha de contar sobremanera para el rechazo de observaciones.

Sin embargo el primer intento de dar un criterio de rechazo basado en algun tipo de razonamiento probabilistico es debido a Peirce (1.852) El argumento de Peirce es reproducido por Chauvenet (1.863) el cual da una regla similar basada en un argumento mas simple. Stone hacia 1.868 introduce un test de rechazo basado en un módulo de descuido o negligencia, lo cual suponia un error al tomar las observaciones el experimentador. Pero podemos decir que el trabajo más serio en estos inicios es el Glaisher (1.872) " On the rejection of discordant observations " en el que trabaja ya con observaciones procedentes de poblaciones normales univariantes dando un procedimiento de medias ponderadas. Un año mas tarde Stone hace una critica a este metodo dando un procedimiento de ponderación ., basado en la maximización de la función de verosimilitud y posteriormente este mismo criterio fue dado tambien por Edgeworth en 1.883.

En esta misma linea podemos citar el trabajo de Wrigth de 1.884 " A treatise on the adjustment of observations by the method of least squares " en el que rechaza las observaciones que se desvian de la media en mas de tres veces la desviación tipica; y el procedimiento de Goodwin (1.913) que rechaza una observacion en una muestra de tamaño n si la desviación de la media de las restantes n - 1 observaciones excede cuatro veces a la desviación promedio de las n - 1 observaciones

Hemos de señalar que todos estos procedimientos en cierto modo rudimentarios adolecen de un defecto general que consiste en no distinguir la varianza poblacional y la varianza muestral. Es IRWIN (1.925) quien da el primer procedimiento para detección de outliers en el que ya

distingue entre varianza muestral y varianza poblacional. Para el caso en que σ es conocida propone el test estadistico,

$$\frac{[x_{(n)} - x_{(n-1)}]}{\sigma} \quad \text{y} \quad \frac{[x_{(n-1)} - x_{(n-2)}]}{\sigma}$$

donde $x_{(i)}$ denota el estadistico ordenado de rango i . Conviene citar tambien a STUDENT (1.927) que da un criterio basado en el recorrido muestral y THOMPSON (1.935) que da un test exacto studentizado,

$$\frac{x - \bar{x}}{S}$$

A partir de este año los estadisticos prestan una mayor atención a estos problemas y las técnicas de detección de outliers unidimensionales se hacen mas rigurosas.

GRUBBS (1.950) da un criterio basado en el cociente entre la suma de cuadrados de desviaciones para una muestra reducida y la suma de cuadrados de las desviaciones para lamuestra completa,DIXON (1.950) da varias reglas basadas en la razon entre la observacion "sospechosa" y su mas proxima y el recorrido o rango muestral y DAVID , PEARSON y HARTLEY (1.954) dan un test para la detección de outliers en poblaciones normales univariantes basado en la razon entre el recorrido muestral y la desviación tipica muestral.

KUDO (1.956) , da unas ciertas reglas para la detección de outliers en poblaciones normales univariantes tanto para el caso de conocer la varianza poblacional como cuando no se conoce. Dichos metodos se basan en el ordenamiento de lás observaciones. Kudo describe cinco criterios por los que una regla de rechazo puede considerarse òptima.

Como caso particular de este trabajo de Kudo se obtiene el estadistico de PEARSON y CHANDRASEKAR (1.936) que ha tenido una gran importancia

en el desarrollo de las reglas de rechazo para outliers, probando que este estadistico es óptimo segun los criterios por él dados y demostrando que el de NAIR (1.948) no posee esta propiedad de optimalidad.

En el año 1.960 , aparece un trabajo de ANSCOMBE que ha tenido una importancia fundamental en el desarrollo ulterior de esta teoria. Comienza dando los tipos de variabilidad que se pueden presentar en un conjunto de datos y a los que ya hemos hecho referencia en el apartado anterior, introduciendo a continuación dos conceptos que el denomina " premium " y " protection " con los que trata de dar un criterio para poder estudiar y comparar todas las reglas de detección de outliers. Define el " premium " como el porcentaje en que se incrementa la varianza del error de estimación debido al uso del criterio de detección cuando de hecho todas las observaciones proceden de una misma población normal y la " protection " como la reduccion en la varianza o error cuadratico medio cuando aparecen observaciones outliers.

ANSCOMBE Y TUKEY (1.963) definen el residuo de una observación de la siguiente forma.

(residuo) = (valor observado) - (valor ajustado)

En este trabajo, dan algunas técnicas de tipo gráfico y otras de tipo numérico para discutir los residuos que resultan de las medias por filas y columnas en un analisis de varianza de dos factores o más. Afirman que : " La razon mas importante para el calculo de los residuos es detectar outliers, observaciones que tienen un gran residuo en comparación con la mayor parte de las otras y esto debe ser tratado especialmente ". En este mismo trabajo dan ciertos estadisticos para la detección de outliers.

Esta definición de residuo es generalizada por COX y SNELL (1.968), indicando los campos de aplicabilidad de esta definición. Estudian las

propiedades de los residuos, haciendo especial hincapié en los residuos de poblaciones Binomial y Poisson; es interesante la discusión que de dicho trabajo hacen destacados especialistas en el tema como , Anscombe F. J., Barnett V., Mallows C. L. , Pearce S. C., Harrison P. J., entre otros.

Basandose en los trabajos de Anscombe , TIAO y GUTTMAN (1.967) trata de dar una estimación para la media de una población normal unidimensional a partir de una muestra de tamaño n, con la posibilidad de que en dicha muestra existan uno o más outliers. Dan unas reglas de rechazo basadas en el analisis de los residuos, que suponen estan correlacionados. Tambien hacen un estudio del " premium" y " protection " de los metodos de detección de outliers que proponen.

En esta linea se encuentra el trabajo de ANDREWS (1.971), que basandose en la distribución conocida de los residuos en el modelo de regresión lineal, da tests de significación exactos, aplicando a continuación dichos resultados para contrastar la existencia de uno o mas outliers.

TIETJE Y MOORE (1.972), basandose en la idea de outliers dada por KENDALL Y BUCKLAND (1.957) definen el efecto de "enmascaramiento" , como la imposibilidad de una regla de rechazo de identificar una observación outliers en presencia de varios valores "sospechosos", ilustrando dicho efecto con un ejemplo sobre isotopos del uranio, aplicandole el estadistico de GRUBBS (1.950).

BROWN (1.975) da tests estadisticos para la detección de outliers basandose en los residuos, que resultan en el analisis de varianza de dos factores, pero a diferencia de otros tests basados en residuos, Brown no se basa en el valor del residuo, sino en el signo de los residuos a que da lugar cada una de las filas y columnas. Obteniendo

así un estadistico que se distribuye aproximadamente como una ley χ^2.

En el año 1.975 ROSNER publica un articulo de bastante importancia en la detección de outliers en muestras univariantes, pues en él compara la potencia de varias reglas de rechazo.

Comienza exponiendo el problema de detección de outliers en muestras de poblaciones normales. Así afirma que.

" Si $x_1, x_2, \ldots, x_n$ es una muestra aleatoria de una muestra compuesta de dos subconjuntos $J_1 = \{ x_{i1}, x_{i2}, \ldots, x_{im} \}$, $J_2 = \{ x_{j1}, x_{j2}, \ldots, x_{jk} \}$ donde $x_{iq} \in N(\mu , \sigma^2)$ $q = 1,2,\ldots,m$ y $x_{jl} \quad N(\mu_{jl} , \sigma^2_{jl})$ $l = 1,2,\ldots,k$ donde todos los parametros son desconocidos. Llamaremos J_1 el subconjunto de las verdaderas observaciones y a J_2 subconjunto de los outliers ($0 \leq k < n$ y $m + k = n$). El problema de la detección de outliers consiste en identificar los subconjuntos J_1 y J_2".

Generaliza cuatro tipos de estadisticos para la detección de un outliers, modificando despùes los estadisticos para que sirvan para detectar mas de un outliers. Estos cuatro estadisticos son.

Desviación extrema studentizada (ESD).

$$ESD = \max_{i=1,\ldots,n} \frac{|x_i - \bar{x}|}{S}$$

Recorrido studentizado (STR).

$$STR = \frac{(x_{(n)} - x_{(1)})}{S}$$

Kurtosis (KUR)

$$KUR = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2}$$

R - estadistico (RST)

$$RST = \max_{i=1,\dots,n} \frac{|x_i - a|}{b}$$

donde a y b son la media y varianza muestral corregidas.

Y de estos estadisticos estudia su potencia mediante procedimientos de simulación, llegando a las conclusiones siguientes:

1) Entre los procedimientos para la detección de mas de un outliers el procedimiento STR es claramente el inferior.
2) Los otros tres procedimientos se pueden considerar iguales, aunque el ESD es ligeramente mejor.
3) Para efectos computacionales da el siguiente orden de menos a mas complejos, RST,ESD,STR,KUR.
4) El procedimiento preferido derá el ESD, que es generalmente mejor y computacionalmente es muy razonable.

Hemos dado hasta aquí un breve esquema sobre reglas de detección de outliers para poblaciones normales univariantes. Una de las lineas de investigación en este campo del Analisis de Datos consiste en encontrar reglas de rechazo para muestras procedentes de poblaciones con diferentes leyes de probabilidad. Así el trabajo de SHAPIRO - WILK (1.972) para el caso de una muestra de tipo exponencial o el más reciente de COLLET (1.980) para conjunto de datos extraidos de una población con distribución de Von Mises.

1.2.2. Outliers en poblaciones multivariantes.

Según acabamos de ver la mayoria de los criterios de detección de outliers, para datos unidimensionales están basados en estadisticos ordenados o en los tamaños de los residuos.

En el caso de poblaciones multidimensionales surgen problemas no existentes en el caso anterior y que hacen aumentar la complejidad del estudio de reglas de rechazo para datos multidimensionales.Así por ejemplo; la imposibilidad de ordenar las observaciones, que una observación puede ser considerada outliers porque exista una gran variabilidad en una de sus componentes o porque existan pequeñas variabilidades en todas sus componentes etc. Conviene tambien señalar aunque parezca obvio que una muestra p - variante no puede ser estudiada como un conjunto de p - muestras univariantes pues se perderia el sentido de dependencia o correlación entre las variables e incluso pueden aparecer outliers unidimensionales y sin embargo el vector del que es componente no lo sea.

Como GNANADESIKAN y KETTERING (1.972) señalan, " La complejidad del caso multivariante ,lleva consigo que sea infructuoso buscar procedimientos de detección de outliers , que sean válidos para todas las situaciones. Mas razonable parece buscar procedimientos de detección de outliers contra situaciones de tipo especifico, debiendose construir un gran conjunto de técnicas con diferentes sensibilidades, de forma qu un outlier para una cierta situación puede no serlo para otra."

Podemos decir que las primeras publicaciones sobre detección de outliers en este campo, son una generalización del trabajo realizado por FERGUSON en 1.961 sobre detección de outliers en poblaciones normales univariantes. Todos ellos tienen como base el exponente de

la distribución normal multivariante

$$Q(x,\mu,\Sigma) = Q(x) = (X-\mu)'\Sigma^{-1}(X-\mu)$$

y se considera este estadistico como una distancia dentro del espacio de las observaciones normales multivariantes. Dicha distancia está estudiada bajo los supuestos de que se conozcan los dos parametros μ y Σ uno de ellos o ninguno de los dos.

Para todos estos casos se realiza un ordenamiento de las observaciones basado en la distancia antes descrita y se toma generalmente como estadistico para decidir si una observación es o no outliers.

$$T = \max_i Q(x_i)$$

Así si μ y Σ son conocidos, tendremos estadisticos ordenados de una distribución gamma, y en este caso se compara el estadistico T con un punto critico dado por un nivel de significación fijado de antemano. Dicho punto critico podemos obtenerlo de las tablas dadas por GUPTA (1.960), para la distribución de los estadisticos ordenados de una distribución gamma.

En el caso de que μ y Σ o uno de los dos no es conocido la distribución de T no está determinada de forma exacta, sin embargo SIOTANI (1.959) da una tabulación de una distribución aproximada a la de T.

Quizás el estudio mas riguroso sobre detección de outliers, es el realizado por WILKS (1.963), quien da un estadistico para detectar k outliers k = 1,2,..... , basandose para ello en ideas geométricas. Así si se tiene una distribución normal p - dimensional, de la que extraemos una muestra de tamaño n (n > p), como estadistico para k = 1 comsidera lo que él llama razón de dispersión uni - outliers , y que

representa por

$$R_i = \frac{|S^{(i)}_{(n-1)}|}{|S_{(n)}|} \qquad i = 1, 2, \ldots, n$$

siendo la razón de dispersión k - outliers

$$R_i = \frac{|S^{(i)}_{(n-k)}|}{|S_{(n)}|} \qquad i = 1, 2, \ldots, \binom{n}{k}$$

donde $S_{(n)}$ es la matriz de sumas de cuadrados y sumas de productos de n observaciones muestrales y $S^{(i)}_{(n-1)}$ es la matriz de s. c. y s. p. de n -1 observaciones de la muestra de tamaño n donde se ha quitado la observación i - ésima.

El estadistico R_i se puede interpretar como la razón de los volúmenes entre paralelotopos, ya que WILKS (1.962), relaciona $|S_{(n)}|$ con el volumen de los paralelotopos formados con p - puntos de los n y el centro de gravedad de la muestra $\bar{X}' = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)$.

Con este estadistico Wilks considera que el mayor "candidato" a ser un outliers es aquel elemento muestral que alcanza el $\min_i R_i$ y será considerado outliers si es menor que un cierto punto critico , determinado bajo un nivel de significación dado, mediante la distribución del $\min_i R_i$.

En 1.972 GNANADESIKAN Y KETTERING realizan un trabajo de recopilación, en el que uno de sus puntos trata sobre detección de outliers en poblaciones normales multivariantes . Estas técnicas estan basadas en procedimientos gráficos entre los que podemos citar uno para detección de outliers en analisis discriminante y el basado en el analisis de las componentes principales.

Sobre este ultimo método HAWKINS (1.974) hace algunas modificaciones

realizando aplicaciones sobre datos geológicos y dando una comparación entre diferentes estadisticos que surgen a lo largo de su trabajo.

Es obvio que una forma gráfica para detectar outliers en distribuciones bidimensionales o tridimensionales seria representar la nube de puntos de las observaciones y ver que valores de la variable se aleja de la misma, sin embargo para dimensiones mayores esto nos resulta imposible, ANDREWS (1.972) propone una técnica para representar observaciones p - dimensionales sin ninguna restricción para p. Afirma que un dato p - dimensional $X' = (X_1, X_2, \ldots, X_p)$ define una función.

$$f_x(t) = \frac{X_1}{\sqrt{2}} + X_2 \operatorname{sen} t + X_3 \cos t + X_4 \operatorname{Sen}. 2t + X_5 \cos. 2t + \ldots$$

siendo dibujada esta función en el recorrido $-\pi < t < \pi$ .Andrews aplica este metodo a un trabajo de antropologia. Está representación gráfica de puntos p - dimensionales es aprovechada por GNANADESIKAN (1.973) para dar otra técnica de tipo gráfico para la detección de outliers.

DEVLIN y otros (1.975) describen un procedimiento para detección de outliers, al utilizar el coeficiente de correlación, basandose en la función influencia dada por HAMPEL (1.974) aunque ellos utilizan la función influencia muestral dada por

$$I_-(y_i, \hat{\theta}) = (n-1)(\hat{\theta} - \hat{\theta}_{-i}) \quad i = 1, 2, \ldots, n.$$

Este metodo tambien está basado en representaciones gráficas de esta función influencia. En este mismo trabajo podemos ver una aplicación al ejemplo que toman de FISHER (1.936) sobre la longitud y anchura de los sepalos de cincuenta plantas de Iris Setosa.

Por ultimo para acabar esta breve reseña histórica citaremos las obras de GNANADESIKAN (1.977) en la que expone diversas técnicas para la detección de outliers multivariantes y la más reciente de BARNETT (1.978) dedicada toda ella al estudio de outliers en el Analisis Estadistico de Datos.

CAPITULO SEGUNDO

DETECCION DE OUTLIERS MEDIANTE LA FUNCION INFLUENCIA

CONTENIDO.

2.1. INTRODUCCION.

El objetivo de este capitulo se centra en la obtención de criterios para la deteccion de outliers, basandonos en el concepto de funcion influencia introducida por HAMPEL (1.974). Estos criterios estan dados, para situaciones en las que se requiera aplicar una tecnica de analisis discriminante, y han de aplicarse una vez obtenida la muestra; antes del calculo de la funcion discriminante, donde ya no seran incluidas las observaciones consideradas outliers.

A continuacion introducimos algunos de los conceptos, que se van a utilizar a lo largo del capitulo.

ANDERSON y BAHADUR (1.962), dan un procedimiento para discriminar entre dos poblaciones normales multivariantes, $N_p(\mu_1, \Sigma_1)$ y $N_p(\mu_2, \Sigma_2)$ no singulares, con vectores medias y matrices de covarianza distintas. Formalmente el problema se resuelve de la siguiente manera:

Sea b un vector p - dimensional de componentes reales y c una constante real, una observacion x p - dimensional es clasificada como perteneciente a la primera poblacion si $b'x \leq c$ y de la segunda si $b'x > c$.

Las posibles probabilidades de errores en el procedimiento de clasificacion son.

$$\varepsilon_1 = P\left[X \in N_p(\mu_2, \Sigma_2) / X \in N_p(\mu_1, \Sigma_1)\right] = 1 - F_{N_1(0,1)}\left(\frac{c - b'\mu_1}{(b'\Sigma_1 b)^{1/2}}\right)$$

$$\varepsilon_2 = P\left[X \in N_p(\mu_1, \Sigma_1) / X \in N_p(\mu_2, \Sigma_2)\right] = 1 - F_{N_1(0,1)}\left(\frac{b'\mu_2 - c}{(b'\Sigma_2 b)^{1/2}}\right)$$

Se trata de minimizar las probabilidades ε_1 y ε_2 o lo que es lo mismo maximizar los argumentos.

$$Z_1 = \frac{c - b'\mu_1}{(b'\Sigma_1 b)^{1/2}} \qquad Z_2 = \frac{b'\mu_2 - c}{(b'\Sigma_2 b)^{1/2}}$$

Anderson y Bahadur demuestran que el procedimiento definido por.

$$b = (\lambda_1 \Sigma_1 + \lambda_2 \Sigma_2)^{-1} (\mu_2 - \mu_1)$$

con

$$C = b'\mu_2 - \lambda_2 b' \Sigma_2 b = b'\mu_1 + \lambda_1 b' \Sigma_1 b$$

para λ_1 y λ_2 arbitrarios tales que $\lambda_1 \Sigma_1 + \lambda_2 \Sigma_2$ sea definida positiva, es un procedimiento admisible.

El problema surge al determinar λ_1 y λ_2 de forma que se cumplan las condiciones anteriores. Para su resolucion existen varios metodos. Nos basaremos en el procedimiento minimax que consiste en igualar Z_1 y Z_2 por lo que las probabilidades de clasificacion correcta serian mayores que $1/2$ y se verifica $\lambda_1 = 1 - \lambda_2 = \lambda$.

$$0 = Z_1^2 - Z_2^2 = b' [\lambda^2 \Sigma_1 - (1-\lambda)^2 \Sigma_2] b$$

y esta ecuacion tiene solucion unica ya que Z_1^2 es creciente con λ y Z_2^2 es decreciente con λ .

Por otro lado CHERNOFF (1.972 , 1973) utiliza como medida para discriminar entre dos poblaciones $N_p(\mu_1 , \Sigma_1)$, $N_p(\mu_2 , \Sigma_2)$ no singulares , la distancia S^2 .

$$S^2 = \lambda(1-\lambda)\, \delta' (\lambda \Sigma_1 + (1-\lambda)\Sigma_2)\, \delta$$

donde t verifica las mismas condiciones que las del procedimiento minimax descrito anteriormente y por δ representamos el vector diferencia

$$\delta = \mu_2 - \mu_1$$

Otro concepto de gran importancia a lo largo de este capitulo es el de funcion influencia que podriamos definir en la forma siguiente.

Sea Ω un espacio metrico separable y completo, sea T una funcion vectorial definida en un subconjunto del espacio de todas las medidas de probabilidad sobre Ω y con valores en el espacio euclideo k - - dimensional $\mathbb{R}^K$ y sea F una funcion del dominio de T . Si notamos δ_w la distribucion degenerada en un punto arbitrario $w \in \Omega$, la curva de influencia de T en F se define, mediante el limite.

$$CI_{T,F}(w) = \lim_{\varepsilon \to 0} \left\{ \frac{T[(1-\varepsilon)F + \varepsilon \delta_w] - T(F)}{\varepsilon} \right\}$$

si este limite existe para cada $w \in \Omega$.

La funcion influencia trata pues de reflejar la posible "influencia" o "perturbacion" que experimentan los parametros poblacionales, bajo la hipotesis de que la variable proceda de una poblacion con funcion de distribución $(1-\varepsilon)F + \varepsilon \delta_w$ en lugar de F . $CI_{TF}(\cdot)$ va a depender del punto en el que suponemos está definida la distribución degenerada, por lo que dicha funcion puede considerarse como una trasformacion de la variable aleatoria, y tendrá por tanto una distribución de probabilidad.

El procedimiento a seguir consiste en determinar la distribución de la funcion influencia bajo la hipotesis de que no existen outliers es decir que la variable no se distribuye segun la distribucion $(1+\varepsilon)F + \varepsilon \delta_w$. Para lo cual será necesario que estudiemos, como son perturbados los parametros poblacionales al considerar la distribución $(1-\varepsilon)F + \varepsilon \delta_w$, calculando a continuacion la funcion influencia CI_{TF} y por ultimo su distribución.

2.2. CALCULO DE LOS PARAMETROS CARACTERISTICOS DE LA DISTRIBUCION PERTURBADA.

Sean dos poblaciones $N_p(\mu_1, \Sigma_1)$ y $N_p(\mu_2, \Sigma_2)$ no singulares.

Representemos por $F_1^*(w)$, la funcion de distribucion perturbada de la funcion de distribución $F_1(w)$ de la poblacion $N_p(\mu_1, \Sigma_1)$. Definida por

$$F_1^*(w) = (1-\varepsilon)F_1(w) + \varepsilon \delta_w$$

donde $\varepsilon \in \mathbb{R}$ con $0 \leq \varepsilon \leq 1$.

Pasemos ahora a calcular los parametros caracteristicos de la distribución F_1^* que van a ser necesarios para la determinación de la curva influencia.

