UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMATICAS

"IDENTIFICACION DE OUTLIERS EN MUESTRAS MULTIVARIANTES"

Josè Luis Pèrez Diez de los Rios

Memoria dirigida por:

Prof. Dr. D. Antonio Pascual Acosta

Prof. Dr. D. Joaquin Muñoz García

UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMATICAS

"IDENTIFICACION DE OUTLIERS EN MUESTRAS MULTIVARIANTES"

por

JOSE LUIS PEREZ DIEZ DE LOS RIOS

Visado en Sevilla a

15 de Junio de 1987

Fdo.: Prof. Dr. D. Antonio Pascual Acosta

Fdo.: Prof. Dr. D. Joaquin Muñoz García

Memoria presentada para aspirar al grado de Doctor en Ciencias Matemáticas

Sevilla, Junio 1987

Josè Luis Pèrez Diez de los Rios

CONTENIDO

INTRODUCCION

INTRODUCCION.

En el análisis estadístico de datos, es frecuente que el investigador se encuentre con observaciones que parecen ser inconsistentes con el resto de los datos. Estas observaciones anómalas, inconsistentes, o que parecen separarse de la masa principal de datos se denominan outliers. De una forma intuitiva, se puede decir que un outlier es una observación que se desvia del resto de las observaciones.

En todo análisis estadístico, para evitar el riesgo de perturbación por la presencia de observaciones outliers, es necesario la utilización de tècnicas estadísticas que se encuentren protegidas frente a este tipo de observaciones (mètodos robustos) o bien proceder antes del estudio estadístico a la identificación de aquellas.

Para evitar los posibles errores que pueden generar las observaciones outliers en las investigaciones que se realicen, a partir de una masa de datos, existen dos tipos de tècnicas: las tècnicas de acomodación, que tienen por objeto la construcción de mètodos estadisticos robustos que no se dejen influenciar por la presencia de outliers en las muestras, y las tècnicas de identificación que tienen por objeto la construcción de mètodos estadísticos que permitan la determinación e identificación de los posibles outliers en el conjunto de las observaciones.

Aunque ambas tècnicas están directamente orientadas a resolver el tipo de problema a que da lugar la presencia de outliers en una muestra, sin embargo su naturaleza, contenido metodológico y aplicación práctica son muy diferentes.

Esta memoria está dedicada al estudio e investigación de problemas y mètodos relacionados con el segundo grupo de tales tècnicas: las tècnicas de identificación. Los primeros indicios sobre estas tècnicas datan de mediados del siglo XVIII y hasta mediados del siglo XX estaban dedicadas exclusivamnete a la identificación de outliers en muestras univariantes. Es a partir de entonces cuando aparecen los primeros mètodos sobre identificación de outliers en muestras multivariantes, siendo aún hoy en dia muy escasos los mètodos que resuelven este problema de identificación de outliers en muestras multidimensionales.

Asi, Gnanadesikan y Kettering (1972) señalan, que debido a la complejidad del caso multivariante resulta infructuoso buscar procedimientos de detección de outliers que sean válidos para todas las situaciones, dado que un outlier para una cierta situación puede no serlo para otra, por lo que es más aconsejable construir un gran conjunto de tècnicas con diferentes sensibilidades.

En el primer capitulo, se hace una breve reseña histórica de la evolución del tèrmino outlier desde sus inicios hasta nuestros dias, indicando además la dificultad que entraña en el caso multivariante detectar a priori que determinadas observaciones puedan ser outliers. Se destacan, a continuación, las diferencias entre tècnicas de acomodación y tècnicas de identificación, y se plantea el problema de la identificación de outliers como un contraste de hipótesis, establecièndose la hipótesis nula, y los distintos tipos de hipótesis alternativas que se pueden presentar. Asimismo se lleva a cabo una revisión de las tècnicas existentes para la identificación de outliers en muestras multivariantes, poniendose de manifiesto que la mayoría de ellas son generalizaciones de mètodos o ideas desarrolladas para el caso univariante o, tècnicas basadas en representaciones gráficas.

Se aborda en el primer apartado del capitulo segundo el denominado efecto de enmascaramiento que se puede presentar cuando en la muestra existe más de una observación outlier, efecto que se evita aplicando de forma secuencial aquellas tècnicas que vayan identificando bloques de outliers simultaneamente.

Se propone un mètodo para la identificación de outliers en muestras multivariantes basado en una R-ordenación de las observaciones muestrales, a partir de sus distancias al vector de medias, siendo esta distancia la inducida por una norma vectorial.

Este procedimiento general, se ha particularizado para el caso de la 1-norma, tabulando la distribución en el muestreo del estadístico resultante mediante simulación, dando lugar a las tablas de valores criticos del test de hipótesis. Este procedimiento se podria llevar a cabo de la misma forma con cualquier otra norma vectorial.

Se completa este capítulo con una subrutina para la aplicación del mètodo propuesto, donde se muestra a su vez el algoritmo de aplicación de dicha tècnica, y se incluye un caso práctico.

El último capítulo de esta memoria aborda el problema de la detección e identificación de outliers en muestras multivariantes desde una perspectiva diferente a la utilizada en el capítulo anterior, para llegar a un mètodo o tècnica que no tiene antecedentes en el caso univariante.

Comienza el capítulo definiendo una distancia entre matrices, que se aplica al caso de matrices de su mas de cuadrados y sumas de productos de observaciones muestrales multidimensionales, que permite construir el estadístico básico para el mètodo de identificación. A continuación se procede a determinar la distribución de dicho estadístico.

Por último se plantea el contraste de hipótesis tanto para el caso de que se trate de identificar un único outlier, como para la identificación de más de un outlier, especificando en ambos la región crítica correspondiente.

1.- EL PROBLEMA DE LAS OBSERVACIONES OUTLIERS EN MUESTRAS MULTIVARIANTES.

1.1. El tèrmino outlier. Reseña histórica

1.2. Tècnicas para el tratamiento de los outliers.

1.3. Tècnicas de identificación de outliers en muestras multivariantes.

1.1. EL TERMINO OUTLIER. RESEÑA HISTORICA.

Cuando se trabaja con observaciones muestrales, no se puede garantizar que todas las observaciones sean totalmente manifestaciones del fenómeno bajo estudio. De una forma intuitiva, la fiabilidad de una observación se refleja por su relación con la otras observaciones que se obtuvieron bajo condiciones similares. Así, cualquier investigador que trabaje con datos reales se suele encontrar con observaciones o grupos de observaciones que, en su opinión, parecen ser inconsistentes con el resto de los datos, por ser valores muy pequeños o muy grandes en comparación con el resto de los datos. Estas observaciones anómalas, inconsistentes, o que parecen separarse de la masa principal de datos han sido denominadas "outliers", "observaciones discordantes", "valores sorprendentes", "observaciones contaminantes", "valores atípicos", etc. por mencionar solo algunos de los têrminos que se han utilizado para identificar tales observaciones a lo largo de los años.

De una forma intuitiva, se puede decir que un outlier es una observación que se desvía, en algún sentido, del resto de las observaciones, lo que induce a sospechar que fue generada por un mecanismo diferente al resto de los datos.

El interès por estas observaciones anómalas data de los primeros intentos de obtener conclusiones a partir de un conjunto de datos estadísticos. Los primeros indicios se deben a Boscovich (1755) y Bernoulli (1777), lo que indica que la práctica de descartar las observaciones que parecen discordantes se realizaba ya hace mas de 200 años. En el siglo XIX aparecen los primeros trabajos orientados a desarrollar mètodos estadísticos "objetivos" para tratar con los outliers, como son los de Peirce (1852), Stone (1868), Glaisher (1872), Edgeworth (1887), etc., y a pesar de lo mucho que se ha escrito sobre el tema, el concepto de outlier sigue siendo tan vago hoy en dia como lo fue 200 años atrás. Así, Edgeworth (1887) escribía:

"Las observaciones discordantes se pueden definir como aquellas que parecen diferir, con respecto a su ley de frecuencias, de las otras observaciones con las que estan combinadas".

Ochenta y dos años despues, Grubbs (1969) establece que:

"Un outlier es una observación que parece desviarse marcadamente de los otros elementos de la muestra en la cual se encuentra".

En la mayoría de los trabajos publicados hasta la fecha han aparecido definiciones del tèrmino outlier, siendo todas tan vagas como las dos anteriores:

"Los outliers son observaciones que parecen ser demasiado grandes o demasiado pequeñas, comparadas con el resto de las observaciones" (Gumbel, 1960).

"En una muestra extraida de una cierta población aparecen una o varias observaciones que, sorprendentemente, se encuentran alejadas del grupo principal de datos" (Ferguson, 1961).

"Los outliers son observaciones que tienen residuos tan grandes, en comparación con las otras, que sugieren que deben tener un tratamiento especial, siendo el residuo de una observación la diferencia entre el valor observado y el valor ajustado" (Anscombe y Tukey, 1963).

"Los outliers son observaciones o conjuntos de observaciones que parecen ser inconsistentes con el resto de la muestra" (Barnett y Lewis, 1984).

Beckman y Cook (1983) distinguen entre observación discordante: "cualquier observación que el investigador considera sorprendente o discrepante"; y observación contaminante: "cualquier observación que no es una realización de la distribución en estudio". Denominan outlier a un colectivo que alude a observaciones contaminantes o discordantes.

A la vista de estas definiciones, se puede decir que el concepto de outlier es un concepto subjetivo despues de observar los datos. Historicamente, los mètodos estadísticos "objetivos" para tratar con outliers se emplearon despues de identificar los outliers a travès de una inspección visual de los datos.

Collet y Lewis (1976) realizan un informe sobre los resultados de un experimento para investigar la naturaleza subjetiva de la decisión de designar una observación como outlier, concluyendo que la percepción de un outlier depende de la forma de presentación de los datos (al azar, graficamente, ordenados), de la experiencia del investigador, y de la escala utilizada para la presentación de los mismos: conforme aumenta la escala, las observaciones extremas tienden a parecer más discrepantes.

La mayor parte de lo dicho anteriormente se podría aplicar a muestras aleatorias de tamaño pequeño o moderado. En grandes conjuntos de datos, muestras multivariantes, análisis de regresión, etc., una impresión visual de los datos puede resultar imposible.

Como Gnanadesikan y Kettering (1972) hacen notar, los outliers en muestras multivariantes no tienen una manifestación tan clara como observaciones "que se alejan de los límites de la muestra". En datos bivariantes se pueden percibir observaciones sospechosas si se realiza un diagrama de dispersión, observando aquellos elementos muestrales que caen fuera del conjunto principal de observaciones, pero en muestras de datos con más de dos dimensiones no se perciben tan facilmente, al no poder realizar tales representaciones gráficas.

La idea de alejamiento, inevitablemente, lleva asociada la consideración de una distancia y a su vez alguna forma de ordenación de las observaciones. En muestras multivariantes, a lo sumo, se podria conseguir el definir algun principio de subordenación de los elementos muestrales, que permitiera esa percepción visual de los datos sorprendentes. Barnett (1976) estudia distintas formas de subordenación y el papel que juegan en el Análisis Multivariante.

Así puès, se hace entonces necesario aplicar algún tipo de criterio objetivo para la identificación de los outliers, en lugar de una inspección visual de los datos.

1.2. TECNICAS PARA EL TRATAMIENTO DE LOS OUTLIERS.

Como se puede deducir de las definiciones del tèrmino outlier expuestas en el apartado anterior, los outliers son observaciones que parecen haber sido producidas por un fenómeno o mecanismo aleatorio diferente al que generó el resto de las mismas. Parece evidente que si no se tiene en cuenta la posible presencia de outliers en las muestras recogidas para la investigación o estudio de un fenómeno aleatorio, las conclusiones que se deriven de tal investigación pueden ser erróneas.

Para subsanar los posibles errores que puede generar la presencia de outliers, existen diversas tècnicas que se pueden integrar en dos grandes grupos:

- Tècnicas de acomodación
- Tècnicas de identificación

Las tècnicas de acomodación consisten en construir mètodos estadísticos que no se dejen influenciar en sus resultados por la presencia de outliers en las muestras. Por contra, las tècnicas de identificación proporcionan mètodos estadísticos que permiten determinar e identificar los posibles outliers en una muestra.

- Tècnicas de acomodación.

Las tècnicas de acomodación incluyen los mètodos robustos de estimación de los parámetros desconocidos de la distribución teórica de la población, así como los mètodos robustos de contraste de hipótesis relativos a estos parámetros. El interès de estos mètodos robustos se debe a que los estimadores y funciones test que se obtienen mantienen determinadas propiedades estadísticas bajo diferentes distribuciones poblacionales.

En el caso multivariante existen pocas tècnicas de acomodación de outliers, y las que existen, generalmente tienen una justificación más intuitiva que teórica. Entre los distintos trabajos sobre acomodación de outliers en muestras multivariantes se pueden destacar los siguientes.

Golub, Guttman y Dutter (1973) consideran que el vector de observaciones $x=(x_1,x_2,\dots,x_n)'$ esta descrito por el modelo lineal general normal

$$x = Ab+e$$

siendo b un vector de q parámetros, A una matriz nxq de coeficientes conocidos (de rango total) y e un vector de residuos de dimensión n, que se supone distribuido según una ley Normal de vector de medias cero y matriz de varianzas-covariazas de la forma $\sigma^2 I_n$, con I_n la matriz identidad de orden n. Estos autores proponen reglas basadas en la Winsorización de los residuos.

Gnanadesikan y Kettenring (1972) tratan sobre la estimacion robusta de medidas de localización y dispersión en modelos multidimensionales. En particular tratan sobre la estimación robusta del vector de medias y la matriz de varianzas-covarianzas. Proponen varios estimadores robustos del vector de medias, siendo la mayoría de ellos el vector formado por las estimaciones robustas de las distintas componentes del vector de medias, tales como el vector de medianas muestrales, medias Winsorizadas, etc. Para la estimación robusta de la matriz de varianzas-covarianzas, sugieren en primer lugar R-ordenar la muestra multivariante $x_1, x_2, \ldots, x_n$ en tèrminos de su distancia euclídea a un estimador robusto del vector de medias x^*,

$$(x_i - x^*)'(x_i - x^*) \qquad i=1,2,\ldots,n$$

seleccionando a continuación un subconjunto de observaciones cuyas distancias a x^* sean las menores, y con este subconjunto se calcula la matriz

$$A_o = \Sigma \ (x_i - x^*)(x_i - x^*)'$$

Se R-ordena nuevamente la muestra completa pero ahora en tèrminos de la forma cuadrática

$$(x_i - x^*)' A^{-1} (x_i - x^*) \qquad i=1,2,\ldots,n$$

y se eliminan aquellas observaciones que corresponden a los mayores valores de la forma cuadrática, siendo entonces la estimación robusta de la matriz de varianzas-covarianzas una matriz proporcional a la matriz de sumas de cuadrados y sumas de productos de las restantes observaciones.

Devlin, Gnanadesikan y Kettenring (1975), utilizando la relación

$$\mathrm{Cov}(X_1,X_2) = \frac{1}{4}[\mathrm{Var}(X_1+ X_2) - \mathrm{Var}(X_1- X_2)]$$

proponen como estimador robusto de la covarianza

$$S^*_{12} = \frac{1}{4}(\hat{\sigma}^{*2}_1 - \hat{\sigma}^{*2}_2)$$

donde $\hat{\sigma}^{*2}_1$ y $\hat{\sigma}^{*2}_2$ son estimadores robustos de las varianzas de $X_1+ X_2$ y $X_1- X_2$ respectivamente, y a partir de S^*_{12} una forma natural de definir un estimador robusto del coeficiente de correlación entre X_1 y X_2 seria

$$r^*_{12} = \frac{S^*_{12}}{\sqrt{S^*_{11}\, S^*_{22}}}$$

donde S^*_{11} y S^*_{22} son estimadores robustos de las varianzas de X_1 y X_2, respectivamente.

- Tècnicas de identificación.

Como ya se indicó al comienzo de este apartado las tècnicas de identificación tienen un objetivo muy distinto al de las tècnicas de acomodación descritas anteriormente; su objetivo es el determinar e identificar los posibles outliers en una muestra.

Collet y Lewis (1976) indican que la identificación de outliers consta de dos etapas. Una primera etapa subjetiva, en la que el investigador, a la vista de los datos obtenidos, indicará si existen o no observaciones sospechosas de ser outliers. Y una segunda etapa objetiva, en la que mediante el uso de mètodos estadísticos se comprueba si tales observaciones sospechosas se pueden considerar outliers. Esta segunda etapa conduce, generalmente, a la realización de contrastes de hipótesis sobre las observaciones muestrales.

En estos contrastes de hipótesis, la hipótesis nula establece que los elementos de la muestra proceden todos de una misma población, la cual es equivalente a la no existencia de outliers en la muestra.

A pesar de la simplicidad de la hipótesis nula, la formulación de la hipótesis alternativa no es tarea fácil, ya que en ella habría que especificar un modelo que permita explicar la presencia de outliers.

Barnett y Lewis (1984) proponen la siguiente clasificación de las hipótesis alternativas:

- Alternativa determinística
- Alternativa inherente
- Alternativa de mixtura
- Alternativa de deslizamiento

La alternativa determinística se utiliza para cubrir algunos casos de outliers producidos por errores de medida, transcripción, etc. Esto es, cuando en la muestra aparecen una o varias observaciones que claramente resultan ser un error de medida o transcripción. En este caso no es necesaria la realización de ningún contraste estadístico. La hipótesis nula se rechaza en favor de la alternativa que indica que estas observaciones erróneas no se han obtenido de la misma población que el resto de las observaciones.

