MATEMÁTICAS Y ESTADÍSTICA

LCT

ANÁLISIS EXPLORATORIO DE DATOS: NUEVAS TÉCNICAS ESTADÍSTICAS

M. FREIXA I BLANXART
L. SALAFRANCA I COSIALLS
J. GUÀRDIA I OLMOS
R. FERRER I PUIG
J. TURBANY I OSET

PPU

ANÁLISIS EXPLORATORIO DE DATOS: NUEVAS TÉCNICAS ESTADÍSTICAS

MONTSERRAT FREIXA I BLANXART
LLUIS SALAFRANCA I COSIALLS
JOAN GUÀRIDIA I OLMOS
RAMON FERRER I PUIG
JAUME TURBANY I OSET
Profesores de Estadística del Departamento de
Metodología de las Ciencias del Comportamiento
de la Universidad de Barcelona

ANÁLISIS EXPLORATORIO DE DATOS: NUEVAS TÉCNICAS ESTADÍSTICAS

LCT-36

PPU
Barcelona, 1992

Colección: LCT-36
(Letras, Ciencias y Técnica)

Primera edición, 1992

Promociones y Publicaciones Universitarias, S.A.
Marqués del Campo Sagrado, 16. 08015 Barcelona
Tfno. (93) 442-03-91 Fax. (93) 442-14-01

I.S.B.N.: 84-7665-179-1
D.L.: L-424-1992

Imprime: Poblagràfic, S.A. Av. Estació, s/n. La Pobla de Segur (Lleida)

Nuestro agradecimiento a Josep María Domènech i Massons, Catedrático de Estadística de la Universidad Autónoma de Barcelona y a Joan Manel Batista i Foguet, Profesor Titular de Estadística de la Universidad de Barcelona por habernos introducido en el tema

ÍNDICE

PRESENTATION

La philosophie de l'analyse exploratoire des données développée par John W. Tukey a profondément marquée le monde de l'analyse des données de ces dernières décennies. Ce n'est qu'avec la publication de l'ouvrage phare *Exploratory Data Analysis EDA* en 1977, qu'un public plus large a pris conscience de l'importance de cette approche de l'analyse de données, une approche qui s'intéresse à la pratique des outils statistiques mis au service des utilisateurs. Aujourd'hui les travaux de Tukey et de son école ont trouvé un très large écho. En particulier l'approche exploratoire s'est avérée être une approche idéale pour enseigner les concepts et les outils statistiques. Pour s'en convaincre il suffit d'examiner la loste des ouvrages introductifs à la statistique qui s'appuient sur l'exploration.[1] En même temps l'exploration a engendré toute une série de logiciels qui s'appuient sur l'exploration (S,DataDesk,EDA) et toute une série d'outils exploratoires ont été introduits dans les logiciels plus conventionnels (SPSS,SAS etc.).[2]

Bien que certains ont tendance à reduire la contribution de Tukey à quelques outils qualifiés d'exploratoires (boxplot et stemleaf), l'oeuvre de Tukey doit être compris d'abord et avant tout comme une philosophie de l'exploration, une critique constructive du développement de la pratique statistique, une statistique qui *a appris à confirmer de façon très*

1. Par example les ouvrages de Erickson (1977), Marsh (1988) ou Hartwig (1978).
2. Le langage statistique S a été développé au laboratoire BELL; DataDesk est un logiciel pour MacIntosh développé par Velleman; le logiciel EDA a été développé à l'Université de Genève.

précise dans des circonstances hautement spécifiques,[3] mais qui a commencé à négliger les outils descriptifs; une statistique qui s'est fortement formalisée, sans toujours garder le constact étroit indispensable avec les applications et les utilisateurs.

Déjà en 1971, les psychologues Cooley et Lohnes, dans la préface de leur *Multivariate Data Analysis (1971)* remercient Tukey de leur avoir donné une identité professionnelle comme analystes de données[4] et une préoccupation propre, à l'écoute des problèmes de la pratique de leur discipline, au lieu de se sentir entre toutes les chaises, ni statisticiens, ni mathématiciens, ni informaticiens, et marginaux dans leur discipline, car peu ou pas compris par leur collègues vu leur spécialisation méthodologique.

L'exploration s'insère dans une tradition des analyse de données (data analysts) et s'appuie en même temps sur un grand nombre de recherchers empiriques et de simulations quant à la performance des statistiques classiques.[5] Ces études engendrent un corpus de connaissances concernant les propriétés des outils connus ou moins bien connus et stimulent des recherches pour trouver de nouveaux outils ayant les qualités requises pour les utilisateurs et activant leur créativité.

L'approche exploratoire peut être caractérisée par les traits suivants:

- Une *attitude* ouverte à l'égard des outils qui ne sont pas pris comme des recettes toutes faites, mais comme des propositions susceptibles de modification et d'adaptation en fonction du problème à traiter.
- Une *approche* orientée vers la compréhension, l'observation d'indices (pas nécessairement attendus), à la place d'une approche orientée exclusivement vers la conclusion et la décision.
- L'exploration est un travail de détective (*numerical detective work*) o il y a *interaction* constante entre les données et les connaissances de l'utilisateurs (*intégration des connaissances*)..

3. *Then they (les statisciens) learned to confirm exactly, to confirm a few things exactly, each under very specific circunstances.* Tukey 1977, p. 3.

4. Ils se réfèrect à *The Future of Data Analysis* (1961), où Tukey vante les vertus d'une orientation pratique de la statistique et appelle les praticiens à joindre leurs efforts pour rendre les outils statistiques plus performants dans une multitude de situations de recherche différentes.

5. Par example la fameuse Princeton Robustness Study (Andrews, Bickel, Hampel, Huber, Rogers, Tukey 1972) est une étude de simulation d'un grand ensemble d'estimateurs du centre d'une distribution.

- Les outils statistiques reposent sur toute une série d'hypothèses qui dans la pratique ne se vérifient pas toujours et qui par rapport à bien des outils ne sont pas vérifiables. Il s'agit donc de s'en rendre compte (*diagnostic*) et d'expliciter ces hypothèses, ainsi que de développer des outils appropriés à des situations o ces hypothèses ne se vérifient pas (*résistance et robustesse*).
- Les outils statistiques ne sont pas figés, il faut les adapter aux besoins de la recherche (*fléxibilité*). Chaque utilisateur est en mesure de créer —à son niveau— ses propes outils, pourvu qu'il soit prêt a réfléchir sur son problème et qu'il ait une vision critique et constructive à l'egard de ces outils (*créativité*).
- La nature du processus de recherche est par définition itérative (*itérativité*). Il s'agit de présenter les données de façon à faire apparaître une structure qui peut être interprétée en fonction de l'objet de l'analyse et de ce que l'on sait déjà des données en question. Mais comme toute structure est définie par rapport à des critères statistiques (méthode) et extra-statistiques (substance), plusieurs structures sont possibles, dont il s'agira de déterminer *l'adequation* à l'objectif poursuivi. Toute structure interprétée simplifie, réduit; il est donc essentiel d'indiquer en même temps ce qui n'est pas décrit par la structure, pour se concentrer dans un deuxième temps sur les *résidus* de la première description. Comme une description des résidus n'est pas toujours complète, on s'intéressera aux résidus de cette description de résidus, et ainsi de suite.
- Une orientation très marquée vers les outils *graphiques*, forçant le chercheur à voir aussi bien la structure que les déviations et les exceptions par rapport à cette structure (*révélation et interaction*). La recherche d'une description simple par transformations appropriée de variables (*Réexpression*) tout en se donnant les moyens de *diagnostiquer* les défauts.
- Une approche qui privilégie les outils et les explications multiples (*multiplicité*).

Cet ouvrage aborde six thèmes spécifiques de l'exploration. Le premier chapitre s'intéresse aux outils descriptifs de base, tels que les boîtes à pattes (Boxplots) et les branchages (Stemleaf). Ces outils, aujourd'hui

largement connus (par les biais des logiciels statistiques et les ouvrages modernes sur la statistique descriptive), illustrent parfaitement l'esprit des outils exploratoires: l'importance des graphiques, l'intérêt porté aux valeurs extrêmes, les mérites de la clarté conceptuelle qui produit des outils à la fois simples et puissants, l'accent mis sur les outils complémentaires et l'adaptabilité des outils à différents besoins.

Les transformations (chapitre 2), réexpressions dans la terminologie exploratoire, occupent une place clé dans la boîte à outils de l'analyste, car une bonne transformation permet souvent de mieux décrire et analyser une variable ou une relation entre variables. L'exploration offre dans ce domaine un cadre cohérent permettant de diagnostiquer des problêmes tels que l'absence de symetrie ou les relations non-linéaires et d'y remédier, afin de rendre les données accessibles à l'analyse à l'aide d'outils exploratoires ou plus conventionnels qui exigent la symétrie ou la forme linéaire de la relation.

La droite résistance, un autre exemple d'un outil exploratoire par excellence, met en évidence le caractère problématique de la droite usuelle obtenue par moindres carrés (sensibilité aux valeurs extrêmes, etc) tout en proposant une alternative. L'accent mis sur l'analyse des résidus illustre bien la démarche exploratoire générale o il s'agit essentiellement de décomposer les données en une partie structurée (ici la description par la droite) et une partie non structurée (ici les résidus par rapport a la droite) pour ensuite étudier les résidus afin de les décrire et structurer, donc de produire de nouveaux résidus qui à leur tour peuvent être décrits et engendrent à leur tour des résidus... Cette démarche met en évidance très clairement ce caractère interactif et itérative de l'exploration, une démarche à constraster avec le caractère plutôt linéaire de l'approche classique.

Cette démarche itérative est également à la base des méthodes expliquées dans les deux chapitres suivants, à savoir le polissage de tableaux par médiane (median polish) et le lissage des série (chronologiques) par médianes mobiles.

Enfin le dernier chapitre étudie les prolongements que trouvent les questions soulevées par les outils simples par rapport aux méthodes statistiques classiques. Le thème de l'estimation robuste esquisse une voie qui

permet de trouver une solution aux hypothèses souvent trop contraignantes et utopiques qui sont à la base des méthodes classiques.

Vu l'intérêt que présente l'exploration pour l'utilisateur, l'enseignement et l'étudiant des outils statistiques, on ne peut que se réjouir de l'apparition de cette ouvrage en langue espagnole. Je suis certain que les idées et concepts présentées dans le présent ouvrage stimuleront l'intérêt pour les outils exploratoires et robustes et serviront de point de départ d'une réflexion critique sur les outils statistiques.

Genève, en février 1992
Eugène Horber
Faculté des Sciences Economiques et Sociales
Département de Science Politique
Université de Genève

Bibliographie

Andrews, D.F.; Bickel, F.R.; Hampel, Huber, Rogers, Tukey, J.W. (1972). *Robust Estimates of Locations, Survey and Advances.* Princeton: Princeton UP.

Cooley, W.W. & Lohnes, P.R. (1971). *Multivariate Data Analysis.* New York: Wiley.

Erickson, B. & Nosanchuck, T.A. (1977). *Understanding Data: An introduction to exploratory and confirmatory data analysis for students in the Social Sciences.* Milton Keynes: Open University Press.

Hartwig, F. & Dearing, B.E. (1979). *Exploratory Data Analysis.* Berverly Hills: Sage.

Marsh, C. (1988). *Exploring Data. An introduction to Data Analysis for Social Scientist.* Oxford: Polity Press.

Mosteller, F. & Tukey, J. (1977). *Data Analysis and Regression. A second course in statistics.* New York: Addison-Wesley.

Tukey, J.W. (1962). The Future of Data Analysis. *Annals of Mathematical Statistics*, 33.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Reading Mass: Addison & Wesley.

Velleman, P.F. & Hoaglin, D.C. (1981). *ABC of EDA: Applications, Basics and Computing of Exploratory Data Analysis.* Boston: Duxbury Press.

PRESENTACIÓN

Una de las innovaciones estadísticas que ha causado un mayor impacto dentro del ámbito de las Ciencias sociales y del comportamiento ha sido, sin duda alguna, el reciente desarrollo del Análisis Exploratorio de los Datos y, es notorio constatar que, en la actualidad, existe una creciente utilización y demanda de esta nueva modalidad de análisis.

El impulso inicial de este conjunto de técnicas ingeniosas se debe al esfuerzo de John Tukey quien, en su ya clásico trabajo *de 1977 Exploratory Data Analysis*, expuso los principios fundamentales. De hecho, el Análisis Exploratorio de Datos se inspira en una filosofía de carácter práctico según la cual, contrariamente a los enfoques de corte clásico, los datos son los que en última instancia guían la selección de modelos adecuados al tiempo que se minimiza la asunción de presupuestos previos. De acuerdo, por tanto, con este nuevo enfoque, el analista trata de desvelar los patrones y las estructuras que se subyacen a los datos, sin que por ello sea necesario asumir un conjunto de postulados previamente definidos y altamente restrictivos.

De forma resumida, destacaremos que los principales aspectos que caracterizan este nuevo enfoque analítico son fundamentalmente cuatro. En primer lugar, mediante las representaciones visuales es fácil descubrir el modo de comportarse de los datos, así como las posibles estructuras que presentan. Esta constituye, a nuestro entender, una de las más importantes

innovaciones que dicha técnica aporta, en el sentido que exige un constante uso de visualizaciones gráficas. En segundo lugar, estas métodos requieren que la atención del analista se centre en los residuales o lo que queda después de haber aplicado algún tipo de análisis. De ahí el protagonismo que adquieren los residuales, como procediemientos para detectar estructuras o patrones que por lo general pasan desapercibidos con los análisis más convencionales. Se tiene, en tercer lugar, que mediante transformaciones matemáticas simples, como por ejemplo, el logaritmo y la raíz cuadrada, los análisis no sólo se simplifican sino que adquieren una mayor claridad. Por último, el carácter resistente, propio de estos métodos, garantizan el hecho de que valores de datos extraños o poco corrientes no influyan indebidamente los resultados de un análisis (Velleman y Hoaglin, 1981).

Sin duda alguna, la principal novedad del Análisis Exploratorio de Datos es la especial forma de presentar el conjunto de datos (*batch*) en representaciones visuales o gráficos como, por ejemplo, el de tronco y hoja (*stem-and-leaf*). Este tipo de representaciones no sólo permiten descubrir los patrones específicos o de tendencia, sino que a su vez, ponen de manifiesto aspectos sosprendentes, insospechados, y a veces divertidos, que de otra forma pasarían totalmente inadvertidos. Junto a estos procedimientos de representación visual deben también destacarse, por su utilidad en la descripción más precisa de las formas o patrones que toman los datos, los gráficos de valores de letra así como sistemas intermedios representacionales o los diagramas de cajas (boxplots). El lector encontrará en este texto, un detallado y preciso análisis de estos sistemas representacionales que permiten detectar, de forma simple y rápida, el conjunto de índices descriptivos necesarios para una correcta comprensión de la estructura de los datos.

No es menos importante destacar, dentro de esta nueva filosofía de análisis, el método de *línea resistente* propia de aquellas situaciones donde los valores de una variable de respuestas son representados contra los valores de un factor explicativo. Paralelamente a los sistemas de ajuste lineal, dentro del ámbito de la estadística clásica, el Análisis Exploratorio de Datos plantea un nuevo procedimiento de ajuste basado en la línea resistente que, a diferencia del método clásico de la regresión mínima cuadrada, protege el modelo contra aquellos casos atípicos o raros capaces de distorsionar el análisis.

Ahora bien, hemos de destacar, en relación a este nuevo enfoque analítico, que contariamente a los análisis más tradicionales que empiezan con la propuesta de un modelo o estructura para la descripción de los datos, el Análisis Exploratorio de los Datos permite extraer dichas estructuras a partir de las representaciones visuales o gráficas. Es decir, es un procedimiento que posee más bien un carácter inductivo, en virtud del cual el ajuste se lleva a cabo despúes del conocimiento previo de su estructura. Como es obvio, al igual que cualquier otra clase de análisis se pretende conseguir una adecuada descripción de los datos mediante un ajuste. De este modo los residuales constituyen la diferencia entre los datos observados y los valores ajustados. Ahora bien, los residuales no son considerados, desde este perspectiva, como simples errores, sino como pistas o indicios del proceso de los datos que se halla escondido detrás de los patrones específicos. El tratamiento estadístico del ajuste de modelos de línea resistente, de una dificultad matemática intermedia, están abordados excelentemente en dicho texto tanto por la claridad expositiva de los conceptos como por la adecuación de los ejemplos propuestos.

La extensión de los modelos de ajuste, cuando los patrones no son lineales, requieren técnicas más sofisticadas conocidas por el nombre genérico de suavizado. Sobre todo en aquellas situaciones donde los valores de la variable independiente, x, se pretenan en intervalos regulares e igualmente espaciados. En estos casos, la atención del analista recae primordialmente sobre la variable de respuesta o variable y. Un ejemplo típico son los las secuencias de datos que suelen generarse en función de la variable tiempo como por ejemplo, la tasa de desempleo, cantidad de accidentes, etc., que suelen registrarse en términos de puntos temporales de intervalos constantes. Cuandos los datos son registrados para cada uno de los puntos sucesivos del tiempo, estas clases de registros suelen recibir el nombre de series temporales. Sin que por ello, se requiere necesariamente la dimensión temporales para la obtención de datos de secuencia, la técnica del suavizado permite desvelar la estructura de tendencia de un conjunto de datos, en términos de curvas no lineales. Evidentemente el ámbito de las técnica de suavizado o alisados, constituye uno de los campos más prometedores dentro del ámbito de las ciencias comportamentales, y he de admitir, como se destaca adecuadamente en el texto, que la mayoría de las veces la solución mediante papel y lápiz es sumamente difícil. Por esta razón, es de destacar el esfuerzo realizado por lo autores del texto en

dar una visión exacta de la temática, así como las diferentes soluciones a fin de tener una conocimiento cabal del suavizado para el ajuste de procesos de datos no lineales.

Por último, como queda explícitamente recogido en este texto, mediante las tablas de datos de doble entrada el Análisis Exploratorio de Datos aporta un conjunto de técnicas resistentes y robustas, para el estudiar las relaciones entre dos o más variables. Es decir, aquellas situaciones en las que, análogamente al trabajo experimental, se analiza la relación entre factores cualitativos independientes y la variable de respuesta o dependiente. La ventaja de este conjunto de técnicas para el análisis de esquemas factoriales sigue siendo la misma: No es necesario asumir los presupuestos que se hallan implícitos en los modelos clásicos de Análisis de la Variancia. En efecto, mediante por ejemplo, el ajuste de medianas en posibe detectar el caracter aditivo o no, de los efectos factoriales.

El libro termina con un capítulo que recogen las propiedades de los estimadores estadísticos, planteando los métodos de estimación robusta y las técnicas de generación de nuevas muestras a partir de la original (Jackknife y Bootstrap).

En suma se trata de una manual, claro, sistemático y completo en el que se afronta por primer vez, dentro del ámbito castellano parlante, la problemática del Análisis Exploratorio de Datos y ofrece al lector la posibilidad de entrar en contacto con el alcance y la aplicación de este conjunto de métodos que se erigen como una alternativa, a tener en cuenta, a los procedimientos estadísticos de corte más clásico y tradicional. Espero que, gracias a esta notable aportación, el Análisis Exploratório de Datos consiga el eco y la resonancia que merece y, alcance una gran difusión dentro de nuestro contexto cultural.

Jaume Arnau i Gras
Catedrático de Metodología de las
Ciencias del Comportamiento
U. B.

INTRODUCCIÓN

"Exploratory data analysis is detective work—numerical detective work—or counting detective work—or graphical detective work"

Tukey, 1977 (pág. 1)

Tukey en su libro "Exploratory DATA ANALYSIS" (1977), E.D.A., desarrolla una serie de nuevas técnicas gráficas y analíticas para conseguir **un conocimiento previo de los datos a analizar siempre desde una perspectiva exploratoria.**

El análisis exploratorio de datos, y que ya puede considerarse una nueva rama de la Estadística, aunque David Cox (1978) (pág. 5) afirma que es sólo una extensión de la Estadística Descriptiva y Gráfica, propugna **un cambio de actitud y de enfoque metodológico ante el análisis de datos.**

El E.D.A. postula que previamente a cualquier análisis de datos es necesario un examen visual de estos. Antes de analizar los datos es preciso "mirárselos", "entenderlos" y "reflexionar" sobre ellos.

La Estadística Descriptiva clásica se ocupa en recoger, ordenar y representar los datos, normalmente en forma de tablas y agrupando los datos en intervalos para representarlos gráficamente y calcula estadísticos basados principalmente en la distancia y con datos centrados en la media.

El E.D.A. tiene los mismos objetivos, pero pretende además detectar anomalías o errores en las distribuciones univariantes de los datos. También intenta descubrir en los datos patrones o modelos. Para ello incorpora nuevas técnicas gráficas y busca estadísticos resistentes y robustos basados principalmente en el orden y centrados en la mediana.

Dado que casi todos los análisis estadísticos avanzados se basan en los primeros análisis es importante que estos esten bien hechos, y sean adecuados. Por ejemplo, si Vd. hace un análisis complejo que necesita como primeros estadísticos las medias de las variables y, alguna serie de datos tiene valores alejados, la media no será la medida de tendencia central más representativa de la distribución. O, si Vd. hace un análisis sofisticado que supone que la relación entre las variables es lineal y en realidad no lo es. O si, simplemente Vd. ha cometido errores al entrar los datos en el ordenador u otras tantas situaciones que se pueden dar Vd. extraerá conclusiones de los contrastes y de las inferencias que fallarán por no haber tomado precauciones en este primer análisis.

En efecto, la Estadística Clásica se ha viciado o anquilosado, por así decirlo, en dos aspectos:

a) Parte casi siempre de hipótesis gaussianas, muchas veces imposibles de verificar y presupone además que los errores y las fluctuaciones aleatorias de los valores empíricos se encuentran simétricamente repartidos alrededor de un valor central.

b) Por el uso casi exclusivo de modelos lineales cuando analiza relaciones entre las variables.

Todo ello se aplica además de forma que nos atreveriamos a calificar de rutinaria, sin detenerse en comprobar los supuestos o a plantearse su conveniencia o adecuación de aplicación a cada serie de datos.

Esto sería aún justificable en Ciencias muy avanzadas, en las que los modelos teóricos estuviesen ya muy estructurados y se conociera la normalidad de las variables, pero desafortunadamente en Ciencias Humanas, en donde la variabilidad de los datos y especialmente por el hecho de que los datos obtenidos son, en la mayoría de los casos difícilmente repetibles, hacen que estos presupuestos se cumplan con seguridad pocas veces. En consecuencia, el nuevo enfoque metodológico propugna enten-

der el análisis de datos "como un ciclo repetitivo en el que los modelos y los datos se suceden alternativamente hasta alcanzar el mejor ajuste" (Mallow y Tukey, 1982).

En la última década han aparecido varias publicaciones (Tukey, 1977; Hoaglin, Masteller y Tukey, 1983; Erickson y Nosanchuk, 1979; Velleman y Hoaglin, 1981 o Peña, 1986) que propugnan la nueva perspectiva, es decir que no sólo los modelos generan datos, sino que los datos deben ser utilizados en la generación de modelos para su análisis.

> *"La Metodología Estadística se concibe como un proceso iterativo de aprendizaje en lugar de concebirse como una aplicación única y directa de un determinado procedimiento óptimo"*
>
> Peña, 1986 (pág. 37)

Las técnicas exploratorias resistentes y robustas que Vd. encontrará en este libro le ayudarán a resolver todos estos problemas. Su sencillez y rápidez de cálculo las hacen sumamente útiles para explorar distribuciones univariantes y las posibles relaciones entre variables. Estas técnicas pueden tener muchas otras aplicaciones que el lector irá descubriendo al avanzar el libro. Sin embargo, para Harwig y Dearing (1979) el análisis exploratorio de datos es:

> *"Exploratory data analysis is interactive and iterative"*
>
> Hartwig y Dearing, 1979 (pag. 76)

Para Everitt y Dunn (1983) el E.D.A. constituye un paso preliminar al análisis de datos y hacen hincapié en que el E.D.A. es necesario en el análisis multivariante, en el que se tratan gran cantidad de datos por computadora, muchas veces sin tener en cuenta este análisis preliminar de datos. Por ejemplo, en distribuciones univariantes, el E.D.A. nos informará sobre:

a) La localización, la desviación y la forma de la distribución de los datos.
b) La simetría o asimetría de los mismos.

c) El número y localización de agujeros (vacíos) y puntas en la distribución.

d) Presencia y número de valores alejados.

El mismo E.D.A., a pesar de ser una teoría reciente, tiene ya sus detractores. Así Good (1983), profesor de Estadística del Instituto Politécnico de Virginia (E.E.U.U.), en un artículo que como su propio nombre indica tiene más de filosófico que de estadístico, "The philosophy of exploratory data analysis", indica que el E.D.A. es más un arte o incluso una suma de trucos que una ciencia (pág. 283).

No obstante, al final de su trabajo reconoce que los puntos importantes del E.D.A. son:

1) La presentación de los datos.
2) El reconocimiento del modelo.
3) La formulación de hipótesis.
4) Buscar otras hipótesis que sean más explicativas.
5) Racionalización del error de Tipo II.

De acuerdo con lo propuesto por Hoaglin, Mosteller y Tukey (1983) y de Welleman y Hoaglin (1981) reconocen la existencia de cinco componentes principales del E.D.A.:

1) Sus **representaciones gráficas** nos revelan visualmente el comportamiento de los datos y la estructura del conjunto.
2) Pone mucha atención al análisis de **residuales**, es decir en las diferencias que hay entre los datos reales y el resultado de su ajuste a un modelo previamente determinado o subyacente.
3) **Utiliza la Transformación de los datos**, que consiste en encontrar la escala que más simplifique y clarifique el análisis, como, por ejemplo, con el uso de funciones matemáticas simples como raiz cuadrada, logaritmos etc.
4) Valora **la resistencia**, propiedad que presentan algunos estadísticos que les hace poco sensibles a la influencia de uno o varios valores sensiblemente distantes de la mayoria de los valores de la distribución.

5) Busca estadísticos **robustos** propiedad que presentan algunos estadisticos que les hace poco sensibles a desviaciones de los supuestos básicos.

Con respecto a la estadística descriptiva clásica el E.D.A:

1) Cálcula índices descriptivos resistentes y robustos.
2) La descripción no se efectua en base a un solo estadístico sino emplea varios índices a la vez.
3) Preferencia por los resumenes visuales a los simplemente numéricos.

Cuando buscamos relaciones entre variables el E.D.A. es especialmente adecuado en Ciencias Sociales, Humanas y de la Salud donde los modelos sustantivos son complejos y las variables han sido medidas en todo tipo de escalas, nominal, ordinal, de intervalo y de razón y los datos están sujetos a gran variabilidad.

En Psicología los analisis mediante E.D.A. ayudan a descubrir tendencias, patrones de conducta, conductas diferenciales, formación de actitudes y evaluación del cambio. En Historia y Lingüística es útil para descubrir indicadores de cambio histórico o lingüístico. Los economistas, sociólogos y pedagogos deben emplear técnicas E.D.A. antes de confirmar sus complicados modelos. El análisis de datos mediante técnicas E.D.A. puede hacer revelaciones "esenciales" en la investigación en Medicina. En la empresa el E.D.A. aporta datos significativos sobre rendimiento y control de calidad. En general, al estadístico e investigador el E.D.A. le ayuda a hacer análisis confirmatorios válidos, evita hacer inferencias erróneas o mentir en Estadistica . Según Tukey el E.D.A. actua de protección al usuario estadístico.

El Análisis Exploratorio de Datos (detective) busca muchas pruebas al Analisis Confirmatorio (juez) a fin de que emita veredictos lo más fiables posibles. En análisis confirmatorios no significativos el E.D.A. da una información de gran valor para mantener, variar o cancelar alguna variable en la investigación o para reorientarla o , si ello es preciso, para generar nuevas hipótesis.

En Análisis Confirmatorios con resultados significativos el E.D.A. ayuda a conocer y hace que, averiguaciones sobre las causas, sean más rápidas

y eficaces. A veces el investigador no busca diferencias significativas, porque ya las conoce, sino que busca cambios, patrones, tendencias etc.

Cabe destacar, asimismo, que las técnicas E.D.A. no sólo constituyen un complemento a las técnicas estadísticas clásicas si no también una valiosa alternativa en caso de incumplimiento de alguna condición de aplicación, puesto que no son tan restrictivas en sus supuestos.

En realidad, el investigador necesita usar las técnicas estadísticas exploratórias y confirmatórias. Las técnicas exploratórias ayudan a comprobar las condiciones de aplicación de las pruebas de hipótesis, a detectar errores o valores anómalos, a buscar la mejor transformación cuando es necesaria etc. etc.. En general, dan una visión distinta, previa pero complementaria, a la confirmatoria. Todo ello repercute en una mejor calidad del análisis de datos globalmente entendido.

1. ORGANIZACIÓN, REDUCCIÓN Y REPRESENTACIÓN DE DATOS

1.1. INTRODUCCIÓN

Como se ha comentado en la Introducción general a la presente obra, el conjunto de técnicas agrupado bajo la denominación de **Análisis Exploratorio de Datos (EDA)** se ha desarrollado desde un punto de vista eminentemente pragmático, sobre todo en lo que se refiere a la reducción del número de condiciones previas para la aplicación de determinadas técnicas, fundamental y limitante en la Estadística que se venía desarrollando hasta el momento; con ello, la selección del modelo al que se intentarán ajustar los datos puede quedar a cargo de su propia configuración, que conducirá al analista de forma paulatina hasta el más idóneo en cada caso particular.

Por tanto, el primer paso propuesto por la perspectiva EDA, que se presupone asimismo en el proceso de análisis clásico, aunque la falta de énfasis en este punto ha llevado en muchas ocasiones a obviar, es la descripción organizada de los datos, con dos características sobresalientes: potenciación del uso de gráficos y "dicotomización" de los datos en centrales y extremos, todo ello desarrollado desde una perspectiva sumamente flexible en su aplicación, y a cuyo efecto se han desarrollado nuevos índices

descriptivos, algunos de los cuáles se incluyen ya actualmente entre los clásicos propuestos por los manuales de Estadística general.

Los usuarios de las técnicas descriptivas clásicas saben que en éstas priman los índices de tendencia central y de dispersión, considerándose como secundarios los de forma y derivados gráficos que pueden obtenerse. Se trata, por tanto, de un tipo de descripción específicamente adecuada a las distribuciones simétricas, en general, y que sigan el modelo de la distribución normal, en particular.

El enfoque del EDA añade una recuperación y potenciación de los índices de forma, y la utilización de los gráficos prácticamente como un índice más puesto que, como afirman, *una gráfica bien realizada puede ser más informativa que un conjunto de números.* Puede hablarse, por tanto, de un cierto **cambio de actitud en la descripción de los datos.**

En base a lo anterior, desarrollaremos este Capítulo en cuatro apartados, correspondientes a los índices descriptivos generados por la perspectiva EDA:

- **Localización** (*Location*), que se corresponderían a los índices de posición y tendencia central clásicos, indicando los valores límite y promedios de la distribución.
- **Dispersión** (*Spread*), para definir agrupación o disgregación en la distribución. Cuanto menor sea su valor, más información aportarán los índices de localización.
- **Forma** (*Shape*), para evaluar la situación de los datos desde ejes verticales (simetría) y horizontales (curtosis).
- **Gráficos**, que mostrarán las agrupaciones internas de los valores e indicarán los índices que mejor representan a la distribución entre los anteriores.