Así la media de esta distribución será.

$$\mu_1(F_1^*) = \int w \, dF_1^*(w) = \mu_1 + \varepsilon(W - \mu_1)$$

Por otro lado la matriz de covarianza de la poblacion perturbada, que notaremos por $\Sigma_1(F_1^*)$, toma la forma

$$\Sigma_1(F_1^*) = \int (w-\mu_1)(w-\mu_1)' \, dF_1^*(w) = (1-\varepsilon)\Sigma_1 + \varepsilon(w-\mu_1)(w-\mu_1)'$$

Si notamos

$$W - \mu_1 = X$$

tendremos

$$\mu_1(F_1^*) = \mu_1 + \varepsilon X$$

$$\Sigma_1(F_1^*) = (1-\varepsilon)\Sigma_1 + \varepsilon X X'$$

Veamos a continuación la forma que tendria la matriz $\Sigma = \lambda\Sigma_1 + (1-\lambda)\Sigma_2$, bajo la distribucion F_1^*, ya que dicha matriz va a ser utilizada en desarrollos posteriores. Así

$$\Sigma(F_1^*) = \lambda\Sigma_1(F_1^*) + (1-\lambda)\Sigma_2 = \Sigma - \lambda\varepsilon\Sigma_1 + \varepsilon\lambda X X'$$

cuya inversa es:

$$\left\{\Sigma(F_1^*)\right\}^{-1} = \left\{\Sigma - \lambda\varepsilon\Sigma_1 + \varepsilon\lambda X X'\right\}^{-1} = \left\{(\Sigma - \varepsilon\lambda\Sigma_1) \cdot \right.$$

$$\left. \cdot \left(I_p + \varepsilon\lambda(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X X'\right)\right\}^{-1} = \left(I_p + \varepsilon\lambda(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X X'\right)^{-1}\left(\Sigma - \varepsilon\lambda\Sigma_1\right)^{-1}$$

y aplicando la igualdad matricial (BELLMAN 1.965)

$$(I + AB)^{-1} = I - A(I + AB)^{-1}B$$

se obtiene

$$\left\{\Sigma(F_1^*)\right\}^{-1} = \left\{I_p - \varepsilon\lambda(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X \left(I_p + \varepsilon\lambda X'(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X'\right\} \times \right.$$

$$\times \left(\Sigma - \varepsilon\lambda\Sigma_1\right)^{-1} =$$

$$= \frac{(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} + \varepsilon\lambda X'(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X (\Sigma - \varepsilon\lambda\Sigma_1)^{-1}}{1 + \varepsilon\lambda X'(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X} -$$

$$- \frac{\varepsilon\lambda(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X X'(\Sigma - \varepsilon\lambda\Sigma_1)^{-1}}{1 + \varepsilon\lambda X'(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} X} \qquad (1)$$

Desarrollando (1) en serie de Mac - Laurin en funcion de ε hasta el termino de primer orden ya que los restantes vendrian dados en función de ε^H $(H > 1)$, y estos terminos se harian cero al efectuar el calculo de la funcion influencia, resulta

$$\left\{\Sigma(F_1^*)\right\}^{-1} = \left(I_p + \varepsilon\lambda\Sigma^{-1}\Sigma_1\right)\Sigma^{-1} - \varepsilon\lambda\Sigma^{-1} X X'\Sigma^{-1}$$

habiendose utilizado para ello

$$\frac{d}{d\varepsilon}(\Sigma - \varepsilon\lambda\Sigma_1)^{-1} = (\Sigma - \varepsilon\lambda\Sigma_1)^{-1}\lambda\Sigma_1(\Sigma - \varepsilon\lambda\Sigma_1)^{-1}$$

Por ultimo el vector δ se trasforma a su vez en

$$\delta(F_1^*) = \mu_2 - \mu_1(F_1^*) = \delta - \varepsilon X$$

2.3. FUNCION INFLUENCIA PARA LA DISTANCIA S^2.

En este apartedo se estudia en primer lugar como seria esta distancia S^2, bajo la hipotesis de que la distribucion de la poblacion es de la forma $(1-\varepsilon)F_1(w)+\varepsilon\delta_w$, para ello bastará utilizar los parametros perturbados, que hemos obtenido en el apartado 2.2. A continuación determinamos la funcion influencia para este parametro, y por ultimo calculamos la distribucion de dicha funcion influencia.

2.3.1. Calculo de $S^2(F_1^*)$.

Por definición

$$S^2 = \lambda(1-\lambda)\delta'\Sigma^{-1}\delta$$

y basandonos en los resultados del apartado anterior

$$S^2(F_1^*) = \lambda(1-\lambda)(\delta-\varepsilon x)'\{(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1}-\varepsilon\lambda\Sigma^{-1}xx'\Sigma^{-1}\}$$

$$\cdot(\delta-\varepsilon x) = \lambda(1-\lambda)\{\delta'(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1}\delta - \varepsilon\lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}\delta -$$

$$- \varepsilon x'(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1}\delta + \varepsilon^2\lambda x'\Sigma^{-1}xx'\Sigma^{-1}\delta -$$

$$- \varepsilon\delta'(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1}x + \varepsilon^2\lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}x +$$

$$+ \varepsilon^2 x'(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1}x - \varepsilon^3\lambda x'\Sigma^{-1}xx'\Sigma^{-1}x.$$

Y debido a que en el calculo de la funcion influencia tomaremos limite en ε para cuando $\varepsilon \longrightarrow 0$ podemos despreciar en $S^2(F_1^*)$, los terminos con potencias de ε mayores que la unidad. Por lo que esta distancia perturbada quedaria

$$S^2(F_1^*) = \lambda(1-\lambda)\{\delta'\Sigma^{-1}\delta + \varepsilon\lambda(\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta) -$$

$$- \varepsilon\lambda(\delta'\Sigma^{-1}x)^2 - 2\varepsilon\delta'\Sigma^{-1}x\}$$

2.3.2. Funcion influencia para S^2.

$$CI_{S^2,F_1}(w) = \lim_{\varepsilon\to 0}\left\{\frac{S^2(F_1^*) - S^2(F_1)}{\varepsilon}\right\} = \lambda(1-\lambda)\cdot$$

$$\cdot\{\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta - \lambda(\delta'\Sigma^{-1}x)^2 - 2\delta'\Sigma^{-1}x\}$$

Y si realizamos el cambio de variable $y = \delta'\Sigma^{-1}x$ nos resultaria

$$CI_{S^2,F_1}(w) = \lambda(1-\lambda)\{\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta - \lambda y^2 - 2y\}$$

2.3.3. Distribucion de $CI_{S^2F_1}$.

Calculamos ahora la distribución de la variable $CI_{S^2F_1}(w)$, bajo la hipotesis de que $W \in N_p(\mu_1, \Sigma_1)$. Esta distribución que será utilizada como veremos en el apartado 2.6 para la deteccion de outliers.

Si $W \in N_p(\mu_1, \Sigma_1)$

entonces

$$X \in N_p(0, \Sigma_1) \quad e \quad Y \; N_1(0, \delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta)$$

Y si llamamos Z a la variable Y tipificada se tiene

$$Z = \frac{Y}{(\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta)^{1/2}}$$

La funcion influencia en función de Z nos quedaria de la forma

$$CI_{S^2,F_1}(W) = \lambda(1-\lambda)\{\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta - \lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta Z^2 -$$

$$- 2(\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta)^{1/2}Z\} = \lambda(1-\lambda)\{\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta + \frac{1}{\lambda} -$$

$$- \lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta\left(Z + \frac{1}{\lambda(\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta)^{1/2}}\right)^2\}$$

y llamando

$$CI'_{S^2,F_1}(W) = \left\{Z + \frac{1}{\lambda(\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta)^{1/2}}\right\}^2$$

y dado que $Z \; N_1(0, 1)$ tendremos que

$$CI'_{S^2,F_1}(W) \in \chi^2\left(1, \frac{1}{t^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta}\right)$$

A continuacion damos un lema con la intención de aproximar la ley chi - cuadrado no centrada mediante una distribución gamma.

2.3.3.1. Lema . Sea X una variable aleatoria distribuida segun una ley chi - cuadrado no centrada con r grados de libertad y parametro de descentralización η . La distribucion de la variable cX donde c es una constante real cualquiera, se puede aproximar por una ley gamma

$$G\left(\frac{(r+\eta)^2}{2(r+2\eta)}\,;\,\frac{r+\eta}{2(r+2\eta)c}\right)$$

La demostración es trivial basta tener en cuenta la aproximación de PATNAIK (1949).

Por tanto la funcion influencia de S^2 nos quedaria de la forma

$$CI_{S^2;F_1}(w) = \lambda(1-\lambda)\left\{\lambda\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta + \frac{1}{\lambda}\right\} - CI'_{S^2F_1}(w)$$

donde $CI'_{S^2;F_1}(w)$ se distribuirá en virtud del lema amterior como una gamma

$$CI'_{S^2;F_1}(w) \in G\left(\frac{(\lambda^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta+1)^2}{2\lambda^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta(\lambda^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta+2)};\ \frac{\lambda^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta+1}{2\lambda^2(1-\lambda)\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta(\lambda^2\delta'\Sigma^{-1}\Sigma\Sigma^{-1}}\right.$$

2.4. FUNCION INFLUENCIA PARA LA MEDIA DISCRIMINANTE.

Obtenemos en este apartado la funcion influencia y su distribución, para la media discriminante de la primera población $b'\mu_1$ y para la de la segunda $b'\mu_2$.

2.4.1 Calculo de $b'\mu_1(F_1^*)$ y $b'\mu_2(F_2^*)$.

Teniendo en cuenta los resultados del apartado 2.2

$$b'\mu_1(F_1^*) = \delta'(F_1^*)\{\Sigma(F_1^*)\}^{-1}\mu_1(F_1^*) = (\delta-\varepsilon x)'\{(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1} -$$

$$-\varepsilon\lambda\Sigma^{-1}xx'\Sigma^{-1}\}(\mu_1+\varepsilon x) = \delta'(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1}\mu_1 -$$

$$-\varepsilon\lambda\,\delta'\Sigma^{-1}xx'\Sigma^{-1}\mu_1 - \varepsilon x'\{I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1\}\Sigma^{-1}\mu_1 +$$

$$+\varepsilon^2\lambda\,x'\Sigma^{-1}xx'\Sigma^{-1}\mu_1 + \varepsilon\,\delta'\{I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1\}\Sigma^{-1}x -$$

$$-\varepsilon^2\lambda\,\delta'\Sigma^{-1}xx'\Sigma^{-1}x - \varepsilon^2 x'\{I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1\}\Sigma^{-1}x +$$

$$+\varepsilon^3\lambda\,x'\Sigma^{-1}xx'\Sigma^{-1}x$$

$$b'\mu_2(F_1^*) = \delta'(F_1^*)\{\Sigma(F_1^*)\}^{-1}\mu_2 = (\delta+\varepsilon x)'\{(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1} -$$

$$- \varepsilon\lambda xx'\Sigma^{-1}\}\mu_2 = \delta'(I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1)\Sigma^{-1}\mu_2 - \varepsilon\lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}\mu_2 -$$

$$- \varepsilon x'\{I_p+\varepsilon\lambda\Sigma^{-1}\Sigma_1\}\Sigma^{-1}\mu_2 + \varepsilon^2\lambda x'\Sigma^{-1}xx'\Sigma^{-1}\mu_2 .$$

y dada la definición de funcion influencia podemos despreciar, los terminos en ε de orden mayor que uno. Luego las expresiones anteriores nos quedarian.

$$b'\mu_1(F_1^*) = \delta'\Sigma^{-1}\mu_1 + \varepsilon\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1 - \varepsilon\lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}\mu_1 -$$
$$- \varepsilon x'\Sigma^{-1}\mu_1 + \varepsilon\delta'\Sigma^{-1}x$$

$$b'\mu_2(F_1^*)' = \delta'\Sigma^{-1}\mu_2 + \varepsilon\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_2 - \varepsilon\lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}\mu_2 -$$
$$- \varepsilon x'\Sigma^{-1}\mu_2 .$$

2.4.2. Funcion influencia para $b\mu_1$ y $b\mu_2$.

Aplicando la definición dada de función influencia resulta

$$CI_{b'\mu_1,F_1}(w) = \lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1 - \lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}\mu_1 - x'\Sigma^{-1}\mu_1 + \delta'\Sigma^{-1}x$$

$$CI_{b'\mu_2,F_1}(w) = \lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_2 - \lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}\mu_2 - x'\Sigma^{-1}\mu_2 .$$

y notando

$$y = \delta'\Sigma^{-1}x \qquad \text{y} \qquad U_i = \mu_i'\Sigma^{-1}x \quad i = 1,2$$

se obtiene

y $$CI_{b'\mu_1,F_1}(w) = \lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1 - \lambda yU_1 - U_1 + y$$

$$CI_{b'\mu_2,F_1}(w) = \lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_2 - \lambda yU_2 - U_2 .$$

2.4.3. Distribución de $CI_{E/\mu_1,E}$ y $CI_{E/\mu_2,E}$.

En este apartado vamos a calcular la distribución de la función influencia de las medias discrminantes, distribuciones que seran necesarias para la deteccion de outliers.

Para ello necesitamos de los siguientes lemas:

2.4.3.1. Lema. Sea V un vector aleatorio bidimensional distribuido segun una $N_2(\mu, \Sigma)$ no singular. Entonces la funcion generatriz de momentos de la variable ξ obtenida al multiplicar las componentes del vector aleatorio V viene dada por.

$$\varphi_\xi(\theta) = \left(1 - \frac{2\theta\sigma_1\sigma_2(1+\rho)}{2}\right)^{-\frac{1}{2}} \exp\left\{\frac{\frac{1}{2}\frac{(1-\rho)(\mu_2\sigma_1+\mu_1\sigma_2)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}\theta\frac{\sigma_1\sigma_2(1+\rho)}{2}}{1-2\theta\frac{\sigma_1\sigma_2(1+\rho)}{2}}\right\} \times$$

$$\times\left(1 - 2(-\theta)\frac{\sigma_1\sigma_2(1-\rho)}{2}\right)^{-1/2} \exp\left\{\frac{\frac{1}{2}\frac{(1+\rho)(\mu_2\sigma_1-\mu_1\sigma_2)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}(-\theta)\frac{\sigma_1\sigma_2(1-\rho)}{2}}{1-2(-\theta)\frac{\sigma_1\sigma_2(1-\rho)}{2}}\right\}$$

donde ρ es el coeficiente de correlaciòn entre las componentes del vector .

En efecto :

$$V = \begin{pmatrix} X \\ Y \end{pmatrix} \in N_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}; \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

Calculemos la funcion generatriz de momentos de la variable $\xi = X.Y$.

$$\varphi_\xi(\theta) = E[e^{\theta\xi}] = E[e^{\theta XY}] = E_Y[E_X[e^{(\theta Y)X}/Y=y]]$$

$$E_X[e^{(\theta Y)X}/Y=y] = \exp\left\{\theta y\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right) + \frac{1}{2}\sigma_1^2(1-\rho^2)\theta^2y^2\right\}$$

Luego

$$\varphi_{z^2}(\theta) = E_y\left[\exp.\left\{\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)\right)\theta y + \frac{1}{2}\sigma_1^2(1-\rho^2)\theta^2 y^2\right\}\right]$$

ya que

$$y \in N_1(\mu_2, \sigma_2)$$

se tiene

$$\varphi_{z^2}(\theta) = \frac{\exp.\left\{-\frac{1}{2}\frac{\mu_2^2}{\sigma_2^2}\right\}\cdot\exp.\left\{\frac{1}{2}\frac{\left(\theta\mu_1 + \frac{\mu_2}{\sigma_2^2} - \rho\frac{\sigma_1}{\sigma_2}\mu_2\theta\right)^2}{\frac{1}{\sigma_2^2} - \sigma_1^2(1-\rho^2)\theta^2 - 2\rho\frac{\sigma_1}{\sigma_2}\theta}\right\}}{\sigma_2\sqrt{\frac{1}{\sigma_2^2} - 2\rho\frac{\sigma_1}{\sigma_2}\theta - \sigma_1^2(1-\rho^2)\theta^2}}$$

Realizando las siguientes descomposiciones.

$$1 - 2\rho\sigma_1\sigma_2\theta - \sigma_1^2\sigma_2^2(1-\rho^2)\theta^2 = \left(1 - \sigma_1\sigma_2\theta(1+\rho)\right)\left(1 + \sigma_1\sigma_2\theta(1-\rho)\right)$$

y

$$\frac{1}{2}\frac{\left(\theta\mu_1 + \frac{\mu_2}{\sigma_2^2} - \rho\frac{\sigma_1}{\sigma_2}\mu_2\theta\right)^2}{\frac{1}{\sigma_2^2} - \sigma_1^2(1-\rho^2)\theta^2 - 2\rho\frac{\sigma_1}{\sigma_2}\theta} - \frac{1}{2}\frac{\mu_2^2}{\sigma_2^2} = \frac{2\rho\sigma_1\sigma_2\mu_1\mu_2 - \mu_1^2\sigma_2^2 - \mu_2^2\sigma_1^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)} +$$

$$+\frac{(1-\rho)(\mu_1\sigma_2 + \mu_2\sigma_1)^2}{4\sigma_1^2\sigma_2^2(1-\rho^2)(1 - \theta\sigma_1\sigma_2(1+\rho))} + \frac{(1+\rho)(\mu_2\sigma_1 - \mu_1\sigma_2)^2}{4\sigma_1^2\sigma_2^2(1-\rho^2)(1 + \theta\sigma_1\sigma_2(1-\rho))}$$

La funcion generatriz de momentos que se obtiene es.

$$\varphi_{z^2}(\theta) = \left(1 - \frac{2\theta\sigma_1\sigma_2(1+\rho)}{2}\right)^{-1/2}\exp.\left\{\frac{\frac{1}{2}\frac{(1-\rho)(\mu_2\sigma_1 + \mu_1\sigma_2)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}\cdot\frac{\theta\sigma_1\sigma_2(1+\rho)}{2}}{1 - 2\theta\frac{\sigma_1\sigma_2(1+\rho)}{2}}\right\}\cdot$$

$$\times\left(1 - 2(-\theta)\frac{\sigma_1\sigma_2(1-\rho)}{2}\right)^{-1/2}\exp.\left\{\frac{\frac{1}{2}\frac{(1+\rho)(\mu_2\sigma_1 - \mu_1\sigma_2)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}(-\theta)\frac{\sigma_1\sigma_2(1-\rho)}{2}}{1 - 2(-\theta)\frac{\sigma_1\sigma_2(1-\rho)}{2}}\right\}$$

2.4.3.2. Lema. Sea X una variable aleatoria distribuida segun una chi cuadrado no centrada con un grado de libertad y parametro de descentralización $\frac{(1-\rho)(\mu_2\sigma_1 + \mu_1\sigma_2)^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)}$ y sea Y una variable aleatoria independiente de la anterior y distribuida segun una chi cuadrado no centrada con un grado de libertad y parametro de descentralización $\frac{(1+\rho)(\mu_2\sigma_1 - \mu_1\sigma_2)^2}{2\sigma_1^2\sigma_2^2(1-\rho^2)}$. Entonces la variable.

$$H = \frac{\sigma_1\sigma_2(1+\rho)}{2} X - \frac{\sigma_1\sigma_2(1-\rho)}{2} Y$$

tiene como funcion generatriz de momentos.

$$\varphi_H(\theta) = \left(1 - \frac{2\theta\sigma_1\sigma_2(1+\rho)}{2}\right)^{-1/2} \exp\left\{\frac{\frac{1}{2}\,\frac{(1-\rho)(\mu_2\sigma_1+\mu_1\sigma_2)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}\,\theta\cdot\frac{\sigma_1\sigma_2(1+\rho)}{2}}{1 - 2\theta\,\frac{\sigma_1\sigma_2(1+\rho)}{2}}\right\} \times$$

$$\times \left(1 - 2(-\theta)\,\frac{\sigma_1\sigma_2(1-\rho)}{2}\right) \exp\left\{\frac{\frac{1}{2}\,\frac{(1+\rho)(\mu_2\sigma_1-\mu_1\sigma_2)^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}\,(-\theta)\,\frac{\sigma_1\sigma_2(1-\rho)}{2}}{1 - 2(-\theta)\,\frac{\sigma_1\sigma_2(1-\rho)}{2}}\right\}$$

La demostracion es trivial, basta utilizar la funcion generatriz de momentos de una distribucion chi - cuadrado no centrada (PATEL 1976)

2.4.3.3. Lema. Sea U una variable aleatoria distribuida segun una $\chi^2(1,\eta_1)$ y sea V una variable aleatoria distribuida según una $\chi^2(1,\eta_2)$ independiente estocasticamente de U . Sean r,s dos constantes reales positivas arbitrarias. Con estas condiciones la funcion de densidad de la variable aleatoria T = rU - sV viene dada por:

$$h(t) = \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \frac{e^{-\eta_1/2}\left(\frac{\eta_1}{2}\right)^k e^{-\eta_2/2}\left(\frac{\eta_2}{2}\right)^j}{k!\quad j!} \int_0^{\infty} \frac{g_{\chi^2_{2k+1}}\left(\frac{t+x}{r}\right)}{r} \cdot \frac{g_{\chi^2_{2j+1}}\left(\frac{x}{s}\right)}{s}\,dx$$

donde g_{2i+1} es la funcion de densidad de una variable chi cuadrado centrada con $2i+1$ grados de libertad.

Ademas la funcion de densidad de T puede aproximarse por.

$$h(t) = \begin{cases} \left(\frac{s}{r+s}\right)^{1/2} e^{-\eta_2/2\left(1-\frac{s}{r+s}\right)} \frac{1}{s} f_{\chi^2(1,\eta_2)}\left(\frac{-t}{s}\right) & t \le 0 \\ \left(\frac{r}{r+s}\right)^{1/2} e^{-\eta_2/2\left(1-\frac{r}{r+s}\right)} \frac{1}{r} f_{\chi^2(1,\eta_1)}\left(\frac{t}{r}\right) & t \ge 0 \end{cases}$$

En efecto:

Para hallar la funcion de densidad de la variable aleatoria T , hemos de calcular la convolución de las variables rU y sV.

Por ser

$$U \in \chi^2(1, \eta_1)$$

esta puede expresarse como mixturas de variables distribuides segun chi cuadrados centradas (JOHNSON , 1970). Asi la funcion de densidad de U se puede expresar

$$f_U(\cdot) = \sum_{K=0}^{\infty} e^{-\eta_1/2} \frac{(\eta_1/2)^K}{K!} \cdot g_{\chi^2(2K+1)}(\cdot)$$

donde $g_{\chi^2(2k+1)}(\cdot)$ es la funcion de densidad de una variable aleatoria chi - cuadrado centrada con $2K+1$ grados de libertad.

Por tanto $Z = r.U$ tendrá como función de densidad

$$f_{r.U}(z) = \sum_{k=0}^{\infty} e^{-\eta/2} \frac{(\eta_1/2)^k}{k!} \frac{1}{r} g_{\chi^2(2k+1)}\left(\frac{z}{r}\right)$$

Analogamente se podrian expresar la funcion de densidad de V y sV. Luego la funcion de densidad de T vendrá dada por

$$h(t) = \int_0^{\infty} f_{rU}(t+x) f_{sV}(x)\, dx = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} \frac{e^{-\eta_1/2} (\eta_1/2)^k e^{-\eta_2/2} (\eta_2/2)^j}{k!\quad j!} \int_0^{\infty} \frac{g_{\chi^2_{2k+1}}\left(\frac{t+x}{r}\right) g_{\chi^2_{2j+1}}\left(\frac{x}{s}\right)}{r.s}$$

y la integral que aparece en esta ultima expresión representa la funciòn de densidad de la variable $T' = rU' - sV'$ donde $U' \in \chi^2_{(2k+1)}$ y $V' \in \chi^2_{(2j+1)}$ que viene expresada por

$$f_{2k+1,2j+1}(t) = \begin{cases} \dfrac{c(2k+1, 2j+1)}{\Gamma\left(\frac{2k+1}{2}\right)} t^{k+j} e^{-\frac{t}{2r}} U\left(\frac{2j+1}{2}, k+j, \frac{r+s}{2r.s} t\right) & t \geq 0 \\ \dfrac{c(2k+1, 2j+1)}{\Gamma\left(\frac{2j+1}{2}\right)} (-t)^{k+j} e^{\frac{t}{2s}} U\left(\frac{2k+1}{2}, k+j, \left[-\frac{r+s}{2r.s} t\right]\right) & t \leq 0 \end{cases}$$

donde $c^{-1}(2k+1, 2j+1) = 2^{k+j+1} r^{\frac{2k+1}{2}} s^{\frac{2j+1}{2}}$ y $U(a, b, z)$ es la conocida funcion de Kummer (ABRAMOWITZ , 1972).