La alternativa inherente fija un modelo poblacional que explica la masa de datos completa. Así pués, con esta alternativa no se identifican outliers, sino que se determina un modelo probabilístico bajo el cual, en la muestra no existen observaciones outliers.

La alternativa de mixtura establece que los elementos de la muestra provienen de una mixtura de distribuciones, es decir, si la función de distribución de la población bajo la hipótesis nula es F, la alternativa de mixtura establece que la función de distribución de la población es $(1-p)F+pG$, con $p\in(0,1)$. Esto es, cualquier observación proviene de la población descrita por G con probabilidad p (generalmente pequeña).

Por último, la alternativa de deslizamiento es la más común que se utiliza para un modelo generador de outliers. En su forma más usual, la alternativa de deslizamiento establece que todas las observaciones, salvo k de ellas, proceden de una misma población, mientras que las k observaciones restantes proceden de una población modificada en la que uno de los parámetros de localización o escala, ha sido desplazado en su valor.

En la práctica, el fijar alguna de las hipótesis alternativas descritas anteriormente resulta dificil, por lo que las hipótesis que generalmente se contrastan son:

H_0: Todas las observaciones proceden de la misma población

H_1: Existen k observaciones que no proceden de la misma población que el resto $\qquad k=1,2,\dots,[n/2]$

Una vez que se han planteado las hipótesis, la siguiente etapa que aparece en una tècnica de identificación es la construcción de un estadístico que permita aceptar o rechazar la hipótesis nula. En este estadístico siempre intervienen aquellas observaciones que parecen ser outliers, y la mayor dificultad se presenta a la hora de calcular la distribución del estadístico bajo la hipótesis nula, ya que en este estadístico suelen intervenir variables aleatorias dependientes, y hay que recurrir al calculo de distribuciones aproximadas o a la utilización de la simulación para la determinación de los valores críticos del estadístico.

1.3. TECNICAS DE IDENTIFICACION DE OUTLIERS EN MUESTRAS MULTIVARIANTES.

La complejidad de la manipulación de los datos y distribuciones multivariantes ha hecho que en las tècnicas de identificación de outliers se suponga que el modelo básico es el modelo Normal multivariante, debido al hecho de que la gran mayoria de los mètodos multivariantes se basan en esta hipótesis de Normalidad.

Sea

$$X = \begin{bmatrix} x_{11}, x_{12}, \dots, x_{1n} \\ x_{21}, x_{22}, \dots, x_{2n} \\ \dots\dots\dots\dots \\ \dots\dots\dots\dots \\ x_{p1}, x_{p2}, \dots, x_{pn} \end{bmatrix}$$

una muestra de tamaño n de una poblacion $N_p(\mu,\Sigma)$. Representemos por

$$x_i = (x_{1i}, x_{2i}, \dots, x_{pi})' \qquad i=1,2,\dots,n$$

las columnas de la matriz X, por

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p)'$$

el vector de medias muestrales, y por

$$S = \begin{bmatrix} s_{11}, s_{12}, \dots, s_{1p} \\ s_{21}, s_{22}, \dots, s_{2p} \\ \dots\dots\dots\dots \\ \dots\dots\dots\dots \\ s_{p1}, s_{p2}, \dots, s_{pp} \end{bmatrix}$$

la matriz de varianzas-covarianzas de la muestra.

Algunas de las tècnicas propuestas para la identificación de outliers multivariantes se basan en asignar a cada observación un único valor numèrico, y establecer de esta forma una subordenación de las observaciones muestrales. Es lo que Barnett (1976) denomina R-ordenación. Así, Siotani (1959) propone el asignar a cada observación x_i el valor

$$Q_i = (x_i - \bar{x})'\Sigma^{-1}(x_i - \bar{x}) \qquad i=1,2,\dots,n$$

en el caso de que la matriz Σ de varianzas-covarianzas, sea conocida, y en el caso de que dicha matriz sea desconocida, asignarle el valor

$$Q'_i = (x_i - \bar{x})'S^{-1}(x_i - \bar{x}) \qquad i=1,2,\dots,n$$

Una vez asignados estos valores a las observaciones muestrales, propone como estadístico para la identificación de un único outlier los siguientes:

$$Q_{(n)} = \max_{1\le i\le n} Q_i$$

o

$$Q'_{(n)} = \max_{1\le i\le n} Q'_i$$

según que la matriz de varianzas-covarianzas Σ sea conocida o desconocida, respectivamente.

Wilks (1963) propone un mètodo para la identificación de k ($1\le k\le [n/2]$) outliers simultaneamente, en el caso en que tanto el vector de medias como la matriz de varianzas-covarianzas poblacionales son desconocidos

Para la identificación de un único outlier define en primer lugar los estadísticos

$$\mathcal{R}_j = \frac{|A^{(j)}|}{|A|} \qquad j=1,2,\ldots,n$$

donde

$$A_{(j)} = \sum_{\substack{i=1 \\ i\neq j}}^{n} (x_i - x)(x_i - x)' \qquad j=1,2,\ldots,n$$

y

$$A = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'$$

y utiliza como estadístico para la identificación de un outlier

$$r_1 = \min_{1\leq j\leq n} \mathcal{R}_j$$

Así, Wilks también establece una R-ordenación de las observaciones muestrales (Barnett, 1976).

Para la identificación de k outliers (k>1) generaliza este último estadístico considerando el mínimo de los $\binom{n}{k}$ posibles valores de

$$\mathcal{R}^{(I)} = \frac{|A^{(I)}|}{|A|}$$

siendo $I=\{i_1,i_2,...,i_k\}$ un conjunto de índices, y $A^{(I)}$ la matriz de sumas de cuadrados y sumas de productos obtenida al eliminar las k observaciones con subíndice en el conjunto I.

Wilks discute con detalle el problema de la identificación de uno y dos outliers, proporcionando tablas de valores críticos aproximados mediante la desigualdad de Bonferroni, para muestras de tamaño menor o igual que 500, y poblaciones con dimensión de 1 a 5, comparandolo con el caso unidimensional dado por Grubbs (1950). Tambièn discute y estudia el caso de tres o más outliers, pero no proporciona tablas de valores críticos para esta situación.

Estos mètodos propuestos por Siotani (1959) y Wilks (1963) son generalizaciones multivariantes de los trabajos de Thompson (1935), Pearson y Chandra Sekar (1936), Nair (1948) y Grubbs (1950).

Schwager y Margolin (1982) consideran que la matriz de observaciones se puede especificar mediante la ecuación matricial

$$X= \mu E_{1n}+U$$

donde la matriz X tiene n columnas independientes e identicamente distribuidas $x_1,x_2,...,x_n$, μ es el vector de p medias desconocidas, E_{1n} es un vector fila de n elementos todos iguales a 1, y las columnas de la matriz U (pxn) son independientes e identicamente distribuidas según una ley $N_p(0,\Sigma)$, con Σ desconocida. Suponen también que n>p para asegurar que μ y Σ son estimables.

Para incorporar la posibilidad de presencia de outliers, consideran el modelo multivariante de media deslizada

$$X= \mu E_{1n}+\Delta^* A^* + U$$

En este modelo E_{1n}, μ y U tienen la misma definición que anteriormente, n>p, Δ^* es un escalar no negativo y A^* es una matriz pxn tal que:

1.- $\|A^*\| = (\Sigma\, a_{ij})^{½} = 1$, salvo que $\Delta^* = 0$, en cuyo caso $A^*= 0$.

2.- Más de la mitad de las filas de A^* son nulas

En este modelo, que es la generalización multivariante del propuesto por Ferguson (1961), la observación x_i es un outlier si la i-esima fila de A^* es no nula.

Schwager y Margolin obtienen el mejor test localmente invariante para las hipótesis

$$H_0: \Delta^* = 0$$

$$H_1: \Delta^* > 0$$

probando que dicho test es equivalente a rechazar la hipótesis nula cuando el coeficente de curtosis definido por Mardia (1970) es suficientemente grande.

Guttman (1973) propone un tècnica Bayesiana para la identificación de outliers univariantes, la cual extiende en el mismo trabajo al caso multivariante. Como en el caso univariante, se prueba que la media de la distribución a posteriori es una función de pesos que indican observaciones outliers.

Se han propuesto otros mètodos para la identificación de outliers multivariantes que no tienen equivalente en el caso univariante, como son los que se detallan a continuación.

Gnanadesikan (1973) muestra el posible uso de la tècnica de representación gráfica de observaciones multivariantes, debida a Andrews (1972), para la identificación de outliers. Andrews (1972) propone representar para cada observación

$$x_i = (x_{1i}, x_{2i}, \ldots, x_{pi})' \qquad i=1,2,\ldots,n$$

la función

$$f_{x_i}(t)=\frac{1}{\sqrt{2}}x_{1i}+x_{2i}\text{sen } t+x_{3i}\cos t+x_{4i}\text{sen } 2t+\ldots \qquad i=1,\ldots,n$$

en el intervalo $(-\pi,\pi)$. De esta forma se tendrían representados los n puntos p-dimensionales mediante n curvas en el espacio bidimensional. El mètodo de identificación es similar a los restantes procedimientos gráficos.

Hawkins (1974) propone cuatro estadísticos para la identificación de outliers multivariantes, utilizando las componentes principales. Partiendo de una población $N_p(\mu,\Sigma)$ en la que ambos parámetros son conocidos, construye los valores

$$z_i=\frac{w_i}{\sqrt{\lambda_i}} \qquad i=1,2,\ldots,n$$

donde $w=(w_1,w_2,\ldots,w_p)'$ es el vector residuo de las componentes principales y $\lambda_1,\lambda_2,\ldots,\lambda_p$ son los autovalores de Σ, y define los estadísticos

$$T_1=\sum_{i=1}^{p} z_i^2$$

$$T_2=\sum_{i=1}^{k} z_i^2$$

$$T_3=\max_{1\le i\le k}|z_i|$$

$$T_4=\max_{1\le i\le k}|v_i|$$

donde $V = DX$, siendo D la matriz obtenida al realizar una rotación varimax con las primeras k filas de B dejando las restantes filas invariantes. Los elementos de la matriz B son de la forma

$$b_{ij} = \frac{p_{ij}}{\sqrt{\lambda_i}} \qquad i,j=1,2,\ldots,p$$

siendo p_{ij} los elementos de la matriz de vectores propios.

A continuación estudia el caso en que los parámetros poblacionales son desconocidos, estimándolos mediante $\bar{x}$ y S, y haciendo un estudio comparativo de estos estadísticos mediante una aplicación en el campo de la Geología.

Gnanadesikan y Kettering (1972) utilizan medidas univariantes basadas en las observaciones x_j de la forma

$$(x_j - \bar{x})' S^b (x_j - \bar{x}) \qquad j=1,2,\ldots,n$$

y

$$\frac{(x_j - \bar{x})' S^b (x_j - \bar{x})}{(x_j - \bar{x})' (x_j - \bar{x})} \qquad j=1,2,\ldots,n$$

tomando el exponente b los valores -1, 0, 1, y mediante representaciones gráficas de estos valores identifican los outliers, que corresponden a valores extremos de estos estadísticos. Algunos casos particulares de estos estadísticos son:

a) $$q_j^2 = (x_j - \bar{x})'(x_j - \bar{x}) = \frac{n}{n-1}[tr(S) - tr(S^{(j)})]$$

donde el superíndice (j) indica que en el cálculo se ha omitido la observación j. Este estadístico, cuadrado de la distancia euclídea a la media, es sensible a los outliers que afectan a la escala.

b) $$t_j^2 = (x_j - \bar{x})'S(x_j - \bar{x}) = \sum_{i=1}^{p} c_i[l_i'(x_j - \bar{x})]^2$$

donde c_i es el autovalor de S asociado al autovector l_i Esta medida es sensible a los outliers que afectan a la orientación y escala de las primeras componentes principales.

c) $$d_j^2 = (x_j - \bar{x})'S^{-1}(x_j - \bar{x}) = \sum_{i=1}^{p} c_i^{-1}[l_i'(x_j - \bar{x})]^2$$

que se utiliza para detectar observaciones que se alejen del diagrama de dispersión.

d) $$u_j^2 = \frac{(x_j - \bar{x})'S(x_j - \bar{x})}{(x_j - \bar{x})'(x_j - \bar{x})} = \sum_{i=1}^{p} c_i \left[\frac{l_i'(x_j - \bar{x})}{\|x_j - \bar{x}\|}\right]^2$$

donde $\|\cdot\|$ representa la norma euclídea. Este estadístico es similar a t_j^2, excepto que hace más ènfasis en la orientación y menos en la escala.

e) $$v_j^2 = \frac{(x_j - \bar{x})'S^{-1}(x_j - \bar{x})}{(x_j - \bar{x})'(x_j - \bar{x})} = \sum_{i=1}^{p} c_i^{-1} \left[\frac{l_i'(x_j - \bar{x})}{\|x_j - \bar{x}\|}\right]^2$$

que mide las contribuciones relativas de las observaciones en las orientaciones de las últimas componentes principales.

Devlin, Gnanadesikan y Kettering (1975) hacen uso de la función influencia para identificar posibles outliers que afecten a la correlación en datos bivariantes. En una distribución multivariante dependiente de de un parametro θ definen la función influencia muestral de la observación x_j para el estimador $\hat{\theta}$ de θ mediante la relación

$$I_{-}(x_j,\hat{\theta}) = (n-1)(\hat{\theta} - \hat{\theta}_{-j}) \qquad j=1,2,\ldots,n$$

donde $\hat{\theta}_{-j}$ es el estimador de θ obtenido al no considerar la observación x_j en $\hat{\theta}$ y teniendo en cuenta que el tamaño de la muestra se reduce en una unidad. Estos autores aplican esta función influencia a datos bivariantes $x_j = (x_{1j},x_{2j})'$, y al coeficiente de correlación muestral $\hat{\theta} = r$. Para la identificación de los outliers y para ver como afectan al valor del coeficiente de correlación muestral proponen la representación gráfica de las observaciones mediante un diagrama de dispersión, junto con determinadas curvas de nivel de la función influencia muestral. Aquellas observaciones que tienen mayor influencia en el valor del coeficiente de correlación son las consideradas outliers. La influencia de una observación se determina a traves de la curva de nivel que esta más próxima a ella.

2.- TECNICA DE IDENTIFICACION BASADA EN LA R-ORDENACION DE UNA MUESTRA MULTIVARIANTE

2.1. INTRODUCCION.

En la mayoría de los problemas prácticos es posible que en la muestra exista más de una observación sospechosa de ser outlier. En este caso algunos autores proponen la aplicación secuencial de la tècnica específica que se usa para identificar un único outlier.

El procedimiento secuencial consiste en aplicar la tècnica de identificación de un único outlier a la muestra completa, y si en dicha aplicación alguna observación se identifica como outlier, se elimina de la muestra y se vuelve a aplicar la tècnica al conjunto de datos resultante. Asi se continuaría hasta llegar a un conjunto de observaciones en el que la aplicación de la tècnica de identificación indique que no existen más observaciones outliers.

Sin embargo, otros autores no estan de acuerdo con esta tècnica secuencial. Entre estos se pueden destacar a Pearson y Chandra Sekar (1936), Murphy (1951), Mc Millan (1971), Tietjen y Moore (1972), etc. Este desacuerdo con el procedimiento secuencial se debe a que muchas veces las posibles observaciones outliers forman un grupo homogêneo entre sí, y en tal situación las tècnicas de identificación tienden a enmascarar la presencia de dichas observaciones outliers. Esta insensibilidad de los procedimientos secuenciales, que indica la no existencia de outliers cuando en realidad existen, se denomina segun Murphy (1951) "Efecto de enmascaramiento de tipo A".

Un mètodo frecuentemente utilizado para evitar este efecto de enmascaramiento consiste en estimar en primer lugar el nùmero de outliers presentes en la muestra, y a continuación aplicar una tècnica para identificar simultaneamente este conjunto de outliers.

Este procedimiento puede conducir a que aparezca el llamado "Efecto de enmascaramiento de tipo B", efecto que tiene lugar cuando el procedimiento tiende a detectar mayor numero de outliers de los que realmente existen en la muestra, es decir, en este caso se pueden considerar como outliers observaciones que en realidad no lo son.

Un mètodo que evite estos efectos de enmascaramiento consiste en aplicar de forma secuencial, y siempre al conjunto de datos inicial, tècnicas que vayan identificando bloques de outliers simultaneamente.

Tietjen y Moore (1972) estudian el problema de la identificación simultanea de varios outliers, con objeto de eliminar el efecto de enmascaramiento. Para ello proponen dos estadísticos para la identificación de k outliers simultaneamente, basados en el estadístico de Grubbs (1950). En primer lugar consideran el caso más frecuente que es el identificar k outliers en los extremos de una muestra de tamaño n.

Si $x_{(1)}<x_{(2)}< \dots <x_{(n)}$ son los valores muestrales en orden creciente, el estadístico que proponen para contrastar la hipótesis nula de que todas las observaciones proceden de la misma población Normal, frente a la alternativa de que las k observaciones mayores son outliers, viene dado por:

$$L_k=\frac{\sum_{i=1}^{n-k}(x_{(i)}-\bar{x}_k)^2}{\sum_{i=1}^{n}(x_{(i)}-\bar{x}\ )^2}$$

con $\bar{x}$ la media muestral, y

$$\bar{x}_k = \frac{1}{n-k}\sum_{i=1}^{n-k}x_{(i)}$$

la media de las n-k observaciones menores.

