UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMATICAS

ANALISIS CUALITATIVO

DE DATOS ESTADISTICOS

Memoria dirigida por:

Prof.Dr.D. Joaquin Muñoz Garcia

Prof.Dr.D. Antonio Pascual Acosta

Memoria presentada por:

Juan Luis Moreno Rebollo

Autorizo la entrega de esta
memoria

UNIVERSIDAD DE SEVILLA

FACULTAD DE MATEMATICAS

ANALISIS CUALITATIVO DE DATOS ESTADISTICOS

Visado en Sevilla a
21 de Mayo de 1987

Fdo.Prof.Dr.D. Joaquin Muñoz Garcia

Fdo.Prof.Dr.D. Antonio Pascual Acosta

Memoria que presenta Juan L. Moreno Rebollo para optar al grado de Doctor en Ciencias Matematicas

Fdo. Juan Luis Moreno Rebollo

INDICE

INTRODUCCION

Hacia la segunda mitad del siglo XVIII , ciertos astronomos aconsejan la eliminacion de las observaciones extremas (valores maximo y minimo) resultantes de sus cuantificaciones experimentales, argumentando para ello, que estas observaciones pueden perturbar las inferencias a realizar.

La idea basica de esta practica experimental, puede considerarse como una primera aproximacion a un problema que posteriormente ha adquirido una gran importancia dentro de la estadistica, como es, el analisis de la calidad de las observaciones experimentales, lo cual es basico para garantizar la maxima fiabilidad de las conclusiones.

Sin embargo, el procedimiento empleado para la depuracion de los datos, antes referido,no es el mas correcto, pues se basa unicamente en la subjetividad del experimentador. Por esta razon, diversos autores se dedican a la busqueda de procedimientos que corrijan en cierta medida la falta de objetividad, y no es hasta la primera mitad del siglo XX cuando aparecen metodos con cierto fundamento teorico. A partir de entonces, surgen numerosas tecnicas, que presentan, entre otros, el inconveniente de carecer de una base formal comun , y que dan lugar a lo que hoy en dia se conoce como "Analisis Estadisticos de Outliers" .

En el Capitulo I , se introduce un esquema que permite analizar de un modo generico los distintos errores que pueden afectar a las observaciones experimentales. Asimismo, se destacan los rasgos e inconvenientes principales que presentan las distintas definiciones que sobre el termino outlier han dado diversos autores,y se propone una nueva definicion para el mismo. Por ultimo, se tipifica la problematica que presentan las Tecnicas de Identificacion de Outliers.

Con el objeto de realizar un analisis cualitativo de los datos experimentales, en el Capitulo II, se propone un marco general -basado en la idea de que "el comportamiento anomalo de una observacion depende en gran medida

de su relacion con las restantes observaciones obtenidas bajo condiciones similares y del objetivo que se persigue en la experimentacion"- , que permite obtener una funcion que cuantifica las desviaciones de las observaciones entre si, respecto de un criterio dependiente del analisis que se desea realizar con los datos.

El esquema propuesto en el Capitulo II, aplicado al caso particular del Analisis de Outliers, presenta, como se recoge en las conclusiones del Capitulo III, ciertas ventajas respecto del enfoque clasico, entre las que cabe destacar, que algunos estadisticos en los que se fundamentan las Tecnicas de Identificacion de Outliers, se obtienen a partir de una base formal comun.

CAPITULO I

OUTLIERS .

PROBLEMATICA DE LAS TECNICAS DE IDENTIFICACION .

INTRODUCCION

El estudio de los fenomenos naturales lleva consigo un proceso de codificacion de los elementos de la naturaleza.

El conjunto de codigos resultantes da lugar a lo que se conoce como masa de datos, a partir de la cual se pretende obtener conclusiones sobre el fenomeno que se analiza. Sin embargo, los datos experimentales, no deben considerarse en ningun caso , como valores absolutamente fidedignos del experimento en cuestion, ya que pueden estar afectados por diversos errores o variabilidades. Por tanto, una de las principales etapas de cualquier analisis estadisco de datos, es la de investigar la calidad de las observaciones, ya que si no se detectan los

errores existentes en las mismas, los procedimientos estadisticos que se apliquen posteriormente pueden conducir a conclusiones inexactas.

Para abordar este problema, en este capitulo, se introduce un esquema que permite analizar, de un modo generico, los diversos errores que pueden presentarse en las observaciones resultantes de una experimentacion.

Los errores que se definen el epigrafe 1.1 , dan lugar a diferentes tipos de observaciones erroneas, sin embargo, se concluye que desde un punto de vista practico, al menos inicialmente, resulta imposible tal diferenciacion, por lo que se las denomina, globalmente, con el termino anglosajon OUTLIER.

Se recogen los rasgos principales que caracterizan las definiciones que de outlier han dado diversos autores, indicandose los inconvenientes que presentan, por lo que se propone una nueva definicion para este termino.

Por ultimo, se realiza un estudio general de las distintas tecnicas existentes para abordar el problema de la presencia de outliers en una masa de datos experimentales, analizando con especial interes las Tecnicas de Identificacion, y la problematica actualmente asociada a las mismas.

1.1 ERRORES EN LAS OBSERVACIONES EXPERIMENTALES

La estadistica se puede considerar como aquella ciencia que permite extraer informacion relevante y util de un conjunto de datos experimentales.

Al analizar esta definicion,se han de tener presentes las afirmaciones que hacen diversos autores , entre los que se pueden citar Anscombe (1960),Beckman-Cook (1983), en el sentido de que las observaciones resultantes de una experimentacion no deben considerarse en ningun caso como valores absolutamente fidedignos del experimento en cuestion , ya que estas pueden estar afectadas por diversos errores o variabilidades.Por esta razon existen autores como Anscombe (1960),Hampel (1968),Grubbs (1969), Hawkins (1980) , Beckman y Cook (1983), Barnett y Lewis (1984),Hampel et al (1986) que se preocupan del analisis de las diferentes fuentes de variabilidad o error que pueden influir en los valores muestrales. Sin embargo , cabe resaltar que entre estos analisis,unos son bastante incompletos y poco formales en sus consideraciones y otros exponen situaciones demasiado particulares.

Esto motiva que se introduzca un esquema que permita analizar de un modo mas generico y preciso, los posibles errores que se pueden presentar en una experimentacion y la forma en que estos influyen en las observaciones de que dispone el estadistico.

La realizacion de un experimento lleva consigo una tripleta de elementos,que se representara por (Ω ,X,E), donde Ω es la fuente o naturaleza que provee los elementos w_i , i=1,...,n y X es el medio por el cual el experimentador, E, cuantifica o transforma los elementos w_i , i=1,...,n en las observaciones x_i ,i=1,...,n de las que dispondra el estadistico para analizar el experimento.

Segun este esquema , es posible afirmar , que las variabilidades o errores que pueden influir en las observaciones se pueden clasificar en:

- Variabilidad de la fuente o naturaleza
- Error del medio
- Error del experimentador

1.1.1 VARIABILIDAD DE LA FUENTE O NATURALEZA

Es la que se manifiesta en las observaciones al recorrer las mismas el soporte de la poblacion bajo estudio . Tal variacion ha de ser considerada como un comportamiento natural de los datos.

1.1.2 ERROR DEL MEDIO.

Se produce cuando no se dispone de la tecnica adecuada,o cuando no existe un procedimiento que permita realizar de forma exacta la transformacion del espacio empirico al espacio matematico.

En este error se incluye , por ejemplo, el ocasionado por el redondeo forzoso que se ha de realizar cuando se trabaja con variables continuas.

1.1.3 ERROR DEL EXPERIMENTADOR

Este error es atribuible al propio experimentador y se puede presentar de distintas formas:

- Error de informacion:Se comete al suponer un modelo matematico no adecuado o preciso para la poblacion, o al admitir informacion o hipotesis iniciales erroneas.

- Error de planificacion: Por el cual el experimentador no llega a delimitar de forma correcta el espacio Ω y realiza el experimento sobre un espacio distinto $\Omega \cup \Omega'$.

  En este tipo de errror se contempla el ocasionado al tomar muestras sesgadas o al incluir individuos no pertenecientes a la poblacion bajo estudio.

- Error de realizacion: Se comete al llevar a cabo valoraciones erroneas de las transformaciones de los elementos de .

  En este tipo de error se incluyen las transcripciones erroneas de datos , falsas lecturas,etc..

1.2 OUTLIERS

Las variabilidades o errores definidos en el epigrafe 1.1 se suelen presentar en los datos resultantes de una experimentacion por lo que en toda muestra cabe distinguir dos tipos de observaciones ,segun esten afectadas o no por los mismos . A las primeras , que constituyen el objeto de nuestro estudio ,se les ha denominado de muy diversas formas en la bibliografia existente sobre el tema.

Definiendolas segun el tipo de variabilidad o error que presentan, se adopta la siguiente terminologia:

Observacion atipica: Aquel valor $x \in X$ que presenta una gran variabilidad de tipo inherente.

Observacion Erronea: Aquella observacion que presenta un gran error de medio y/o un gran error del experimentador,el cual puede manifestarse a traves de uno o varios de los errores definidos anteriormente.

La posible presencia de estas observaciones en una experimentacion , hace que algunos autores , entre ellos Chatfield (1985), destaquen,entre las principales etapas que se han de seguir en un analisis estadistico de datos, la de investigar la estructura y calidad de las observa-

vaciones , asi como la de realizar un examen inicial de las mismas , ya que si no se detectan los errores presentes en las observaciones y el estadistico actua suponiendo la ausencia de ellos , los procedimientos estadisticos que se apliquen pueden conducir a conclusiones inexactas.

En el analisis a realizar sobre la calidad de los datos , seria deseable que se distinguiesen las observaciones atipicas de las erroneas , ya que las primeras carecerian de error . Sin embargo , desde un punto de vista practico tal distincion es imposible ya que como muy bien dice Anscombe (1960) , esto es materia de conjetura.

Por tanto , aunque desde un punto de vista teorico se pueda admitir la diferencia entre los distintos tipos de observaciones definidas anteriormente , e incluso se podrian definir muchos mas,segun los errores citados, en la practica solo se considerara de un unico tipo de observaciones , para la que se adoptara el termino anglosajon de OUTLIER.

Antes de definir observacion outlier , se destacan y discuten los rasgos que comunmente se presentan en las definiciones mas utilizadas sobre la misma.

Generalmente en las definiciones dadas se caracteriza el outlier mediante su desviacion del resto de las

observaciones muestrales, Miller (1981) , y en ocasiones aparecen ademas algunas de las siguientes peculiaridades:

a) Intervienen los diversos tipos de error que pueden afectar a las observaciones . En este sentido se pueden citar las dadas por Grubbs (1950) ,Kendall y Buckland (1957) , Anscombe (1960) , Hawkins (1980), Beckman y Cook (1983).

b) Admiten la subjetividad del experimentador para identificarlas . Entre estas se encuentran las de Grubbs (1950) , Kendal y Buckland (1957) , Ferguson (1961),Grubbs (1969),Hawkins (1980), Beckman y Cook (1983),Barnett y Lewis (1984).

La idea de caracterizar al outlier mediante su desviacion del resto de los elementos que componen la muestra parece evidente, ya que , como indican Beckman y Cook (1983):

> "... la fiabilidad probable de una observacion se refleja mediante su relacion con otras observaciones que se obtienen bajo condiciones similares...".

Con respecto a la primera de las peculiaridades , parece necesario que los errores aparezcan en la definicion, de forma implicita o explicita , sin embargo , una definicion formal de este concepto no deberia contener referencia alguna a la influencia subjetiva del experi-

mentador , ya que la presencia de terminos como "despertar sospechas" , "parecer inconsistente" ,... se han de considerar como una reminiscencia de los albores de este problema . Basta citar , a modo de ejemplo , a Boscovich (1755) , quien para determinar la elipticidad de la tierra rechazaba las dos observaciones extremas resultantes de la experimentacion , sin criterio alguno , unicamente por ser la menor y la mayor de las observaciones.

El problema de la subjetividad se pone de manifiesto en trabajos , como el Collet y Lewis (1976) , donde se afirma:

> "... existen comportamientos diferentes entre los individuos en su reaccion a valores sorprendentes e incluso entre los juicios realizados por un mismo individuo en distintas ocasiones...".

Es mas, en las "situaciones mas estructuradas", segun la clasificacion realizada por Barnett y Lewis (1984), o cuando se observan datos multivariantes , resulta en general imposible la influencia subjetiva del experimentador sobre la apreciacion de observaciones outliers.

Esto hace que la definicion que se adopte sobre este termino evite cualquier grado de subjetividad. De hecho, existen autores que ponen en entredicho la influencia del experimentador . Asi Beckman y Cook (1983) muestran

su preocupacion hacia definiciones que dependan de una nocion tan subjetiva como "parecer soprendente".

Por ultimo, la definicion tambien deberia reflejar el hecho de que una observacion sera outlier para un determinado modelo, metodo o criterio , baste citar para ello lo apuntado por Gnanadesikan y Kettering (1972):

> "... la complejidad del caso multivariante sugiere que seria conveniente investigar procedimientos que ofrezcan proteccion contra cualquier tipo de outlier , aunque una aproximacion mas razonable seria la de construir procedimientos que permitan proteger los datos ante situaciones especificas...".

Ademas, un outlier para modelo , metodo o criterio no lo es necesariamente para otro.

Asi pues , teniendo en cuenta las consideraciones realizadas se propone la siguiente definicion para el termino outlier:

> OUTLIER es aquella observacion que siendo atipica y/o erronea , se desvia marcadamente del comportamiento general de la masa de datos ,respecto del criterio (modelo, metodo o situacion) que se desea analizar en los mismos.

Con el termino anglosajon de INLIERS se denominaran a las observaciones no caracterizadas como outliers.

De lo anteriormente expuesto se deduce que es necesario estudiar la presencia de outliers dentro de la muestra, ya que estas observaciones pueden distorsionar, de forma considerable , estimaciones , tests u otro tipo de analisis estadistico que se desee realizar sobre los datos ; incluso, existen bastantes situaciones, donde la caracterizacion de estos es mas importante que los errores que ocasionan.

A continuacion se estudian las diferentes formas de abordar la posible presencia de estas observaciones en una masa de datos.

1.3 METODOS DE ANALISIS DE OUTLIERS.

Los diversos metodos para analizar la presencia de outliers en un conjunto de datos son clasificados por Beckman y Cook (1983) en:

- Metodos de Acomodacion.
- Metodos de Identificacion.

La utilizacion de un metodo u otro dependera , del objetivo o de la informacion que se desee obtener de la muestra resultante de la experimentacion.

1.3.1 METODOS DE ACOMODACION.

Dentro de estos metodos se engloban dos formas distintas de abordar el problema:

a) Dotar a la fuente o naturaleza de una estructura o modelo matematico que explique todos los valores muestrales.

   Este tipo de procedimiento puede dar lugar a la determinacion de una estructura o modelo matematico erroneo , en caso de que las observaciones outliers fuesen producidas por efecto de la variabilidad inherente.

b) Utilizar procedimientos estadisticos que no sean influenciables por la presencia de outliers (Metodos Robustos).

   Desde esta perspectiva se presupone que todos los outliers son observaciones erroneas, sin admitir la posibilidad de que existan observaciones con cierto grado de variabilidad inherente,las cuales desde un punto de vista teorico han de ser consideradas.

Asi pues, puede considerarse que la mayor poblematica que plantean estos metodos esta contenida en la conjetura de Anscombe (1960), citada previamente.

Ademas estos procedimientos presentan el inconveniente de que requieren de una gran informacion para poder

cumplir con sus objetivos , informacion generalmente no disponible.

1.3.2 METODOS DE IDENTIFICACION.

Dento de estos metodos , se engloban todos aquellos procedimientos que se caracterizan por las dos etapas siguientes:

I. Determinacion de la observacion u observaciones outliers.

II. Analisis pormenorizado de cada outlier , y en conjunto de todos ellos , con el objetivo de determinar las variabilidades o errores presentes , lo que permitiria llegar a una o varias de las siguientes conclusiones:

   a) Mantener o rechazar las observaciones detectadas como miembros del conjunto inicial de datos.

   b) Confirmar el modelo matematico caso que este hubiese sido supuesto inicialmente o determinar las desviaciones que se presenten sobre las hipotesis realizadas sobre la poblacion.

   c) Identificar rasgos de interes practico que en un principio son desconocidos por el investigador.

d) Determinar comportamientos singulares en la experimentacion.

Las dos ultimas situaciones originan en muchos casos que los outliers tengan mas importancia que la muestra en si.

No obstante , se ha de reconocer ,que a veces se tienen experimentos en los que no se llega a conclusiones como las expuestas , bien por la conjetura de Anscombe (1960) , o bien porque no es posible el analisis de los outliers , lo cual puede ocurrir por muy diversas razones , por ejemplo , en aquellos experimentos donde los elementos del espacio empirico no vuelven a su estado incial tras la experimentacion . En estos casos , se tiende a rechazar las observaciones determinadas en la Etapa I , fundamentandose para ello en las afirmaciones realizadas al respecto por diversos autores, entre ellos Hampel et al (1986):

> "... el peligro causado por el outlier es mucho mayor que el peligro de perdida de eficiencia que motiva su rechazo en caso de que sea inlier...".

Esta argumentacion tambien seria valida en la situacion descrita en segundo lugar dentro de los Metodos de Acomodacion.

El objetivo fundamental del estudio es dar un procedimiento de tipo general para la construccion de tecnicas de identificacion de outliers , por lo que a continuacion se estudian las diferentes tecnicas que se encuadran dentro de los Metodos de Identificacion de outliers , y los principales inconvenientes que presentan estos metodos actualmente.

1.4 METODOS DE IDENTIFICACION.

En estas tecnicas se han distinguido dos etapas diferentes , sin embargo , se analizara unicamente la primera de ellas , aquella en la que se determinan las observaciones outliers , ya que la segunda es mas apropiada para el experimentador que para el estadistico.