Y utilizando ahora la aproximacion dada por PRESS (1966), resulta para la funcion $h(t)$ la expresion

$$h(t)=\begin{cases}\displaystyle\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\frac{e^{-\eta_1/2}(\eta_1/2)^k}{k!}\,\frac{e^{-\eta_2/2}(\eta_2/2)^j}{j!}\,\frac{c(2k+1,2j+1)}{\Gamma\left(\frac{2j+1}{2}\right)}(-t)^{k+j}e^{\frac{t}{2s}}\left(-\frac{r+s}{2rs}t\right)^{-(k+1/2)} & t<0\\[2ex] \displaystyle\sum_{k=0}^{\infty}\sum_{j=0}^{\infty}\frac{e^{-\eta_1/2}(\eta_1/2)^k}{k!}\,\frac{e^{-\eta_2/2}(\eta_2/2)^j}{j!}\,\frac{c(2k+1,2j+1)}{\Gamma\left(\frac{2k+1}{2}\right)}t^{k+j}e^{-\frac{t}{2r}}\left(\frac{r+s}{2rs}t\right)^{-(j+1/2)} & t\geq 0\end{cases}$$

y mediante sucesivos calculos se obtiene finalmente:

$$h(t)=\begin{cases}\left(\frac{s}{r+s}\right)^{1/2}e^{-\eta_1/2\left(1-\frac{s}{r+s}\right)}\,\frac{1}{s}\,f_{\chi^2(1,\eta_2)}\left(\frac{-t}{s}\right) & t<0\\[2ex] \left(\frac{r}{r+s}\right)^{1/2}e^{-\eta_2/2\left(1-\frac{r}{r+s}\right)}\,\frac{1}{r}\,f_{\chi^2(1,\eta_1)}\left(\frac{t}{r}\right) & t\geq 0\end{cases}$$

Veamos ahora la distribución de las funciones influencias para las medias discriminantes. En el epigrafe 2.4.2 se ha obtenido que

$$CI_{B\mu_1,F_1}(w)=\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1-(\lambda y+1)(U_1-\tfrac{1}{2}\lambda)-\tfrac{1}{2}\lambda$$

$$CI_{B\mu_2,F_1}(w)=\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_2-(\lambda y+1)U_2$$

Ahora bien bajo la hipotesis inicial de que w no es outlier, las variables y , U_1 y U_2 se distribuyen

$$y\in N_1(0,\ \delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta)$$

$$U_1\in N_1\left(0,\ \mu_1'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1\right)$$

$$U_2\in N_1\left(0,\ \mu_2'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_2\right)$$

Consideremos las funciones

$$CI'_{B\mu_1,F_1}(w)=\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1+\frac{1}{\lambda}-CI_{B\mu_1,F_1}(w)=$$
$$=(\lambda y+1)(U_1-\tfrac{1}{2}\lambda)$$

$$CI'_{b'\mu_2,F_1}(w) = \lambda\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_2 - CI_{b'\mu_2,F_1}(w) =$$

$$= (\lambda Y + 1)\,U_2$$

Ya que si toda combinación lineal de las componentes de un vector aleatorio es una variable aleatoria normal, dicho vector tambien se distribuye normalmente (DUMAS DE RAULY (1.966)), por la dfinicion de las variables

$$(\lambda Y + 1) \quad y \quad (U_1 - 1/\lambda)$$

y en virtud de este teorema de caracterizacion , se tiene

$$\begin{pmatrix} \lambda Y + 1 \\ U_1 - 1/\lambda \end{pmatrix} \in N_2\left(\begin{pmatrix} 1 \\ -1/\lambda \end{pmatrix}; \begin{pmatrix} \lambda^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta & \lambda\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1 \\ \lambda\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1 & \mu_1'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1 \end{pmatrix}\right)$$

luego se verifican las condiciones de los lemas anteriormente demostrados y por tanto la funcion de densidad de $CI'_{b'\mu_1,F_1}(w)$ será:

$$h(CI') = \begin{cases} \left(\frac{1-\rho}{2}\right)^{1/2} \exp\left\{-\frac{1}{2}\left[\frac{(\sigma_2 - \sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}\right]\frac{1+\rho}{2}\right\} \times \frac{2}{\sigma_1\sigma_2(1-\rho)} \cdot f_{\chi^2}\left(1, \frac{(\sigma_2+\sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}\right)\left(\frac{-2CI'_{b'\mu_1,F_1}}{\sigma_1\sigma_2(1-\rho)}\right) \\ \qquad CI'_{b'\mu_1,F_1}(w) < 0 \\ \left(\frac{1+\rho}{2}\right)^{1/2} \exp\left\{-\frac{1}{2}\left[\frac{(\sigma_1/\lambda+\sigma_2)^2}{2\sigma_1^2\sigma_2^2(1-\rho)}\right]\frac{1-\rho}{2}\right\} \times \frac{2}{\sigma_1\sigma_2(1+\rho)} f_{\chi^2}\left(1, \frac{(\sigma_2-\sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}\right)\left(\frac{2CI'_{b'\mu_1,F_1}}{\sigma_1\sigma_2(1+\rho)}\right) \\ \qquad CI'_{b'\mu_1,F_1}(w) \geq 0 \end{cases}$$

donde

$$\sigma_1 = \sqrt{\lambda^2\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta}$$

$$\sigma_2 = \sqrt{\mu_1'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1}$$

$$\rho = \frac{\lambda\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\mu_1}{\sigma_1\,\sigma_2}$$

Analogamente todas las combinaciones lineales de las variables

$$(tY + 1) \quad y \quad U_2$$

son normales , luego

$$\begin{pmatrix} tY+1 \\ U_2 \end{pmatrix} \in N_2\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix} ; \begin{pmatrix} \lambda^2 \delta' \Sigma^{-1} \Sigma_1 \Sigma^{-1} \delta & \lambda \delta' \Sigma^{-1} \Sigma_1 \Sigma^{-1} \mu_2 \\ \lambda \delta' \Sigma^{-1} \Sigma_1 \Sigma^{-1} \mu_2 & \mu_2' \Sigma^{-1} \Sigma_1 \Sigma^{-1} \mu_2 \end{pmatrix}\right)$$

y por los lemas anteriores, la funcion de densidad de $CI'_{b'\mu_2,F_1}(w)$ es

$$h(CI') = \begin{cases} \left(\frac{1-\rho}{2}\right)^{1/2} \exp\left\{-\frac{1}{2}\left[\frac{1}{2\sigma_1^2(1+\rho)}\right]\frac{1+\rho}{2}\right\} \frac{2}{\sigma_1\sigma_2(1-\rho)} f_{\chi^2}\left(1, \frac{1}{2\sigma_1^2(1-\rho)}\right)\left(\frac{-2\,CI'_{b'\mu_2,F_1}}{\sigma_1\sigma_2(1-\rho)}\right) \\ CI'_{b'\mu_2,F_1}(w) < 0 \\ \\ \left(\frac{1+\rho}{2}\right)^{1/2} \exp\left\{-\frac{1}{2}\left[\frac{1}{2\sigma_1^2(1-\rho)}\right]\frac{1-\rho}{2}\right\} \frac{2}{\sigma_1\sigma_2(1+\rho)} f_{\chi^2}\left(1, \frac{1}{2\sigma_1^2(1+\rho)}\right)\left(\frac{2\,CI'_{b'\mu_2,F_1}}{\sigma_1\sigma_2(1+\rho)}\right) \\ CI'_{b'\mu_2,F_1} \geq 0 \end{cases}$$

donde

$$\sigma_1 = \sqrt{\lambda^2 \delta' \Sigma^{-1} \Sigma_1 \Sigma^{-1} \delta}$$

$$\sigma_2 = \sqrt{\mu_2' \Sigma^{-1} \Sigma_1 \Sigma^{-1} \mu_2}$$

$$\rho = \frac{\lambda \delta' \Sigma^{-1} \Sigma_1 \Sigma^{-1} \mu_2}{\sigma_1 \sigma_2}$$

Y teniendo en cuenta el lema 2.3.3.1. la distribución de estas variables se pueden aproximar

$$CI'_{b'\mu_1,F_1}(w) \alpha G\left(\frac{\left(1 + \frac{(\sigma_1/\lambda + \sigma_2)^2}{2\sigma_1^2\sigma_2^2(1-\rho)}\right)^2}{2\left(1 + 2\frac{(\sigma_2 + \sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}\right)} ; \frac{1 + \frac{(\sigma_1/\lambda + \sigma_2)^2}{2\sigma_1^2\sigma_2^2(1-\rho)}}{2\left(1 + 2\frac{(\sigma_1/\lambda + \sigma_2)^2}{2\sigma_1^2\sigma_2^2(1-\rho)}\right)\frac{\sigma_1\sigma_2(1-\rho)}{2}}\right)$$

$$CI'_{b'\mu_1,F_1}(w) < 0$$

$$CI'_{b'\mu_1,F_1}(w) \alpha G\left(\frac{\left(1 + \frac{(\sigma_2 - \sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}\right)^2}{2\left(1 + 2\frac{(\sigma_2 - \sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}\right)} ; \frac{1 + \frac{(\sigma_2 - \sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}}{2\left(1 + 2\frac{(\sigma_2 - \sigma_1/\lambda)^2}{2\sigma_1^2\sigma_2^2(1+\rho)}\right)\frac{\sigma_1\sigma_2(1+\rho)}{2}}\right)$$

$$CI'_{b'\mu_1,F_1}(w) \geq 0$$

$$CI'_{b;\mu_2,F_3}(w) \propto G_1\left(\frac{\left(1+\frac{1}{2\sigma_1^2(1-\rho)}\right)^2}{2\left(1+2\frac{1}{2\sigma_1^2(1-\rho)}\right)} \; ; \; \frac{1+\frac{1}{2\sigma_1^2(1-\rho)}}{2\left(1+2\frac{1}{2\sigma_1^2(1-\rho)}\right)^{\frac{\sigma_1\sigma_2(1-\rho)}{2}}} \right)$$

$$CI'_{b;\mu_2,F_3}(w) < 0$$

$$CI'_{b;\mu_2,F_3}(w) \propto G_1\left(\frac{\left(1+\frac{1}{2\sigma_1^2(1+\rho)}\right)^2}{2\left(1+2\frac{1}{2\sigma_1^2(1+\rho)}\right)} \; ; \; \frac{1+\frac{1}{2\sigma_1^2(1-\rho)}}{2\left(1+2\frac{1}{2\sigma_1^2(1+\rho)}\right)^{\frac{\sigma_1\sigma_2(1+\rho)}{2}}} \right)$$

$$CI'_{b;\mu_2,F_3}(w) > 0$$

2.5. FUNCION INFLUENCIA PARA LOS COEFICIENTES DE LA FUNCION DISCRIMINANTE .

Al igual que en los apartados anteriores se estudia la funcion influencia para el producto $b'b$, o modulo del vector b y a continuación determinamos su distribución.

2.5.1. Calculo de $b'b(F_1^*)$.

Se tiene que:

$$b'b(F_1^*) = \delta'(F_1^*)\left\{\Sigma(F_1^*)\right\}^{-1}\left\{\Sigma(F_1^*)\right\}^{-1}\delta(F_1^*) = \delta'\Sigma^{-1}\Sigma^{-1}\delta +$$
$$+\varepsilon\lambda\,\delta'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta - \varepsilon t\,\delta'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta - \varepsilon\,\delta'\Sigma^{-1}\Sigma^{-1}x -$$
$$-\varepsilon^2\lambda\,\delta'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}x + \varepsilon^2\lambda\,\delta'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}x +$$
$$+\varepsilon\lambda\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}\delta + \varepsilon^2\lambda^2\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta -$$
$$-\varepsilon^2\lambda^2\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta - \varepsilon^2\lambda\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}x -$$
$$-\varepsilon^3\lambda^2\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}x + \varepsilon^3\lambda^2\,\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}x -$$
$$-\varepsilon\lambda\,\delta'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}\delta - \varepsilon^2\lambda^2\,\delta'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta +$$
$$+\varepsilon^2\lambda^2\,\delta'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta + \varepsilon^3 t\,\delta'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}x +$$

$$+\varepsilon^2\lambda^2\delta'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}x-\varepsilon^3\lambda^2\delta'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}x-$$

$$-\varepsilon x'\Sigma^{-1}\Sigma^{-1}\delta-\varepsilon^2\lambda x'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta+\varepsilon^2\lambda x'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta$$

$$+\varepsilon^2 x'\Sigma^{-1}\Sigma^{-1}x+\varepsilon^3\lambda x'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}x-\varepsilon^3\lambda x'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}x-$$

$$-\varepsilon^2\lambda x'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}\delta-\varepsilon^3\lambda^2 x'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta+$$

$$+\varepsilon^3\lambda^2 x'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta+\varepsilon^3\lambda x'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}x+$$

$$+\varepsilon^4\lambda^2 x'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}x-\varepsilon^4\lambda^2 x'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}xx'\cdot$$

$$\cdot\Sigma^{-1}x+\varepsilon^2\lambda x'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}\delta+\varepsilon^3\lambda^2 x'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}$$

$$\Sigma_1\Sigma^{-1}\delta-\varepsilon^3\lambda^2 x'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta-\varepsilon^3\lambda x'\Sigma^{-1}xx'$$

$$\Sigma^{-1}\Sigma^{-1}x-\varepsilon^4\lambda^2 x'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}x+$$

$$+\varepsilon^4\lambda^2 x'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}x$$

y despreciando las potencias en ε de orden mayor que uno, ya que estas no intervienen en la función influencia. Se obtiene

$$b'b(F_1^*)=\delta'\Sigma^{-1}\Sigma^{-1}\delta+\varepsilon\lambda\delta'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta-\varepsilon\lambda\delta'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta-$$

$$-\varepsilon\delta'\Sigma^{-1}\Sigma^{-1}x+\varepsilon\lambda\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\Sigma^{-1}\delta-\varepsilon\lambda\delta'\Sigma^{-1}xx'\Sigma^{-1}\Sigma^{-1}\delta-$$

$$-\varepsilon x'\Sigma^{-1}\Sigma^{-1}\delta$$

2.5.2.- Funcion influencia para $b'b$.

$$CI_{b'b,F_1}(w)=\lim_{\varepsilon\to 0}\left\{\frac{b'b(F_1^*)-b'b(F_1)}{\varepsilon}\right\}=2\left[\lambda\delta'\Sigma^{-1}\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta-\right.$$

$$\left.-\lambda\delta'\Sigma^{-1}\Sigma^{-1}xx'\Sigma^{-1}\delta-\delta'\Sigma^{-1}\Sigma^{-1}x\right]$$

Y si notamos

$$y=\delta'\Sigma^{-1}X=X'b \quad (1) \quad \text{y} \quad z=b'\Sigma^{-1}X \quad (2)$$

obtenemos que la funcion influencia se puede representar por

$$CI_{\hat{b}b_1F_1}(w) = 2[\lambda b'\Sigma^{-1}\Sigma_1 b - (\lambda y+1)\ell]$$

2.5.3. Distribución de $CI_{\hat{b}'\hat{b}_1F_1}(w)$.

Bajo la hipotesis de que w no es outlier, las variables y y ℓ definidas en (1) y (2) se distribuyen

$$y \in N_1(0, \delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta)$$
$$N_1(0, b'\Sigma^{-1}\Sigma_1\Sigma^{-1}b)$$

Consideremos la función

$$CI'_{\hat{b}b_1F_1}(w) = (\lambda y+1)\ell = \lambda b'\Sigma^{-1}\Sigma_1 b - \frac{CI_{\hat{b}b_1F_1}(w)}{2}$$

Ya que cualquier combinación lineal de las variables $(\lambda y+1)$ y ℓ es normal univariante, en virtud del teorema de caracterización de la ley normal se verifica que

$$\begin{pmatrix}\lambda y+1\\ \ell\end{pmatrix} \in N_2\left(\begin{pmatrix}1\\0\end{pmatrix}; \begin{pmatrix}\lambda^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta & \lambda b'\Sigma^{-1}\Sigma_1 b\\ \lambda b'\Sigma^{-1}\Sigma_1 b & b'\Sigma^{-1}\Sigma_1\Sigma^{-1}b\end{pmatrix}\right)$$

Estamos asi en las condiciones de los lemas 2.4.3.1 , 2.4.3.2 y 2.4.3.3 y por el lema 2.3.3.1 , se tendra que la variable se distribuye

$$CI'_{\hat{b}b_1F_1}(w) \propto G\left(\frac{\left(1+\frac{1}{2\sigma_1^2(1-\rho)}\right)^2}{2\left(1+2\frac{1}{2\sigma_1^2(1-\rho)}\right)} ; \frac{1+\frac{1}{2\sigma_1^2(1-\rho)}}{2\left(1+2\frac{1}{2\sigma_1^2(1-\rho)}\right)}\frac{\sigma_1\sigma_2(1-\rho)}{2}\right)$$

$$CI'_{\hat{b}b_1F_1}(w) < 0$$

$$CI'_{\hat{b}b_1F_1}(w) \propto G\left(\frac{\left(1+\frac{1}{2\sigma_1^2(1+\rho)}\right)^2}{2\left(1+2\frac{1}{2\sigma_1^2(1+\rho)}\right)} ; \frac{1+\frac{1}{2\sigma_1^2(1+\rho)}}{2\left(1+2\frac{1}{2\sigma_1^2(1+\rho)}\right)}\frac{\sigma_1\sigma_2(1+\rho)}{2}\right)$$

donde

$$\sigma_1 = \sqrt{\lambda^2\delta'\Sigma^{-1}\Sigma_1\Sigma^{-1}\delta}$$

$$\sigma_2 = \sqrt{b'\Sigma^{-1}\Sigma_1\Sigma^{-1}b}$$

$$\rho = \frac{\lambda b'\Sigma^{-1}\Sigma_1 b}{\sigma_1\ \sigma_2}$$

2.6. DETECCION DE OUTLIERS.

El estudio realizado anteriormente, está hecho suponiendo conocidos los parametros poblacionales, sin embargo en las aplicaciones prácticas no vamos a conocer dichos parametros, por lo que tendremos que recurrir a estimaciones de los mismos, mediante la muestra que obtengamos. Las estimaciones generalmente las haremos por el metodo de maxima verosimiltud, por aquello de que es el metodo de estimación con mejores propiedades asintóticas.

Por tanto si vamos a utilizar estimaciones de los parametros poblacionales, tendremos que calcular estimaciones de la función influencia y en virtud del teorema de Zehna (ROHATGI , 1.976), sabemos que la estimación de maxima verosimilitud de la función influencia , será la función influencia de los estimadores de maxima verosimilitud de los parametros poblacionales.

Con esto tendremos determinada la función influencia para los parametros muestrales. La distribución de la función influencia muestral se distribuirá asintóticamente como la de la función influencia poblacional, de la que conocemos su distribución y parametros.

Hemos determinado la distribución de la función influencia poblacional bajo la hipótesis de que la variable $W \in N_p(\mu_1, \Sigma_1)$ es decir bajo la hipotesis de que la variable no es outliers. Por esta razon tendremos que la distribución de la función influencia muestral estará calculada bajo la hipótesis de que las n observaciones muestrales obtenidas no son outliers, debiendo exigirse que el tamaño de la muestra sea suficientemente grande pues estamos trabajando con procedimientos asintóticos.

Señalemos que el calculo de la función influencia, se ha realizado suponiendo que existe algún punto W en el cual hay definida una distribución degenerada. Se han extraido n observaciones por lo que tendremos una muestra de tamaño n de la variable función influencia, obtenida al ir suponiendo cada una de las observaciones como aquel punto en el que existe dicha distribución degenerada. Luego si se cumple la hipótesis nula de que las n observaciones pertenecen a una $N_p(\mu_1, \Sigma_1)$ tendremos que dichas observaciones se ajustarán a la distribución muestral de la función influencia.

Como se ha comprobado en los apartados anteriores la distribución de las funciones influencia son de tipo gamma. Por lo que proponemos para cada función un método gráfico para la detección de outliers basados en las representaciones gráficas de distribuciomes gammas obtenidas mediante procedimientos computacionales por WILK, GNANADESIKAN y HUYETT (1.962)

CAPITULO TERCERO

DETECCION DE OUTLIERS MEDIANTE EL COCIENTE DE VEROSIMILITUDES

CONTENIDO.

3.1. INTRODUCCION

En el capitulo anterior hemos dado un metodo para la deteccion de outliers en poblaciones normales, cuando se supone que estas proceden de una población con función de distribución dada por $(1-\varepsilon)F(x) + \varepsilon\,\delta(x)$, y además para una técnica determinada como es el analisis discriminante con vectores medias y matrices de covarianza distintas, para funciones discriminantes lineales.

En este capitulo lo que estudiamos es un criterio para detectar outliers cuando se ha extraido una muestra procedente de una población $N_p(\mu, \Sigma)$ no singular.

Para ello se supone que dada una muestra de tamaño n siempre es posible afirmar que en dicha muestra existen k elementos con $k \geqslant \left[\frac{n}{2}\right]$, que no son outliers es decir que son observeciones de la población que estamos muestreando, pues en caso contrario la muestra no seria representativa.

El criterio que se propone para la deteccion de outliers es el siguiente: Mediante un muestreo aleatorio simple extraemos k observaciones de entre las n de la muestra y sobre estas observaciones hacemos la suposición de que todas son de la población $N_p(\mu, \Sigma)$ y con cada una de las n - k observaciones restantes, realizamos un contraste de hipotesis para comprobar si cada una de estas observaciones es o no outlier.

Este contraste obtenido por el procedimiento de razon de verosimilitud, se basa en un estadistico cuya distribución es una ley F de Snedecor.

Al final del capitulo se incluye un programa de ordenador, para la deteccion de outliers , mediante el metodo dado en este capitulo.

NOTA: Con $\left[\frac{n}{2}\right]$ indicamos parte entera de $\frac{n}{2}$.

3.2. CONTRASTE DE COCIENTE DE VEROSIMILITUDES.

Sea $X_1, X_2, \ldots, X_k, Y$ una muestra aleatoria simple procedente de una poblacion normal p - dimensional no singular. Supongamos que $X_1, X_2, \ldots, X_k$ es una muestra de tamaño k extraida de una población $N_p(\mu, \Sigma)$ no singular.

Se quiere contrastar la hipotesis de que la observacion Y se ha extraido de una población $N_p(\mu, \Sigma)$ no singular frente a la hipotesis alternativa de que Y pertenece a una población $N_p(\mu_1, a^2\Sigma)$ con a ($a \neq 1$) una constante real no nula y fijada de antemano.

Utilizaremos para contrastar esta hipotesis el procedimiento del cociente de verosimilitudes.

Bajo la hipotesis nula, la distribución conjunta de la muestra aleatoria de tamaño k+1, es.

$$f(x_1, x_2, \ldots, x_k, y) = \left[\prod_{i=1}^{k} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x_i-\mu)'\Sigma^{-1}(x_i-\mu)\right\}\right] \cdot (2\pi)^{-p/2} |\Sigma|^{-1/2} \cdot \exp\left\{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right\}$$

Donde $\{X_i\}_{i=1,2,\ldots,k}$, Y y μ son vectores cuyas componentes X_{ji}, y_j y μ_j con j / $1 \leq j \leq p$ toman sus valores en el intervalo $(-\infty, \infty)$ y por ultimo Σ es una matriz simétrica definida positiva, que representaremos por $\Sigma > 0$.

Esta distribución conjunta tambien se puede expresar de la forma (KSHIRSAGAR, 1.972).

$$f(x_1, x_2, \ldots, x_k, y) = (2\pi)^{-\frac{(k+1)p}{2}} |\Sigma|^{-\frac{k+1}{2}} \cdot \exp\left\{-\frac{1}{2} tr.\Sigma^{-1}\left[\{(X|Y) - \mu E_{1,k+1}\}\{(X|Y) - \mu E_{1,k+1}\}'\right]\right\} \quad (1)$$

donde

$$(X|Y) = \left(\begin{array}{cccc|c} X_{11} & X_{12} & \ldots & X_{1k} & y_1 \\ X_{21} & X_{22} & \ldots & X_{2k} & y_2 \\ \ldots & \ldots & \ldots & \ldots & \vdots \\ X_{p1} & X_{p2} & \ldots & X_{pk} & y_p \end{array}\right) \quad \text{y} \quad E_{1,k+1} = (1, 1, \underbrace{\ldots}_{k+1}, 1)$$

Y los parametros μ y Σ toman sus valores en el espacio paramétrico Ω ,

$$\Omega = \{\mu, \Sigma \mid -\infty < \mu < \infty, \Sigma > 0\}$$

siendo $p + \frac{p(p+1)}{2}$ el numero de parametros desconocidos y que estimaremos mediante el metodo de maxima verosimilitud.

De (1) se deduce que el logaritmo de la funcion de verosimilitud es

$$\ln L(\Omega) = \ln_{\Omega} f(x_1, x_2, \ldots, x_K, y) = -\frac{(k+1)p}{2} \ln 2\pi - \frac{k+1}{2} \ln |\Sigma| -$$

$$- \frac{1}{2} \operatorname{tr} \Sigma^{-1} S_X - \frac{k}{2} \operatorname{tr} \Sigma^{-1} (\bar{x} - \mu)(\bar{x} - \mu)' - \frac{1}{2} \operatorname{tr} \Sigma^{-1} (y - \mu)(y - \mu)' \quad (2)$$

donde

$$\bar{X} = \frac{1}{K} X E_{K,1}$$

y S_X es la matriz de suma de cuadrados y suma de productos de la matriz de datos X y viene dada por

$$S_X = X(I - \tfrac{1}{K} E_{KK}) X'$$

Las ecuaciones de verosimilitud vendrian dadas por.

$$\frac{\partial \ln L}{\partial \mu} = 0 \quad (3)$$

$$\frac{\partial \ln L}{\partial \sigma_{ij}} = 0 \,, \; i = 1, 2, \ldots, p \,; \; j \geq i \quad (4)$$

donde σ_{ij} es el elemento que está en la i- ésima fila y j - ésima columna de la matriz Σ .