Observemos, para ejemplificarlo de forma muy simple, dos muestras simuladas:

Tabla 1.1

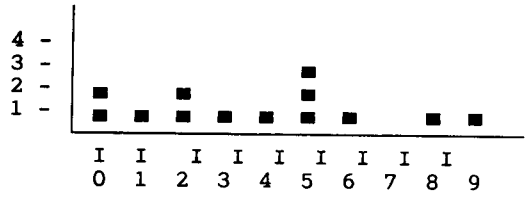
MUEST.1	MUEST.2
10	10
10	10
11	11
12	12
12	12
13	13
14	14
15	15
15	15
15	15
16	16
18	18
19	57

a simple vista se aprecia que la única diferencia entre ellas reside en los dos últimos datos. Si efectuamos un análisis mediante los índices clásicos podríamos obtener, entre otros, los siguientes estadísticos y gráficos:

Tabla 1.2

	MUESTRA 1	MUESTRA 2
n	13	13
$\overline{X}$	13.85	16.77
Md	14	14
Mo	15	15
S^2	8.14	151.86
Simetría	0.31	3.37
Curtosis	-0.688	11.77

MUESTRA 1



MUESTRA 2

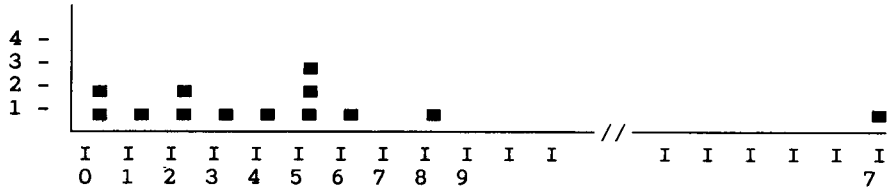


Figura 1.1

puede comprobarse que, si bien ambas distribuciones de datos son muy parecidas, los estadísticos y gráficos obtenidos no mantienen esta semejanza, debido a que casi todos ellos describen de forma óptima aquellas distribuciones que siguen el modelo de la curva normal, es decir: unimodales, simétricas y mesocúrticas.

En este caso, provocado por un solo valor muy alejado del resto, se han visto afectados los índices de tendencia central, dispersión y forma, y no así los de posición, si consideramos entre éstos a la Mediana y la Moda. El enfoque propuesto por el EDA pretende **obtener índices robustos y resistentes** que se vean poco afectados por estos elementos extraños.

Se considera a los índices como resistentes si muestran poca sensibilidad ante la presencia de valores anómalos (alejados del núcleo central de la distribución), entre los que el enfoque clásico tiene a la Mediana por ejemplo, en tanto que puede proponerse a la ampliamente utilizada

Media aritmética como ejemplo de índice poco resistente. Un estadístico resistente mostrará pocas variaciones ante la sustitución de los valores originales por otros muy diferentes, en base a su focalización en la parte central o relativamente agrupada de la distribución respecto de sus valores alejados.

Se denominan índices o estadísticos robustos a los que muestran poca sensibilidad ante las desviaciones de los supuestos inherentes a modelos probabilísticos (respecto a su forma, por ejemplo), por lo que ni la Mediana ni la Media aritmética pueden considerarse robustos, aunque sí lo serán los derivados de las Medias Recortadas que trataremos más adelante. Se trata, en definitiva, de diseñar técnicas cuyas hipótesis de aplicación sean poco restrictivas.

1.2. ÍNDICES DE LOCALIZACIÓN

Iniciaremos nuestra exposición con la descripción de índices encaminados a resumir los datos brutos obtenidos para su análisis. Los sumarios obtenidos son de gran utilidad no sólo para muestras grandes, como es obvio, sino para la descripción de muestras compuestas por un número relativamente pequeño de registros. No esperaremos obtener detalles de los mismos sino configuraciones generales que hagan más específica la tarea estadística posterior y nos encaminen hacia la descripción configurativa general del "*batch*", denominación propuesta por el EDA para la muestra.[1]

Debido a la prioridad que concede el enfoque EDA a la resistencia y robustez, los algoritmos de cálculo de estadísticos se realizan a partir de los centiles, como índices de posición poco afectados por los datos alejados del grupo central, y de entre ellos destacamos especialmente a

1. El término "muestra" es sustituido habitualmente por los creadores de estas técnicas por el de "lote" (del inglés "*batch*") (Erickson & Nosanchuk, 1979) que supone únicamente que los datos a analizar comparten una característica común, que los define como conjunto.

la mediana por su uso frecuente en la estadística descriptiva clásica. Con ellos trabajaremos en base tanto a su difundido uso en Estadística, como a la poca diferenciación que presentan respecto a la propia generación de "pseudo-centiles" realizada desde el enfoque EDA que, como más adelante se detallará, consiste básicamente en subdividir en partes iguales cada mitad de la muestra, y así sucesivamente hasta fraccionarlo en el número de valores que, en opinión del analista, describa el lote en toda extensión. Volviendo al ejemplo propuesto en el apartado anterior, los centiles en base a los que trabajaremos son:

Tabla 1.3

	MUESTRA 1	MUESTRA 2
C_{10}	10	10
C_{25}	11.5	11.5
C_{50}	14	14
C_{75}	15.5	15.5
C_{90}	18.6	41.4

en donde puede observarse que el único índice afectado es el centil 90, y aún en este caso, no de forma tan manifiesta como en la Tabla 1.2 de estadísticos clásicos.

Es importante reiterar que las muestras de datos a analizar no precisan la asunción de muchas de las condiciones previas que conllevan la mayoría de análisis diseñados por la estadística clásica. La delimitación de estos datos y sus promedios definitorios se obtendrán del cálculo de los siguientes índices:

1.2.1. PROMEDIO DE CUARTILES *(Q)*

Se calcula mediante:

$$\overline{Q} = \frac{C_{25} + C_{75}}{2} \tag{1.1}$$

que, como puede observarse, consiste en la suma promediada del primer y tercer cuartiles,[2] recogiendo por tanto el 50% central de los valores y revertiendo en el valor de la Md; en otras palabras, se elimina la influencia de los valores extremos, característica que observaremos también en la mayoría de los siguientes indicadores.

1.2.2. TRIMEDIA

Se define la Trimedia (Tri) como la distancia media entre la Md y $\overline{Q}$, es decir, eliminando un 25% de las observaciones de cada extremo y promediando los límites que las determinan respecto del valor central Md, que es ponderada por tanto desde dos puntos. En este sentido, pueden considerarse casos particulares parecidos a la Trimedia, denominados, en general, Medias Recortadas (*Trimmed Mean*), como muy bien indican Batista & Valls (1985a), las cuales serán objeto de nuestra atención en el capítulo sexto de este volúmen. En general, puede identificarse una determinada media recortada por la proporción de individuos que se excluye en su cálculo, punto que lleva a cuestionar, el uso de la media aritmética clásica en los intervalos de probabilidad o confianza, en los que podría resultar útil "recalcular" este índice, una vez delimitado el (1-α)% central de los datos. Se calcula mediante:

$$TRI = \frac{Md + \overline{Q}}{2} = \frac{C_{25} + (2 \cdot Md) + C_{75}}{4} \qquad (1.2)$$

1.2.3. CENTRIMEDIA O MEDIA INTERCUARTÍLICA (*Midmean, interquartile mean)*

Se calcula promediando todos los valores **entre** el primer y tercer cuartiles.

2. o del primer y tercer "cuarto", como se definirá más adelante.

$$MID = \frac{X_{iC25+1} + \cdots + X_{iC75-1}}{n_i} \tag{1.3}$$

Se trata por tanto de la media aritmética del 50% central de la distribución (dividimos la suma de sus valores entre el número de éstos). En el cálculo no deben incluirse los valores repetidos, y debe procurarse que el número de éstos a un lado y otro de la Md sean los mismos, es decir, ni debe ser un número impar (para conseguirlo puede optarse por incluir uno de los valores repetidos en el "lado" que presente un valor menos).[3]

En estos índices destaca el uso del 50% central de los datos, y en especial de la Md. Si los valores se hallan agrupados, el valor de ésta será muy semejante al de la Media aritmética clásica en tanto que ante la presencia de algunos valores muy alejados, la Md reflejará mejor el valor promedio del grupo. Este tipo de estrategias se utilizan, por ejemplo en muchos deportes de competición, como la gimnasia o el submarinismo, cuyas puntuaciones máxima y mínima otorgadas por los jueces son eliminadas al calcular el promedio de cada individuo, en previsión de posibles penalizaciones o sobrevaloraciones idiosincráticas.

1.3. ÍNDICES DE DISPERSIÓN

1.3.1. AMPLITUD INTER-CUARTÍLICA *(IQR) (interquartile range)*

También denominada dispersión media (*midspread*) o diferencia entre cuartiles (dq), se calcula mediante:

$$IQR = C_{75} - C_{25} \tag{1.4}$$

que proporciona a la vez resistencia, facilidad de cálculo, referencia a la

3. en el ejemplo que hemos presentado pueden eliminarse el primer y último valor (12 y 25).

mitad de la distribución y facilidad de interpretación, por lo que su uso es muy generalizado.[4]

1.3.2. MEDIANA DE LAS DESVIACIONES ABSOLUTAS *(MAD)*

Se calcula mediante:

$$MAD = Md|X_i - Md| \tag{1.5}$$

es decir, obteniendo la Md de las diferencias, en valor absoluto, respecto de la Md. El proceso de cálculo supone pues obtener, en primer lugar, el valor de la Md de la muestra y, a continuación, obtener las diferencias en valor absoluto de cada uno de los valores respecto de aquella. Una vez transformada así la distribución de los datos originales en distribución centrada respecto de la Md, se reordenan éstos y se obtiene su Md, que denominaremos MAD. En nuestro ejemplo inicial sería:

Tabla 1.4

							Md						
Muestra 1:	10	10	11	12	12	13	14	15	15	15	16	18	19
Muestra 2:	10	10	11	12	12	13	14	15	15	15	16	18	57
$\|X_i - Md\|$ 1:	4	4	3	2	2	1	0	1	1	1	2	4	5
$\|X_i - Md\|$ 2:	4	4	3	2	2	1	0	1	1	1	2	4	43
reordenamos:	0	1	1	1	1	2	2	2	3	4	4	4	5
reordenamos:	0	1	1	1	1	2	2	2	3	4	4	4	43
							MAD						

4. puede calcularse asimismo diferenciando los "cuartos", que se definirán más adelante, denominándose entonces *fourth-spread*. También se utilizará al comparar diversos lotes entre sí, al escoger una escala apropiada de medida (logarítmica, raíz cuadrada, etc.) de las observaciones.

1.3.3. ESTANDARIZACIÓN DE *IQR* Y *MAD*

A fin de aproximar los dos últimos estadísticos presentados a la distribución normal, pueden calcularse sus valores estandarizados, que se denominarán *pseudo-desviación estándar* de los respectivos valores, dado que si la distribución que estamos describiendo se ajustara a la curva normal, estos valores coincidirían con su S. Su cálculo se lleva a cabo mediante

$$PSD_{IQR} = \frac{IQR}{1.349} \tag{1.6}$$

$$PSD_{MAD} = \frac{MAD}{0.6745} \tag{1.7}$$

Obsérvese que 1.349 y 0.6745 son los valores, en puntuaciones Z, que delimitan el 50% central de la distribución normal. $(p(0.25) = Z(0.6745); (0.6745 \cdot 2) = 1.349)$, por tanto, con el valor de probabilidad 0.6745 estamos delimitando un 25% de la distribución hacia cada extremo, es decir, estamos trabajando con una dispersión de 1.349 σ.

1.3.4. COEFICIENTE DE VARIACIÓN CUARTÍLICO *(CVc)*

Se calcula mediante:

$$CVc = \frac{\frac{IQR}{2}}{\overline{Q}} = \frac{C_{75} - C_{25}}{C_{75} + C_{25}} \tag{1.8}$$

Se trata de un índice de dispersión relativa, que viene a sustituir al Coeficiente de Variación clásico, de escasa robustez, permitiendo asimismo comparar dispersiones entre distribuciones, independientemente de sus unidades de medida.

1.4. ÍNDICES DE FORMA

En ellos reside una de las principales aportaciones de la estrategia descriptiva EDA, siendo complementada su información por los índices gráficos que trataremos a continuación. Como ya hemos comentado, con distribuciones asimétricas o multimodales los índices clásicos son poco precisos, manifestándose principalmente esta carencia en los índices de forma, que aumentan su robustez mediante los siguientes algoritmos de cálculo:

1.4.1. ÍNDICE DE SIMETRÍA DE YULE

Se calcula mediante:

$$H_1 = \frac{C_{25} + C_{75} - (2 \cdot Md)}{2 \cdot Md} \tag{1.9}$$

que, de nuevo observamos, hace referencia a la simetría del 50% central de la distribución. Su interpretación se realiza como sigue:

- $*H_1 = 0 \longrightarrow$ Distribución simétrica
- $*H_1 > 0 \longrightarrow$ Asimetría positiva (sesgo o carencia de datos en la mitad superior de la distribución).
- $*H_1 < 0 \longrightarrow$ Asimetría negativa (sesgo o carencia de datos en la mitad inferior de la distribución).

1.4.2. ÍNDICE DE SIMETRÍA DE KELLY

Se calcula mediante:

$$H_2 = Md - \frac{C_{10} + C_{90}}{2} \tag{1.10}$$

que, como se puede observar, indica la simetría de la distribución en sus extremos o colas. Por ello puede ser de gran utilidad obtener ambos índices, H_1 y H_2, en base a la complementariedad de su información. Dado que el índice de Kelly depende de las unidades de medida (al emplear la Md), es conveniente transformarlo en el siguiente índice:

$$H_3 = \frac{C_{10} + C_{90} - (2 \cdot Md)}{2 \cdot Md} = \frac{-H_2}{Md} \tag{1.11}$$

interpretándose de esta forma al igual que se hace con el índice de Yule, en base a su adimensionalidad.

1.4.3. COEFICIENTE DE CURTOSIS

Se calcula mediante:

$$K_2 = \frac{C_{90} - C_{10}}{1.9(C_{75} - C_{25})} \tag{1.12}$$

o bien empleando octiles:[5]

$$K_1 = \frac{C_{87.5} - C_{12.5}}{1.7(C_{75} - C_{25})} \tag{1.13}$$

En nuestro ejemplo, el centil 12.5 se corresponde al valor 10 en ambas muestras, en tanto que el centil 87.5 al 18.25 en la muestra 1 y al valor 27.75 en la muestra 2.

Ambos índices se hallan centrados sobre el valor 1, por lo que su interpretación es:

$*K_2$ ó $K_1 = 1 \longrightarrow$ Distribución Mesocúrtica
$*K_2$ ó $K_1 > 1 \longrightarrow$ Distribución Leptocúrtica
$*K_2$ ó $K_1 < 1 \longrightarrow$ Distribución Platicúrtica

5. centiles que dividen a la muestra en 8 partes iguales.

En general, para finalizar la presentación de estos índices, si llevamos a cabo la descripción de una distribución mediante estadísticos del EDA, no es estrictamente necesario que se cumplan las condiciones de aplicación que precisa la Estadística clásica, pero sí que puede afirmarse que es más conveniente, puesto que sus índices tendrán mayor fuerza. Evidentemente, caso de no ajustarse la muestra a las características mencionadas, los índices EDA ven aumentada su potencia.

En el pequeño ejemplo que hemos presentado, los estadísticos calculados siguiendo algoritmos EDA serían:

Tabla 1.5

	MUESTRA 1	MUESTRA 2
$\overline{Q}$	13.5	13.5
TRI	13.75	13.75
MID	13.8	13.8
IQR	4	4
MAD	2	2
$PSDiqr$	2.97	2.97
$PSDmad$	2.97	2.97
CVc	0.15	0.15
H_1	-0.03	-0.03
H_2	-0.3	-11.7
H_3	0.02	0.84
K_1	1.21	2.61
K_2	1.13	4.13

Los índices estadísticos expuestos hasta este punto derivan no sólo de un punto de vista ordinal, y por tanto potenciador de su resistencia, sino de un estudio a fondo de sus características de robustez, que no trataremos dado el enfoque eminentemente aplicado de este primer Capítulo. Sin embargo, sí es conveniente reseñar la división que realizan los autores que los han propuesto entre estimadores **w, M** y **L**. Muy superficialmente, la diferenciación entre los dos primeros reside en el número de iteraciones que se realizan en su cálculo, una en los estimadores w y las necesarias hasta producirse convergencia en los estimadores M. Los estimadores L

son combinaciones lineales de estadísticos ordinales, incluyendo por tanto la Media aritmética, Mediana, y Medias recortadas.[6]

1.5. GRÁFICOS

1.5.1. TRONCO Y HOJAS (Stem and leaf)

Supongamos la siguiente muestra de valores: 112, 112, 115, 212, 213, 213, 215, 342, 358, 361, 362, 383, 433, 436, 438, 513 y 568. Una disposición gráfica de estilo clásico podría obtenerse de un diagrama de barras, por ejemplo, que los agruparía por intervalos como los siguientes:

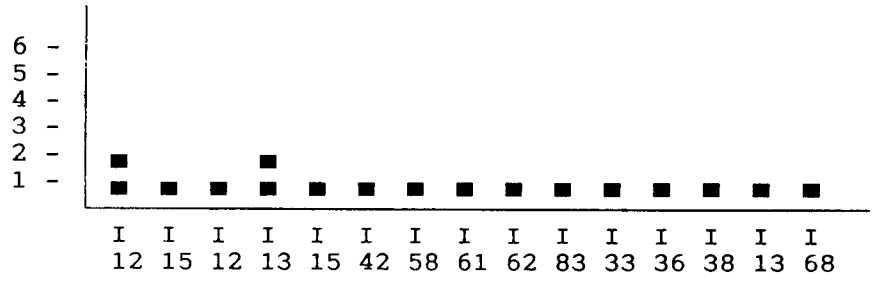


Figura 1.2

6. En el Capítulo 6 de la presente obra se amplian los L y M estimadores. Se recomienda asimismo los Capítulos 9, 10 y 11 de Hoaglin & cols. (1983), entre otros, para su ampliación.

o bien:

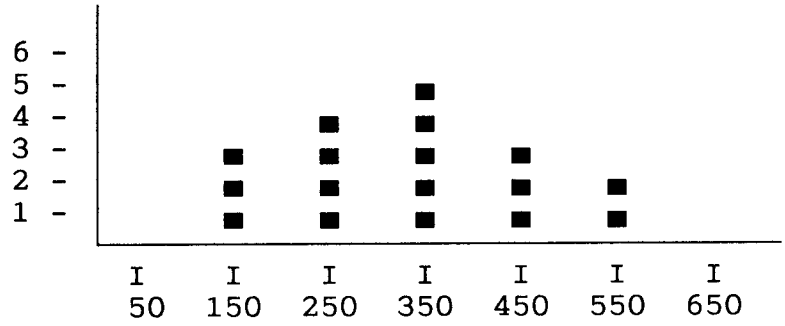


Figura 1.3

donde pueden observarse las variaciones en la información que de ellos puede obtenerse como consecuencia, en primer lugar, de la inclusión de distintos valores en un mismo intervalo, cuyos extremos o marca de clase pueden no ser representativos de la distribución interna de dichos valores o simplemente no reflejar sus posibles sub-agrupaciones (en el segundo caso, debido a la amplitud de la distribución, se han utilizado intervalos de 1/5 de rango muestral) y, en segundo lugar, de la propia elección arbitraria del intervalo escogido, que en el primer ejemplo se convierte en un mero símil de la lista original de dígitos. Al ser pocos los datos ejemplificados no se ha utilizado un diagrama de puntos, que ve disminuida su utilidad cuando el número de éstos es considerable. Por otro lado presenta particularidades que es importante conservar: la ordenación y el agrupamiento de datos similares pueden ayudar a establecer la configuración general de la distribución, aunque, en éste tipo de gráfico, sea a costa de eliminar el detalle o identificación precisa de los datos originales. Téngase en cuenta, como se ha podido deducir de los estadísticos EDA, que la ordenación de los datos, y el cálculo de índices a partir del presupuesto de ordenación, incrementa la resistencia de éstos.

El enfoque EDA propone la utilización de representaciones gráficas que potencien la "visualización" de la información, no sólo en su aspecto básicamente cualitativo, sino cuantitativo, conservando en lo posible los propios valores numéricos. Este gráfico debe pues mantener, por una parte, las características de un histograma y, por otra, las de una tabulación inicial. El primer "híbrido" entre Tabla y Gráfico que comenta-

remos se denomina **Stem-and-leaf** cuya traducción literal, tronco-y-hoja, expresa de forma clara su propósito: la parte más relevante significativa o destacada de un valor, en el contexto de toda una serie, es generalmente su primera cifra, es el *tronco* de donde partirán las *ramas* que definirán con más precisión su forma. Se trata, en definitiva, de centrar el dato por su parte más "pesada", de forma parecida a la que habitualmente empleamos cuando, hablando de cantidades monetarias, las tratamos en general de "millones", "miles", "centenares", etc.

Esta parte "definidora" del grupo, es la que utilizaremos como "tronco", situándola en forma de columna ordenada a intervalos constantes, desde el valor más bajo hasta el más alto registrado, se hallen éstos presentes o no.

La unidad que nos proporcionará la precisión informativa que podía perderse en los gráficos clásicos es la hoja o parte más variable a lo largo de la distribución, que transcribiremos horizontal y ordenadamente, en la fila definida por su valor troncal; podemos considerar como hoja al dígito que sigue inmediatamente al tronco y despreciar el resto de dígitos posibles o bien, siendo el caso en el que el tronco seleccionado deje como hojas a más de un dígito, puede optarse por incluir a todos ellos, o por truncar los datos originales en la unidad que siga al tronco, más que redondear los valores, puesto que de esta manera es más fácil la recuperación del valor original, en los casos que hagan necesaria la identificación de determinados datos, dependiendo del criterio e intenciones de quien esté realizando el gráfico. En este punto es interesante remarcar que estamos trabajando en un análisis **exploratorio** de los datos, por lo que su uso se circunscribe, básicamente, al analista; éste puede realizar infinidad de combinaciones hasta hallar la que en su opinión sea más informativa, clara o comprensible para sí mismo, y puede servir de guía para seleccionar los gráficos que presentará al público al que vaya destinado la exposición de los resultados.

Obsérvese en este punto que el primer paso para construir un diagrama de tronco-y-hoja consiste en la presentación **ordenada** de los datos, obligada por la propia forma del gráfico, que facilitará en gran medida el cálculo manual de los índices descriptivos reseñados en apartados anteriores. Nos hallamos por tanto ante un **enfoque transversal** del análisis, punto que no debe olvidarse en ningún momento si se aplica el EDA sobre datos

obtenidos de un diseño longitudinal, apartado que se tratará en Capítulos posteriores.

El número de "ramas"[7] en un gráfico de tronco y hojas puede escogerse bien a criterio del propio analista, si éste tiene cierta experiencia en su uso, bien guiándose de reglas prácticas que aconsejan iniciarlo con un máximo de $L = 10 \cdot \log_{10} n$, regla sugerida por Dixon & Kronmal (1965) para el dibujo de histogramas, que proporciona buenos resultados cuando $20 \leq n \leq 300$ (con lo que, para $n = 20$ emplearíamos el mínimo de 13 líneas, hasta las 25 aconsejadas cuando $n = 300$). La amplitud del intérvalo se hallaría entonces dividiendo el rango entre el número de líneas, y redondeándolo hasta la potencia de 10 más próxima. Obsérvese que la diferencia respecto del histograma en este punto reside en que, en tanto que este último debe aproximar los datos a una supuesta función de densidad, el objetivo en el diagrama de tronco-y-hoja es producir un intervalo 2, 5 o 10 veces potencia de 10, para facilitar la detección de patrones no prefijados en los datos y facilitar el acceso individual a éstos. Si el tamaño muestral es ≤ 50, Velleman (1976) sugiere emplear $L = 2\sqrt{n}$ para no complicar el diagrama con un número excesivo de líneas. Un estudio detallado de ambos (Hoaglin & cols. 1983) muestra como ambos tipos de cálculo intersectan en n= 100, con lo que podría sugerirse, a nivel eminentemente aplicado, emplear la regla logarítmica cuando $n > 100$ y la de raiz cuadrada cuando sea inferior. Finalmente, Sturges (1926) sugirió que, cuando n es potencia de 2, se usara $L = 10 \cdot \log_2 n$, que Hoaglin & cols. (1983) encuentran satisfactoria cuando n oscila entre 30 y 40 individuos. Con ello comprobamos, una vez más, la versatilidad del diagrama de tronco-y-hoja, concluyendo que la propia experiencia de cada analista le llevará a seleccionar un valor determinado, más teniendo en cuenta que con los modernos algoritmos implementados en paquetes informatizados, pueden probarse varias opciones en pocos minutos, y escoger la más conveniente en cada caso.

Si se observan valores muy alejados del núcleo central, que precisarían de un gráfico excesivamente alargado, puede optarse por truncar la línea central del gráfico, al igual que se lleva a cabo en los histogramas clásicos, o bien por incluir una línea cuyo tronco consista en la palabra "altos" o

7. Unicamente hemos hallado esta denominación, tras su uso interno por parte de los autores, en la obra de Horber (1990). ("rameaux" en el contexto de Stem & Leaf, traducido como "Branchage"). Su adopción nos parece contextualmente apropiada como analogía de "fila".

"bajos" según sea el extremo por el que se hayan registrado estos valores, y especificarlos por completo en la parte correspondiente a las hojas, entre paréntesis.

A uno de los lados del diagrama acostumbran a situarse, en forma de columna, los valores correspondientes a la frecuencia acumulada, con una particularidad: la adición se efectúa en doble dirección, es decir, empezando simultáneamente por los extremos superior e inferior del gráfico, hasta llegar a la fila que contenga el valor de la Md, en el que constará como valor la frecuencia absoluta de dicha fila entre paréntesis, indicando la "rama" que divide la muestra en dos subconjuntos de igual tamaño. Esta columna proporciona lo que en EDA se denomina "*depth*" (profundidad o distancia) del dato, puesto que viene a ser un indicador del intervalo que le separa del extremo más próximo. Se corresponde con su valor ordinal, eligiendo el menor entre los que le correspondan desde cada uno de los extremos de la muestra, excepción hecha evidentemente con la Md, cuya distancia será el número total de valores más 1, dividido por 2, lo que nos proporciona la fórmula general para calcular las distancias:

$$\frac{[\text{distancia previa}] + 1}{2} \tag{1.14}$$

Evidentemente, la máxima profundidad corresponderá al valor de la Md. Como comprobación de la buena ejecución del gráfico, cuando éste se lleva a cabo "manualmente", es conveniente sumar las frecuencias anterior y posterior a la filas en que se encuentra la Md, junto con ésta; dicho sumatorio debe ser evidentemente igual al tamaño muestral.

Otro punto a remarcar es la utilidad de incluir en el gráfico la unidad empleada, junto con un ejemplo de su transcripción.

En el ejemplo planteado en la página 36 que estamos trabajando, una posible solución sería:

F	Tr.	Hojas				
3	1	1	1	1		
7	2	1	1	1	1	
(5)	3	4	5	6	6	8
5	4	3	3	3		
2	5	1	6			

$n = 17$ Unidad= 100; 1| 1 = 110-119

Figura 1.4

donde observamos que la mejora respecto del segundo histograma es que, si bien su perfil se asemeja al de este último, la información proporcionada al mantener los valores originales es mucho mayor.

Podemos resumir los pasos para la realización de un diagrama de tronco-y-hoja en:

1.– escoger el intervalo de unidades a representar en el tronco, intentando que éste cubra la totalidad de los datos a representar. Puede resultar útil realizar más de un diagrama, empleando distintas unidades.

2.– dibujar la línea vertical, situando las unidades seleccionadas en orden creciente o decreciente, según el tipo de datos a representar, intentando evitar la confusión visual resultante de representar distribuciones "crecientes" en orden inverso a su aumento.

3.– anotar al principio o final del diagrama la unidad representada en el tronco, con un ejemplo de su transcripción.

4.– caso de hallarse valores sumamente alejados del grupo central de los datos, puede optarse por:

   a) "partir" el tronco, al igual que se ha realizado con el diagrama de barras de la Fig. 1.1 para la segunda muestra

   b) hacerlos constar en la primera y/o última "ramas" del gráfico, en cuyo tronco figurará "altos" o "bajos" según sea el caso

c) indicar estos valores, anotándolos bajo la descripción de la unidad empleada, a fin de no distorsionar excesivamente el gráfico.

5.– anotar los valores de frecuencias absolutas y acumuladas al margen, indicando la fila en la que se halla la *Md*. Es aconsejable efectuar un recuento de las frecuencias para asegurar la presencia de todos y cada uno de los datos originales.

A fin de remarcar los posibles saltos ("*gaps*" en su acepción anglosajona) en la distribución, pueden emplearse una serie de recursos, el más simple de los cuáles consiste en diferenciar el primer dígito de las hojas, de forma que, si éste está comprendido entre el 0 y el 4, la línea se inicie con el signo "*", y si su valor está comprendido del 5 al 9, con un "o", en nuestro ejemplo mantendríamos los centenares como tronco y sustituiríamos las decenas de la siguiente forma:

3	1	*	2	2	5	
3	1	o				
7	2	*	2	3	3	5
7	2	o				
8	3	*	2			
(4)	3	o	8	1	2	3
5	4	*	3	6	8	
2	4	o				
2	5	*	3			
1	5	o	8			

Unidad= 100; 1*|2 = 102-142

Figura 1.5

con lo que podemos apreciar mejor los saltos que en la figura anterior. Se trata, en definitiva, de disminuir la amplitud de intérvalo a fin de amplificar la aportación de cada valor o de la ausencia de éstos en puntos determinados.

Evidentemente pueden llevarse a cabo subdivisiones más sensibles, siendo la mas usual dividir en cinco partes cada valor troncal:

0-1: ∗
2-3: *T* (dígitos cuya denominación inglesa es Two y Three)
4-5: *F* (Four y Five en inglés)
6-7: *S* (por Six y Seven)
8-9: o

resultando en nuestro caso:

3	1	∗	2	2	5	
		T				
		F				
		S				
	1	o				
7	2	∗	2	3	3	5
		T				
		F				
		S				
	2	o				
	3	∗				
		T				
(2)		*F*	2	8		
8	*S*	1	2			
6	3	o	3			
	4	∗				
5		*T*	3	6	8	
		F				
		S				
	4	o				
2	5	∗	3			
		T				
		F				
1		*S*	8			
	5	o				

Unidad= 100; 1∗|2 = 102-112

Figura 1.6

que para lo menguado de nuestro ejemplo quizás resultara excesivo. Para una matización aún más precisa debería modificarse el tronco, constando

éste de 2 valores (centenas y decenas en nuestro ejemplo) en lugar de uno solo.

Si los datos son pocos, o se hallan distribuidos a lo largo de muchas unidades básicas de descripción (tronco), pueden agruparse dichas unidades, situándolas, por ejemplo, en pares, con lo que el tronco, en lugar de indicar una única unidad (p. ej. 0 ó 1) se representará como "0,1" abarcando simultáneamente dos de ellas, y separando las "hojas" correspondientes a cada unidad mediante ":". Dado pues la gran flexibilidad del gráfico, no cabe detenerse en discutir cuál es el idóneo en cada caso, dependiendo de la intención de su autor al intentar mostrarnos determinadas facetas de la distribución su concreción final, que nunca será buena ni mala sino, en todo caso, discutible. Y para terminar con este párrafo destinado a sugerir otras posibilidades, comentar la posibilidad de efectuar diagramas con escalas nominales, situando las frecuencias o escalas cuantitativas en el tronco y las categorías (generalmente en abreviaturas) como hojas, tal y como presentan Erickson & Nosanchuk (1979). La única limitación del diagrama de tronco y hoja puede residir en el número de datos a representar, puesto que de tratarse de una cantidad elevada su eficacia se vería menguada, por lo que en estos casos es conveniente re-muestrear la distribución original a fin de que los valores representados no sobrepasen el centenar, aproximadamente.