Los procedimientos existentes para determinar observaciones outliers se pueden clasificar en dos grandes grupos:

- Metodos o Procedimientos Formales
- Metodos o Procedimientos no Formales o Graficos

1.4.1 METODOS O PROCEDIMIENTOS FORMALES.

Estos metodos se caracterizan fundamentalmente , por la realizacion de un contraste de hipotesis , donde la hipotesis nula o hipotesis de trabajo supone que la muestra obtenida de la experimentacion no contiene

outlier alguno . Para aceptar o rechazar esta hipotesis se construye una funcion test , que permite , en base al nivel del mismo , decidir sobre la presencia o no de outliers en los datos.

No obstante , para una misma hipotesis de trabajo , suelen presentarse situaciones en las que es necesario elegir entre diversas funciones test , por lo que , al igual que en la teoria clasica de contrastes de hipotesis , es necesario introducir algunas medidas que permitan elegir entre dichas funciones . Asi , se define la potencia del test, de forma similar a como se hace en la teoria clasica de Neymann - Pearson , y por tanto a cada hipotesis de trabajo se le ha de asociar una hipotesis alternativa que fija el modelo que provoca la presencia de outliers en la muestra.

Barnett y Lewis (1984) , estudian diferentes tipos de hipotesis alternativas en caso de que la hipotesis nula fije la pertenencia de las observaciones a un modelo distribucional.

En estos contrastes de hipotesis ,ademas de la potencia , aparecen otras medidas para la comparacion de las funciones test . Asi David y Paulson (1965) definen diversos criterios para el caso de una observacion outlier , los cuales han sido analizados por diversos autores , entre ellos Hawkins (1978,1980) , quien indica la importancia y utilidad de algunos de ellos .

Beckman y Cook (1983) , sugieren los siguientes parametros , para la comparacion de tests para la identificacion de outliers , que son validos para el caso en que exista mas de una de estas observaciones:

Sea Z el numero de outliers correctamente identificados , sea Y el numero de no outliers declarados como tales (numero de falsos positivos) , y sea k el numero de outliers existentes en la muestra.

i.) La probabilidad de que no existan falsos positivos y todos los outliers sean detectados :

$$P(Z = k \; ; \; Y = 0)$$

ii.) La probabilidad de ningun falso positivo y como minimo un outlier:

$$P(Z > 0 \; ; \; Y = 0)$$

iii.) La proporcion esperada de outliers detectados:

$$E(Z/n)$$

donde n es el numero de elementos que componen la muestra.

iv.) La probabilidad de al menos un falso positivo :

$$P(Y > 0)$$

v.) El numero esperado de falsos positivos:

$$E(Y)$$

vi.) La proporcion entre el valor esperado de falsos positivos y el valor esperado de todas las observa-

ciones declaradas outliers:

$$E(Y)/(E(Z)+E(Y))$$

Pero al igual que ocurre en la teoria clasica de los tests , no siempre es posible encontrar funciones con propiedades optimas globales sobre los parametros definidos , por esta razon se suelen buscar tests con condiciones de optimalidad local o insesgadez o que satisfagan ciertas condiciones de invarianza o bien se utilizan procedimientos que tengan buenas propiedades , como ocurre con el principio de razon de verosimilitud.

Los Metodos Formales asi expuestos , presentan el inconveniente de que el rechazo de la hipotesis nula no determina que observacion u observaciones son outliers , ya que la hipotesis alternativa unicamente formula el modelo que seguirian estas observaciones , caso de existir . Esto hace que se introduzcan las denominadas Hipotesis o Modelos Etiquetados y las Hipotesis o Modelos no Etiquetados.

HIPOTESIS O MODELOS ETIQUETADOS.

Son aquellos en que la hipotesis alternativa especifica las observaciones que se consideraran outliers en caso de rechazar la hipotesis nula.

HIPOTESIS O MODELOS NO ETIQUETADOS.

Son los que no expresan de forma explicita las observaciones que se admitiran como outliers en caso de rechazar la hipotesis de trabajo . En estos se suele indicar el numero de outliers a detectar , o una cota superior del mismo.Para la determinacion de esta cota se utilizan ,entre otros, los denominados procedimientos de saltos.

1.4.2 METODOS O PROCEDIMIENTOS NO FORMALES O GRAFICOS .

Estos metodos tienen como objetivo la representacion de los datos en un sistema de ejes cartesianos bi - o - tridimensional , utilizando procedimientos muy diversos. El mas usual consiste en transformar cada observacion mediante una funcion con valores en la recta real.

La funcion o estadistico se elige de acuerdo con el rasgo o propiedad que se desee analizar en los datos experimentales , considerandose outliers aquellas observaciones cuya representacion grafica se desvia relevantemente de la masa de datos.

1.5 PROBLEMATICA DE LOS METODOS DE IDENTIFICACION.

En este apartado se analizan los inconvenientes , teoricos y practicos , que presentan los dos metodos englobados en este tipo de tecnicas.

1.5.1 -DE LOS METODOS FORMALES.

En estos se presentan muy diversos problemas , que se analizaran atendiendo a los elementos caracteristicos de los mismos.

- RESPECTO DE LA HIPOTESIS ALTERNATIVA . Como ya se ha indicado , esta hipotesis fija el modelo que explica los outliers en caso de que estos existan.

  Esto resulta muy dificil por las razones que se exponen a continuacion:

  i.) Porque no todos los outliers han de proceder de un mismo modelo , ya que esto seria equivalente a afirmar que todos ellos presentan el mismo tipo de error.

  Esta problematica ha sido abordada por diversos autores , para el caso particular de modelos poblacionales univariantes , entre ellos Kale (1976), Rosado (1984),los cuales contrastan la hipotesis nula frente a una hipotesis alternativa en la que recogen distintos modelos , todos ellos con la misma forma funcional , aunque de distintos parametros.

Rosado (1984) , reconoce que resulta dificultoso realizar generalizaciones en todos los sentidos posibles.

En el caso de modelos poblacionales multivariantes , o de situaciones mas estructuradas,la problematica que se plantea es aun mayor ya que en el primero de estos casos se han estudiado pocos modelos alternativos y en situaciones mas estructuradas , a la diversidad de modelos existentes , se les une el conjunto de condiciones iniciales que lleva consigo cada uno de ellos.

ii.) Aun en el caso en que todos los outliers pudiesen explicarse por un unico modelo , cabria preguntarse por la unicidad de este.

iii.) La incertidumbre existente sobre la eleccion de la hipotesis alternativa reduce la utilidad de cualquier propiedad de optimalidad para los tests , ya que estas dependen de la forma que se adopte para dicha hipotesis.

Estos problemas hacen que la mayor parte de las tecnicas formales esten referidas unicamente a la hipotesis nula.

- RESPECTO A LA HIPOTESIS ALTERNATIVA ETIQUETADA. Los modelos etiquetados obligan a determinar , a priori , las posibles observaciones outliers. Esto puede considerarse como una reminiscencia del caracter subjetivo con que se trataba esta teoria en sus inicios.

  En este sentido se puede citar a Rosado (1984) , el cual indica que una de las principales aportaciones del metodo GAN (Generativo con Alternativa Natural) que el estudia radica en el hecho de que las observaciones consideradas outliers se obtienen a posteriori , con lo que se consigue retirar toda subjetividad en la seleccion de observaciones a testar como posibles outliers.

- RESPECTO A LA HIPOTESIS NULA. El problema fundamental es , al igual que en en el caso de la hipotesis alternativa, fijarla ya que generalmente no se tiene un conocimiento suficientemente preciso sobre el modelo o estructura matematica que siguen los datos.

  En este sentido se puede citar la afirmacion de Barnett (1984):

  "... un outlier para una distribucion normal puede no serlo para una distribucion Cauchy..."

Por tanto es necesario tener gran certeza sobre el modelo que se establece en esta hipotesis ya que se podria producir el denominado error de informacion.

- RESPECTO AL ESTADISTICO O TEST. La problematica que se plantea en este punto esta muy relacionada, e incluso podria incluirse en el apartado anterior , ya que la distribucion del estadistico a que da lugar la funcion test , se calcula basandose en el modelo que se propone en la hipotesis nula.

  El mayor inconveniente radica en el hecho de que generalmente no es posible calcular la distribucion exacta del estadistico , por lo que debe aproximarse , utilizandose para ello diversos metodos , siendo los mas comunes:

  a) Metodo Conservativo: La aproximacion de la distribucion del estadistico da lugar a una region critica conservativa , lo que presenta el inconveniente de que observaciones outliers pueden ser declaradas como inliers.

  b) Metodo Asintotico: En este caso se considera la distribucion asintotica del estadistico.

Mas en esta situacion los outliers pierden parte de su importancia ya que estas observaciones cuando mas interesa ser estudiadas es en el caso de muestras de tamano reducido , que es donde ejercen su mayor efecto perturbador.

1.5.2 - DE LOS METODOS NO FORMALES O GRAFICOS .

Estas tecnicas presentan el inconveniente de que no existe una estructura teorica comun a todas ellas , ya que lo unico que se exige a las funciones que se utilizan en estos procedimientos es que tengan un fundamento relacionado con el problema que se desea analizar.

El problema fundamental , en casi todas las tecnicas que se engloban dentro de este apartado , es que no indican ni direccion ni magnitud de la desviacion del outlier respecto de la masa de datos.

1.5.3 EFECTO ENMASCARAMIENTO.

En distintos epigrafes se ha indicado que pueden presentarse situaciones en las que se desee analizar la presencia de mas de un outlier en la masa de datos , lo que da lugar a lo que algunos autores denominan problema de multiples outliers.

Este problema es abordado de formas muy diversas desde el punto de vista de las tecnicas formales y presenta inconvenientes adicionales a los ya destacados para las mismas.

La primera forma de abordar el problema de multiples outliers consiste en la aplicacion sucesiva de procedimientos de identificacion de un outlier , eliminando de la masa de datos la observacion declarada como tal y aplicando la misma tecnica de forma sucesiva a la muestra restringida hasta aceptar la hipotesis de trabajo.

Este procedimiento plantea como inconveniente lo que se denomina efecto de enmascaramiento de tipo A: considerar inliers a ciertas observaciones outliers.

El efecto de enmascaramiento de tipo A fue citado por primera vez por Pearson y Chandrasekar (1936) , quienes afirman sobre su criterio:

> "... este criterio parece eficiente unicamente en caso de que exista una unica observacion outlier...".

Este efecto causa la disminucion de la potencia del metodo considerado, y presenta, ademas, un inconveniente de tipo metodologico , ya que el nivel de significacion referido a la hipotesis de trabajo (todas las observaciones son inliers), no sera el mismo que se le asigne al conjunto de observaciones declaradas como outliers.

Estos inconvenientes , hacen que algunos autores propongan otros tipos de procedimientos para abordar el problema de la presencia de multiples outliers.

Asi , se sugiere un procedimiento iterativo , que en una etapa preliminar selecciona las posibles observaciones outliers , de donde se extrae la observacion que menos sospechas despierta . A esta observacion se le aplica una tecnica de identificacion de un outlier en la muestra restringida (omitiendo los posibles outliers). En caso de aceptar la hipotesis nula , se incluiria esta observacion en la muestra restringida y se procederia a la seleccion de una nueva observacion de entre los restantes posibles outliers , aplicandose de nuevo el procedimiento anterior. En el momento en que se rechace la hipotesis nula, se concluye que la observacion que estaba siendo considerada , junto a las que aun no lo habian sido , son outliers.

Este procedimiento evita el problema del enmascaramiento de tipo A , pero presenta el inconveniente de que es necesario, en cada etapa, seleccionar la observacion a estudiar , lo cual quizas sea factible en el caso de datos univariantes pero no asi en el caso multivariante o en las denominadas situaciones mas estructuradas.

Otros tipos de procedimientos que evitan el efecto de enmascaramiento de tipo A , son las denominadas tecnicas de bloques , las que se aplican para la determinacion de

un numero fijo de outliers.

Desde un punto de vista metodologico , estos ultimos procedimientos , necesitan fijar inicialmente el numero de observaciones a testar como outliers , lo cual puede realizarse mediante los denominados procedimientos de saltos.

Fieller (1976) , indica que estos procedimientos para la identificacion de multiples outliers pueden presentar el denominado efecto de enmascaramiento de tipo B: considerar outliers observaciones que no lo son.

Ademas, tanto la determinacion del numero de outliers a testar , como el efecto de enmascaramiento de tipo B , afectan a las probabilidades de error de tipo I y tipo II.

CAPITULO II

APROXIMACION FORMAL AL ANALISIS CUALITATIVO

DE DATOS

--- Introduccion .

2.1 Propiedades de la relacion criterio .

2.2 Funciones de C-similitud . Seleccion .

2.3 Invarianza.

INTRODUCCION

En el capitulo anterior, se pone de manifiesto, que una de las etapas fundamentales de cualquier analisis estadistico, es el estudio de la calidad de los datos experimentales, con el fin de garantizar la maxima fiabilidad en las conclusiones que se obtengan del mismo.

Ademas, se ha reseñado, que el comportamiento anomalo de una observacion, depende en gran medida de su relacion con las restantes observaciones obtenidas bajo condiciones similares, y del analisis que se pretende llevar a cabo sobre las mismas.

Para responder a estas cuestiones, en este capitulo, se plantea el analisis del comportamiento de los elementos de un conjunto -expresado mediante relaciones de

preferencia entre los mismos- respecto de un criterio establecido, que depende del estudio que se desee realizar.

Por ello, este capitulo se centra en el estudio de la existencia de una funcion, bajo diversas restricciones, basadas en las relaciones definidas entre los elementos del conjunto inicial, proponiendose un procedimiento de seleccion en aquellos casos en que no quede garantizada la unicidad de la misma.

Por ultimo, se aborda el problema de la invarianza, con el objeto de que los resultados obtenidos no dependan del sistema de coordenadas elegido para representar las observaciones experimentales.

2.1 PROPIEDADES DE LA RELACION CRITERIO

Dado $X \subseteq R^k$, sea C un criterio que permite comparar los elementos de X , mediante una relacion binaria C_X , para la que dados $x, x' \in X$, $x\ C_X\ x'$ se interpreta como

x es "al menos tan preferido como" x' , respecto del criterio C , considerados como elementos de X.

A partir de las relaciones binarias C_X ,se consideran las siguientes definiciones:

Definicion 2.1.1.

Dado $X \subseteq R^k$, $x,x' \in X$, se dira que :

x es " preferido " a x' respecto del criterio C , considerados como elementos de X , $x +C_X x'$, si :

$$x +C_X x' \iff x\ C_X\ x' \text{ y } x'\ \not C_X\ x \qquad (2.1)$$

Definicion 2.1.2.

Dado $X \subseteq R^k$, $x,x' \in X$, se dira que :

x y x' son "equivalentes" respecto del criterio C, considerados como elementos de X , $x \sim C_X x'$, si :

$$x \sim C_X x' \iff x\ C_X\ x' \text{ y } x'\ C_X\ x \qquad (2.2)$$

Definicion 2.1.3.

Un criterio C se dira que es evaluable sobre $X \subseteq R^k$, si C_X es fuertemente completa :

$$\forall\ x,x' \in X, \Longrightarrow \quad x\ C_X\ x' \text{ o } x'\ C_X\ x$$

Las relaciones C_X permiten introducir los conjuntos que se recogen en la siguiente

Definicion 2.1.4.

Sea C un criterio evaluable sobre $X \subseteq R^k$.

Dado $x \in X$, se define:

i) El conjunto C_X-dominante de x , $DC_X(x)$, como :

$$DC_X(x)=\{x' \in X \ /\ x'\ C_X\ x\}$$

ii) El conjunto C_X-dominado de x , $dC_X(x)$, como :

$$dC_X(x)=\{x' \in X \ /\ x\ C_X\ x'\}$$

Analogamente, se definen los conjuntos C_X-dominante

y C_X-dominado estrictos, que se representaran mediante $D(+C_X(x))$ y $d(+C_X(x))$, respectivamente.

Como generalizacion se introduce:

Definicion 2.1.5.

Sea C un criterio evaluable sobre $X \subseteq R^k$.

Dado $X' \subseteq X$, se define :

i) El conjunto C_X-dominante de X', $DC_X(X')$, como :

$$DC_X(X') = \{x \in X \ / \ x \ C_X \ x' \ , \ \forall \ x' \in X'\}$$

ii) El conjunto C_X-dominado de X', $dC_X(X')$, como :

$$dC_X(X') = \{x \in X \ / \ x' \ C_X \ x \ , \ \forall \ x' \in X'\}$$

Proposicion 2.1.1.

Dado un criterio C, evaluable, sobre X R^k y $X' \subseteq X$, se verifica :

a) $$DC_X(X') = \bigcap_{x' \in X'} DC_X(x')$$

b) $$dC_X(X') = \bigcap_{x' \in X'} dC_X(x')$$

Estas definiciones permiten introducir:

Definicion 2.1.6.

Dado un criterio evaluable ,C, sobre X R^k y X' X , se dira que X' esta C_X-dominado superiormente (inferiormente) sii $DC_X(X') \neq \emptyset$ $(dC_X(X') \neq \emptyset)$.

Si $DC_X(X') \cap X' \neq \emptyset$ ($dC_X(X') \cap X' \neq \emptyset$), se considera que X' esta propiamente C_X-dominado superiormente (inferiormente).

Se dira que X' esta C_X-dominado (propiamente) si lo esta superior (propiamente) e inferiormente (propiamente).

Definicion 2.1.7.

Sea C un criterio evaluable sobre $X \subseteq R^k$, y $X' \subseteq X$. Entonces :

i) Si X' esta C_X-dominado superiormente , el elemento $x^* \in DC_X(X')$ es C_X-dominante minimal (C_X D m) de X' sii $\forall\, x_D \in DC_X(X') \Longrightarrow x_D\ C_X\ x^*$

ii) Si X' esta C_X-dominado inferiormente , el elemento $x_* \in dC_X(X')$ es C_X-dominado maximal (C_X d M) de X' sii $\forall\, x_d \in dC_X(X') \Longrightarrow x_*\ C_X\ x_d$

Proposicion 2.1.2.

Sea C un criterio evaluable sobre $X \subseteq R^k$ y sea $X' \subseteq X$. Entonces :

i) si x_1^*, x_2^* son elementos (C_X D m) de X' $\Longrightarrow x_1^* \sim_{C_X} x_2^*$

ii) si x_*^1, x_*^2 son elementos (C_X d m) de X' $\Longrightarrow x_*^1 \sim_{C_X} x_*^2$

El resultado se deduce de forma inmediata a partir de la definicion 2.1.7.