Escribiendo $\hat{\mu}$ por μ y $\hat{\Sigma}$ por Σ en estas ecuaciones para distinguir los estimadores de los prametros originales, de (3) se obtiene

$$\frac{\partial \ln L}{\partial \mu} = K \hat{\Sigma}^{-1} (\bar{x} - \hat{\mu}) + \hat{\Sigma}^{-1} (y - \hat{\mu}) = 0 \quad (5)$$

y de (4) se obtiene

$$\frac{\partial \ln L}{\partial \sigma_{ij}} = -\frac{k+1}{2} (2 - \delta_{ij}) \hat{\sigma}^{ij} + \frac{1}{2} \operatorname{tr} \frac{\partial \hat{\Sigma}}{\partial \hat{\sigma}_{ij}} \hat{\Sigma}^{-1} S_X \hat{\Sigma}^{-1} +$$

$$+ \frac{n}{2} \operatorname{tr} \frac{\partial \hat{\Sigma}}{\partial \hat{\sigma}_{ij}} \hat{\Sigma}^{-1} (\bar{x} - \hat{\mu})(\bar{x} - \hat{\mu})' \hat{\Sigma}^{-1} + \frac{1}{2} \operatorname{tr} \frac{\partial \hat{\Sigma}}{\partial \hat{\sigma}_{ij}} (y - \hat{\mu})(y - \hat{\mu})' \hat{\Sigma}^{-1} =$$

$$= 0 \quad (6)$$

habiendose utilizado los siguientes resultados de analisis matricial (BELLMAN , 1.965)

$$\frac{\partial \log|\Sigma|}{\partial \sigma_{ij}} = (2-\delta_{ij})\,\sigma^{ij} \qquad i,j = 1,2,\ldots,p$$

donde σ^{ij} es el elemento que está en la i - ésima fila y j - ésima columna de Σ^{-1} y δ_{ij} es la delta de Kronecker. Y

$$\frac{\partial \Sigma^{-1}}{\partial \sigma_{ij}} = -\Sigma^{-1}\frac{\partial \Sigma}{\partial \sigma_{ij}}\Sigma^{-1}$$

Multiplicando a la izquierda por $\hat{\Sigma}$, de la expresión (5) se deduce.

$$\hat{\mu} = \frac{k\bar{x}+y}{k+1} \qquad (7)$$

que no es mas que el vector media muestral de las k + 1 observaciones.

Y despejando en (6) $\hat{\sigma}^{ij}$ se obtiene

$$\hat{\sigma}^{ij} = \frac{1}{(k+1)(2-\delta_{ij})}\left[\frac{1}{2}\,tr.\,\frac{\partial \hat{\Sigma}}{\partial \hat{\sigma}_{ij}}\,\hat{\Sigma}^{-1}\left[S_x + k(\bar{x}-\hat{\mu})(\bar{x}-\hat{\mu})' + (y-\hat{\mu})(y-\hat{\mu})'\right]\hat{\Sigma}^{-1}\right]$$

luego

$$\hat{\Sigma}^{-1} = \frac{1}{k+1}\,\hat{\Sigma}^{-1}\left[S_x + k(\bar{x}-\hat{\mu})(\bar{x}-\hat{\mu})' + (y-\hat{\mu})(y-\hat{\mu})'\right]\hat{\Sigma}^{-1}$$

y multiplicando a izquierda y derecha por $\hat{\Sigma}$,

$$\hat{\Sigma} = \frac{1}{k+1}\left[S_x + k(\bar{x}-\hat{\mu})(\bar{x}-\hat{\mu})' + (y-\hat{\mu})(y-\hat{\mu})'\right]$$

y sustituyendo $\hat{\mu}$ por su valor dado en (7) se obtiene finalmente

$$\hat{\Sigma} = \frac{1}{k+1}\,S_{(k+1)} \qquad (8)$$

donde

$$S_{(k+1)} = (X|y)\left(I - \frac{1}{k+1}E_{k+1,k+1}\right)(X|y)'$$

que es la matriz de suma de cuadrados y suma de productos de las k + 1 observaciones p - dimensionales $X_1, X_2, \ldots, X_k, y$.

Para comprobar que estos estimadores maximizan la función de verosimilitud, bastaria aplicar el procedimiento de WATSON (1.964).

Bajo la hipotesis alternativa la función de densidad conjunta de la muestra aleatoria viene dada por (GIRI , 1.977).

$$f(X_1, X_2, \ldots, X_k, Y) = \left[\prod_{i=1}^{k} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp.\left\{-\frac{1}{2}(X_i-\mu)'\Sigma^{-1}(X_i-\mu)\right\}\right] \times$$

$$\times (2\pi)^{-p/2} |a^2\Sigma|^{-1/2} \exp.\left\{-\frac{1}{2}(Y-\mu_1)' a^{-2}\Sigma^{-1}(Y-\mu_1)\right\} \qquad (9)$$

y los parametros μ , μ_1 y Σ toman sus valores en el espacio paramétrico Ω_1.

$$\Omega_1 = \left\{\mu, \mu_1, \Sigma \;\middle|\; -\infty < \mu < \infty , -\infty < \mu_1 < \infty , \Sigma > 0\right\}$$

siendo $2p + \frac{p(p+1)}{2}$ el numero de parametros desconocidos y que estimaremos mediante el método de maxima verosimilitud.

Tomando logaritmos en (9) se deduce.

$$\ln L(\Omega_1) = \ln_{\Omega_1} f(X_1, \ldots, X_k, Y) = -\frac{(k+1)p}{2} \ln 2\pi - \frac{k+1}{2} \ln|\Sigma| - \frac{1}{2} tr.\Sigma^{-1}S_x$$

$$- \frac{k}{2} tr.\Sigma^{-1}(\bar{X}-\mu)(\bar{X}-\mu)' - \frac{1}{2} \ln a^{2p} - \frac{1}{2} tr.a^{-2}\Sigma^{-1}(Y-\mu_1)(Y-\mu_1)' \qquad (10)$$

y las ecuaciones de verosimilitud vendrian dadas por el sistema

$$\frac{\partial \ln L}{\partial \mu} = k\hat{\Sigma}^{-1}(\bar{X}-\hat{\mu}) = 0 \qquad (11)$$

$$\frac{\partial \ln L}{\partial \mu_1} = a^{-2}\hat{\Sigma}^{-1}(Y-\mu_1) = 0 \qquad (12)$$

$$\frac{\partial \ln L}{\partial \sigma_{ij}} = -\frac{k+1}{2} \frac{\partial}{\partial \hat{\sigma}_{ij}} \ln|\hat{\Sigma}| - \frac{1}{2} tr.\frac{\partial \hat{\Sigma}^{-1}}{\partial \hat{\sigma}_{ij}} S_x - \frac{k}{2} tr.\frac{\partial \hat{\Sigma}^{-1}}{\partial \hat{\sigma}_{ij}} (\bar{X}-\hat{\mu})(\bar{X}-\hat{\mu})' -$$

$$- \frac{1}{2} tr.a^{-2}\frac{\partial \hat{\Sigma}^{-1}}{\partial \hat{\sigma}_{ij}} (Y-\hat{\mu}_1)(Y-\hat{\mu}_1)' = 0 \qquad (13)$$

De (11) multiplicando por $\hat{\Sigma}$ a la izquierda se tiene.

$$\hat{\mu} = \bar{X} \qquad (14)$$

De (12) multiplicando por $a^2\hat{\Sigma}$ a la izquierda se tiene

$$\hat{\mu}_1 = Y \qquad (15)$$

y sustituyendo en (13) obtenemos

$$-\frac{k+1}{2}\frac{\partial}{\partial \hat{\sigma}_{ij}} \ln|\hat{\Sigma}| - \frac{1}{2} tr.\frac{\partial \hat{\Sigma}^{-1}}{\partial \hat{\sigma}_{ij}} S_X = 0$$

y por ultimo

$$\hat{\Sigma} = \frac{1}{k+1} S_X \qquad (16)$$

Veamos ahora que estos estimadores maximizan la función de verosimilitud

$$\ln L(\hat{\Omega}_1) - \ln L(\Omega_1) = -\frac{k+1}{2}\ln\left|\frac{1}{k+1} S_X\right| - \frac{(k+1)p}{2} + \frac{k+1}{2}\ln|\Sigma| +$$
$$+ \frac{1}{2} tr.\Sigma^{-1}S_X + \frac{k}{2} tr.\Sigma^{-1}(\bar{x}-\mu)(\bar{x}-\mu)' + \frac{1}{2} tr.\, a^{-2}\Sigma^{-1}(y-\mu_1)(y-\mu_1)'$$

y como

$$tr.\,\Sigma^{-1}(\bar{x}-\mu)(\bar{x}-\mu)' = (\bar{x}-\mu)'\Sigma^{-1}(\bar{x}-\mu)$$

y

$$tr.\,\Sigma^{-1}(y-\mu_1)(y-\mu_1)' = (y-\mu_1)'\Sigma^{-1}(y-\mu_1)$$

son definidas positivas se tiene

$$\ln L(\hat{\Omega}_1) - \ln L(\Omega_1) \geq -\frac{k+1}{2}\ln\left|\frac{1}{k+1} S_X\right| - \frac{(k+1)p}{2} + \frac{k+1}{2}\ln|\Sigma| +$$
$$+ \frac{1}{2} tr.\Sigma^{-1}S_X = -\frac{k+1}{2}\ln\frac{\left|\frac{1}{k+1} S_X\right|}{|\Sigma|} - \frac{(k+1)p}{2} + \frac{1}{2} tr.\Sigma^{-1}S_X =$$
$$= -\frac{k+1}{2}\ln\left|\frac{1}{k+1} S_X \Sigma^{-1}\right| - \frac{(k+1)p}{2} + \frac{1}{2} tr.\Sigma^{-1}S_X \qquad (17)$$

Notemos por $\lambda_1, \lambda_2, \ldots\ldots, \lambda_p$ los autovalores de la matriz $\frac{1}{k+1} S_X \Sigma^{-1}$ entonces.

$$\left|\frac{1}{k+1} S_X \Sigma^{-1}\right| = \prod_{i=1}^{p} \lambda_i \quad \text{y} \quad \frac{1}{k+1} tr.\Sigma^{-1}S_X = \sum_{i=1}^{p} \lambda_i$$

y sustituyendo en (17)

$$\ln L(\hat{\Omega}_1) - \ln L(\Omega_1) \geq -\frac{k+1}{2}\sum_{i=1}^{p}\ln\lambda_i - \frac{(k+1)p}{2} + \frac{1}{2}(k+1)\sum_{i=1}^{p}\lambda_i =$$

$$= \frac{k+1}{2}\left[\sum_{i=1}^{p}\left(-\ln \lambda_i - 1 + \lambda_i\right)\right] \qquad (18)$$

ya que para cualquier t real no negativo se tiene

$$t - \ln t \geq 1$$

de (18) deducimos

$$\ln h(\hat{\Omega}_j) - \ln h(\Omega_j) \geq 0$$

luego

$$\ln h(\hat{\Omega}_j) \geq \ln h(\Omega_j)$$

y esto es cierto para cualquier punto $(\hat{\mu}, \hat{\mu}_j, \hat{\Sigma}) \in \Omega_j$.

Sustituyendo ahora los valores obtenidos para $\hat{\mu}$, $\hat{\Sigma}$, en (7) y (8) en la expresión (2) tendremos.

$$\max_{\Omega} \ln h(\Omega) = -\frac{(k+1)p}{2}\ln 2\pi - \frac{k+1}{2}\ln\left|\frac{1}{k+1} S_{(k+1)}\right| - \frac{1}{2} tr(k+1) S_{(k+1)}^{-1} S_{(k+1)}$$

$$= -\frac{(k+1)p}{2}\ln 2\pi - \frac{(k+1)}{2}\ln\left|\frac{1}{k+1} S_{(k+1)}\right| - \frac{(k+1)p}{2} \qquad (19)$$

Analogamente sustituyendo los valores obtenidos para $\hat{\mu}$, $\hat{\mu}_j$, y $\hat{\Sigma}$ en (14) , (15) , y (16) en la expresión (10), obtenemos.

$$\max_{\Omega_j} \ln h(\Omega_j) = -\frac{(k+1)p}{2}\ln 2\pi - \frac{k+1}{2}\ln\left|\frac{1}{k+1} S_x\right| - \frac{(k+1)p}{2} - \frac{1}{2}\ln a^{2p} \qquad (20)$$

El criterio de razón de verosimilitud para contrastar la hipotesis nula es.

$$\lambda = \frac{\max_{\Omega} h(\Omega)}{\max_{\Omega_j} h(\Omega_j)}$$

de (19) y (20) se deduce

$$\lambda = \frac{|S_x|^{\frac{k+1}{2}} a^{p/2}}{|S_{(k+1)}|^{\frac{k+1}{2}}}$$

por lo que

$$\lambda_j = \frac{|S_x|}{|S_{(k+1)}|}$$

Ahora bien

$$S_{(k+1)} = S_x + \frac{k}{k+1}(\bar{x}-y)(\bar{x}-y)'$$

por tanto

$$\lambda_1 = \frac{|S_x|}{\left|S_x + \frac{k}{k+1}(\bar{x}-y)(\bar{x}-y)'\right|}$$

y basandonos en la igualdad (AITKEN , 1.965)

$$|A + qq'| = |A|(1 + q'A^{-1}q)$$

$$\lambda_1 = \frac{1}{1 + \frac{k}{k+1}(y-\bar{x})'S_x^{-1}(y-\bar{x})}$$

Por lo que la región critica para éste contraste viene dada por:

$$\frac{1}{1 + \frac{k}{k+1}(y-\bar{x})'S_x^{-1}(y-\bar{x})} \leq \lambda_0$$

que tambien puede expresarse en la forma

$$\frac{1-\lambda_0}{\lambda_0} \leq \frac{k}{k+1}(y-\bar{x})'S_x^{-1}(y-\bar{x})$$

Calculemos ahora la distribución del estadistico

$$\frac{k}{k+1}(y-\bar{x})'S_x^{-1}(y-\bar{x})$$

bajo la hipotesis nula de que $X_1, X_2, \ldots, X_k, y$ es una muestra aleatoria simple de tamaño k+1 extraida de una población que se distribuye segun una ley $N_p(\mu, \Sigma)$ no singular.

Luego

$$y \in N_p(\mu, \Sigma) \quad , \quad X_i \in N_p(\mu, \Sigma) \quad \forall i \quad i = 1, 2, \ldots, k$$

y por tanto (ANDERSON , 1.958)

$$\bar{X} \in N_p\left(\mu, \frac{1}{k}\Sigma\right)$$

de donde

$$y - \bar{X} \in N_p\left(0, \frac{k+1}{k}\Sigma\right)$$

y

$$\sqrt{\frac{k}{k+1}}(y-\bar{x}) \in N_p(0, \Sigma)$$

Por otro lado para $K > p$ (KSHIRSAGAR , 1.972)

$$S_X \in W_p(k-1, \Sigma)$$

y como $(y - \bar{X})$ y S_X son independientes (WILKS , 1.962) , el estadistico

$$\frac{k}{k+1}(y - \bar{X})' S_X^{-1} (y - \bar{X}) \equiv \frac{T_p^2}{k-1}$$

donde T_p^2 es una variable que sigue una distribución de Hotelling centrada p - dimensional con k - 1 grados de libertad.

Por otro lado segun (KSHIRSAGAR , 1.972)

" Si $U \in N_p(0, \Sigma)$ y D es independiente de U y distribuida segun $W_p(f, \Sigma)$ la distribución de la variable

$$T_p^2 = f\, U' D^{-1} U'$$

es una distribución de Hotelling centrada p - dimensional con f grados de libertad y

$$\frac{f - p + 1}{p} \cdot \frac{T_p^2}{f}$$

se distribuye segun una ley F de Snedecor con p y f - p + 1 grados de libertad "

Por tanto el estadistico

$$\frac{K-p}{p} \frac{k}{k+1} (y - \bar{X})' S_X^{-1} (y - \bar{X})$$

se distribuye segun una F de Snedecor con p y k - p grados de libertad

Por lo para un nivel de significación α prefijado la región critica será:

$$\frac{K-p}{p} \cdot \frac{k}{K+1} (y - \bar{X})' S_X^{-1} (y - \bar{X}) \geq F_{1-\alpha, p, k-p}$$

siendo $F_{1-\alpha, p, k-p}$ el cuantil de orden $1 - \alpha$, es decir

$$P\left[F_{p, k-p} \geq F_{1-\alpha, p, k-p} \right] = \alpha$$

3.3. DETECCION DE OUTLIERS.

Consideremos una población que se distribuye segun una ley $N_p(\mu, \Sigma)$ no singular de la que hemos extraido una muestra de tamaño n. Se trata de detectar cuáles de estas observaciones son outliers.

Como ya vimos en la introducción, dada una muestra de tamaño n siempre es posible afirmar que en dicha muestra existen k elementos con $k > \left[\frac{n}{2}\right]$ que no son outliers, es decir son observaciones de la población $N_p(\mu, \Sigma)$ que estamos muestreando.

El metodo que proponemos para la detección de outliers es el siguiente:

Mediante un muestreo aleatorio simple, extraemos k observaciones de entre las n de la muestra. Sobre estas k observaciones hacemos la hipótesis de que todas ellas pertenecen a la población $N_p(\mu, \Sigma)$ no singular. A continuación calculamos el vector media de estas k observaciones $\bar{X}$ y la matriz de suma de cuadrados y suma de productos S_X de estas observaciones e impondremos a k la condición $k > p$ para que la distribución de S_X sea no degenerada.

Para cada una de las n - k observaciones restantes y, calculamos el valor del estadistico

$$\frac{k-p}{p} \cdot \frac{k}{k+1} (y - \bar{X})' S_X^{-1} (y - \bar{X})$$

Fijado un nivel de significación α, se busca el valor del cuantil $F_{1-\alpha, p, k-p}$ tal que

$$P\left[\mathcal{F} \geq F_{1-\alpha, p, k-p}\right] = \alpha$$

donde $\mathcal{F}$ se distribuye segun una ley F de Snedecor con p y k - p grados de libertad.

Para cada una de las n - k observaciones y la regla de decisión

a seguir sería.

Si

$$F_{1-\alpha,p,k-p} < \frac{k-p}{p} \cdot \frac{k}{k+1} (y-\bar{X}) S_x^{-1} (y-\bar{X})$$

se rechaza la hipotesis nula de que $y \in N_p(\mu, \Sigma)$ aceptandose por tanto que dicha observación es outlier.

Midamos ahora el riesgo de error que cometemos, al suponer que las k observaciones obtenidas mediante el muestreo aleatorio simple, $X_1, X_2, \ldots, X_k$ no son outliers.

Sea Z la variable aleatoria que representa el numero de outliers en una muestra de tamaño n, Z podrá tomar los valores $0, 1, \ldots, [\frac{n}{2}]$

Si notamos por θ la probabilidad de ser outlier

$$P[Z=r] = \frac{\binom{n}{r}\theta^r(1-\theta)^{n-r}}{\sum_{j=0}^{[n/2]}\binom{n}{j}\theta^j(1-\theta)^{n-j}}$$

Notemos por A el suceso de que ninguna de las k observaciones extraidas de la muestra es outlier. Entonces

$$P[A] = \sum_{r=0}^{[n/2]} P[A/Z=r] \cdot P[Z=r] =$$

$$= \sum_{r=0}^{[\frac{n}{2}]} \frac{\binom{r}{0}\binom{n-r}{k}}{\binom{n}{k}} \times \frac{\binom{n}{r}\theta^r(1-\theta)^{n-r}}{\sum_{j=0}^{[n/2]}\binom{n}{j}\theta^j(1-\theta)^{n-j}} =$$

$$= \frac{1}{\sum_{j=0}^{[n/2]}\binom{n}{j}\theta^j(1-\theta)^{n-j}} \sum_{r=0}^{[n/2]} \binom{n-k}{r}\theta^r(1-\theta)^{n-r} =$$

$$= (1-\alpha)^k \frac{\sum_{r=0}^{[n/2]}\binom{n-k}{r}\theta^r(1-\theta)^{n-k-r}}{\sum_{j=0}^{[n/2]}\binom{n}{j}\theta^j(1-\theta)^{n-j}}$$

Solo quedaría dar una estimación de la probabilidad θ de ser outlier, para ello pueden seguirse los siguientes criterios.

1º .- Considerar la probabilidad θ de ser outlier como la probabilidad de que el estadistico caiga en la región critica. Es decir tomar θ como el nivel de significación del contraste

2º .- Estimar θ mediante la proporción entre el numero de observaciones que detectamos como outliers y el numero de total de observaciones.

3º .- La estimación de θ dada por SHAH (1.966)

$$\hat{\theta} = \frac{a_3 - \left(\left[\frac{n}{2}\right] + 1\right) a_2}{(n-2)a_2 + \frac{1}{2}(n-1)\left(n - \left[\frac{n}{2}\right]\right) - n(n-2)a_1 - n\left[\frac{n}{2}\right]}$$

donde

$$a_i = \frac{\sum_{j=1}^{n} r_j^i}{n}$$

4º .- Elegir como probabilidad de ser outlier la que se desprende de los trabajos de DEVLIN y otros (1.975) , GNANADESIKAN (1.977) , BARNETT (1.978) etc. , que consideran esta probabilidad del orden de 1/n , siendo n el tamaño de la muestra.

3.4. APLICACION.

Aplicamos este metodo al siguiente ejemplo dado por Barnett (1.978) donde se representan las edades y salarios estos ultimos expresados en libras esterlinas , de 55 de los 374 miembros de la sección de estudiantes del Instituto de Ingenieria Electrica.