La región crítica para este contraste viene dada por

$$L_k < L_{k,\alpha}$$

Para la hipótesis alternativa de que los outliers son las k menores observaciones, el estadístico anterior toma la forma:

$$L_k^* = \frac{\sum_{i=k+1}^{n} (x_{(i)} - \bar{x}_k^*)^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2}$$

con

$$\bar{x}_k^* = \frac{1}{n-k} \sum_{i=k+1}^{n} x_{(i)}$$

la media de las n-k observaciones mayores.

La región crítica para este contraste viene dada por

$$L_k^* < L_{k,\alpha}^*$$

Para k=1,2 se demuestra (Tietjen y Moore,1972) que L_1, L_1^*, L_2 y L_2^* coinciden con los estadísticos propuestos por Grubbs (1950) para la identificación de uno o dos outliers.

Ahora bien, generalmente las observaciones sospechosas de ser outliers no se encuentran en uno de los extremos de la muestra ordenada, sino que, en general, estas observaciones sospechosas de ser outliers se encuentran en ambos extremos de la muestra. En este caso, en el que existen valores muy grandes y valores muy pequeños, los estadísticos L_k y L_k^* no son útiles, y Tietjen y Moore proponen un nuevo estadístico deducido del siguiente razonamiento. Sean $x_1, x_2, \ldots, x_n$ los n valores muestrales, y sea $\bar{x}$ la media muestral. Se consideran los valores r_i definidos por

$$r_i = |x_i - \bar{x}| \qquad i=1,2,...,n$$

y sea z_i **la observación x_i cuyo r_i** asociado es el i-èsimo mayor. De esta forma z_1 es la observación más próxima a la media y z_n es la observación más alejada de la media. El estadístico utilizado para contrastar la hipótesis nula de que todas las observaciones proceden de la misma población Normal, frente a la alternativa de que existen k outliers en la muestra viene dado por:

$$E_k = \frac{\sum_{i=1}^{n-k} (z_i - \bar{z}_k)^2}{\sum_{i=1}^{n} (z_i - \bar{z})^2}$$

donde

$$\bar{z}_k = \frac{1}{n-k} \sum_{i=1}^{n-k} z_i$$

es la media muestral de las n-k observaciones mas próximas a la media, y $\bar{z} = \bar{x}$.

La región crítica para este contraste viene dada por

$$E_k < E_{k,\alpha}$$

Mediante un procedimiento de simulación, Tietjen y Moore han tabulado los valores críticos $L_{k,\alpha}$, $L^*_{k,\alpha}$ y $E_{k,\alpha}$ de los estadísticos L_k, L^*_k y E_k, los cuales comparan con los valores exactos proporcionados por Grubbs en 1950 en los casos k=1 y k=2, llegando a resultados muy similares que validan el procedimiento de simulación utilizado en la tabulación de los valores críticos de los estadísticos.

Barnett y Lewis (1984) proponen una forma alternativa del estadístico E_k, que viene dada por:

$$E_k = \frac{\sum_{i=1}^{n-k} (r_{(i)} - \bar{r}_{n-k})^2}{\sum_{i=1}^{n} (r_{(i)} - \bar{r})^2}$$

con $r_i = |x_i - \bar{x}|$, $r_{(1)} < r_{(2)} < \ldots < r_{(n)}$ los valores ordenados de los r_i, y

$$\bar{r}_{n-k} = \frac{1}{n-k} \sum_{i=1}^{n-k} r_{(i)}$$

Según estos autores este procedimiento permite un calculo más pragmático.

2.2. TECNICA DE IDENTIFICACION BASADA EN LA R-ORDENACION DE UNA MUESTRA.

En el caso de una muestra multivariante el problema del enmascaramiento tambiên se puede presentar, y en este apartado se propone una generalización del estadistico E_k propuesto por Tietjen y Moore (1972) en la forma alternativa de Barnett y Lewis (1984).

Como ya se ha dicho en la introducción del capítulo, Barnett y Lewis (1984) consideran para el caso univariante el estadístico

$$E_k = \frac{\sum_{j=1}^{n-k} (r_{(j)} - \bar{r}_{n-k})^2}{\sum_{j=1}^{n} (r_{(j)} - \bar{r})^2}$$

donde los valores $r_i = |x_i - \bar{x}|$ representan las distancias de cada una de las observaciones a la media muestral. La idea básica de este estadístico es considerar como posibles observaciones outliers las correspondientes a los k mayores r_i.

La extensión de esta idea al caso multivariante se puede realizar tomando como valores r_i la distancia de la observacion x_i al vector de medias $\bar{x}$, consisiderando la distancia inducida por una norma vectorial. Es decir, considerar

$$r_i = \|x_i - \bar{x}\| = d(x_i, \bar{x})$$

siendo x_i, $i=1,2,\ldots,n$ las columnas de una matriz X de orden pxn que representa una muestra multivariante de una población de dimensión p.

En este caso para cada norma que se considere se obtendría un estadístico distinto. En este trabajo se ha considerado la 1-norma, definida por

$$\|x_i - \bar{x}\|_1 = \max_{1\le j\le p} |x_{ji} - \bar{x}_j|$$

que en el caso unidimensional coincide con el estadístico propuesto por Barnett y Lewis (1984), y por ello se considera que es una adecuada generalización del mismo. Esto permite que en el caso multivariante el estadístico que se propone tome la forma

$$T_k = \frac{\sum_{j=1}^{n-k} (r_{(j)} - \bar{r}_{n-k})^2}{\sum_{j=1}^{n} (r_{(j)} - \bar{r})^2}$$

donde $r_i = \|x_i - \bar{x}\|_1$, $r_{(1)} < r_{(2)} < \ldots < r_{(n)}$ y

$$\bar{r}_{n-k} = \frac{1}{n-k} \sum_{j=1}^{n-k} r_{(j)}$$

como en el caso unidimensional.

Esta definición de los r_i permite realizar una subordenación de los elementos de la muestra, del tipo que Barnett (1976) denomina R-ordenación, lo que faculta para poder denominar a aquellas observaciones cuyos r_i asociados son los menores, observaciones más "pequeñas" (observaciones más próximas a la media), y denominar observaciones más "grandes" a aquellas cuyos r_i asociados son los mayores (observaciones más alejadas de la media).

Se puede observar que el numerador del estadístico T_k es proporcional a la varianza de las distancias de las n-k observaciones mas próximas al vector de medias, mientras que el denominador es proporcional a la varianza de las distancias de las n observaciones al vector de medias. Asi, la región crítica adecuada para contrastar la hipótesis nula de que no existen outliers en la muestra, frente a la alternativa de que existen k outliers es de la forma

$$T_k < T_{k,\alpha}$$

siendo $T_{k,\alpha}$ el percentil de orden α de la distribución de T_k, distribución que será objeto de análisis y estudio en el apartado siguiente.

2.3. DISTRIBUCION DEL ESTADISTICO T_k.

Recordando la expresión del estadistico T_k:

$$T_k = \frac{\Sigma\,(r_{(i)} - \bar{r}_{n-k})^2}{\Sigma\,(r_{(i)} - \bar{r}\ )^2}$$

donde $r_{(i)}$ es la sucesión de valores ordenados de

$$r_i = \|x_i - \bar{x}\|_i \qquad\qquad i=1,2,...,n$$

siendo x_i la observación i-èsima de una muestra aleatoria procedente de una población multivariante de dimensión p, **se trata ahora de determinar la correspondiente** distribución muestral del estadístico T_k. Para la determinación de dicha distribución muestral, se ha supuesto que la muestra procede de una población Normal multivariante, por ser el caso que más frecuentemente se presenta en las aplicaciones.

Abordar este problema mediante los procedimientos analíticos usuales en la Estadística Matemática, puede llevar a un sin número de dificultades de dificil superación, y aún en el caso más favorable de encontrar la expresión analitica de la distribución muestral, es seguro que habrá que recurrir finalmente a procedimientos de cálculo numèrico para determinar los valores de los percentiles $T_{k,\alpha}$ que en definitiva es el objeto último de nuestro interès.

A continuaciuón se desarrolla en en este trabajo una tècnica de simulación de la distribución del estadístico T_k a fin de determinar los valores de sus percentiles.

El procedimiento de simulación desarrollado consiste basicamente de las siguientes etapas:

a) Generación de muestras aleatorias de tamaño n, es decir, matrices de dimensiones pxn de una población Normal Multivariante de dimensión p.

b) Calcular para cada una de estas muestras el valor del etadístico T_k.

c) Determinar la distribución muestral de los valores T_k observados.

d) Estimar los percentiles de la distribución muestral teórica de T_k a partir de los percentiles de la distribución muestral.

Para el cálculo de los percentiles de T_k se ha de tener en cuenta que la distribución de T_k va a depender de tres parámetros:

n: tamaño de la muestra

p: dimensión de la población

k: número de outliers

como se aprecia al observar la expresión del estadístico. Así puès habrá que tabular los percentiles de T_k para cada 3-upla de valores (n,p,k).

Se han considerado valores de

$$n = 3(1)20$$

valores de

$$p = 1(1)5$$

(teniendo en cuenta la restricción de que $p<n$, para que la distribución sea no degenerada) y valores de

$$k = 1(1)[n/2]$$

y para cada uno de estos valores se han generado 10000 muestras aleatorias Normales p-dimensionales con vector de medias cero y matriz de varianzas-covarianzas identidad, calculando para cada una de ellas el valor del estadístico.

Así puès, el desarrollo de la primera etapa de la tècnica de simulación consiste en fijar los valores de los parámetros n, p y k, y generar 10000 muestras aleatorias Normales p-dimensionales de tamaño n. Para la generación de muestras aleatorias de una distribución Normal, existen una gran cantidad de mètodos, como pueden verse en Fishman (1978), Kennedy y Gentle (1980) y Rubinstein (1981). En este trabajo se ha utilizado el mètodo basado en la transformación de Box-Muller (1958) por ser el único mètodo exacto para la generación de muestras Normales.

Box y Muller demuestran que si U_1 y U_2 son dos variables aleatorias independientes e identicamente distribuidas según una ley Uniforme en el intervalo (0,1), entonces las variables aleatorias Z_1 y Z_2 definidas por la transformación

$$Z_1 = (-2 \ln U_1)^{1/2} \cos 2\pi U_2$$

$$Z_2 = (-2 \ln U_1)^{1/2} \operatorname{sen} 2\pi U_2$$

son independientes e identicamente distribuidas según una ley N(0,1).

Luego, el algoritmo de generación de muestras Normales se puede resumir en los dos siguientes pasos:

1.- Generar dos números aleatorios U_1 y U_2 utilizando un generador de números aleatorios.

2.- Calcular Z_1 y Z_2 simultaneamente, sustituyendo los valores U_1 y U_2 en las ecuaciones que definen la transformación de Box y Muller.

Este mètodo de generar muestras N(0,1), además de ser el único mètodo exacto, es el único que permite obtener un valor N(0,1) por cada valor U(0,1) generado, lo cual hace que el tiempo de proceso se reduzca notablemente.

Una vez generada cada una de las muestras, el cálculo del valor del estadístico, objeto de la siguiente etapa, no entraña ninguna dificultad.

La tercera etapa consiste basicamente en la ordenación de los 10000 valores del estadístico obtenidos en las etapas anteriores. En general, la ordenación de una serie de valores es un proceso que requiere una gran cantidad de tiempo, y más en este caso en el que hay que ordenar 10000 valores para cada 3-upla de valores (n,p,k). Por ello, se ha utilizado el mètodo denominado Quicksort, que es el algoritmo de ordenación más eficiente para el caso de un gran número de valores. La descripción detallada de este algoritmo y la comparación de su eficiencia con la de otros algoritmos de ordenación puede verse en Aho, Hopcroft y Ullman (1983).

Los resultados obtenidos mediante este proceso de simulación se muestran a continuación en las tablas siguientes. Respecto a la estabilidad del procedimiento seguido y consiguiente validez de los valores tabulados hay que señalar que realizadas al azar repeticiones del proceso no se han apreciado variaciones significativas en los resultados obtenidos.

2.4. PERCENTILES DE LA DISTRIBUCION DEL ESTADISTICO T_k.

A continuación se muestran los resultados obtenidos en el proceso de simulación de la distribución del estadístico T_k.

Estos resultados se refieren a los percentiles de la distribución del estadístico T_k, los cuales se encuentran dispuestos en tablas de cuatro entradas.

Dichas entradas se corresponden con los parámetros de la distribución del estadístico y el nivel de significación del contraste:

- p: dimensión de la población
- n: tamaño de la muestra
- k: número de outliers
- α: nivel de significación

Se han tabulado los percentiles de T_k para los niveles de significación más usuales,

0.005, 0.01, 0.025, 0.05, 0.1

dimensión de la población de 1 a 5, tamaños de muestra de max(3,p+1) a 20, y número de outliers de 1 a [n/2].

p = 1

n	k	α 0.005	0.010	0.025	0.050	0.100
3	1	0.0004	0.0004	0.0024	0.0064	0.0244
4	1	0.0107	0.0197	0.0447	0.0857	0.1667
	2	0.0000	0.0006	0.0006	0.0006	0.0016
5	1	0.0302	0.0442	0.0782	0.1182	0.1642
	2	0.0006	0.0016	0.0036	0.0076	0.0166
6	1	0.0557	0.0767	0.1137	0.1567	0.2227
	2	0.0040	0.0070	0.0150	0.0240	0.0420
	3	0.0005	0.0005	0.0015	0.0035	0.0065
7	1	0.0853	0.1059	0.1398	0.1841	0.2467
	2	0.0131	0.0186	0.0301	0.0441	0.0665
	3	0.0021	0.0031	0.0065	0.0105	0.0191
8	1	0.1015	0.1278	0.1684	0.2188	0.2838
	2	0.0203	0.0290	0.0461	0.0667	0.1001
	3	0.0056	0.0080	0.0154	0.0236	0.0373
	4	0.0013	0.0019	0.0036	0.0062	0.0113
9	1	0.1160	0.1444	0.2037	0.2546	0.3201
	2	0.0357	0.0481	0.0684	0.0916	0.1292
	3	0.0111	0.0162	0.0260	0.0371	0.0555
	4	0.0036	0.0050	0.0087	0.0138	0.0220
10	1	0.1339	0.1647	0.2152	0.2729	0.3442
	2	0.0467	0.0606	0.0872	0.1142	0.1543
	3	0.0191	0.0261	0.0373	0.0517	0.0749
	4	0.0692	0.0092	0.0152	0.0221	0.0345
	5	0.0020	0.0030	0.0053	0.0083	0.0136
11	1	0.1618	0.1955	0.2500	0.3040	0.3763
	2	0.0619	0.0807	0.1112	0.1432	0.1842
	3	0.0286	0.0373	0.0526	0.0717	0.0976
	4	0.0123	0.0165	0.0249	0.0345	0.0490
	5	0.0054	0.0070	0.0106	0.0156	0.0237
12	1	0.1852	0.2204	0.2758	0.3297	0.3968
	2	0.0766	0.1009	0.1286	0.1619	0.2065
	3	0.0366	0.0471	0.0657	0.0875	0.1152
	4	0.0184	0.0238	0.0340	0.0455	0.0627
	5	0.0080	0.0104	0.0157	0.0227	0.0327
	6	0.0035	0.0045	0.0077	0.0108	0.0166

p = 1

n	k	α 0.005	0.010	0.025	0.050	0.100
13	1	0.2009	0.2396	0.2943	0.3542	0.4196
	2	0.0888	0.1116	0.1474	0.1820	0.2286
	3	0.0505	0.0608	0.0799	0.1011	0.1343
	4	0.0272	0.0351	0.0460	0.0595	0.0792
	5	0.0133	0.0165	0.0245	0.0327	0.0456
	6	0.0057	0.0078	0.0118	0.0167	0.0240
14	1	0.2176	0.2595	0.3110	0.3693	0.4420
	2	0.1098	0.1316	0.1675	0.2021	0.2505
	3	0.0615	0.0761	0.0989	0.1238	0.1548
	4	0.0339	0.0422	0.0555	0.0714	0.0950
	5	0.0170	0.0223	0.0311	0.0412	0.0563
	6	0.0088	0.0120	0.0178	0.0242	0.0329
	7	0.0042	0.0057	0.0088	0.0128	0.0183
15	1	0.2490	0.2716	0.3474	0.3963	0.4691
	2	0.1322	0.1538	0.1893	0.2289	0.2775
	3	0.0873	0.0969	0.1148	0.1396	0.1745
	4	0.0479	0.0587	0.0748	0.0869	0.1107
	5	0.0268	0.0333	0.0448	0.0544	0.0692
	6	0.0141	0.0171	0.0241	0.0323	0.0435
	7	0.0076	0.0095	0.0134	0.0181	0.0251
16	1	0.2632	0.2964	0.3582	0.4109	0.4806
	2	0.1394	0.1599	0.2003	0.2365	0.2897
	3	0.0796	0.0980	0.1240	0.1498	0.1848
	4	0.0470	0.0591	0.0782	0.0971	0.1223
	5	0.0307	0.0363	0.0476	0.0627	0.0822
	6	0.0185	0.0227	0.0305	0.0387	0.0525
	7	0.0104	0.0129	0.0181	0.0239	0.0328
	8	0.0056	0.0071	0.0104	0.0143	0.0203
17	1	0.2768	0.3137	0.3751	0.4337	0.4983
	2	0.1545	0.1729	0.2182	0.2607	0.3102
	3	0.0972	0.1136	0.1396	0.1685	0.2036
	4	0.0616	0.0723	0.0915	0.1127	0.1398
	5	0.0375	0.0451	0.0618	0.0761	0.0951
	6	0.0232	0.0283	0.0374	0.0479	0.0627
	7	0.0145	0.0181	0.0236	0.0313	0.0415
	8	0.0082	0.0102	0.0143	0.0191	0.0257