Resumiendo, mediante el diagrama de tronco-y-hoja podemos obtener y/o observar fácilmente:

- **Rango** que cubren los datos.
- **Localización** de los valores centrales de la distribución.
- **Concentraciones** o agrupaciones de valores.
- **Identificación** de valores poco o muy frecuentes.
- "Gaps" o **lagunas** en los que no se registrado valores.
- Aproximación visual a la **dispersión** y **simetría**.
- Valores notablemente desviados del conjunto (**anomalías**).

Si se trabaja con dos distribuciones o sub-grupos de forma simultánea, es de gran utilidad representar ambas de forma simultánea (como se realiza

más adelante en este mismo Capítulo), puede emplearse el denominado Tronco-y-hoja "espalda contra espalda" (*back-to-back*) (Erickson & Novanchuk, 1979).

Veamos, para concluir con este apartado, los diagramas de tronco-y-hoja que proporciona el SPSS/PC+ (ver. 4.0) para los datos correspondientes a las tres muestras presentadas hasta este punto:

MUESTRA 1a					MUESTRA 1b	
Frequency	Stem	&	Leaf			Frequency
3.00	100	*	1	*	001	3.00
3.00	322	*t*	1	*t*	223	3.00
4.00	5554	*f*	1	*f*	4555	4.00
1.00	6	*s*	1	*s*	6	1.00
2.00	98	·	1	·	8	1.00
			E		(57)	1.00

Stem width: 10

Each leaf: 1 case(s)

MUESTRA 2			
Frequency	Stem	&	Leaf
3.00	1	·	111
4.00	2	·	1111
5.00	3	·	45668
3.00	4	·	333
2.00	5	·	16

Stem width: 100

Each leaf: 1 case(s)

Fig. 1.7

Finalmente indicar que, a pesar de que por razones didácticas hemos ubicado este apartado a continuación de los índices numéricos, en la práctica es generalmente aconsejable llevar a cabo su aplicación de forma previa, puesto que la información que de ellos se derive, puede servir de

orientación en la selección de los indicadores anteriormente expuestos. Lo mismo se aplica a las dos herramientas de análisis que siguen.

1.5.2. GRÁFICA DE CENTILES

Consiste simplemente en dibujar los puntos correspondientes a la intersección de los pares de valores registrados en orden creciente, en el eje de abcisas, y los valores de sus correspondientes centiles, en el eje de ordenadas. Se trata, con los ejes invertidos, de la función de distribución empírica de la variable analizada.

Su utilidad (Batista & Valls, 1985b) puede resumirse en la localización visual de puntos de interés (Q_1, Md, Q_3) de los que pueden derivarse estadísticos ya comentados (IQR, etc), la indicación de las posibles concentraciones en los datos, que se manifestarán en disminuciones en la pendiente del gráfico (y viceversa para los "saltos" en la continuidad de la variable), y de la simetría de la distribución, indicada por las desviaciones que se observen respecto de la recta dibujada con los pares

$$\begin{aligned} v_i &= Md - x_i \\ u_i &= x_{n+1-i} - Md \end{aligned} \tag{1.15}$$

respecto de la recta $u = v$. Las desviaciones por encima o por debajo de ésta recta indicarán sesgos hacia la derecha o la izquierda en la distribución, en base a que, en una distribución simétrica, los puntos por encima y por debajo de la Md siguen un patrón idéntico de dispersión respecto de ésta.

Para las muestras simuladas empleadas como primer ejemplo, sus gráficas de centiles adoptarían la siguiente forma:

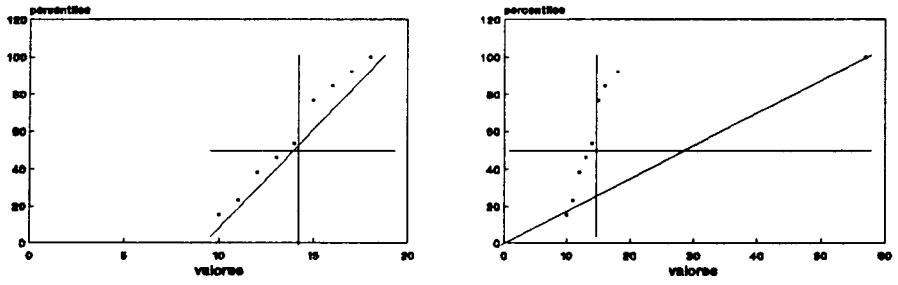


Fig. 1.8

Y para el ejemplo que se ha empleado en la exposición del gráfico de tronco-y-hoja, tendría la forma que aparece en la Fig. 1.9.

Como se habrá observado, se han añadido a la gráfica de centiles 2 rectas perpendiculares a cada uno de los ejes, partiendo del valor de la Md en la distribución analizada y del C_{50} en el eje de centiles que, caso de intersectar con la diagonal que representaría a una muestra cuyos valores equidistarán de forma perfecta, proporcionando una indicación visual de su simetría. La desviación de la diagonal por encima o por debajo de la intersección $Md - C_{50}$ sería indicativa de asimetrías negativas o positivas respectivamente y, evidentemente, la posición de la perpendicular Md respecto de la paralela al eje de abcisas C_{50} tendrá una interpretación análoga a la que se comentará en el apartado siguiente para la recta que parte del valor de la Md.

En los ejemplos anteriores, las representaciones gráficas obtenidas mediante diferenciación de los valores respecto de sus Medianas aparecen en las Figs. 1.10 y 1.11 para las dos muestras utilizadas hasta el momento.

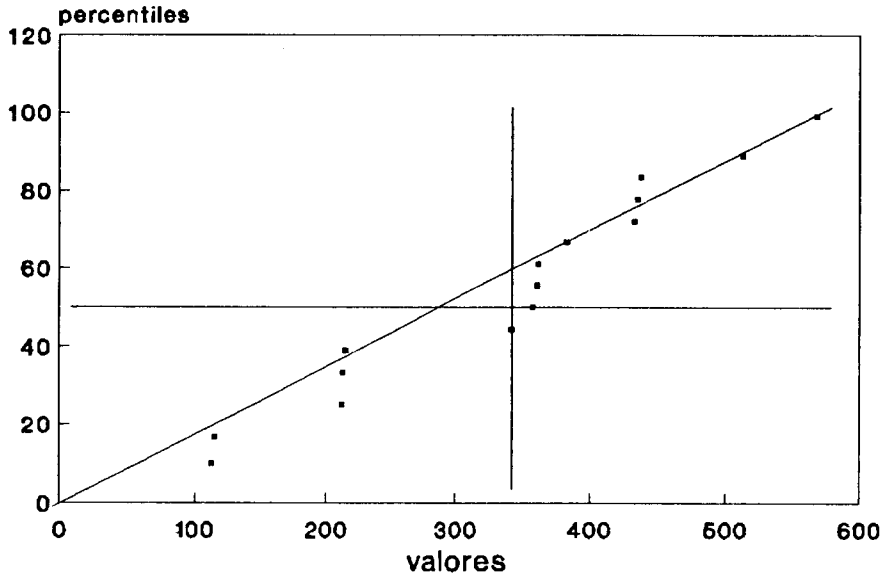


Figura 1.9

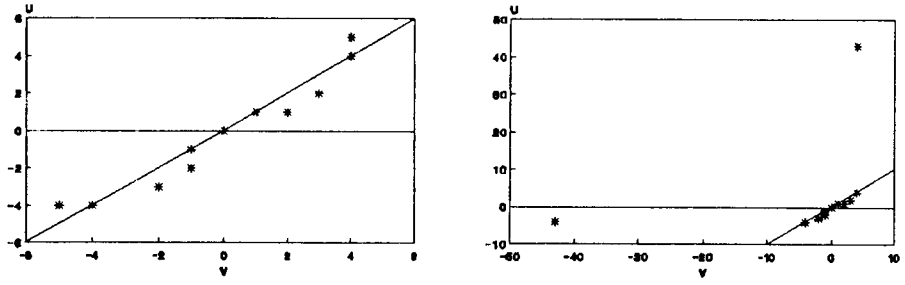


Figura 1.10

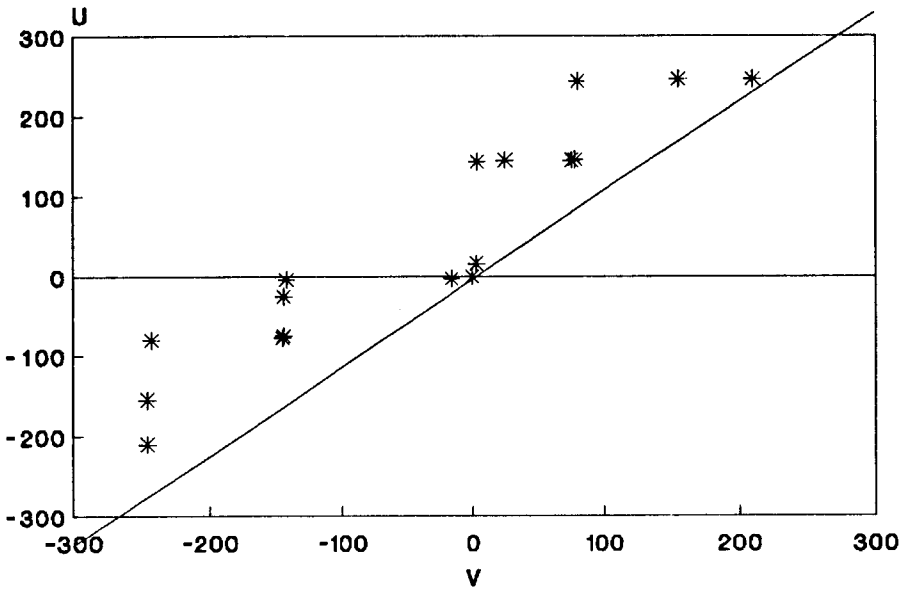


Figura 1.11

1.5.3. DIAGRAMA DE CAJA (*Box Plot*)

Denominado por algunos autores (Erickson & Nosanchuk (1979) por ejemplo) "Box-and-Dot Plot", es decir, gráfico de caja y punto, intenta proporcionar gráficamente los valores de los datos pero sin que éstos aparezcan con todo detalle, a fin de no perder su configuración espacial; para ello mantiene únicamente sus valores indicativos principales, que pueden resumirse en:

- **Localizaciones**
- **Agrupaciones** significativas de valores

- Zonas en las que predomine la **dispersión**
- Relación entre las dos anteriores
- Referencia visual de la **simetría central** y de los **extremos**.
- Referencia visual de la **curtosis**, relacionando longitud de la Caja y de las patillas.
- **Longitud** de la colas.
- **Rango**
- Outliers, **anomalías** o valores significativamente alejados del grupo central de datos.

Como puede deducirse, la información proporcionada por el Diagrama de Caja es de gran utilidad si se complementa con la del diagrama de Tronco-y-hoja, puesto que en este caso prestamos especial atención al núcleo central de los datos y a las colas, que compararemos con las de una distribución de Laplace-Gauss de forma muy simple, como se verá a continuación.

Debido a que estas agrupaciones están generalmente mejor definidas por los métodos estadísticos clásicos y, en su defecto, son fácilmente observables, se potencian las posibles desviaciones hacia los extremos de las distribuciones, en base a la hipotética aportación de claves importantes para su explicación. Es decir, en el gráfico podrán observarse, por una parte, la agrupación central de la distribución, que es la que acostumbra a seguir una forma parecida a la distribución de Laplace-Gauss y, por otra, las anomalías respecto de dicha distribución, que acostumbran a situarse en sus colas.

La definición de anomalía toma como base la curva normal. Si nuestra muestra sigue este último modelo se cumple que:

$$IQR = 1.35\sigma \tag{1.16}$$

Por ello (Batista & Valls, 1985b) (Tukey, 1977), si por encima y por debajo de los cuartiles 3 y 1 añadimos un segmento de 1.5 IQR, obtendremos un intervalo que excluirá el 7 ‰ de los valores (3.5 ‰ a

cada extremo). Estos valores son los denominados exteriores ("*ouside observations*") o "*outliers*", que no tienen necesariamente que ser valores anómalos para una determinada muestra. Como anomalías extremas ("*far outside observations*") o "*far outliers*" se fijan aquellos valores que estén más alejados de $3 \cdot IQR$ a partir del primer y tercer cuartiles, cuya probabilidad bilateral de $2 \cdot 10^{-6}$ puede ser considerada evidentemente como anómala.

Los valores superior e inferior en la distribución, que no son *outliers* son denominados por Tukey "adyacentes", es decir, cercanos a ser considerados como outliers. Es importante su consideración puesto que no siempre observaremos valores outliers a estudiar para decidir su pertenencia a nuestra muestra

La elaboración de un Box-Plot se inicia con el cálculo del denominado gráfico de valores-letra.

1.5.4. GRÁFICOS DE VALORES-LETRA

Esta técnica analiza el conjunto de datos con especial atención a los valores que configuran sus límites aportando, entre otros, el denominado **Resumen en 5 letras**, que tal y como fue propuesto por Tukey (1977) se elaboraba en base a:

Mediana
"Cuarto" superior e inferior
Límite superior e inferior

en los que se ha traducido el término inglés "*hinge*" (bisagra) por cuarto, en base a que su cálculo consiste en hallar el punto que divide a cada una de las mitades definidas por la Md en dos partes iguales, es decir, se trata de la Md de cada mitad de la distribución, que se corresponde generalmente con el primer y tercer cuartiles, pero cuyo cálculo, $\frac{1}{2}(1+\text{nº}$ datos), puede hacerlo variar ligeramente. A efectos prácticos, sobretodo si se trabaja con un volumen relativamente grande de datos y se cuenta con un ordenador que proporcione los centiles de la muestra, las posibles diferencias respecto de los cuartiles pueden, en nuestra opinión, obviarse.

La precisión del gráfico puede incrementarse trabajando con el Resumen en 7 letras:

Mediana
"Cuarto" superior e inferior
"Octavo" superior e inferior
Límite superior e inferior

Los "octavos" son, como el lector habrá supuesto, aquellos valores que vuelven a dividir en dos partes iguales a los "cuartos", es decir, equidistantes de la Md y el límite superior o inferior. Evidentemente, si el número de datos a trabajar es suficientemente amplio, pueden llevarse a cabo sucesivas subdivisiones hasta que el analista crea conveniente. Recuérdese el algoritmo de cálculo expuesto anteriormente.[8]

Para su calculo:

1) se ordenan los datos de menor a mayor (para lo que puede aprovecharse el stem-and-leaf)

2) definimos la distancia (*depth*) de cada dato con respecto a su extremo más próximo (superior o inferior), que se corresponderá con la columna izquierda del stem-and-leaf). Los primeros valores de cada extremo tienen un valor de distancia 1, los siguientes 2, etc. Para fijar con precisión el valor de la Md existen dos procedimientos:

   a) Si n es impar, encontraremos $n-1$ distancias emparejadas y un dato sin emparejar, que corresponderá al valor que ocupa la posición $\frac{1}{2}(n+1)$, equivalente a $d(Md)$.

   b) Si n es par se designa como Md, por convención, el valor que ocupe la posición $d(Md) = \frac{1}{2}(n+1) - \frac{1}{2}$

3) Una vez hallado el valor de la Md, calcularemos los valores que dividen en 2 partes iguales a cada una de las mitades que aquella define, denominados por diferentes autores como "*hinges*" o

8. Aunque Tukey (1977) y Hoaglin & cols (1983) proporcionan algoritmos más complejos y exactos, creemos que el expuesto anteriormente es suficiente a los propósitos del presente libro, opinión asimismo manifestada por los últimos autores (Op. cit. pág. 48).

"*fourths*", que aquí hemos traducido como "cuartos". Su cálculo se lleva a cabo hallando la parte entera de $\frac{1}{2}(Md+1)$. Es decir, serán los valores que ocupen las posiciones $d(H) = \frac{1}{2}([d(Md)]+1)$ por encima y por debajo de la Md previamente calculada. Su diferencia respecto de los cuartiles tradicionales reside en la mayor proximidad que guardan estos últimos respecto de la Md, aunque a efectos prácticos creemos que puede aceptarse su uso.

Hasta aquí hemos calculado los valores que necesitamos para elaborar un gráfico de 5 letras; si queremos obtener un Box-Plot más definido optaremos por el gráfico de 7 letras, para lo que calcularemos los valores intermedios de los cuartos que han quedado a los extremos de la distribución, denominados "octavos" (*eighths*), y que abreviaremos con la letra "E", despreciando la parte fraccionaria del cálculo. Serán, por tanto, los valores que ocuparán las posiciones $d(E) = \frac{1}{2}([d(H)]+1)$ mas allá de los H calculados anteriormente.

Para resumir estos cálculos, podemos dibujar un gráfico como el siguiente, en forma de U invertida, denominado gráfico de valor-letra (*letter-value display*), que para 5 valores será:

n			
$d(Md)$		Md	
$d(H)$	Hi		Hs
1	Li		Ls

Punto Medio	Amplitud
Md	
Hx	Ha

donde:

L: límites del batch

i: inferior $\quad$ a: absoluta

s: superior $\quad$ x: medio

Figura 1.9: Gráfico de 5 letras

y para 7 valores:

n				Punto Medio	Amplitud
$d(Md)$		Md		Md	
$d(F)$	Hi		Hs	Hx	Ha
$d(E)$	Ei		Es	Ex	Ea
1	Li		Ls		

Figura 1.10: Gráfico de 7 letras

que podemos dibujar sobre una escala vertical, que abarque el rango de la distribución a representar, con los valores obtenidos en el gráfico de valores-letra:

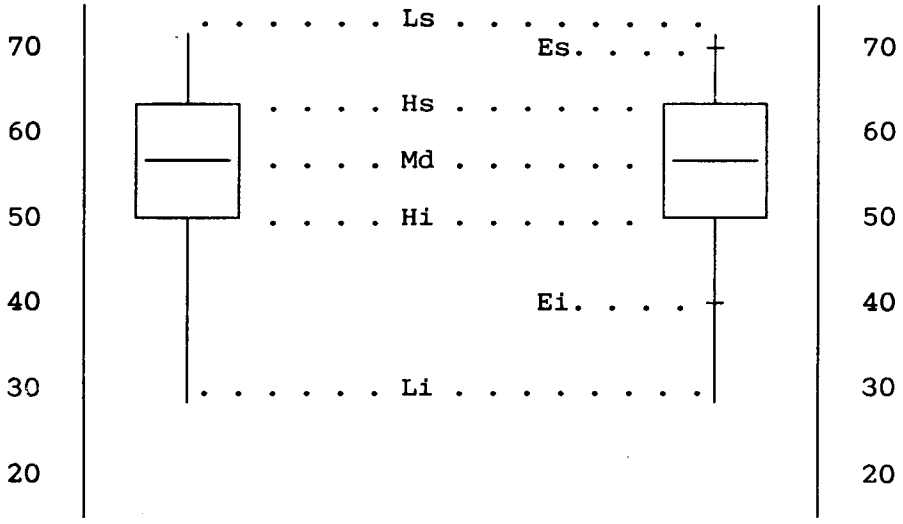


Figura 1.12

Esta figura, denominada "*Box-and-whiskers plot*" o gráfico de caja y patillas, puede complementarse incluyendo, si los hubiera, los valores que se apartan de este núcleo central de la distribución. Para ello identificare-

mos en primer lugar los límites interiores (*inner fences*: "cercas" "vallas" o, en cierto sentido "límites" interiores) que nos indicarán qué valores pueden considerarse como **Exteriores** (*outside*) y cuáles como **Remotos** (*far out*):

- Calculemos en primer lugar la amplitud de los cuartos (*H-spread*) diferenciando los valores calculados para ambas "*hinges*". Este valor se situará sobre el gráfico de valores-letra
- Fijemos un valor de paso o salto (*step*): 1.5 veces la amplitud de cuarto.
- Los límites interiores estarán un paso hacia el exterior de las bisagras.
- Los límites exteriores estarán dos pasos hacia el exterior de las bisagras (es decir, un paso más allá de los límites interiores).
- El valor que se halla en cada extremo, pero aún dentro, de los límites interiores se denomina "**adyacente**".
- Los valores entre cada límite interior y su límite exterior correspondiente se denominan **exteriores** (*outside*).
- Los que se hallen más allá de los límites exteriores son los "**remotos**".

Tukey (1977) recomienda escribir cada uno de estos valores exteriores y remotos bajo sus límites correspondientes, para así destacarlos, en el que pasa a denominarse gráfico de letras límite (*fenced letter display*):

$$\begin{aligned} \text{inferior} &= Hi - (1.5 \cdot Ha) \\ \text{superior} &= Hs + (1.5 \cdot Ha) \end{aligned} \tag{1.17}$$

y los límites exteriores (*outer fences*):

$$\begin{aligned} \text{inferior} &= Hi - (3 \cdot Ha) \\ \text{superior} &= Hs + (3 \cdot Ha) \end{aligned} \tag{1.18}$$

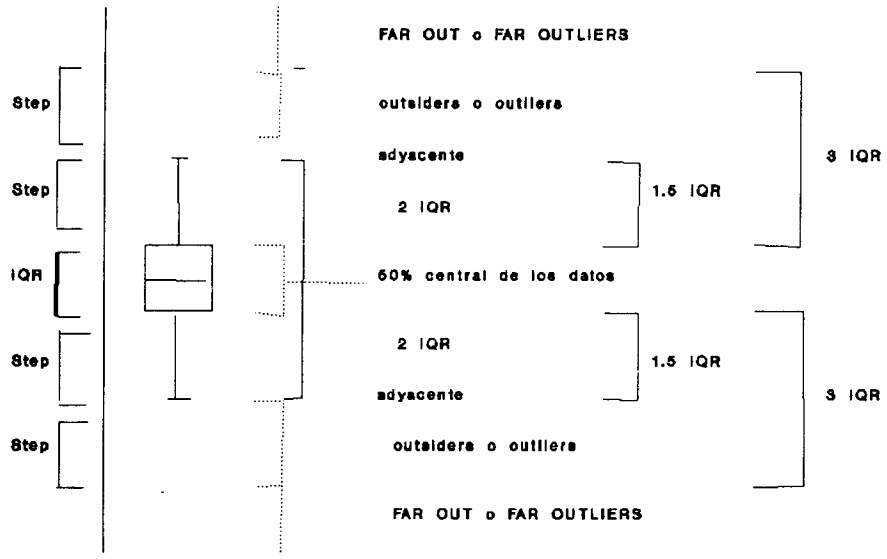


Figura 1.13

Finalmente, para dibujar el Box-plot:

- dibujaremos la caja central, mediante 3 líneas perpendiculares a la escala, en las posiciones correspondientes al primer y tercer cuartiles y la Md.
- uniremos estas 3 líneas con dos paralelas, que formarán la Caja.
- de sus lados superior e inferior trazaremos una línea ("*whisker*" o "patilla") que la una con los valores que limitan con los exteriores, cuyos dígitos indicaremos por escrito al lado.
- señalaremos todos los valores entre los límites interiores y exteriores con un punto y sus valores o etiquetas respectivas a su lado.

- y los valores exteriores a los límites exteriores con un asterisco y su valor o etiqueta al lado (Tukey (1977) recomienda transcribir en mayúsculas).

La interpretación de ciertos valores como exteriores (*outside*) se ha adoptado a partir del establecimiento de límites (E_i y E_s), criterio aplicable principalmente cuando se supone que la distribución muestral se aproxima a una curva de Laplace-Gauss. El cuarto corresponde al punto que deja mas allá el 25% de los valores, por lo que en una distribución normal se corresponderían a los valores $\mu \pm 0.6745\sigma$, o en otras palabras, a un amplitud intercuartílica de 1.349σ. Al multiplicar este valor por 1.5 y añadirle esta profundidad a los límites fijados para los cuartos (H), los límites para los valores anómalos se han situado en $\mu \pm 2.698[0.6745 + (1.349 \cdot 1.5)]$, cuya área exterior es de 0.00349 en cada extremo, por lo que consideramos anómalos aquellos valores cuya probabilidad es igual o inferior a 0.00698; el área del Diagrama de Caja que contiene datos considerados como representativos de un determinado lote es, por tanto, del 99.3% del total de individuos. La frecuencia de aparición de valores anómalos (*outliers*), en estudios de simulación realizados por Hoaglin & cols (1981) es de 0.6 para muestras de $n = 30$. Dicha probabilidad aumentaría si nos basáramos en distribuciones t de Student, por ejemplo, cuyos extremos son más pronunciados. En el caso de distribuciones χ^2, considerando aquellas con tendencia a la simetría (desde 5 hasta 20 grados de libertad) Hoaglin & cols. (1983) observan que el límite inferior que define a los valores anómalos acostumbra a situarse en un valor menor que el mínimo dato muestral, lo que podría proporcionar un indicador más de simetría: la aparición de valores anómalos a un lado u otro del gráfico.

Una forma de obtener una valor resistente de IQR es calcular qué desviación estándar tendría una distribución normal cuyo IQR fuese similar al nuestro. Para ello dividiremos el IQR por 1.349, valor denominado **$H-$pseudosigma**, que debe aproximarse al de la S muestral de una distribución que se ajuste a una ley normal (aunque la distribución no siga la ley de Laplace-Gauss, podemos usar la $H-$pseudosigma como índice de dispersión, teniendo en cuenta que será más resistente que la S clásica). En general, Tukey (1977) recomienda indicar todos los valores externos o remotos siempre y cuando su identificación no cree confusión en el gráfico, cosa que puede suceder cuando éstos excedan de 6 valores en un extremo.

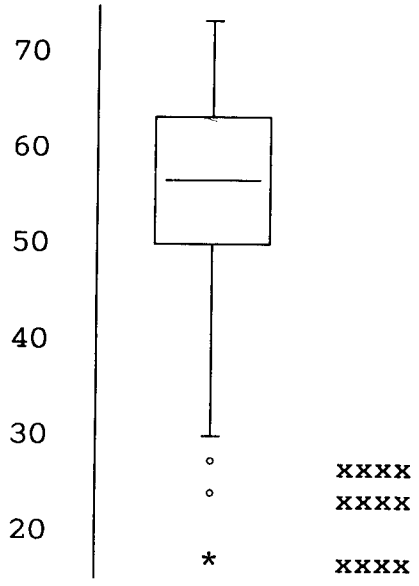


Figura 1.14

A partir de un Box-Plot podemos obtener con suma facilidad una serie de índices descriptivos de la muestra, mediante inspección visual o cálculos muy simples:

- **Rango**: diferencia entre ambos extremos del gráfico
- IQR: altura de la "caja"
- Md: como indicador resistente de agrupación o tendencia central
- TRI: que corresponderá al punto medio entre la semidistancia intercuartílica y la línea de la Md
- **Simetría**: central (en la caja) observando la posición de la Md respecto de los cuartiles, y de las colas, comparando las longitudes respectivas de las "patillas"

Tal y como se ha comentado anteriormente, si el número de individuos en la muestra es suficientemente grande, puede representarse ésta mediante gráficos de 7 valores-letra, 9 valores letra, etc. a criterio del analista.

El diagrama de Caja presenta como inconveniente la ausencia de indicadores ante distribuciones que no sean unimodales, aunque es especialmente

útil en el estudio comparativo de varias distribuciones o de subgrupos en una misma muestra. Para ello se sitúan los diversos gráficos en paralelo con una misma línea de referencia común, que permita observar sus diferencias en localización y dispersión principalmente, e incluso en simetría e importancia de las colas. En estos casos puede ser útil realizar algún tipo de transformación en los datos, como se verá en el siguiente Capítulo, a fin de conseguir unos niveles de variabilidad comparables entre sí. Por otra parte, el propio diagrama de Caja puede indicar la mejor transformación a aplicar sobre los datos cuando ésta sea necesaria, aspecto que será asimismo tratado en el próximo Capítulo.

Otro inconveniente a remarcar es el efecto de distorsión sobre la interpretación perceptiva del Diagrama de Caja que puede inducir la relación entre las respectivas longitudes de la caja y las patillas. Behrens & cols. (1990) han llevado a cabo un interesante estudio sobre el particular, relacionándolo con la similitud que guardan respecto de los estímulos empleados en estudios sobre la ilusión Baldwin (Pressey & Smith, 1986).