En muchas situaciones, las relaciones binarias , C_X , consideradas sobre los conjuntos X de R^k son transitivas, por lo que se da la siguiente

Definicion 2.1.8.

Un criterio C se dira T-evaluable sobre $X \subseteq R^k$, si la relacion binaria C_X define un orden debil en X .

Proposicion 2.1.3.

Si C es uncriterio T-evaluable sobre X , la relacion dada mediante (2.2) define una relacion de equivalencia en dicho conjunto.

Asi , si se representa por $\hat{x} = \{x' \in X / x' \sim_{C_X} x\}$, y por $\hat{X} = \{ \hat{x} / x \in X \}$, la relacion definida en este conjunto mediante :

$\hat{x}\ C_X\ \hat{x}' \iff x\ C_X\ x'$ es un orden simple en $\hat{X}$.

Proposicion 2.1.4.

Sea C un criterio T-evaluable sobre $X \subseteq R^k$ y $\hat{X}$ su conjunto cociente , entonces si existe una clase (C_X D m) o (C_X d M) de $\hat{X}' \subseteq \hat{X}$, es unica.

Este resultado se deduce a partir de la proposicion 2.1.2.

Proposicion 2.1.5.

Si C es un criterio T-evaluable sobre X, de modo que $\#(\hat{X})$ es finito, X esta propiamente C_X-dominado.

2.2 FUNCIONES DE C-SIMILITUD . SELECCION

Las relaciones binarias C_X definidas sobre los subconjuntos de R^k, unicamente determinan el sentido de preferencia existente entre los elementos de dichos conjuntos respecto del criterio en cuestion, sin embargo , no permiten evaluar si los elementos presentan un comporta-

miento similar frente al criterio bajo estudio. Por esta razon,se define a continuacion una funcion (real), sobre el producto cartesiano X x X.

Definicion 2.2.1.

Sea C un criterio T-evaluable sobre $X \subseteq R^k$. Se dira que una funcion

$$SC_X : X \times X \text{ -----} > R$$

es una C-similitud sobre X si verifica:

S1. $SC_X(x,x') \geq 0 \quad \forall\, x,x' \in X$

S2. $SC_X(x,x') = 0$ sii $x \sim_{C_X} x'$

S3. $SC_X(x,x') = SC_X(x',x) \quad \forall\, x,x' \in X$

S4. $\forall\, x,x',x'' \in X$ / $x\ C_X\ x'\ C_X\ x''$ se verifica:

S41. $SC_X(x,x'') \leq SC_X(x,x') + SC_X(x',x'')$

S42. $SC_X(x,x'') \geq \max\{ SC_X(x,x') , SC_X(x',x'')\}$

Esta funcion refleja la similaridad de cada par de elementos de X respecto de C y cumple las siguientes propiedades:

Propiedad 1.

SC_X verifica la desigualdad triangular.

Demostracion:

Dados $x,x',x'' \in X$, se presenta una de las siguientes situaciones:

$x\ C_X\ x'\ C_X\ x''$, para la que se cumple la afirmacion debido a (S41)

$x\ C_X\ x''\ C_X\ x' \Rightarrow (S42) \Rightarrow SC_X(x,x'') \leq SC_X(x,x')$

$x' \; C_X \; x \; C_X \; x'' \;\; ==>(S42)==> \;\; SC_X(x,x'') \leq SC_X(x',x'')$

De donde se deduce :

$$SC_X(x,x'') \leq SC_X(x,x') + SC_X(x',x'')$$

para cualesquiera $x,x',x'' \in X$

Propiedad 2.

Dados $x,x' \in X$:

$$x \sim C_X \; x' \;\; <==> \;\; SC_X(x,y) = SC_X(x',y) \quad \forall \; y \in X \; .$$

Demostracion:

Si $x \sim C_X \; x'$, se tiene

$$SC_X(x,y) \leq SC_X(x,x') + SC_X(x',y) \leq SC_X(x',x) + SC_X(x,y) \;\; ==>$$

$$==> \;\; SC_X(x,y) = SC_X(x',y)$$

El reciproco es evidente.

Propiedad 3.

La funcion $SC_{\hat{X}} : \hat{X} \times \hat{X} \text{ ------> } R$

definida como

$$SC_{\hat{X}}(\hat{x},\hat{y}) = SC_X(x,y)$$

esta bien definida y $(\hat{X} \; , \; SC_{\hat{X}})$ es un espacio metrico.

Propiedad 4.

Si $X' \subseteq X$ es un conjunto propiamente C_X-dominado, $\forall \; x,x' \in X'$, se verifica:

$$SC_X(x,x') \leq SC_X(x_d,x_D) \quad \forall \; (x_d,x_D) \in dC_X(X') \times DC_X(X')$$

y ademas

$$SC_X(x_*,x^*) = \sup_{x,x' \in X'} \{ SC_X(x,x') \}$$

donde x^* , x_* , son los elementos (C_X D m) y (C_X d M) de X', respectivamente.

En ocasiones puede ser de especial interes comparar la similitud de los elementos de X respecto de un valor fijado $x_o \in X$, por lo que se introduce :

Definicion 2.2.2.

Sea C un criterio T-evaluable sobre $X \subseteq R^k$ y SC_X una funcion de C-similitud sobre X . Dado $x_o \in X$, se define :

$$SC_{X,x_o} : X \text{ -----} > R$$

$$SC_{X,x_o}(x) = SC_X(x,x_o)$$

Esta permite enunciar

Propiedad 5.

Si $X' \subseteq X$ es un conjunto C_X-dominado , se verifica:

$$SC_{X,x_D}(x) < SC_{X,x_D}(x') \implies x_D \; C_X \; x + C_X \; x' \; C_X \; x_d \implies$$

$$\implies SC_{X,x_d}(x) \geq SC_{X,x_d}(x')$$

$$\forall (x_d , x_D) \in dC_X(X') \times DC_X(X') \quad y \quad \forall x,x' \in X$$

Existen situaciones en las que la funcion SC_X no recoge la informacion que proporciona la relacion binaria ya que la C-similitud de dos elementos cualesquiera no refleja de que modo estan relacionados los mismos respecto del criterio C , salvo que dichos elementos sean C_X -equivalentes.

Por esta razon se introduce la siguiente

Definicion 2.2.3.

Sea C un criterio T-evaluable sobre $X \subseteq R^k$. Se dira que una funcion

$$SSC_X : X \times X \text{ -----> } R$$

es una C-similitud signada sobre X si verifica :

SS1. $SSC_X(x,x') \geq 0$ sii $x\ C_X\ x'$

SS2. $SSC_X(x,x') = - SSC_X(x',x) \quad \forall\ x,x' \in X$

SS3. $\forall\ x,x',x'' \in X$ / $x\ C_X\ x'\ C_X\ x''$

SS31. $SSC_X(x,x'') \leq SSC_X(x,x') + SSC_X(x',x'')$

SS32. $SSC_X(x,x'') \geq \max \{SSC_X(x,x'), SSC_X(x',x'')\}$

La funcion SSC_X verifica las siguientes propiedades:

Propiedad 6.

Si SSC_X es una funcion de C-similitud signada sobre X , entonces dados $x,x' \in X$:

$$SSC_X(x,x') = 0 \text{ sii } x\ {\sim}C_X\ x'.$$

Demostracion:

$$SSC_X(x,x')=0 \iff SSC_X(x,x')=0 \text{ y } SSC_X(x',x) = 0$$

$$\iff x\ C_X\ x' \quad y \quad x'\ C_X\ x \iff \quad x\ {\sim}C_X\ x' .$$

Propiedad 7.

Dados $x,x' \in X$. Entonces :

$$x\ {\sim}C_X\ x' \iff SSC_X(x,y) = SSC_X(x',y) \quad \forall\ y \in X.$$

Demostracion:

Si $x \sim C_X x'$, los dos casos que se pueden presentar: $x' \sim C_X x \; C_X \; y$; $y \; C_X \; x \sim C_X x'$

$x' \sim C_X x \; C_X \; y \implies SSC_X(x',y) \leq SSC_X(x',x)+SSC_X(x,y) \leq$

$\leq \; SSC_X(x,x')+SSC_X(x',y)$ por tanto

$$SSC_X(x,y) = SSC_X(x',y)$$

Analogamente se analizaria el otro caso.

El reciproco es inmediato a partir de la propiedad 6.

Propiedad 8.

La funcion $SSC_{\hat{X}} : \hat{X} \times \hat{X} \dashrightarrow R$

$$SSC_{\hat{X}}(\hat{x},\hat{y}) = SSC_X(x,y)$$

esta bien definida.

Propiedad 9.

Si $X' \subseteq X$, esta propiamente C_X-dominado, se verifica

$$SSC_X(x_d,x_D) \leq SSC_X(x,x') \leq SSC_X(x_D,x_d)$$

$\forall \; x,x' \in X \;$ y $\; \forall \; (x_d,x_D) \in dC_X(X') \times DC_X(X')$.

Ademas

$$SSC_X(x^*,x_*) = \sup_{x,x' \in X'} \{SSC_X(x,x')\} = - SSC_X(x_*,x^*) = \ldots$$

$$\ldots = \inf_{x,x' \in X'} \{SSC_X(x,x')\}$$

siendo x^*, x_* , elementos $(C_X \; D \; m)$ y $(C_X \; d \; M)$, de X' respectivamente.

Las funciones introducidas en las definiciones 2.2.1 y 2.2.3 permiten definir una relacion binaria sobre X x X , tal como se recoge a continuacion.

Definicion 2.2.4.

Dados $x_i \in X$, i=1,2,3,4 , se define la relacion binaria C'_X sobre X x X como :

$$(x_1,x_2)\ C'_X\ (x_3,x_4) \quad \text{sii} \quad SC_X(x_1,x_2) \leq SC_X(x_3,x_4) \qquad (2.3)$$

Esta definicion refleja que el par (x_1,x_2) es " al menos tan similar" como el par (x_3,x_4) respecto del criterio C , considerados como elementos de X .

Esta relacion permite introducir las siguientes:

Definicion 2.2.5.

Dados $x_i \in X$, i=1,2,3,4, se dira que el par (x_1,x_2) es "mas similar" que (x_3,x_4) respecto del criterio C , considerados como elementos de X

$$(x_1,x_2)\ +C'_X\ (x_3,x_4), \text{ sii:} \qquad (2.4)$$

$$(x_1,x_2)\ C'_X\ (x_3,x_4) \quad y \quad (x_3,x_4)\ C'_X\ (x_1,x_2) \quad <==>$$

$$<==> SC_X(x_1,x_2) < SC_X(x_3,x_4)$$

Definicion 2.2.6.

Dados $x_i \in X$, i=1,2,3,4, se dira que el par (x_1,x_2) es "equivalente" al par (x_3,x_4) respecto del criterio C , considerados como elementos de X

$$(x_1,x_2)\ \sim C'_X\ (x_3,x_4)\ , \text{ sii:} \qquad (2.5)$$

$$(x_1,x_2)\ C'_X\ (x_3,x_4) \quad y \quad (x_3,x_4)\ C'_X\ (x_1,x_2) \quad <==>$$

$$\langle==\rangle \; SC_X(x_1,x_2) = SC_X(x_3,x_4)$$

Proposicion 2.2.1.

i) C'_X define un un orden debil sobre X x X.

ii) $+C'_X$ define un orden debil estricto sobre X x X

iii) $\sim C'_X$ define una relacion de equivalencia sobre X x X .

Las relaciones dadas mediante (2.3),(2.4) y (2.5) se pueden definir de la misma forma a traves de una funcion de similitud signada , para la que se puede enunciar una proposicion analoga a la proposicion 2.2.1. Ademas , en esta situacion se verifica

Proposicion 2.2.2.

Las relaciones C'_X ,$+C'_X$,$\sim C'_X$, definidas a traves de SSC_X mediante (2.3) , (2.4) y (2.5) verifican el axioma de signo reverso, es decir:

$$(x_1,x_2)\ C'_X\ (x_3,x_4) \langle==\rangle (x_4,x_3)\ C'_X\ (x_2,x_1)$$

$$(x_1,x_2)\ +C'_X\ (x_3,x_4) \langle==\rangle (x_4,x_3)\ +C'_X\ (x_2,x_1)$$

$$(x_1,x_2)\ \sim C'_X\ (x_3,x_4) \langle==\rangle (x_4,x_3)\ \sim C'_X\ (x_2,x_1)$$

En este caso se ha de destacar que la relacion C'_X lleva implicita la relacion C_X , ya que :

$$(x,x)\ C'_X\ (x,y) \langle==\rangle SSC_X(x,x) \leq SSC_X(x,y) \langle==\rangle$$

$$\langle==\rangle SSC_X(x,y) \geq 0 \langle==\rangle x\ C_X\ y.$$

Para la posterior aplicacion practica de esta teoria es deseable disponer de una funcion sobre X en lugar de sobre X x X que refleje el sentido de preferencia establecido en las relaciones C_X y C'_X .

Para ello se estudiara bajo que condiciones existe una funcion

$$IC_X : X \dashrightarrow R$$

que verifique

$$SSC_X(x,y) \leq SSC_X(z,w) \iff IC_X(x)-IC_X(y) \leq IC_X(z)-IC_X(w)$$

ya que entonces se tendria que

$$y\ C_X\ x \iff SSC_X(y,x) \geq SSC_X(x,x) \iff IC_X(y) \geq IC_X(x)$$

y por tanto se recoge la informacion que proporcionan las relaciones .

Las funciones de similitud y similitud signadas introducidas en las definiciones 2.2.1 y 2.2.3 , incluyen como casos particulares a las definidas por:

$$SC_X(x,x') = \begin{cases} k & \text{si } x +C_X\ x' \text{ o } x' +C_X\ x \\ 0 & \text{si } x \sim C_X\ x' \end{cases}$$

$$SSC_X(x,x') = \begin{cases} k & \text{si } x +C_X\ x' \\ 0 & \text{si } x \sim C_X\ x' \\ -k & \text{si } x' +C_X\ x \end{cases}$$

siendo k una constante positiva.

Estas funciones no recogerian de forma deseable toda la informacion que proporciona la relacion binaria , por lo que se introduce la siguiente

Definicion 2.2.7.

Una funcion de C-similitud (signada), se dira aditiva, si cumple:

$$\forall\, x,x',x'' \in X \;/\; x\, C_X\, x'\, C_X\, x'' \;\Longrightarrow$$

$$\Longrightarrow SC_X(x,x'') = SC_X(x,x') + SC_X(x',x'')$$

Teorema 2.2.1.

Si SSC_X es una funcion de C-similitud signada aditiva sobre X , y C'_X , $+C'_X$, $^{\sim}C'_X$, son las relaciones dadas por (2.3) , (2.4) , (2.5) , a traves de SSC_X ,y se verifican las condiciones :

T1. $\forall\, x_i, x'_i \in X$, i=1,2,3 tales que

$$\left.\begin{array}{l}(x_1,x_2)\; C'_X\; (x'_1,x'_2)\\ \\ (x_2,x_3)\; C'_X\; (x'_2,x'_3)\end{array}\right\} \Longrightarrow (x_1,x_3)\; C'_X\; (x'_1,x'_3)$$

T2. $\forall\, y, x_i \in X$, i=1,2,3,4 tales que

$$\left.\begin{array}{l}(x_1,x_2)\; C'_X\; (x_3,x_4)\\ \\ (x_3,x_4)\; C'_X\; (y,y)\end{array}\right\} \Longrightarrow$$

$$\Longrightarrow \exists\, z_1, z_2 \in X \text{ de forma que}$$

$$(x_1,z_1)\; {}^{\sim}C'_X\; (x_3,x_4) \qquad y \qquad (z_2,x_2)\; {}^{\sim}C'_X\; (x_3,x_4)$$

Entonces :

$$\exists\ IC_X : X \dashrightarrow R \quad \text{tal que} \quad \forall\ x_i \in X\ ,i=1,2,3,4 \tag{2.6}$$

$$(x_1,x_2)\ +C'_X\ (x_3,x_4) \Longleftrightarrow IC_X(x_1)-IC_X(x_2) < IC_X(x_3)-IC_X(x_4)$$

y ademas se verifica que si

$$IC^{\star}_X : X \dashrightarrow R$$

es otra funcion que cumple (2.6) :

$$IC^{\star}_X = \alpha\ IC_X + \beta \quad / \alpha > 0$$

Demostracion:

De los resultados obtenidos en las Proposiciones 2.2.1 y 2.2.2 , y de T1, T2, se deduce que el par $(X, +C'_X)$, tiene estructura de diferencia algebraica, de lo que se obtiene el resultado propuesto a partir de Krantz et al (1971).

Corolario 2.2.1.

En las condiciones del teorema anterior se verifica:

i) $(x_1,x_2)\ {}^{\sim}C'_X\ (x_3,x_4)$ sii $IC_X(x_1)-IC_X(x_2)=IC_X(x_3)-IC_X(x_4)$

ii) $(x_1,x_2)\ C'_X\ (x_3,x_4)$ sii $IC_X(x_1)-IC_X(x_2) \leq IC_X(x_3)-IC_X(x_4)$

iii) $x\ C_X\ y$ sii $IC_X(x) \geq IC_X(y)$

iv) $x\ +C_X\ y$ sii $IC_X(x) > IC_X(y)$

v) $x\ {}^{\sim}C_X\ y$ sii $IC_X(x) = IC_X(y)$

La importancia de estos resultados radica en el hecho, de que la funcion IC_X es unica salvo transformacion lineal, lo que significa que incorpora dos constantes (un cero arbitrario y una unidad arbitraria).

En el caso particular en que X estuviese C_X-dominado, se podria considerar:

$$IC_X(DC_X(X)) \text{ -----> } 1$$
$$IC_X(dC_X(X)) \text{ -----> } 0$$

con lo que $IC_X(x) \in [0,1]$ $\forall x \in X$, lo que se puede interpretar como un conjunto difuso sobre X , definido a traves del criterio C.