p = 1

n	k	α 0.005	0.010	0.025	0.050	0.100
18	1	0.2908	0.3307	0.3964	0.4512	0.5139
	2	0.1650	0.1938	0.2374	0.2743	0.3270
	3	0.1057	0.1227	0.1524	0.1828	0.2224
	4	0.0724	0.0856	0.1066	0.1273	0.1547
	5	0.0450	0.0537	0.0706	0.0853	0.1071
	6	0.0292	0.0345	0.0471	0.0570	0.0730
	7	0.0192	0.0229	0.0305	0.0390	0.0500
	8	0.0116	0.0144	0.0194	0.0252	0.0332
	9	0.0064	0.0080	0.0118	0.0155	0.0214
19	1	0.3154	0.3519	0.4081	0.4651	0.5262
	2	0.1793	0.2042	0.2508	0.2940	0.3428
	3	0.1204	0.1391	0.1702	0.1990	0.2341
	4	0.0767	0.0903	0.1127	0.1367	0.1673
	5	0.0482	0.0597	0.0782	0.0954	0.1184
	6	0.0353	0.0421	0.0536	0.0655	0.0825
	7	0.0241	0.0290	0.0371	0.0471	0.0594
	8	0.0143	0.0177	0.0241	0.0301	0.0395
	9	0.0084	0.0112	0.0156	0.0202	0.0269
20	1	0.3281	0.3651	0.4275	0.4797	0.5419
	2	0.2017	0.2280	0.2742	0.3112	0.3583
	3	0.1258	0.1450	0.1831	0.2175	0.2549
	4	0.0827	0.0981	0.1259	0.1514	0.1823
	5	0.0606	0.0698	0.0891	0.1064	0.1317
	6	0.0398	0.0470	0.0604	0.0747	0.0933
	7	0.0275	0.0339	0.0438	0.0536	0.0681
	8	0.0186	0.0222	0.0296	0.0374	0.0485
	9	0.0125	0.0149	0.0201	0.0254	0.0331
	10	0.0076	0.0097	0.0131	0.0170	0.0226

p = 2

α

n	k	0.005	0.010	0.025	0.050	0.100
3	1	0.0004	0.0004	0.0014	0.0054	0.0194
4	1	0.0056	0.0126	0.0326	0.0576	0.1106
	2	0.0002	0.0002	0.0002	0.0012	0.0022
5	1	0.0257	0.0427	0.0737	0.1137	0.1717
	2	0.0013	0.0023	0.0053	0.0123	0.0253
6	1	0.0502	0.0722	0.1182	0.1632	0.2282
	2	0.0073	0.0109	0.0219	0.0370	0.0609
	3	0.0005	0.0011	0.0026	0.0055	0.0119
7	1	0.0817	0.1034	0.1527	0.2034	0.2771
	2	0.0192	0.0252	0.0451	0.0656	0.0991
	3	0.0035	0.0057	0.0121	0.0194	0.0333
8	1	0.1125	0.1413	0.1928	0.2523	0.3255
	2	0.0368	0.0463	0.0689	0.0963	0.1408
	3	0.0113	0.0149	0.0260	0.0387	0.0578
	4	0.0025	0.0038	0.0079	0.0124	0.0208
9	1	0.1336	0.1608	0.2193	0.2788	0.3531
	2	0.0490	0.0651	0.0922	0.1251	0.1709
	3	0.0210	0.0273	0.0419	0.0606	0.0854
	4	0.0065	0.0094	0.0164	0.0247	0.0390
10	1	0.1664	0.1955	0.2512	0.3135	0.3854
	2	0.0707	0.0880	0.1191	0.1493	0.1965
	3	0.0334	0.0412	0.0584	0.0789	0.1077
	4	0.0137	0.0195	0.0295	0.0401	0.0575
	5	0.0049	0.0071	0.0112	0.0172	0.0263
11	1	0.1832	0.2220	0.2823	0.3416	0.4180
	2	0.0837	0.1082	0.1424	0.1795	0.2316
	3	0.0465	0.0601	0.0792	0.1012	0.1332
	4	0.0204	0.0272	0.0417	0.0551	0.0757
	5	0.0097	0.0135	0.0205	0.0295	0.0414
12	1	0.2160	0.2498	0.3096	0.3699	0.4417
	2	0.1152	0.1368	0.1723	0.2099	0.2577
	3	0.0616	0.0732	0.0970	0.1219	0.1549
	4	0.0326	0.0410	0.0566	0.0717	0.0951
	5	0.0173	0.0223	0.0315	0.0402	0.0558
	6	0.0074	0.0098	0.0163	0.0229	0.0322

p = 2

n	k	0.005	0.010	0.025	0.050	0.100
				α		
13	1	0.2467	0.2787	0.3428	0.3939	0.4678
	2	0.1301	0.1533	0.1927	0.2299	0.2793
	3	0.0719	0.0896	0.1154	0.1430	0.1783
	4	0.0407	0.0507	0.0694	0.0875	0.1131
	5	0.0230	0.0302	0.0422	0.0546	0.0720
	6	0.0117	0.0158	0.0233	0.0320	0.0435
14	1	0.2420	0.2847	0.3557	0.4161	0.4835
	2	0.1463	0.1739	0.2183	0.2566	0.3100
	3	0.0876	0.1072	0.1367	0.1628	0.2006
	4	0.0515	0.0643	0.0857	0.1054	0.1326
	5	0.0324	0.0415	0.0536	0.0698	0.0890
	6	0.0176	0.0242	0.0339	0.0434	0.0570
	7	0.0097	0.0129	0.0194	0.0262	0.0353
15	1	0.2831	0.3208	0.3788	0.4435	0.5086
	2	0.1576	0.1892	0.2319	0.2740	0.3268
	3	0.0996	0.1151	0.1461	0.1772	0.2141
	4	0.0687	0.0810	0.1022	0.1219	0.1507
	5	0.0409	0.0492	0.0643	0.0821	0.1047
	6	0.0266	0.0323	0.0413	0.0541	0.0714
	7	0.0138	0.0185	0.0257	0.0342	0.0460
16	1	0.3007	0.3353	0.4060	0.4600	0.5255
	2	0.1803	0.2047	0.2453	0.2881	0.3374
	3	0.1118	0.1344	0.1660	0.1983	0.2393
	4	0.0757	0.0888	0.1122	0.1365	0.1671
	5	0.0488	0.0581	0.0773	0.0958	0.1190
	6	0.0335	0.0407	0.0537	0.0658	0.0833
	7	0.0202	0.0252	0.0347	0.0451	0.0582
	8	0.0119	0.0146	0.0212	0.0285	0.0379
17	1	0.3112	0.3515	0.4259	0.4836	0.5414
	2	0.1960	0.2197	0.2680	0.3079	0.3584
	3	0.1280	0.1470	0.1831	0.2181	0.2582
	4	0.0870	0.1045	0.1289	0.1537	0.1869
	5	0.0601	0.0733	0.0912	0.1096	0.1338
	6	0.0434	0.0512	0.0638	0.0779	0.0973
	7	0.0269	0.0316	0.0420	0.0527	0.0682
	8	0.0172	0.0213	0.0282	0.0364	0.0474

p = 2

n	k	0.005	0.010	0.025	0.050	0.100
18	1	0.3396	0.3745	0.4386	0.4940	0.5559
	2	0.2073	0.2334	0.2858	0.3253	0.3788
	3	0.1492	0.1680	0.2032	0.2353	0.2741
	4	0.1025	0.1178	0.1422	0.1672	0.2001
	5	0.0706	0.0812	0.1021	0.1217	0.1493
	6	0.0495	0.0579	0.0730	0.0896	0.1106
	7	0.0336	0.0396	0.0511	0.0637	0.0801
	8	0.0221	0.0274	0.0363	0.0455	0.0585
	9	0.0147	0.0178	0.0236	0.0305	0.0404
19	1	0.3470	0.3909	0.4513	0.5057	0.5667
	2	0.2343	0.2616	0.3073	0.3452	0.3934
	3	0.1602	0.1775	0.2145	0.2470	0.2888
	4	0.1114	0.1311	0.1572	0.1807	0.2148
	5	0.0791	0.0920	0.1129	0.1361	0.1619
	6	0.0609	0.0698	0.0857	0.1008	0.1243
	7	0.0395	0.0481	0.0609	0.0746	0.0916
	8	0.0294	0.0343	0.0437	0.0542	0.0680
	9	0.0200	0.0236	0.0304	0.0385	0.0492
20	1	0.3589	0.4002	0.4607	0.5197	0.5825
	2	0.2432	0.2760	0.3139	0.3569	0.4098
	3	0.1681	0.1939	0.2318	0.2636	0.3069
	4	0.1200	0.1384	0.1706	0.1968	0.2300
	5	0.0915	0.1051	0.1266	0.1480	0.1762
	6	0.0685	0.0800	0.0948	0.1111	0.1346
	7	0.0489	0.0582	0.0725	0.0845	0.1026
	8	0.0366	0.0425	0.0528	0.0627	0.0773
	9	0.0254	0.0293	0.0376	0.0461	0.0582
	10	0.0171	0.0201	0.0268	0.0333	0.0423

p = 3

n	k	α = 0.005	0.010	0.025	0.050	0.100
4	1	0.0065	0.0115	0.0275	0.0525	0.0985
	2	0.0005	0.0005	0.0015	0.0035	0.0155
5	1	0.0231	0.0401	0.0671	0.1051	0.1651
	2	0.0017	0.0027	0.0057	0.0127	0.0267
6	1	0.0506	0.0701	0.1106	0.1573	0.2265
	2	0.0076	0.0119	0.0240	0.0394	0.0647
	3	0.0005	0.0010	0.0027	0.0060	0.0131
7	1	0.0768	0.1086	0.1537	0.2071	0.2839
	2	0.0200	0.0290	0.0475	0.0702	0.1042
	3	0.0040	0.0067	0.0130	0.0216	0.0355
8	1	0.1092	0.1356	0.1892	0.2518	0.3285
	2	0.0356	0.0477	0.0714	0.0998	0.1417
	3	0.0113	0.0157	0.0280	0.0411	0.0622
	4	0.0025	0.0046	0.0086	0.0141	0.0239
9	1	0.1348	0.1602	0.2248	0.2902	0.3656
	2	0.0536	0.0708	0.0990	0.1311	0.1789
	3	0.0206	0.0296	0.0438	0.0625	0.0889
	4	0.0077	0.0112	0.0185	0.0270	0.0420
10	1	0.1725	0.2010	0.2604	0.3208	0.3993
	2	0.0789	0.0921	0.1267	0.1590	0.2072
	3	0.0369	0.0476	0.0664	0.0883	0.1176
	4	0.0161	0.0206	0.0311	0.0431	0.0613
	5	0.0057	0.0078	0.0140	0.0202	0.0310
11	1	0.1982	0.2388	0.3001	0.3582	0.4288
	2	0.0923	0.1130	0.1517	0.1905	0.2411
	3	0.0516	0.0644	0.0849	0.1083	0.1404
	4	0.0265	0.0332	0.0466	0.0615	0.0834
	5	0.0107	0.0153	0.0239	0.0328	0.0464
12	1	0.2208	0.2599	0.3199	0.3754	0.4515
	2	0.1036	0.1335	0.1749	0.2170	0.2671
	3	0.0651	0.0752	0.1034	0.1333	0.1699
	4	0.0350	0.0409	0.0602	0.0778	0.1029
	5	0.0178	0.0234	0.0335	0.0448	0.0622
	6	0.0083	0.0119	0.0179	0.0254	0.0364

$p = 3$

n	k	0.005	0.010	0.025	0.050	0.100
				α		
13	1	0.2369	0.2773	0.3430	0.4025	0.4774
	2	0.1373	0.1586	0.2025	0.2464	0.2979
	3	0.0854	0.1028	0.1268	0.1556	0.1915
	4	0.0474	0.0563	0.0766	0.0970	0.1229
	5	0.0270	0.0329	0.0459	0.0595	0.0789
	6	0.0137	0.0189	0.0267	0.0357	0.0496
14	1	0.2684	0.3083	0.3693	0.4270	0.4978
	2	0.1560	0.1763	0.2231	0.2612	0.3176
	3	0.0950	0.1116	0.1417	0.1713	0.2135
	4	0.0591	0.0724	0.0931	0.1138	0.1449
	5	0.0385	0.0451	0.0597	0.0755	0.0972
	6	0.0215	0.0270	0.0382	0.0485	0.0639
	7	0.0116	0.0154	0.0229	0.0303	0.0400
15	1	0.2864	0.3269	0.3896	0.4464	0.5138
	2	0.1740	0.2019	0.2457	0.2863	0.3365
	3	0.1021	0.1238	0.1561	0.1872	0.2289
	4	0.0728	0.0888	0.1118	0.1327	0.1643
	5	0.0455	0.0558	0.0713	0.0892	0.1128
	6	0.0280	0.0349	0.0474	0.0595	0.0773
	7	0.0158	0.0204	0.0302	0.0399	0.0534
16	1	0.3076	0.3471	0.4133	0.4698	0.5344
	2	0.1825	0.2145	0.2612	0.3061	0.3576
	3	0.1230	0.1458	0.1812	0.2107	0.2499
	4	0.0850	0.0996	0.1232	0.1487	0.1806
	5	0.0605	0.0678	0.0858	0.1046	0.1301
	6	0.0359	0.0437	0.0571	0.0737	0.0923
	7	0.0229	0.0289	0.0392	0.0491	0.0642
	8	0.0147	0.0182	0.0254	0.0329	0.0442
17	1	0.3264	0.3612	0.4268	0.4853	0.5482
	2	0.1984	0.2295	0.2780	0.3193	0.3720
	3	0.1451	0.1633	0.1974	0.2311	0.2710
	4	0.0974	0.1120	0.1366	0.1620	0.1954
	5	0.0663	0.0803	0.0981	0.1196	0.1449
	6	0.0458	0.0539	0.0711	0.0859	0.1058
	7	0.0290	0.0356	0.0471	0.0591	0.0752
	8	0.0188	0.0237	0.0326	0.0414	0.0538

p = 3

n	k	α 0.005	0.010	0.025	0.050	0.100
18	1	0.3386	0.3856	0.4517	0.5057	0.5664
	2	0.2206	0.2545	0.2901	0.3338	0.3874
	3	0.1466	0.1751	0.2102	0.2462	0.2876
	4	0.1041	0.1187	0.1480	0.1754	0.2080
	5	0.0794	0.0903	0.1105	0.1314	0.1623
	6	0.0541	0.0644	0.0806	0.0979	0.1194
	7	0.0390	0.0470	0.0596	0.0731	0.0900
	8	0.0255	0.0306	0.0402	0.0506	0.0646
	9	0.0167	0.0213	0.0287	0.0366	0.0474
19	1	0.3579	0.4052	0.4702	0.5217	0.5800
	2	0.2319	0.2625	0.3120	0.3563	0.4065
	3	0.1626	0.1884	0.2266	0.2611	0.3034
	4	0.1222	0.1407	0.1680	0.1954	0.2295
	5	0.0892	0.1010	0.1231	0.1454	0.1739
	6	0.0634	0.0738	0.0921	0.1101	0.1338
	7	0.0464	0.0532	0.0685	0.0843	0.1017
	8	0.0338	0.0398	0.0507	0.0623	0.0772
	9	0.0223	0.0267	0.0348	0.0441	0.0570
20	1	0.3736	0.4110	0.4726	0.5294	0.5897
	2	0.2579	0.2791	0.3293	0.3740	0.4227
	3	0.1727	0.2012	0.2408	0.2753	0.3171
	4	0.1294	0.1486	0.1780	0.2067	0.2420
	5	0.0973	0.1128	0.1349	0.1570	0.1897
	6	0.0648	0.0821	0.1023	0.1204	0.1455
	7	0.0523	0.0602	0.0750	0.0927	0.1121
	8	0.0392	0.0452	0.0572	0.0705	0.0873
	9	0.0283	0.0334	0.0425	0.0530	0.0666
	10	0.0181	0.0221	0.0297	0.0383	0.0489