En las dos muestras propuestas al inicio del Capítulo, el diagrama de Caja, proporcionado por el paquete SPSS/PC+ (v. 4.0) adoptaría la forma:

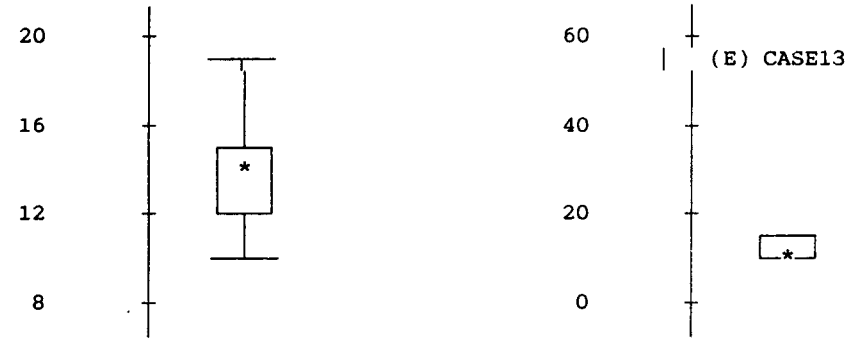
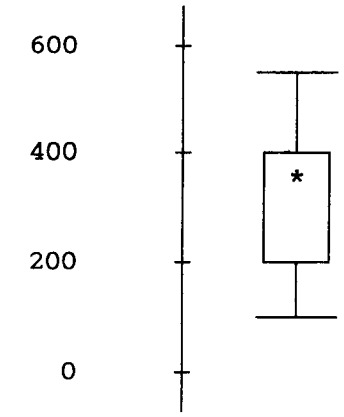


Figura 1.15

En tanto que para la muestra propuesta al inicio del apartado sobre Stem-and-leaf:



N of Cases 17.00

Symbol Key: * Median (O) Outlier (E) Extreme

Figura 1.16

1.6. PRESENTACIÓN DE UN EJEMPLO

Veamos a continuación un ejemplo con mayor número de datos, que supondremos obtenido de las calificaciones en una asignatura por un grupo de niños y niñas:

Tabla 1.6

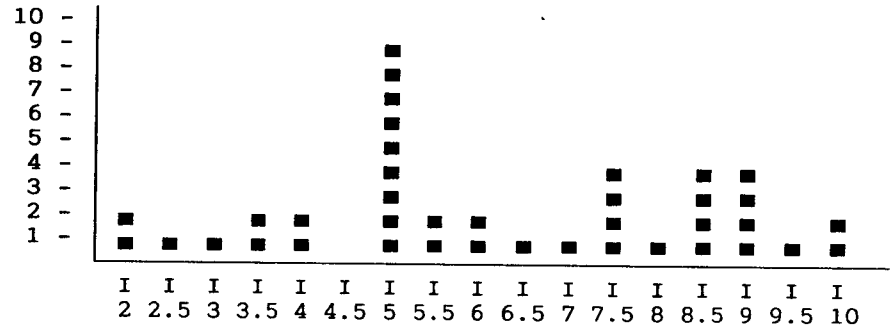
NIÑOS				NIÑAS			
2.1	5.0	6.1	8.7	2.2	4.0	5.0	7.5
2.2	5.0	6.5	8.7	2.3	4.0	5.2	7.5
2.7	5.0	7.2	8.8	2.5	4.2	5.3	8.2
3.2	5.0	7.3	8.8	2.9	4.3	5.5	9.0
3.5	5.0	7.5	9.2	3.0	4.4	5.6	10.0
3.7	5.1	7.5	9.2	3.5	4.5	6.0	10.0
4.0	5.2	7.7	9.5	3.5	5.0	6.0	10.0
4.0	5.3	8.2	10.0	3.5	5.0	6.0	
5.0	5.4	8.4	10.0	3.7	5.0	6.2	
5.0	6.0	8.5		3.8	5.0	7.2	

describiendo la muestra mediante algunos estadísticos tradicionales:

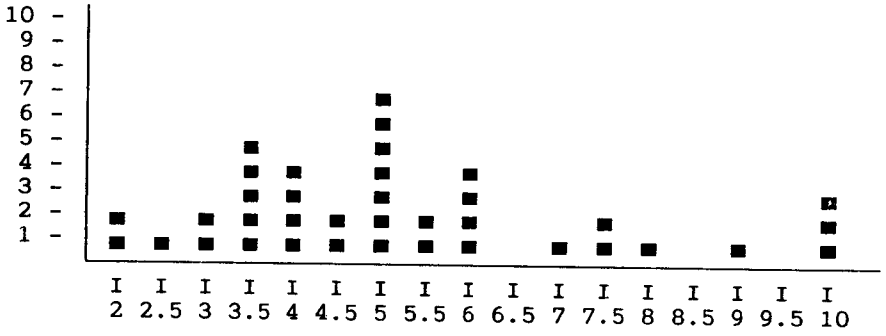
Tabla 1.7

	NIÑOS	NIÑAS
n	39	37
$\overline{X}$	6.29	5.31
Md	6	5
Mo	5	5
S^2	5.17	4.54
Simetría	-0.05	0.83
Curtosis	-1.1	0.1

y gráficamente:



Niños



Niñas

Figura 1.17

Los centiles necesarios para calcular los estadísticos propuestos por el EDA son:

Tabla 1.8

	NIÑOS	NIÑAS
C_{10}	3.2	2.82
$C_{12.5}$	3.5	2.98
C_{25}	5	3.75
C_{50}	6	5
C_{75}	8.5	6.1
$C_{87.5}$	9.2	8.4
C_{90}	9.2	9.2

que nos permiten llegar a los siguientes índices:

Tabla 1.9

	NIÑOS	NIÑAS
$\overline{Q}$	6.75	4.93
TRI	6.38	4.97
MID	6.23	4.52
IQR	3.5	2.35
MAD	2	1.2
$PSDiqr$	2.56	1.67
$PSDmad$	2.97	1.78
CVq	0.26	0.24
$H1$	0.125	-0.015
$H2$	-0.2	-1.01
$H3$	0.03	0.2
$K1$	0.96	1.36
$K2$	0.9	1.43

El diagrama de tronco-y-hoja para el conjunto de los datos, sin diferenciar por sexos es:

Frequency	Stem	&	Leaf
7.00	2	·	1223579
9.00	3	·	025555778
8.00	4	·	00002345
20.00	5	·	00000000000012233456
7.00	6	·	0000125
8.00	7	·	22355557
8.00	8	·	22457788
4.00	9	·	0225
5.00	10	·	00000
Stem width:	1.0		
Each leaf:	1 case(s)		

Figura 1.18

y su diagrama de caja:

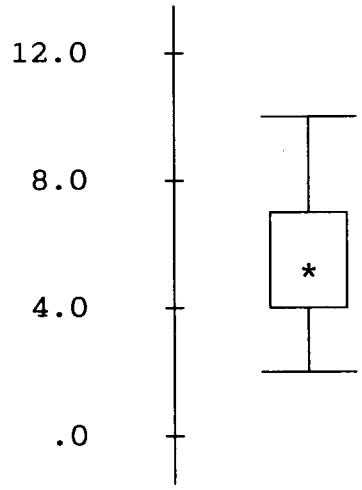


Figura 1.19

A continuación podríamos dibujar el stem-and-leaf para los dos grupos, de la siguiente forma:

Fa	Fi	Niños		Niñas	Fi	Fa
3	3	7 2 1	2	2 3 5 9	4	4
6	3	7 5 2	3	0 5 5 5 7 8	6	10
8	2	0 0	4	0 0 2 3 4 5	6	16
19	11	4 3 2 1 0 0 0 0 0 0 0	5	0 0 0 0 0 2 3 5 6 9	(9)	
(3)	3	5 1 0	6	0 0 0 2	4	12
17	5	7 5 5 3 2 7	2	5 5	3	8
12	7	8 8 7 7 5 4 2 8	2		1	5
5	3	5 2 2	9	0	1	4
2	2	0 0 10	0	0 0	3	3

(Unidad: 0.1 = 1/10 de punto; 2|2 representa 2.2)

Figura 1.20

Las gráficas de centiles obtenidas serían:

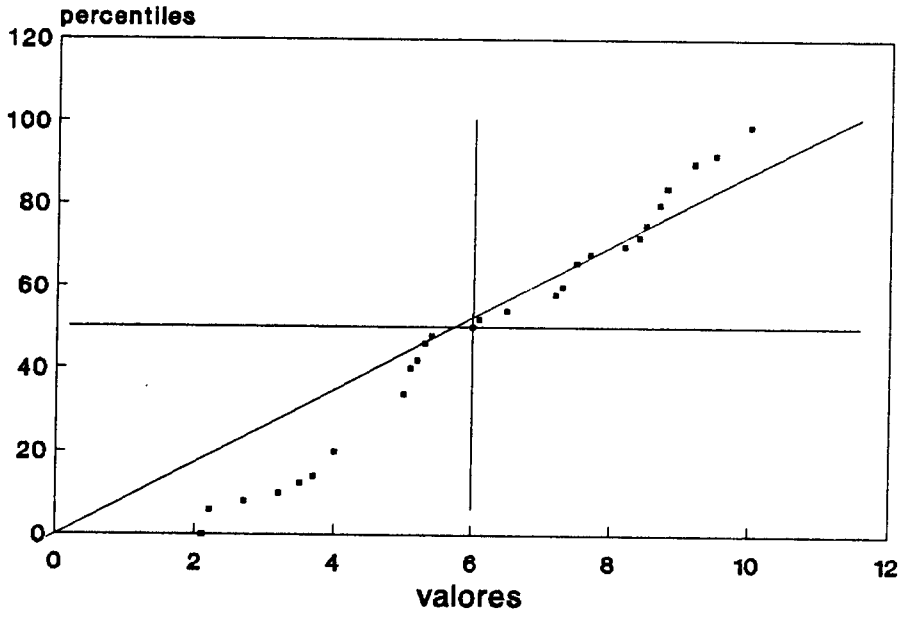


Figura 1.21

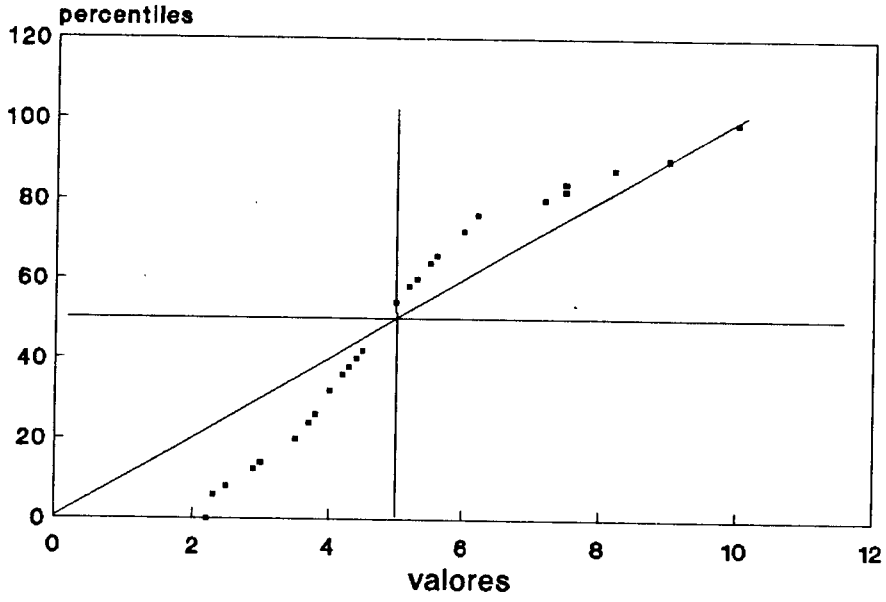


Figura 1.22

Ajustando a las rectas u y v para observar su simetría:

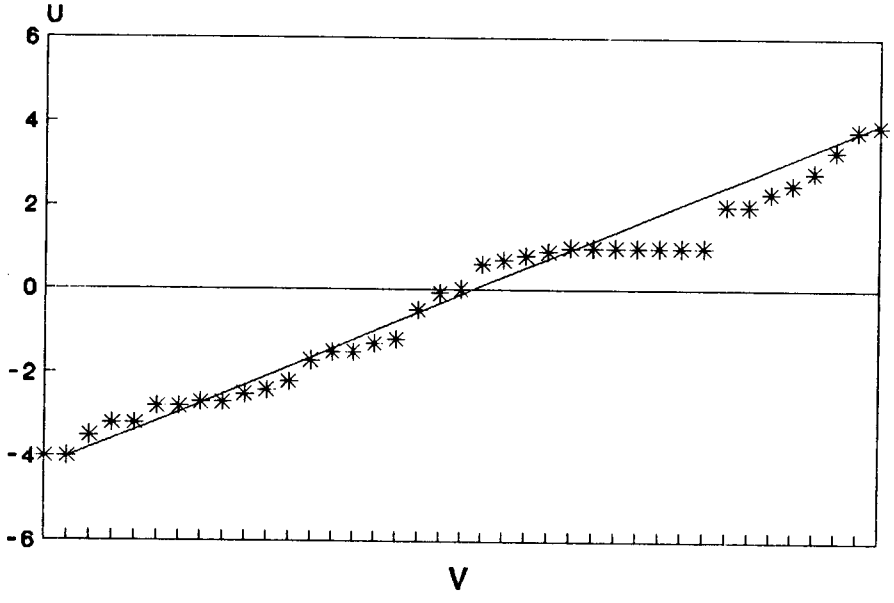


Figura 1.23

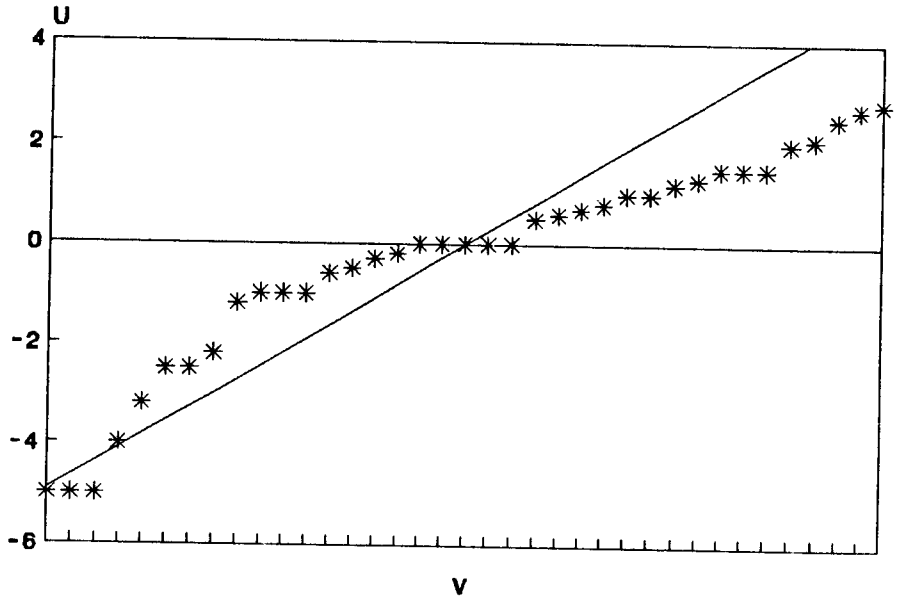


Figura 1.24

y los Box-Plots mediante SPSS/PC+ (v. 4.0) en el programa para MANOVA:[9]

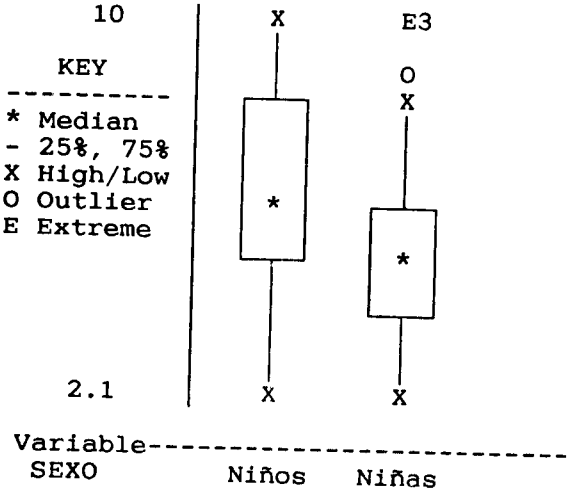


Figura 1.25

9. Obsérvense las diferencias terminológicas adoptadas en el paquete estadístico para etiquetar valores anómalos en la distribución.

o bien, mediante el SPSS/PC+ v3.1

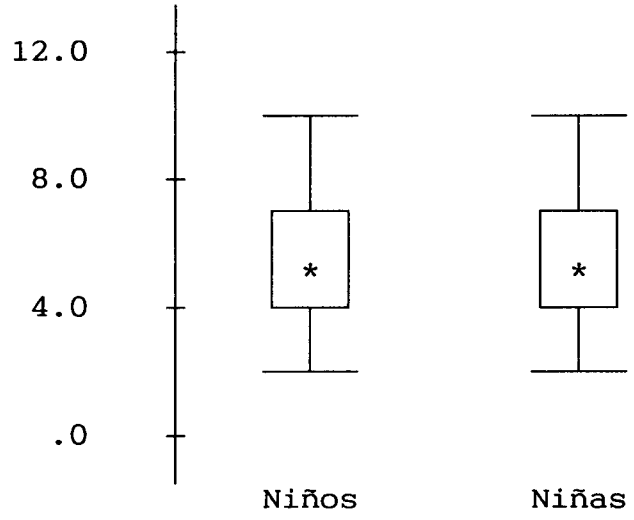


Figura 1.26

2. TRANSFORMACIÓN DE LAS VARIABLES

2.1. INTRODUCCIÓN

El capítulo anterior nos ha introducido en una serie de técnicas que nos permitirán llevar a cabo representaciones visuales de los datos. Así mismo hemos promovido escepticismo frente a los índices resúmen que caracterizan la tendencia central o la dispersión de los datos, basándonos esencialmente en la poca resistencia que tienen la mayoría de los estadísticos habituales al sintetizar distribuciones asimétricas, o en presencia de valores alejados.

Frecuentemente, los datos, en la métrica original en que fueron recogidos, se distribuyen de forma asimétrica, incluyen valores alejados, presentan variabilidad no constante o exhiben un patrón no lineal de relación entre variables.

De hecho, el principal objetivo de todos estos métodos exploratórios, consiste en evaluar en que medida las distribuciones empíricas que se consideran padecen alguno de los problemas mencionados.

La utilización de una **transformación** de la escala de medida, en lugar de las puntuaciones directas, puede solventar simultáneamente la mayor parte de estos inconvenientes.

Presentaremos en este capítulo aquellas transformaciones de los datos más extensamente aplicadas, y que se conocen como **transformaciones de potencia** (Box y Cox, 1964). Aunque también se pueden considerar transformaciones las codificaciones o categorizaciones, o sea crear una variable cualitativa a partir de valores cuantitativos, no las estudiaremos en este apartado ya que en este caso se pierde el carácter inyectivo de la relación binaria entre el conjunto de partida y el conjunto de valores reexpresados, mientras que en las transformaciones de potencia existe una relación funcional que permite el paso de un conjunto a otro simplemente conociendo la función que los relaciona.

2.2. TRANSFORMACIONES DE POTENCIA

Dentro de las infinitas posibilidades de reexpresión de un conjunto original de datos, unas de las más sencillas y ampliamente utilizadas son las denominadas transformaciones de potencia. Debemos clarificar, por otra parte, que el origen de estas transformaciones no se encuentra en el E.D.A., ya que algunas de ellas, sobretodo la transformación logarítmica, ya son normalmente utilizadas por la estadística clásica, sobretodo por economistas y psicofisiólogos (Levey, 1980), dada la naturaleza de los datos por ellos registrados. Por tanto, la aportación del E.D.A. consiste en facilitar la elección de la mejor transformación a realizar en cada serie de datos, siempre que, con la ayuda de técnicas gráficas (gráficos de tronco y hojas y de caja) hayamos observado un marcado sesgo en nuestra distribución original.

Como se ha indicado anteriormente, el propósito que perseguimos con la transformación es conseguir la simetrización de los datos originales. Podríamos intentar encontar transformaciones más o menos complicadas, al estilo de la transformación hiperbólica de Fisher para el coeficiente de correlación de Pearson, para conseguir la normalización o estandarización de los datos. Pero este tremendo esfuerzo no nos merece la pena, ya que

consiguiendo simetria en la distribución es posible utilizar prácticamente todas la pruebas estadísticas que exigen normalidad de los datos, al ser estas suficientemente robustas.

Es importante observar, por otra parte, que las transformaciones las realizamos con los datos registrados en la escala original, nunca con los datos previamente estandarizados, ya que la reexpresión, además de la forma, también nos afectará la tendencia central y la dispersión de los datos. Por otra parte la previa estandarización imposibilitaría algunas de las transformaciones, ya que ,por ejemplo, no está definido el logaritmo de valores negativos, o dos puntuaciones iguales pero de signo diferente nos producirian el mismo valor si utilizamos la transformación $y = x^2$.

Formalmente definiremos una transformación de la siguiente forma:

Sea T una función ,que aplicada sobre una serie $x_1, x_2, \ldots, x_n$, sustituye cada valor x_i por un nuevo valor $T(x_i)$ de tal manera que la serie quedará convertida en $T(x_1), T(x_2), \ldots, T(x_n)$.

Las transformaciones de potencia, que son las que nos ocupan, son simples reexpresiones que cumplen las siguientes propiedades:

1.– Conservan el orden de los datos en las series originales, lo único que se modificará es la distancia entre ellos.

2.– Preservan los valores letras, excepto por pequeñas diferencias debido al redondeo. Así, por ejemplo, el valor transformado de la mediana seguirá siendo la mediana de la nueva serie.

3.– Son funciones contínuas, lo que garantiza que si los puntos estaban muy cerca en la serie original, también lo estarán en la serie reexpresada.

4.– Normalmente vienen especificadas por funciones elementales, esto es, pueden realizarse rápidamente con las mas simples calculadoras de bolsillo.

En general, las transformaciones de potencia tienen la siguiente forma:

$$T(x_i) = \begin{array}{ll} ax_i^p + b & \text{si } (p <> 0) \\ c \log x_i + d & \text{si } (p = 0) \end{array} \qquad (2.1)$$

donde a, b, c, d y p son números reales, siendo importante para que se cumplan las condiciones anteriormente mencionadas que $a > 0$ si $p > 0$ y $a < 0$ cuando $p < 0$.

Normalmente es suficiente, para conseguir la simetria de los datos originales, la utilización de un subconjunto más simple y fácil de aplicar, quedando, por tanto, de la siguiente forma:

$$T(x_i) = \begin{array}{ll} x_i^p & \text{si } (p > 0) \\ \log x_i & \text{si } (p = 0) \\ -x_i^p & \text{si } (p < 0) \end{array} \qquad (2.2)$$

Basándose en estas últimas Tukey (1.977) propone la siguiente Escala de Transformaciones, suficiente en la mayoría de los casos:

p	-2	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1	2	3
trans	$-1/x^2$	$-1/x$	$-1/\sqrt{x}$	$\log x$	$\sqrt{x}$	x	x^2	x^3

corrección del sesgo positivo — sesgo neg.

no altera la forma

En ciencias sociales son mas usuales las transformaciones que simetrizan los datos que proceden de distribuciones con tendencia a ampliarse hacia la derecha, estas funciones cóncavas actuarán dispersando los valores de la izquierda y concentrando las anomalías de la derecha de la serie. Las dos familias de funciones cóncavas mas extendidas y de efecto más suave son las raices cuadradas y las logarítmicas.[1] Mucho más fuerte es el efecto de la transformación $-1/x$, aunque esta reexpresión, a menudo simetriza correctamente la parte central de la distribución, pero hace que los extremos sean asimétricos. El hecho de utilizar el valor negativo del recíproco cumple la función de poder conservar el orden original de los datos transformados. En efecto, si tenemos los valores 1 y 2, al reexpresarlos utilizando el recíproco, 1/1 es mayor que 1/2, si utilizamos

1. De hecho, la función logarítmica en base 10 asigna las mismas distancias a aquellos valores cuyos cocientes relativos sean idénticos, así por ejemplo:
como $10^4/10^2 = 10^6/10^4 = \cdots\cdots = 10^2$
y $\log(10^2) = 2$
por tanto $\log(10^x/10^y) = \log 10^x - \log 10^y = x - y$.

el negativo, -1 es menor que -0,5, conservando, de esta manera, el orden original.

Obviamente las funciones convexas provocarán el efecto opuesto, magnificando las distancias de la parte superior del recorrido de la escala original, mientras reducirán relativamente las de la parte inferior. Entre éstas son las transformaciones potenciales y exponenciales las más conocidas. Algunas veces será necesario añadir o sustraer alguna constante si existieran números negativos en la serie original. Veamos, con la ayuda de un simple ejemplo, como actúan las transformaciones potenciales. Sea la serie de números 2, 3, 4, si utilizamos la transformación x^2, obtendremos los valores 4, 9, 16. Mientras entre cada valor en la escala original hay una distáncia de 1 unidad, en la serie rexpresada vemos como aumentan las distáncias entre los valores altos de la distribución. Si utilizamos la transformación x^3 este efecto todavía será mayor. Los valores en este caso serán 8, 27 y 64. Donde la distancia entre el 1er. y 2.º valor es de 19 unidades y entre el 2.º y el 3er. valor de 37 unidades.

2.3. TRANSFORMACIONES LINEALES

Consideraremos brevemente esta reexpresión de los datos originales, que implica únicamente una traslación del origen y un cambio uniforme en la escala, que dependerán respectivamente de los valores de las constantes b y a de la ecuación 2.1 siendo $p = 1$, permaneciendo la distribución estadística de la variable inalterable, ya que se conservan las distancias relativas en los datos transformados.

a) Cambio en el origen

   Si hemos observado una variable X y la transformamos mediante la adición de una constante b

$$X' = X + b$$

es fácil observar que en lo único en que se verá afectada es en la traslación del origen una distancia igual precisamente a la constante b (fig. 2.1). De manera que si a todas las puntuaciones originales se les ha añadido un valor igual a b la media que obtendremos después de la transformación es igual a la media de la variable original más la constante

$$\overline{X}' = \overline{X} + b$$

en efecto

$$\begin{aligned} \overline{X}' &= 1/n\Sigma X_i' = 1/n\Sigma(X_i + b) = 1/n(\Sigma X_i + nb) = \\ &= \overline{X} + b \end{aligned}$$

por otra parte el hecho de añadir una constante a cada una de nuestras observaciones originales, no afectará a la dispersión que originalmente presenta la variable

$$\begin{aligned} S_{x'}^2 &= S_x^2 \\ S_{x'}^2 &= 1/(n-1)\Sigma(X_i' - \overline{X}')^2 = \\ &= 1/(n-1)\Sigma((X_i + b) - (\overline{X} + b))^2 = \\ &= 1/(n-1)\Sigma(X_i - \overline{X})^2 = S_x^2 \end{aligned}$$

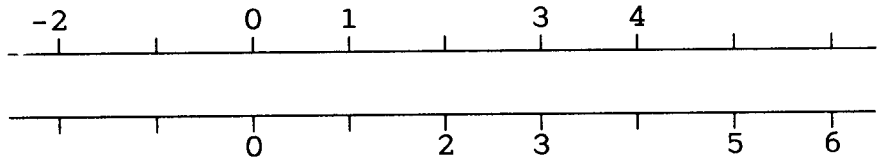


Fig. 2.1. Cambio de origen, $b = 2$.

b) Cambio de escala

Supongamos en este caso que a cada una de las observaciones originales se la multiplica por una constante a

$$X' = aX$$

en este caso obtendremos que la media de la variable así originada es a veces la media de la variable observada originalmente (ver fig. 2.2). Pero en este caso también la variabilidad se verá afectada, siendo

$$\begin{aligned} S^2_{x'} &= a^2 S^2_x \\ S^2_{x'} &= 1/(n-1)\Sigma(X'_i - \overline{X}')^2 = 1/(n-1)\Sigma(aX_i - a\overline{X})^2 = \\ &= a^2/(n-1)\Sigma(X_i - \overline{X})^2 = a^2 S^2_x \end{aligned}$$

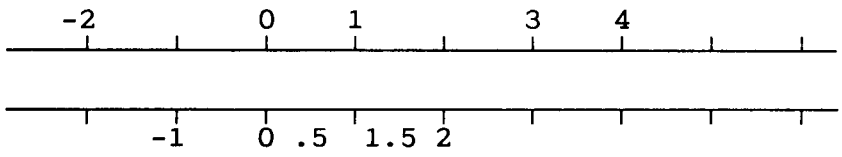


Fig. 2.2. Cambio de escala, a = 1/2.

Por tanto si combinamos los dos efectos en una sola expresión obtendremos los que se producirán en una transformación lineal del estilo

$$X'_i = aX_i + b$$

en la que tendremos que

$$\begin{aligned} \overline{X}' &= a\overline{X} + b \quad \text{y} \\ S^2_{x'} &= a^2 S^2_x \end{aligned}$$

Veamos un ejemplo, supongamos que la variable Z se distribuye normalmente con media cero y desviación típica unitaria. Consideremos ahora la transformación lineal X de Z definida por

$$X = \sigma Z + \mu \qquad \text{donde } \sigma, \mu \in \mathbf{R}$$

la distribución de la variable X, por tanto será tambien normal con media μ y desviación tipica σ. Obsérvese que estamos siguiendo precisamente el proceso inverso al que realizariamos al **estandarizar** una variable para reducirla a una distribución con media 0 y varianza 1. Por tanto la estandarización o tipificación de variables, que nos permite la utilización de una unica tabla para la distribución normal, consiste en deshacer una transformación lineal.

$$X = \sigma Z + \mu \quad \Longrightarrow \quad Z = \frac{X - \mu}{\sigma}$$

Para apreciar el efecto que la transformación tiene sobre los datos originales, consideremos una muestra aleatória de la variable Z. Escogeremos de una tabla de números aleatórios 10 secuencias de tres dígitos, consideremos estos como valores de probabilidad $P(Z \geq Z_i)$, o simplemente $P(Z_i)$, de obtener valores iguales o superiores a Z_i. A partir de la función de densidad de Z y conociendo $P(Z_i)$ calcularemos los diferentes valores Z_i, y transformando estos, obtendremos los valores X_i, utilizando por ejemplo, las constantes $\mu = 50$ y $\sigma = 10$. Estos pasos estan representados en la siguiente tabla

Tabla 2.1 Puntuaciones directas Z_i y su transformación X_i

Digitos al azar	$P(Z_i)$	Z_i	X_i	orden	
544	0.544	-0.1105	48.895	6	
632	0.632	-0.3372	46.628	4	
266	0.266	0.625	56.25	9	
265	0.265	0.628	56.28	10	
905	0.905	-1.3106	36.894	2	
706	0.706	-0.5417	44.583	3	$Q1$
397	0.397	0.2611	52.611	7	
936	0.936	-1.522	34.78	1	
567	0.567	-0.1687	48.313	5	
382	0.382	0.3002	53.002	8	$Q3$

Calcularemos algunos indices de tendencia central y de dispersión y veremos como se conserva la relación

Tabla 2.2 Índices resúmen de las variables Z y X

Original		Transformada	
$\overline{Z} =$	-0.21764	$\overline{X} =$	47.8236
$Mdnz =$	-0.1396	$Mdnx =$	48.604
$Sz =$	0.74174	$Sx =$	7.4174
$IQRz =$	0.8419	$IQRx =$	8.419
$Wz =$	2.15	$Wx =$	21.5

La utilización de transformaciones lineales obedece exclusivamente a motivos de conveniencia y facilidad de interpretación, pero no a la necesidad de producir un cambio esencial en la configuración de la variable. Así especificamos la temperatura en grados Celsius (º C), en lugar de hacerlo en grados Fahrenheit (º F)

$$C = 5/9(F - 32) = 5/9F - 160/9$$

de este modo se asigna el cero al valor de la temperatura en la cual el agua congela, referencia que facilita la vulgarización de la escala centígrada. Los paises anglosajones para medir el peso se sirven de una escala cuya unidad és la libra (Lb), sin embargo para una mayor comprensión de esta magnitud, generalmente nosotros la transformamos a nuestra escala en Kilogramos (Kgr)

$$Kgr. = 0.45Lb.$$

Ya se ha mencionado que esta transformación, consistente en añadir, substraer, multiplicar o dividir por una constante, no altera la forma de la distribución, solo sus valores numéricos, por lo que preservará además del orden de las observaciones las distancias relativas entre ellas. Por ello, el gráfico de los datos originales frente a los transformados exhibe una línea recta, lo que ilustra geométricamente que esta transformación

conservará todos los estadísticos habituales, basados en el orden o en la distancia.

2.4. TRANSFORMACIONES NO LINEALES

En el apartado anterior, hemos visto como el nivel y la dispersión de un conjunto de datos pueden variar al añadir y multiplicar por sendas constantes, cada una de las observaciones originales. Geométricamente, esta transformación se traduce en una línea recta, cuya pendiente e intersección con el eje de abcisas estan relacionadas respectivamente con la media y desviación tipo de los datos transformados. En particular, cuando el conjunto original tiene media cero y desviación tipo unitaria, ambos valores coinciden exactamente con el de estos estadísticos.

La no linealidad puede considerarse bajo distintos puntos de vista, entre ellos, sin duda el más importante, es el que distingue las transformaciones según su **monotonicidad**. Así, se denominan monótonas aquellas transformaciones en las que para todo el recorrido de la variable en la métrica original, los valores transformados siempre aumentan o disminuyen, aunque la razón de crecimiento o decrecimiento pueda variar. Un caso particular de estas, serian las reexpresiones lineales del apartado anterior, en las que esta ratio permanece constante, sea cual sea el intervalo que se considere en la variable original, preservándose de esta manera la configuración interna de los datos.

2.4.1. TRANSFORMACIONES MONÓTONAS NO LINEALES

A diferencia de las lineales, las transformaciones que estudiaremos a continuación, producen ratios de crecimiento o decrecimiento variables. Sin embargo, esto no debe ser óbice para que la información facilitada

en la métrica original sea transmitida fielmente por la serie transformada. Para ello, se requiere que las transformaciones que se puedan realizar, cumplan una serie de propiedades:

1) **Simplicidad**: esta propiedad se refiere al efecto que produce la transformación en los datos originales. Evidentemente, las más simples serán las lineales, ya que solo afectan al valor numérico. Las funciones monótonas no lineales alteran, además, las distancias relativas, pero conservando el orden. Esta propiedad en ningún caso se refiere a las operaciones matemáticas implicadas, aunque, logicamente, se preferirán aquellas funciones que son de uso común, instaladas en las calculadoras de bolsillo.

2) **Continuidad**: garantizando, de este modo, solo cambios deseables en las distancias relativas entre los puntos de la nueva escala.

3) **Monotonicidad**: de manera que la función preserve, como mínimo, el orden, y por tanto todos los estadísticos basados en éste. Solo las transformaciones no monótonas alteran la ordenación original de los datos.

4) **Derivabilidad**: característica esta que asegura la ausencia de brusquedades, que podrían motivar la invalidez de la reexpresión.