El teorema da lugar a una condicion suficiente para la existencia de una funcion verificando (2.6) , sin embargo , cuando X es un subconjunto de R^k , con cardinal finito , es posible establecer un teorema que proporciona una condicion necesaria y suficiente para la existencia de dicha funcion, para lo cual es necesario dar previamente los siguientes resultados.

Sea $X \subset R^k$/ $X \subseteq \prod_{i=1}^{k} X_i$; $X_i \subset R$ con $\#(X_i) < +\infty$,y sea $X^m = X \times \overset{m}{\ldots} \times X$.

Definicion 2.2.8.

En X^m , $m > 1$, dados x^i, $y^i \in X$ $i=1,..,m$ se define la relacion E_m , por :

$(x^1,\ldots,x^m)$ E_m $(y^1,\ldots,y^m)$ sii $(x_i^1,\ldots,x_i^m)$ es una permutacion de $(y_i^1,\ldots,y_i^m)$ para $i=1,\ldots,k$.

La relacion asi definida es una relacion de equivalencia.

Lema 2.2.1.

Sea C un criterio T-evaluable en $X \subset R^k$, y sea $X \subseteq \prod_{i=1}^{k} X_i$; $\#(X_i) < +\infty$ / $x_{ij} = x_{ij'}$, $\forall\, j \neq j'$; i=1,..,k.

Entonces :

(2.7) $\exists\, u_1,\ldots,u_k$ / $u_i : X_i \dashrightarrow R$, verificando

$$x +C_X y \;\Longrightarrow\; \sum_{i=1}^{k} u_i(x_i) > \sum_{i=1}^{k} u_i(y_i) \qquad \text{sii}$$

(2.8) $\forall\, x^1,\ldots,x^m , y^1,\ldots,y^m \in X$, m = 2,3... se cumple :

$$[(x^1,\ldots,x^m)\ E_m\ (y^1,\ldots,y^m) \;/\; x^j +C_X y^j \quad \text{ó}$$

$$x^j = y^j \ , \ j=1,\ldots,m-1] \Longrightarrow\ x^m +C_X y^m$$

Demostracion:

Suponiendo (2.7) y dado $(x^1,\ldots,x^m)\ E_m\ (y^1,\ldots,y^m)$ /

$$x^j +C_X y^j \quad \text{o} \quad x^j = y^j \ , \ j < m \Longrightarrow$$

$$\sum_{j=1}^{m-1} \sum_{i=1}^{k} u_i\,(x_i^j) \geq \sum_{j=1}^{m-1} \sum_{i=1}^{k} u_i(y_i^j) \quad . \qquad (2.9)$$

Mas como $(x^1,\ldots,x^m)\ E_m(y^1,\ldots,y^m) \Longrightarrow$

$$\Longrightarrow \quad \sum_{j=1}^{m} \sum_{i=1}^{k} u_i\,(x_i^j) = \sum_{j=1}^{m} \sum_{i=1}^{k} u_i(y_i^j) \qquad (2.10)$$

Mediante (2.9) y (2.10) se deduce:

$$\sum_{i=1}^{k} u_i\,(x_i^m) \leq \sum_{i=1}^{k} u_i(y_i^m) \Longrightarrow x^m +C_X\ y^m$$

Reciprocamente:

Dado $X_i = \{ x_{ij} \ , j=1,\ldots,n_i \}$, $i = 1,\ldots,k$ se considera el vector

$$U = \{ u_i(x_{ij}) \quad i=1,..,k \ ; \ j=1,..,n_i \}. \qquad (2.11)$$

Sea $H = \#\{ (x,y) \ / \ x +C_X y \ , \ x,y \in X \}$, por lo que se puede expresar

$$\{ (x,y) \ / \ x +C_X y \} = \{ (x^1,y^1) \ ,\ldots, \ (x^H,y^H) \}$$

Dado (x^h,y^h) , $h = 1,\ldots,H$ se definen los vectores

$a_x^h = \{ (a_x^h)_{ij} \quad i=1,..,k \ ; \ j=1,..,n_i \}$ donde

$$(a_x^h)_{ij} = \begin{cases} 1 & \text{si} \quad x_{ij} = x_i^h \\ 0 & \text{en otro caso} \end{cases}$$

y analogamente

$a_y^h = \{ (a_y^h)_{ij} \quad i=1,..,k \ ; \ j=1,..,n_i \}$ donde

$$(a_y^h)_{ij} = \begin{cases} 1 & \text{si} \quad x_{ij} = y_i^h \\ 0 & \text{en otro caso} \end{cases}$$

(2.12)

Sea $a_{xy}^h = a_x^h - a_y^h = \{(a_{xy}^h)_{ij} \quad i=1,..,k \ ; \ j=1,..,n_i \}$

por lo que

$$(a_{xy}^h)_{ij} = \begin{cases} 1 \text{ si } x_{ij} = x_i^h \ ; \ x_{ij} \neq y_i^h \\ -1 \text{ si } x_{ij} \neq x_i^h \ ; \ x_{ij} = y_i^h \\ 0 \text{ si } \begin{cases} (x_{ij} = x_i^h \ ; \ x_{ij} = y_i^h) \\ \text{ó} \\ (x_{ij} \neq x_i^h \ ; \ x_{ij} \neq y_i^h) \end{cases} \end{cases}$$

A partir de los vectores definidos en (2.11) y (2.12), se considera el sistema

$$U' \; a^h_{xy} \; > 0 \qquad h = 1,..,H \tag{2.13}$$

por lo que si este sistema tiene solucion en U , se verifica que

$$\exists \; u_i : X_i \dashrightarrow R \quad , \; i=1,..,k$$

tales que

$$x \; +C_X \; y \quad \Longrightarrow \quad \sum_{i=1}^{k} u_i(x_i) \; > \; \sum_{i=1}^{k} u_i(y_i)$$

En caso de que no existiese solucion para dicho sistema , el Teorema de la Alternativa (Tucker 1956) ,asegura

$\exists \; r_1 \; , \; \ldots , \; r_H \geq 0$ y enteros , no todos nulos , tales que

$$\sum_{h=1}^{H} r_h \; (a^h_{xy} \;)_{ij} = 0 \qquad i=1,..,k \; ; \; j=1,.., \; n_i \; .$$

Es decir :

$$\sum_{h=1}^{H} r_h \; (a^h_{x} \;)_{ij} \; = \; \sum_{h=1}^{H} r_h \; (a^h_{y} \;)_{ij} \qquad i=1,..,k; \; j=1,..,n_i$$

de donde se deduce que

$$(r_1 \; x^{1,}s \; ,..,r_H \; x^{H,}s) \; E_{r_1+..+r_H} \; (\; r_1 \; y^{1,}s \; ,..,r_H \; y^{H,}s)$$

lo que condradice la hipotesis inicial ya que

$$x^i \; +C_X \; y^i \qquad i \; = \; 1,..,H$$

Por tanto ha de existir solucion para el sistema (2.13)

Lema 2.2.2.

Sea C un criterio T-evaluable en $X \subset R^k$, y sea $X \subseteq \prod_{i=1}^{k} X_i$, con $\#(X_i) < +\infty$.

Entonces:

$$\exists\ u_1,\ldots,u_k \quad / \quad u_i : X_i \dashrightarrow R \ , \qquad (2.14)$$

verificando

$$x \ +C_X \ y \iff \sum_{i=1}^{k} u_i(x_i) > \sum_{i=1}^{k} u_i(y_i) \qquad \text{sii}$$

$$\forall\ x^1,\ldots,x^m \ , \ y^1,\ldots,y^m \in X \ , \ m=2,3... \qquad (2.15)$$

se cumple:

$$[(x^1,..,x^m)\ E_m\ (y^1,\ldots,y^m) \ / \ x^j\ C_X\ y^j,\ j=1,\ldots,m-1] \iff$$

$$\iff x^m \ +C_X \ y^m$$

Demostracion:

Suponiendo (2.14) y dado $(x^1,\ldots,x^m)\ E_m\ (y^1,\ldots,y^m)$ tal que $x^j\ C_X\ y^j$, $j < m$, se tiene :

$$\sum_{j=1}^{m-1} \sum_{i=1}^{k} u_i\ (x_i^j\ ) \geq \sum_{j=1}^{m-1} \sum_{i=1}^{k} u_i(y_i^j\ ) \quad ,$$

y ya que

$$(x^1,\ldots,x^m)\ E_m\ (y^1,\ldots,y^m) \Longrightarrow$$

$$\Longrightarrow \sum_{j=1}^{m} \sum_{i=1}^{k} u_i\ (x_i^j\ ) = \sum_{j=1}^{m} \sum_{i=1}^{k} u_i(y_i^j\ )$$

Por tanto:

$$\sum_{i=1}^{k} u_i\ (x_i^m) \leq \sum_{i=1}^{k} u_i(y_i^m) \iff x^m \ +C_X \ y^m$$

Reciprocamente

Dado $X_i = \{ x_{ij} \ , \ j = 1,\ldots,n_i \}$ $i = 1,\ldots,k$ se considera el vector

$$U = \{ u_i(x_{ij}) \quad i=1,..,k \ ; \ j=1,..,n_i \}.$$

Sea $H = \#\{ (x,y) \ / \ x +C_X y \ ; \ x,y \in X \}$ y considerese el conjunto de pares $\{ (x,y) \ / \ x \sim C_X y \ , \ x \neq y\}$ donde de los pares (x,y) ,(y,x) solo se considera uno de ellos y sea M-H el cardinal de este conjunto.

En estas condiciones se tienen los conjuntos :

$$\{ (x^i,y^i) \quad i=1,\ldots,H \ / \ x^i +C_X y^i \}$$

$$\{ (x^i,y^i) \quad i=H+1,\ldots,M \ / \ x^i \sim C_X y^i \}$$

Para los pares (x^i, y^i) , $h=1,\ldots,M$, se definen los vectores

$$a^h_x \quad , \quad a^h_y \quad , \quad a^h_{xy}$$

de la misma forma que en el lema 2.2.1.

A partir de estos vectores se considera el sistema

$$U' \ a^h_{xy} > 0 \qquad h=1,..,H$$

$$U' \ a^h_{xy} = 0 \qquad h=H+1,..,M$$

Si este sistema tiene solucion , se verifica :

$$\exists \ u_i : X_i \dashrightarrow R \ , \ i=1,..,k$$

tales que

$$x +C_X y \quad \Longrightarrow \quad \sum_{i=1}^{k} u_i(x_i) > \sum_{i=1}^{k} u_i(y_i)$$

$$x \sim C_X\ y \quad \Longrightarrow \quad \sum_{i=1}^{k} u_i(x_i) = \sum_{i=1}^{k} u_i(y_i)$$

Si el sistema no admite solucion, a partir del Teorema de la Alternativa (Tucker 1956), se deduce

$\exists\ r_h$, $h=1,\ldots,M$ / $r_h \geq 0$, $h=1,\ldots,H$, no todos nulos de forma que

$$\sum_{h=1}^{H} r_h \left(a^h_{xy}\right)_{ij} = 0 \qquad i=1,..,k\ ,\ j=1,..,n_i$$

Debido a que los valores $\left(a^h_{xy}\right)_{ij} \in \{ 1,0,-1 \}$, puede considerarse que $r_h \in Z$, ademas , pueden considerarse positivos, ya que si alguno de ellos no lo fuese, bastaria reemplazar el par (x,y) por el (y,x) y r_h por $-r_h$, con lo cual se llegaria a una contradiccion similar a la del lema 2.2.1.

Teorema 2.2.2.

Sea C un criterio T-evaluable en $X \subset R^k$ con $\#(X) < +\infty$, y sea SSC_X una funcion de C-similitud signada sobre X . Entonces

$\exists\ IC_X$: X -----> R verificando (2.16)

$(x,y)\ +C'_X\ (z,w) \Longrightarrow IC_X(x)-IC_X(y) < IC_X(z)-IC_X(w)$

sii

para $m=2,3,\ldots$ dados $x^i,y^i,z^i,w^i \in X$, $i=1,\ldots,m$ cumpliendo

I) $x^1,\ldots,x^m\ ,w^1,\ldots,w^m$ es una permutacion de $y^1,..,y^m,z^1,\ldots,z^m$

II) $(x^i,y^i) +C'_X (z^i,w^i)$ $i=1,\ldots,m-1$

se tiene :

$$(z^m,w^m)\ C'_X\ (x^m,y^m)$$

Demostracion:

Si se verifica (2.16) :

dado que $(x^j,y^j) +C'_X (z^j,w^j)$, $j < m$

se tiene que:

$$IC_X(x^j)-IC_X(y^j) < IC_X(z^j)-IC_X(w^j)\ ,\ j=1,\ldots,m-1$$

y como ademas:

$$\sum_{j=1}^{m} [IC_X(x^j)-IC_X(y^j)] = \sum_{j=1}^{m} [IC_X(z^j)-IC_X(w^j)]$$

se obtiene que :

$$IC_X(x^m) - IC_X(y^m) > IC_X(z^m) - IC_X(w^m)$$

y por tanto

$$(x^m,y^m) +C'_X (z^m,w^m).$$

Reciprocamente :

Dado $X = \{x_1,\ldots,x_N\}$, se considera el vector

$$U = \{ u(x_i)\ ,\ i=1,..,N \}$$

y el conjunto

$$A = \{[(x,y),(z,w)] / (x,y) +C'_X (z,w)\ ,\ x,y,z,w \in X \}$$

Ya que $\#(X) < +\infty \Longrightarrow \#(A) = H < +\infty$, por lo que se puede considerar

$$A = \{ [(x^h ,y^h)\ ,\ (z^h ,w^h)]\ ,\ h=1,\ldots,H \}$$

Dado (x^h, y^h) , (z^h, w^h) , se definen los vectores

$$(a_x^h) = \{ (a_x^h)_i \ , \ i=1,..,N \}$$

$$(a_y^h) = \{ (a_y^h)_i \ , \ i=1,..,N \}$$

$$(a_z^h) = \{ (a_z^h)_i \ , \ i=1,..,N \}$$

$$(a_w^h) = \{ (a_w^h)_i \ , \ i=1,..,N \}$$

donde

$$(a_x^h)_i = \begin{cases} 1 & \text{si} \quad x^h = x_i \\ 0 & \text{si} \quad x^h \neq x_i \end{cases}$$

$$(a_y^h)_i = \begin{cases} 1 & \text{si} \quad y^h = x_i \\ 0 & \text{si} \quad y^h \neq x_i \end{cases}$$

$$(a_z^h)_i = \begin{cases} 1 & \text{si} \quad z^h = x_i \\ 0 & \text{si} \quad z^h \neq x_i \end{cases}$$

$$(a_w^h)_i = \begin{cases} 1 & \text{si} \quad w^h = x_i \\ 0 & \text{si} \quad w^h \neq x_i \end{cases}$$

y

$$a_{xyzw}^h = a_y^h + a_z^h - a_x^h - a_w^h$$

A partir de estos vectores se considera el sistema

$$U' \ a_{xyzw}^h \ > 0 \quad , \ h=1,..,H \qquad (2.17)$$

Si este sistema admite solucion , entonces :

$$\exists \ IC_X : X \text{ -----} > R$$

verificando:

$$(x,y) +C'_X (z,w) \Longrightarrow IC_X(x)-IC_X(y) < IC_X(z)-IC_X(w)$$

para lo cual basta considerar $IC_X(x_i) = u(x_i)$.

Si no existiese solucion para el sistema (2.17) , por Tucker (1956), existirian $r_h \in Z^+$, no todos nulos , tales que :

$$\sum_{h=1}^{H} r_h \, (a^h_{xyzw})_j = 0 \qquad j=1,\ldots,N \qquad -$$

y de

$(x^h,y^h) +C'_X (z^h,w^h)$,se deduce que existe una secuencia:

$$(x^1,y^1) +C'_X (z^1,w^1) \quad , \quad \ldots \quad , \quad (x^m,y^m) +C'_X (z^m,w^m)$$

donde $x^1,\ldots,x^m,w^1,\ldots,w^m$ es una permutacion de $y^1,\ldots,y^m, z^1,\ldots,z^m$.

Con lo cual, si $m > 1$, esta condicion contradice la hipotesis de partida y si $m = 1$, entonces:

$$(x,y) +C'_X (x,y) \quad \text{ó} \quad (x,x) +C'_X (y,y)$$

que tambien contradice la hipotesis inicial .

Por tanto ha de existir solucion para el sistema (2.17).

Teorema 2.2.3.

Sea C un criterio T-evaluable en $X \subset R^k$ con $\#(X) < +\infty$, y sea SSC_X una funcion de C-similitud signada sobre X . Entonces

(2.18)

$$\exists\ IC_X: X \dashrightarrow R \qquad \text{verificando}$$

$$(x,y) +C'_X (z,w) \Longleftrightarrow IC_X(x)-IC_X(y) < IC_X(z)-IC_X(w)$$

sii (2.19)

para $m=2,3,\ldots$, dados $x^i,y^i,z^i,w^i \in X$, $i=1,\ldots,m$

cumpliendo

I') $x^1,\ldots,x^m,w^1,\ldots,w^m$ es una permutacion de $y^1,\ldots,y^m,z^1,\ldots,z^m$

II') $(x^i,y^i)\ C'_X\ (z^i,w^i)\quad i=1,\ldots,m-1$

se tiene :

$$(z^m,w^m)\ C'_X\ (x^m,y^m)$$

Demostracion:

Si se verifica (2.18) :

$$IC_X(x^j)-IC_X(y^j) \leq IC_X(z^j)-IC_X(w^j)\ ,\ j=1,\ldots,m-1$$

y como ademas:

$$\sum_{j=1}^{m} [IC_X(x^j)-IC_X(y^j)] = \sum_{j=1}^{m} [IC_X(z^j)-IC_X(w^j)]$$

se obtiene que :

$$IC_X(x^m) - IC_X(y^m) \geq IC_X(z^m) - IC_X(w^m)$$

y por tanto

$$(x^m,y^m)\ +C'_X\ (z^m,w^m).$$

Reciprocamente:

De la condicion (2.19) , en el caso bidimensional, se deduce:

$$\left.\begin{array}{l} [(x^1,y^1),\ldots,(x^m,y^m)]\ E_m\ [(z^1,w^1),\ldots,(z^m,w^m)] \\ (x^j,y^j)\ C'_X\ (z^j,w^j)\ ,j=1,\ldots,m-1 \end{array}\right\} \Longrightarrow$$

$$\Longrightarrow (x^m, y^m) +C'_X (z^m, w^m)$$ y por lema 2.2.2 :

$\exists\ u_1 , u_2 : X \dashrightarrow R$ de modo que

$$(x,y) +C'_X (z,w) \Longleftrightarrow u_1(x)+u_2(y) < u_1(z)+u_2(w)$$

y de (2.19)

$$(x,y) +C'_X (z,w) \Longleftrightarrow (w,z) +C'_X (y,x)$$

por lo que

$$(x,y) +C'_X (z,w) \Longleftrightarrow u_1(w)+u_2(z) < u_1(y)+u_2(x)$$

Por tanto definiendo

$$IC_X(x) = u_1(x) - u_2(x)$$

se tiene que

$$(x,y) +C'_X (z,w) \Longleftrightarrow IC_X(x)-IC_X(y) < IC_X(z)-IC_X(w) \quad .$$

En el caso en que no exista IC_X verificando las condiciones impuestas en los teoremas 2.2.1, 2.2.2, 2.2.3, se construira una funcion que refleje unicamente el sentido de preferencia establecido en la relacion C_X .