p = 4

n	k	0.005	0.010	0.025	0.050	0.100
5	1	0.0272	0.0382	0.0672	0.1052	0.1662
	2	0.0016	0.0026	0.0066	0.0126	0.0266
6	1	0.0507	0.0700	0.1059	0.1533	0.2223
	2	0.0077	0.0114	0.0247	0.0400	0.0677
	3	0.0005	0.0011	0.0030	0.0061	0.0139
7	1	0.0728	0.1095	0.1519	0.2119	0.2984
	2	0.0208	0.0334	0.0533	0.0745	0.1099
	3	0.0052	0.0083	0.0150	0.0229	0.0392
8	1	0.1092	0.1396	0.1872	0.2507	0.3325
	2	0.0377	0.0453	0.0769	0.1037	0.1439
	3	0.0097	0.0156	0.0278	0.0414	0.0641
	4	0.0025	0.0046	0.0089	0.0147	0.0258
9	1	0.1291	0.1621	0.2277	0.2936	0.3705
	2	0.0574	0.0756	0.1071	0.1406	0.1830
	3	0.0224	0.0316	0.0459	0.0657	0.0927
	4	0.0087	0.0126	0.0193	0.0290	0.0444
10	1	0.1680	0.2007	0.2599	0.3241	0.4014
	2	0.0752	0.0919	0.1257	0.1632	0.2108
	3	0.0386	0.0486	0.0688	0.0910	0.1207
	4	0.0165	0.0221	0.0335	0.0477	0.0661
	5	0.0056	0.0086	0.0139	0.0208	0.0317
11	1	0.1932	0.2261	0.2873	0.3479	0.4305
	2	0.0968	0.1205	0.1583	0.1925	0.2465
	3	0.0523	0.0643	0.0842	0.1095	0.1438
	4	0.0264	0.0336	0.0466	0.0622	0.0850
	5	0.0116	0.0154	0.0235	0.0341	0.0497
12	1	0.2216	0.2603	0.3264	0.3824	0.4583
	2	0.1149	0.1428	0.1802	0.2195	0.2717
	3	0.0633	0.0782	0.1054	0.1315	0.1659
	4	0.0337	0.0471	0.0605	0.0801	0.1065
	5	0.0194	0.0239	0.0354	0.0474	0.0650
	6	0.0086	0.0122	0.0183	0.0253	0.0374

p = 4

n	k	α 0.005	0.010	0.025	0.050	0.100
13	1	0.2560	0.2837	0.3505	0.4119	0.4808
	2	0.1377	0.1631	0.2087	0.2449	0.2964
	3	0.0847	0.0986	0.1275	0.1534	0.1946
	4	0.0496	0.0607	0.0798	0.0998	0.1278
	5	0.0290	0.0351	0.0490	0.0642	0.0838
	6	0.0157	0.0204	0.0284	0.0385	0.0533
14	1	0.2759	0.3183	0.3797	0.4367	0.4993
	2	0.1563	0.1829	0.2276	0.2675	0.3194
	3	0.0989	0.1167	0.1461	0.1756	0.2172
	4	0.0588	0.0711	0.0924	0.1160	0.1454
	5	0.0361	0.0464	0.0618	0.0771	0.1003
	6	0.0217	0.0285	0.0389	0.0499	0.0671
	7	0.0112	0.0156	0.0225	0.0307	0.0423
15	1	0.2818	0.3214	0.3917	0.4506	0.5151
	2	0.1715	0.1969	0.2434	0.2802	0.3352
	3	0.1121	0.1321	0.1611	0.1941	0.2368
	4	0.0699	0.0846	0.1100	0.1349	0.1655
	5	0.0446	0.0556	0.0721	0.0911	0.1154
	6	0.0296	0.0352	0.0472	0.0611	0.0801
	7	0.0163	0.0216	0.0314	0.0408	0.0546
16	1	0.3055	0.3483	0.4152	0.4734	0.5386
	2	0.1869	0.2163	0.2583	0.3059	0.3586
	3	0.1303	0.1459	0.1793	0.2139	0.2528
	4	0.0833	0.0972	0.1251	0.1486	0.1831
	5	0.0564	0.0679	0.0857	0.1062	0.1327
	6	0.0387	0.0468	0.0601	0.0750	0.0942
	7	0.0253	0.0309	0.0396	0.0504	0.0671
	8	0.0152	0.0191	0.0264	0.0342	0.0457
17	1	0.3151	0.3560	0.4325	0.4928	0.5565
	2	0.2098	0.2336	0.2815	0.3246	0.3764
	3	0.1422	0.1612	0.1952	0.2279	0.2707
	4	0.0932	0.1127	0.1379	0.1641	0.1986
	5	0.0665	0.0799	0.0990	0.1199	0.1493
	6	0.0457	0.0550	0.0727	0.0884	0.1084
	7	0.0319	0.0366	0.0492	0.0627	0.0795
	8	0.0190	0.0248	0.0339	0.0433	0.0569

p = 4

n	k	0.005	0.010	0.025	0.050	0.100
18	1	0.3253	0.3757	0.4402	0.4977	0.5647
	2	0.2353	0.2636	0.3100	0.3473	0.3974
	3	0.1515	0.1748	0.2146	0.2492	0.2922
	4	0.1102	0.1258	0.1550	0.1809	0.2166
	5	0.0786	0.0931	0.1146	0.1359	0.1661
	6	0.0553	0.0644	0.0822	0.1005	0.1226
	7	0.0407	0.0473	0.0606	0.0732	0.0911
	8	0.0258	0.0331	0.0436	0.0549	0.0694
	9	0.0172	0.0204	0.0294	0.0374	0.0497
19	1	0.3699	0.4112	0.4659	0.5149	0.5787
	2	0.2385	0.2629	0.3138	0.3594	0.4114
	3	0.1715	0.1929	0.2299	0.2639	0.3071
	4	0.1208	0.1398	0.1696	0.1993	0.2371
	5	0.0898	0.1026	0.1284	0.1488	0.1784
	6	0.0667	0.0770	0.0950	0.1143	0.1382
	7	0.0455	0.0531	0.0688	0.0835	0.1036
	8	0.0328	0.0403	0.0504	0.0626	0.0793
	9	0.0240	0.0283	0.0355	0.0457	0.0584
20	1	0.3721	0.4104	0.4699	0.5302	0.5915
	2	0.2536	0.2853	0.3273	0.3691	0.4257
	3	0.1888	0.2093	0.2460	0.2777	0.3190
	4	0.1390	0.1563	0.1815	0.2094	0.2478
	5	0.0980	0.1158	0.1394	0.1618	0.1917
	6	0.0738	0.0848	0.1040	0.1238	0.1499
	7	0.0595	0.0670	0.0808	0.0974	0.1182
	8	0.0406	0.0488	0.0615	0.0739	0.0899
	9	0.0284	0.0334	0.0435	0.0547	0.0694
	10	0.0211	0.0254	0.0322	0.0399	0.0511

p = 5

n	k	α 0.005	0.010	0.025	0.050	0.100
6	1	0.0476	0.0669	0.1081	0.1555	0.2245
	2	0.0072	0.0127	0.0257	0.0413	0.0671
	3	0.0006	0.0012	0.0033	0.0063	0.0140
7	1	0.0761	0.1063	0.1581	0.2058	0.2863
	2	0.0217	0.0294	0.0501	0.0724	0.1080
	3	0.0042	0.0068	0.0135	0.0216	0.0367
8	1	0.1182	0.1485	0.1988	0.2606	0.3336
	2	0.0415	0.0535	0.0770	0.1046	0.1485
	3	0.0123	0.0181	0.0299	0.0439	0.0671
	4	0.0029	0.0045	0.0083	0.0143	0.0248
9	1	0.1426	0.1706	0.2338	0.2918	0.3652
	2	0.0549	0.0727	0.1020	0.1318	0.1788
	3	0.0205	0.0287	0.0470	0.0665	0.0939
	4	0.0810	0.0121	0.0189	0.0283	0.0435
10	1	0.1584	0.1980	0.2574	0.3203	0.3939
	2	0.0797	0.0979	0.1347	0.1696	0.2156
	3	0.0337	0.0461	0.0660	0.0880	0.1194
	4	0.0153	0.0205	0.0325	0.0448	0.0646
	5	0.0064	0.0087	0.0136	0.0213	0.0324
11	1	0.1850	0.2258	0.2899	0.3535	0.4299
	2	0.0976	0.1178	0.1574	0.1962	0.2423
	3	0.0537	0.0652	0.0890	0.1145	0.1464
	4	0.0239	0.0309	0.0472	0.0641	0.0884
	5	0.0103	0.0146	0.0243	0.0343	0.0500
12	1	0.2001	0.2522	0.3166	0.3773	0.4526
	2	0.1140	0.1397	0.1800	0.2188	0.2698
	3	0.0642	0.0781	0.1068	0.1355	0.1701
	4	0.0376	0.0463	0.0622	0.0790	0.1063
	5	0.0202	0.0248	0.0370	0.0503	0.0678
	6	0.0081	0.0119	0.0187	0.0270	0.0394
13	1	0.2536	0.2926	0.3504	0.4139	0.4823
	2	0.1335	0.1608	0.2037	0.2446	0.2974
	3	0.0767	0.0957	0.1276	0.1549	0.1955
	4	0.0482	0.0616	0.0807	0.0993	0.1299
	5	0.0254	0.0343	0.0482	0.0639	0.0843
	6	0.0132	0.0187	0.0273	0.0385	0.0527

p = 5

n	k	0.005	0.010	0.025	0.050	0.100
		α				
14	1	0.2590	0.3005	0.3654	0.4302	0.4995
	2	0.1565	0.1761	0.2230	0.2648	0.3169
	3	0.1012	0.1183	0.1491	0.1790	0.2160
	4	0.0553	0.0711	0.0939	0.1159	0.1472
	5	0.0375	0.0449	0.0614	0.0761	0.0994
	6	0.0223	0.0286	0.0395	0.0509	0.0679
	7	0.0120	0.0160	0.0233	0.0305	0.0419
15	1	0.2925	0.3306	0.3927	0.4558	0.5269
	2	0.1709	0.1987	0.2427	0.2855	0.3409
	3	0.1123	0.1351	0.1687	0.1947	0.2390
	4	0.0689	0.0858	0.1144	0.1366	0.1695
	5	0.0466	0.0574	0.0748	0.0925	0.1164
	6	0.0266	0.0366	0.0489	0.0625	0.0830
	7	0.0170	0.0232	0.0318	0.0411	0.0557
16	1	0.2945	0.3365	0.4057	0.4667	0.5348
	2	0.1894	0.2176	0.2651	0.3063	0.3591
	3	0.1246	0.1501	0.1826	0.2143	0.2561
	4	0.0859	0.0972	0.1266	0.1518	0.1847
	5	0.0582	0.0699	0.0881	0.1081	0.1344
	6	0.0365	0.0453	0.0597	0.0745	0.0948
	7	0.0244	0.0304	0.0409	0.0527	0.0681
	8	0.0149	0.0177	0.0263	0.0345	0.0465
17	1	0.3234	0.3667	0.4342	0.4895	0.5544
	2	0.2131	0.2381	0.2843	0.3244	0.3753
	3	0.1394	0.1579	0.1939	0.2285	0.2728
	4	0.0987	0.1165	0.1442	0.1704	0.2022
	5	0.0650	0.0778	0.1007	0.1216	0.1483
	6	0.0472	0.0557	0.0724	0.0886	0.1109
	7	0.0313	0.0390	0.0493	0.0628	0.0805
	8	0.0204	0.0252	0.0343	0.0437	0.0573
18	1	0.3384	0.3753	0.4463	0.5013	0.5647
	2	0.2213	0.2567	0.2976	0.3463	0.3957
	3	0.1555	0.1812	0.2163	0.2500	0.2916
	4	0.1084	0.1259	0.1572	0.1866	0.2194
	5	0.0759	0.0868	0.1099	0.1322	0.1609
	6	0.0570	0.0675	0.0852	0.1024	0.1253
	7	0.0369	0.0450	0.0605	0.0752	0.0934
	8	0.0254	0.0324	0.0436	0.0542	0.0697
	9	0.0165	0.0207	0.0293	0.0383	0.0496

p = 5

n	k	α 0.005	0.010	0.025	0.050	0.100
19	1	0.3637	0.4007	0.4687	0.5276	0.5818
	2	0.2473	0.2756	0.3198	0.3624	0.4121
	3	0.1644	0.1920	0.2289	0.2659	0.3068
	4	0.1195	0.1371	0.1678	0.1935	0.2336
	5	0.0913	0.1043	0.1285	0.1511	0.1807
	6	0.0627	0.0727	0.0929	0.1140	0.1385
	7	0.0481	0.0553	0.0710	0.0867	0.1073
	8	0.0343	0.0411	0.0527	0.0650	0.0812
	9	0.0229	0.0289	0.0372	0.0463	0.0586
20	1	0.3721	0.4202	0.4803	0.5332	0.5917
	2	0.2572	0.2822	0.3291	0.3743	0.4257
	3	0.1792	0.2017	0.2436	0.2793	0.3231
	4	0.1296	0.1493	0.1805	0.2111	0.2480
	5	0.1024	0.1151	0.1404	0.1645	0.1948
	6	0.0737	0.0866	0.1081	0.1256	0.1502
	7	0.0532	0.0633	0.0808	0.0978	0.1189
	8	0.0390	0.0489	0.0617	0.0752	0.0926
	9	0.0280	0.0339	0.0444	0.0553	0.0707
	10	0.0187	0.0238	0.0329	0.0413	0.0533

2.5. SUBRUTINA DE CALCULO DEL ESTADISTICO T_k.

Se completa este segundo capitulo con una subrutina en BASIC para el cálculo del valor del estadístico T_k, y la identificación de las observaciones outliers en el caso de que existan.

El motivo de la elección del BASIC como lenguaje de programación se debe a que, hoy en dia, es el que está más al alcance de cualquier investigador, debido a la proliferación de los ordenadores personales, ya sean micro o miniordenadores. En todo caso la traducción de esta rutina a otro lenguaje de alto nivel, como FORTRAN, PASCAL, etc., es inmediata, sin más que tener unos mínimos conocimientos de la sintáxis y estructura de estos lenguajes.

Las variables de entrada y salida de esta rutina están suficientemente detalladas en las líneas de comentario que figuran al comienzo de la misma.

En el listado adjunto se han incluido asimismo líneas de comentario, que indican en cada momento el cálculo que se está realizando, para facilitar aún más la traducción a otro lenguaje de programación.

```
64000 REM *******************************************************
64010 REM          Subrutina para el cálculo del estadístico Tk. *
64020 REM                                                        *
64030 REM   Datos de entrada:                                    *
64040 REM                                                        *
64050 REM          P: Dimensión de la población.                 *
64060 REM          N: Tamaño de la muestra.                      *
64070 REM          X: Matriz PxN, cuyas columnas son las observa *
64080 REM             nes muestrales.                            *
64090 REM          K: Número de outliers.                        *
64100 REM                                                        *
64110 REM   Datos de salida:                                     *
64120 REM                                                        *
64130 REM         XM: Vector de medias de dimensión P.           *
64140 REM          R: Vector de distancias de  cada  observación *
64150 REM             al vector de medias.                       *
64160 REM          N: Vector de indices, cuyos k últimos elemen- *
64170 REM             tos indican las columnas de X que posible- *
64180 REM             mente sean outliers.                       *
64190 REM         TK: Valor del estadístico                      *
64200 REM *******************************************************
64210 REM
64220 DIM XM(P),R(N),N(N)
64230 REM ******************************
64240 REM  Cálculo del vector de medias. *
64250 REM ******************************
64260 FOR I=1 TO P
64270 XM(I)=0
64280 FOR J=1 TO N
64290 XM(I)=XM(I)+X(I,J)
64300 NEXT J
64310 XM(I)=XM(I)/N
64320 NEXT I
64330 REM *********************************************************
64340 REM Cálculo de las distancias de cada observación al vector *
64350 REM de medias, e inicialización del vector de indices.      *
64360 REM *********************************************************
64370 FOR I=1 TO N
64380 R(I)=ABS(X(1,I)-XM(1))
64390 N(I)=I
64400 FOR J=2 TO P
64410 A=ABS(X(J,I)-XM(J))
64420 IF A>R(I) THEN R(I)=A
64430 NEXT J
64440 NEXT I
64450 REM ***********************************
64460 REM  Ordenación del vector de distancias. *
64470 REM ***********************************
64480 FOR I=1 TO N-1
64490 FOR J=I+1 TO N
64500 IF R(I)<=R(J) THEN 64530
64510 RR=R(I):R(I)=R(J):R(J)=RR
64520 NN=N(I):N(I)=N(J):N(J)=NN
64530 NEXT J
64540 NEXT I
```

```
64550 REM ***************************************************************
64560 REM  Cálculo de la media de las N-K menores distancias y de *
64570 REM  la media de las N distancias.                            *
64580 REM ***************************************************************
64590 RMK=0
64600 FOR I=1 TO N-K
64610 RMK=RMK+R(I)
64620 NEXT I
64630 RM=RMK
64640 RMK=RMK/(N-K)
64650 FOR I=N-K+1 TO N
64660 RM=RM+R(I)
64670 NEXT I
64680 RM=RM/N
64690 REM ************************************
64700 REM  Cálculo del valor del estadístico. *
64710 REM ************************************
64720 T1=0
64730 T2=0
64740 FOR I=1 TO N-K
64750 T1=T1+(R(I)-RMK)^2
64760 NEXT I
64770 FOR I=1 TO N
64780 T2=T2+(R(I)-RM)^2
64790 NEXT I
64800 TK=T1/T2
64810 RETURN
```

2.6. CASO PRACTICO.

A continuación se aplica el mètodo propuesto a los siguientes datos extraidos de la publicación del Banco de Bilbao "Renta Nacional de España 1981".

V. A. B. por empleo (Año 1981)
(millones de pesetas)

Comunidades Autónomas	Transportes y Comunicaciones	Ahorro, Banca y Seguros	Enseñanza y Sanidad
Andalucía	1.56	2.11	1.38
Aragón	1.64	2.70	1.42
Asturias	1.68	2.60	1.49
Baleares	1.96	2.20	1.65
Canarias	1.80	2.38	1.51
Cantabria	1.77	2.45	1.44
Castilla - La Mancha	1.28	2.43	1.28
Castilla - León	1.54	2.75	1.36
Cataluña	1.87	2.79	1.74
Extremadura	1.27	2.37	1.25
Galicia	1.40	2.54	1.38
Madrid	1.97	2.62	1.86
Murcia	1.67	2.21	1.56
Navarra	1.56	2.44	1.43
Pais Vasco	1.65	2.67	1.48
La Rioja	1.64	3.34	1.42
Valencia	1.71	2.34	1.62

Fuente: Banco de Bilbao. Renta Nacional de España 1981.