Existen ejemplos de transformaciones no lineales, que son utilizadas popularmente, incluso más que las mediciones en la métrica natural. En efecto, la medida del sonido es un típico ejemplo de escala transformada no linealmente. La escala física original mide la intensidad del sonido en ciclos por segundo (cps), pero la escala musical, más usual que la anterior, introduce la unidad de octava, adjudicando idéntico significado, en octavas, a intervalos desiguales en cps.

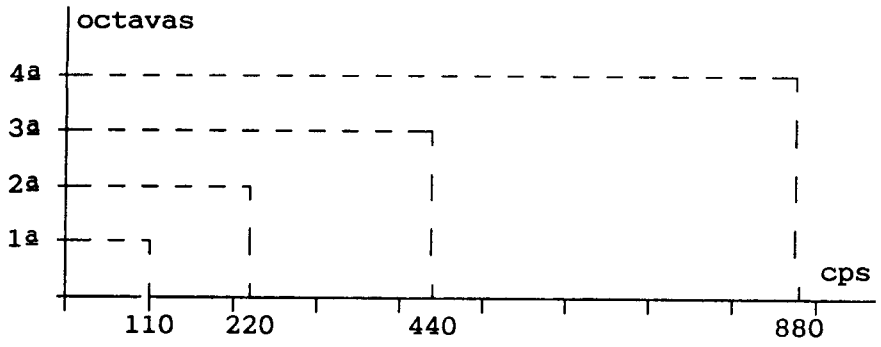


Fig. 2.3 Relación entre ciclos por segundo y octavas musicales.

Básicamente, y limitándonos al caso univariable, transformaremos los datos originales siempre que exista una fuerte asimetría y encontremos valores extremos en alguna de las colas de la distribución.

La simetría de la distribución es importante por diversas razones:

1.– El centro de las distribuciones simétricas no tiene ambiguedad en su definición, pues coinciden media, mediana y moda, cuando se trata de distribuciones unimodales. En este caso cualquiera de los estadísticos de tendencia central representa un resumen aceptable del conjunto de las observaciones.

2.– Es mucho más interpretable e informativa.

3.– Muchos de los estadísticos utilizados normalmente son buenos indices de tendencia central o de variabilidad de los datos, únicamente si la distribución goza de simetría. Además, la mayor parte de los métodos estadísticos habituales son robustos frente a desviaciones de la normalidad, siempre y cuando los datos conserven la simetría.

Por otra parte, es fácil encontar distribuciones de datos donde se produzcan los conocidos efectos "suelo" y "techo", en los cuales se observa una fuerte asimetría a la derecha e izquierda, respectivamente. Generalmente es más corriente observar el efecto "suelo", inherentes a los datos expresados en escalas de razón, que hace que a menudo observemos en estas distribuciones una mayor dispersión hacia los valores altos.

Este tipo de distribuciones tiene un crecimiento mucho más rápido hacia el final por no estar éste acotado, mientras sí que lo está el principio. Precisan, por tanto, de una transformación que actúe en sentido inverso, es decir, oprima los valores elevados y extienda relativamente los reducidos.

Utilizaremos un ejemplo para observar como afectan diferentes transformaciones no lineales realizadas sobre un mismo conjunto de datos. En la figura 2.4 tenemos una tabla en la que se detalla el consumo de helados, expresado en litros por habitante y año, de diversos paises durante el año 1990.

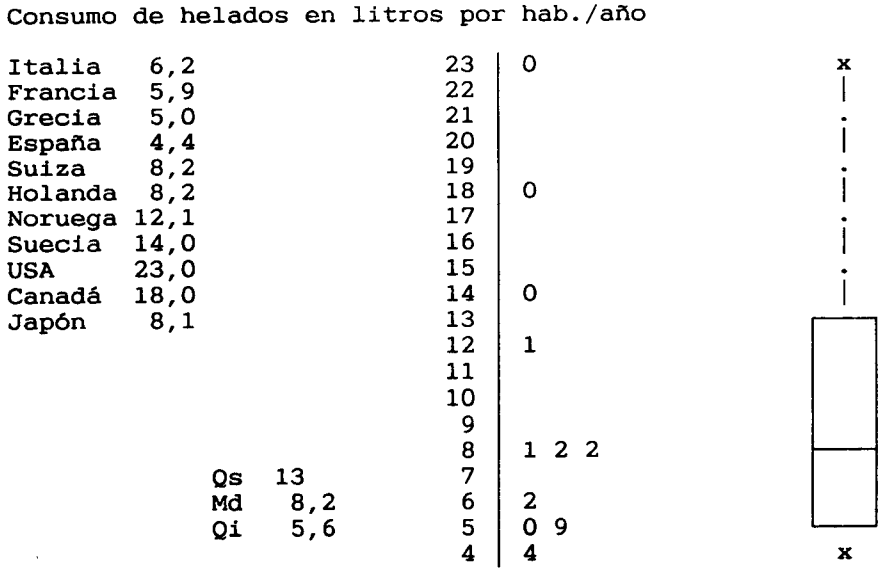


Fig. 2.4 Tabla de datos, gráfico de tronco y hojas, diagrama de caja y principales índices de la serie (Fuente: El País, 26/5/91)

A continuación realizaremos paso a paso las operaciones necesarias para realizar las diferentes transformaciones:

1.– Calcularemos la mediana, los cuartiles y los extremos de la distribución de los datos, representaremos el diagrama de tronco y hojas y el diagrama de caja. De esta manera podremos apreciar si la

distribución es o no simétrica. En la fig. 2.4 tenemos representados estos gráficos para el ejemplo. Podemos observar como la distribución se extiende hacia los valores altos, por tanto presenta un sesgo positivo.

2.– Según lo expresado anteriormente, conocemos que las transformaciones que corrigen este tipo de asimetría son, por orden de menor a mayor corrección: la raiz cuadrada, la logarítmica, el negativo del inverso de la raiz cuadrada, la recíproca negativa, etc.

3.– Reexpresaremos los datos originales utilizando las cuatro transformaciones mencionadas en el apartado anterior, calculando en cada caso los mismos índices y representaciones gráficas obtenidos en el paso 1. En las figuras 2.5, 2.6, 2.7, y 2.8, estan realizadas estas operaciones para cada una de las transformaciones utilizadas.

PAISES	x	$\sqrt{x}$
USA	(23)	4,8
Canadá	(18)	4,24
Suecia	(14)	3,74
Noruega	(12,1)	3,48
Suiza	(8,2)	2,86
Holanda	(8,2)	2,86
Japón	(8,1)	2,85
Italia	(6,2)	2,49
Francia	(5,9)	2,43
Grecia	(5)	2,24
España	(4,4)	2,1

Qs	3,61
Md	2,86
Qi	2,46

*	80
s	
f	
t	24
4	
*	
s	74
f	48
t	
3	
*	85,86,86
s	
f	43,49
t	24
2	10

Fig. 2.5 Serie reexpresada mediante $\sqrt{x}$, gráfico de tronco y hojas, diagrama de caja y principales índices.

PAISES	x	log x
USA	(23)	1,36
Canadá	(18)	1,26
Suecia	(14)	1,15
Noruega	(12,1)	1,08
Suiza	(8,2)	0,91
Holanda	(8,2)	0,91
Japón	(8,1)	0,91
Italia	(6,2)	0,79
Francia	(5,9)	0,77
Grecia	(5)	0,7
España	(4,4)	0,64

```
1,3 | 6          x
1,2 | 6          |
1,1 | 5          |
1   | 8          |
0,9 | 1,1,1      |
0,8 |            |
0,7 | 0,7,9      |
0,6 | 4          x
```

Qs 1,11
Md 0,91
Qi 0,74

Fig. 2.6 Serie reexpresada mediante $\log x$, gráfico de tronco y hojas, diagrama de caja y principales índices.

PAISES	x	$-1/\sqrt{x}$
USA	(23)	-0,21
Canadá	(18)	-0,24
Suecia	(14)	-0,27
Noruega	(12,1)	-0,29
Suiza	(8,2)	-0,35
Holanda	(8,2)	-0,35
Japón	(8,1)	-0,35
Italia	(6,2)	-0,4
Francia	(5,9)	-0,41
Grecia	(5)	-0,45
España	(4,4)	-0,48

```
-0,2 | 1          x
   t |            |
   f | 4          .
   s | 7          |
   * | 9          |
-0,3 |            |
   t |            |
   f | 5,5,5      |
   s |            |
   * |            |
-0,4 | 1,0        |
   t |            |
   f | 5          .
   s |            |
   * | 8          x
```

Qs -0,28
Md -0,35
Qi -0,41

Fig. 2.7 Serie reexpresada mediante $-1/\sqrt{x}$, gráfico de tronco y hojas, diagrama de caja y principales índices.

PAISES	x	-1/x
USA	(23)	-0,043
Canadá	(18)	-0,055
Suecia	(14)	-0,071
Noruega	(12,1)	-0,083
Suiza	(8,2)	-0,122
Holanda	(8,2)	-0,122
Japón	(8,1)	-0,123
Italia	(6,2)	-0,161
Francia	(5,9)	-0.169
Grecia	(5)	-0,2
España	(4,4)	-0,227

Qs	-0,077
Md	-0,122
Qi	-0,165

-0,04	3	x
-0,05	5	|
-0,06		.
-0,07	1	|
-0,08	3	
-0,09		
-0,1		
-0,11		
-0,12	3,2,2	
-0,13		
-0,14		
-0,15		
-0,16	9,1	
-0,17		|
-0,18		.
-0,19		|
-0,2	0	.
-0,21		|
-0,22	7	x

Fig. 2.8 Serie reexpresada mediante $-1/x$, gráfico de tronco y hojas, diagrama de caja y principales índices.

4.– Compararemos los efectos producidos por las diferentes transformaciones, escogiendo aquella que consiga simetrizar mejor los datos. En el ejemplo podemos observar que la raiz cuadrada, o sea la transformación más suave de las cuatro empleadas no llega a conseguir el objetivo de simetrizar la serie, mientras que la reexpresión recíproca negativa sobretransforma los datos originales, cambiando el sesgo positivo por uno negativo en los extremos. Las transformaciones que mejor se comportan en nuestro caso son la logarítmica y la $-1/\sqrt{x}$, sobretodo esta última que consigue simetrizar incluso los extremos de la distribución, y no solo la parte central. En caso de que nos encontremos dos transformaciones que actúen de manera similar, utilizaremos la menos fuerte.

2.4.2. TRANSFORMACIONES PARA PROMOVER SIMETRÍA

El método que hemos utilizado hasta el momento, ir tanteando como afectan las diferentes transformaciones hasta encontrar la que funciona mejor, es adecuado cuando la serie original es reducida. Si el tamaño de nuestra muestra es muy grande se convierte en una empresa excesivamente pesada, si solo contamos con la ayuda de una simple calculadora. Es por ello que, a continuación, vamos a exponer una técnica gráfica que nos ayudará a encontrar el valor de p (la **potencia** de la transformación), o sea, que reexpresión será la más adecuada, sin necesidad de tener que probarlas todas.

Una manera de comprobar la simetría de una distribución, es calcular los promedios de los denominados valores letra (ver cap. 1). Calculados los promedios[2] para todos los pares de valores letra, podemos examinar la simetría de la serie. Si ésta es perfectamente simétrica todos ellos coincidirán con el valor de la mediana. Si, en cambio, esta es asimétrica por la derecha estos valores promedio irán aumentando conforme vayamos disminuyendo la profundidad a la que han sido calculados, actuando de forma inversa si el sesgo es negativo.

2. El cálculo del promedio de los valores letra se realiza fácilmente de la siguiente manera:

$$midS = \frac{1}{2}(X_s + X_i)$$

siendo S cualquier valor letra que expresemos (mantendremos la nomenclatura anglosajona), y X_s e X_i representan los valores letra superior e inferior.

El cálculo de la profundidad a la que se calculan estos valores letra se realiza de la siguiente forma:

$$d(S) = \frac{([d(S^{-1})] + 1)}{2}$$

donde $[d(S^{-1})]$ representa el valor truncado de la profundidad de la pareja de valores letra anterior a la que queremos calcular.

así, si tenemos una serie lo suficientemente larga:

$$\begin{aligned} d(Md) &= (n+1)/2 && \text{localización} && \text{mediana} \\ d(H) &= ([d(Md)]+1)/2 && \text{"} && \text{cuartiles} \\ d(E) &= ([d(H)]+1)/2 && \text{"} && \text{octavos} \\ &\vdots \end{aligned}$$

Veamos un caso práctico. En la tabla 2.3 tenemos la población de las 40 comarcas, exceptuando el Barcelonés, de Catalunya, según el censo realizado en 1986.

Tabla 2.3 Población de las Comarcas Catalanas según el censo de 1986. (unidad = 1.000 habitantes) (Fuente: Geografía Comarcal de Catalunya, coleccionable del diario Avuí 1991)

Comarca	Población	Comarca	Población
Alta Ribagorça	4	Berguedà	41
Pallars Sobirà	5	Garrotxa	45
Val d'Aran	6	Montsià	54
Priorat	10	Baix Ebre	64
Solsonès	11	Alt Penedès	67
Cerdanya	12	Garraf	72
Terra Alta	13	Anoia	80
Pallars Jussà	14	Baix Empordà	84
Segarra	17	Alt Empordà	85
Conca de Barberà	18	Selva	91
Alt Urgell	19	Osona	115
Garrigues	20	Gironès	122
Pla de l'Estany	21	Baix Camp	124
Ribera d'Ebre	24	Tarragonès	149
Ripollès	28	Bages	150
Pla d'Urgell	29	Segrià	159
Urgell	30	Vallès Oriental	240
Baix Penedès	33	Maresme	270
Alt Camp	34	Baix Llobregat	583
Noguera	36	Vallès Occidental	621

En la Tabla 2.4 se hallan calculados los valores letra y sus promedios para el conjunto de datos del ejemplo.

Tabla 2.4 Valores letra y sus promedios correspondientes, de la población de las comarcas catalanas

		Población Comarcas		
	$n = 40$			
valor letra	profundidad	X_i	$midS$	X_s
Md	20,5		38,5	
H	10,5	18,5	60,75	103
E	5,5	11,5	83	154,5
D	3	6	138	270
C	2	5	294	583
	1	4	312,5	621

Observemos como se han calculado estos valores, por ejemplo los correspondientes a los valores letra E (octavos)

$$d(E) = ([d(H)] + 1)/2 = (10 + 1)/2 = 5,5$$

por lo tanto los valores correspondientes a la letra E serán los que ocupen la posición 5,5, a partir de la cola inferior y superior de los datos ordenados. El promedio entre estos dos valores será:

$$midE = \frac{1}{2}(X_s + X_i) = \frac{1}{2}(11,5 + 154,5) = 83$$

En nuestro ejemplo podemos observar, claramente, como los promedios de los valores letra aumentan conforme disminuimos la profundidad de estos, por tanto, esto quiere decir que la distribución está sesgada positivamente.

Para la construcción del gráfico de transformación se definen las siguientes expresiones:

$$\frac{(X_s - Md)^2 + (Md - X_i)^2}{4Md} \tag{2.3}$$

y

$$\frac{X_i + X_s}{2} - Md \tag{2.4}$$

estas expresiones como se observa se pueden calcular para cada uno de los valores letra de los que dispongamos. Construiremos un gráfico

bidimensional representando los valores de la expresión (2.3) en el eje de abcisas y las de la expresión (2.4) en el eje de ordenadas. Vemos como la fórmula (2.4) representa la distancia entre el promedio de los valores letra y la mediana, si la serie es simétrica, esta distáncia siempre será igual a 0. Si no ocurre esto, y el gráfico que resulta es aproximadamente lineal, podemos calcular la **pendiente** de la recta formada por los puntos representados. Pues bien, 1 - pendiente nos dará una estimación de p, o sea el valor de la potencia de la transformación más indicada para corregir la asimetría de los datos, según la fórmula siguiente:

$$T(X_i) = kX_i^p \tag{2.5}$$

teniendo en cuenta que si $p > 0$, la constante k tomará valor 1, mientras que si $p < 0$ entonces k = - 1. Recordemos que esta estrategia se utiliza para mantener el ordenamiento original de los datos. Si $p = 0$, la transformación logarítmica será la más adecuada.

Mediante el cociente:

$$\frac{\dfrac{X_i + X_s}{2} - Md}{\dfrac{(X_s - Md)^2 + (Md - X_i)^2}{4MD}} = \text{pendiente} \tag{2.6}$$

obtenemos la pendiente de la recta que pasa por cada punto representado y el origen de coordenadas. Una estimación rápida de la pendiente de la recta que pasa más cerca de todos los puntos, es la mediana de todas las pendientes de esta forma calculadas. Una vez realizado este proceso es facil calcular la potencia de la transformación utilizando:

$$\text{potencia} = 1 - \text{pendiente} \tag{2.7}$$

En la tabla 2.5 están calculados estos valores para el ejemplo de las comarcas catalanas.

Tabla 2.5 Coordenadas y estimación de la pendiente para cada uno de los valores letra del ejemplo de la población de las comarcas catalanas

Letra	X_i	X_s	$\frac{X_i+X_s}{2} - Md$	$\frac{(X_s-Md)^2+(Md-X_i)^2}{4Md}$	pendiente
H	18,5	103	22,25	29,61	0,75
E	11,5	154,5	44,5	92,11	0,48
D	6	270	99,5	354,86	0,28
C	5	583	255,5	1.932,48	0,13

Observando la sexta columna de la Tabla 2.5 donde aparecen las estimaciones de la pendiente de la recta a partir de cada uno de los valores letra, vemos que la mediana de estas estimaciones se situa en 0,38, por tanto la potencia de la transformación necesaria será aproximadamente 0,62. En la práctica se suele redondear a las potencias sencillas más próximas, en este caso se podrían utilizar las transformaciones correspondientes a p = 0,5 o incluso $p = 0$. No obstante en la tabla 2.6 tenemos los resultados de los valores letra y sus correspondientes promedios con los datos transformados mediante $\sqrt[3]{x^2}(p = 0,66)$; $\sqrt{x}(p = 0,5)$; $\sqrt[3]{x}(p = 0,33)$ y $\log x(p = 0)$.

En nuestro ejemplo vemos como las transformaciones $\sqrt[3]{x^2}$ y $\sqrt{x}$ no son suficientemente fuertes para conseguir la simetría de la distribución, siendo necesaria la más potente de las cuatro utilizadas, la logarítmica, para conseguir, si no una simetría perfecta, si en unos niveles aceptables. En la fig. 2.9 tenemos representados el diagrama de tronco y hojas de los datos originales y el conseguido, una vez estos han sido reexpresados por sus respectivos logaritmos.

Tabla 2.6 Valores promedio con los datos reexpresados utilizando las transformaciones $\sqrt[3]{x^2}$, $\sqrt{x}$, $\sqrt[3]{x}$, $\log x$

X_i	$\sqrt[3]{x^2}$ $midS$	X_s	letra	X_i	$\sqrt{x}$ $midS$	X_s
	11,395	Md		6,2		
6,995	14,467	21,94	H	4,3	7,21	10,13
5,095	16,942	28,79	E	3,39	7,91	12,43
3,3	22,535	41,77	D	2,45	9,44	16,43
2,92	36,355	69,79	C	2,24	13,19	24,14
2,52	37,655	72,79		2	13,46	24,92

.../...

.../...

	$\sqrt[3]{x}$				$\log x$	
X_i	$midS$	X_s	letra	X_i	$midS$	X_s
	3,38		Md		1,59	
2,65	3,66	4,68	H	1,26	1,63	2,01
2,25	3,8	5,36	E	1,06	1,62	2,19
1,82	4,14	6,46	D	0,78	1,6	2,43
1,71	5,03	8,35	C	0,7	1,73	2,76
1,59	5,06	8,53		0,6	1,69	2,79

Población comarcas		Transf. logarítmica	
0	4,5,6	0	
1	0,1,2,3,4,7,8,9	t	
2	0,1,4,8,9	f	
3	0,3,4,6	s	6,7
4	1,5	*	8
5	4	1	0,0,1,1,1
6	4,7	t	2,2,3,3,3
7	2	f	4,4,5,5,5,5
8	0,4,5	s	6,6,6,7
9	1	*	8,8,9,9,9,9
10		2	0,1,1,1
11	5	t	2,2,2
12	2,4	f	4,4
13		s	
14	9	*	8,8
15	0,9	3	
valores extremos		(unidad = 0,1)	
240,270,583,621			
(unidad = 1000 hab.)			

Fig. 2.9 Gráfico de tronco y hojas de los datos originales y de su reexpresión logarítmica

2.4.3. TRANSFORMACIONES PARA CONSEGUIR DISPERSIÓN ESTABLE

Un importante campo de aplicación, de las transformaciones, es la situación en que una variable cualitativa o atributo nos divide el grupo original de observaciones en varios subgrupos. Piénsese que, en este caso, en la mayoría de las ocasiones el nivel o tendencia central de la variable de interés, dentro de cada grupo, afecta también a la variabilidad presentada por los datos. Por tanto será necesario encontrar una transformación, que además de promover la simetría dentro de cada grupo, consiga que presenten una dispersión similar entre ellos, aunque los niveles difieran significativamente.

Podemos resumir las ventajas de esta supuesta transformación ideal de la siguiente forma:

1.– Se podrán estudiar mejor los datos transformados por comparación y exploración visual.
2.– Los datos de esta manera reexpresados nos permitirán la utilización de técnicas confirmatorias clásicas, sobre todo aquellas que precisan homogeneidad de varianzas entre los diferentes grupos.
3.– Conseguiremos simetrizar y por tanto eliminar valores alejados en cada una de las series individuales.

Es importante destacar que, en esta situación, la característica más importante es conseguir que todos o la mayoría de los grupos presenten una dispersión balanceada, aunque esto suponga la imposibilidad de conseguir una perfecta simetría.

El primer paso a seguir es establecer la dependencia entre el nivel de la serie y su variabilidad. En la mayoría de los casos encontraremos que, conforme aumenta el nivel crece la dispersión de la serie, en este caso será necesario aplicar una transformación de raiz cuadrada, logarítmica, reciproco inversa, etc. Estas transformaciones actuarán reduciendo la variabilidad de los subgrupos con mayor tendencia central. En caso contrario, si la dispersión disminuye conforme aumenta el nivel, será necesario realizar alguna transformación de potencia superior a 1, en función de la fuerza de esta dependencia presentada por los datos.

Para ayudarnos en la correcta selección de la mejor transformación utilizaremos el **gráfico de logaritmos de nivel y dispersión.** Veamos con la ayuda de un ejemplo su utilización.

Tenemos en la fig. 2.10 los datos correspondientes a la evolución de la producción industrial, expresada en porcentajes, de 10 paises industrializados, durante el cuatrienio 1986 - 1989. En esta misma tabla se encuentran representados los diagramas de tronco y hojas, además de algunos indices resumen calculados en cada uno de los subgrupos. En la fig. 2.11 se encuentran comparados los diagramas de caja, correspondientes a cada año.

PAIS	1986	1987	1988	1989
BENELUX	1,1	0,7	4,9	7,4
RFA	2,2	0,2	3,7	5,2
ESPAÑA	3,1	4,6	3,1	4,5
FRANCIA	0,9	1,9	4,6	4,1
IRLANDA	2,2	8,9	10,7	11,6
ITALIA	4,1	2,6	6,9	3,9
PORTUGAL	5,7	2,4	6,2	5,2
REINO UNIDO	2,4	3,3	3,6	0,4
USA	2,9	6,1	5,8	2,9
JAPÓN	0,1	3,0	9,8	8,1

1986		1987		1988		1989	
0	1,9	0	2,7			0	4
1	1	1	9			1	
2	2,2,4,9	2	4,6			2	9
3	1	3	0,3	3	1,6,7	3	9
4	1	4	6	4	6,9	4	1,5
5	7	5		5	8	5	2,2
		6	1	6	2,9	6	
		7		7		7	4
		8	9	8		8	1
				9	8	9	
				10	7	10	
						11	6

Es	5,7	8,9	10,7	11,6
Qs	3,1	4,6	6,9	7,4
Md	2,3	2,8	5,3	4,8
Qi	1,1	1,9	3,7	3,9
Ei	0,1	0,2	3,1	0,4
IQR	2	2,7	3,8	3,5

Fig. 2.10 Datos de la evolución de la producción industrial de los paises en cada año, diagrama de tronco y hojas y principales indices resumen dentro de cada sugbrupo. (Fuente: El País, 16/6/91)

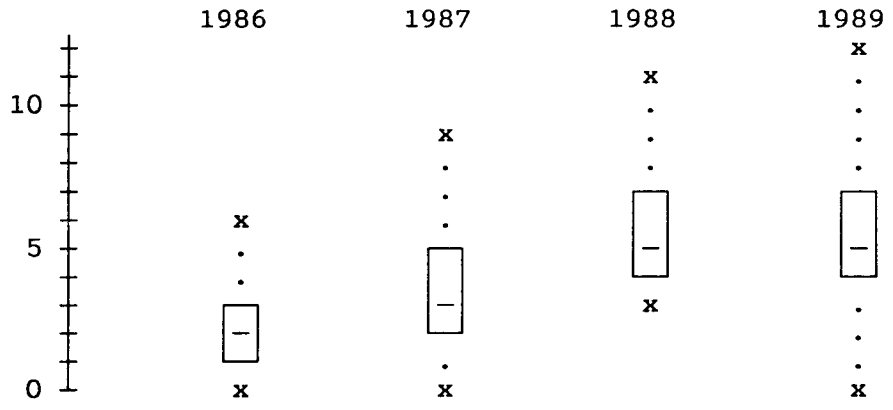


Fig. 2.11 Comparación de las dispersiones de cada grupo de valores.

Para poder representar nuestro gráfico será necesario calcular los logaritmos de la mediana y del rango intercuartílico de cada uno de los subgrupos. Estos resultados los tenemos expresados en la tabla 2.7. Construiremos un grafico representando en el eje de abcisas los logaritmos de las medianas de cada grupo, mientras que, situaremos en el eje de ordenadas el logaritmo del rango intercuartílico, tal como observamos en la Fig. 2.12 para los datos de nuestro ejemplo.

Tabla 2.7 Logaritmos del nivel y la dispersión para cada uno de los subgrupos.

Logaritmo	1986	1987	1988	1989
Md	0,36	0,45	0,72	0,68
IQR	0,30	0,43	0,58	0,54

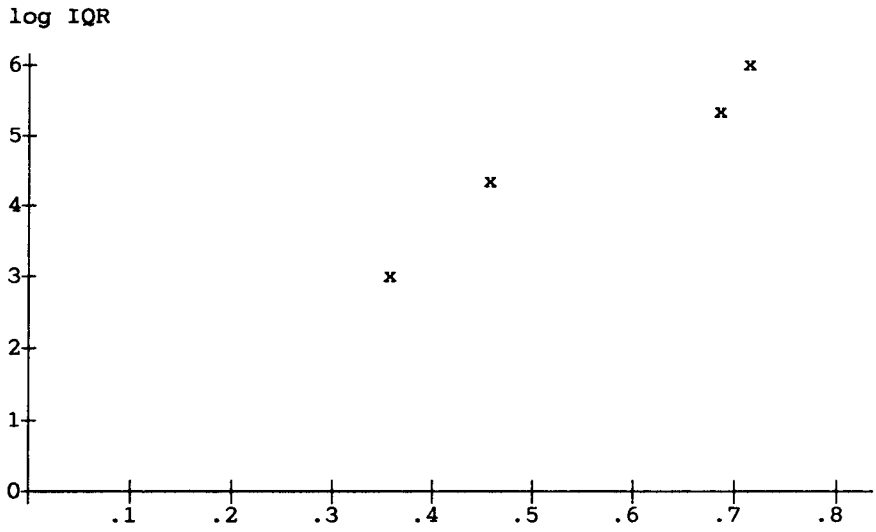


Fig 2.12 Gráfica de logaritmos de nivel y dispersión.

Representaremos en el gráfico la recta que pasando por el origen de coordenadas mejor ajuste a la nube de puntos formada por cada uno de los subgrupos. Precisamente, la pendiente de esta recta nos ayudará en la elección de la mejor transformación que tendremos que aplicar a nuestros datos. Para poder hacer una estimación aproximada de esta pendiente, utilizaremos los valores de las coordenadas de los dos puntos que delimitan esta recta, o sea el más próximo y el más alejado del origen de coordenadas. En nuestro ejemplo estos dos puntos corresponden a los subgrupos del año 1986 y 1988. La pendiente se calcula de forma directa mediante la siguiente expresión:

$$\text{pendiente} = \frac{\log IQR\,\text{sup.} - \log IQR\,\text{inf.}}{\log Md\,\text{sup.} - \log Md\,\text{inf.}} \qquad (2.8)$$

escogiendo la mejor transformación conforme a la siguiente escala:

pend.:	- 2	-3/2	-1	0	1/2	1	3/2	2
transf.:	antilog x	x^3	x^2	x	$\sqrt{x}$	$\log x$	$-1/x$	$-1/x^2$

siguiendo con nuestro ejemplo:

$$\text{pendiente} = \frac{0,58 - 0,30}{0,72 - 0,36} = 0,78$$

por tanto observaremos como actúan las transformaciones logarítmica y de raiz cuadrada. En la tabla 2.8 se hallan calculadas las reexpresiones de los indices resúmen originales de la fig. 2.10 de nuestro ejemplo.

Tabla 2.8 Indices resúmen de las series reexpresadas mediante $\sqrt{x}$ y $\log x$.

$\sqrt{x}$								
	Es	.....	2,39	....	2,98	....	3,27	....	3,41
	Qs	.....	1,76	....	2,14	....	2,63	....	2,72
	Md	.....	1,52	....	1,67	....	2,30	....	2,19
	Qi	.....	1,05	....	1,38	....	1,92	....	1,97
	Ei	.....	0,32	....	0,45	....	1,76	....	0,63
$\log x$								
	Es	.....	0,76	....	0,95	....	1,03	....	1,06
	Qs	.....	0,49	....	0,66	....	0,84	....	0,87
	Md	.....	0,36	....	0,45	....	0,72	....	0,68
	Qi	.....	0,04	....	0,28	....	0,57	....	0,59
	Ei	.....	−1,00	....	−0,70	....	0,49	....	−0,40

En nuestro ejemplo se observa que la transformación que mejor consigue homogeneizar las dispersiones, como mínimo en la parte central de las distribuciones es el de raiz cuadrada, no ocurre lo mismo si tenemos en cuenta todo el conjunto de la distribución, además, exceptuando los años 87 y 88, en los otros dos no consigue simetrizar la serie. La transformación logarímica, por su parte, promueve una mayor simetría en las series, pero maximiza las diferencias en las dispersiones.

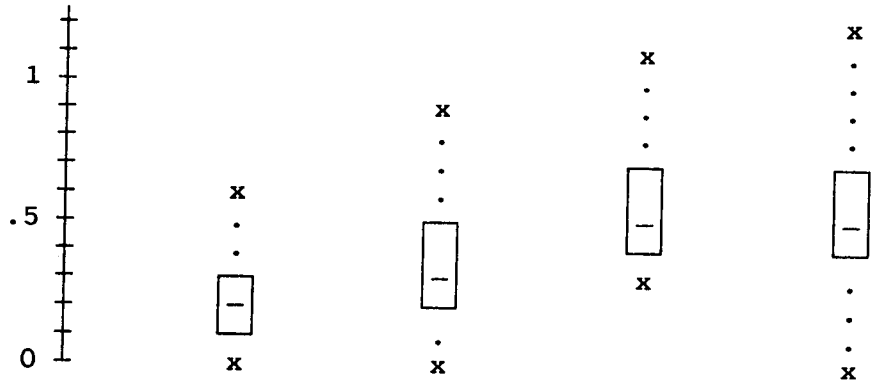


Fig. 2.13 Comparación de las dispersiones de cada grupo con los datos reexpresados logarítmicamente.