Definicion 2.2.9.

Sea C un criterio sobre $X \subseteq R^k$, que define una relacion binaria C_X . Se dira que una funcion

$$IC_X : X \dashrightarrow R \ , \ \text{es :}$$

i) Estrictamente C_X-creciente si :

$$x +C_X y \Longleftrightarrow IC_X(x) > IC_X(y)$$

ii) Estrictamente C_X-decreciente si :

$$x\ C_X\ y \Longleftrightarrow IC_X(x) < IC_X(y)$$

A continuacion se recogen una serie de resultados que caracterizan distintas situaciones en las que existen funciones estrictamente C_X-crecientes (decrecientes).

Teorema 2.2.4.

Sea C un criterio T-evaluable sobre $X \subset R^k$, con #(X) finito. Entonces:

$$IC_X : X \text{ -----} > R$$

estrictamente C_X-creciente.

Demostracion:

Basta considerar $\#(d+C_X(x)) = IC_X(x)$

Teorema 2.2.5.

Sea C un criterio T-evaluable sobre $X \subset R^k$, con #(X) finito, e IC_X, estrictamente C_X-creciente.

Entonces una funcion

$$f : X \text{ -----} > R$$

es estrictamente C_X-creciente sii:

$\exists \emptyset : IC_X(X) \text{ -----} > R$, estrictamente creciente / $\emptyset \cdot IC_X = f$

Demostracion:

Si $f = \emptyset\ IC_X$, con $\emptyset$ estrictamente creciente, entonces

$x +C_X x' \langle == \rangle IC_X(x) > IC_X(x') \langle == \rangle \emptyset \cdot IC_X(x) > \emptyset \cdot IC_X(x')$

Reciprocamente, dada $f : X \text{ -----} > R$, estrictamente C_X-creciente se define

$$\emptyset : IC_X(X) \text{ -----} > R$$

$$\emptyset[IC_X(x)] = f(x) \ ,$$

la cual es estrictamente creciente , ya que

$$IC_X(x) > IC_X(x') \iff f(x) > f(x') \iff$$

$$\iff \emptyset[IC_X(x)] > \emptyset[IC_X(x')]$$

Teorema 2.2.6.

Si C es un criterio T-evaluable sobre $X \subset R^k$, con X numerable , existe una funcion estrictamente C_X-creciente .

Demostracion:

$$\text{Sea } r_{ij} = \begin{cases} 1 \text{ si } x_i +C_X x_j \\ \\ 0 \text{ en otro caso} \end{cases} \qquad \text{con } i,j \in N$$

y considerese $IC_X(x_i) = \sum_{j=1}^{\infty} (1/2)^j r_{ij}$

de donde se deduce que si x_i , $x_k \in X$

$$x_i +C_X x_k \iff IC_X(x_i) > IC_X(x_k)$$

Teorema 2.2.7.

Sea C un criterio T-evaluable sobre $X \subset R^k$, con X numerable, e IC_X estrictamente C_X-creciente.

Entonces una funcion

$$f : X \dashrightarrow R \quad , \quad \text{es estrictamente}$$

C_X-creciente , sii:

$\exists \emptyset$: $IC_X(X) \dashrightarrow R$, estrictamente creciente / $\emptyset \cdot IC_X = f$

La demostracion de este resultado es analoga a la del teorema 2.2.5.

Corolario 2.2.6.

Sea C un criterio T-evaluable sobre $X \subseteq R^k$.

Si $X/\sim C_X$ es numerable :

$$\exists\ IC_X : X \dashrightarrow R ,$$

estrictamente C_X-creciente.

En el caso en que se desee analizar la existencia de una funcion estrictamente C_X-decreciente, los resultados obtenidos son igualmente validos , ya que si IC_X es estrictamente C_X-creciente se tendra que $-IC_X$ es estrictamente C_X-decreciente.

En algunos casos, puede ser de interes el estudio de funciones bajo condiciones mas debiles, por lo que se da la siguiente

Definicion 2.2.10.

Sea C un criterio sobre $X \subseteq R^k$, que define una relacion binaria C_X . Se dira que una funcion

$$IC_X : X \dashrightarrow R , \quad \text{es :}$$

i) C_X-creciente si $x\ C_X\ y \Longrightarrow IC_X(x) \geq IC_X(y)$

ii) C_X-decreciente si $x\ C_X\ y \Longrightarrow IC_X(x) \leq IC_X(y)$

iii) C_X-orden convexa si

$$x\ C_X\ z\ C_X\ y \Longrightarrow IC_X(z) \geq \min\{IC_X(x), IC_X(y)\}$$

La existencia de los distintos tipos de funciones introducidas en la definicion 2.2.10 esta garantizada, ya

que las funciones constantes son C_X-crecientes (decrecientes) y C_X-orden convexas, siendo de interes los siguientes resultados

Proposicion 2.2.3.

Dados un criterio T-evaluable sobre $X \subseteq R^k$, y $\{ f_j \}_{j \in J}$, una familia de funciones reales definidas sobre X, sea $f = \inf_{j \in J} \{ f_j \}$

Entonces si :

i) f_j es C_X-creciente (C_X-decreciente) $\forall\, j \in J$ ==> f es C_X-creciente (C_X-decreciente)

ii) f_j es C_X-orden convexa $\forall\, j \in J$ ==> f es C_X-orden convexa.

Definicion 2.2.11.

Sea C un criterio T-evaluable sobre $X \subseteq R^k$ y

$$f : X \dashrightarrow R .$$

i) Se dira que $f^* : X \dashrightarrow R$ es el cierre C_X-creciente de f y se representara por $CC_X(f)$, si verifica:

C1 . Es C_X-creciente

C2 . $f^*(x) \geq f(x) \quad \forall\, x \in X$

C3 . $\forall\, f'$, C_X-creciente / cumple C1,C2

$$f'(x) \geq f^*(x) \quad \forall\, x \in X$$

ii) Se dira que $f_*: X \dashrightarrow R$ es el cierre C_X-decreciente de f y se representara por $DC_X(f)$, si verifica:

D1 . Es C_X-decreciente

D2 . $f_\star(x) \geq f(x) \quad \forall\, x \in X$

D3 . $\forall\, f''$, C_X-decreciente / cumple D1,D2

$f''(x) \geq f_\star(x) \quad \forall\, x \in X$

iii) Se dira que $\varphi : X \dashrightarrow R$ es el cierre C_X-orden convexo de f , y se representara por $OcC_X(f)$, si verifica:

OC1 . Es C_X-o.convexa

OC2 . $\varphi(x) \geq f(x) \quad \forall\, x \in X$

OC3 . $\forall \varphi'$, C_X-o.convexa / cumple OC1,OC2

$\varphi'(x) \geq \varphi(x) \quad \forall\, x \in X$

Proposicion 2.2.4.

Dados un criterio T-evaluable sobre $X \subseteq R^k$ y una funcion

$f : X \dashrightarrow R$, se verifica:

i) $CC_X(f)(x) = \sup \{ f(y) / x\ C_X\ y\} = \sup_{y \in dC_X(x)} \{ f(y)\}$

ii) $DC_X(f)(x) = \sup \{ f(y) / y\ C_X\ x\} = \sup_{y \in DC_X(x)} \{ f(y)\}$

iii) $OcC_X(f)(x) = \sup\{\min \{f(x_1),f(x_2)\} / x_2\ C_X\ x\ C_X\ x_1\}$

$= \sup_{(x_1,x_2) \in DC_X(x) \times dC_X(x)} \{ \min \{ f(x_1), f(x_2)\}$

Demostracion.

i) Sea $f_o : X \dashrightarrow R$, con

$f_o(x) = \sup \{ f(y) / x\ C_X\ y \}$

Como se verifica $x\ C_X\ x\ \forall x \Longrightarrow f(x) \leq f_o(x)$

y como ademas

$$x_2\ C_X\ x_1 \Longrightarrow dC_X(x_1) \subseteq dC_X(x_2) \Longrightarrow f_o(x_1) \leq f_o(x_2)$$

se concluye que f_o es C_X-creciente.

Ademas para cualquier $f' : X \dashrightarrow R$, C_X-creciente con $f(x) \leq f'(x)\ \forall x$, se verifica :

$$x\ C_X\ y \Longrightarrow f'(x) \geq f'(y) \geq f(y)$$

de donde se deduce que

$$f_o = CC_X(f)$$

ii) Se demostraria de forma analoga.

iii) Sea $f_1 : X \dashrightarrow R$, con

$$f_1(x) = \sup\{ \min\{ f(x_1), f(x_2)\} / \ x_2\ C_X\ x\ C_X\ x_1 \}$$

Evidentemente $f_1(x) \geq f(x)\ \forall x$.

Sea $\theta \in R$, tal que

$\theta < \min\{ f_1(x), f_1(y)\} \Longrightarrow \exists\ x_1, x_2$ tales que

$x_2\ C_X\ x\ C_X\ x_1$ con $\min\{ f(x_1), f(x_2)\} > \theta$

De forma analoga $\exists\ y_1, y_2$ tales que

$y_2\ C_X\ y\ C_X\ y_1$ con $\min\{ f(y_1), f(y_2)\} > \theta$

Por tanto para

$$x,y,z \in X / \ y\ C_X\ z\ C_X\ x$$

se verifica

$$y_2\ C_X\ y\ C_X\ z\ C_X\ x\ C_X\ x_1 \Longrightarrow$$

$f_1(z) \geq \min\{ f(x_1), f(y_2)\} > \theta$

para cualquier $\theta / \ \theta < \min\{ f_1(x), f_1(y) \}$

obteniendose que

$f_1(z) \geq \min \{ f_1(x), f_1(y)\} \Longrightarrow \quad f_1$ es C_X-o.convexa.

Para cualquier f'' C_X-o.convexa , verificando

$$f''(x) \geq f(x) \quad \forall\, x \in X ,$$

$x_2\ C_X\ x\ C_X\ x_1 \Longrightarrow f''(x) \geq \min \{ f''(x_1), f''(x_2)\} \geq$

$\geq \min \{ f(x_1), f(x_2)\} \Longrightarrow f''(x) \geq f_1(x) \quad \forall\, x \in X ,$

concluyendo pues que $f_1 = OcC_X(f)$.

Lema 2.2.3.

Dado un criterio T-evaluable , C , sobre $X \subseteq R^k$, y f_1 , f_2 , tales que f_1 es C_X-creciente y f_2 es C_X-decreciente, se verifica :

$f = \min \{ f_1, f_2 \}$ es C_X-o.convexa.

Demostracion:

Si $x\ C_X\ z\ C_X\ y$ entonces

$f_1(z) \geq f_1(y) \geq f(y)$; $f_2(z) \geq f_2(x) \geq f(x)$

y por tanto

$$f(z) \geq \min \{ f(x), f(y) \} \Longrightarrow$$

es C_X-o.convexa.

Teorema 2.2.8.

$$OcC_X(f) = \min \{ CC_X(f), dC_X(f)\}$$

Demostracion:

$\varphi = \min \{ CC_X(f), dC_X(f)\}$ es C_X-o.convexa por lema 2.2.3 y es evidente que $f(x) \leq \varphi(x) \quad \forall\, x \in X \Longrightarrow$

$\Longrightarrow Oc C_X(f)(x) \leq \varphi(x) \quad \forall x \in X$.

Si $\exists x \in X / Oc C_X(f)(x) \neq \varphi(x) \Longrightarrow$

$\Longrightarrow \exists \theta / Oc C_X(f)(x) < \theta < \varphi(x) \leq CC_X(f)(x) \Longrightarrow$

$\Longrightarrow \exists y_1 / x\, C_X\, y_1$ con $f(y_1) > \theta$

y como $\varphi(x) \leq dC_X(f)(x) \Longrightarrow \exists y_2 / y_2\, C_X\, x$ con $f(y_2) > \theta$, se deduce :

$y_2\, C_X\, x\, C_X\, y_1 \Longrightarrow Oc C_X(f)(x) \geq \min \{ f(y_2), f(y_1) \} > \theta$

por lo que $Oc C_X(f) = \varphi$.

Teorema 2.2.9.

Sea C un criterio T-evaluable sobre $X \subseteq R^k$ y sea

$f : X \dashrightarrow R$.

Entonces f es C_X-o.convexa sii existen dos funciones f_1 , C_X-creciente , y f_2 , C_X-decreciente de forma que :

$$f = \min \{ f_1 , f_2 \}.$$

Demostracion:

Si f es C_X-o.convexa $\Longrightarrow$ $f = Oc C_X(f) \Longrightarrow$

$\Longrightarrow$ $f = \min \{ CC_X(f), dC_X(f) \}$

Reciprocamente , si $f = \min \{ f_1, f_2 \}$, con f_1 C_X-creciente y f_2 C_X-decreciente, del lema 2.2.3 , se deduce que f es C_X-o.convexa.

Para el analisis de las situaciones experimentales , el estadistico dispone de un conjunto finito de observaciones $X \subset R^k$, con $\#(X) = n$.

En este caso , la existencia de una funcion

$$IC_X : X \dashrightarrow R$$

verificando la condicion

$$(x,y) +C'_X (z,w) \langle == \rangle IC_X(x)-IC_X(y) \langle IC_X(z)-IC_X(w)$$

viene determinada por los teoremas 2.2.1 y 2.2.3 .

En cualquier caso siempre existiran C_X-crecientes , C_X-decrecientes y C_X-o.convexas ,y la existencia de funciones estrictamente C_X-crecientes (decrecientes) ,queda garantizada por el teorema 2.2.4 , en el caso en que el criterio C sea T-evaluable.

Sin embargo , debido a la falta de unicidad de estas funciones, sera necesario utilizar algun procedimiento de seleccion entre las mismas, mediante el cual se refleje como se desvian las observaciones del comportamiento general de la masa de datos, respecto del criterio bajo estudio.

El procedimiento de seleccion que se utiliza en el Capitulo III , viene expresado a traves de la optimizacion de un funcional real,que depende de las observaciones y del conjunto de funciones que se desea analizar.

Asi , si se denota por $\mathfrak{X} = \{ X \subset R^k \ / \ \#(X) = n \}$,

$$F : \mathfrak{X} \times \mathcal{M} \dashrightarrow R$$

$$F(X,\mu) = F(x_1,..,x_n, \mu(x_1),.., \mu(x_n))$$

donde $\mathcal{M}$ representa al conjunto de funciones que se desea analizar.

2.3 INVARIANZA.

Admitiendo el caracter subjetivo del criterio C , y del funcional F , seria deseable que las conclusiones o resultados que se obtengan no dependan del sistema de coordenadas elegido para representar los elementos de , por lo que a continuacion se estudia este problema .

En el estudio que se realiza , se supone que g es una funcion

$$g : R^k \dashrightarrow R^k \quad \text{uno a uno}$$

En estas condiciones se dan las siguientes

Definicion 2.3.1.

Un criterio , C , T-evaluable sobre los elementos de $\mathfrak{X}$, se dira que es g-invariante si verifica

$$x\ C_X\ x' \iff g(x)\ C_{g(X)}\ g(x') \quad \forall\ x,x' \in X\ ,\ \forall\ X \in \mathfrak{X}$$

Se denotara por

$$I_{\mathfrak{X}}(C)=\{\ g\ /\ C\ \text{es g-invariante}\ \}$$

Definicion 2.3.2.

Sea C un criterio T-evaluable sobre los elementos de $\mathfrak{X}$, y $SSC(\mathfrak{X}) = \{\ SSC_X\ /\ X \in \mathfrak{X}\ \}$, donde SSC_X es una funcion de C-similitud signada sobre X.

$SSC(\mathfrak{X})$ es g-invariante , si

$$(x,y)\ C'_X\ (z,w) \iff (g(x),g(y))\ C'_{g(X)}\ (g(z),g(w))$$

para cualesquiera $x,y,z,w \in X$, $\forall\ X \in \mathfrak{X}$.

Se representara por ISSC($\mathfrak{X}$) al conjunto

ISSC($\mathfrak{X}$) = { g / SSC($\mathfrak{X}$) es g-invariante }

Proposicion 2.3.1.

$$\text{Si } g \in \text{ISSC}(\mathfrak{X}) \Longrightarrow g \in I_{\mathfrak{X}}(C)$$

Demostracion:

$$x\ C_X\ y \Longleftrightarrow (x,x)\ C'_X\ (x,y) \Longleftrightarrow$$

$$(g(x),g(x))\ C'_{g(X)}(g(x),g(y)) \Longleftrightarrow g(x)\ C_{g(X)}\ g(y)$$

Proposicion 2.3.2.