En primer lugar se contrasta la normalidad multivariante de las observaciones mediante el test de Mardia (Mardia, Kent y Bibby, 1979), obteniendose los resultados que se exponen a continuación.

TEST DE MARDIA

$\chi^2 = 13.6150$ gdl = 10 $\chi^2_{10,0.95}=18.3070$

$Z = -0.2416$ $z_{0.975}= 1.9600$

A la vista de estos resultados se puede considerar que los datos anteriores representan una muestra extraida de una población Normal Multidimensional.

En la tabla siguiente se muestran los resultados de la aplicación del mètodo propuesto para la identificació. de outliers, el cual se ha aplicado secuencialmente para detectar bloques de 1, 2 y 3 outliers, resultando significativos unicamente los valores del estadístico para 1 y 2 outliers. Asimismo se indican las observaciones outliers detectadas.

IDENTIFICACION DE OUTLIERS

N. outliers	Estadístico	Valor crítico	Observaciones
1	0.3741	0.4853	16
2	0.3078	0.3193	16 1
3	0.2614	0.2311	

3.- DISTANCIA ENTRE MATRICES DE SUMAS DE CUADRADOS Y SUMAS DE PRODUCTOS.

3.1. INTRODUCCION.

En este capitulo se aborda el problema de la la detección e identificación de outliers en muestras multivariantes desde una perspectiva diferente a la utilizada en el capitulo anterior, y que no tiene antecedente en el caso unidimensional.

La caracteristica muestral en la que se va a basar el mètodo que se propone es la matriz de sumas de cuadrados y sumas de productos de observaciones muestrales. El estadístico básico utilizado en la tècnica de identificación se basa en la distancia inducida por la norma espectral de matrices, y a partir de èl se construye el utilizado en el mètodo de identificación.

Se determina la distribución en el muestreo del estadístico básico, y se analiza el caso de detección de un outlier, asimismo se construyen estadísticos para la detección de más de un outlier con objeto de eliminar el efecto de enmascaramiento definido en el capitulo anterior.

3.2. METRICA EN EL ESPACIO VECTORIAL DE LAS MATRICES.

Sea $\mathcal{M}$ el espacio vectorial de las matrices de dimensiones mxn sobre el cuerpo de los números reales, y sea $(E, \|\ \|_E)$ un espacio vectorial normado de dimensión n. Se verifica que la función

$$\|\ \| : \mathcal{M} \longrightarrow R^{+}$$

definida mediante

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|_E}{\|x\|_E} = \max_{\|x\|_E = 1} \|Ax\|_E \qquad A \in \mathcal{M} \ , \ x \in E$$

es una norma en el espacio vectorial $\mathcal{M}$, consistente con la norma vectorial $\|\ \|_E$ definida sobre E.

Sea $A \in \mathcal{M}$. Se denominan valores singulares de A, y se representan por $\sigma_i(A)$, a las raices cuadradas positivas de los autovalores positivos de A'A:

$$\sigma_i(A) = +\sqrt{\lambda_i(A'A)} \qquad \lambda_i(A'A) > 0 \qquad i = 1, 2, \ldots n$$

siendo $\lambda_i(A'A)$ los autovalores de A'A.

Se denomina norma espectral, o 2-norma, de una matriz $A \in \mathcal{M}$ a la definida mediante

$$\|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2 \qquad x \in E$$

siendo $\|x\|_2 = (x'x)^{1/2}$ la norma euclidea. Para la norma así definida, se verifica que

$$\|A\|_2 = \max_{1 \leq i \leq n} \sigma_i(A) \qquad A \in \mathcal{M}$$

(Stewart, 1973)

Si $A,B \in \mathcal{M}$, entonces la función

$$d: \mathcal{M} \times \mathcal{M} \longrightarrow R^{+}$$

definida por

$$d(A,B) = \max_{1 \leq i \leq n} \sigma_i (A-B)$$

es una mètrica. La comprobación es fácil, sin más que tener en cuenta las propiedades de la norma espectral.

Si la matriz A-B, con $A,B \in \mathcal{M}$, es simètrica y definida positiva, los autovalores de $(A-B)'(A-B)$ son iguales a los cuadrados de los autovalores de A-B, y teniendo en cuenta la definición de valor singular y la definición de la mètrica anterior, se deduce que

$$d(A,B) = \max_{1 \leq i \leq n} \lambda_i (A-B)$$

A lo largo de este capítulo se supondrá que los autovalores de una matriz A se tienen ordenados de mayor a menor, es decir

$$\lambda_1(A) \geq \lambda_2(A) \geq \dots$$

y, por tanto, la mètrica definida anteriormente se puede expresar en la forma

$$d(A,B) = \lambda_1 (A-B)$$

El objetivo va a ser aplicar estos resultados a las matrices de sumas de cuadrados y sumas de productos de una muestra multivariante con el fin de obtener una tècnica de identificación de outliers en muestras extraidas de poblaciones Normales multivariantes.

3.3. CONSTRUCCION DEL ESTADISTICO BASICO.

Sea X una muestra aleatoria procedente de una población $N_p(\mu,\Sigma)$ no singular

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

El vector de medias muestrales, $x_{(n)}$, y la matriz de sumas de cuadrados y sumas de productos, $S_{(n)}$, **vienen dados por:**

$$\bar{x}_{(n)} = X\frac{1}{n}E_{n1}$$

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

con I_n **la matriz identidad de orden n,** y E_{kh} representa una matriz kxh cuyos elementos son todos iguales a 1.

En las expresiones anteriores, el subíndice (n) representa el número de observaciones utilizadas en el cálculo del vector de medias y la matriz de sumas de cuadrados y sumas de productos.

La matriz $S_{(n)}$ es simètrica por su propia construcción, y además $S_{(n)}$ tiene distribucion de Wishart de dimensión p con n-1 grados de libertad y matriz asociada Σ (Kshirsagar, 1972). Esto último se expresará simbolicamente por $S(n) \in W_p(n-1,\Sigma)$

Si $n>p$, $S_{(n)}$ es definida positiva con probabilidad 1.

Sean $X_{(n-r)}$, $X_{(r)}$ dos matrices formadas por las n-r primeras columnas de X y por las restantes r columnas de X, respectivamente. Esto es, $X_{(n-r)}, X_{(r)}$ representan dos submuestras de la muestra X, de tamaños n-r y r, respectivamente. Se tiene por tanto

$$X = [X_{(n-r)} | X_{(r)}]$$

por lo que

$$S_{(n)} = X\,(I_n - \tfrac{1}{n} E_{nn})\,X' =$$
$$= [X_{(n-r)} | X_{(r)}]\,(I_n - \tfrac{1}{n} E_{nn})\,[X_{(n-r)} | X_{(r)}]'$$

Descomponiendo ahora las matrices I_n y E_{nn} de forma análoga, se tiene

$$S_{(n)} = X_{(n-r)}X'_{(n-r)} + X_{(r)}X'_{(r)} - \tfrac{1}{n}(X_{(n-r)}E_{n-r,r}X'_{(n-r)} +$$
$$+ X_{(n-r)}E_{n-r,r}X'_{(r)} + X_{(r)}E_{r,n-r}X'_{(n-r)} + X_{(r)}E_{rr}X'_{(r)})$$

Sumando y restando en el segundo miembro de la expresión anterior

$$\frac{1}{n-r} X_{(n-r)}E_{n-r,n-r}X'_{(n-r)} + \frac{1}{r} X_{(r)}E_{rr}X'_{(r)}$$

se tiene

$$S_{(n)} = S_{(n-r)} + S_{(r)} + (n-r)\,\bar{x}_{(n-r)}\bar{x}'_{(n-r)} -$$

$$-\frac{(n-r)^2}{n}\bar{x}_{(n-r)}\bar{x}'_{(n-r)} + r\,\bar{x}_{(r)}\bar{x}'_{(r)} - \frac{r^2}{n}\bar{x}_{(r)}\bar{x}'_{(r)} -$$

$$-\frac{(n-r)r}{n}\bar{x}_{(r)}\bar{x}'_{(n-r)} - \frac{(n-r)r}{n}\bar{x}_{(n-r)}\bar{x}'_{(r)} =$$

$$= S_{(n-r)} + S_{(r)} + \frac{r(n-r)}{n}(\bar{x}_{(n-r)} - \bar{x}_{(r)})(\bar{x}_{(n-r)} - \bar{x}_{(r)})'$$

Luego la matriz de sumas de cuadrados y sumas de productos se puede expresar en la forma

$$S_{(n)} = S_{(n-r)} + S_{(r)} + \frac{r(n-r)}{n}(\bar{x}_{(n-r)} - \bar{x}_{(r)})(\bar{x}_{(n-r)} - \bar{x}_{(r)})'$$

Teniendo en cuenta ahora que

$$\bar{x}_{(n-r)} \in N_p(\mu, \frac{1}{n-r}\Sigma) \qquad y \qquad \bar{x}_{(r)} \in Np(\mu, \frac{1}{r}\Sigma)$$

y que ambas son independientes (Kshirsagar, 1972) se deduce que

$$\sqrt{\frac{r(n-r)}{n}}\,(\bar{x}_{(n-r)} - \bar{x}_{(r)}) \in N_p(0, \Sigma)$$

y

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

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

De Kshirsagar (1972), $S_{(r)} \in W_p(r-1,\Sigma)$, y al ser $S_{(r)}$ y

$$\frac{r(n-r)}{n}(\bar{x}_{(n-r)}-\bar{x}_{(r)})(\bar{x}_{(n-r)}-\bar{x}_{(r)})'$$

independientes (Wilks, 1962), se verifica que

$$S^* = S_{(r)} + \frac{r(n-r)}{n}(\bar{x}_{(n-r)}-\bar{x}_{(r)})(\bar{x}_{(n-r)}-\bar{x}_{(r)})' \in Wp(r,\Sigma)$$

De aquí se deduce que $S_{(n)}-S_{(n-r)}$ es una matriz simètrica y definida positiva, y teniendo en cuenta la la mètrica definida anteriormente

$$d(S_{(n)},S_{(n-r)}) = \lambda_1(S_{(n)} - S_{(n-r)}) = \lambda_1(S^*)$$

con $\lambda_1(S^*)$ el mayor autovalor de S^*.

La tècnica de identificación de outliers va a tener como estadístico básico

$$T_r = d(S_{(n)},S_{(n-r)}) = \lambda_1(S^*)$$

que representa la distancia entre la matriz de sumas de cuadrados y sumas de productos calculada con las n observaciones y la matriz de sumas de cuadrados y sumas de productos cuando se eliminan r de esas observaciones.

3.4. DISTRIBUCION DEL ESTADISTICO BASICO.

Para calcular la distribución del estadístico

$$T_r = \lambda_1(S^*)$$

se han de dar en primer lugar las siguientes definiciones:

Sea k un entero positivo. Se denomina partición de k, y se representa por $\kappa = (k_1, k_2, \ldots)$, a un conjunto de enteros no negativos, $k_1, k_2, \ldots$, tales que $\sum_i k_i = k$ y que, por convenio, se supone que $k_1 \geq k_2 \geq \ldots$

Si $\kappa = (k_1, k_2, \ldots)$ y $\lambda = (l_1, l_2, \ldots)$ son dos particiones de un mismo entero k, se escribe $\kappa > \lambda$ si se verifica que $k_i > l_i$ para el primer índice i en que k_i es distinto de l_i.

Sean $\kappa = (k_1, k_2, \ldots, k_m)$ y $\lambda = (l_1, l_2, \ldots, l_m)$ dos particiones de un entero positivo k, y sean $y_1, y_2, \ldots, y_m$ m variables. Si $\kappa > \lambda$ se dice que el monomio

$$y_1^{k_1} \; y_2^{k_2} \ldots\ldots \; y_m^{k_m}$$

es de mayor grado que el monomio

$$y_1^{l_1} \; y_2^{l_2} \ldots\ldots \; y_m^{l_m}$$

Sea A una matriz cuadrada simètrica de orden p, con autovalores $\lambda_1, \lambda_2, \ldots, \lambda_p$, y sea $\kappa = (k_1, k_2, \ldots, k_p)$ una partición de un entero positivo k en no más de p partes. El polinomio zonal de A correspondiente a la particion κ , $C_\kappa(A)$, es un polinomio homogèneo y simètrico de grado k en los autovalores $\lambda_1,\lambda_2,\ldots,\lambda_p$, tal que:

(i) El tèrmino de mayor grado de $C_\kappa(A)$ es

$$\lambda_1^{k_1}\ldots\ldots\lambda_p^{k_p}$$

es decir,

$$C_\kappa(A) = d_\kappa \lambda_1^{k_1}\ldots\lambda_p^{k_p} + \text{tèrminos de menor grado}$$

donde d_κ es una constante real.

(ii) $C_\kappa(A)$ es una autofunción del operador diferencial Δ_A dado por

$$\Delta_A = \sum_{i=1}^{p} \lambda_i^2 \frac{\partial^2}{\partial\lambda_i^2} + \sum_{i=1}^{p} \sum_{\substack{j=1 \\ j\neq i}}^{p} \frac{\lambda_i^2}{\lambda_i-\lambda_j}\frac{\partial}{\partial\lambda_i}$$

(iii) Cuando κ varia a travès de todas las particiones de k, los polinomios zonales tienen coeficiente unidad en el desarrollo de $(\mathrm{tr}\ A)^k$, es decir,

$$(\mathrm{tr}\ A)^k = (\lambda_1+\lambda_2+\ \ldots\ +\lambda_p)^k = \sum_\kappa C_\kappa(A)$$

Las propiedades fundamentales de los polinomios zonales se encuentran en los trabajos de James (1960, 1961, 1964, 1968, 1973, 1976), Constantine (1963, 1968), Kushner y Meisner (1984).

Se define la función hipergeomètrica de argumento matricial como

$$ {}_rF_s(a_1,a_2,\ldots,a_r;b_1,b_2,\ldots,b_s;A) = $$
$$ = \sum_{k=0}^{\infty} \sum_{\kappa} \frac{(a_1)_\kappa \cdots\cdots (a_r)_\kappa}{(b_1)_\kappa \cdots\cdots (b_s)_\kappa} \frac{C_\kappa(A)}{k!} $$

donde $\sum_{\kappa}$ denota la suma a través de todas las particiones $\kappa = (k_1,k_2,\ldots,k_p)$, $k_1 \geq k_2 \geq \ldots\ldots \geq k_p \geq 0$, de k, $C_\kappa(A)$ es el polinomio zonal de la matriz A correspondiente a la partición κ, y $(a)_\kappa$ son los coeficientes hipergeomètricos generalizados dados por:

$$ (a)_\kappa = \prod_{i=1}^{p} \left(a - \frac{1}{2}(i-1)\right)_{k_i} $$

siendo

$$ (a)_{k_i} = a(a+1)\ldots\ldots(a+k_i-1), \qquad (a)_0 = 1 $$

Herz (1955), Constantine (1963).

Como el estadístico básico definido en el apartado anterior es el máximo autovalor de una matriz con distribución de Wishart, se va a obtener a continuación la distribución de dicho máximo autovalor.

Sea S una matriz con distribución $W_p(n,\Sigma)$, y autovalores $\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_p \geq 0$. James (1960), obtiene la función de densidad conjunta de los autovalores $\lambda_1,\lambda_2,\ldots,\lambda_p$, la cual viene dada por:

$$ f(\lambda_1,\lambda_2,\ldots,\lambda_p) = H^* \sum_{k=0}^{\infty} \sum_{\kappa} \frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{C_\kappa(I_p)} |\Lambda|^{\frac{n-p+1}{2}} C_\kappa(\Lambda) \prod_{i<j} (\lambda_i - \lambda_j) $$

para $\lambda_1 > \lambda_2 \ldots . > \lambda_p > 0$, donde $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ y

$$H^* = \frac{\pi^{p^2/2} \, [\Gamma_p(p/2) \; \Gamma_p(n/2)]^{-1}}{|\Sigma|^{n/2} \, 2^{np/2}}$$

siendo Γ_p la función Gamma multivariante.