2.4.4. TRANSFORMACIONES COMPARADAS

El problema más importante que nos encontramos al realizar transformaciones, que nos ayudan a simetrizar las series o a conseguir igualdad de dispersión en el caso de múltiples grupos, es la pérdida de la escala original en la que fueron recogidos los datos. En efecto, esto nos lleva muchas veces a tener que interpretar unos datos que se hallan expresados en una unidad que carece totalmente de sentido.

Volvamos al ejemplo de la fig. 2.4, en el cual intentábamos encontrar la transformación que mejor simetrizara los datos del consumo de helados en litros por habitante y año. Naturalmente al utilizar la reexpresión las unidades originales litros . hab/año pasan a ser $\sqrt{\text{litros}}$ · hab/año o log litros · hab/año. Este hecho hace que sea difícil la comparación entre la serie original de datos y la serie transformada, para evitar este problema podemos realizar sobre la serie transformada una transformación lineal del estilo:

$$Z_i = aT(X_i) + b \tag{2.9}$$

que obligue a dicha reexpresión a modificar el origen y la escala, para

que esta se parezca lo más posible a la escala original de los datos que estamos analizando. Esto mismo nos ayudará, por otra parte, a comparar los datos transformados con los originales y estudiar de esta manera los efectos producidos por la reexpresión.
Para conseguir nuestro objetivo tendremos que estimar "a", la pendiente de la recta que relaciona los datos con la transformación efectuada. Una manera sencilla de calcularla, aunque no la más exacta, será la de escoger dos puntos x_1 y x_2, no demasiado extremos, y calcular a y b de la siguiente manera:

$$\begin{aligned} z_1 &= aT(x_1) + b = x_1 \\ &\text{y} \\ z_2 &= aT(x_2) + b = x_2 \end{aligned} \tag{2.10}$$

de esta forma

$$a = \frac{x_1 - x_2}{T(x_1) - T(x_2)} \tag{2.11}$$

Veamos el efecto que tiene esta combinación de transformaciones en el ejemplo propuesto. Realizaremos esta transformación comparada sobre la transformación logarítmica. Escojamos, por ejemplo x_1 = 14 y x_2 = 6,2, los valores originales de Suecia e Italia respectivamente. Aplicando las anteriores fórmulas:

$$\begin{aligned} a &= \frac{14 - 6,2}{\log(14) - \log(6,2)} = 21,67 \\ b &= 14 - 21,67 \log 14 = -10,92 \end{aligned}$$

Escogeremos los valores de los enteros mas cercanos, ya que esto supondrá una transformación mas simple, b = -11 y a = 22. Por tanto la transformación lineal a realizar será:

$$Z_i = 22 \log X_i - 11$$

En la tabla 2.9 vemos comparados los valores en la escala original, la transformación logarítmica y la combinada logaritmico-lineal.

Tabla 2.9 Valores originales, transformación logarítmica y transformación comparada.

PAISES	x	$\log x$	$22 \log x - 11$
USA	23	1,36	18,92
Canadá	18	1,26	16,72
Suecia	14	1,15	14,3
Noruega	12,1	1,08	12,76
Suiza	8,2	0,91	9,02
Holanda	8,2	0,91	9,02
Japón	8,1	0,91	9,02
Italia	6,2	0,79	6,38
Francia	5,9	0,77	5,94
Grecia	5	0,7	4,4
España	4,4	0,64	3,08
	$Es = 23$	1,36	18,92
	$Qs = 13$	1,11	13,53
	$Md = 8,2$	0,91	9,02
	$Qi = 5,6$	0,74	6,16
	$Ei = 4,4$	0,64	3,08

La transformación lineal realizada sobre la serie transformada, mantiene la forma de la distribución, por tanto permite observar el efecto que produce la reexpresión, posibilitando, además, una fácil comparación visual entre la serie original y la transformada.

Hasta el momento hemos presentado diferentes estrategias para escoger la **transformación más adecuada** a nuestros datos, no obstante, será necesario actuar con "prudencia", ya que es relativamente sencillo falsear los datos originales utilizando las transformaciones.

Será necesario siempre, después de la transformación, realizar un estudio detallado de los posibles valores alejados. Es posible, que al realizar cualquier transformación, los valores alejados:

a) dejen de serlo en la serie transformada.

b) continúen siendo valores extremos, incluso después de la reexpresión.

c) puedan aparecer nuevos valores extremos, que no aparecían en las puntuaciones brutas.

Por tanto, siempre después de la reexpresión, observaremos las colas de la distribución, comprobando como actúan diferentes transformaciones, y escogiendo la que mejor consiga cumplir con nuestros objetivos, sin deformar exageradamente la información original.

2.5. TRANSFORMACIONES DE LAS VARIABLES TRATADAS MEDIANTE INTERVALOS

En este capítulo hemos creido conveniente incorporar un apartado dedicado al estudio de un tipo de gráfico especialmente destacado en el ámbito de las técnicas E.D.A.. Ello se justifica en base al hecho de que su construcción pasa por someter a los datos (frecuencias individuales de cada intervalo) a una transformación de carácter simple, buscando con ello tres puntos de interés:

a) Estudio gráfico más suavizado de la distribución univariable

b) Estudio de los residuales con respecto a un modelo de probabilidad teórico

c) Utilizarlo como una prueba de conformidad a modelos teóricos de probabilidad

De este modo, como decíamos, hemos creido más conveniente incorporar tal estrategia como un apartado peculiar de las transformaciones.

No es que se pretenda redescubrir la utilidad del **Histograma** de frecuencias como gráfico tradicional del análisis de datos; sino que se propone emplearlo de una forma especial de modo que podamos extraer aún más información de la habitual. Hemos destinado una gran parte del primer capítulo de este texto a diseñar y presentar distintos gráficos para la descripción univariable. Ello, sin embargo, no hace inútil las representaciones gráficas más clásicas, puesto que en algunos casos siguen siendo irrenunciables.

Ciertamente, la representación de una variable con una amplitud muy elevada mediante un diagrama de Tronco y Hojas puede ser bastante larga y compleja y si a ello le unimos un tamaño de muestra muy grande, su elaboración no será nada rápida. Parece lógico recurrir al típico histograma de frecuencias transformando los datos en intervalos y asi, reducir el rango. Los mismos comentarios pueden efectuarse si se trata de una variable continua, en la que se justifica más claramente el empleo del histograma de frecuencias.

De ahí, pues, que nos planteemos seguir a Tukey (1971, 1977) en el estudio exploratório más avanzado del Histograma de frecuencia, y de algunos indicadores que de el se desprenden para el **análisis de residuales y el ajuste a modelos de probabilidad teóricos.**

2.5.1. DIAGRAMA DE RAIZ CUADRADA

Una de las primeras cuestiones que se plantean al establecer intervalos para describir gráficamente una variable, es la posibilidad de que la excesiva pérdida de información (pocos intervalos) conviertan al histograma en excesivamente deformado con respecto a la distribución original. Por contra, la situación no mejora sustancialmente, puesto que un exceso de intervalos lleva a la no solución de la complejidad de la representación original.

De forma complementaria, debe recordarse que en algunos casos, la amplitud del intervalo no se mantiene constante a lo largo del dominio de la variable a representar, es decir, que no será suficiente con definir la

frecuencia de observaciones por intervalo, sino que la representación gráfica final deberá tener en cuenta esta circunstancia.
Las siguientes figuras muestran dos histogramas, uno de ellos con igual amplitud de intervalo y el segundo con distinta amplitud.

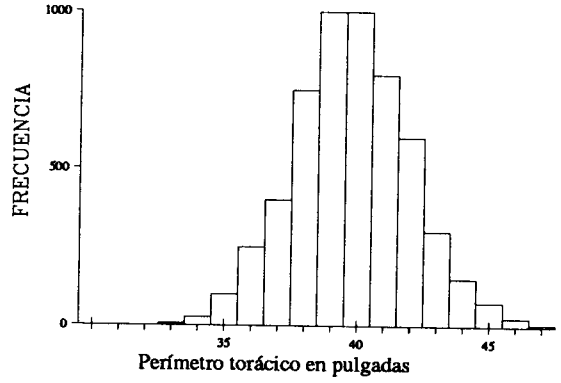


Fig. 2.14 Medidas pectorales de soldados escoceses (Tomado de Tukey, 1977; pág. 260)

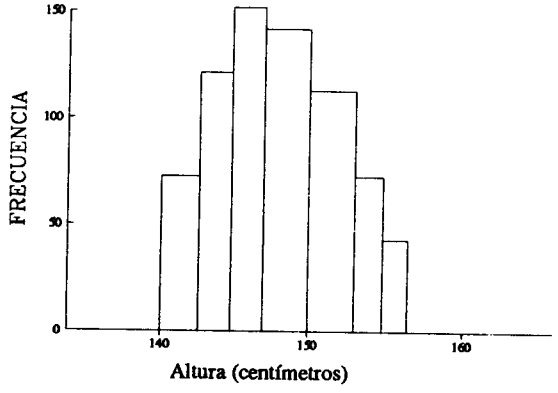


Fig. 2.15 Altura de mujeres de Bangladesh (Tomado de Tukey, 1977; pág. 262)

Tukey sugiere modificar el histograma de frecuencias tradicional en el sentido de no utilizar en el eje de ordenadas las frecuencias individuales de los intervalos, sino poner en su lugar lo que denominamos "**raiz cuadrada de la densidad del intervalo**", definiéndose esta densidad (d_i) con la siguiente expresión:

$$d_i = n_i / W_i \tag{2.12}$$

siendo W_i la amplitud del intervalo $(X_i - X_{i-1})$
n_i la frecuencia individual del intervalo.

Situar los valores de $\sqrt{d_i}$ en el eje de ordenadas origina lo que se denomina Diagrama de raiz cuadrada, que tiene como objetivo el de suavizar la forma de la distribución para facilitar un mejor análisis gráfico. En apartados posteriores abordaremos otras transformaciones para el estudio de residuales y de pruebas de conformidad.
Obviamente, cuando la variable es discreta y de rango bajo, no es preciso la utilización de intervalos, lo cual determina que W_i sea la diferencia entre valores sucesivos de la variable. En el caso más clásico, el valor de $W_i = 1$, lo que determina que $\sqrt{d_i} = \sqrt{n_i}$.
Veamos que efecto produce en unos datos la aplicación de la estrategia del diagrama de raiz cuadrada. La tabla número 2.10 contiene los pesos en gramos de una muestra de 200 recién nacidos (Adaptado de Schwartz, 1985).

Tabla núm. 2.10.: Peso en gramos de recien nacidos. (Adaptado de Schwartz, 1985).

INTERVALOS	n_i	$\sqrt{d_i}$
2100 - 2299	3	0.1227
2300 - 2499	5	0.1584
2500 - 2699	9	0.2126
2700 - 2899	13	0.2555
2900 - 3099	16	0.2835
3100 - 3299	30	0.3882
3300 - 3499	41	0.4539
3500 - 3699	36	0.4253
3700 - 3899	17	0.2922
3900 - 4099	14	0.2651
4100 - 4299	8	0.1549
4300 - 4499	3	0.1227
4500 - 4699	2	0.1002
4700 - 4899	1	0.0709
4900 - 5099	1	0.0709
5100 - 5299	1	0.0709
	$\Sigma = 200$	

Las representaciones gráficas que se obtienen con n_i y con $\sqrt{d_i}$ muestran claramente el efecto de suavizado al que nos referíamos. Las figuras 2.16 y 2.17 muestran este aspecto.

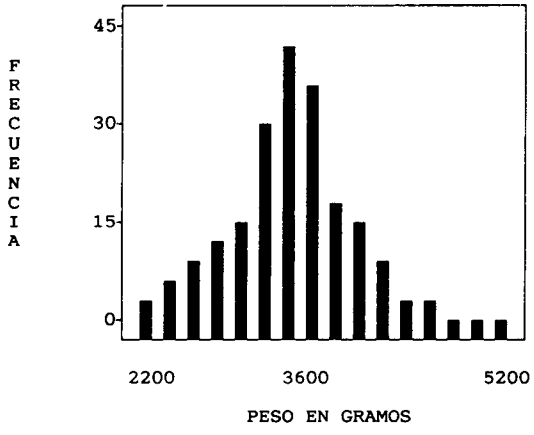


Fig. 2.16.: Histograma de frecuencias con n_i.

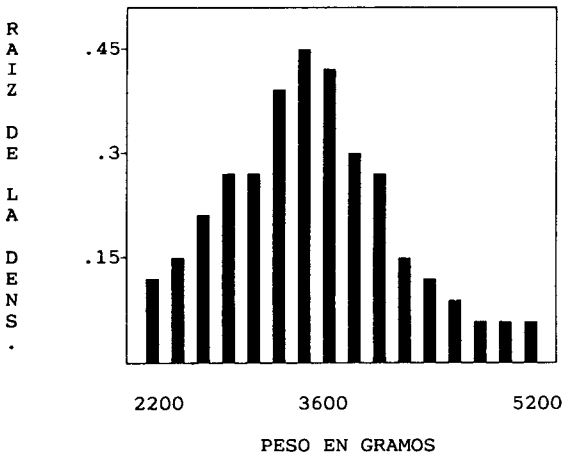


Fig. 2.17.: Histograma de frecuencias con $\sqrt{d_i}$.

No creemos que sea necesario incidir más en el efecto que se produce, puesto que la inspección gráfica de ambas figuras pone de manifiesto la suavización de la forma de la distribución original.

2.5.2. RESIDUALES DE DOBLE RAIZ

Se ha mencionado la utilización de esta estrategia como instrumento para evaluar residuales. Concretamente presentamos ahora el empleo de esa "raiz cuadrada" para evaluar el posible ajuste entre una distribución teórica y una observada. En general, estas pruebas de bondad de ajuste se basan en una expresión general de significación del residual (desajuste) definido del siguiente modo:

Residual = [Dato Observado] - [Dato Ajustado]

Para simplificarlo partiremos de una expresión más corta:

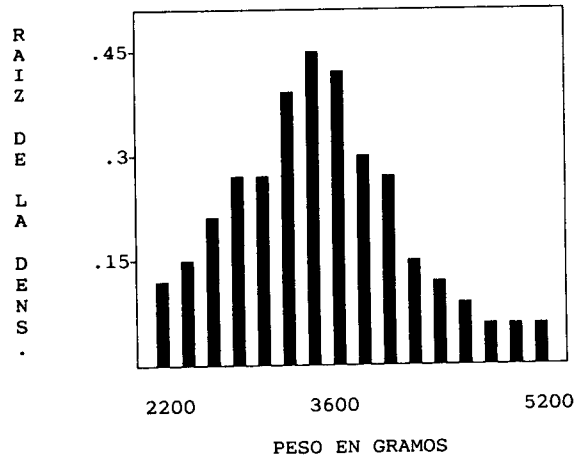


Fig. 2.17.: Histograma de frecuencias con $\sqrt{d_i}$.

No creemos que sea necesario incidir más en el efecto que se produce, puesto que la inspección gráfica de ambas figuras pone de manifiesto la suavización de la forma de la distribución original.

2.5.2. RESIDUALES DE DOBLE RAIZ

Se ha mencionado la utilización de esta estrategia como instrumento para evaluar residuales. Concretamente presentamos ahora el empleo de esa "raiz cuadrada" para evaluar el posible ajuste entre una distribución teórica y una observada. En general, estas pruebas de bondad de ajuste se basan en una expresión general de significación del residual (desajuste) definido del siguiente modo:

Residual = [Dato Observado] - [Dato Ajustado]

Para simplificarlo partiremos de una expresión más corta:

$$R = X - X' \tag{2.13}$$

siendo X la frecuencia observada y X' la frecuencia ajustada (teórica) al modelo teórico que se propone. Como decíamos, la significación del valor de "r" informa de la posible bondad de ajuste de la distribución observada con respecto a la teórica. Como es obvio, pues, tratamos de aceptar la hipótesis nula como evidencia estadística del ajuste.

Tukey (1977) mantiene el criterio de la raiz cuadrada en este tipo de situaciones y propone una reexpresión del residual en términos de raiz cuadrada. Asi,

$$R = \sqrt{X} - \sqrt{X'} \tag{2.14}$$

La obtención de unos datos, tanto observados como ajustados, no se basa en la simple raiz cuadrada de los mismos, sino que se fundamente en someter a los datos a una transformación basada en ese criterio. De este modo, se transforman los datos del siguiente modo:

Si $X <> 0$ se transforma en $\sqrt{(2 + 4(X))}$

Si $X = 0$ se transforma en un 1.

Por su parte X' se transforma aplicando a sus valores la siguiente expresión:

$$\sqrt{(1 + 4(X'))}$$

Efectuadas estas operaciones definimos la distribución de los DRR (Double-Root Residual) denominados "**Residuales de raiz doble**" del siguiente modo:

$$\begin{aligned} DRRi &= \sqrt{(2 + 4(X))} - \sqrt{(1 + 4(X'))} \quad \text{cuando } X_i <> 0 \text{ ó} \\ DRRi &= 1 - \sqrt{(1 + 4(X'))} \qquad \text{cuando } X_i = 0 \end{aligned} \tag{2.15}$$

Los valores de DRR, que como es fácil de establecer, parten de la estrategia general del diagrama de raiz cuadrada, se ajustan a una ley normal

reducida y estandarizada, con lo cual aquellos valores que superen un valor de $\pm$ 1.96 pueden considerarse como distintos de 0 y, consecuentemente, rechazar la hipótesis del ajuste al modelo teórico.

Como ejemplo de la utilización de los DRR como prueba de bondad de ajuste, presentamos unos datos sobre el número de ensayos correctos de entre 50 intentos en pruebas de precisión motriz que consigue una muestra de 100 sujetos con lesiones cerebrales graves. Se supone que el rendimiento de estos sujetos no está sujeto a su patología, y que el rendimiento debe, por tanto, seguir una distribución equiprobable entre los distintos intervalos. La tabla número 2.11 muestra las anteriores operaciones aplicadas a este caso concreto.

Tabla núm. 2.11.: Operaciones y resultados para el cálculo de los DRR.

Intervalo	X	X'	$\sqrt{(2+4(X))}$	$\sqrt{(1+4(X'))}$	DRR
0 - 5	6	10	5.09	6.40	-1.31
6 - 10	8	10	5.83	6.40	-0.57
11 - 15	11	10	6.78	6.40	0.38
16 - 20	12	10	7.07	6.40	0.67
21 - 25	20	10	9.05	6.40	2.65
26 - 30	14	10	7.61	6.40	1.21
31 - 35	10	10	6.48	6.40	0.08
36 - 40	9	10	6.16	6.40	-0.24
41 - 45	6	10	5.09	6.40	-1.31
46 - 50	4	10	4.24	6.40	-2.16
	$\Sigma = 100$	$\Sigma = 100$			

Los valores de DRR ponen de manifiesto la existencia de dos residuales de dobre raiz con valores superiores a $\pm$1.96. Concretamente los residuales del quinto y último intervalo valen 2.65 y -2.16 respectivamente. Con ello podemos concluir que la hipótesis de la equiprobabilidad de los intervalos no puede ser mantenida en nuestra distribución de frecuencias observadas.

Como es lógico, el análisis puede hacerse más exhaustivo a través del estudio de los signos y valores de esos residuales; pero ello no es intrínsecamente distinto a cualquiera de las pruebas de bondad de ajuste clásicas, por lo que no es necesario insistir en este aspecto.

La utilización de los DRR planteada en el apartado anterior adquiere su máxima relevancia cuando se aplican a una prueba de bondad de ajuste con relación a la distribución normal. A partir de lo visto hasta el momento, la cuestión fundamental radica en utilizar el modelo de la curva normal para establecer los valores de X' para cada uno de los intervalos definidos. Una vez obtenidos esos valores, será necesario repetir el proceso del apartado anterior y significar los valores de DRR.

Es suficientemente conocida la expresión que refleja la función de densidad de la distribución normal reducida y estandarizada,

$$f(z) = (1/\sqrt{2\pi}) \cdot e^{(-z^2/2)} \tag{2.16}$$

que si se quiere establecer en términos de variables centradas y para todo el intervalo, basta modificar la expresión anterior en el siguiente sentido:

$$(N/s) \cdot f(z) = (N/\sqrt{2\pi}) \cdot e^{[-(X-\mu)^2/2\sigma^2]} \tag{2.17}$$

Esta última expresión pone de manifiesto una de las prácticas normales en este tipo de situaciones. Se establecen pruebas de normalidad empleando como parámetros de la distribución teórica las estimaciones puntuales de la media y desviación muestrales. Tratándose de técnicas E.D.A., no parecerá extraño que se propongan valores distintos para μ y σ. En concreto Tukey (1977) propone los siguientes valores resistentes:

$$\begin{aligned} \mu &= \frac{1}{2}(H_i + H_s) \\ \sigma &= (H_s - H_i)/1.349 \end{aligned} \tag{2.18}$$

siendo H_i el cuartil 1 y H_s el cuartil 3. Los valores de H_i y H_s (Hinges es el término original para designar esos puntos "bisagra") mantienen todas las características propias de las técnicas E.D.A., las cuales se centran, como ya se ha comentado, en utilizar índices resistentes y básicamente de posición.

La cuestión, pues, reside en el cálculo de H_i y H_s para establecer los parámetros de la distribución teórica y, de este modo, hallar los valor de X' que, recordemos el apartado anterior, son las frecuencias esperadas según el modelo teórico en cada uno de los intervalos. El valor 1.349 constante que se utiliza en el cálculo de σ viene dado en función del valor de z que corresponde a H_i y a H_s en la distribución normal, siendo este $\pm$ 0.6745 [$p(z \geq 0.6745)$=0.25] que dado de forma bilateral fijan el valor constante del denominador (0.6745 · 2=1.349).

Los pasos necesarios para establecer H_i y H_s son los siguientes (Tukey, 1977):

a) Determinar en que intervalo se hallan H_i y H_s

$$d(H) = \{[(N+1)/2] + 1\}/2 \tag{2.19}$$

siendo $d(H)$ el número de sujetos que corresponden a ese 25% propio de las colas de los cuartiles. Suponiendo que H_i se encuentre en el intervalo determinado entre $[X_i - X_{i+1}]$ se puede plantear que:

b) Cálculo de H_i

$$H_i = X_i + \{d(H) - [(n_0 + n_1 + \cdots + n_i) - 0.5]/n_{i+1}\} \cdot W_i \tag{2.20}$$

siendo W_i la amplitud del intervalo fijado
$d(H)$ el número de sujetos acumulados hasta el intervalo siguiente al de "i"
X_i la marca de clase del intervalo fijado por $d(H)$
n_i la frecuencia observada en el intervalo "i".

c) Cálculo de H_s: suponiendo que H_s se encuentra en el intervalo $[X_s - X_{s+1}]$ podemos establecer análogamente al apartado (b) la siguiente expresión:

$$H_s = X_s + \{d(H) - [(n_{s+1} + \cdots + n_{k+1}) - 0.5]/n_{s+1} - n_s\} \cdot W_i \tag{2.21}$$

De este modo, el procedimiento de los DRR se basa en la obtención de X' a partir de la función de la curva normal estimando los valores de μ y de σ a partir de H_i y de H_s. La obtención de los DRR sigue los mismos pasos que se han descrito en el apartado anterior.

Veamos una aplicación de este método con los datos de la tabla número 2.10, en la que se muestran los pesos en gramos de 200 recién nacidos. Analizaremos los DRR que surgen del análisis de la normalidad de esa distribución. La tabla 2.12 muestra los datos originales, incorporándo la marca de clase (punto central de cada intervalo) para reconocerlos.

Tabla núm. 2.12.: Pesos en gramos de 200 recién nacidos (Tomado de Schwartz, 1985).

INTERVALO(i)	MARCA DE CLASE	FREC.(X)	
0		0	
	2100		
1		3	
	2300		
2		5	
	2500		
3		9	
	2700		
4		13	
	2900		
5		16	
	3100	→	
6		30	En este intervalo se situa H_i.
	3300		
7		41	
	3500		
8		36	
	3700	→	
9		17	En este intervalo se situa H_s.
	3900		
10		14	
	4100		
11		8	
	4300		
12		3	
	4500		
13		2	
	4700		
14		1	
	4900		
15		1	
	5100		
16		1	
	5300		
17		0	
		Σ=200	

En primer lugar establecemos el lugar (intervalo) en el que se encuentra H_i y H_s:

$$d(H) = [(200/2) + 1]/2 = 50 \quad (25\% \cdot 200 = 50)$$

Definimos en que intervalos por la dos colas suponen una frecuencia acumulada de 50 sujetos:

Cola Inferior:

$$n_0 + n_1 + n_2 + n_3 + n_4 + n_5 = 46$$
$$n_0 + n_1 + n_2 + n_3 + n_4 + n_5 + n_6 = 76$$

supone, pues que entre el intervalo 5 y 6 se encuentra H_i.

Cola superior

$$n_9 + n_{10} + n_{11} + n_{12} + n_{13} + n_{14} + n_{15} + n_{16} + n_{17} = 47$$
$$n_8 + n_9 + n_{10} + n_{11} + n_{12} + n_{13} + n_{14} + n_{15} + n_{16} + n_{17} = 83$$

supone, pues que entre el intervalo 8 y 9 se encuentra H_i. A la vista de estos datos, recordando que hemos definido en estos datos que $W_i = 200$, podemos definir H_i y H_s del siguiente modo:

$$\begin{aligned} H_i &= X_i + \{d(H) - [(n_0 + n_1 + \cdots + n_5) - 0.5]/n_6\} \cdot W_i = \\ &= 3100 + \{[(50 - 46 - 0.5)/76]\} \cdot 200 = 3109.21 \\ H_s &= X_s + \{d(H) - [(n_9 + \cdots + n_{17}) - 0.5]/n_{17}\} \cdot W_i = \\ &= 3700 + \{[(50 - 47 - 0.5)/(83 - 47)]\} \cdot 200 = 3713.89 \end{aligned}$$

A la vista de estos resultados, el cálculo de μ y σ es ya muy simple:

$$\begin{aligned} \mu &= \frac{1}{2}(3109.21 + 3713.89) = 3411.55 \\ \sigma &= (3713.89 - 3109.21)/1.349 = 448.24 \end{aligned}$$

Con estos valores como parámetros, convertimos cada marca de clase en puntuación estandarizada, en el valor de probabilidad que corresponde en

la función de distribución y en la frecuencia esperada para cada intervalo (Tabla número 2.13).

Tabla núm. 2.13.: Valores esperados para cada intervalo según el modelo de la curva normal.

INT.(i)	MAR.CLASE	FREC.(X)	Z	$F(Z)$	$X' = [p(z_i) - p(z_{i+1})] \cdot N$
0		0			0.34
	2100		-2.93	0.0017	
1		3			0.98
	2300		-2.48	0.0066	
2		5			2.92
	2500		-2.03	0.0212	
3		9			6.94
	2700		-1.59	0.0559	
4		13			14.24
	2900		-1.14	0.1271	
5		16			23.60
	3100		-0.69	0.2451	
6		30			31.24
	3300		-0.25	0.4013	
7		41			35.60
	3500		0.20	0.5793	
8		36			31.92
	3700		0.64	0.7389	
9		17			24.64
	3900		1.09	0.8621	
10		14			14.98
	4100		1.53	0.9370	
11		8			7.82
	4300		1.98	0.9761	
12		3			3.28
	4500		2.43	0.9925	
13		2			1.08
	4700		2.87	0.9979	
14		1			0.08
	4900		3.32	0.9983	
15		1			0.16
	5100		3.76	0.9991	
16		1			0.16
	5300		4.21	0.9999	
17		0			0.02
		Σ=200			Σ=200

Con los valores de X y de X' estamos en condiciones de aplicar las transformaciones propuestas por Tukey para el cálculo de los DRR. La siguiente tabla muestra los cálculos necesarios para ello:

Tabla núm. 2.14. Cálculo de los DRR.

FREC.(X)	FREC.ESP.(X')	$\sqrt{(2+4X)}$	$\sqrt{(1+4X')}$	DRR
0	0.34	1.00	1.53	-0.53
3	0.98	3.74	2.21	1.53
5	2.92	4.69	3.56	1.13
9	6.94	6.16	5.36	0.80
13	14.24	7.34	7.61	-0.27
16	23.60	8.12	9.76	-1.64
30	31.24	11.04	11.22	-0.18
41	35.60	12.88	11.97	0.91
36	31.92	12.08	11.34	0.74
17	24.64	8.36	9.97	-1.61
14	14.98	7.61	7.80	-0.19
8	7.82	5.83	5.68	0.15
3	3.28	3.74	3.75	-0.01
2	1.08	3.16	2.30	0.86
1	0.08	2.44	1.14	1.30
1	0.16	2.44	1.28	1.16
1	0.16	2.44	1.28	1.16
0	0.02	1.00	1.03	-0.03

De los valores hallados para los distintos DRR se desprende que se puede aceptar la normalidad de la distribución con μ y σ definidas a partir de H_i y H_s. El argumento que facilita tal conclusión se basa, como se ha comentado con anterioridad, en que -1.96 $\leq$ DRRi $\leq$ +1.96, lo cual implica que todos los residuales de doble raiz son iguales a 0 con un nivel de confianza del 95%.

2.5.2.2. DIAGRAMA SUSPENDIDO DE RAIZ CUADRADA

En el apartado anterior hemos analizado la estrategia que, dentro del marco de las técnicas E.D.A., se puede emplear para la evaluación de la normalidad de la distribución. Tukey (1971) propone otra alternativa, basada directamente en el diagrama de raiz cuadrada, que permite en un mismo gráfico evaluar el ajuste al modelo teórico elegido, así como analizar los residuales que se generen de tal comparación. En la figura siguiente se muestra el diagrama de raiz cuadrada que surge de los datos de la tabla número 2.10, superponiendo en el mismo la curva normal que se genera a partir de μ=3411.55 y σ=448.24.

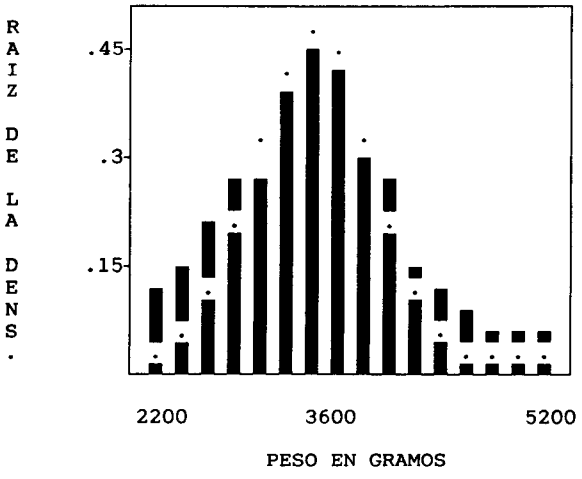


Fig. 2.18. Histograma de frecuencias con $\sqrt{d_i}$ con superposición de la curva normal.

Como se observa, los residuales con respecto al modelo quedan claramente definidos de forma gráfica. Tukey propone situar los valores originales en linea con el modelo, girar el gráfico y evaluar los residuales con respecto al eje de abcisas. Para ello el diagrama de raiz cuadrada debe quedar suspendido, es decir, con su máximo orientado hacia abajo, de forma que los residuales se situen en la horizontal de los ejes. La siguiente figura muestra este efecto.

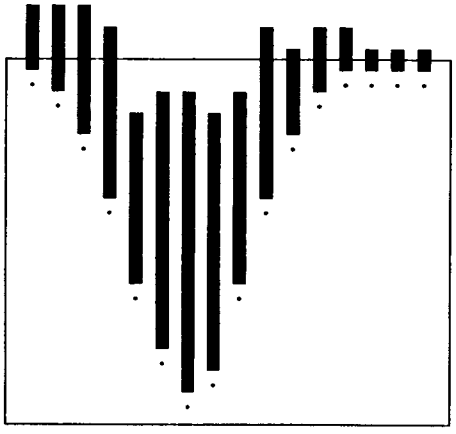


Fig. 2.19. Diagrama suspendido de raiz cuadrada

En este gráfico se pone de manifiesto el valor de los residuales que se generan con respecto al modelo teórico. Evidentemente la inspección gráfica debe venir acompañada de la significación de esos residuales, de forma que se puede adoptar una clara actitud con respecto a la normalidad de la distribución.