Si C es un criterio T-evaluable sobre los elementos de $\mathfrak{X}$ y $\varphi \subset I_{\mathfrak{X}}(C)$, siempre existe un grupo G, cuya ley interna es la composicion de funciones, verificando

$$\varphi \subset G \subset I_{\mathfrak{X}}(C)$$

Demostracion:

En efecto , $\forall\ g,g' \in \varphi \Longrightarrow g \cdot g' ,\ g^{-1} \in I_{\mathfrak{X}}(C)$

ya que $\forall\ x,x' \in X$, $\forall\ X \in \mathfrak{X}$:

i) $x\ C_X\ x' \Longleftrightarrow g'(x)\ C_{g'(X)}\ g'(x') \Longleftrightarrow gg'(x)\ C_{gg'(X)}\ gg'(x')$

ii) $g^{-1}(x)\ C_{g^{-1}(X)}\ g^{-1}(x') \Longleftrightarrow gg^{-1}(x)\ C_{gg^{-1}(X)}\ gg^{-1}(x') \Longleftrightarrow$

$\Longleftrightarrow x\ C_X\ x'$

de donde se deduce de forma inmediata el resultado de la proposicion.

Proposicion 2.3.3.

Dado SSC($\mathfrak{X}$) y Ψ un subconjunto de ISSC($\mathfrak{X}$), existe un grupo G , cuya ley interna es la composicion de funciones ,verificando

$$\Psi \subset G \subset ISSC(\mathfrak{X})$$

La demostracion es analoga a la de la proposicion 2.3.2.

Al igual que se han introducido los conceptos de de invarianza de un criterio y de una funcion de C-similitud signada, para que las relaciones de preferencia definidas no dependan del sistema de coordenadas elegido, tambien es deseable que las funciones IC_X , que caracterizan a las distintas relaciones, fuesen tambien independientes del sistema de coordenadas .

Con tal objetivo se introduce:

Definicion 2.3.3.

Sea C un criterio T-evaluable sobre los elementos de $\mathfrak{X}$, e $IC(\mathfrak{X}) = \{ IC_X / X \in \mathfrak{X} \}$ donde

$$IC_X : X \dashrightarrow R$$

Se dira que $IC(\mathfrak{X})$ es g-invariante si verifica :

$$IC_X(x) = IC_{g(X)}(g(x)) \quad \forall x \in X , \ \forall X \in \mathfrak{X} .$$

Se denotara por $IIC(\mathfrak{X})$ al conjunto de funciones

$$IIC(\mathfrak{X}) = \{ g / IC(\mathfrak{X}) \text{ es g-invariante } \}$$

Proposicion 2.3.4.

Si C es un criterio T-evaluable sobre los elementos de $\mathfrak{X}$ y $\Psi \subset IIC(\mathfrak{X})$, existe un grupo de transformaciones G , cuya ley interna es la composicion de funciones , que verifica :

$$\Psi \subset G \subset IIC(\mathfrak{X})$$

Proposicion 2.3.5.

Sea $IC(\mathfrak{X}) = \{IC_X / X \in \mathfrak{X}\}$, de forma que IC_X es estrictamente C_X-creciente (C_X-decreciente).

Entonces :

$$g \in IIC(\mathfrak{X}) \Longrightarrow g \in I_{\mathfrak{X}}(C).$$

Demostracion:

$$x\ C_X\ x' \Longleftrightarrow IC_X(x) \geq IC_X(x') \Longleftrightarrow$$

$$\Longleftrightarrow IC_{g(X)}(g(x)) \geq IC_{g(X)}(g(x')) \Longleftrightarrow g(x)\ C_{g(X)}\ g(x')$$

Proposicion 2.3.6.

Dado $SSC(\mathfrak{X})$, sea $IC(\mathfrak{X})$ verificando :

$$(x,y)\ C'_X\ (z,w) \Longleftrightarrow IC_X(x) - IC_X(y) < IC_X(z) - IC_X(w)$$

$$\forall\ x,y,z,w \in X \ ; \ \forall\ X \in \mathfrak{X}$$

Entonces : $g \in IIC(\mathfrak{X}) \Longrightarrow g \in ISSC(\mathfrak{X})$

Ya que las funciones IC_X , que se utilizaran en el capitulo siguiente se obtienen mediante la optimizacion de un funcional F, es logico estudiar bajo que condiciones dichas funciones son invariantes .

El teorema esta referido al caso de funciones estrictamente C_X-crecientes pero en el caso de estrictamente C_X-decrecientes se obtiene un resultado analogo.

Teorema 2.3.1.

Sea un criterio C , T-evaluable sobre los elementos de $\mathfrak{X}$, y $g \in I_{\mathfrak{X}}(C)$, de forma que :

$F(g(X), \mu) = g^{*} F(X, \mu)$ con g^{*} estrictamente creciente.

Entonces $IC^{*}(\mathfrak{X}) = \{ IC_X^{*} / X \in \mathfrak{X} \}$ es g-invariante si IC_X^{*} se determina a traves de la maximizacion (minimizacion) del funcional F , sobre la clase de funciones estrictamente C_X-crecientes.

Demostracion:

Si se denota por CC_X a la clase de funciones estrictamente C_X -crecientes , se tiene

$$\left.\begin{array}{l} \text{Max} \quad F(g(X), \mu) \\ \mu \in CC_{g(X)} \end{array}\right\} \equiv \begin{array}{l} \text{Max} \quad F(g(x_1), .., g(x_n), \mu_1, \ldots, \mu_n) \\ \text{sujeto a} \\ g(x_i)\ C_{g(X)}\ g(x_j) \iff \mu_i \geq \mu_j \end{array}$$

que debido a las hipotesis es equivalente a

$$\left.\begin{array}{l} \text{Max} \quad g^{*} F(x_1, \ldots, x_n, \mu_1, \ldots, \mu_n) \\ \text{sujeto a} \\ x_i\ C_X\ x_j \iff \mu_i \geq \mu_j \end{array}\right\} \equiv \begin{array}{l} \text{Max} \quad F(X, \mu) \\ \mu \in CC_X \end{array}$$

Teorema 2.3.2.

Sea C un criterio T-evaluable sobre los elementos de $\mathfrak{X}$, y $g \in ISSC(\mathfrak{X})$ de forma que

$$F(g(X), \mu) = g^{*} F(X, \mu)$$

con g^{*} estrictamente creciente.

Entonces $g \in IIC^{*}(\mathfrak{X})$, si IC_X^{*} se determina a traves de

$$\text{Max} \quad F(x_1,\ldots,x_n, \mu_1,\ldots,\mu_n)$$

sujeto a

$$(x_i,x_j) + C'_X (x_h,x_k) \iff \mu_i - \mu_j < \mu_h - \mu_k$$

La demostracion de este resultado es analoga a la del Teorema 2.3.1

CAPITULO III

ANALISIS CUALITATIVO DE DATOS .

TECNICAS DE IDENTIFICACION DE OUTLIERS .

-- Introduccion

3.1 Estructura Basica. Aproximacion General.

3.2 Criterio de Dispersion Central

3.2.1 Outliers

3.2.2 Caso de Parametros Poblacionales Conocidos

3.3 Criterio de Dispersion Basado en el Recorrido

3.3.1 Outliers

3.3.2 Caso de Parametros Conocidos

3.4 Criterio Natural (Origen Conocido)

3.4.1 Outliers

3.5 Generalizacion al Caso de mas de una Observacion

3.5.1 Criterio de Dispersion Central

3.5.2 Criterio de Dispersion Basado en el Recorrido

3.5.3 Criterio Natural

3.6 Criterio de Dispersion Central Multivariante

INTRODUCCION

En el capitulo I se ha destacado la problematica que presentan, actualmente, las Tecnicas de Identificacion de Outliers.

En este capitulo, se aplican los resultados obtenidos en el capitulo anterior, con el objeto de obtener algunos de los estadisticos que son de especial interes dentro de las tecnicas de deteccion de outliers.

En el apartado de conclusiones, se refleja como con el planteamiento propuesto se corrigen algunos de los defectos que hoy en dia aun persisten en la teoria de Outliers.

3.1 ESTRUCTURA BASICA. APROXIMACION GENERAL.

Sea un conjunto $X \subset R^k$ / $\#(X) = n$, en el que se desea comparar sus elementos a traves de un criterio C , el cual determina una relacion binaria C_X definida por:

$$x \; C_X \; x' \iff C_X(x) \leq C_X(x') \qquad (3.1)$$

con $C_X : X \dashrightarrow R$.

Esta relacion binaria da lugar a los conjuntos C_X-dominantes:

$$DC_X(x) = \{ x' \in X \; / \; C_X(x') \leq C_X(x) \} \qquad x \in X$$

y C_X-dominados:

$$dC_X(x) = \{ x' \in X \; / \; C_X(x) \leq C_X(x') \} \qquad x \in X$$

introducidos en la definicion 2.1.4.

Cualquier subconjunto $X' \subseteq X$, estara propiamente C_X-dominado , y los elementos $(C_X \; D \; m)$ y $(C_X \; d \; M)$ de X' , que se representaran por x^* y x_* , respectivamente,son aquellos que verifican :

$$\underset{x \in X'}{\text{Min}} \; C_X(x) = C_X(x^*)$$

$$\underset{x \in X'}{\text{Max}} \; C_X(x) = C_X(x_*)$$

Ademas el criterio C es T-evaluable sobre X,y las clases de equivalencia estan formadas por aquellos elementos sobre los que la funcion C_X es constante.

De forma natural; la relacion binaria dada en (3.1) , da lugar a las funciones:

$$SC_X : X \times X \longrightarrow R \qquad (3.2)$$

$$SC_X(x,x') = |C_X(x') - C_X(x)|$$

y

$$SSC_X : X \times X \longrightarrow R \qquad (3.3)$$

$$SSC_X(x,x') = C_X(x') - C_X(x)$$

Proposicion 3.1.1.

Las funciones SC_X y SSC_X , dadas en (3.2) y (3.3),son funciones de C-similitud y C-similitud signada sobre X ,respectivamente, que son aditivas, en el sentisentido de la definicion 2.2.7 ,a las que se denominaran funciones de C-similitud y C-similitud signada naturales.

La demostracion de esta proposicion es evidente a partir de las definiciones dadas de SC_X y SSC_X .

De forma analoga a (3.2) y (3.3),pueden definirse las funciones de C-similitud y C-similitud signada , tipificadas como:

$$SC_X(x,x') = [\,|C_X(x') - C_X(x)|\,] \,/\, [C_X(x_*) - C_X(x^*)]$$

$$SSC_X(x,x') = [\,C_X(x') - C_X(x)\,] \,/\, [C(x_*) - C(x^*)]$$

La funcion de C-similitud tipificada puede interpretarse como una relacion difusa definida en X x X ,a traves del criterio C.

Proposicion 3.1.2.

Dada la funcion de C-similitud signada natural , existe

$$IC_X : X \text{ -----} > R$$

verificando:

$$(x,y)\ C'_X\ (z,w) \iff IC_X(y)-IC_X(x) \leq IC_X(w)-IC_X(z) \quad (3.4)$$

siendo C'_X ,la relacion introducida en la definicion 2.2.4.

El resultado de esta proposicion es inmediato, ya que seria suficiente considerar $IC_X = C_X$.

Proposicion 3.1.3.

Para la funcion de C-similitud signada natural , se verifica :

$\forall\ x_i , x'_i \in X$, i=1,2,3 , tales que

$$\left.\begin{array}{l}(x_1,x_2)\ C'_X\ (x'_1,x'_2)\\ (x_2,x_3)\ C'_X\ (x'_2,x'_3)\end{array}\right\} \Longrightarrow \quad (x_1,x_3)\ C'_X\ (x'_1,x'_3)$$

que es la condicion T1 del Teorema 2.2.1.

Sin embargo :

Dados $y,x_i \in X$, i=1,2,3,4 , tales que

$$(x_1,x_2)\ C'_X\ (x_3,x_4) \quad ; \quad (x_3,x_4)\ C'_X\ (y,y)$$

no esta garantizada la existencia de $z_1 , z_2 \in X$, cumpliendo :

$$(x_1,z_1)\ {\sim}C'_X\ (x_3,x_4) \quad ; \quad (z_2,x_2)\ {\sim}C'_X\ (x_3,x_4)$$

por tanto , la condicion T2 del Teorema 2.2.1 , no siempre se verifica ,por lo que no queda garantizada la uni-

cidad salvo transformaciones lineales de IC_X .

Debido a la falta de unicidad , es necesario considerar algun criterio de optimalidad que permita seleccionar entre las diversas funciones que verifican (3.4).

En los casos que se analizan posteriormente , se considera como funcion IC_X optima aquella que :

$$\text{Min} \sum_{i=1}^{n} \{ [IC_X(x_i)]^2 [KC_X - C_X(x_i)] + [1-IC_X(x_i)]^2 C_X(x_i) \}$$

sujeto a

$$(x,y)\ C'_X\ (z,w) \iff IC_X(y)-IC_X(x) \leq IC_X(w)-IC_X(z)$$

siendo KC_X ,una constante positiva que depende del conjunto X y del criterio C , que se consideran.

Este criterio de optimalidad propuesto puede considerarse como una modificacion del dado por Dunn (1974).

El optimo IC_X^* , sin restricciones ,viene dado por :

$$IC_X^*(x) = C_X(x) / KC_X \qquad x \in X$$

el cual verifica la restriccion impuesta en (3.4) . En efecto:

$$(x,y)\ C'_X\ (z,w) \iff C_X(y)-C_X(x) \leq C_X(w)-C_X(z) \iff$$

$$IC_X^*(y)-IC_X^*(x) \leq IC_X^*(w)-IC_X^*(z)$$

En el caso en que $IC_X^*(x) \in [0,1]$, esta funcion puede interpretarse como un conjunto difuso definido sobre X a traves del criterio C.

Aunque de forma general,no esta garantizada la unicidad (salvo transformaciones lineales) de la funcion IC_X, existen situaciones en las que se verifica la condicion T2 del Teorema 2.2.1 , por ejemplo , cuando la funcion C_X , es tal que :

$$C_X(X) = \{ 0,1,\ldots,r \; ; \; r \leq n-1 \}$$

o de forma mas general

$$C_X(X) = \{ \delta.i + K \; ; \; i = 0,\ldots,r \; ; \; r \leq n-1 \}$$

en cuyo caso queda garantizada la unicidad salvo trans-transformaciones lineales de IC_X .

Es de destacar que esta funcion no seria valida para los propositos de la memoria , ya que en esta situacion no seria posible determinar que elementos se desvian marcadamente del comportamiento general de la masa de datos,ya que los rangos son robustos frente a la presencia de outliers.

A continuacion , se van a obtener mediante la aplicacion del procedimiento estudiado , algunos estadisticos de gran relevancia dentro de las actuales tecnicas de identificacion de outliers.

3.2 CRITERIO DE DISPERSION CENTRAL

Sea $X \subset R$ / $\#(X) = n$, sobre el que se desea analizar el comportamiento de sus elementos respecto del criterio C , que define la relacion binaria :

$x\ C_X\ x'$ sii x esta al menos tan proximo a $\bar{x}$ como x' .

Esta relacion puede expresarse mediante

$$x\ C_X\ x' \quad <==> \quad (x - \bar{x})^2 \leq (x' - \bar{x})^2 \qquad (3.5)$$

o bien por

$$x\ C_X\ x' \quad <==> \quad |x - \bar{x}| \leq |x' - \bar{x}| \qquad (3.6)$$

Evidentemente , (3.5) y (3.6) definen la misma relacion binaria sobre el conjunto X.

En el analisis que se realiza a continuacion se trabaja con la forma (3.5).

En este caso los conjuntos C_X-dominantes y C_X-dominados son :

$$DC_X(x) = \{\ x' \in X\ /\ (x' - \bar{x})^2 \leq (x - \bar{x})^2\ \}$$

$$dC_X(x) = \{\ x' \in X\ /\ (x - \bar{x})^2 \leq (x' - \bar{x})^2\ \}$$

respectivamente.

Como se indico en el epigrafe 3.1 , todo $X' \subseteq X$ esta propiamente C_X-dominado , y ademas, en este caso , x_* , el elemento (C_X d M) de X , verifica:

$$(x_* - \bar{x})^2 = \underset{x \in X}{\text{Max}} \{(x - \bar{x})^2\} = \text{Max}\ \{(x_{(1)} - \bar{x})^2, (x_{(n)} - \bar{x})^2\}$$

donde $x_{(1)}$, $x_{(n)}$, son , respectivamente, la menor y la mayor de las observaciones , por lo que el elemento $(C_X$ d M) de X , sera $x_{(1)}$ ó $x_{(n)}$.

Como ya se indico en epigrafe 3.1 , el criterio C es T-evaluable sobre X ,y en este caso particular, las clases de equivalencia estan formadas por elementos equidistantes a la media.

Para esta situacion , las funciones de C-similitud y C-similitud signada , naturales , introducidas mediante (3.2) y (3.3) vienen dadas por:

$$SC_X(x,x') = |(x - \bar{x})^2 - (x' - \bar{x})^2|$$

$$SSC_X(x,x') = (x' - \bar{x})^2 - (x - \bar{x})^2$$

Como quedo reflejado en el epigrafe anterior , en general la funcion de similitud signada natural , no verifica la condicion T2 del Teorema 2.2.1 , por lo que no quedaria garantizada la unicidad de IC_X .

El optimo se seleccionara mediante :

(3.7)

$$\text{Min} \sum_{i=1}^{n} \{ [IC(x_i)]^2(nS_X^2 - (n-1)S_i^2) + [1-IC_X(x_i)]^2(n-1)S_i^2 \}$$

sujeto a

$$(x_j,x_k)\ C'_X\ (x_l,x_m) \iff IC_X(x_j)-IC_X(x_k) \leq IC_X(x_l)-IC_X(x_m)$$

con

$$nS_X^2 = \sum_{p=1}^{n} (x_p - \bar{x})^2 \quad ; \quad (n-1)S_i^2 = \sum_{\substack{p=1 \\ p\neq i}}^{n} (x_p - \bar{x}_i)^2$$

y $\quad \bar{x}_i = [\sum_{\substack{p=1 \\ p \neq i}}^{n} x_p]/(n-1)$

Debido a que la funcion de C-simlilitud signada natural es invariante por traslacion y escala, y ademas

$$F(.,a\,x_i+b,\ldots,IC_X(a\,x_i+b),.) = a^2\,F(.,x_i,\ldots,IC_X(x_i),.)$$

siendo F el funcional a optimizar en (3.7) , a partir del Teorema 2.3.2 , se concluye que el optimo del funcional es invariante por localizacion y escala.