La distribución de λ_1 vendrá dada entonces por la función de densidad marginal de λ_1 en la anterior distribución conjunta. Antes de pasar a calcular dicha **función de densidad marginal, se realiza el siguiente** cambio de variable

$$\lambda_1 = \lambda_1$$
$$l_i = \frac{\lambda_i}{\lambda_1} \qquad i=2,3,\ldots,p$$

La función de densidad conjunta de la nueva va riable aleatoria $(\lambda_1, l_2, \ldots, l_p)$ vendrá dada por:

$$g(\lambda_1, l_2, \ldots, l_p) = f(\lambda_1, \lambda_1 l_2, \ldots, \lambda_1 l_p) \; |J|$$

siendo J el Jacobiano de la transformación inversa, el cual viene dado por:

$$\lambda_1^{p-1}$$

y, teniendo en cuenta que:

$$|\Lambda| = \lambda_1^p \; |\Lambda_1| \qquad \text{con } \Lambda_1 = \text{diag}(l_2, l_3, \ldots, l_p)$$

$$\prod_{i<j}(\lambda_i-\lambda_j) = \prod_{j=2}^{p}(1-l_j)\prod_{i<j}(l_i-l_j)\,\lambda_1^{p(p-1)/2}$$

$$C_\kappa(\Lambda) = \lambda_1^k C_\kappa(\Lambda_1^p) \qquad \text{con } \Lambda_1^p = \mathrm{diag}(1,l_2,\ldots,l_p)$$

se tiene que la función de densidad conjunta de las nue vas variables $\lambda_1, l_2,\ldots,l_p$, viene dada por:

$$g(\lambda_1,l_2,\ldots,l_p)=H^*\sum_{k=0}^{\infty}\sum_{\kappa}\frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{k!\,C_\kappa(I_p)}\;\lambda_1^{\frac{pn-p(p-1)}{2}}\;|\Lambda_1|^{\frac{n-p-1}{2}}\cdot$$

$$\cdot\, C_\kappa(\Lambda_1^p)\;\lambda_1^k\;\lambda_1^{p(p-1)/2}\prod_{j=2}^{p}(1-l_j)\prod_{i<j}(l_i-l_j)\;\lambda_1^{p-1}$$

para $\lambda_1>0$, $1>l_2>l_3>\ldots>l_p>0$.

La función de densidad de λ_1 se obtendrá integrando esta última función de densidad conjunta respecto de $l_2,l_3,\ldots,l_p$:

$$f(\lambda_1) = H^*\sum_{k=0}^{\infty}\sum_{\kappa}\qquad\lambda_1^{p\frac{n}{2}k-1}$$

$$\int_{1>l_2>\ldots>l_p>0}|\Lambda_1|^{\frac{n}{2}-\frac{p+1}{2}}\,C_\kappa(\Lambda_1^p)\prod_{j=2}^{p}(1-l_j)\prod_{i<j}(l_i-l_j)\;dl_2\ldots dl_p$$

Para calcular esta última integral se hace uso del siguiente resultado (Sugiyama, 1966):

$$\int_{1>a_1>\ldots>a_p>0}|A|^{t-\frac{p+1}{2}}\,|I-A|^{u-\frac{p+1}{2}}\,C_\kappa(A)\prod_{i<j}(a_i-a_j)\;da_1\ldots da_p =$$

$$=\frac{\Gamma_p(p/2)}{\pi^{p^2/2}}\,\frac{\Gamma_p(t,\kappa)\;\Gamma_p(u)}{\Gamma_p(t+u,\kappa)}\,C_\kappa(I_p)$$

con $A=diag(a_1,a_2,\dots,a_p)$ y $1>a_1>a_2>\dots>a_p>0$.

Si $u=(p+1)/2$, la integral anterior se expresa de la forma:

$$\int_{1>a_1>\dots>a_p>0} |A|^{t-\frac{p+1}{2}} C_\kappa(A) \prod_{i<j} (a_i - a_j)\, da_1\dots da_p$$

y mediante el cambio de variable:

$$a_1 = a_1$$
$$l_i = \frac{a_i}{a_1} \qquad i=2,3,\dots,p$$

cuyo Jacobiano es:

$$|J| = a_1^{p-1}$$

la integral anterior se puede expresar como:

$$\int_0^1 a_1^{pt+k-1}\, da_1 \cdot \Big(\int_{1>l_2>\dots>l_p>0} |\Lambda_1|^{t-\frac{p+1}{2}} C_\kappa(\Lambda_1^p) \prod_{j=2}^{p} (1-l_j) \prod_{i<j} (l_i-l_j)\, dl_2\dots dl_p$$

y mediante la igualdad:

$$\int_0^1 a_1^{pt+k-1}\, da_1 = \frac{1}{pt+k}$$

se tiene que:

$$\int_{1>l_2>\dots>l_p>0} |\Lambda_1|^{t-\frac{p+1}{2}} C_\kappa(\Lambda_1^p) \prod_{j=2}^{p} (1-l_j) \prod_{i<j} (l_i-l_j)\, dl_2\dots dl_p =$$

$$= \frac{(pt+k)\, \Gamma_p(p/2)}{\pi^{p^2/2}} \, \frac{\Gamma_p(t,\kappa)\, \Gamma_p((p+1)/2)}{\Gamma_p(t+\frac{p+1}{2},\kappa)} C_\kappa(I_p)$$

y, haciendo $t=n/2$, se tiene que:

$$f(\lambda_1)=\frac{\pi^{p^2/2}[\Gamma_p(p/2)\ \Gamma_p(n/2)]^{-1}}{|\Sigma|^{n/2}\ 2^{np/2}}\ \frac{\Gamma_p(p/2)\ \Gamma_p((p+1)/2)}{\pi^{p^2/2}}\ .$$

$$\cdot\sum_{k=0}^{\infty}\sum_{\kappa}\frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{k!\,C_\kappa(I_p)}\ \lambda_1^{p\frac{n}{2}+k-1}\ \frac{(p\frac{n}{2}+k)\ \Gamma_p(\frac{n}{2}+\kappa)}{\Gamma_p(\frac{n+p+1}{2},\kappa)}\quad C_\kappa(I_p)\ =$$

$$=\frac{\Gamma_p(\frac{p+1}{2})}{|\Sigma|^{n/2}\ 2^{np/2}\ \Gamma_p(\frac{n}{2})}\ \sum_{k=0}^{\infty}\sum_{\kappa}\frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{k!}\cdot\frac{(p\frac{n}{2}+k)\ \Gamma_p(\frac{n}{2},\kappa)}{\Gamma_p(\frac{n+p+1}{2},\kappa)}\ \lambda_1^{p\frac{n}{2}+k-1}$$

que también se puede expresar

$$f(\lambda_1)\ =\ H\sum_{k=0}^{\infty}\sum_{\kappa}\ (p\frac{n}{2}+k)\ \frac{(\frac{n}{2})_\kappa}{(\frac{n+p+1}{2})_\kappa}\ \frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{k!}\ \lambda_1^{\frac{np}{2}+k-1}\qquad \lambda_1>0$$

con

$$H=\frac{\Gamma_p((p+1)/2)}{2^{np/2}\quad |\Sigma|^{n/2}\quad \Gamma_p((n+p+1)/2)}$$

Integrando esta última función de densidad, se obtiene la función de distribución de λ_1.

$$F(\lambda_1)\ =\ \int_0^1 f(t)\ dt\ =$$

$$=\ H\sum_{k=0}^{\infty}\sum_{\kappa}\ (pn/2+k)\frac{(\frac{n}{2})_\kappa}{(\frac{n+p-1}{2})_\kappa}\ \frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{k!}\ \int_0^1 t^{\frac{np}{2}+k-1}\ dt\ =$$

$$=\ H\sum_{k=0}^{\infty}\sum_{\kappa}\frac{(\frac{n}{2})_\kappa}{(\frac{n+p+1}{2})_\kappa}\ \frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{k!}\ \lambda_1^{\frac{np}{2}+k}\ =$$

$$=\ H\ \lambda_1^{np/2}\ \sum_{k=0}^{\infty}\sum_{\kappa}\frac{(\frac{n}{2})_\kappa}{(\frac{n+p+1}{2})_\kappa}\ \frac{C_\kappa(-\frac{1}{2}\lambda_1\Sigma^{-1})}{k!}$$

Luego

$$F(\lambda_1) = H\ \lambda_1^{np/2}\ {}_1F_1(n/2;(n+p+1)/2;\tfrac{1}{2}\lambda_1\Sigma^{-1}) \qquad \lambda_1 > 0$$

y, por las propiedades de las funciones hipergeomètricas de argumento matricial,

$$F(\lambda_1) = H\ \lambda_1^{np/2}\ \mathrm{etr}(\tfrac{-1}{2}\lambda_1\Sigma^{-1})\ {}_1F_1(\tfrac{p+1}{2};\tfrac{n+p+1}{2};\tfrac{1}{2}\lambda_1\Sigma^{-1}) \qquad \lambda_1 > 0$$

De los resultados anteriores, se deduce que la función de densidad y la función de distribución del estadístico

$$T_r = \lambda_1(S^*)$$

vienen dadas, respectivamente, por:

$$f(t) = H \sum_{k=0}^{\infty} \sum_{\kappa} (rp/2+k) \frac{(r/2)_\kappa}{(\frac{r+p+1}{2})_\kappa} \frac{C_\kappa(-\frac{1}{2}\Sigma^{-1})}{k!} t^{\frac{rp}{2}+k-1} \qquad t>0$$

$$F(t) = H\ t^{rp/2}\ \mathrm{etr}(-\tfrac{1}{2}t\Sigma^{-1})\ {}_1F_1(\tfrac{p+1}{2};\tfrac{r+p+1}{2};\tfrac{1}{2}t\Sigma^{-1}) \qquad t>0$$

con

$$H = \frac{\Gamma_p((p+1)/2)}{2^{rp/2}\ |\Sigma|^{r/2}\ \Gamma_p((r+p+1)/2)}$$

3.5. TECNICA DE IDENTIFICACION DE OUTLIERS.

En este apartado se aplican a la identificación de outliers los resultados obtenidos hasta ahora. Para la identificación de un outlier en una muestra de tamaño n, $x_1, x_2, \ldots, x_n$, extraida de una población Normal $N_p(\mu,\Sigma)$, se establecen las hipótesis:

$$H_0\colon\ x_i \in N_p(\mu,\Sigma) \qquad \forall\, i,\ i=1,2,\ldots,n$$
$$H_1\colon\ \exists\ x_i \notin N_p(\mu,\Sigma)$$

El estadístico a utilizar para realizar el contraste es:

$$T_1^{*} = \max_{1\le i\le n} T_{1,i}$$

con

$$T_{1,i} = d(S_{(n-1)}^{(i)}, S_{(n)}) \qquad i=1,2,\ldots,n$$

que bajo la hipótesis nula verificará

$$d(S_{(n-1)}^{(i)}, S_{(n)}) \le q_1 \qquad \forall\, i,\ i=1,2,\ldots,n$$

para un cierto $q_1>0$, $q_1 \in R$, por lo que

$$\max_{1\le i\le n}(d(S_{(n-1)}^{(i)}, S_{(n)})) \le q_1$$

Así, la región crítica correspondiente a este contraste es

$$T_1^* = \max_{1\le i\le n} T_{1,i} > T_\alpha$$

con $T\alpha$ solución de la ecuación

$$P(\max_{1\le i\le n} T_{1,i} > T_\alpha) = \alpha$$

la que se obtiene mediante la aproximación

$$P(\max_{1\le i\le n} T_{1,i} > T_\alpha) \le n\ P(T_1 > T_\alpha)$$

basada en la desigualdad de Bonferroni (Rohatgi, 1976). Así, T_α es la solución de

$$n\ P(T_1 > T_\alpha) = \alpha$$

o bien

$$P(T_1 > T_\alpha) = \alpha/n$$

Por tanto, el punto crítico T_α se obtiene a partir de la distribución de T_1.

El rechazo de la hipótesis nula lleva a afirmar que en la muestra existe al menos una observación outlier. La identificación de la observación que da lugar al rechazo de Ho se consigue determinando el estadístico $T_{1,i}$ que toma el valor máximo en la expresión de T_1^*.

Este procedimiento lleva a determinar la presencia de una observación outlier, pero a veces ocurre que en una muestra de tamaño n existe más de una observación outlier. Para resolver esta situación, se propone el siguiente mètodo que generaliza al anterior.

Para la identificaión de r outliers en una muestra de tamaño n, se establecen las hipótesis:

$$H_0: \ x_i \in N_p(\mu,\Sigma) \qquad \forall\ i,\ i=1,2,\ldots,n$$

$$H_1: \ \exists\ x_{i_j} \notin N_p(\mu,\Sigma) \qquad j=1,2,\ldots,r \quad r<[n/2]$$

El estadístico a utilizar en este contraste de hipótesis es

$$T_r^{*}=\max_{(i_1\ldots i_r)} T_{r,i_1\ldots i_r}$$

con

$$T_{r,i_1\ldots i_r}= d(S_{(n-r)}^{(i_1\ldots i_r)},S_{(n)})$$

donde $(i_1,\ldots,i_r)$ es cualquiera de las $\binom{n}{r}$ r-uplas que se pueden formar con los enteros $1,2,\ldots,n$.

De esta forma se evita el efecto de enmascaramiento, ya que se identifican bloques de outliers simultaneamente.

Como en el caso anterior, bajo la hipótesis nula se verificará

$$d(S^{(i_1\ldots i_r)}_{(n-r)}, S_{(n)}) \leq q_r \qquad \forall\ (i_1,\ldots,i_r)$$

para un cierto $q_r>0$, $q_r \in R$, por lo que

$$\max_{(i_1\ldots i_r)} d(S^{(i_1\ldots i_r)}_{(n-r)}, S_{(n)}) \leq q_r$$

Así, la región crítica correspondiente a este contraste es

$$T^*_r = \max_{(i_1\ldots i_r)} T_{r,i_1,\ldots,i_r} > T_\alpha$$

con T_α solución de la ecuación

$$P(\max_{(i_1\ldots i_r)} T_{r,i_1,\ldots,i_r} > T_\alpha)=\alpha$$

y mediante la desigualdad de Bonferroni, T_α se determinará a partir de la ecuación

$$P(T_r > T_\alpha) = \alpha/\binom{n}{r}$$

Así, T_α se determina a partir de la distribución de T_r.

El rechazar la hipótesis nula conduce a afirmar que en la muestra existen al menos r observaciones outliers. La identificación de estas observaciones se se consigue mediante el estadístico $T_{r,i_1,\dots,i_r}$ que alcanza el máximo en la expresión de T_r^*.

3.6. DETERMINACION DEL PUNTO CRITICO BAJO DISTINTOS SUPUESTOS POBLACIONALES.

Al estudiar una tècnica de identificación de outliers en un modelo Normal p-dimensional hay que tener presente la casuística que se puede presentar sobre los parámetros que caracterizan a dicho modelo poblacional, ya que ello va a dar lugar a diferentes mètodos para la tabulación o determinación del punto crítico.

Del estudio realizado se concluye que la distribución es independiente del vector de medias de dicha población. Es por ello que solo se estudiarán posibles formas de la matriz de varianzas-covarianzas Σ.

Así, para la determinación del punto crítico T_α mediante

$$P(T_1 > T_\alpha) = \alpha/n$$

o mediante

$$P(T_r > T_\alpha) = \alpha / \binom{n}{r}$$

se distinguen los siguientes casos que se pueden presentar sobre Σ.

A) $\Sigma = I_p$.

En este caso, para la determinación de las soluciones de las ecuaciones anteriores, se pueden utilizar las tablas realizadas por Thompson (1962) y Sugiyama (1968).

Las tablas de Thompson proporcionan los percentiles de ordenes

$$0.95,\ 0.975,\ 0.99,\ 0.995$$

de la distribución de λ_1 para p=2 y valores de

$$n=2(1)20(2)30(5)50(10)100$$

con cinco cifras significativas.

Sugiyama proporciona los percentiles de orden

$$0.95 \text{ y } 0.99$$

de la distribución de λ_1/n para

$$n=2(2)50 \quad \text{y} \quad p=2,3,4$$

Hanumara y Thompson (1968) proporcionan valores aproximados de los percentiles de ordenes

$$0.95,\ 0.975,\ 0.99,\ 0.995$$

de la distribución de λ_1, para valores de

$p=2(1)10$ y $n=p(1)10(5)30(10)300$

Pearson y Hartley (1971) proponen unas relaciones aproximadas entre los percentiles de λ_1, que se verifican para todos los valores de n y p. Estas relaciones son:

$$\lambda_{1,0.9} \doteq \lambda_{1,0.95} - \lambda_{1,0.99}$$

$$\lambda_{1,0.975} \doteq 0.55\ \lambda_{1,0.95} + 0.45\ \lambda_{1,0.99}$$

$$\lambda_{1,0.995} \doteq 1.39\ \lambda_{1,0.99} - 0.39\ \lambda_{1,0.95}$$

donde $\lambda_{1,\alpha}$ denota el percentil de orden α de la distribución de λ_1.

Según cada caso concreto se utilizarán unas u otras tablas, teniendo en cuenta los distintos valores de p y n.

B) Σ conocida.

Para esta situación, el cálculo de la solución de las ecuaciones que proporcionan los valores críticos se realiza mediante la función de distribución de T_r, obtenida anteriormente. Para ello, el cálculo de los polinomios zonales se puede hacer a travès de la subrutina POLLY, Mc Laren, (1976).

C) Σ desconocida.

Bajo este supuesto, el cálculo del punto crítico no se puede realizar directamente a partir de la función de distribución de T_r obtenida anteriormente.

Para este cálculo, se propone realizar primeramente la acomodación de outliers a la matriz de varianzas - covarianzas, mediante mètodos de estimación robustos en los que los estimadores mantienen propiedades estadísticas deseables, aún cuando la muestra posiblemente incluya observaciones outliers. Esto es, se trata de estimar Σ mediante estimadores que no sean sensibles a la presencia de outliers. A continuación se exponen dos mètodos de estimación robusta de Σ.

El primer procedimiento que se propone es el dado por Gnanadesikan y Kettenring (1972).

Sea x^* un estimador robusto de localización, y sea

$$d_i = (x_i - x^*)'(x_i - x^*) \qquad i=1,2,\ldots,n$$

el cuadrado de la distancia euclídea de cada observación muestral a dicho estimador robusto.