Recordemos a este respecto, que el diagrama de raiz cuadrada se basa en la "densidad del intervalo" (d_i) o más concretamente en la raiz cuadrada de ese valor. Así pues, podemos generar un residual de raiz cuadrada que cumpla el papel de los DRR definidos anteriormente. De este modo definimos los siguientes valores:

$$d_i = n_i/W_i \qquad d'_i = n'_i/W_i$$

siendo W_i la amplitud del intervalo
n_i la frecuencia individual observada en el intervalo
n'_i la frecuencia individual esperada en el intervalo
d_i la densidad del intervalo
d'_i la densidad estimada del intervalo.

Con estos valores podemos definir unos nuevos residuales:

$$r = \sqrt{d_i} - \sqrt{d'_i} \tag{2.22}$$

que podriamos denominar como residuales del diagrama suspendido. Su definición no es la misma que la que hemos planteado para los valores DRR, a pesar de que ambos tipos de residuales mantienen una relación que se puede reflejar en la siguiente expresión:

$$(DRR_i/2\sqrt{W_i}) \approx [\sqrt{d_i} - \sqrt{d'_i}] \tag{2.23}$$

Por ello, la significación de los residuales del diagrama suspendido de raiz cuadrada no puede efectuarse mediante el mismo criterio que el aplicado a los DRR, que recordemos que se trata de la prueba de hipótesis clásica basada en la distribución normal estandarizada. En este caso, consideramos como distintos de 0 aquellos residuales que superen el intervalo en torno a cero establecio por el valor de $\pm(1/\sqrt{W_i})$.

Veamos estos cálculos aplicados a nuestros datos de la tabla número 2.10, aprovechando algunos cálculos parciales del apartado anterior.

Tabla núm. 2.15. Cálculo de los residuales del diagrama suspendido de raiz cuadrada.

FREC.OBS	FREC.ESP.	$\sqrt{d_i}$	$\sqrt{d'_i}$	residuales
0	0.34	0	0.0412	-0.0412
3	0.98	0.1225	0.0700	0.0525
5	2.92	0.1581	0.1208	0.0373
9	6.94	0.2121	0.1863	0.0258
13	14.24	0.2549	0.2668	-0.0119
16	23.60	0.2828	0.3435	-0.0607
30	31.24	0.3873	0.3952	-0.0079

.../...

.../...

41	35.60	0.4528	0.4219	0.0309
36	31.92	0.4243	0.3995	0.0248
17	24.64	0.2915	0.3510	-0.0595
14	14.98	0.2646	0.2737	-0.0091
8	7.82	0.2000	0.1977	0.0023
3	3.28	0.1225	0.1281	-0.0056
2	1.08	0.1000	0.0735	0.0265
1	0.08	0.0707	0.0200	0.0507
1	0.16	0.0707	0.0283	0.0424
1	0.16	0.0707	0.0283	0.0424
0	0.02	0	0.0100	-0.0100

Los valores de los residuales se situan todos ellos dentro del intervalo fijado por $\pm 1/\sqrt{W_i}$, ya que:

$$1/\sqrt{W_i} = 1/\sqrt{200} = 0.0707$$

$$-0.0707 < r_i < 0.0707$$

Así pues, podemos aceptar la normalidad de la distribución puesto que todos los residuales del diagrama suspendido no son distintos de cero. A tal conclusión ya habiamos llegado con el análisis de los DRR; simplemente se plantea una nueva estrategia en el marco de las técnicas E.D.A.

El estudio de estos residuales puede verse completado con la representación gráfica de los mismos, estableciendo las bandas de confianzas que delimitan la aceptación de la normalidad. La siguiente figura muestra este tratamiento con respecto a los residuales.

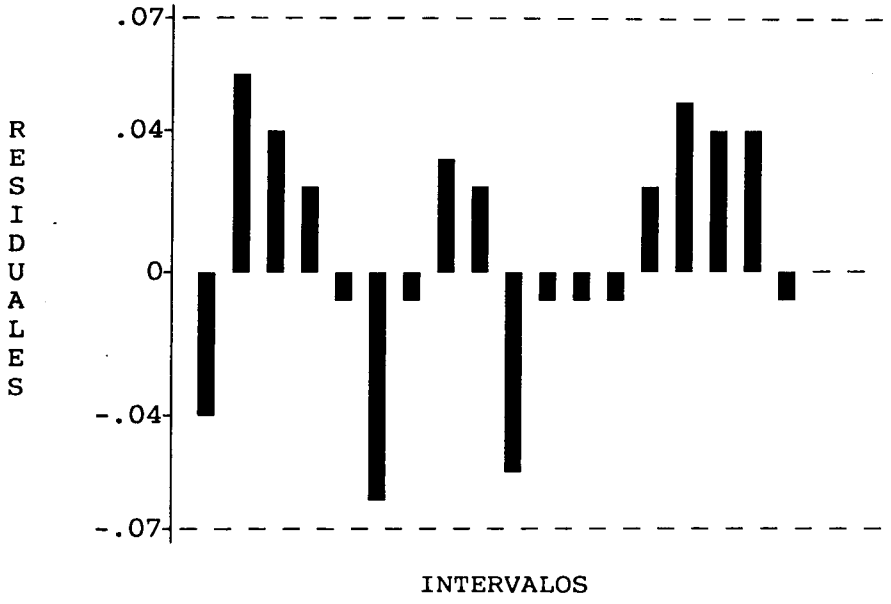


Fig. 2.20. Análisis gráfico de los residuales del diagrama suspendido

La linea discontinua muestra las bandas de confianza en la que consideramos los residuales como no diferentes de 0. Como se observa, todos los residuales cumplen la condición necesaria para la aceptación de la hipótesis nula.

De esta forma presentamos las distintas utilidades y características de los diagramas de raiz cuadrada en lo que se refiere a la descripción univariable y a las estrategias de bondad de ajuste insertas en el mismo ámbito de actuación exploratorio.

3. LÍNEA RESISTENTE

3.1. INTRODUCCIÓN

En la mayoría de los ámbitos cientificos, y las Ciencias Sociales no son una excepción, el estudio de la relación entre variables se ha abordado desde una perspectiva estadística especialmente rígida. De hecho, a nadie extraña encontrar el análisis de la relación entre dos variables cuantitativas mediante el ajuste de un modelo de regresión lineal. No se trata aqui de poner en tela de juicio tal proceder sino el de marcar y señalar el peligro que ello implica puesto que en la mayoría de los casos, los datos aportados no ofrecen información de algunos de los puntos importantes en todo ajuste lineal. Por ejemplo, no es frecuente encontrar información acerca del comportamiento de los residuales ni, como mínimo, evaluar gráficamente la viabilidad de la linealidad en la relación para ser modelizada de esta forma.

Como mucho, se facilitan datos acerca del ajuste del modelo propuesto y de su utilidad ya sea descriptiva o predictiva. Por ejemplo, no es infrecuente, tratar de ajustar un modelo lineal a la nube de puntos planteada en la figura número 3.1.

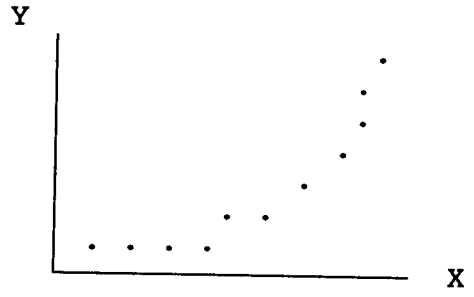


Fig. 3.1. Nube de puntos simulada

Los datos que se obtienen del ajuste del modelo de regresión simple con los puntajes de la figura 3.1 muestran un valor de F=43.65 (P=0.0002); con lo cual no es difícil que en la mayoría de ocasiones se concluyera con la aceptación de la linealidad como modelo de relación entre las variables y, lo que puede ser aún más comprometido, establecer la inferencia poblacional en base al mismo criterio.

Debemos aceptar como necesario el hecho de la exploración como paso previo a muchos de nuestros análisis y comprometer nuestros resultados con informaciones de carácter teórico que permitan efectuar afirmaciones más allá del mero ajuste estadístico de modelos. Lo que nos proponemos abordar en este capítulo supone la propuesta que el Análisis Exploratorio de Datos ha efectuado por lo que se refiere a la exploración de la relación entre dos variables cuantitativas y sobre los mecanismos para obtener información empírica, al margen de cualquier supuesto confirmatorio, para llegar a la decisión de la existencia o no de relación lineal entre las variables. No se trata, pues, de plantear una técnica alternativa al modelo de la regresión simple; a pesar de que podría desempeñar ese papel en algunas ocasiones, sino la de asegurar que la técnica confirmatoria será aplicada dentro de la no vulneración del supuesto de linealidad entre las variables.

Asi justificada, trataremos de plasmar en los siguientes apartados las características fundamentales y el desarrollo de lo que se ha denominado

Línea Resistente o Línea de Tukey. En términos descriptivos generales se trata de un suavizador lineal resistente. Es una línea recta obtenida de la relación de las medianas cruzadas del primer y último tercio de los casos a lo largo de los valores de X (como variable independiente). Veamos como se plantean sus distintos elementos y usos.

3.2. ASPECTOS GENERALES DE LA LÍNEA RESISTENTE BIVARIABLE

Tukey (1977) muestra, siguiendo el esquema clásico para ello, la relación que se desprende del ajuste de una recta a la nube de puntos bivariable. La siguiente gráfica bastará para poner de manifiesto este aspecto.

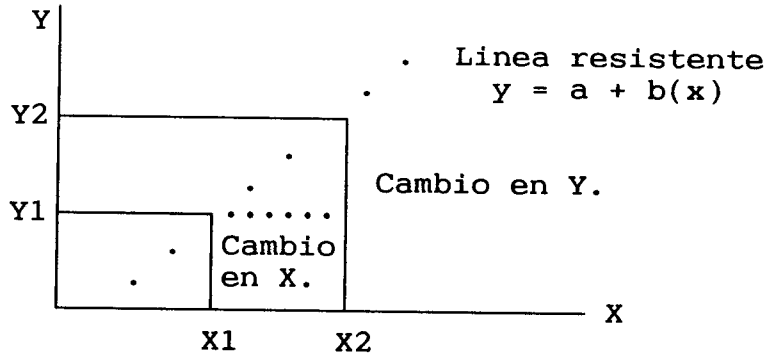


Fig. 3.2. Ajuste lineal

La expresión general que define esa relación se puede plantear del siguiente modo:

$$Y = b(X) \tag{3.1}$$

siendo b, como siempre, el coeficiente (pendiente) que determina el cambio que se da en Y por el cambio en una unidad en X.

En efecto, la evidencia empírica que se obtiene mediante el valor de "b", se relaciona con el tipo de función que se establece en la nube. De este modo, si definimos dos pares cualesquiera $[(Y_i, X_i); (Y_j, X_j)]$ de puntos de la nube podemos establecer dos incrementos en los rangos de Y e X respectivamente:

$$\Delta Y = (Y_j - Y_i) \qquad (3.2.1)$$

$$\Delta X = (X_j - X_i) \qquad (3.2.2)$$

De forma esquemática, podemos establecer que el comportamiento de esos incrementos permite determinar, de manera más exhaustiva, el tipo de función concreta que caracteriza la nube de puntos. De esta forma se pueden diferenciar dos funciones primitivas:

a) Función estrictamente creciente en un punto:

Se caracterizan, básicamente, por la igualdad en los signos de ambos incrementos. Con ello, se puede establecer que:

$$\text{Si } \Delta X > 0 \longrightarrow \Delta Y > 0 \qquad (3.3.1)$$

$$\text{Si } \Delta X < 0 \longrightarrow \Delta Y < 0 \qquad (3.3.2)$$

Dada esta igualdad en los signos de los incrementos, es evidente que:

$$(\Delta Y / \Delta X) > 0 \qquad (3.4)$$

Este tipo de funciones se pueden representar según la siguiente figura:

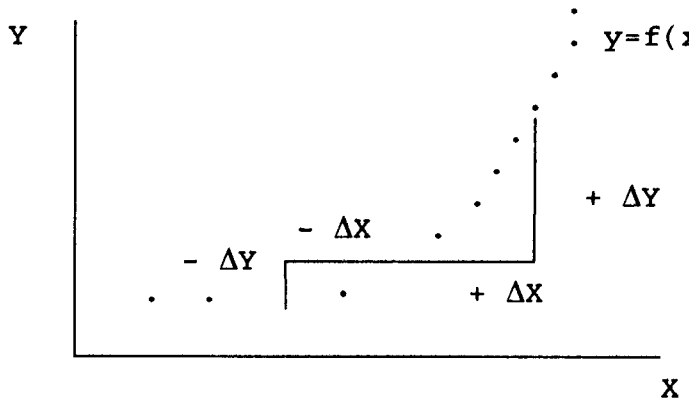


Fig. 3.3. Función estrictamente creciente en un punto.

b) Función estrictamente decreciente en un punto:

En este caso, los signos de los incrementos no se matienen iguales entre Y e X, es decir, siguiendo con la formulación anterior, podemos establecer que:

$$\text{Si } \Delta X > 0 \longrightarrow \Delta Y < 0 \qquad (3.5.1)$$

$$\text{Si } \Delta X < 0 \longrightarrow \Delta Y > 0 \qquad (3.5.2)$$

En consecuencia, la razón entre incrementos adopta siempre la siguiente estructura:

$$(\Delta Y / \Delta X) < 0 \qquad (3.6)$$

Gráficamente, se puede representar como en la siguiente figura:

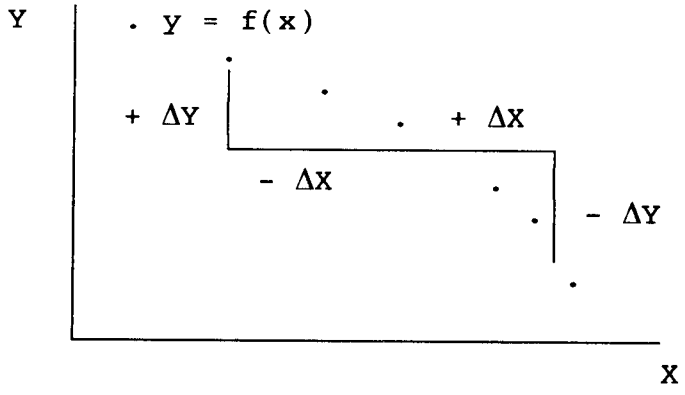


Fig.3.4. Función estrictamente decreciente en un punto.

Una vez revisados estos aspectos generales es fácil derivar de ello que la expresión $Y = b(X)$ puede expresarse en términos de incrementos (o decrementos) en el dominio de la función:

$$Y - Y1 = b(X - X_1) \qquad (3.7.1)$$

o si se prefiere

$$Y = Y_1 + b(X - X_1) \qquad (3.7.2)$$

En esa expresión se mantiene la idea inicial de la figura 3.2, puesto que en el caso de que $(X = X_1)$ tiene como consecuencia que

$(Y = Y_1)$. Igualmente, Tukey aisla, como es lógico, la expresión de cálculo de "b" a partir de esa expresión:

$$b = (Y - Y_1)/(X - X_1) \tag{3.8}$$

Tal formulación se mantiene sea cual sea el par de puntos que se seleccionaran. Por ejemplo, por seguir con la notación de la figura 3.2, si $(X = X_2)$ lleva a concretar que $(Y = Y_2)$. En efecto, la expresión se plantearía del siguiente modo:

$$(Y - Y_1) = b(X_2 - X_1) \tag{3.9}$$

Tal como se ha definido "b" en esta última expresión podría subtituirse por:

$$b = (Y_2 - Y_1)/(X_2 - X_1) \tag{3.10}$$

y tal definición substituida en la expresión general sirve para poner de manifiesto la igualdad descrita anteriormente:

$$\begin{aligned} (Y - Y_1) &= [(Y_2 - Y_1)/(X_2 - X_1)] \cdot (X_2 - X_1) \\ (Y - Y_1) &= (Y_2 - Y_1) \\ Y &= Y_2 \end{aligned} \tag{3.11}$$

Por insistir en lo ya planteado, ésto no supone ninguna novedad especial. El problema real reside en la selección de los valores que determinan la recta final. El criterio de los mínimos cuadrados propone un criterio en la selección de esos puntos; pero lo que aqui plantearemos será la propuesta de Tukey de establecer esa recta a partir de los puntos de la nube determinados por los valores de las medianas de cada uno de los grupos que estableceremos. Recordemos que la expresión general de la recta que perseguimos establecer adopta la forma clásica:

$$Y = a + bX \tag{3.12}$$

y que el valor de "a" representa el valor de Y cuando X=0 y el valor de "b" supone la ya comentada pendiente de la recta.

3.2.1. CÁLCULO DE LOS COEFICIENTES DE LA LÍNEA RESISTENTE

Siguiendo las características generales que presentan las técnicas E.D.A., la línea resistente está intimamente ligada con los estadísticos resistentes. En concreto, el uso de la Mediana será el elemento fundamental a tal efecto. Para la obtención de los valores de "a" y de "b", será necesario seguir los pasos que a continuación se detallan:

1.– División en tercios del rango de X.

Como primera actuación debemos establecer tres tercios en el rango de X, ordenando previamente los valores de X. Cada uno de los tercios establecidos deben contener, en lo posible, el mismo número de puntos de la nube inicial. La figura siguiente muestra está partición en una nube de puntos simulada:

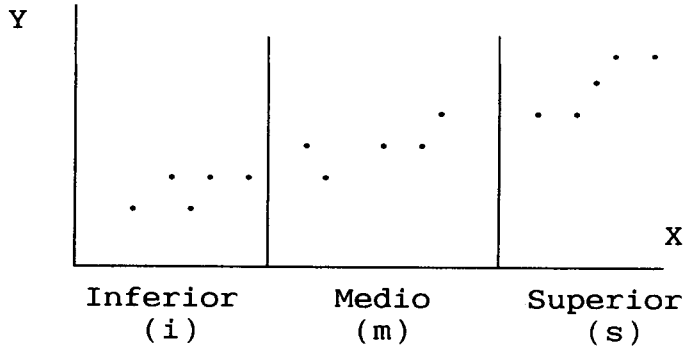


Fig. 3.5. División en tercios del rango de X.

Estos tercios en X es importante que cumplan unas determinadas concidiones que maximizan el establecimiento de la línea resistente. Podemos destacar:

a) Los tercios extremos (inferior y superior) deberan contener al menos una tercera parte de los puntos de la nube. El tercio central es el menos relevante.

b) Cada tercio extremo tendrá un rango en X que sea menor de la mitad del rango total de X.

c) Cuando se den varios puntos en el mismo valor de X, serán tratados en el mismo tercio.

d) Deberemos tender a trabajar con el máximo número de puntos en los tercios extremos.

Distribuir los valores de X en tres tercios puede, en función del número de puntos de la nube, ofrecer tercios de distinto tamaño, es decir, tercios no equilibrados con respecto al número de puntos que incluyen. Velleman y Hoaglin (1981) ofrecen el siguiente cuadro guia a este respecto:

Cuadro 3.1. Distribución de los tamaños de los tercios según el tamaño de la muestra (Velleman y Hoaglin, 1981).

TERCIOS	Si $n = 3k$	Si $n = 3k + 1$	Si $n = 3k + 2$
Inferior	k	k	$k + 1$
Medio	k	$k + 1$	k
Superior	k	k	$k + 1$

2.– Cálculo de los Puntos resumen

Por **punto resumen** se entiende la mediana, tanto en Y como en X para cada uno de los tercios. De forma que, obtenidos esos valores, disponemos de seis valores resumen en base a las medianas halladas:

Tercio Inferior	X_i	Y_i
Tercio Medio	X_m	Y_m
Tercio Superior	X_s	Y_s

siendo "X_j" la mediana de los valores de X en el tercio "j" e "Y_j" la mediana de los valores de Y incluidos en el tercio "j". Gráficamente podría adoptar un aspecto parecido al siguiente:

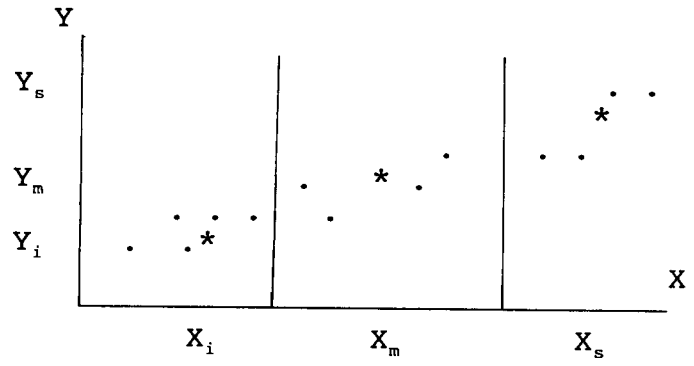


Fig. 3.6. Puntos (*) que representan los resumenes efectuados a partir de las Medianas de cada uno de los tercios.

En este proceso debe tenerse en cuenta que los valores del resumen no tienen por que coincidir con valores de la muestra evaluada, y que pueden cometerse algunos errores de selección que son graves en lo que se refiere a la obtención de la línea resistente. Por ejemplo,

a) Se ordenan los valores según Y en lugar de hacerlo con X.

b) Se obtienen los puntos resumen de forma conjunta . No tercio a tercio.

c) Se selecciona un valor de Y_m que mantiene una relación lineal entre Y_i e Y_s.

Una vez establecidos esos seis valores, solo resta determinar los valores de la constante "a" y de la pendiente de la recta "b" siguiendo para ello las siguientes expresiones:

$$b = (Y_s - Y_i)/(X_s - X_i) \qquad (3.13)$$

$$a = 1/3(a_i + a_m + a_s) \qquad (3.14)$$

siendo $a_i = Y_i - bX_i$
$a_m = Y_m - bX_m$
$a_s = Y_s - bX_s$

De este modo establecemos los coeficientes que integran la expresión general de la línea resistente derivada de la nube de puntos inicial.

Como es obvio, tal estrategia genera, a su vez, residuales en base a la diferencia entre los valores de Y y aquellos que ofrece la línea resistente. De este modo, los valores que surgen de $R = Y - (a + bX)$ deben ser objeto de un estudio exhaustivo que abordaremos con más profundidad en los apartados siguientes. Tales valores son, pues, determinados de forma clásica, del mismo modo que en modelos lineales confirmatorios, tal como se muestra en la siguiente figura:

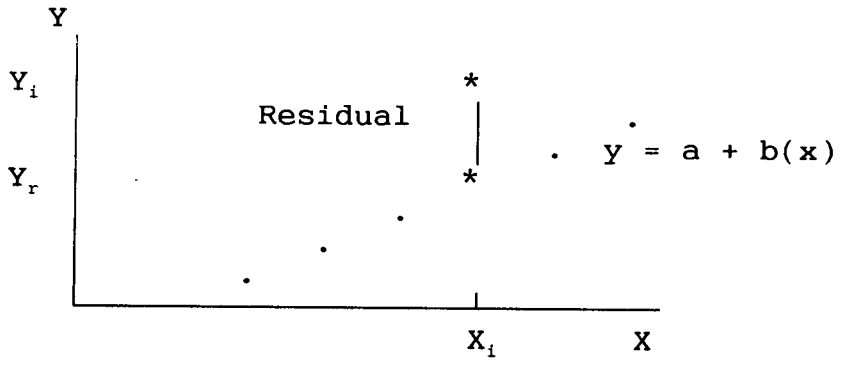


Fig. 3.7. Representación gráfica del residual entre el valor original de $Y(Y_i)$ y el determinado por la línea resistente (Y_r).

3.2.2. OBTENCIÓN DE LA LÍNEA RESISTENTE EN UNA MUESTRA

Veamos como el proceso anterior se aplica a unos datos concretos. En una muestra de 18 sujetos, se han evaluado sus rendimientos en una tarea motriz en base al número de ensayos correctos efectuados por los sujetos. Se pretende estudiar la relación de este rendimiento con el tiempo en

minutos que se dedicó a ensayar este tipo de tareas. De este modo, la nube de puntos que se obtiene es la siguiente:

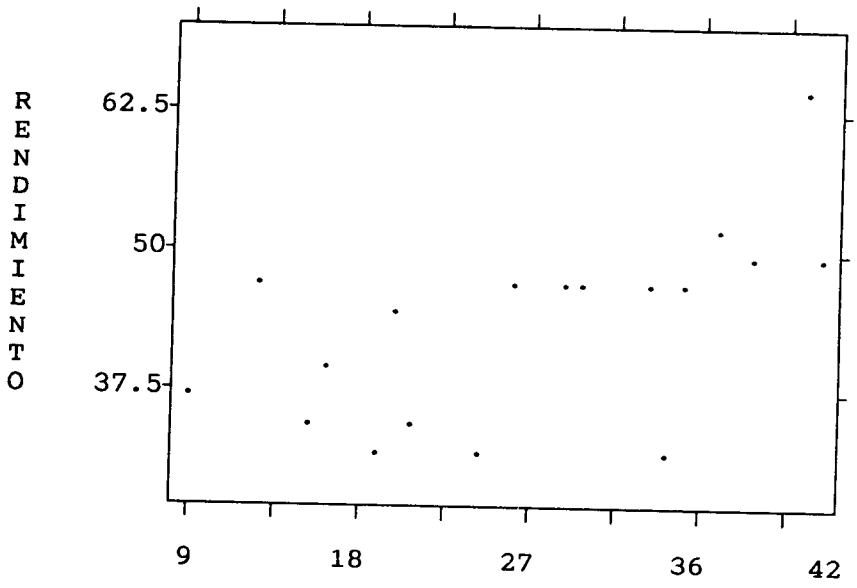


Fig. 3.8. Nube de puntos de la muestra de 18 sujetos.

Con objeto de presentar paso a paso los cálculos necesarios para el establecimiento de la línea resistente presentamos a continuación la tabla de datos directos, ya ordenados según los valores de X y separados por cada uno de los tercios (cada tercio con un total de 6 puntos):

Tabla núm. 3.1. Datos recogidos en una muestra de 18 sujetos.

SUJETO	MINUTOS ENSAYO (X)	RENDIMIENTO (Y)	TERCIO
1	9	37	
2	13	47	
3	15	36	Inferior
4	16	40	
5	19	32	
6	20	45	
.....			

.../...

.../...

SUJETO	MINUTOS ENSAYO (X)	RENDIMIENTO (Y)	TERCIO
7	21	35	
8	24	33	
9	26	48	Medio
10	29	48	
11	30	47	
12	33	48	
.....			
13	34	33	
14	35	48	
15	37	52	Superior
16	39	50	
17	41	65	
18	42	49	

El cálculo, de acuerdo con lo visto anteriormente (Tukey, 1977), de los puntos resumen muestra los siguientes valores:

Tercio Inferior	Tercio Medio	Tercio Superior
X_i= 15.5	X_m = 27.5	X_s = 38.0
Y_i = 38.5	Y_m = 47.5	Y_s = 49.5

Estos valores determinan, como hemos señalado anteriormente, los puntos necesarios para el establecimiento del valor de "a" y de "b". Tales puntos quedan reflejados en la siguiente figura:

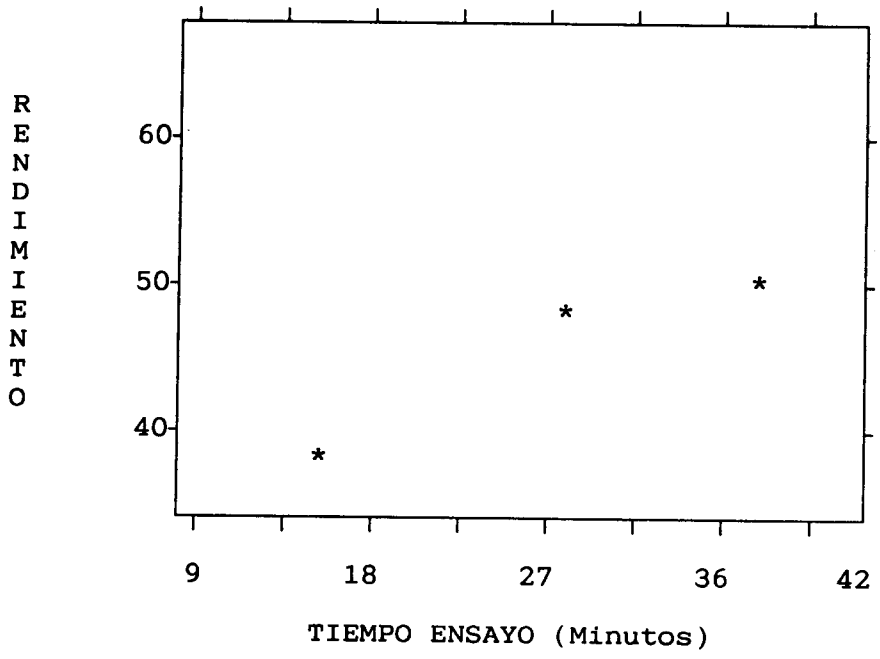


Figura 3.9. Puntos definidos por las Medianas en cada uno de los tercios.

Aplicando las expresiones correspondientes para el cálculo de "a" y de "b" obtendremos los siguientes valores:

$$b = (49.5 - 38.5)/(38.0 - 15.5) = 0.49$$

$$\begin{aligned} a_i &= 38.5 - (0.49 \cdot 15.5) = 30.91 \\ a_m &= 47.5 - (0.49 \cdot 27.5) = 34.03 \\ a_s &= 49.5 - (0.49 \cdot 38.0) = 30.88 \end{aligned}$$

$$a = 1/3(30.91 + 34.03 + 30.88) = 31.94$$

En consecuencia, los valores de los componentes de la recta la definen del siguiente modo:

$$Y = 31.94 + 0.49(X) \qquad (3.15)$$

Esta expresión, pues, constituye la línea resistente ajustada a la nube de puntos presentada anteriormente. Como ya se ha comentado su utilidad es eminentemente exploratória y su papel en un análisis de datos es el de evaluar la posibilidad del ajuste lineal confirmatorio. Esa posibilidad depende, evidentemente, de la cuantía de los residuales que genera la

línea resistente hallada. Estos residuales definidos por:

$$r_i = Y_i - (a + bX_i) \tag{3.16}$$

aportan una evaluación empírica del ajuste de un futuro modelo lineal. El tratamiento de los residuales será objeto del siguiente apartado.

3.2.3. ANÁLISIS DE LOS RESIDUALES

Como se ha comentado con anterioridad, la evaluación de los **residuales** nos aporta información acerca del ajuste de la línea reistente y del posible ajuste de un modelo lineal confirmatorio. A este respecto, los residuales, definidos según la expresión anterior, pueden ser conceptualizados mediante diferentes acepciones:

a) Como diferencias del ajuste entre X e Y.

b) Parte de Y no explicada por X.

c) Variación de Y con efectos de X no controlados.

d) Diferencia entre el valor de Y y la predicción de la ecuación.

Sea cual sea, la utilidad última a la que se destine el residual, no hay dudas de que su evaluación reporta una información especialmente relevante. Debe aclararse que, a diferencia de los modelos confirmatorios, la línea resistente no utiliza los residuales como vehículo propicio para la evaluación de los supuestos del modelo; puesto que en nuestro caso no existen supuestos a evaluar. De forma general, el residual permitirá, tal como veremos a continuación, una mejor aproximación de carácter exploratorio a la nube de puntos bivariable.