El optimo del funcional F , sin restricciones , viene dado por :

$$IC_X^{\star}(x_i) = \frac{(n-1)\,S_i^2}{n\,S_X^2}$$

el cual verifica la restriccion , ya que:

$$IC_X^{\star}(x_j)-IC_X^{\star}(x_k) \leq IC_X^{\star}(x_l)-IC_X^{\star}(x_m) \langle==\rangle S_j^2 - S_k^2 \leq S_l^2 - S_m^2$$

$$\langle==\rangle (x_k - \bar{x})^2 - (x_j - \bar{x})^2 \leq (x_m - \bar{x})^2 - (x_l - \bar{x})^2 \quad \langle==\rangle$$

$$\langle==\rangle (x_j, x_k)\ C_X'\ (x_l, x_m)$$

debido a que

$$n\,S_X^2 = (n-1)\,S_i^2 + [(n-1)/n]\,(x_i - \bar{x}_i)^2$$

y

$$(x_i - \bar{x})^2 = [(n-1)/n]^2\,(x_i - \bar{x}_i)^2$$

Se ha de notar :

I) $IC_X^{\star}(x) \in [0,1] \quad \forall\, x \in X$, lo que se puede interpre-

tar como un conjunto difuso definido sobre X, mediante el criterio C.

II) $IC^{\star}_{X}(x) = 1 \quad sii \quad x = \bar{x}$

III) En el caso particular de $n=2$

$S_1^2 = (x_2 - \bar{x}_1)^2 = (x_2 - x_2)^2 = 0$, y analogamente

$S_2^2 = 0$, por lo que $IC^{\star}_{X}(x_i) = 0$, $i=1,2$

IV) Dado $\alpha \in (0,1)$

$$IC^{\star}_{X}(x_j) \leq \alpha \iff S_j^2 \leq [\alpha / (1-\alpha)][n/(n-1)^2](x_j - \bar{x})^2$$

y en el caso particular $\alpha = 1/2$

$$IC^{\star}_{X}(x_j) \leq 1/2 \iff \sum_{x \neq x_j} (x - \bar{x})^2 \leq [(n+1)/(n-1)](x_j - \bar{x})^2$$

lo que reflejaria que dicha observacion tiene una gran influencia sobre la dispersion de la masa de datos.

3.2.1 OUTLIERS.

Grubbs (1950), propone los estadisticos :

$$T_1 = [(n-1)\, S^2_{(1)}] \,/\, [n\, S^2_X]$$

$$T_2 = [(n-1)\, S^2_{(n)}] \,/\, [n\, S^2_X]$$

para detectar la presencia de un outlier inferior o superior, respectivamente , en una poblacion normal con media y varianza desconocidas.

Es de destacar que

$T_1 = IC^{\star}_{X}(x_{(1)})$ y $T_2 = IC^{\star}_{X}(x_{(n)})$

Ademas, Tietjen y Moore (1972) proponen el estadistico

$$T_3 = \text{Min} \{ T_1 , T_2 \}$$

para detectar la presencia de un outlier superior o inferior en una poblacion normal de media y varianza desconocidas, observandose que

$$T_3 = \text{Min} \{ IC_X^{\star}(x) / x \in X \}$$

3.2.2 CASO DE PARAMETROS POBLACIONALES CONOCIDOS.

En el caso en que se dispusiese de alguna informacion a priori, sobre la poblacion de la que se ha extraido el el conjunto de observaciones X , seria conveniente, hacer uso de la misma.

Asi , en el caso en que la media de la poblacion , μ_o , sea conocida, la relacion dada por (3.5), podria modificarse, considerando :

$$x \; C_X \; x' \quad \langle == \rangle \quad (x - \mu_o)^2 \leq (x' - \mu_o)^2$$

Es de destacar, que en este caso ,la relacion binaria C_X es independiente de X.

Los resultados obtenidos en el epigrafe 3.2.1 serian serian facilmente generalizables en esta nueva situacion, obteniendose los estadisticos equivalentes a T_1, T_2 y T_3 , correspondientes al caso de media conocida .

Es de resaltar que $IC_X^{\star}$ si depende del conjunto de observaciones X .

3.3 CRITERIO DE DISPERSION BASADO EN EL RECORRIDO

Sea $X \subset R$ / $\#(X) = n$, sobre el que se desea analizar el comportamiento de sus elementos respecto del criterio C , que define la relacion binaria :

$x\ C_X\ x'$ sii el recorrido de $X - \{x\}$ es mayor ó igual que el de $X - \{x'\}$.

La relacion definida de esta forma , puede expresarse mediante:

$$x_i\ C_X\ x_j \iff R_i \geq R_j \qquad (3.8)$$

donde $R_i = \underset{h \neq i}{\text{Max}} \{ x_h \} - \underset{h \neq i}{\text{Min}} \{ x_h \}$

Si se denota por $R_{(i)}$ al recorrido de $X - \{x_{(i)}\}$ siendo $x_{(i)}$, la i-esima menor de las observaciones ,se tiene :

$$R_{(1)} = x_{(n)} - x_{(2)}$$
$$R_{(i)} = x_{(n)} - x_{(1)} \qquad i=2,\ldots,n-1$$
$$R_{(n)} = x_{(n-1)} - x_{(1)}$$

con lo cual , la relacion definida en (3.8) , da lugar a que los conjuntos C_X-dominantes y C_X-dominados, introducidos en la def. 2.1.4 , verifiquen :

$$x_{(i)} \in DC_X(x_{(j)}) \qquad i= 2,..,n-1 \ ; \ j= 1,..,n$$
$$dC_X(x_{(i)}) = X \qquad i= 2,..,n-1$$

Como quedo reflejado en el epigrafe 3.1 ,todo subconjunto $X' \subseteq X$, esta propiamente C_X-dominado , y x_* , elemento (C_X d M) de X , verifica :

$$R_{x_*} = \text{Min} \{ R_i , i=1,..,n\} = \text{Min} \{ R_{(1)} , R_{(n)} \}$$

por lo que x_* sera $x_{(1)}$ ó $x_{(n)}$, dependiendo de los valores de $R_{(1)}$, $R_{(n)}$.

El criterio C , es T-evaluable sobre X , y ademas , elementos $x_{(i)}$, i = 2,...,n-1 , son C_X-equivalentes .

Por tanto el criterio C, efectua una particion de X , en a lo sumo tres clases de equivalencia.

Para este criterio , las funciones de C-similitud y C-similitud signada, naturales,introducidas mediante las expresiones (3.2) y (3.3), vienen dadas por :

$$SC_X : X \times X \longrightarrow R$$
$$SC_X(x_i,x_j) = |R_i - R_j|$$

y

$$SSC_X : X \times X \longrightarrow R$$
$$SSC_X(x_i,x_j) = R_i - R_j$$

por lo que :

$$SC_X(x_{(i)},x_{(j)}) = \begin{cases} 0 & \text{si } 1 < i,j < n \text{ ó } i=j \\ x_{(2)} - x_{(1)} & \text{si } 1 < i < n \text{ ; } j=1 \\ x_{(n)} - x_{(n-1)} & \text{si } 1 < i < n \text{ ; } j=n \\ |x_{(n)} - x_{(n-1)} - x_{(2)} + x_{(1)}| & \text{si } i=1 \text{ ; } j=n \end{cases}$$

Analogamente , se expresa la funcion de C-similitud signada natural .

De forma general , se puede afirmar que la funcion de similitud signada natural , no verifica la condicion T2

del teorema 2.2.1 ,cumpliendose en el caso particular en que

$$R_{(1)} = 2\,R_{(n)} - R \qquad \text{ó} \qquad R_{(n)} = 2\,R_{(1)} - R$$

Debido pues a que no esta garantizada la unicidad de la funcion IC_X , se seleccionara el optimo a partir de:

$$\text{Min} \quad \sum_{i=1}^{n} \{\, [IC_X(x_i)]^2 [\, R - R_i] + [1-IC_X(x_i)]^2\, R_i \,\}$$

sujeto a

$$(x_j,x_k)\ C'_X\ (x_l,x_m) \iff IC_X(x_j)-IC_X(x_k) \leq IC_X(x_l)-IC_X(x_m)$$

El optimo de este problema ,sera invariante por localizacion y escala, ya que la funcion de C-similitud signada natural lo es , y ademas

$$R_{aX+b} = a\,R_X \qquad a > 0$$

por lo que la invarianza se deduce partir del Teorema 2.3.2.

El optimo del problema anterior , viene dado por :

$$IC^{\star}_X(x_j) = \frac{R_j}{R} = \frac{R_j}{x_{(n)} - x_{(1)}}$$

el cual verifica la restriccion , como se deduce de forma evidente .

Se ha de notar :

(I) $IC^{\star}_X(x) \in [0,1] \qquad \forall\, x \in X$

por lo que se puede interpretar como un conjunto difuso definido sobre X , a traves del criterio C.

(II) Para $n > 2$

$$IC^{\star}_{X}(x_{(i)}) = 1 \qquad i = 2,\ldots,n-1$$

En el caso particular $n = 2$:

$$IC^{\star}_{X}(x_{i}) = 0 \qquad i = 1,2$$

3.3.1 OUTLIERS.

(I)

$$IC^{\star}_{X}(x_{(1)}) = \frac{x_{(n)} - x_{(2)}}{x_{(n)} - x_{(1)}}$$ por lo que

$1 - IC^{\star}_{X}(x_{(1)})$ coincide con el estadistico dado por Dixon (1950,1951),para testar la presencia de un outlier inferior en una muestra extraida de una poblacion exponencial de origen desconocido.

(II)

$$IC^{\star}_{X}(x_{(n)}) = \frac{x_{(n-1)} - x_{(1)}}{x_{(n)} - x_{(1)}}$$ por lo que

$1 - IC^{\star}_{X}(x_{(n)})$ coincide con el estadistico dado por Dixon (1950,1951) , para testar la presencia de un outlier superior en la misma situacion de (I) y para testar la presencia de un outlier superior en una muestra extraida de una poblacion normal de varianza desconocida.

(III) Max $\{ 1 - IC^{\star}_{X}(x_{(1)}) , 1 - IC^{\star}_{X}(x_{(n)}) \}$ coindice con el estadistico propuesto por King (1953) , para detectar la presencia de un extremo outlier en una

muestra extraida de una poblacion normal de varianza desconocida.

(IV) Los estadisticos recogidos en (I) , (II) y (III) se utilizan tambien para testar la presencia de outliers en muestras extraidas de poblaciones uniformes .

3.3.2 CASO DE PARAMETROS CONOCIDOS.

Si el origen de las observaciones fuese un valor $\underline{a}$ conocido , y se considerase :

$$R = x_{(n)} - a$$

$$R_{(i)} = x_{(n)} - a \qquad i = 1,\ldots,n-1$$

$$R_{(n)} = x_{(n-1)} - a$$

se puede realizar un analisis similar al efectuado anteriormente , llegandose a la conclusion :

$$IC_X^{\star}(x_{(i)}) = 1 \qquad i = 1,\ldots,n-1$$

$$IC_X^{\star}(x_{(n)}) = \frac{x_{(n-1)} - a}{x_{(n)} - a}$$

y por tanto $1 - IC_X^{\star}(x_{(n)})$, coincide con el estadistico utilizado para testar la presencia de un outlier superior en una muestra extraida de una poblacion exponencial de origen $\underline{a}$ conocido .

3.4 CRITERIO NATURAL (ORIGEN CONOCIDO)

Sea un conjunto $X \subset R$ / $\#(X) = n$, $x > a$, $\forall x \in X$ sobre el que se desea analizar el comportamiento de sus elementos respecto de un criterio C , que define sobre X la relacion binaria determinada por la ordenacion natural de las observaciones :

$$x \; C_X \; x' \iff x \leq x' \qquad (x \; C_X \; x' \iff x-a \leq x'-a)$$

Es de destacar , que la relacion binaria expresada de esta forma ,es independiente de X y da lugar a los conjuntos C_X-dominantes y C_X-dominados :

$$DC_X(x) = \{ x' \in X \; / \; x' \leq x \}$$

$$dC_X(x) = \{ x' \in X \; / \; x \leq x' \}$$

y ademas , los elementos (C_X D m) y (C_X d M) de X , son $x_{(1)}$ y $x_{(n)}$, respectivamente.

Evidentemente , el criterio C es T-evaluable sobre X y las clases de equivalencia estan formadas por elementos identicos .

En este caso , las funciones de C-similitud y C-similitud signada naturales , son :

$$SC_X(x,x') = x - x'$$

$$SSC_X(x,x') = |x' - x|$$

De forma general , la funcion de C-similitud signada natural no verifica la condicion T2 del teorema 2.2.1 , sin embargo , dicha condicion si se verificara para el caso en que :

$$x_{(i)} = x_{(1)} + (i-1)\delta \quad , \quad \delta > 0$$

Si se selecciona el optimo a partir de :

$$\text{Min} \sum_{i=1}^{n} \{ [IC_X(x_i)]^2 [\sum_{\substack{j=1 \\ j \neq i}}^{n} (x_j - a)] + [1-IC_X(x_i)]^2 (x_i - a) \}$$

sujeto a

$$(x_j, x_k) \; C'_X \; (x_l, x_m) \iff IC_X(x_k) - IC_X(x_j) \leq IC_X(x_m) - IC_X(x_l)$$

dicho optimo es invariante por localizacion y escala , segun se deduce del Teorema 2.3.2 , y viene dado por :

$$IC^{\star}_X(x_i) = \frac{x_i - a}{\sum_{j=1}^{n} (x_j - a)}$$

y como $IC^{\star}_X(x) \in [0,1]$, este puede interpretarse como un conjunto difuso , definido sobre X , a traves del criterio C.

Es de destacar , que aunque la relacion C_X , no depende de X , $IC^{\star}_X$ si que es funcion del conjunto X.

3.4.1 OUTLIERS

(I)

$$IC^{\star}_X(x_{(1)}) = \frac{x_{(1)} - a}{\sum_{j=1}^{n} (x_j - a)}$$

es el estadistico que estudia Lewis y Fieller (1979) , para testar la presencia de un outlier inferior , en una una muestra procedente de una poblacion Gamma , con origen a conocido.

(II)
$$IC^{*}_{X}(x_{(n)}) = \frac{x_{(n)} - a}{\sum_{j=1}^{n} (x_j - a)}$$

es el estadistico que se recoge en los trabajos de Fisher (1929) , Cochran (1941), para testar la presencia de un outlier superior en una poblacion de las mismas caracteristicas del caso (I) .

3.5 GENERALIZACION AL CASO DE MAS DE UNA OBSERVACION.

En los casos analizados , se ha estudiado el comportamiento de cada elemento del conjunto de observaciones X , respecto de un determinado criterio C . Sin embargo , en ocasiones puede resultar de interes el estudio , del comportamiento de varios elementos de X , respecto de C , de forma simultanea . Para ello se propone el siguiente esquema :

Dado $X = \{ x_1 , \ldots , x_n \}$,se considera el conjunto $X^s = X \times \overset{s}{\ldots} \times X$, al cual es posible extender de forma natural los criterios estudiados anteriormente .

3.5.1 GENERALIZACION DEL CRITERIO DE DISPERSION CENTRAL

Dado $X \subset R$ / $\#(X) = n$,sea $X^s = X \times \overset{s}{\ldots} \times X$, $s < n$, sobre el que se desea analizar el comportamiento de sus elementos respecto del criterio C , que determina la relacion :

$\tilde{x}_i = (x_{i1},\ldots,x_{is})\ C_{X^s}\ (x_{j1},\ldots,x_{js}) = \tilde{x}_j$ sii la

la varianza del conjunto $X - \{ x_{ih}\ ,\ h=1,..,s \}$ es mayor o igual que la de $X - \{ x_{jh}\ ,\ h=1,..,s \}$, es decir :

$$\tilde{x}_i\ C_{X^s}\ \tilde{x}_j \quad \langle == \rangle \quad S^2_{i1,..,is} \geq S^2_{j1,..,js}$$

donde $$S^2_{i1,..,is} = (1/n_i) \sum_{\substack{p=1 \\ p \neq i_h\ ,\ h=1,..,s}}^{n} (x_p - \bar{x}_i)^2$$

con $n_i = \#(X - \{ x_{ih}\ ,\ h=1,..,s \})$

y $$\bar{x}_i = (1/n_i) \sum_{\substack{p=1 \\ p \neq i_h\ ,\ h=1,..,s}}^{n} x_p$$

Es de notar que la relacion definida de esta forma generaliza a la introducida en el epigrafe 3.2 , ya que dados $x\ ,\ x' \in X$:

$$x\ C_X\ x' \quad \langle == \rangle \quad (x,\overset{s}{..},x)\ C_{X^s}\ (x',\overset{s}{..},x')$$

incluso , generalizaria a las relaciones que se definan de igual forma sobre X^r , $r < s$.

De igual forma a como se hizo en los casos ya analizados , se determinan los conjuntos C_{X^s}-dominantes y C_{X^s}-dominados y todo subconjunto de X^s esta propiamente C_{X^s}-dominado , y la relacion es T-evaluable.

Las funciones de C-similitud y C-similitud signada

naturales se definen de igual forma , y el funcional a optimizar , como generalizacion del dado en el epigrafe 3.2 , viene expresado por :

$$\sum_{\tilde{x}_i \in X^s} \{ [IC_{X^s}(\tilde{x}_i)]^2 (nS_X^2 - n_i S_i^2) + [1-IC_{X^s}(\tilde{x}_i)]^2 n_i S_i^2 \}$$

sujeto a

$$(\tilde{x}_j, \tilde{x}_k)\ C'_{X^s}(\tilde{x}_l, \tilde{x}_m) <==> IC_{X^s}(\tilde{x}_j) - IC_{X^s}(\tilde{x}_k) < IC_{X^s}(\tilde{x}_l) - IC_{X^s}(\tilde{x}_m)$$

El optimo de este funcional es invariante por localizacion y escala, segun se dedujo en el epigrafe 3.2 , y viene dado por :

$$IC^{\star}_{X^s}(\tilde{x}_i) = \frac{n_i\ S^2_{il,..,is}}{n\ S_X^2}$$

el cual puede interpretarse como un conjunto difuso definido sobre X^s , a traves del criterio C , ya que $IC^{\star}_{X^s}(\tilde{x}_i) \in [0,1]$, $\forall\ \tilde{x}_i \in X^s$

OUTLIERS

(I) Debido a que C_{X^s} generaliza a C_X , segun se ha visto , pueden obtenerse como casos particulares los resultados del epigrafe 3.2.1 .