27.67	2930	25.33	2300
23.42	2330	25.25	2200
24.67	2480	27.92	3500
27.92	4100	29.50	3600
26.92	2500	21.17	1470
28.92	3380	25.67	2690
26.08	2720	28.42	2860
29.92	4930	26.00	3000
29.50	3020	27.42	3100
22.42	1970	26.58	2600
23.42	1700	25.50	2250
28.00	3100	29.25	3600
23.00	1950	26.00	2750
25.17	2320	29.33	3500
26.58	2750	26.25	3400
22.75	1960	26.83	4500
25.00	2300	27.92	2800
30.00	4120	27.08	3610
23.50	3900	28.33	3100
26.58	5200	28.33	2900
28.25	3200	30.00	3600

28.25	3600	29.92	3610
24.67	2030	30.58	4200
25.42	3520	23.83	3050
22.67	1900	26.33	2760
25.92	3230	26.83	4000
25.25	2500	28.25	3100
25.58	3020		

Para un nivel de significación α = 0.05 y para k = 10 , mediante el siguiente programa de ordenador obtenemos los elementos OUTLIERS.

```
      PARAMETER N=5,K=10,JI=5,M=2,NREA=4
      EXTERNAL A1
C     ------------------------------------------------------------------
C NEN. DE DATOS, KEN. DE EXCLUSIONES, JI= No. INICIAL DE ALEATORIZACION
C     NE DIMENSION
C     ------------------------------------------------------------------
      DIMENSION DATOS(M,N),EXCL (M,K),IEXCL(K),C(K,K),PR(M,K),S(M,M),
     * Y(M,NK),XSEK(M),YK(M,NK),P(M),SINV(M,M),V(61,171)
C     ------------------------------------------------------------------
C LECTURA DE UN DATO M-DIMENSIONAL EN CADA TARJETA E IMPRESION
C     ------------------------------------------------------------------
      READ1,((DATOS(I,J),I=1,M),J=1,N)
    1 FORMAT(2F30.2)
      PRINT1,((DATOS(I,J),I=1,M),J=1,N)
      PRINT 23
   23 FORMAT(/,/)
C     ------------------------------------------------------------------
C ELECCION DE K DATOS ALEATORIOS DE LA MUESTRA SIN REEMPLAZAMIENTO E IMPRESION
C     ------------------------------------------------------------------
      JJ=JI
      DO 2 I=1,K
   32 IEXCL(I)=ENRAND(JJ)
      IEXCL(I)=IEXCL(I)-N*(IEXCL(I)/N)
      IF(I.EQ.1) GOTO 2
      I1=I-1
      DO 22 J=1,I1
      IF(IEXCL(I).EQ.IEXCL(J)) GOTO 32
   22 CONTINUE
    2 CONTINUE
      PRINT 3,IEXCL
    3 FORMAT(' INDICES ALEATORIOS'7X,10(1X,I9),/)
      DO 4 I=1,K
      L=IEXCL(I)
      DO 4 J=1,M
    4 EXCL(J,I)=DATOS(J,L)
      PRINT1,EXCL
      PRINT 23
C     ------------------------------------------------------------------
C CALCULO DE LA MATRIZ DE SUMA DE CUADRADOS Y PRODUCTOS DE K OBSERVACIONES SK
C (EN EL PROGRAMA, LA MATRIZ SK ES S)
C     ------------------------------------------------------------------
      II=0
      JJ=0
      DO 6 I=1,K
      DO 5 J=1,K
    5 C(I,J)=-1./K
      DO 6 I=1,K
    6 C(I,I)=C(I,I)+1
      DO 7 I=1,M
      DO 7 J=1,K
      PR(I,J)=0
      DO 7 L=1,K
    7 PR(I,J)=PR(I,J)+ EXCL(I,L)*C(L,J)
      DO 8 I=1,M
      DO 8 J=1,M
      S(I,J)=0
      DO 8 L=1,K
    8 S(I,J)=S(I,J)+PR(I,L)* EXCL(J,L)
C     ------------------------------------------------------------------
C IMPRESION DE LA MATRIZ SK
C     ------------------------------------------------------------------
      PRINT 301
  301 FORMAT(' MATRIZ SK',/)
      PRINT9,S
    9 FORMAT(1X,E18.9)
      PRINT 23
C     ------------------------------------------------------------------
C CALCULO DE LA MATRIZ SK(-1), INVERSA DE SK, E IMPRESION
C (EN EL PROGRAMA, LA MATRIZ SK(-1) ES SINV)
C     ------------------------------------------------------------------
      PRINT 302
  302 FORMAT( ' MATRIZ SK(-1)',/)
      DET=S(1,1)*S(2,2)-S(1,2)*S(2,1)
      IF((S(1,2)).EQ.(S(2,1))) GOTO 200
      PRINT201, S(1,2),S(2,1)
  201 FORMAT( ' S12='E18.9,'S21='E18.9,/)
  200 IF(DET.EQ.0) STOP1
      SINV(1,1)=S(2,2)/DET
      SINV(2,2)=S(1,1)/DET
      SINV(1,2)=-S(1,2)/DET
      SINV(2,1)=-S(2,1)/DET
      PRINT9,SINV
```

```
      PRINT 22
C     ------------------------------------------------------------------
C FORMACION DEL VECTOR DE DATOS NO ELEGIDOS E IMPRESION
C (EN EL PROGRAMA,EL VECTOR DE DATOS NO ELEGIDOS ES Y)
C     ------------------------------------------------------------------
      IO=0
      DO 10 I=1,N
      A=1
      DO 11 J=1,K
   11 A=A*(I-IEXCL(J))
      IF(A.EQ.0) GOTO 12
      DO 13 L=1,M
   13 Y(L,I-IO)=DATOS(L,I)
      GOTO 10
   12 IO=IO+1
   10 CONTINUE
      PRINT 303
  303 FORMAT( ' N-K DATOS NO ELEGIDOS',/)
      PRINT210,(I,(Y(J,I),J=1,M),I=1,NK)
  310 FORMAT(I5,2F20.0)
      PRINT 23
C     ------------------------------------------------------------------
C CALCULO DEL ESTADISTICO
C     ------------------------------------------------------------------
C     ------------------------------------------------------------------
C CALCULO DEL VECTOR MEDIA E IMPRESION
C (EN EL PROGRAMA, EL VECTOR MEDIA ES  XMEDK)
C     ------------------------------------------------------------------
      DO 14 I=1,M
   14 XMEDK(I)=0
      DO 15 I=1,M
      DO 15 J=1,K
   15 XMEDK(I)=XMEDK(I)+EXCL(I,J)
      DO 16 I=1,M
   16 XMEDK(I)=XMEDK(I)/K
      PRINT 210
  210 FORMAT(1X,'VECTOR MEDIA',/)
      PRINT1,XMEDK
      PRINT 23
C     ------------------------------------------------------------------
C CALCULO DEL VECTOR DIFERENCIA DEL VECTOR DE DATOS NO ELEGIDOS Y EL VECTOR
C MEDIA, E IMPRESION)
C (EN EL PROGRAMA, ES YK)
C     ------------------------------------------------------------------
      DO 17 I=1,M
      DO 17 J=1,NK
   17 YK(I,J)=Y(I,J)-XMEDK(I)
      PRINT 211
  211 FORMAT(1X,'VECTOR DIFERENCIA DEL VECTOR DE DATOS NO ELEGIDOS Y EL
     * VECTOR MEDIA',/)
      PRINT1,YK
      PRINT 23
C     ------------------------------------------------------------------
C CALCULO DEL F EXPERIMENTAL PARA CADA DATO NO ELEGIDO, E IMPRESION
C (EN EL PROGRAMA ES FEXP)
C     ------------------------------------------------------------------
      PRINT 212
  212 FORMAT(1X,'CALCULO DEL ESTADISTICO PARA LOS N-K DATOS NO ELEGIDOS'
     *,/)
      DO 18 J=1,NK
      DO 19 I=1,M
      P(I)=0
      DO 19 L=1,M
```

```
      10 P(I)=P(I)+YK(L,J)*SINV(L,I)
         FEXP=E
         DO 20  I=1,M
      20 FEXP=FEXP+P(I)*YK(I,J)
         FEXP=FEXP*(K-M)*K/(M*(K+1))
         IF(FEXP.GT.FTEST) GOTO 500
         V1='       '
         V2='       '
         GOTO 501
  500 V1='OUTLIE'
         V2='R      '
  501 PRINT 21,J,FEXP,V1,V2
   21 FORMAT(1X,'DATO NO ELEGIDO NUMERO =',I4,9X,'F.EXPERIMENTAL=',F10.6,
     *9X,2A8)
         PRINT 23
   23 CONTINUE
C     ------------------------------------------------------------------
C GRAFICA DE LOS DATOS
C (EN EL PROGRAMA, LOS DATOS NO ELEGIDOS VIENEN REPRESENTADOS POR '*', Y LOS
C  DATOS ELEGIDOS, POR '0')
C     ------------------------------------------------------------------
         DO 503 I=1,61
         DO 503 J=1,121
  503 V(I,J)=' '
         DO 505 I=1,121
  505 V(1,I)='-'
         DO 506 I=1,61
  506 V(I,1)='I'
         DO 502 I=1,NK
         II=(60*(Y(1,I)-20))/11+1
         JJ=(120*(Y(2,I)-1400))/3800+1
  502 V(II,JJ)='*'
         DO 504  I=1,K
         II=(60*(EXCL(1,I)-20))/11+1
         JJ=(120*(EXCL(2,I)-1400))/3800+1
  504 V(II,JJ)='0'
         PRINT 507
  507 FORMAT(1H1)
         DO 508 I=1,61
         PRINT 508,(V(I,J),J=1,121)
  508 FORMAT(1X,121A1)
  503 CONTINUE
         STOP
         END

NO OF COMPILATION:        NO  DIAGNOSTICS.
```

```
27.67          2370.00
23.42          2370.00
24.67          2460.00
27.92          4100.00
28.92          2600.00
28.92          3380.00
28.08          2720.00
29.92          4930.00
29.50          3020.00
22.42          1370.00
23.42          1700.00
28.00          3100.00
23.00          1950.00
25.17          2320.00
26.58          2750.00
22.75          1960.00
25.00          2300.00
30.00          4120.00
23.50          2900.00
26.58          5200.00
```

28.25	3200.00
28.33	2300.00
25.25	2200.00
27.92	3500.00
29.50	3600.00
31.17	1470.00
25.67	2050.00
28.42	2860.00
26.00	3000.00
27.42	3300.00
26.58	2600.00
25.50	2250.00
29.25	3600.00
28.00	2750.00
29.33	3500.00
26.25	3400.00
26.83	4500.00
27.92	2800.00
27.08	3510.00
28.33	3100.00
28.33	2900.00
30.00	3500.00
28.25	3600.00
24.67	2030.00
25.42	3520.00
22.67	1300.00
25.92	2230.00
25.25	2500.00
25.58	3020.00
29.92	3510.00
30.58	4200.00
23.83	3050.00
26.33	2760.00
26.83	4000.00
28.25	3100.00

INDICES ALEATORIOS

28 53 5 37 41 33 25 1 45 14

28.42	2860.00
26.33	2760.00
25.92	2500.00
26.83	4000.00
28.33	2900.00
29.25	3500.00
29.58	3500.00
27.67	2030.00
25.42	3520.00
25.17	2320.00

MATRIZ SK

.7040181E+02	.2267727E+04
.7267759E+04	.3822893E+07

MATRIZ SK(-1)

SIDE	.3267759E+04 .2267727E+04

-.2479363 9-01	-.3117862 54-04
-.7112266 48 5-04	.2980477 08-06

N-K DATOS NO ELEGIDOS

1	23.	2330.
2	29.	2480.
3	28.	4100.
4	29.	3380.
5	28.	2720.
6	30.	4970.
7	29.	3020.
8	22.	1970.
9	23.	1700.
10	26.	3100.
11	23.	1950.
12	27.	2750.
13	23.	1960.
14	29.	2300.
15	30.	4120.
16	23.	3300.
17	27.	5200.
18	28.	3200.
19	25.	2300.
20	29.	2200.
21	28.	3500.
22	21.	1470.
23	26.	2600.
24	29.	3000.
25	27.	3100.
26	27.	2600.
27	25.	2250.
28	26.	2750.
29	29.	3500.
30	26.	3400.
31	28.	2600.
32	27.	3610.
33	28.	3100.
34	30.	3600.
35	28.	3600.
36	25.	2030.
37	23.	1900.
38	26.	3230.
39	29.	2500.
40	26.	3020.
41	30.	3610.
42	31.	4200.
43	24.	3050.
44	27.	4000.
45	28.	3100.

VECTOR MEDIA

27.38	3149.00

VECTOR DIFERENCIA DEL VECTOR DE DATOS NO ELEGIDOS Y EL VECTOR MEDIA

-3.06	-619.00
-2.71	-599.00
.54	951.00
1.54	2?1.00
-1.30	-429.00
2.54	1781.00
2.12	-129.00
-4.96	-1179.00
-3.96	-1949.00
.63	-49.00
-4.38	-1109.00
-.80	-799.00
-4.63	-1159.00
-2.38	-849.00
2.62	971.00
-3.88	751.00
-.80	2051.00
.87	51.00
-2.05	-899.00
-2.13	-949.00
.54	351.00
-6.21	-1679.00
-1.71	-499.00
-1.38	-149.00
.04	-49.00
-.80	-549.00
-1.88	-899.00
-1.38	-399.00
1.95	351.00
-1.13	251.00
.54	-249.00
-.30	451.00
.95	-49.00
2.62	451.00
.87	451.00
-2.71	-1119.00
-4.71	-1249.00
-1.46	51.00
-2.13	-549.00
-1.80	-129.00
2.54	451.00
3.20	1051.00
-3.55	-99.00
-.55	851.00
.87	-49.00

CALCULO DEL ESTADISTICO PARA LOS N-K DATOS NO ELEGIDOS

DATO NO ELEGIDO NUMERO = 1 F.EXPERIMENTAL= 2.346495

DATO NO ELEGIDO NUMERO = 2 F.EXPERIMENTAL= 1.470264

DATO NO ELEGIDO NUMERO = 3 F.EXPERIMENTAL= .870424

DATO NO ELEGIDO NUMERO = 4	F.EXPERIMENTAL= .424213	
DATO NO ELEGIDO NUMERO = 5	F.EXPERIMENTAL= .386245	
DATO NO ELEGIDO NUMERO = 6	F.EXPERIMENTAL= 3.474880	
DATO NO ELEGIDO NUMERO = 7	F.EXPERIMENTAL= .973179	
DATO NO ELEGIDO NUMERO = 8	F.EXPERIMENTAL= 4.792631	OUTLIER
DATO NO ELEGIDO NUMERO = 9	F.EXPERIMENTAL= 3.836191	
DATO NO ELEGIDO NUMERO = 10	F.EXPERIMENTAL= .081686	
DATO NO ELEGIDO NUMERO = 11	F.EXPERIMENTAL= 3.941431	
DATO NO ELEGIDO NUMERO = 12	F.EXPERIMENTAL= .212847	
DATO NO ELEGIDO NUMERO = 13	F.EXPERIMENTAL= 4.2[illegible]946	
DATO NO ELEGIDO NUMERO = 14	F.EXPERIMENTAL= 1.350327	
DATO NO ELEGIDO NUMERO = 15	F.EXPERIMENTAL= 1.690949	
DATO NO ELEGIDO NUMERO = 16	F.EXPERIMENTAL= 4.117308	
DATO NO ELEGIDO NUMERO = 17	F.EXPERIMENTAL= 4.780483	OUTLIER
DATO NO ELEGIDO NUMERO = 18	F.EXPERIMENTAL= .135750	
DATO NO ELEGIDO NUMERO = 19	F.EXPERIMENTAL= 1.144202	

DATO NO ELEGIDO NUMERO = 20 F.EXPERIMENTAL= 1.327676

DATO NO ELEGIDO NUMERO = 21 F.EXPERIMENTAL= .137692

DATO NO ELEGIDO NUMERO = 22 F.EXPERIMENTAL= 7.877045 OUTLIER

DATO NO ELEGIDO NUMERO = 23 F.EXPERIMENTAL= .537029

DATO NO ELEGIDO NUMERO = 24 F.EXPERIMENTAL= .341429

DATO NO ELEGIDO NUMERO = 25 F.EXPERIMENTAL= .003032

DATO NO ELEGIDO NUMERO = 26 F.EXPERIMENTAL= .370384

DATO NO ELEGIDO NUMERO = 27 F.EXPERIMENTAL= 1.116899

DATO NO ELEGIDO NUMERO = 28 F.EXPERIMENTAL= .402613

DATO NO ELEGIDO NUMERO = 29 F.EXPERIMENTAL= .597444

DATO NO ELEGIDO NUMERO = 30 F.EXPERIMENTAL= .377331

DATO NO ELEGIDO NUMERO = 31 F.EXPERIMENTAL= .221208

DATO NO ELEGIDO NUMERO = 32 F.EXPERIMENTAL= .255784

DATO NO ELEGIDO NUMERO = 33 F.EXPERIMENTAL= .183707

DATO NO ELEGIDO NUMERO = 34 F.EXPERIMENTAL= 1.245900

DATO NO ELEGIDO NUMERO = 35 F.EXPERIMENTAL= .251620

DATO NO ELEGIDO NUMERO = : 36 F.EXPERIMENTAL= 1.997141

DATO NO ELEGIDO NUMERO = : 37 F.EXPERIMENTAL= 4.495047 OUTLIER

DATO NO ELEGIDO NUMERO = : 38 F.EXPERIMENTAL= .442510

DATO NO ELEGIDO NUMERO = : 39 F.EXPERIMENTAL= .904372

DATO NO ELEGIDO NUMERO = : 40 F.EXPERIMENTAL= .595268

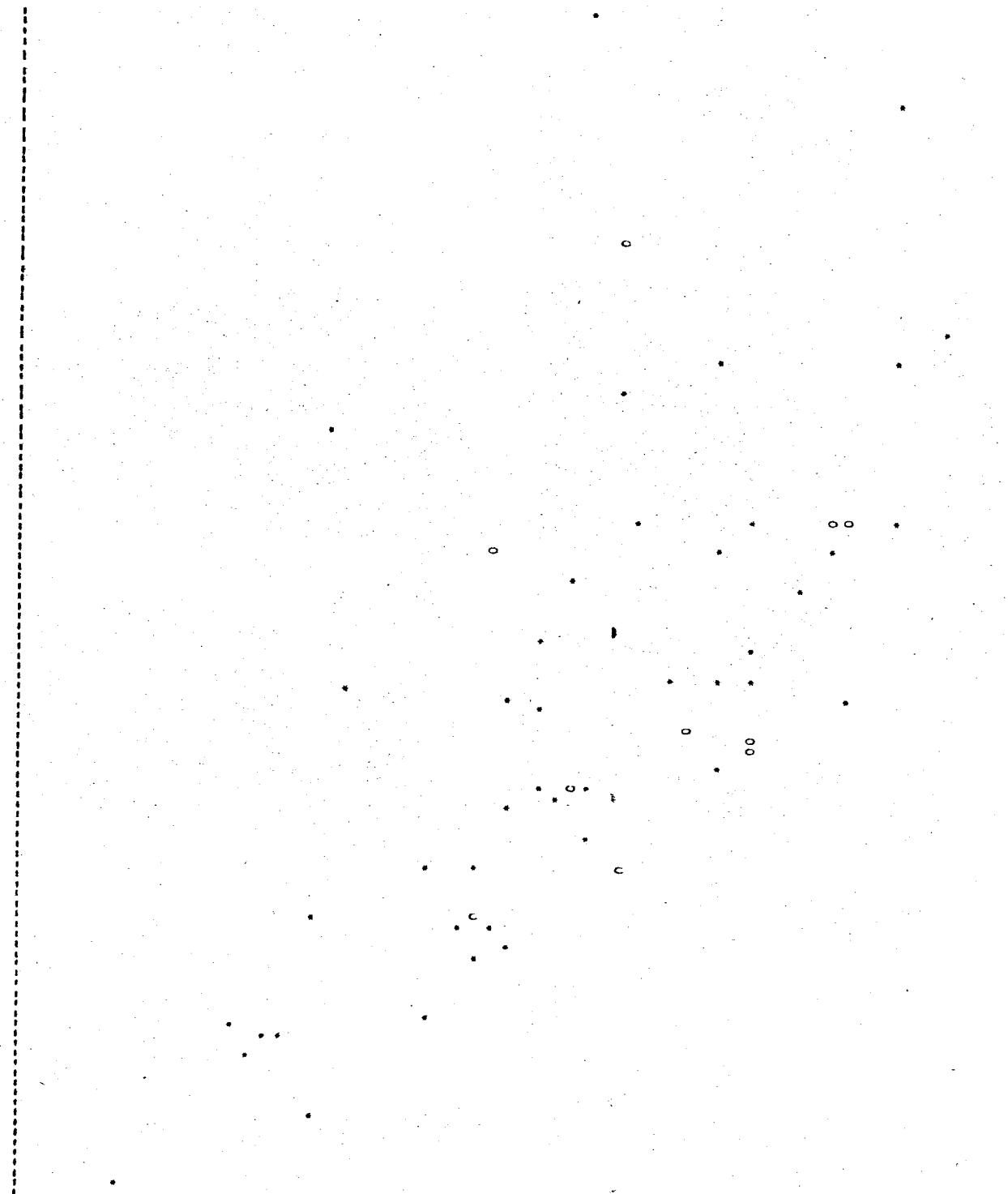
DATO NO ELEGIDO NUMERO = : 41 F.EXPERIMENTAL= 1.178906

DATO NO ELEGIDO NUMERO = : 42 F.EXPERIMENTAL= 2.317540

DATO NO ELEGIDO NUMERO = : 43 F.EXPERIMENTAL= 2.340552

DATO NO ELEGIDO NUMERO = : 44 F.EXPERIMENTAL= .302792

DATO NO ELEGIDO NUMERO = : 45 F.EXPERIMENTAL= .155150



CAPITULO CUARTO

DETECCION DE OUTLIERS BASADO EN LA DISTANCIA ENTRE MATRICES SIMETRICAS Y DEFINIDAS POSITIVAS

CONTENIDO.

4.1. INTRODUCCION.

En este capitulo damos un procedimiento para la detección de outliers en muestras procedentes de poblaciones $N_2(\mu, \Sigma)$, utilizando como estadistico el máximo del cuadrado de la distancia entre matrices de sumas de cuadrados y sumas de productos de observaciones muestrales.

Se comienza calculando la distancia entre matrices simétricas y definidas positivas, comprobandose que dicha distancia se puede definir entre matrices de sumas de cuadrados y suma de productos de observaciones muestrales.

A continuación obtenemos la función de densidad del cuadrado de la distancia, acotandose dicha función por otra que puede ser tabulada.

Finalizamos el capitulo exponiendo un criterio para la detección de outliers que se basa en el estadistico ordenado

$$\max_i d^2(S_{(n)}, S^{(i)}_{(n-k)}) \quad i = 1, 2, \ldots, \binom{n}{k}$$

4.2. DISTANCIA ENTRE MATRICES DE SUMAS DE CUADRADOS Y SUMAS DE PRODUCTOS.

En este apartado vamos a calcular la distancia entre matrices simètricas y definidas positivas, si se utiliza como forma métrica diferencial la dada por MAAS (1.955). A continuación se comprueba que dicha distancia se puede definir entre las matrices de sumas de cuadrados y sumas de productos (s. c. y s. p.) de observaciones muestrales. Encontrando finalmente la distribución del estadistico distancia.

4.2.1. Distancia geodésica entre matrices simétricas definidas positivas.

4.2.1.1. Definición. Sea H un espacio conexo sobre el que se ha

definido una forma métrica diferencial $(ds)^2$; la distancia entre dos puntos $x, y \in H$ viene dada por

$$d(x,y) = \inf_{\mathfrak{X}} \int_x^y ds$$

donde $\mathfrak{X}$ es el conjunto de todos los posibles caminos que unen X con Y.

La función $d(X,Y)$ verifica las siguientes propiedades (KOBAYASHI, NOMIZU , 1.963).

$$d(X,Y) \geq 0$$

$$d(X,Y) = 0 \Longleftrightarrow X = Y$$

$$d(X,Y) = d(Y,X)$$

$$d(X,Z) \leq d(X,Y) + d(Y,Z)$$

La topologia definida por d es equivalente a la topologia inicial del espacio H.

Las matrices simétricas definidas positivas de dimensión $p \times p$, forman un cono convexo en el espacio euclideo $\mathbb{R}^{\frac{1}{2}p(p+1)}$. Utilizando ahora la forma métrica diferencial para matrices simétricas definidas positivas dada por Maas, podemos enunciar el siguiente teorema.

4.2.1.1. Teorema .- Sea A y B dos matrices simétricas definidas positivas. Tomando como forma diferencial la dada por Maas, la distancia geodésica entre estas matrices viene dada por

$$d(A,B) = \left(\sum_{i=1}^{p} (\log \lambda_i)^2 \right)^{1/2}$$

donde λ_i son las raices de la ecuación determinante $|B - \lambda A| = 0$

En efecto :

La distancia entre dos matrices A y B vendrá dada por

$$d(A,B) = \inf_{\mathfrak{X}} \int_A^B ds$$

donde $\mathfrak{X}$ es el conjunto de todos los posibles caminos que unen A con B y ds es la forma métrica diferencial.

Por otro lado se sabe por Analisis Matricial que (BELLMAN , 1965).

" Dadas dos matrices simétricas reales A y B , con A definida positiva, existe una matriz no singular T tal que

$$TAT' = I \quad y \quad TBT' = \Lambda$$

donde Λ es una matriz diagonal y si B es definida positiva los elementos que componen la matriz Λ , $\lambda_i > 0 \ \forall i$ "

Como $(ds)^2$ es invariante por trasformaciones de congruencias podemos escribir,

$$d(A,B) = \inf_{\mathfrak{X}} \int_{I_p}^{\Lambda} ds$$

donde

$$(ds)^2 = \sum_{i=1}^{p} \left(\frac{d\lambda_i}{\lambda_i}\right)^2 + \sum_{i<j}^{p} \frac{(\lambda_i - \lambda_j)^2}{\lambda_i \lambda_j} (dz_{ij})^2$$

Haciendo el cambio $\eta_i = \log. \lambda_i \quad i=1,2,\ldots,p$ se tiene

$$d(A,B) = \inf_{\mathfrak{X}} \int_{0}^{\eta} ds$$

donde ahora

$$(ds)^2 = \sum_{i=1}^{p} (d\eta_i)^2 + \sum_{i<j}^{p} \left(e^{\eta_i - \eta_i} + e^{-(\eta_i - \eta_i)} - 2 e^{\frac{\eta_i - \eta_i}{2}} e^{-\frac{\eta_i - \eta_i}{2}}\right) (dz_{ij})^2 =$$

$$= \sum_{i=1}^{p} (d\eta_i)^2 + \sum_{i<j}^{p} \left(\operatorname{senh}.\left(\tfrac{1}{2}(\eta_i - \eta_j)\right)\right)^2 (dz_{ij})^2$$

Por tanto

$$d(A,B) = \inf_{\mathfrak{X}} \int_{0}^{\eta} ds = \inf_{\mathfrak{X}} \int_{0}^{\eta} \sqrt{\sum_{i=0}^{p} (d\eta_i)^2 + \sum_{i<j}^{p} \left(\operatorname{senh}.\left(\tfrac{1}{2}(\eta_i - \eta_j)\right)\right)^2 (dz_{ij})^2} >$$

$$\geq \inf_{x} \int_0^{\eta} \left(\sum_{i=0}^{p} (d\eta_i)^2 \right)^{1/2}$$

Y esto ultimo se hace minimo (1) si como distancia tomamos la euclidea entre el punto (0,0,....,0) y (η_1 , η_2 ,...., η_p).