Se ordénan los valores $d_1, d_2, \ldots, d_n$, en orden creciente, y se selecciona la fracción $1-\alpha$ de observaciones que proporcionan los $(1-\alpha)n$ menores valores de d_i. Sea J el subconjunto de estas observaciones, y

$$A_o = \sum_{x_i \in J} (x_i - x^*)(x_i - x^*)'$$

la matriz de sumas de cuadrados y sumas de productos obtenida a partir de estas observaciones. La fracción α de observaciones no incluidas en el cálculo de A_o debe ser los suficientemente pequeña como para que A_o sea no singular.

A continuación se ordenan las n observaciones $x_1, x_2, \ldots, x_n$ en tèrminos de la forma cuadrática

$$d_i = (x_i - x^*)' A_0^{-1} (x_i - x^*) \qquad i=1,2,\ldots,n$$

y se selecciona la fracción $1-\beta$ de observaciones correspondientes a los $(1-\beta)n$ menores valores de d_i. Sea J^* el subconjunto de tales observaciones. La estimación robusta de la matriz de varianzas-covarianzas se obtiene entonces mediante

$$\hat{\Sigma} = \frac{k}{n(1-\beta)} \sum_{x_h \in J^*} (x_h - x^*)(x_h - x^*)'$$

donde k es una constante que haga que el estimador sea suficientemente insesgado, y β debe ser tal que

$$n(1-\beta) > p$$

y $\hat{\Sigma}$ sea no singular. Generalmente se toma $\alpha=\beta$

A continuación se propone otro procedimiento para obtener una estimación robusta de la matriz de varianzas-covarianzas Σ.

Este mètodo consiste en realizar primeramente una R-ordenación (Barnett, 1976) de las observaciones, de la siguiente forma:

Se dirá que la observación $x_j=(x_{1j},x_{2j},\ldots,x_{pj})'$ es menor que la observación $x_k = (x_{1k},x_{2k},\ldots,x_{pk})'$ si $\|S^{(j)}\|_\infty < \|S^{(k)}\|_\infty$, siendo $S^{(j)}$ y $S^{(k)}$ las matrices de sumas de cuadrados y sumas de productos de la muestra, calculadas sin considerar las observaciones j y k, respectivamente.

Una vez ordenadas las observaciones muestrales según este principio de R-ordenación, se seleccionan las $n(1-\alpha)$ observaciones menores para calcular la matriz de varianzas-covarianzas

$$A = \frac{1}{n(1-\alpha)}\sum_{x_l} (x_l - x^*)(x_l - x^*)'$$

siendo x^* la media muestral obtenida mediante las observaciones seleccionadas para el cálculo de la matriz A.

Para garantizar que A sea no singular, la fracción α de observaciones no consideradas en su cálculo deberá cumplir

$$n(1-\alpha) > p$$

o lo que es lo mismo,

$$\alpha < n(1 - p/n)$$

Una vez obtenida la estimación robusta de Σ por cualquiera de los dos procedimientos anteriores, se procedería como en el apartado B para la obtención del punto crítico.

3.7. SUBRUTINA DE CALCULO DEL ESTADISTICO T_r.

A continuación se proporciona una subrutina para el cálculo del estadístco T_r, y la identificación de los posibles outliers. Al igual que en el capítulo anterior, esta rutina está codificada en BASIC, y su codificación en otro lenguaje tambièn es inmediata.

Para el cálculo del máximo autovalor de la matriz S^* se utiliza el mètodo de la potencia, como lo describe Searle (1982), normalizando en cada iteración el vector aproximación del autovector asociado al máximo autovalor, con objeto de que los valores de sus componentes no varien mucho de una iteración a otra y así disminuir en lo posible los errores de redondeo.

Al igual que la rutina del capítulo anterior, la que se expone a continuación está autodocumentada.

```
63000 REM ***************************************************
63010 REM       Subrutina para el cálculo del estadístico Tr.      *
63020 REM                                                          *
63030 REM   Datos de entrada:                                      *
63040 REM                                                          *
63050 REM          P: Dimensión de la población.                   *
63060 REM          N: Tamaño de la muestra.                        *
63070 REM          X: Matriz PxN cuyas columnas son las observa-   *
63080 REM             ciones muestrales.                           *
63090 REM          R: Número de outliers.                          *
63100 REM                                                          *
63110 REM   Datos de salida:                                       *
63120 REM                                                          *
63130 REM         TR: Valor del estadístico.                       *
63140 REM          N: Vector de índices de dimensión R, cuyos      *
63150 REM             elementos indican las columnas de X que      *
63160 REM             posiblemente sean outliers.                  *
63170 REM ***************************************************
63180 REM
63190 DIM IND(R),XMN(P),XMR(P),XMNR(P),S(P,P),W0(P),W1(P),W2(P)
63200 DIM N(R)
63210 REM ******************************
63220 REM  Cálculo del vector n·x̄(n).  *
63230 REM ******************************
63240 FOR I=1 TO P
63250 XMN(I)=0
63260 FOR J=1 TO N
63270 XMN(I)=XMN(I)+X(I,J)
63280 NEXT J
63290 NEXT I
63300 REM ****************
63310 REM   Cálculo de (n r) *
63320 REM ****************
63330 NC=1
63340 FOR I=1 TO R
63350 NC=NC*(N-I+1)/I
63360 NEXT I
63370 REM **************************************************
63380 REM              Cálculo del valor del estadístico.      *
63390 REM **************************************************
63400 IFAULT=0
63410 FOR L=1 TO NC
63420 GOSUB 63800
```

```
63430 REM ***************************
63440 REM  Cálculo de X̄(r) y X̄(n-r). *
63450 REM ***************************
63460 FOR I=1 TO P
63470 XMR(I)=0
63480 FOR J=1 TO R
63490 K=IND(J)
63500 XMR(I)=XMR(I)+X(I,K)
63510 NEXT J
63520 XMNR(I)=(XMN(I)-XMR(I))/(N-R)
63530 XMR(I)=XMR(I)/R
63540 NEXT I
63550 REM ****************
63560 REM  Cálculo de S*. *
63570 REM ****************
63580 FOR I=1 TO P
63590 FOR J=I TO P
63600 S(I,J)=0
63610 FOR K=1 TO R
63620 II=IND(K)
63630 S(I,J)=S(I,J)+(X(I,II)-XMR(I))*(X(J,II)-XMR(J))
63640 NEXT K
63650 S(I,J)=S(I,J)+R*(N-R)*(XMNR(I)-XMR(I))*(XMNR(J)-XMR(J))/N
63660 S(I,J)=S(J,I)
63670 NEXT J
63680 NEXT I
63690 REM ***********************************
63700 REM  Cálculo del valor del estadístico. *
63710 REM ***********************************
63720 GOSUB 64160
63730 IF EIGMAX<TR THEN 63780
63740 TR=EIGMAX
63750 FOR I=1 TO R
63760 N(I)=IND(I)
63770 NEXT I
63780 NEXT L
63790 RETURN
63800 REM ******************************************************
63810 REM    Subrutina para el cálculo de la siguiente combina- *
63820 REM    ción a una dada.                                   *
63830 REM                                                       *
63840 REM    Datos de entrada:                                  *
63850 REM                                                       *
63860 REM       IFAULT: Variable indicadora.                    *
63870 REM               Si IFAULT=0, primera combinación.       *
63880 REM               Si IFAULT<>0, siguiente combinación.    *
63890 REM                                                       *
63900 REM    Datos de salida:                                   *
63910 REM                                                       *
63920 REM       IND: Vector de dimensión R, cuyos elementos     *
63930 REM            indican la combinación de los enteros      *
63940 REM            1,2,...,N que se ha generado.              *
63950 REM ******************************************************
63960 REM
63970 IF IFAULT<>0 THEN 64030
63980 FOR I=1 TO R
```

```
63990 IND(I)=I
64000 NEXT
64010 IFAULT=1
64020 RETURN
64030 FOR I=1 TO R
64040 I1=R-I+1
64050 IF IND(I1)<N-R+I1 THEN 64090
64060 NEXT
64070 IFAULT=-1
64080 RETURN
64090 IND(I1)=IND(I1)+1
64100 IF I1>=R THEN 64140
64110 FOR I=I1+1 TO R
64120 IND(I)=IND(I-1)+1
64130 NEXT
64140 IFAULT=IFAULT+1
64150 RETURN
64160 REM *****************************************************
64170 REM      Cálculo del máximo autovalor de una matriz      *
64180 REM                                                      *
64190 REM   Datos de entrada:                                  *
64200 REM           S: matriz de orden PxP.                    *
64210 REM           P: dimensión de la matriz.                 *
64220 REM                                                      *
64230 REM   Datos de salida:                                   *
64240 REM           EIGMAX: máximo autovalor de la matriz A.   *
64250 REM *****************************************************
64260 REM **********************
64270 REM  Autovector inicial. *
64280 REM **********************
64290 FOR I=1 TO P
64300 WO(I)=1
64310 NEXT
64320 REM ********************
64330 REM  Proceso iterativo *
64340 REM ********************
64350 FOR I=1 TO P
64360 W1(I)=0
64370 FOR J=1 TO P
64380 W1(I)=W1(I)+S(I,J)*WO(J)
64390 NEXT
64400 NEXT
64410 REM *******************************
64420 REM  Normalización del autovector *
64430 REM *******************************
64440 WMAX=W1(I)
64450 FOR I=2 TO P
64460 IF W1(I)>WMAX THEN WMAX=W1(I)
64470 NEXT
64480 FOR I=1 TO P
64490 W2(I)=W1(I)/WMAX
64500 NEXT
64510 REM **********************************************************
64520 REM  Se comprueba si se ha alcanzado la precisión requerida. *
64530 REM **********************************************************
64540 FOR I=1 TO P
```

```
64550 IF ABS(W2(I)-WO(I))>0.00001 THEN 64620
64560 NEXT
64570 EIGMAX=WMAX
64580 RETURN
64590 REM ***********************
64600 REM  Siguiente iteración. *
64610 REM ***********************
64620 FOR I=1 TO P
64630 WO(I)=W2(I)
64640 NEXT
64650 GOTO 64350
```

REFERENCIAS BIBLIOGRAFICAS

REFERENCIAS BIBILOGRAFICAS

Aho, A.V. - J.E. Hopcroft - J.D. Ullman (1983). Data structures and algorithms.
Addison-Wesley, Massachusetts.

Andrews, D.F. (1972). Plots of high-dimensional data.
Biometrics, 28, 125-136.

Anscombe, F.J. - J.W. Tukey (1963). The examination and analysis of residuals.
Technometrics, 5, 141-160.

Barnett, V.D. (1976). The ordering of multivariate data (with Discussion).
J. Roy. Statist. Soc. A, 139,

Barnett, V.D.- T. Lewis (1984). Outliers in Statistical Data
John Wiley & Sons, New York.

Beckman, R.J. - R.D. Cook (1983). Outlier......s.
Technometrics, 25, 119-163.

Bernoulli, D. (1777). Dijudicatio maxime probabilis plurium observationum discrepantium atque verismillima inductio inde formanda.
Academiae Scientorum Petropolitanae, 1, 3-33.

Boscovich, R.J. (1755). De litteraria expeditione per pontificiam ditionem, et sypnosis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa.
Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii, 4, 353-396.

Box, G.E.P. - M.E. Muller (1958). A note on the generation of random normal deviates.
Ann. Math. Stat., 29, 610-611.

Collet, D. -T. Lewis (1976). The subjetive nature of outlier rejection procedures.
Applied Statistics, 25, 228-237.

Constantine, A.G. (1963). Some noncentral distribution problem in multivariate analysis.
Ann. Math. Statist., 34, 1270-1285.

Constantine, A.G. (1966). The distribution of Hotelling's generalized T^2.
Ann. Math. Statist., 37, 215-225.

Devlin, S.J. - R. Gnanadesikan - J.R. Kettenring (1975). Robust estimation and outlier detection with correlation coefficients.
Biometrika, 62, 531-545.

Edgeworth, F.Y. (1887). On discordant observations.
Philosophical Magazine, 23, 364-375.

Ferguson, T.S. (1961). On the rejection of outliers.
Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1.

Fishman , G.S. (1978). Principles of discrete event simulation.
John Wiley & Sons. New York.

Glaisher, J.W.L. (1873). On the rejection of discordant observations.
Monthly Notices Roy. Astr. Soc., 33, 391-402.

Gnanadesikan, R. (1973). Graphical methods for informal inference in multivariate data analysis.
Bull. Ins. Statis. Ins., 45, 195-206.

Gnanadesikan, R. (1977). Methods for statistical data analysis of multivariate observations.
John Wiley & Sons. New York.

Gnanadesikan, R. - J.R. Kettering (1972). Robust estimates, residuals and outlier detection with multiresponse data.
Biometrics, 28, 81-124.

Grubbs, F.E. (1950). Sample criteria for testing outlying observations.
Ann. Math. Statist., 21, 27-58.

Grubbs, F.E. (1969). Procedures for detecting outlying observations in samples.
Technomterics, 11,1-21.

Gumbel, E.J. (1960). Discussion on rejection of outliers by Anscombe.
Technometrics, 2,

Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity - a Bayesian approach.
Technometrics, 15, 723-738.

Hanumara, R.C. - W.A. Thompson (1968). Percentage points of the extreme root of a Wishart matrix.
Biometrika, 55, 505-512.

Hawkins,D.M. (1974). The detection of errors in multivariate data using principal components.
J. Amer. Statist. Ass., 69, 340-344.

Hawkins, D.M. (1980). Identification of outliers.
Chapman and Hall. London.

Herz, C.S. (1955). Bessel functions of matrix argument.
Ann. Math., 61, 474-523.

James, A.T. (1960). The distribution of the latent roots of the covariance matrix.
Ann. Math. Statist., 31, 151-158.

James, A.T. (1961a). The distribution of noncentral means with know covariance matrix.
Ann. Math. Statist., 32, 874-882.

James, A.T. (1961b). Zonal polynomials of the real positive definite symetric matrices.
Ann. Math., 74, 456-469.

James, A.T. (1964). Distribution of matrix variates and latent roots derived from normal samples.
Ann. Math. Statist., 35, 475-501.

James, A.T. (1968). Calculation of zonal polynomial coefficicients by use of the Laplace-Beltrani operator.
Ann. Math. Statist., 39, 1711-1718.

James, A.T. (1973). The variance information manifold and the functions on it.
En Multivariate Analysis (P.R. Krishnaiah, editor), Vol. III, 157-169. Academic Press. New York.

James, A.T. (1976). Special functions of matrix and single argument in statistics.
En Theory and Applications of Special Functions. (R.A. Askey, editor) 497-520. Academic Press. New York.

Kennedy, W.J. - J. E. Gentle (1980). Statistical computing.
Marcel Dekker. New York.

Kshirsagar, A.M. (1972). Multivariate Analysis.
Marcel Dekker. New York.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications.
Biometrika, 53, 519-530.

Mardia, K.V. - J. T. Kent - J. M. Bibby (1979). Multivariate Analysis.
Academic Press. London.

Nair, K.R. (1948). The distribution of extreme deviate from the sample mean and its studentized form. Biometrika, 35, 118-144.

Pearson, E.S. - C. Chandra Sekar (1936). The efficiency of statisticals tools and a criterion for the rejection of outlying observations. Biometrika, 28, 308-320.

Pearson, E.S. - H.O. Hartley (1971). Biometrika Tables for Statisticians, 2. Cambridge University Press.

Peirce, B. (1852). Criterion for the rejection of doubtful observations. Astr. J., 2, 161-163.

Rubinstein, R.Y. (1981). Simulation and the Monte Carlo method. John Wiley & Sons. New York.

Schwager, S.J. - B. Margolin, B. (1982). Detection of multivariate normal outliers. Ann. Statist., 10, 943-954.

Searle, S.R. (1982). Matrix algebra useful for statistics. John Wiley & Sons. New York.

Siotani, M. (1959). The extreme value of the generalizel distances of the individual points in the multivariate normal sample. Ann. Inst. Statist. Math. Tokyo, 10, 183-208.

Stewart, G.W. (1973). Introduction to matrix computations. Academic Press. London.

Stone, E.J. (1868). On the rejection of discordant observations. Monthly Notices Roy. Astr. Soc., 28, 165-168.

Sugiyama, T. (1966). On the distribution of the largest latent root and corresponding latent vector for principal component analysis. Ann. Math. Statist., 37, 995-1001.

Sugiyama, T. (1968). Percentile points of the largest latent root of a matrix and power calculations for testing the hypothesis $\Sigma = I$. Mimeo Series No. 590. Institute of Statistics, University of North Carolina, Chapel Hill.

Thompson, W.A. (1962). Estimation of dispersion parameters. Journal of Research of the National Bureau of Standards, Series B, 60, 161-164.

Thompson, W.R. (1935). On a criterion for the rejection of observations and the distribution of the ratio of the sample standard deviation.
Ann. Math. Statist., 6, 214-219.

Wilks, S.S. (1962). Mathematical Statistics.
John Wiley & Sons. New York.

Wilks, S.S. (1963). Multivariate statistical outliers.
Sankhya, A, 25, 407-426.

UNIVERSIDAD DE SEVILLA

Reunido el Tribunal integrado por los abajo firmantes en el día de la fecha, para juzgar la Tesis Doctoral de D. Luis L. Perez de los Rios titulada "Identificación de outliers en muestras multivariantes"

acordó otorgarle la calificación de APTO CUM LAUDE

Sevilla, 14 de Julio 1.987

El Vocál, El Vocal, El Vocal,

El Presidente El Secretario, El Doctorado,