(II) Si se representa mediante $(\ldots x_{(i)}, x_{(j)} \ldots)$ a cualquier elemento de X^s con unicamente dos componentes distintas : $x_{(i)}$ y $x_{(j)}$, se tiene que

$$IC^{\star}_{X^s}(\ldots x_{(1)}, x_{(n)} \ldots) = \frac{(n-2)\, S^2_{(1),(n)}}{n\, S^2_X}$$

es el estadistico propuesto por Grubbs (1950) , para detectar la presencia de un outlier inferior y uno superior en una muestra extraida de una poblacion normal de parametros desconocidos.

Analogamente :

$$IC^{\star}_{X^s}(\ldots x_{(1)}, x_{(2)} \ldots) = \frac{(n-2)\, S^2_{(1),(2)}}{n\, S^2_X}$$

$$IC^{\star}_{X^s}(\ldots x_{(n-1)}, x_{(n)} \ldots) = \frac{(n-2)\, S^2_{(n-1),(n)}}{n\, S^2_X}$$

se utilizan para testar la presencia de dos outliers , inferiores y superiores , respectivamente , en la misma situacion antes mencionada.

(III) Si se representa por $(\ldots x_{(n-k+1)}, \ldots, x_{(n)} \ldots)$ a cualquier elemento de X^s , con unicamente k componentes distintas : $x_{(n-k+1)}, \ldots, x_{(n)}$, se obtiene que

$$IC^{\star}_{X^s}(\ldots x_{(n-k+1)},\ldots,x_{(n)}\ \ldots) = \frac{(n-k)\ S^2_{(n-k+1),..,(n)}}{n\ S^2_X}$$

es el estadistico estudiado por Grubbs (1950,1969) , Dixon (1950) , McMillan (1971) , Tietjen y Moore (1972) y Fieller (1976) ,para testar la presencia de k-outliers superiores en las condiciones anteriormente indicadas.

Analogamente se deducirian , como casos particulares los estadisticos utilizados para testar la presencia de k-outliers inferiores ó t-inferiores y k-t superiores (k > t).

Para testar k-outliers , sin especificar direccion , se considera el minimo de $IC^{\star}_{X^s}$, sobre el conjunto de elementos , con k-componentes distintas .

3.5.2 GENERALIZACION DEL CRITERIO DE DISPERSION BASADO EN EL RECORRIDO.

La generalizacion del criterio dado en epigrafe 3.3 , puede realizarse de igual forma en este caso, sin embargo, se prestara atencion especial al caso $s = 2$, que es el que presenta mayor interes desde un punto de vista practico.

Dado $X \subset R$ / $\#(X) = n$, $n > 2$, sea $X^2 = X \times X$, sobre el que se desea analizar el comportamiento de sus

elementos respecto del criterio C que determina la relacion

$(x_i,x_j)\ C_{X^2}\ (x_h,x_k)$ sii el recorrido de $X - \{ x_i, x_j \}$ es mayor o igual que el de $X - \{ x_h, x_k \}$.

Por tanto , la relacion puede expresarse mediante :

$$(x_i,x_j)\ C_{X^2}\ (x_h,x_k) \quad <==> \quad R_{i,j} \geq R_{h,k}$$

siendo $R_{i,j} = \underset{h \neq i,j}{\text{Max}} \{ x_h \} - \underset{h \neq i,j}{\text{Min}} \{ x_h \}$

Al igual que el caso estudiado en el epigrafe 3.5.1, esta relacion generaliza a la estudiada en el epigrafe 3.3

Ademas , si se denota por $R_{(i),(j)}$ al recorrido $X - \{ x_{(i)} , x_{(j)} \}$ se obtiene que :

$$R_{(i),(j)} = R_{(i)} \qquad \text{si} \quad i = j$$

$$R_{(i),(j)} = x_{(n)} - x_{(1)} \qquad \text{si} \quad 1 < i,j < n$$

$$R_{(1),(j)} = \begin{cases} x_{(n)} - x_{(3)} & \text{si} \quad j = 2 \\ x_{(n)} - x_{(2)} & \text{si} \quad j = 3,..,n-1 \\ x_{(n-1)} - x_{(2)} & \text{si} \quad j = n \end{cases}$$

$$R_{(i),(n)} = \begin{cases} x_{(n-1)} - x_{(1)} & \text{si} \quad i = 2,...,n-2 \\ x_{(n-2)} - x_{(1)} & \text{si} \quad i = n-1 \end{cases}$$

Para los conjuntos C_{X^2}-dominantes y C_{X^2}-dominados se verifica :

$$(x_{(i)},x_{(j)}) \in DC_{X^2}(x_{(h)},x_{(k)}) \qquad 1 < i,j < n \ ; \ \forall\ h,k$$

$$dC_{X^2}(x_{(i)},x_{(j)}) = X \qquad 1 < i,j < n$$

y el elemento $\tilde{x}_\star$, (C_{X^2} d M) de X^2 , cumple :

$$R_{\tilde{x}_\star} = \underset{i,j}{\text{Min}} \{ R_{i,j}\} = \text{Min } \{R_{(1),(2)},R_{(n),(n-1)},R_{(1),(n)}\}$$

En esta situacion cabe la posibilidad de que el elemento (C_X d M) de X , para la relacion estudiada en el epigrafe 3.3, no forme parte del par (C_{X^2} d M) de X^2 .

Las funciones de C-similitud y C-similitud signada naturales ,se definen de la forma usual , y el funcional a optimizar , como generalizacion del dado en epigrafe 3.3 , es :

$$\sum_{(x_i,x_j)\in X^2} \{[IC_{X^2}(x_i,x_j)]^2[R - R_{i,j}]+[1-IC_{X^2}(x_i,x_j)]^2R_{i,j}\}$$

sujeto a la restriccion (3.4) .

El optimo de esta funcion es invariante por traslacion y escala , y viene dado por :

$$IC^\star_{X^2}(x_i,x_j) = \frac{R_{i,j}}{R}$$

el cual puede interpretarse , al igual que en los casos anteriores , como un conjunto difuso sobre X^2 .

OUTLIERS

Pueden considerarse como casos particulares de este criterio los resultados obtenidos en el epigrafe 3.3.1 , y ademas :

(I) Los estadisticos propuestos por Dixon (1950)

$$E_1 = \frac{x_{(3)} - x_{(1)}}{x_{(n)} - x_{(1)}}$$

$$E_2 = \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(1)}}$$

$$E_3 = \frac{x_{(n-1)} - x_{(2)}}{x_{(n)} - x_{(1)}}$$

para testar la presencia de dos outliers inferiores, dos outliers superiores y uno superior y otro inferior respectivamente , en una muestra extraida de una poblacion exponencial de origen desconocido , se obtienen a partir de las expresiones

$$1 - IC^{*}_{X^2}(x_{(1)}, x_{(2)})$$

$$1 - IC^{*}_{X^2}(x_{(n-1)}, x_{(2)})$$

$$1 - IC^{*}_{X^2}(x_{(1)}, x_{(n)})$$

(IV) En el caso en que se hubiese realizado el estudio de forma generica sobre X^s , se obtiene , como caso particular :

$$\frac{x_{(n-q)} - x_{(p+1)}}{x_{(n)} - x_{(1)}}$$

que es el valor correspondiente a

$$IC^{*}_{X^{s}}(..,x_{(1)},..,x_{(p)},x_{(n-q+1)},..,x_{(n)},..)$$

y coincide con el estadistico propuesto por Dixon (1950, 1951), para testar la presencia de q-outliers superiores y p-inferiores en la situacion indicada previamente .

3.5.3 GENERALIZACION DEL CRITERIO NATURAL

La generalizacion de este criterio , para el caso en que se se desee estudiar el comportamiento conjunto de una serie de observaciones ,se realizaria de forma similar a las anteriores . Se han de resaltar los casos particulares :

$$E_1 = \frac{x_{(1)} + \ldots + x_{(k)} - k\,a}{\sum_{i=1}^{n} (x_i - a)}$$

$$E_2 = \frac{x_{(n-k+1)} + \ldots + x_{(n)} - k\,a}{\sum_{i=1}^{n} (x_i - a)}$$

que coinciden con los estadisticos tratados en los trabajos de Fieller (1976),Lewis y Fieller (1978),para testar la presencia de k-outliers inferiores ó superiores , respectivamente en una muestra extraida de una poblacion gamma de origen $\underline{a}$ conocido .

3.6 CRITERIO DE DISPERSION CENTRAL MULTIVARIANTE

Dado $X \subset R^k$ / $\#(X) = n$, con $ran\ (X) = k < n$,

$$X = \begin{pmatrix} x_{11} \cdot\cdot & x_{1j} \cdot\cdot & x_{1n} \\ x_{k1} \cdot\cdot & x_{kj} \cdot\cdot & x_{kn} \end{pmatrix} = (x_1 \cdot\cdot x_j \cdot\cdot x_n)$$

las matrices de suma de cuadrados y suma de productos asociadas a X y $X_{(i)} = (x_1,..,x_{i-1},x_{i+1},..x_n)$, que se representaran por $S_{(n)}$ y $S^i_{(n-1)}$, respectivamente , vienen dadas por

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

$$S^i_{(n-1)} = X_{(i)}(I_{n-1} - [1/(n-1)]E_{n-1\ n-1})\ X'_{(i)}$$

donde por I_h se denota a la matriz identidad de orden h y por E_{ab} , a la matriz de dimensiones axb , con todos sus elementos iguales a la unidad .

En esta situacion , se considera la relacion binaria definida sobre X , por :

$$x_i\ C_X\ x_j \iff \det (S^i_{(n-1)}) \geq \det (S^j_{(n-1)}) \qquad (3.9)$$

Esta relacion ,para el caso particular k = 1 , coincide con la estudiada en el epigrafe 3.2 por lo que la relacion dada por (3.9) puede considerarse como una generalizacion del Criterio de Dispersion Central analizado en el epigrafe 3.2.

Los conjuntos C_X-dominantes y C_X-dominados , se construyen de la forma usual y evidentemente , el criterio que da lugar a C_X es T-evaluable sobre X .

Las funciones de C-similitud y C-similitud signada naturales son:

$$SC_X(x_i,x_j) = \left| \det (S^i_{(n-1)}) - \det (S^j_{(n-1)}) \right|$$

$$SSC_X(x_i,x_j) = \det (S^i_{(n-1)}) - \det (S^j_{(n-1)})$$

La seleccion del $IC^{\star}_X$, se realizara mediante la minimizacion de :

$$[IC_X(x_i)]^2[\det(S_{(n)})-\det(S^i_{(n-1)})]+[1-IC_X(x_i)]^2 \det(S^i_{(n-1)})$$

bajo la restriccion

$$(x_j,x_k)\ C'_X\ (x_l,x_m) \iff IC_X(x_j)-IC_X(x_k) \leq IC_X(x_l)-IC_X(x_m)$$

El optimo es invariante por localizacion y escala , y viene dado por :

$$IC^{\star}_X(x_j) = \frac{\det (S^j_{(n-1)})}{\det (S_{(n)})}$$

Si se representa por $\bar{x} = (1/n)\ X\ E_{n\ 1}$, se obtiene

$$IC^{\star}_X(x_j) = \frac{\det (S_{(n)} - \frac{n}{n-1} (x_j - \bar{x}) (x_j - \bar{x})')}{\det (S_{(n)})} = \ldots$$

$$\ldots = 1 - \frac{n}{n-1} (x_j - \bar{x})' S^{-1}_{(n)} (x_j - \bar{x})$$

e $IC^{\star}_{X}$ puede interpretarse como un conjunto difuso sobre X , definido a traves del criterio C .

OUTLIERS

(I) Wilks (1963) , considera

$$\min_{j=1,..,n} \frac{\det (S^{j}_{(n-1)})}{\det (S_{(n)})}$$

(al que denomina "one-outlier scatter ratio") , como estadistico para testar la presencia de un outlier en una muestra extraida de una poblacion normal multivariante de parametros desconocidos.

(II) La generalizacion llevada a cabo en el epigrafe 3.5, puede realizarse tambien en este caso obteniendose :

$$IC^{\star}_{X^{s}}(x_{j1},...,x_{js}) = \frac{\det (S^{j1,..,js}_{(n-s)})}{\det (S_{(n)})}$$

y

Min $IC^{\star}_{X^{s}}(x_{j1},...,x_{js})$ coincide con el estadistico propuesto por Wilks (1963) , para testar en bloque s observaciones outliers en una muestra extraida en las mismas condiciones de (I) .

Siguiendo el procedimiento indicado , se podrian obtener algunos otros estadisticos utilizados en distintas situaciones particulares para la identificacion de outliers.

De la definicion dada para el termino outlier en el Capitulo I ,se deduce que el analisis de las observaciones outliers , puede realizarse mediante el estudio del comportamiento de los distintos elementos de la masa de datos respecto del criterio que se desee analizar en los mismos.

El estudio llevado a cabo en el Capitulo II , posibilita que en el Capitulo III , se obtengan algunos de los estadisticos mas utlizados para la identificacion de outliers.

Algunas de las ventajas que presenta el procedimiento que se propone frente a las tecnicas clasicas de identificacion de outliers son las siguientes :

(I) Los estadisticos se obtienen a traves de un metodo formal.

(II) No se obtienen bajo hipotesis de modelo poblacional alguno.

(III) La consideracion de las posibles observaciones outliers carece de cualquier tipo de subjetividad inicial.

Por ultimo , cabe destacar , que el metodo propuesto depende del criterio que se considere y este debera estar en relacion con el analisis estadistico posterior que se desee realizar con el conjunto de observaciones experimentales.

BIBLIOGRAFIA

Anscombe , F.J. (1960) "Rejection of outliers" .Technometrics,2, 123-147 .

Barnett , V. y Lewis , T. (1984) "Outliers in Statistical Data".2nd Edition. Ed. John Wiley and Sons.

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

Boscovich ,R.J. (1757) "De litteraria expeditione per pontificiam ditionem,et synopsis amplioris operis, ac habentur plura ejus ex exemplaria etiam sensorum impressa". Bononiensi Scientiarum et Artum Instuto Atque Academia Commentarii, 4, 353-396.

Chatfield , C. (1985) "The initial examination of data". Journal Royal Statistical Society . Ser. A, 148, 214-253.

Cochran , W.G. (1941) "The distribution of the largest of a set of variances as a fraction of their total" . Ann. Eugen.,11,47-52.

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

David , H. A. y Paulson , A.S. (1965) "The performance of several tests for outliers" . Biometrika ,52, 429-436.

Dixon , W.J. (1950) "Analysis of extreme values" . Ann. Math. Statis. ,21, 488-506.

Dixon , W.J. (1951) "Ratios involving estreme values" . Ann. Math. Statis. ,22, 68-78.

Dunn , J.C.(1974) "A Fuzzy Relative of the ISODATA Process and its use in detecting Compact Well-Separated Clusters".Journal of Cybernetics,3, 32-57.

Ferguson , T.S. (1961a) "On the rejection of outliers". Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability ,Vol.1, 253-287.

Ferguson , T.S. (1961b) "Rules for rejection of outliers". Rev. Inst. Int. de Statist., 29, 29-43.

Fieller , N.R.J. (1976) "Some Problems related to the Rejection of Outlying Observations".Ph.D.Thesis, University of Sheffield.

Fisher , R.A. (1929) "Tests of significance in harmonic analysis" . Proc. Roy. Soc. A, 125, 54-59.

Gnanadesikan , R. y Kettenring, J.R. (1972) "Robust estimates, residuals and outlier detection with multi response data" . Biometrics ,28, 81-124.

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

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

Hampel F.R. (1968) "Contributions to the Theory of Robust Estimation". Ph.D. dissertation, University of California-Berkeley.

Hampel,R.R.; Ronchetti,E.M.; Rousseeuw,P.J.; Stahel,W.A. (1986) "Robust Statistics.The approach based on influence functions" . Ed. John Wiley and Sons.

Hawkins , D.M. (1978) "Analysis of three tests for one or two outliers". Statistica Neerlandica, 32 , 137-148.

Hawkins , D.M. (1980) "Identificacion of outliers" . Ed. Chapman y Hall.

Kabe ,D.G. (1970) "Testing outliers from an exponential population" . Metrika ,15, 15-18.

Kale , B.K. (1976) "Detection of outliers". Sankhya B , 38, 356-363 .

Kendall ,M.G. y Buckland , W.R. (1957) "A dictionary of Statistical Terms" . Longman.

King ,E.P. (1953) "On some procedures for the rejection of suspected data" . J. Amer. Statist. Ass. ,48, 531-533.

Krantz , D.H. ;Luce , R.D. ;Suppes ,P. and Tversky , A. (1971) "Foundations of Measurement", Vol I , Ed. Academic Press.

Lewis , T. y Fieller , N.R.J. (1979) "A recursive algorithm for null distributions for outliers : I. Gamma samples" . Technometrics ,21, 371-376.

McMillan, R.G. (1971) "Tests for one or two outliers in normal samples with known variance".Technometrics, 13, 75-85.

Miller, R.G. Jr. (1981) "Simultaneous Statistical Inference". 2nd. Edition. Ed. Springer Verlag.

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

Rosado,F.M.F. (1984) "Existencia e Deteccao de outliers. Uma Abordagem Metodologica" . Ph.D. Thesis. Universidade de Lisboa.

Tietjen ,G.L. y Moore , R.H. (1972) "Some Grubbs-type statistics for the detection of several outliers". Technometrics ,14, 583-597.

Tucker,A.W. (1956). "Dual Systems of homogeneous linear relations". IN H.W. Kuhn and A.W. Tucker (Eds.), Linear inequalities and relates systems, Annals

of Mathematics Study 38. Princenton University Press, Princenton, New Jersey.

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. Juan L. Moreno Redollo titulada "Análisis Cualitativo de dato estadístico"

acordó otorgarle la calificación de APTO CUM LAUDE

Sevilla, 8 de Julio 1.987

El Vocal, El Vocal, El Vocal,

El Presidente El Secretario, El Doctorado,