En definitiva

$$d(A,B) = \left(\sum_{i=1}^{p} \eta_i^2 \right)^{1/2}$$

4.2.2. Distribución del estadistico distancia entre matrices de suma de cuadrados y suma de productos.

Aplicaremos el resultado anterior a las matrices de s. c. y s. p. de observaciones muestrales.

Así si

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

es una muestra aleatoria simple procedente de una población $N_p(\mu, \Sigma)$ no singular, entonces la matriz de s. c. y s. p. de observaciones muestrales para esta muestra de tamaño n viene dada por

$$S_{(n)} = X\left(I - \frac{1}{n} E_{nn}\right) X'$$

y sabemos que esta matriz es simétrica por su propia construcción , además si no conocemos los parametros poblacionales, $S_{(n)}$ está distribuida según una ley de Wishart p - dimensional con n - 1 grados de

---

(1) . El minimo se puede alcanzar de la misma forma si como trayectorias tomamos

$$\eta_i(t) = t\eta_i \qquad 0 \leq t \leq 1 \qquad i = 1,2,\ldots,p$$
$$z_{ij}(t) = 0$$

libertad y matriz asociada Σ ($W_p(n-1,\Sigma)$) (ANDERSON, 1.958) y si se tiene que $n > p$ entonce $S_{(n)}$ es definida positiva con probabilidad uno, esto se puede representar como

$$\int_{S>0} W_p(n-1,\Sigma)\, dS = 1$$

Luego podemos afirmar que las matrices de s. c. y s. p. muestrales, cumplen las condiciones exigidas para poder definir la funcion distancia $d(S_{(n-k)}, S_{(n)})$, salvo conjuntos de medida nula, en la forma.

$$d(S_{(n-k)}, S_{(n)}) = \left[\sum_{i=1}^{p} (\log \lambda_i)^2 \right]^{1/2} \qquad (2)$$

donde $S_{(n-k)}$es la matriz de s. c. y s. p. de $n - k$ ($n - k > p$) observaciones , y los λ_i son las soluciones de la ecuación determinante.

$$|S_{(n-k)} - \lambda S_{(n)}| = 0$$

Como estamos trabajando con matrices de tipo aleatorio la distancia definida en (2), podemos considerarla como una variable aleatoria y por tanto podemos hablar de la distribución de la variable distancia y para calcular dicha distribución vamos a necesitar la distribución conjunta de las raices $\lambda_1, \lambda_2, \ldots, \lambda_p$ que determinaremos basandonos en los siguientes resultados.

4.2.2.1. Lema. Sea X una muestra aleatoria simple de tamaño $n(n > p)$ procedente de una población $N_p(\mu, \Sigma)$ no singular. La matriz de s. c. y s. p. de la muestra X , $S_{(n)}$ puede descomponerse en la forma

$$S_{(n)} = S_{(n-k)} + S_{(k)} + \frac{k(n-k)}{n} (\bar{X}_{(n-k)} - \bar{X}_{(k)})(\bar{X}_{(n-k)} - \bar{X}_{(k)})'$$

donde $S_{(n-k)}$es la matriz de s. c. y s. p. de $n - k$ ($n - k > p$) observaciones y $\bar{X}_{(n-k)}$ su vector media, siendo $S_{(k)}$la matriz de s. c. y s. p. de la k ($k > p$) observaciones restantes y $\bar{X}_{(k)}$su vector media.

Además

$$S_{(k)} + \frac{k(n-k)}{n} (\bar{X}_{(n-k)} - \bar{X}_{(k)})(\bar{X}_{(n-k)} - \bar{X}_{(k)}) \in W_p(k, \Sigma)$$

En efecto

$$S_{(n)} = X\left(I - \frac{1}{n} E_{nn}\right)X' = \left(X_{(n-k)} \mid X_{(k)}\right)\left(I - \frac{1}{n} E_{nn}\right)\left(X_{(n-k)} \mid X_{(k)}\right)' =$$

$$= X_{(n-k)} X'_{(n-k)} + X_{(k)} X'_{(k)} - \frac{1}{n}\left(X_{(n-k)} \mid X_{(k)}\right)\begin{pmatrix} E_{n-k,n} \\ E_{k,n} \end{pmatrix}\left(X'_{(n-k)} \mid X'_{(k)}\right) =$$

$$= X_{(n-k)} X'_{(n-k)} + X_{(k)} X'_{(k)} - \frac{1}{n}\left[X_{(n-k)} E_{n-k,n-k} X'_{(n-k)} + \right.$$

$$\left. + X_{(n-k)} E_{(n-k),K} X'_{(k)} + X_{(k)} E_{k,(n-k)} X'_{(n-k)} + X_{(k)} E_{k,k} X'_{(k)}\right]$$

Sumando y restando

$$\frac{1}{n-k} X_{(n-k)} E_{(n-k),(n-k)} X'_{(n-k)} \quad \text{y} \quad \frac{1}{k} X_{(k)} E_{kk} X'_{(k)}$$

se obtiene

$$S_{(n)} = S_{(n-k)} + S_{(k)} + (n-k)\bar{X}_{(n-k)}\bar{X}'_{(n-k)} - \frac{(n-k)^2}{n}\bar{X}_{(n-k)}\bar{X}'_{(n-k)} +$$

$$+ k\bar{X}_{(k)}\bar{X}'_{(k)} - \frac{k^2}{n}\bar{X}_k\bar{X}'_k - \frac{(n-k)\cdot k}{n}\bar{X}_{(k)}\bar{X}'_{(n-k)} - \frac{(n-k)k}{n}\bar{X}_{(n-k)}\bar{X}'_{(k)} =$$

$$= S_{(n-k)} + S_{(k)} + \frac{k(n-k)}{n}\left(\bar{X}_{(n-k)} - \bar{X}_{(k)}\right)\left(\bar{X}_{(n-k)} - \bar{X}_{(k)}\right)'$$

Ahora bien, la matriz de s. c. y s. p. $S_{(k)}$ se distribuye segun una ley $W_p(k-1, \Sigma)$. Por otro lado como:

$$\bar{X}_{(n-k)} \in N_p\left(\mu, \frac{1}{n-k}\Sigma\right) \quad \text{y} \quad \bar{X}_{(k)} \in N_p\left(\mu, \frac{1}{k}\Sigma\right)$$

y ambas son independientes se deduce

$$\sqrt{\frac{k(n-k)}{n}}\left(\bar{X}_{(n-k)} - \bar{X}_{(k)}\right) \in N_p(0, \Sigma)$$

Por tanto

$$\left(\sqrt{\frac{k(n-k)}{n}}\right)^2\left(\bar{X}_{(n-k)} - \bar{X}_{(k)}\right)\left(\bar{X}_{(n-k)} - \bar{X}_{(k)}\right)'$$

se distribuye segun una ley pseudo Wishart p - dimensional com un grado de libertad y matriz asociada Σ.

Por último como

$$S_{(k)} \quad \text{y} \quad \frac{k(n-k)}{n}\left(\bar{X}_{(n-k)} - \bar{X}_{(k)}\right)\left(\bar{X}_{(n-k)} - \bar{X}_{(k)}\right)'$$

son independientes (WILKS , 1.962)

$$S' = S_{(k)} + \frac{k(n-k)}{n}(\bar{X}_{(n-k)} - \bar{X}_{(k)})(\bar{X}_{(n-k)} - \bar{X}_{(k)})'$$

se distribuye segun una ley $W_p(K, \Sigma)$ en virtud de la reproducti-- vidad de esta ley.

Otro resultado que utilizaremos para encontrar la funcion de densidad del estadistico distancia es el dado por ; HSU (1.939) , FISHER (1.939) , ROY (1.939) y KHRISNAIAH (1.978).

4.2.2.2. Lema. Sea S' una matriz p x p que se distribuye segun una ley $W_p(K,\Sigma)$ y $S_{(n-k)}$ una matriz p x p que se distribuye segun una ley $W_p(n-k-1, \Sigma)$ independiente de la anterior y sea $S_{(n)} = S_{(n-k)} + S'$ Entonces la distribución conjunta de las raices de la ecuación deter-- minante.

$$|S_{(n-k)} - \lambda S_{(n)}| = 0$$

es de la forma

$$f(\lambda_1, \lambda_2, \ldots, \lambda_p) = A \prod_{i=1}^{p} \prod_{j=i+1}^{p} (\lambda_j - \lambda_i) \left[\prod_{i=1}^{p} \lambda_i\right]^{\frac{1}{2}(n-k-p-2)} \times \left[\prod_{i=1}^{p} (1-\lambda_i)\right]^{\frac{1}{2}(k-p-1)}$$

donde

$$0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_p \leq 1$$

y

$$A = \pi^{p/2} \prod_{i=1}^{p} \frac{\Gamma\left(\frac{1}{2}(n-i)\right)}{\Gamma\left(\frac{1}{2}(n-k-i)\right)\Gamma\left(\frac{1}{2}(k-i+1)\right)\Gamma\left(\frac{1}{2}(p-i+1)\right)}$$

4.2.2.1. Teorema. Sea X una muestra aleatoria simple de tamaño n (n > 2) extraida de una población con distribución $N_2(\mu, \Sigma)$ no singular. Sea $S_{(n)}$ la matriz de s. c. y s. p. de las observaciones

muestrales y $S_{(n-k)}$ la matriz de s. c. y s. p. de n - k observaciones (k > 2).

La función de densidad del estadistico cuadrado de la distancia

$$U = d^2(S_{(n-k)}, S_{(n)}) = (\log \lambda_1)^2 + (\log \lambda_2)^2$$

donde λ_1 y λ_2 son las soluciones de la ecuación determinante

$$|S_{(n-k)} - \lambda S_{(n)}| = 0$$

viene dada por

$$\begin{cases} A e^{-\sqrt{u}\left(\frac{n-k-2}{2}\right)} \int_{\sqrt{2}-1}^{1} \left(e^{-\sqrt{u}\frac{1-w^2}{1+w^2}} - e^{-\sqrt{u}\frac{2w}{1+w^2}}\right) e^{-\sqrt{u}(n-k-2)w\frac{1-w}{1+w^2}} \frac{1}{1+w^2} \times \\ \qquad \times \left[\left(1-e^{-\sqrt{u}\frac{2w}{1+w^2}}\right)\left(1-e^{-\sqrt{u}\frac{1-w^2}{1+w^2}}\right)\right]^{\frac{k-3}{2}} dw. \quad 0 \le u \le \infty. \\ 0 \qquad\qquad \text{en el resto} \end{cases}$$

siendo

$$A = \frac{\sqrt{\pi}\ \Gamma\left(\frac{1}{2}(n-1)\right) \cdot \Gamma\left(\frac{1}{2}(n-2)\right)}{\Gamma\left(\frac{1}{2}(n-k-1)\right)\Gamma\left(\frac{1}{2}(n-k-2)\right)\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k-1}{2}\right)}$$

En efecto

En virtud del Lema 4.2.2.2.

$$f(\lambda_1, \lambda_2) = A(\lambda_2 - \lambda_1)(\lambda_1 \lambda_2)^{\frac{1}{2}(n-k-4)}\left[(1-\lambda_1)(1-\lambda_2)\right]^{\frac{k-3}{2}} \quad 0 \le \lambda_1 \le \lambda_2 \le 1$$

y

$$A = \frac{\sqrt{\pi}\ \Gamma\left(\frac{1}{2}(n-1)\right)\ \Gamma\left(\frac{1}{2}(n-2)\right)}{\Gamma\left(\frac{1}{2}(n-k-1)\right)\Gamma\left(\frac{1}{2}(n-k-2)\right)\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{k-1}{2}\right)}$$

Haciendo el cambio de variables

$$X = -\ln \lambda_1$$

$$Y = -\ln \lambda_2$$

se obtiene

$$f(x,y) = A(e^{-y} - e^{-x})\, e^{-(x+y)\frac{1}{2}(n-k-2)}\left[(1-e^{-y})(1-e^{-x})\right]^{\frac{k-3}{2}} \quad 0 < y < x < \infty$$

y si hacemos

$$\theta = X^2$$

$$\varphi = Y^2$$

obtendremos:

$$f(\theta,\varphi) = \frac{A}{4}\left(e^{-\sqrt{\varphi}} - e^{-\sqrt{\theta}}\right) e^{-(\sqrt{\theta}+\sqrt{\varphi})\frac{1}{2}(n-k-2)}\left[\left(1-e^{-\sqrt{\varphi}}\right)\left(1-e^{-\sqrt{\theta}}\right)\right]^{\frac{k-3}{2}} (\theta\varphi)^{-1/2}$$

$$0 < \varphi < \theta < \infty$$

y por ultimo realizando la trasformación

$$U = \theta + \varphi$$

$$t = \theta$$

Se obtiene

$$f(u,t) = \frac{A}{4}\left(e^{-\sqrt{u-t}} - e^{-\sqrt{t}}\right) e^{-(\sqrt{t}+\sqrt{u-t})\frac{1}{2}(n-k-2)}\left[\left(1-e^{-\sqrt{u-t}}\right)\left(1-e^{-\sqrt{t}}\right)\right]^{\frac{k-3}{2}} \times \left(t(u-t)\right)^{-1/2}$$

$$0 < U < \infty, \quad \frac{U}{2} < t < U$$

Por lo que la función de densidad marginal de U, viene dada por

$$f(u) = \frac{A}{4}\int_{U/2}^{U}\left(e^{-\sqrt{u-t}} - e^{-\sqrt{t}}\right) e^{-(\sqrt{t}+\sqrt{u-t})\frac{1}{2}(n-k-2)}\left[\left(1-e^{-\sqrt{u-t}}\right)\left(1-e^{-\sqrt{t}}\right)\right]^{\frac{k-3}{2}}\left[t(u-t)\right]^{-1/2} dt$$

$$0 < U < \infty$$

Efectuando ahora el cambio de variable

$$Z = \sqrt{\frac{2(U-t)}{U}}$$

se obtiene

$$f(u) = \frac{A}{4}\int_0^1 \left(e^{-\frac{z}{\sqrt{2}}\sqrt{u}} - e^{-\sqrt{u}\sqrt{1-\frac{z^2}{2}}}\right) e^{-\sqrt{u}\,\frac{n-k-2}{2}\left(\frac{z}{\sqrt{2}}+\sqrt{1-\frac{z^2}{2}}\right)}\left[\left(1-e^{-\sqrt{u}\sqrt{1-\frac{z^2}{2}}}\right)\left(1-e^{-\frac{z}{\sqrt{2}}\sqrt{u}}\right)\right]^{\frac{k-3}{2}} \times \sqrt{2}\left(1-\frac{z^2}{2}\right)^{-1/2} dz$$

$$0 < U < \infty$$

y realizando la trasformación

$$\sqrt{1-\frac{z^2}{2}} = \left(1+\frac{z}{\sqrt{2}}\right) w$$

se deduce finalmente

$$f(u) = A\, e^{-\sqrt{u}\left(\frac{n-k-2}{2}\right)}\int_{\sqrt{2}-1}^{1}\left(e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}} - e^{-\sqrt{u}\,\frac{2w}{1+w^2}}\right) e^{-\sqrt{u}\,(n-k-2)w\,\frac{1-w}{1+w^2}}\left[\left(1-e^{-\sqrt{u}\,\frac{2w}{1-w^2}}\right)\left(1-e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}}\right)\right]^{\frac{k-3}{2}} \times \frac{dw}{1+w^2}$$

$$0 < U < \infty$$

4.3. ACOTACION.

4.3.1. <u>Teorema</u>. La función

$$f(u) = A\, e^{-\sqrt{u}\left(\frac{n-k-2}{2}\right)} \int_{\sqrt{2}-1}^{1} \left(e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}} - e^{-\sqrt{u}\,\frac{2w}{1+w^2}}\right) e^{-\sqrt{u}\,(n-k-2)\,w\,\frac{1-w}{1+w^2}}\, \frac{1}{1+w^2} \cdot \cdot$$
$$\cdot \left[\left(1 - e^{-\sqrt{u}\,\frac{2w}{1+w^2}}\right)\left(1 - e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}}\right)\right]^{\frac{k-3}{2}} dw \qquad 0 < u < \infty$$

siendo

$$A = \frac{\sqrt{\pi}\,\Gamma\left(\frac{1}{2}(n-1)\right)\Gamma\left(\frac{1}{2}(n-2)\right)}{\Gamma\left(\frac{1}{2}(n-k-1)\right)\Gamma\left(\frac{1}{2}(n-k-2)\right)\Gamma\left(\frac{k}{2}\right)\cdot\Gamma\left(\frac{k-1}{2}\right)}$$

está acotada por la función

$$g(u) = A\, e^{-\sqrt{u}\left(\frac{n-k-2}{2}\right)} \left(1 - e^{-\sqrt{u}}\right)^{k-2}$$

En efecto:

Sea

$$y_1 = \left(e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}} - e^{-\sqrt{u}\,\frac{2w}{1+w^2}}\right)$$

$$y_1' = e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}} \sqrt{u}\,\frac{4w}{(1+w^2)^2} + e^{-\sqrt{u}\,\frac{2w}{1+w^2}} \sqrt{u}\,\frac{2-2w^2}{(1+w^2)^2} > 0 \qquad \forall w \in I$$

luego y_1 es una función creciente en el intervalo de integración $I = (\sqrt{2}-1,\ 1)$. Por tanto

$$\sup_I y_1 = \left(1 - e^{-\sqrt{u}}\right)$$

Sea

$$y_2 = e^{-\sqrt{u}\,(n-k-2)\,w\,\frac{1-w}{1+w^2}}$$

$$y_2' = -e^{-\sqrt{u}\,(n-k-2)\,w\,\frac{1-w}{1+w^2}} \sqrt{u}\,(n-k-2)\left[\frac{1-2w-w^2}{(1+w^2)^2}\right]$$

esta derivada primera se anula en los puntos

$$w_1 = -1-\sqrt{2} \qquad w_2 = -1+\sqrt{2}$$

dandose

$$y_2''(w_1) < 0 \qquad y_2''(w_2) > 0$$

Luego

$$\sup_I p.\ y_2 = 1$$

Tomemos

$$y_3 = \left(1 - e^{-\sqrt{u}\,\frac{2w}{1+w^2}}\right)^{\frac{k-3}{2}}$$

$$y_3' = \frac{k-3}{2}\left(1 - e^{-\sqrt{u}\,\frac{2w}{1+w^2}}\right)^{\frac{k-5}{2}} e^{-\sqrt{u}\,\frac{2w}{1+w^2}}\sqrt{u}\,\frac{2-2w^2}{(1+w^2)^2}$$

e $y_3' > 0 \quad \forall w \in (\sqrt{2}-1,\ 1)$, por lo que alcanza el supremo en $W = 1$

así

$$\sup_I p.\ y_3 = \left(1 - e^{-\sqrt{u}}\right)^{\frac{k-3}{2}}$$

Sea

$$y_4 = \left(1 - e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}}\right)^{\frac{k-3}{2}}$$

$$y_4' = \frac{k-3}{2}\left(1 - e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}}\right)^{\frac{k-5}{2}} e^{-\sqrt{u}\,\frac{1-w^2}{1+w^2}}\sqrt{u}\left[\frac{-4w}{(1+w^2)^2}\right]$$

e $y_4' < 0$, por tanto el supremo en el intervalo de integración se alcanza en el extremo inferior del mismo así:

$$\sup_I p.\ y_4 = \left(1 - e^{-\sqrt{u}\,\times\,0.7071}\right)^{\frac{k-3}{2}}$$

Y por ultimo

$$\int_{\sqrt{2}-1}^{1} \frac{dw}{1+w^2} = \text{arc.tg}\, w \Big|_{\sqrt{2}-1}^{1} \leq 0.5$$

Por tanto

$$f(u) = A\int_{\sqrt{2}-1}^{1} y_1 \cdot y_2 \cdot y_3 \cdot y_4 \frac{dw}{1+w^2} \leq A\, e^{-\sqrt{u}\left(\frac{n-k-2}{2}\right)}\left(1 - e^{-\sqrt{u}}\right)\left(1 - e^{-\sqrt{u}}\right)^{\frac{k-3}{2}}$$

$$\times \left(1 - e^{-\sqrt{u}\,\times\,0.7071}\right)^{\frac{k-3}{2}} \times \frac{1}{2}$$

así

$$f(u) \leq A\, e^{-\sqrt{u}\left(\frac{n-k-2}{2}\right)}\left(1 - e^{-\sqrt{u}}\right)^{k-2} = g(u) \quad \forall u \quad 0 < u < \infty$$

4.3.2. Teorema. La funcion $q(x)$ definida mediante la expresión

$$q(x) = \frac{1}{2\sum_{j=0}^{k-2}\left[\binom{k-2}{j}(-1)^j \frac{1}{\left(\frac{n-k-2}{2}+j\right)^2}\right]} e^{-\sqrt{x}\left(\frac{n-k-2}{2}\right)}\left(1-e^{-\sqrt{x}}\right)^{k-2} \quad 0<x<\infty$$

con $k>2$ y $(n-k)>2$ es función de densidad.

En efecto:

Para que $q(x)$ sea función de densidad ha de ocurrir que

$$q(x) \geq 0 \qquad \forall x$$

$$\int_0^\infty q(x)\,dx = 1$$

La primera propiedad es evidente que se cumple. Veamos que es cierta la segunda.

$$\int_0^\infty q(x)\,dx = \frac{1}{2\sum_{j=0}^{k-2}\left[\binom{k-2}{j}(-1)^j \frac{1}{\left(\frac{n-k-2}{2}+j\right)^2}\right]} \int_0^\infty e^{-\sqrt{x}\left(\frac{n-k-2}{2}\right)}\left(1-e^{-\sqrt{x}}\right)^{k-2} dx =$$

$$= \frac{1}{2\sum_{j=0}^{k-2}\left[\binom{k-2}{j}(-1)^j \frac{1}{\left(\frac{n-k-2}{2}+j\right)^2}\right]} \sum_{j=0}^{k-2}(-1)^j\binom{k-2}{j}\int_0^\infty e^{-\sqrt{x}\left(\frac{n-k-2}{2}+j\right)} dx$$

Realizando el cambio de variable

$$\sqrt{x} = y$$

se obtiene

$$\int_0^\infty q(x)\,dx = \frac{1}{2\sum_{j=0}^{k-2}\left[\binom{k-2}{j}(-1)^j \frac{1}{\left(\frac{n-k-2}{2}+j\right)^2}\right]} 2\sum_{j=0}^{k-2}(-1)^j\binom{k-2}{j}\int_0^\infty y\, e^{-y\left(\frac{n-k-2}{2}+j\right)} dy.$$

La función de densidad $g(x)$ presenta la ventaja sobre $f(u)$ de que su función de distribución se puede tabular, aplicando para ello el metodo de Simpson . Y además $F(u)$ domina estocasticamente a $G(x)$.

En las páginas 83 a 87 se incluye un programa de ordenador para las representaciones gráficas de $f(u)$ y $g(x)$, determinando dichas representaciones en el caso particular $n = 55$ y $k = 10$.

A continuación en las páginas 88 y 89 se incluye un programa para la tabulación de $G(x)$ en la situación práctica $n = 15$ y $k = 6$.

Representación gráfica de las funciones $f(u)$ y $g(x)$.

```
      C     ------------------------------------------------------------------
      C COMPARACION Y REPRESENTACION DE LAS FUNCIONES DE DENSIDAD DE LA VARIABLE
      C DISTANCIA AL CUADRADO F(U) Y LA DADA EN EL TEOREMA 4.3.2.
      C (EN EL PROGRAMA, F(U) ES F, Y  LA  DEL TEOREMA ES G)
      C     ------------------------------------------------------------------
            IMPLICIT DOUBLE PRECISION(A-H,O-U,W-Z)
            COMMON N,K,U
            DIMENSION A(2),V(61,81),U1(61),F2(61),G1(61)
            EXTERNAL F1
            LOGICAL REL
            N=5
            U=0
            K=10
            DO 10 I=1,61
              DO 10 J=1,81
         10 V(I,J)=' '
            DO 11 I=1,61
         11 V(I,1)='I'
              DO 12 I=1,81
         12 V(1,I)='-'
            G1(1)=0
            F2(1)=0
            U1(1)=0
      C     ------------------------------------------------------------------
      C CALCULO DE LAS CONSTANTES B Y C
      C     ------------------------------------------------------------------
            B=4./(N-K-2.)**2
            K2=K-2
            DO 1 IR=1,K2
          1 B=B+(-1)**IR*(FACT(K2)/(FACT(IR)*FACT(K2-IR)))*
           *(4./(N-K-2+2*IR))**2)
            C=FACT(N-3)/(4*FACT(K2)*FACT(N-K-3))
      C     ------------------------------------------------------------------
      C CALCULO DE LAS FUNCIONES F Y G
      C     ------------------------------------------------------------------
            A(1)=DSQRT(2.D)-1.D
            A(2)=1.D
            E=1.D-6
            MAXIT=2000
            REL=.TRUE.
              PRINT 9
          9 FORMAT(1X,//,1X,'VALORES COMPARATIVOS DE LAS FUNCIONES F Y G',//)
            DO 2 II=1,60
            U=U+.05D/60
            AINT=SIMPNT(F1,A,E,REL,MAXIT,1.D,47)
            F=C*AINT*DEXP(-DSQRT(U)*(N-K-2)/2.  )
            G=(DEXP(-DSQRT(U)*((N-K-2)/2.)))*(1.-DEXP(-DSQRT(U)))**(K-2))
            G=G/(2*3)
            II=II+1
            JF=(200*F)/10+1
            JG=(200*G)/10+1
            IF(JF.NE.JG) GOTO22
            V(II,JF)='*'
            GOTO 23
         22 V(II,JF)='F'
            V(II,JG)='G'
         23 G1(II)=G
            F2(II)=F
              U1(II)=U
            PRINT6,U,F,G
          6 FORMAT(1X,'U= ',D9.3,10X,'F(U)= ',D12.6,10X,'G(X)= ',D12.6,//)
            GOTO 2
          3 PRINT 7,U
          7 FORMAT(1X,'ERROR EN U=' D9.3)
          2 CONTINUE
      C     ------------------------------------------------------------------
      C GRAFICA DE LAS FUNCIONES DE DENSIDAD F Y G
      C     ------------------------------------------------------------------
            PRINT 31
         31 FORMAT(1H1)
            PRINT 30,((U1(I),F2(I),G1(I),(V(I,J),J=1,81)),I=1,61)
         30 FORMAT(1X,D9.3,1X,D12.6,1X,D12.6,10X,81A1)
            STOP
            END
```

```
END OF COMPILATION:        NO  DIAGNOSTICS.
```

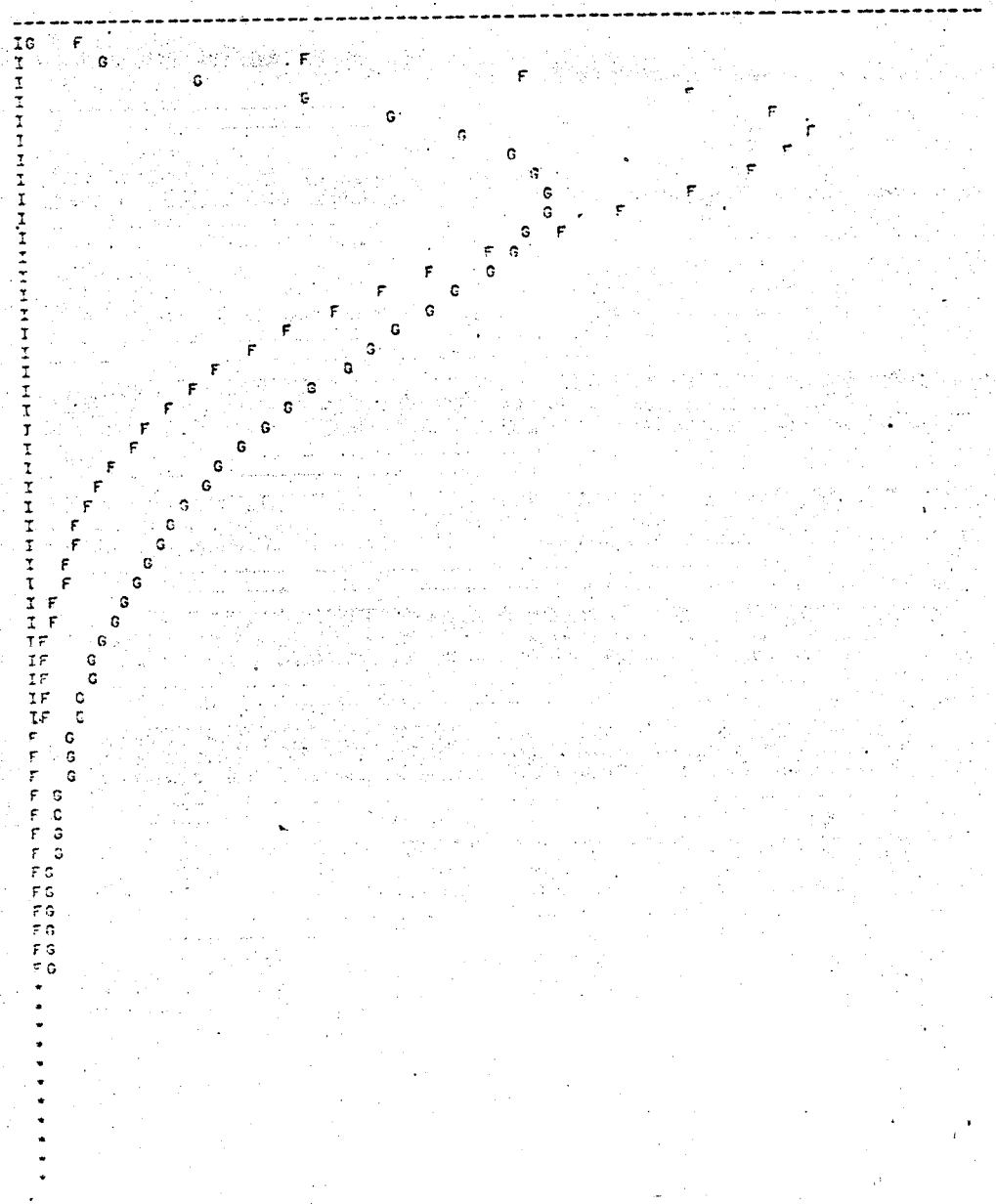
```
      DOUBLE PRECISION FUNCTION SIM3NI(FX,A,E,REL,MAXIT,FK,$)                SIM3NI
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)                                     SIM3NI
      DIMENSION A(2)
      LOGICAL REL
      PREV=0.
      H=(A(2)-A(1))/3.
      X=A(1)
      N=0
      M=3
      S=0.
      DO 3 J=1,MAXIT
      DO 1 I=N,M
      R=3.
      IF(MOD(I,3).EQ.2*N) R=N+1.
      S=S+FX(X,FK)*R
    1 X=X+H
      SIM3NI=S*H*.375D0/(N+1.D0)
      IF(N.EQ.0) GO TO 2
      H=H*.5D0
      M=2*M
      R=SIM3NI-PREV
      IF(REL) R=R/SIM3NI
      IF(DABS(R).LT.E) GO TO 4
    2 PREV=SIM3NI
      N=1
    3 X=A(1)+.5D0*H
      RETURN 7
    4 RETURN
      END                                                                    SIM3NI

END OF COMPILATION:        NO  DIAGNOSTICS.


      DOUBLE PRECISION FUNCTION FACT(M)
      COMMON N,K
      IF(M) 1,3,4
    1 PRINT 2,N,K
    2 FORMAT(' ARGUMENTO NEGATIVO PARA UN FACTORIAL ,N=',I4,' K=',I4)
      STOP
    3 FACT=1
      RETURN
    4 FACT=1
      DO 5 I=1,M
    5 FACT=FACT*I
      RETURN
      END

END OF COMPILATION:        NO  DIAGNOSTICS.


      DOUBLE PRECISION FUNCTION F1(T,FK)
      IMPLICIT DOUBLE PRECISION(A-H,O-U,W-Z)
      COMMON N,K,U
      R=DEXP(-DSQRT(U)*(1. -T**2)/(1. +T**2))
      S=DEXP(-DSQRT(U)*2. *T/(1. +T**2))
      Q=DEXP(-DSQRT(U)*(N-K-2)*T*(1. -T)/(1. +T**2))
      F1=(R-S)*Q*((1. -S)*(1. -R))**((K-3)/2. )/(1. +T**2)
      RETURN
      END

END OF COMPILATION:        NO  DIAGNOSTICS.
```

VALORES COMPARATIVOS DE LAS FUNCIONES F Y G

U= .1 08 -0 01	F(U) = .05 34 32 +0 00	G(X) = .187171 +000
U= .2 17 -0 01	F(U) = .26 75 17 +0 01	G(X) = .787149 +000
U= .3 25 -0 01	F(U) = .47 75 08 +0 01	G(X) = .172039 +001
U= .4 33 -0 01	F(U) = .62 84 37 +0 01	G(X) = .267974 +001
U= .5 42 -0 01	F(U) = .71 09 23 +0 01	G(X) = .350905 +001
U= .6 50 -0 01	F(U) = .73 56 30 +0 01	G(X) = .414455 +001
U= .7 58 -0 01	F(U) = .72 06 72 +0 01	G(X) = .457726 +001
U= .8 67 -0 01	F(U) = .68 03 05 +0 01	G(X) = .482580 +001
U= .975 -0 01	F(U) = .62 57 94 +0 01	G(X) = .492021 +001
U= .1 08 +0 00	F(U) = .56 50 43 +0 01	G(X) = .489295 +001
U= .1 19 +0 00	F(U) = .50 32 72 +0 01	G(X) = .477255 +001
U= .1 30 +0 00	F(U) = .44 37 13 +0 01	G(X) = .458646 +001
U= .1 41 +0 00	F(U) = .38 92 15 +0 01	G(X) = .435558 +001
U= .1 52 +0 00	F(U) = .33 76 89 +0 01	G(X) = .409837 +001
U= .1 62 +0 00	F(U) = .29 24 46 +0 01	G(X) = .782388 +001
U= .1 73 +0 00	F(U) = .25 24 20 +0 01	G(X) = .756430 +001
U= .1 84 +0 00	F(U) = .21 73 27 +0 01	G(X) = .727155 +001
U= .1 95 +0 00	F(U) = .18 67 56 +0 01	G(X) = .700488 +001
U= .2 05 +0 00	F(U) = .16 02 85 +0 01	G(X) = .674984 +001
U= .2 17 +0 00	F(U) = .13 74 38 +0 01	G(X) = .650871 +001
U= .2 27 +0 00	F(U) = .11 77 77 +0 01	G(X) = .628281 +001
U= .2 38 +0 00	F(U) = .10 08 36 +0 01	G(X) = .607272 +001
U= .2 49 +0 00	F(U) = .86 42 75 +0 00	G(X) = .587851 +001
U= .2 60 +0 00	F(U) = .74 03 99 +0 00	G(X) = .569986 +001
U= .2 71 +0 00	F(U) = .63 44 28 +0 00	G(X) = .553619 +001
U= .2 82 +0 00	F(U) = .54 30 19 +0 00	G(X) = .538677 +001
U= .2 92 +0 00	F(U) = .46 53 65 +0 00	G(X) = .525073 +001
U= .3 03 +0 00	F(U) = .40 01 59 +0 00	G(X) = .512718 +001
U= .3 14 +0 00	F(U) = .34 35 60 +0 00	G(X) = .501520 +001
U= .3 25 +0 00	F(U) = .29 51 62 +0 00	G(X) = .491381 +000
U= .3 36 +0 00	F(U) = .25 37 03 +0 00	G(X) = .482237 +000
U= .3 47 +0 00	F(U) = .21 83 32 +0 00	G(X) = .473926 +000
U= .3 57 +0 00	F(U) = .18 79 95 +0 00	G(X) = .466253 +000
U= .3 68 +0 00	F(U) = .16 20 02 +0 00	G(X) = .459174 +000
U= .3 79 +0 00	F(U) = .13 97 17 +0 00	G(X) = .452796 +000

U= .3 90 +0 00	F(U) = .12 05 96 +000	G(X) = .487483+000
U= .4 01 +0 00	F(U) = .10 41 83 +000	G(X) = .434649+000
U= .4 12 +0 00	F(U) = .90 07 69-001	G(X) = .390757+000
U= .4 22 +0 00	F(U) = .77 95 07-001	G(X) = .351319+000
U= .4 33 +0 00	F(U) = .67 51 33-001	G(X) = .315890+000
U= .4 44 +0 00	F(U) = .58 52 36-001	G(X) = .284070+000
U= .4 55 +0 00	F(U) = .50 77 44-001	G(X) = .255492+000
U= .4 66 +0 00	F(U) = .44 08 90-001	G(X) = .229828+000
U= .4 77 +0 00	F(U) = .38 31 64-001	G(X) = .206781+000
U= .4 87 +0 00	F(U) = .33 32 78-001	G(X) = .186083+000
U= .4 98 +0 00	F(U) = .29 01 30-001	G(X) = .167495+000
U= .5 09 +0 00	F(U) = .25 27 80-001	G(X) = .150799+000
U= .5 20 +0 00	F(U) = .22 04 20-001	G(X) = .135801+000
U= .5 31 +0 00	F(U) = .19 23 60-001	G(X) = .122326+000
U= .5 42 +0 00	F(U) = .16 80 09-001	G(X) = .110217+000
U= .5 52 +0 00	F(U) = .14 68 59-001	G(X) = .992349-001
U= .5 53 +0 00	F(U) = .12 84 74-001	G(X) = .895524-001
U= .5 74 +0 00	F(U) = .11 24 78-001	G(X) = .807566-001
U= .5 85 +0 00	F(U) = .98 55 18-002	G(X) = .728462-001
U= .5 96 +0 00	F(U) = .86 41 54-002	G(X) = .657304-001
U= .6 07 +0 00	F(U) = .75 83 15-002	G(X) = .593277-001
U= .6 17 +0 00	F(U) = .66 59 40-002	G(X) = .535681-001
U= .6 28 +0 00	F(U) = .58 52 53-002	G(X) = .483772-001
U= .6 39 +0 00	F(U) = .51 47 20-002	G(X) = .437055-001
U= .6 50 +0 00	F(U) = .45 30 15-002	G(X) = .394973-001

.00 0	.00 00 00	.0 00 00 0
.13 8+00 1	.65 24 32 +0 00	.1 47 17 1+000
.21 7+00 1	.26 75 17 +0 01	.7 87 14 3+000
.32 5+00 1	.47 75 60 +0 01	.1 72 03 8+001
.43 3+00 1	.62 64 37 +0 01	.2 57 57 4+001
.54 2+00 1	.71 04 25 +0 01	.3 50 30 6+001
.65 0+00 1	.77 69 20 +0 01	.4 14 45 5+001
.75 8+00 1	.72 06 72 +0 01	.4 57 72 6+001
.86 7+00 1	.66 63 06 +0 01	.4 82 56 6+001
.97 5+00 1	.63 67 04 +0 01	.4 82 62 1+001
.10 84+01 0	.65 50 43 +0 01	.4 99 24 5+001
.11 9+00 0	.69 32 72 +0 01	.4 77 25 5+001
.13 0+00 0	.44 37 12 +0 01	.4 68 84 6+001
.14 1+00 0	.38 62 15 +0 01	.4 35 65 5+001
.15 2+00 0	.37 76 85 +0 01	.4 09 69 7+001
.16 2+00 0	.39 24 46 +0 01	.3 82 30 8+001
.17 3+00 0	.25 24 20 +0 01	.3 54 67 0+001
.18 4+00 0	.21 73 27 +0 01	.3 27 15 5+001
.19 5+00 0	.19 67 65 +0 01	.3 03 48 9+001
.20 6+00 0	.16 02 69 +0 01	.2 74 98 4+001
.21 7+00 0	.13 74 36 +0 01	.2 50 87 1+001
.22 7+00 0	.11 77 77 +0 01	.2 28 06 1+001
.23 8+00 0	.10 08 96 +0 01	.2 07 27 2+001
.24 9+00 0	.86 42 75 +0 00	.1 87 65 1+001
.25 0+00 0	.74 03 90 +0 00	.1 69 08 6+001
.27 1+00 0	.63 44 20 +0 00	.1 52 61 9+001
.28 2+00 0	.54 35 19 +0 00	.1 37 67 7+001
.29 2+00 0	.46 63 65 +0 00	.1 25 07 3+001
.30 3+00 0	.40 01 93 +0 01	.1 12 71 8+001
.31 4+00 0	.34 35 60 +0 00	.1 01 52 0+001
.32 5+00 0	.29 51 02 +0 00	.9 13 88 1+000
.33 6+00 0	.25 37 63 +0 00	.8 22 37 7+000
.34 7+00 0	.21 83 02 +0 00	.7 39 72 6+000
.35 7+00 0	.18 75 06 +0 00	.6 65 45 7+000
.36 8+00 0	.16 10 02 +0 00	.5 98 12 4+000
.37 9+00 0	.13 87 17 +0 00	.5 37 70 6+000
.39 0+00 0	.12 05 05 +0 00	.4 83 48 3+000
.40 1+00 0	.10 43 83 +0 00	.4 34 64 9+000
.41 2+00 0	.91 07 83-0 01	.3 90 75 7+000
.42 2+00 0	.77 05 47-0 01	.3 51 31 3+000
.43 3+00 0	.67 51 33-0 01	.3 15 98 0+000
.44 4+00 0	.58 52 36-0 01	.2 84 07 0+000
.45 5+00 0	.50 77 44-0 01	.2 55 46 2+000
.46 6+00 0	.44 06 30-0 01	.2 29 90 0+000
.47 7+00 0	.38 21 24-0 01	.2 06 76 1+000
.48 7+00 0	.33 12 78-0 01	.1 86 08 3+000
.49 8+00 0	.29 01 30-0 01	.1 67 45 5+000
.50 9+00 0	.25 27 80-0 01	.1 50 79 7+000
.52 0+00 0	.22 06 20-0 01	.1 35 80 1+000
.53 1+00 0	.19 25 50-0 01	.1 22 32 6+000
.54 2+00 0	.16 80 05-0 01	.1 10 21 7+000
.55 2+00 0	.14 68 59-0 01	.9 92 30 3-001
.56 3+00 0	.12 84 76-0 01	.8 95 52 4+001
.57 4+00 0	.11 24 10-0 01	.8 07 88 6-001
.58 5+00 0	.98 33 14-0 02	.7 29 45 2-001
.59 6+00 0	.85 91 46-0 02	.6 57 30 4-001
.60 7+00 0	.75 03 47-0 02	.5 93 27 7-001
.61 7+00 0	.65 52 01-0 02	.5 35 05 3-001
.62 8+00 0	.57 21 63-0 02	.4 82 77 2-001
.63 9+00 0	.49 97 21-0 02	.4 34 95 4-001
.65 0+00 0	.43 65 15-0 02	.3 92 47 3-001



Tabulación de la función de distribución $G_1(x)$:

```
      DOUBLE PRECISION FUNCTION SIM3NI(FX,A,E,REL,MAXIT,FK,S)                SIM3NI
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)                                     SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     INTEGRACION NUMERICA UTILIZANDO LA REGLA 3/8 DE SIMPSON                SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      DIMENSION A(2)                                                         SIM3NI
      LOGICAL REL                                                            SIM3NI
      PREV=0.                                                                SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     INICIALIZAR H,X,N,M,S                                                  SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      H=(A(2)-A(1))/3.                                                       SIM3NI
      X=A(1)                                                                 SIM3NI
      N=0                                                                    SIM3NI
      M=3                                                                    SIM3NI
      S=0.                                                                   SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     CICLO HASTA EL MAXIMO NUMERO DE EVALUACIONES                           SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      DO 3 J=1,MAXIT                                                         SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     CICLO PARA EL NUMERO DE EVALUACIONES                                   SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      DO 1 I=N,M                                                             SIM3NI
      R=3.                                                                   SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     DETERMINACION DEL COEFICIENTE R                                        SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      IF(MOD(I,3).EQ.2*N) R=N+1.                                             SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     SUMA DE LAS EVALUACIONES DE LA FUNCION                                 SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      S=S+FX(X,FK)*R                                                         SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     INCREMENTAR X                                                          SIM3NI
C-----------------------------------------------------------------------     SIM3NI
    1 X=X+H                                                                  SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     OBTENCION DEL NUEVO VALOR DE INTEGRACION                               SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      SIM3NI=S*H*.375D0/(N+1.D0)                                             SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     TEST PARA EL PRIMER CICLO                                              SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      IF(N.EQ.0) GO TO 2                                                     SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     PARA LOS CICLOS DISTINTOS DEL PRIMERO,MITAD DE H Y DOBLE DE M          SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      H=H*.5D0                                                               SIM3NI
      M=2*M                                                                  SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     REVISION DEL ERROR DE CONTROL                                          SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      R=SIM3NI-PREV                                                          SIM3NI
      IF(REL) R=R/SIM3NI                                                     SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     SI EL ERROR ESTA DENTRO DE LOS LIMITES DADOS,FIN                       SIM3NI
C-----------------------------------------------------------------------     SIM3NI
      IF(DABS(R).LT.E) GO TO 4                                               SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     FIJAR UN NUEVO VALOR DE INTEGRACION                                    SIM3NI
C-----------------------------------------------------------------------     SIM3NI
    2 PREV=SIM3NI                                                            SIM3NI
      N=1                                                                    SIM3NI
C-----------------------------------------------------------------------     SIM3NI
C     OBTENER EL NUEVO LIMITE INFERIOR PARA LA EVALUACION DE LA FUNCION      SIM3NI
C-----------------------------------------------------------------------     SIM3NI
    3 X=A(1)+.5D0*H                                                          SIM3NI
      RETURN 7                                                               SIM3NI
    4 RETURN                                                                 SIM3NI
      END                                                                    SIM3NI

END OF COMPILATION:       NO  DIAGNOSTICS.
```

```
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      EXTERNAL F1
      LOGICAL REL
      COMMON N,K
      N=15
      K=6
      DIMENSION A(2)
      B=9./(N-K-2.)**2
      K2=K-2
      DO 1 IR=1,K2
    1 B=B+(-1)**IR*(FACT(K2)/(FACT(IR)*FACT(K2-IR)))*
     *(4./(N-K-2+2*IR)**2)
      A(1)=0
      A(2)=9
      E=1.D-6
      MAXIT=200
      REL=.TRUE.
      DO 3 I=1,200
      A(2)=A(2)+1.D-1
      AINT=SIMSNI(F1,A,E,REL,MAXIT,1.D+2)
      F=AINT/(2*B)
      PRINT 6,A(2),F
    6 FORMAT(10X,'F(',D9.4,') =',D20.10)
      GO TO 3
    2 PRINT 4,N,K
    4 FORMAT(' ERROR EN LA INTEGRAL PARA N=',I4,' K=',I4)
    3 CONTINUE
      STOP
      END
```

END OF COMPILATION: NO DIAGNOSTICS.

```
      DOUBLE PRECISION FUNCTION FACT(M)
      COMMON N,K
      IF(M) 1,3,4
    1 PRINT 2,N,K
    2 FORMAT(' ARGUMENTO NEGATIVO PARA UN FACTORIAL ,N=',I4,' K=',I4)
      STOP
    3 FACT=1
      RETURN
    4 FACT=1
      DO 5 I=1,M
    5 FACT=FACT*I
      RETURN
      END
```

END OF COMPILATION: NO DIAGNOSTICS.

```
      DOUBLE PRECISION FUNCTION F1(U,FK)
      IMPLICIT DOUBLE PRECISION(A-H,O-Z)
      COMMON N,K
      F1=(DEXP(-DSQRT(U)*((N-K-2.)/2.)))*(1-DEXP(-DSQRT(U)))**(K-2)
      RETURN
      END
```

END OF COMPILATION: NO DIAGNOSTICS.

4.4 DETECCION DE OUTLIERS.

En el apartado 4.2 se ha obtenido la función de densidad de el cuadrado de la distancia entre la matriz de s.c. y s.p. de una muestra aleatoria de tamaño n extraida de una población $N_2(\mu, \Sigma)$ no singular y la matriz de s. c. y s. p. de n - k de estas observaciones.

Debido a que se pueden extraer de $\binom{n}{k}$ formas diferentes estas n - k observaciones de entre las n de la muestra para cada k fijo variando dentro de las condiciones del teorema 4.2.2.1 , obtendriamos matrices de s. c. y s. p. $S_{(n-k)}$. Por otro lado bajo la hipotesis de que los elementos de la muestra de tamaño n pertenecen a la misma población $N_2(\mu, \Sigma)$, se tendrá que $\exists\, q \in \mathbb{R}$, $q > 0$ tal que:

$$d^2\left[S^{(i)}_{(n-k)}, S_{(n)}\right] \leq q \qquad \forall i \quad i = 1, 2, \ldots, \binom{n}{k}$$

y por tanto

$$\max_i d^2\left[S^{(i)}_{(n-k)}, S_{(n)}\right] \leq q$$

Evidentemente si $\max_i d^2\left[S^{(i)}_{(n-k)}, S_{(n)}\right] > q$ esto se interpretaria afirmando que los n elementos que componen la muestra no pertenecen a la misma población. Por otro lado los k elementos que se han suprimido de la muestra para obtener $S_{(n-k)}$ serian los elementos que no pertenecen a la población $N_2(\mu, \Sigma)$ ya que al añadir estos k elementos a la muestra de tamaño n - k se produce una dispersión que hace que la distancia entre dichas matrices sobrepase a esa cota q . Estos k elementos serán los posibles outliers para esta distribución $N_2(\mu, \Sigma)$

Ya que los estadisticos $\left\{d^2\left(S^{(i)}_{(n-k)}, S_{(n)}\right),\ i = 1, 2, \ldots, \binom{n}{k}\right\}$ no son independientes no podemos calcular la distribución del estadistico

ordenado $\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)})$. Por ello recurriremos a hallar una cota superior para la cola de dicha distribución, es decir acotaremos

$$P\left[\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)}) > q\right]$$

Si definimos E_i como el suceso $\{d^2(S^{(i)}_{(n-k)}, S_{(n)}) > q\}$ para $i = 1, 2, \ldots, \binom{n}{k}$ entonces

$$P\left[\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)}) > q\right] = P\left[\bigcup_{i=1}^{\binom{n}{k}} E_i\right] \leq \sum_{i=1}^{\binom{n}{k}} P(E_i) =$$

$$= \binom{n}{k} P(E_i) = \binom{n}{k} P\left[d^2(S^{(i)}_{(n-k)}, S_{(n)}) > q\right]$$

Sin embargo nos interesa trabajar con la variable aleatoria X con función de densidad la dada en el teorema 4.3.2. , cuya distribución que es dominada por la del estadistico $d^2(S_{(n-k)}, S_{(n)})$, hemos tabulado en el apartado 4.3..

Ya que

$$P[X \leq q] \leq P\left[d^2(S_{(n-k)}, S_{(n)}) \leq q\right] \quad \forall q \quad 0 \leq q \leq \infty$$

se deduce

$$P[X > q] \geq P\left[d^2(S_{(n-k)}, S_{(n)}) > q\right]$$

y en definitiva

$$P\left[\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)}) > q_\alpha\right] \leq \binom{n}{k} P\left[d^2(S_{(n-k)}, S_{(n)}) > q_\alpha\right] \leq$$

$$\leq \binom{n}{k} P[X > q_\alpha] = \alpha.$$

En definitiva el proceso a seguir seria el siguiente:

Dado un nivel de significación α se determinaria el cuantil q_α mediante la tabulación que se ha hecho de la distribución de la variable aleatoria X.

A continuación calcularíamos el valor del estadistico $\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)})$ y la regla de decisión seria la siguiente.

Si

$$\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)}) > q_\alpha \quad i = 1, 2, \ldots, \binom{n}{k}$$

los n elementos que componen la muestra no pertenecen a la misma población , siendo considerados como outliers los k elementos que se han suprimido de la muestra para obtener $S_{(n-k)}$.

En las paginas siguientes incluimos un programa de ordenador para calcular el valor del estadistico ordenado $\max_i d^2(S^{(i)}_{(n-k)}, S_{(n)})$ y obtener los k elementos de la muestra excluidos en el calculo de dicho estadistico y que determinariamos si son outliers mediante la regla de decisión dada anteriormente.

Calculo del estadistico maxima distancia y detección de los posibles outliers.

```
      PARAMETER N=15,N4=N-4,N3=N-3,N2=N-2,N1=N-1
      DIMENSION DATOS(2,N),C(N,N),PR(2,N),S(2,2),SINV(2,2),VI(2,2),
     *X1(2,N4),CN4(N4,N4),PR2(2,N4),S2(2,2),A(2,2)
C     ------------------------------------------------------------------
C LECTURA DE UN DATO BIDIMENSIONAL EN CADA TARJETA E IMPRESION
C     ------------------------------------------------------------------
      READ 1,((DATOS(I,J),I=1,2),J=1,N)
      PRINT1,((DATOS(I,J),I=1,2),J=1,N)
    1 FORMAT(2F20.2)
      PRINT 23
   23 FORMAT(/,/)
      D1=0
      J1=0
C     ------------------------------------------------------------------
C CALCULO DE LA MATRIZ DE SUMA DE CUADRADOS Y PRODUCTOS DE N OBSERVACIONES SN
C (EN EL PROGRAMA, LA MATRIZ SN ES S)
C     ------------------------------------------------------------------
      DO 2 I=1,N
      DO 2 J=1,N
    2 C(I,J)=-1./N
      DO 3 I=1,N
    3 C(I,I)=C(I,I)+1
      DO 4 I=1,2
      DO 4 J=1,N
      PR(I,J)=0
      DO 4 K=1,N
    4 PR(I,J)=PR(I,J)+DATOS(I,K)*C(K,J)
      DO 5 I=1,2
      DO 5 J=1,2
      S(I,J)=0
      DO 5 K=1,N
    5 S(I,J)=S(I,J)+PR(I,K)*DATOS(J,K)
C     ------------------------------------------------------------------
C IMPRESION DE LA MATRIZ SN
C     ------------------------------------------------------------------
      PRINT 301
  301 FORMAT(' MATRIZ SN',/)
      PRINT300,((S(I,J),I=1,2),J=1,2)
  300 FORMAT(2E18.9)
      PRINT 23
C     ------------------------------------------------------------------
C COMPROBACION DE QUE LA MATRIZ SN ES INVERTIBLE Y SIMETRICA, E IMPRESION ERROR
C     ------------------------------------------------------------------
      DET=S(1,1)*S(2,2)-S(1,2)*S(2,1)
      IF((S(1,2)).EQ.(S(2,1))) GOTO 200
      PRINT201, S(1,2),S(2,1)
  201 FORMAT(' S12=',E16.9,' S21=',E18.9)
  200 IF(DET.EQ.0) STOP1
C     ------------------------------------------------------------------
C CALCULO DE LA MATRIZ SN(-1), INVERSA DE SN, E IMPRESION
C (EN EL PROGRAMA, SN(-1) ES SINV)
C     ------------------------------------------------------------------
      SINV(1,1)=S(2,2)/DET
      SINV(2,2)=S(1,1)/DET
      SINV(1,2)=-S(1,2)/DET
      SINV(2,1)=-S(2,1)/DET
      PRINT 302
  302 FORMAT(' MATRIZ SN(-1)',/)
      PRINT300,((SINV(I,J),I=1,2),J=1,2)
      PRINT 23
C     ------------------------------------------------------------------
C ENUMERACION UNO A UNO DE TODAS LAS POSIBLES CUATERNAS
C     ------------------------------------------------------------------
      DO 6   I1=1,N3
      I11=I1+1
      DO 6   I2=I11,N2
      I21=I2+1
      DO 6 I3=I21,N1
      I31=I3+1
      DO 6 I4=I31,N
      IO=0
      DO 9 J=1,N
      IF((J-I1)*(J-I2)*(J-I3)*(J-I4)                .EQ.0) GOTO7
      DO 8 M=1,2
    8 X1(M,J-IO)=DATOS(M,J)
      GOTO 9
    7 IO=IO+1
    9 CONTINUE
C     ------------------------------------------------------------------
```

```
C CALCULO DE LA MATRIZ DE SUMA DE CUADRADOS Y PRODUCTOS DE N-4 OBSERVACIONES SN4
C  (EN EL PROGRAMA LA MATRIZ SN4 ES S2 )
C     ------------------------------------------------------------------
      DO 11 I=1,N4
      DO 11 J=1,N4
   11 CN4(I,J)=-1./N4
      DO 12 I=1,N4
   12 CN4(I,I)=CN4(I,I)+1
      DO 13 I=1,2
      DO 13 J=1,N4
      PR2(I,J)=0
      DO 13 K=1,N4
   13 PR2(I,J)=PR2(I,J)+X1(I,K)*CN4(K,J)
      DO 14 I=1,2
      DO 14 J=1,2
      S2(I,J)=0
      DO 14 K=1,N4
   14 S2(I,J)=S2(I,J)+PR2(I,K)*X1(J,K)
C     ------------------------------------------------------------------
C CALCULO DE LA MATRIZ  SN4*SN(-1)
C (EN EL PROGRAMA, LA MATRIZ PRODUCTO SN4*SN(-1) ES A)
C     ------------------------------------------------------------------
      DO 15 I=1,2
      DO 15 J=1,2
      A(I,J)=0
      DO 15 K=1,2
   15 A(I,J)=A(I,J)+S2(I,K)*SINV(K,J)
C     ------------------------------------------------------------------
CALCULO DE LOS AUTOVALORES
C     ------------------------------------------------------------------
      RAD=SQRT((A(1,1)-A(2,2))**2+4.*A(2,1)*A(1,2))
      Y1=(A(1,1)+A(2,2)+RAD)/2.
      Y2=(A(1,1)+A(2,2)-RAD)/2.
C     ------------------------------------------------------------------
C COMPROBACION DE QUE LOS AUTOVALORES ESTAN EN (0,1)
C     ------------------------------------------------------------------
      IF((0.LT.Y1).AND.(Y1.LT.1)) GOTO 16
      PRINT 17,Y1,I1,I2,I3,I4
   17 FORMAT(' Y1 NO ESTA EN (0,1).      Y1=',F15.9,5X,4I9,/)
      GOTO 60
   16 IF((0.LT.Y2).AND.(Y2.LT.1)) GOTO 18
      PRINT 19,Y2,I1,I2,I3,I4
   19 FORMAT(' Y2 NO ESTA EN (0,1).      Y2=',F15.9,5X,4I9,/)
      GOTO 60
C     ------------------------------------------------------------------
C CALCULO DEL ESTADISTICO MAXIMA DISTANCIA E IMPRESION DE LOS DATOS EXCLUIDOS
C(POSIBLES OUTLIERS)
C J1 ES EL CONTADOR DEL NUMERO DE COMPARACIONES
C     ------------------------------------------------------------------
   18 D2=((ALOG(Y1))**2+(ALOG(Y2))**2)
      IF(D2.LT.D1) GOTO 60
      I111=I1
      I222=I2
      I333=I3
      I444=I4
      D1=D2
   60 J1=J1+1
      IF(((J1/1000.)-INT(J1/1000)).GT.(0.000001)) GOTO 80
      PRINT 80,J1,I1,I2,I3,I4,      D1,I111,I222,I333,I444
   80 FORMAT(' ITERACION',I7,' EN',4I3,20X,' MAX. DISTANCIA PARCIAL =',
     *F15.9,' EN',4I3,/)
    6 CONTINUE
      PRINT 20,D1
```

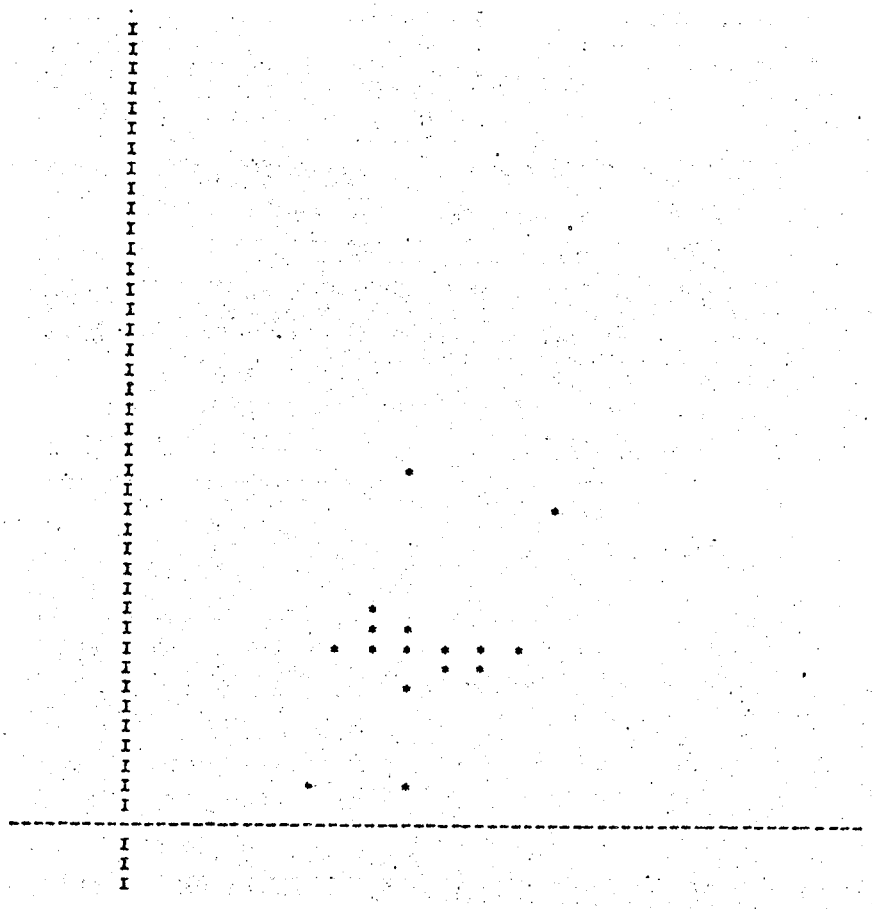
```
      20 FORMAT( ' MAXIMA DISTANCIA ='F15.9,/)
         PRINT 21, DATOS(1,I111),DATOS(1,I222),DATOS(1,I333),DATOS(1,I444)
         PRINT 22, DATOS(2,I111),DATOS(2,I222),DATOS(2,I333),DATOS(2,I444)
      21 FORMAT( ' DATOS EXCLUIDOS EN EL MAXIMO' 10X,4F15.9)
      22 FORMAT(39X,4F15.9)
C     ------------------------------------------------------------------
C GRAFICA DE LOS PUNTOS
C     ------------------------------------------------------------------
         PRINT 103
     103 FORMAT(1H1)
       * DO 101 I=1,61
         DO 105 J=1,121
     105 V(J)=' '
         V(61)='I'
         IF(I.NE.41) GOTO 106
         DO 107 J=1,121
     107 V(J)='-'
     106 DO 102 J=1,N
         IF((DATOS(2,J).GE.(20.25-.5*I)).AND.(DATOS(2,J).LT.(20.75-.5*I)))
        * GOTO 104
         GOTO 102
     104 J1=(10*DATOS(1,J))*121/20
         V(J1)='*'
     102 CONTINUE
         PRINT 108,V
     108 FORMAT(1X,121A1)
     101 CONTINUE
         STOP
         END

END OF COMPILATION:        NO  DIAGNOSTICS.
```

4.5. APLICACION.

Se ha extraido una muestra de tamaño quince de las calificaciones obtenidas en teoria y problemas en una asignatura de tercer curso de la licenciatura de Matematicas, con los siguientes resultados.

5.00	4.00
6.00	8.00
3.00	4.50
4.00	3.50
3.50	5.50
4.00	9.00
4.50	4.50
4.00	5.00
5.50	4.50
4.00	4.50
4.50	4.00
4.00	1.00
3.50	4.50
5.00	4.50
3.50	5.00

Representando en el eje de abscisas las calificaciones de teoria y en el de ordenadas las de problemas, la nube de puntos que se obtiene es de la forma



Se quieren detectar las calificaciones que destacan de entre las de la muestra, es decir las observaciones outliers aplicando para ello el metodo de la distancia entre las matrices de sumas de cuadrados y sumas de productos de observaciones muestrales desarrollado en el apartado anterior.

Mediante el programa dado en las pags. 93 a 95 , para k = 3 se obtiene:

MAXIMA DISTANCIA = 8.6990

DATOS EXCLUIDOS EN EL MAXIMO:

6.00	8.00
4.00	9.00
4.00	1.00

Para un nivel de significación α = 0.01 el percentil de la distribución del estadistico $\max_i d^2(S_{(n)}, S^{(i)}_{(n-k)})$, que viene dado mediante el programa de ordenador de las págs. 88 y 89 es 8.5.

Como la maxima distancia experimental es mayor que este valor critico, se tendrá que para los datos excluidos en el maximo, se rechaza la hipotesis de que las tres observaciones pertenecen a la misma poblaciòn y por tanto serian consideradas outliers.

A continuación comprobamos si pueden existir cuatro observaciones outliers. Mediante el programa de ordenador dado en las paginas 93 , 94 y 95 , obtenemos.

MAXIMA DISTANCIA = 12.6026

DATOS EXCLUIDOS EN EL MAXIMO

6.00	8.00
4.00	9.00
4.00	1.00
4.00	3.50

Para un nivel de significación α = 0.01 el valor teõrico del estadistico es 13.6 y como el experimental es menor rechazamos la hipotesis de que esas cuatro observaciones son outliers.

REFERENCIAS BIBLIOGRAFICAS.

ABRAMOWITZ M. and STEGUN I. A. (1.972). Handbook of mathematical functions. Dover.

AITKEN A. C. (1.965). Determinantes y Matrices. Dossat.

ANDERSON T. W. (1.958). An Introduction to Multivariate Statistical Analysis. Wiley.

ANDERSON T. W. and BAHADUR R.R. (1.962). "Classification into two multivariate normal distributions with different covariance matrices". Annals of Mathematical Statistics (33).

ANDREWS D. F. (1.971). "Significance tests based on residuals" Biometrika (58).

ANDREWS D. F. (1.972). "Plots of high-dimensional data". Biometric (28).

ANSCOMBE F. J. (1.960). "Rejection of outliers". Technometrics (2).

ANSCOMBE F. J. and TUKEY J. W. (1.963). "The examination and analysis of residuals". Technometrics (5).

BARNETT V. and LEWIS T. (L.978). Outliers in Statistical Data. Wiley.

BELLMAN R. (1.965). Introduccion al Analisis Matricial. Reverté.

BROWN B. M. (1.975). "A short-cut test for outliers using residuals". Biometrika (62).

CHERNOFF H. (1.972). "The selection of effective atributes for deciding between hypotheses using linear discriminant functions". In Frontiers of Pattern Recognition, Watanabe (Ed.). Academic Press.

CHERNOFF H. (1.973). "Some measure for discriminating between normal multivariate distributions with unequal covariance matrices". In Multivariate Analysis III, Krishnaiah (Ed.). Academic Press.

COLLET D. (1.980). "Outliers in circular data". Applied Statisque (29).

COX D. R. and SNELL E. J. (1.968). "A general definition of residuals". Journal of the Royal Statistical Society B (30).

DAVID H. A., HARTLEY H. O. and PEARSON E. S. (1.954). "The Distribution of the ratio, in a single normal sample of range of standard deviation". Biometrika (41)

DE RAULY D. (1.966). L'estimation Statistique. Gauthier Villars.

DEVLIN S. J.; GNANADESIKAN R. and KETTERING J. R. (1.975). "Robust estimation and outlier detection with correlation coefficients". Biometrika (62)

DIXON W. J. (1.950). "Analysis of extreme values". Annals of Mathematical Statistics (21).

FERGUSON T. S. (1.961). "On the rejection of outliers". Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol I.

FISHER R. A. (1.936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics (7).

FISHER R. A. (1.939). "The sampling distribution of some statistics obtained from non-linear equations". Annals of Eugenics (9)

GIRI N. C. (1.977). Multivariate Statistical Inference. Academic Press.

GNANADESIKAN R. (1.973). "Graphical methods for informal inference in multivariate data analysis". Bull. Int. Statist. Inst. (45).

GNANADESIKAN R. (1.977). Methods for Statistical Data Analysis of Multivariate Observations. Wiley.

GNANADESIKAN R. and KETTERING J. R. (1.972). "Robust estimates residuals, and outlier detection with multiresponse data". Biometrics (28)

GRUBBS F. E. (1.950). "Sample criteria for testing outlying observations". Annals of Mathematical Statistics (21)

GRUBBS F. E. (1.969). "Procedures for detecting outlyng observations in samples". Technometrics (11).

GUMBELL E. J. (1.960). "Discussion on rejection of outliers by Anscombe". Technometrics (2).

GUPTA S. S. (1.960). "Order Statistics from the Gamma distribution". Technometrics (2).

HAMPEL F. R. (1.974). "The influence curve and its role in robust estimation". Journal of the the American Statistical Association (69).

HAWKINS D. M. (1.974). "The detectionof errors in multivariate data using principal components". Journal of the American Statistical Association (69).

HSU P. L. (1.939). "On the distribution of roots of certain determinantal equations". Annals of Eugenics (9).

IRWIN J. O. (1.925). "On a criterion for the rejection of outlyng observations". Biometrika (17).

JOHNSON N. L. and KOTZ S. (1.969). Distributions in Statistics; Discrete Distributions. Wiley.

JOHNSON N. L. and KOTZ S. (1.970). Distributions in Statistics: Continous univariate distributions, I y II. Wiley

JOHNSON N. L. and KOTZ S. (1.972). Distributions in Statistics: Continous Multivariate Distributions. Wiley.

KENDALL M. G. and BUCKLAND W. R. (1.957). A Dictionary of Statistical Terms. Longman.

KOBAYASHI S. and NOMIZU K. (1.963). Foundation of Differential Geometry, Vol. I. Wiley Interscience.

KRISHNAIAH P. R. (1.978). "Some recent developments on real multivariate distributions". In Developments in Statistics, Krishnaiah (Ed.). Academic Press.

KSHIRSAGAR A. M. (1.972). Multivariate Analysis. Marcel Dekker.

KUDO A. (1.956). "On the testing of outlying observations". Sankhya (17).

MAAS H. (1.955). "Die bestimmung der dirichletreihen mit grosse sencharakteren zu den modulformen n-ten grades". J. Indian Math. Soc. (19).

NAIR K. R. (1.948). "The distribution of the extreme deviate from the sample mean and its studentized form". Biometrika (35).

PATEL J. K.; KAPADIA C. H. and OWEN D. B. (1.976). Handbook of Statistical Distributions. Marcel Dekker.

PATNAIK P. B. (1.949). "The non-central χ^2 and F-distributions and their applications". Biometrika (36).

PEARSON E. S. and CHANDRASEKAR C. (1.936). "The efficiency of statistical tools and a criterion for the rejection of outlying observations". Biometrika (28).

PRESS S. J. (1.966)."Linear combinations of non central chi square variates". Annals of Mathematical Statistics (37).

ROHATGI V. K. (1.976). An Introduction to Probability Theory and Mathematical Statistics. Wiley.

ROSNER B. (1.975). "On the detection of many outliers". Technometrics (11).

ROY S. N. (1.939). "P-statistics or some generalizations in a-analysis of variance appropiate to multivariate problems". Shankya (4).

SHAH S. M. (1.966)."On estimating the parameter of a doubly truncated binomial distribution". Journal of the American Statistical Association (61).

SHAPIRO S. S. and WILK M. B. (1.972). "An analysis of variance test for the exponential distribution". Technometrics (14).

SIOTANI M. (1.959). "The extreme value of the generalised distances of the individual points in the multivariate normal sample". Ann. Inst. Statist. Math. Tokyo (10).

STUDENT (1.927). "Errors of routine analysis". Biometrika (19).

TIAO G. C. and GUTTMAN I. (1.967). "Analysis of outliers with adjusted residuals". Technometrics (9).

TIETJEN G. L. and MOORE R. H. (1.972). "Some Grubbs-Type Statistics for the detection of several outliers". Technometrics (14).

THOMPSON W.R. (1.935). "On a criterion for the rejection of observations and the distribution of the ratio of the desviation to the sample standard desviation", Annals of Mathematical Statistics (6).

WATSON G. W. (1.964). "A note on maximun likelihood". Shankya (26

WILK M. B. GNANADESIKAN R. and HUYETT J. (1.962). "Probability plots for the gamma distribution". Technometrics

WILKS S. S. (1.962). Mathematical Statistics. Wiley.

WILKS S. S. (1.963). "Multivariate Statistical outliers". Sankhya A, (25).