La primera aproximación al estudio de los residuales, y totalmente coherente con el espíritu general de las técnicas E.D.A, se centra en la evaluación gráfica; la cual nos lleva al análisis de las posibles tendencias, errores sistemáticos o sesgos en algún o algunos grupos determinados de sujetos muestrales. Por ejemplo en nuestros datos, la relación de residuales adoptará los siguientes valores:

Tabla núm. 3.2. Relación de datos y residuales

SUJETO	MINUTOS ENSAYO (X)	RENDIMIENTO (Y)	RESIDUAL
1	9	37	.65
2	13	47	8.69
3	15	36	-3.29
4	16	40	.22
5	19	32	-9.25
6	20	45	3.26
.....			
7	21	35	-7.23
8	24	33	-10.70
9	26	48	3.32
10	29	48	1.85
11	30	47	.36
12	33	48	-.11
.....			
13	34	33	-15.60
14	35	48	-1.09
15	37	52	1.93
16	39	50	-1.05
17	41	65	12.97
18	42	49	-3.52

Una de las formas más simples de evaluación gráfica se puede articular mediante los gráficos de caja y de Tronco y Hojas. En nuestro caso, sus formas serían las siguientes:

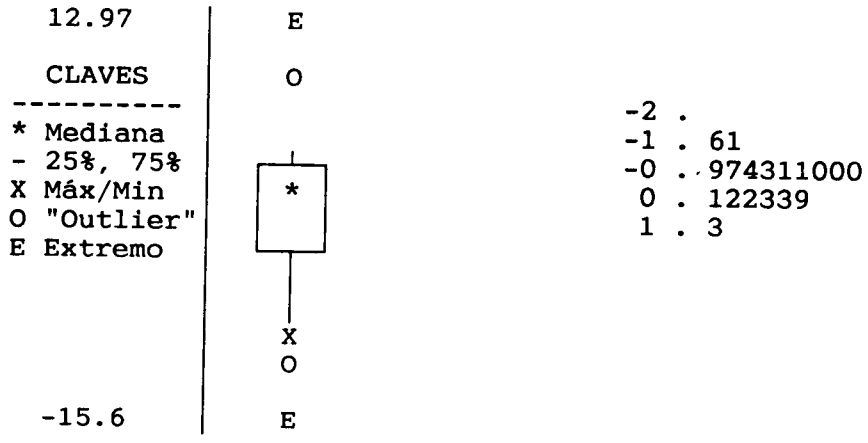


Fig. 3.10 . Diagrama de Caja y Diagrama de Tronco y Hojas de los residuales de la línea resistente.

No es difícil observar la existencia de unos residuales altos en la cola de abajo de las distribución, pues el diagrama de caja así lo detecta. Del mismo modo, puede pensarse, a partir del diagrama de Tronco y Hojas, en una posible "normalidad" del residual, con vistas al tratamiento confirmatorio de los datos.

Con respecto a la existencia de valores extremos y "outliers" de acuerdo con el diagrama de caja, el estudio de la relación de la tabla número 3.2 por lo que se refiere a los valores residuales, se detecta la existencia de residuales elevados en los sujetos 17 y 13, respectivamente "outlier" y extremo. Recordemos que el carácter resistente de la técnica que nos ocupa, hace propicio pensar que, aunque se eliminarán del análisis esos dos sujetos, los valores de los coeficientes de la línea resistente hallada no variarían sustancialmente. En efecto, si rehacemos nuestros análisis eliminando esos dos sujetos, con lo cual dispondríamos de un muestra de 16 sujetos, obtendríamos un línea resistente con la siguiente expresión:

$$Y = 30.31 + 0.54(X) \qquad (3.17)$$

Es fácil comprobar que no existe apenas diferéncias entre esta última linea resistente con la que hemos establecido en primera instáncia (Y = 31.94 + 0.49·X). Por otro lado, la evaluación gráfica de la distribución

de los residuales puede efectuarse mediante la construcción de la nube de puntos entre el valor del residual y los valores de X. Por ejemplo, en nuestro caso tal representación adoptaría la forma que se presenta en la figura 3.11:

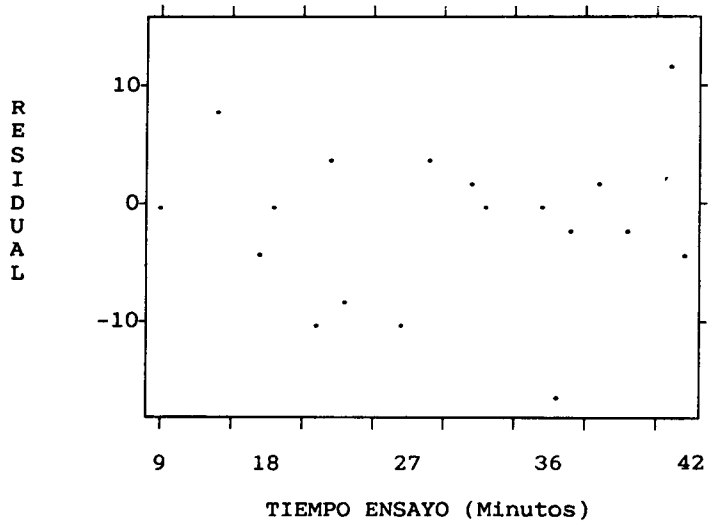


Fig.3.11. Nube de puntos entre los residuales y los valores de X.

Es muy fácil localizar en esa nube de puntos a los dos sujetos (13 y 17) que muestran residuales altos. Del mismo modo, esta representación permite la evaluación de la existencia de algún tipo de tendencia o error sistemático que permita preveer dificultades de este estilo en el comportamiento de los residuales en el modelo confirmatorio.

Johnstone y Velleman (1982) proponen un esquema mínimo para la evaluación de los residuales de la línea resistente, consistente en su empleo para los siguientes objetivos:

a) Analizar gráficamente los residuales por distintas zonas del rango de X, para detectar aspectos parciales del desajuste.

b) Evaluar los signos que presenten como mecanismo de análisis de error no aleatorio.

c) Realizar ajustes resistentes por zonas para detectar comportamientos peculiares de la muestra con respecto a la posible relación entre las dos variables originales.

Se desprende de lo anterior la necesidad de evaluar los residuales en todas y cada una de las posibles fuentes de variabilidad en los mismos. Por ejemplo, supongamos que los sujetos impares de nuestra muestra fueran de sexo masculino y los pares de sexo femenino. No sería difícil establecer la misma evaluación gráfica de la figura 3.11 incorporando esta información y así poder detectar si la agrupación de los sujetos mediante la variable sexo supondría un elemento de distorsión en el posible ajuste posterior. La figura a la que nos referimos sería la siguiente:

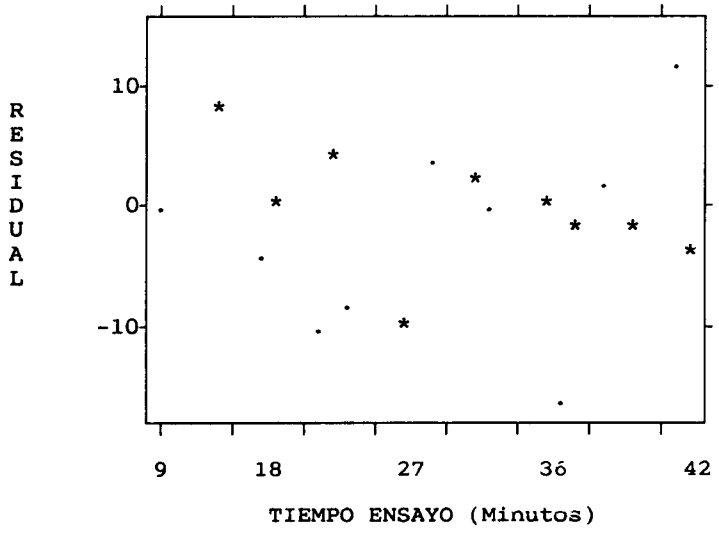


Fig. 3.12. Nube de puntos entre los residuales y los valores de X, diferenciando entre Hombres (.) y Mujeres (*).

Esta gráfica muestra que la incorporación de esta información no permite establecer algún efecto especial, de forma que la variable controlada (sexo) no aporta ningún dato relevante en el comportamiento del residual.

Hasta este momento, hemos revisado el establecimiento de la línea resistente y un repaso a algunas propuestas gráficas de evaluación de los residuales. Sin embargo, parece lógico pensar en la posibilidad de utilizar la información de que se dispone para intentar minimizar el posible desajuste. No se trata de establecer una estrategia que minimize el residual al uso de los mínimos cuadrados ordinarios, sino de intentar ajustar

al máximo los valores de los coeficientes que integran la línea resistente para conseguir dos objetivos básicos:

a) Obtener los valores de "a" y "b" con menor residual.

b) Establecer a partir de ese valor de "b" un indicador del posible ajuste de la línea resistente.

Como sea que la línea resistente es un paso exploratorio previo a la confirmación; ¿Por que no explorar también el ajuste?.

Veamos en el apartado siguiente alguna de las propuestas que a este respecto se han efectuado.

3.2.4. UTILIZACIÓN DE LOS RESIDUALES PARA UN MEJOR AJUSTE DE LA LÍNEA RESISTENTE

En este apartado, como ya se ha comentado, abordaremos una utilización avanzada del residual y su vinculación con los valores de "a" y "b" en la línea resistente.

Como primer paso a considerar, debemos aportar aqui la propuesta de Johnstone y Velleman (1982) relativa a un **índice exploratorio de ajuste**, denominado "δ" y definido por la siguiente expresión:

$$\delta = dq(\text{residuales})/dq(Y) \qquad (3.18)$$

siendo "dq" la distancia entre cuartos de cada uno de los valores referidos en las expresiones (residuales y valores de Y). Este índice fluctua entre 0 y 1 y, evidentemente, si $\delta \approx 0$ se puede pensar en un ajuste adecuado y si $\delta \approx 1$, el ajuste es incorrecto. Por seguir con el ejemplo propuesto anteriormente, nuestra línea resistente muestra un ajuste de $\delta = 0.54$. La propuesta descrita no pasa, por supuesto, por la significación de este valor (entre otras cosas por que se desconece su función de densidad); simplemente se trata de obtener un descriptor del concepto abstracto del "ajuste" de la línea resistente. En nuestro caso, el valor hallado se encuentra en la zona intermedia, lo que permitiría pensar en un ajuste

moderado. Recuérdese la existencia de un par de residuales muy altos (sujetos 13 y 17).

Pero la existencia del residual merece una consideración mayor y una utilización más exhaustiva de la que aqui se ha descrito. Es por ello que abordamos a continuación un par de propuestas de cálculo de "a" y de "b" que incorporan el residual como parte importante del proceso.

En primer lugar planteamos la corrección en el cálculo de "a" y de "b" definida por Velleman y Hoaglin (1981). Estos autores presentan un proceso en el cual se ajusta una línea resistente entre los valores originales de X y los residuales que se obtienen despues de hallar "a" y "b" tal como se ha descrito con anterioridad. En caso de que la pendiente de esta nueva línea resistente (a la que denominaremos b') sea cercana a 0 permitirá aceptar el valor de "b" inicial, puesto que ello significaría la no existencia de relación entre los valores de X originales y los residuales hallados (Véase la figura 3.11 con la que esta propuesta mantiene claras vinculaciones). En caso contrario, es decir que b' sea claramente distinta de 0, se corrige el valor de la pendiente original, sumándole el valor de la pendiente hallada entre X y los residuales (b'). En términos más formales, si b_1 es la pendiente de la línea resistente original y b' es la pendiente de la línea resistente entre los valores de X y los residuales generados a partir de b_1, podemos esquematizar este planteamiento del siguiente modo:

Si $b' \approx 0$ b_1 puede aceptarse como ajustada
Si $b' \not\approx 0$ se establece una nueva pendiente (b_2)
definida por $b_2 = b_1 + b'$

Si se ha llegado a establecer una nueva pendiente (b_2) será necesario repetir este proceso hasta llegar a un valor de (b_i') que muestre unos residuales que sean totalmente independientes de los valores de X. Es decir, debemos repetir este proceso hasta encontrar un valor que cumpla la condición establecida que supone que $(b_i') \approx 0$.

Puede darse el caso de que en algún paso de este proceso se determine un valor positivo de b_i y en el siguiente ese valor sea negativo, es decir que $b_{(i+1)}$ sea negativo. Ello supone determinar que entre esos dos valores se encuentra el valor de "b" que mejor ajusta a la nube. En tal caso, se

propone estimar ese valor de "b" mediante la siguiente expresión:

$$b_{(i+1)} = b_i - \left\{ b'_i \cdot [(b_i - b_{(i-1)})/(b'_i - b'_{(i-1)})] \right\} \qquad (3.19)$$

Por lo que se refiere al valor de "a" se determina su constancia en todos los pasos, es decir se escoge la solución inicial como única. De modo que esta corrección sólo influye en la estimación del valor de "b".

Veamos como se aplicaría este procedimiento en nuestro datos. En la tabla 3.2 disponemos de los valores de X y de los residuales generados por la línea resistente hallada que, recordemos, adopta la expresión:

$$Y = 31.94 + 0.49(X) \qquad (3.20)$$

Con esas dos series de datos podemos establecer, de acuerdo con el procedimiento general de estimación de la línea resistente, los siguientes valores resumen de cada tercio (recuérdese que se tratan de medianas de cada tercio, una vez ordenados los valores según X):

	Tercio Inferior	Tercio Medio	Tercio Superior
Valor X	$X_i = 15.5$	$X_m = 27.5$	$X_s = 38.0$
Residual	$R_i = 0.435$	$R_m = 0.125$	$R_s = -1.07$

No debe sorprender encontar en el tercio superior un valor negativo de mediana del residual, puesto que el orden lo determina X no los residuales, y en consecuencia es factible hallar una mediana en el tercer tercio menor que en el primer o segundo tercio. Con esos datos podemos establecer la línea resistente del primer paso entre X y R:

$$b'_1 = (R_s - R_i)/(X_s - X_i) = -0.067 \qquad (3.21)$$

Con este valor de b'_1 muy cercano a 0, podemos pensar en aceptar la primera solución ($b_1 = 0.49$) como adecuada. En caso de desear un mayor ajuste, corregimos el valor de b de acuerdo con la siguiente fórmula:

$$b_2 = b_1 + b'_1 = 0.49 - 0.067 = 0.423 \qquad (3.22)$$

De este forma, la nueva línea resistente después de este primer paso quedaría con los siguientes valores (recuérdese que el valor de "a" no se ve afectado por esta corrección):

$$Y = 31.94 + 0.423(X) \qquad (3.23)$$

Deberemos repetir el proceso anterior con la nueva línea resistente. Veamos que resultados muestra este procedimiento aplicado a la nueva línea resistente hallada. Con estos valores de "b" y "a" se consiguen unos residuales distintos, obviamente, a los que se detallan en la tabla 3.2 y que a continuación relacionamos (Tabla 3.3)

Tabla 3.3. Relación de residuales de la línea resistente $Y = 31.94 + 0.423(X)$.

Tercio Inferior	Tercio Medio	Tercio Superior
1.25	-5.82	-13.32
9.56	-9.09	1.25
-2.29	5.06	4.41
1.29	3.79	1.56
-7.98	2.37	15.72
4.60	2.10	-.71

En las figuras 3.13 y 3.14 se muestra la exploración gráfica de estos residuales, de forma análoga a la ya planteada en las figuras 3.10 y 3.11. No es necesario efectuar comentario alguno respecto a este punto puesto que el comportamiento de los residuales es prácticamente igual al encontrado en el paso inicial.

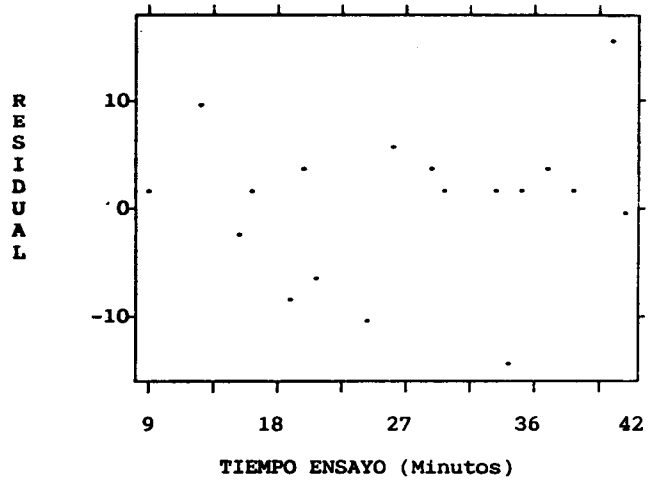


Figura 3.13. Nube de puntos entre los valores de X y los residuales de la segunda línea resistente.

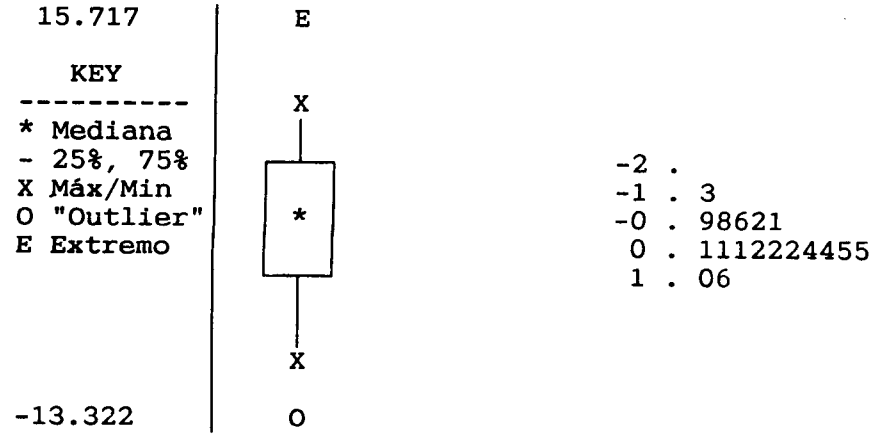


Figura 3.14. Diagrama de caja y Diagrama de Tronco y Hojas de los residuales de la segunda línea resistente.

Con los datos de la tabla 3.3, podemos establecer los siguientes valores resumen:

	Tercio Inferior	Tercio Medio	Tercio Superior
Valor X	X_i = 15.5	X_m = 27.5	X_s = 38.0
Residual	R_i = 1.273	R_m = 2.235	Rs = 1.409

En virtud de esos valores, podemos establecer un nuevo valor de b', en este caso

$$b'_2 = (R_s - R_i)/(X_s - X_i) = 0.006 \qquad (3.24)$$

que es prácticamente igual a 0 y que reduce este valor con respecto a b'_1 (-0.06). Si aceptamos este segundo paso como más ajustado que el primero, el valor de "b_3" final dependerá de este nuevo valor de b'_2.

Tal como se ha dicho con anterioridad, no podemos establecer el valor de b_3 de forma directa a través de la expresión $(b_1 + b'_2)$ puesto que b'_1 y b'_2 son de signo distinto. Asi pues, deberemos establecer ese valor utilizando para ello la estimación definida anteriormente, que en nuestro caso adoptaría la siguiente forma:

$$b_3 \;=\; b_2 - \{b'_2\,[(b_2 - b_1)/(b'_2 - b'_1)]\} =$$

$$= 0.423 - \{0.006[(0.423 - 0.49)/(0.006 + 0.06)]\} =$$
$$= 0.417 \qquad (3.25)$$

Así, la nueva ecuación que surge de este segundo paso se puede plantear del siguiente modo:

$$Y = 31.94 + 0.417(X) \qquad (3.26)$$

Con esta ecuación, y teniendo en cuenta que $b'_2 \approx 0$, podemos aceptarla como la más ajustada a nuestra nube de puntos inicial, con lo cual podemos definir la siguiente tabla resumen global:

Tabla 3.4. Lineas resistentes halladas en los distintos pasos según la técnica de Velleman y Hoaglin (1981).

PASO	LÍNEA	VALORES DE b'
0	$Y = 31.94 + 0.490(X)$	
1	$Y = 31.94 + 0.423(X)$	$b'_1 = -0.067$
2	$Y = 31.94 + 0.417(X)$	$b'_2 = 0.006$

Una forma rápida de evaluar el efecto en el ajuste de este proceso de corrección puede realizarse mediante el índice "δ" que hemos descrito con anterioridad. Para cada ocasión se obtienen los siguientes valores de "δ":

Primera línea $\delta = 0.53672$
Segunda línea $\delta = 0.51792$
Tercera línea $\delta = 0.51016$

Esos valores muestran un acercamiento a un ajuste estable de este indicador y, además, indica que entre la segunda y tercera línea no se dan excesivas mejoras por lo que a los residuales se refiere.

Esta estrategia de corrección intenta explorar la relación entre las dos variables de forma mejorada con respecto al cálculo inicial de "a" y "b".

Como segunda propuesta de obtención de la línea resistente presentaremos la que plantean Emerson y Hoaglin (1985) estableciendo un cálculo iterativo de "a" y "b" a partir, al igual que en el anterior tratamiento, de los valores residuales.

Tal como hemos definido la obtención de "a" y de "b" iniciales; sus valores son consecuencia de lo que podríamos definir como un "ajuste

arbitrario". Por ejemplo, el valor de "a" se genera a partir del valor de Y cuando ($X = 0$). Ello supone un excesivo supuesto si X no adopta jamás el valor 0 o si ese valor es imposible (en el caso de la variable edad se da esta circunstáncia por poner un caso clásico); ya que obviamente el valor ($X = 0$) no pertenece al rango de la variable. Estos autores proponen adoptar un valor de referencia conocido, como la media, la mediana o la mediana de X en el tercio central de la línea resistente, entre otros. Para nuestros propósitos adoptaremos esta última posibilidad, es decir, adoptaremos como punto de referencia el valor de la mediana de X en el tercio central. De este modo estableceremos la ecuación de la línea resistente en los siguiente términos:

$$Y = a + b(X - X_m) \tag{3.27}$$

Se puede destacar claramente el hecho de que tal expresión supone simplemente someter a X a un valor de distáncia respecto al punto de referencia. Esta expresión tiene sólo implicaciones para el establecimiento del valor de "a". Los componentes de la línea resistente quedarían definidos del siguiente modo:

$$b = (Y_s - Y_i)/(X_s - X_i) \tag{3.28}$$

lo cual es ya conocido, mientras que en lo tocante al valor de "a" este queda definido por:

$$a = 1/3\,\{[Y_i - b(X_i - X_m)] + Y_m + [Y_s - b(X_s - X_m)]\} \tag{3.29}$$

Toda vez que el valor de la constante viene determinado por el centrado de X con respecto al punto de referencia (en nuestro caso X_m), la consideración del residual obtenido de la línea resistente asi definida, puede incorporar esta situación, definiéndolo del siguiente modo:

$$R = Y - [a + b(X - X_m)] \tag{3.30}$$

Al igual que Velleman y Hoaglin, estos autores proponen la utilización del residual para cálcular "a" y "b" de forma que se obtenga una estabilidad (convergencia) de esos valores, a la vez que se obtenga una corrección del efecto del residual. La diferencia entre ambas propuestas reside básicamente en que la que hemos presentado en primer lugar respeta la estimación inicial de "a" corrigiendo solo "b", mientras la que ahora abordamos corrige ambos valores. El proceso de corrección, al

que nos hemos referido, se efectua mediante el cálculo de dos nuevos componentes, surgidos de la línea resistente entre los valores de X y los residuales. Estos componentes se denominan δ para la pendiente y γ para la constante. Si δ y γ son cercanos a 0, el proceso de iteración puede detenerse y establecer los valores de "a" y de "b" sumando δ i γ respectivamente a "a" y "b". Los valores de δ y γ en un determinado paso "j" del proceso se pueden definir por las siguientes expresiones:

$$\delta_j = (R_{sj} - R_{ij})/(X_s - X_i) \tag{3.31}$$

$$\gamma_j = 1/3 \cdot \{[R_{ij} - \delta_j(X_i - X_m)] + R_{mj} + [R_{sj} - \delta_j(X_s - X_m)]\} \tag{3.32}$$

En esta propuesta, el concepto de convergencia se asimila con el de estabilidad, lo cual encaja perfectamente con la propuesta anterior con la que, ya se ha dicho, mantiene evidentes paralelismos.

Veamos esta estrategia con nuestros datos. La tabla 3.1 muestra los datos originales con los que empezaremos a trabajar. Así, siguiendo las expresiones generales y en base a los valores de los resúmenes de los tercios, obtenemos:

Tercio Inferior	Tercio Medio	Tercio Superior
$X_i = 15.5$	$X_m = 27.5$	$X_s = 38.0$
$Y_i = 38.5$	$Y_m = 47.5$	$Y_s = 49.5$

y aplicando las fórmulas de cálculo de "a" y "b" propuestas:

$$b = (49.5 - 38.5)/(38.0 - 15.5) = 0.49 \tag{3.33}$$

$$\begin{aligned} a &= 1/3\,\{[38.5 - 0.49(15.5 - 27.5)] + 47.5 + \\ &+ [49.5 - 0.49(38.0 - 27.5)]\} = \\ &= 45.41 \end{aligned} \tag{3.34}$$

Así la primera de las lineas resistentes adopta la siguiente expresión:

$$Y = 45.41 + 0.49(X - X_m) \tag{3.35}$$

mostrando los residuales (R) que se detallan en la siguiente tabla:

Tabla 3.4. Relación de residuales de la línea resistente $Y = 45.41 + 0.49(X - X_m)$.

Tercio Inferior	Tercio Medio	Tercio Superior
.66	-7.22	-15.60
8.70	-10.69	-1.08
-3.28	3.33	1.94
.23	1.86	-1.04
-9.24	.37	12.98
3.27	-.10	-3.51

Con estos valores podemos determinar los tres puntos resumen de los residuales del modo tradicional:

$$R_i = 0.440 \qquad R_m = 0.130 \qquad R_s = -1.065$$

y a partir de ellos establecer los valores de δ y γ para el primer paso de la iteración. De esta forma, y siguiendo para ello las fórmulas planteadas, obtenemos que:

$$\delta = (R_s - R_i)/(X_s - X_i) = -0.067 \tag{3.36}$$

$$\gamma = 1/3 \cdot \{[R_i - \delta(X_i - X_m)] + R_m + [R_s - \delta(X_s - X_m)]\} = \\ = -0.13 \tag{3.37}$$

Dado que tanto δ como γ son cercanos a 0, la corrección es mínima:

$$b_2 = b_1 + \delta = 0.49 - 0.067 = 0.423 \tag{3.38}$$

$$a_2 = a_1 + \gamma = 45.41 - 0.13 = 45.28 \tag{3.39}$$

Caso de que esta corrección diera valores de "b_2" y de "a_2" muy distintos de los hallados inicialmente (b_1 y a_1), el proceso debería repetirse hasta que los valores de δ y γ entre un paso y el siguiente fuera cercano a 0. En nuestro caso, no será preciso reiterar, puesto que los valores son prácticamente estables. A este respecto, nótese como el valor de b_2 según Emerson y Hoaglin en nuestros datos coincide con el valor de b_2 hallado según el esquema de Velleman y Hoaglin. No incideremos aqui en el estudio de los residuales puesto que creemos sería reiterativo e innecesario dada la igualdad entre las lineas resistentes encontradas.

Debe reconocerse que esta última proposición es claramente dificultosa por lo lento de los cálculos y por la posibilidad de que el proceso de

convergencia sea oscilante; es decir que δ y γ vayan adoptando a lo largo de las iteraciones valores con signos opuestos. A este respecto los propios autores remiten al proceso de Velleman y Hoaglin como más rápido y menos complejo en su estructura.

3.3. ANTECEDENTES DE LA LÍNEA RESISTENTE

El planteamiento de la línea resistente de Tukey (1977) no supone, en si mismo, una novedad puesto que algunos autores han propuesto con anterioridad estrategias de modelización lineal parecidas. Estas alternativas no surgieron como una solución contraria a la estimación clásica de los mínimos cuadrados, sino que se plantearon, en su momento, para resolver la estimación de los componentes de la ecuación de regresión en aquellos casos en los que las variables tratadas contienen error de medida (Marsh, 1988; Mosteller, Siegel, Trapiro y Youtz, 1985).

Clásicamente, los modelos lineales conciben a las variable X e Y como exentas de error, siendo el modelo general:

$$Y = \alpha + \beta Y \tag{3.40}$$

siendo α y β valores constantes pero desconocidos. La estimación muestral de los mínimos cuadrados admite la existencia de error en la variable criterio (Y), definiéndo tal situación con la expresión:

$$y = Y + v \tag{3.41}$$

donde "v" es error simétrico con $E(v) = 0$ y variáncia igual a $\sigma^2(v)$. Este tipo de planteamientos típicos de la regresión permiten establecer la ecuación general en términos de (Gujarati, 1987):

$$E(y|x) = \alpha + \beta x \tag{3.42}$$

Esta expresión no contempla la existencia de error en X, con lo cual la situación real de error en ambas variables no se contempla. De hecho, los

mínimos cuadrados pueden modificarse para aceptar esa cuestión (Jonnstone y Velleman, 1982). Si, por ejemplo, se asume el error en X, al igual que en Y, podemos definir que

$$x = X + u \tag{3.43}$$

con $E(u) = 0$ y variáncia del error definida por $\sigma^2(u)$. La expresión general de los mínimos cuadrados minimiza la siguiente formulación:

$$\Sigma\,[y - (a + b \cdot x)]^2 \tag{3.44}$$

pudiéndose replantear para solventar el error de medida en ambas variables, incorporando a esa expresión las variáncias de los errores (Berry, 1984):

$$\Sigma\left\{\left[y - (a + b \cdot x)^2/\sigma^2(v)\right)\right] + \left[(x - x)^2/\sigma^2(u)\right]\right\} \tag{3.45}$$

Sin embargo, la solución a esta cuestión, en la actualidad ampliamente utilizada, provino de las propuestas efectuadas en el mismo sentido de la línea resistente de Tukey. Por ejemplo, Wald (1940) divide el rango de X en dos grupos, definiendo "a" y "b" del siguiente modo:

$$\begin{aligned} b &= [(Y_{m+1} + \cdots + Y_n) - (Y_1 + \cdots + Y_m)] \,/\, [(X_{m+1} + \cdots + \\ &+ X_n) - (X_1 + \cdots + Xm)] \qquad (3.46) \\ a &= \overline{y} - b(\overline{x}) \qquad (3.47) \end{aligned}$$

Como se desprende de lo anterior, la estrategia es muy parecida a la que aqui hemos presentado, con la salvedad de que se divide el rango de X en dos grupos en lugar de tres. Muy parecida a la propuesta de Wald, se dispone de la estrategia de Nair y Shrivastava (1942); los cuales dividen en dos grupos el rango de X, estimando "a" y "b" del siguiente modo:

$$\overline{X}_1 = (X_1 + \cdots + X_{ni})/n_i \quad \overline{Y}_1 = (Y_1 + \cdots + Y_{ni})/n_i \tag{3.48}$$

$$\overline{X}_2 = (X_{ni+1} + \cdots + X_{nj})/n_j - n_i \quad \overline{Y} = (Y_{ni+1} + \cdots + Y_{nj})/n_j - n_i \tag{3.49}$$

y con posterioridad,

$$\begin{aligned} b &= (\overline{y}_2 - \overline{y}_1)/(\overline{x}_2 - \overline{x}_1) \qquad (3.50) \\ a &= \overline{y}_1 - b\overline{x}_1 = \overline{y}_2 - b\overline{x}_2 \qquad (3.51) \end{aligned}$$

Por otra parte, Bartlett (1949) propone la división en tres grupos y estima el valor de "b" con la misma expresión que emplean Nair y Shrivastava (1942). Se diferencia de estos autores en el cálculo de "a" que se define por

$$\overline{a} = \overline{y} - b\overline{x} \tag{3.52}$$

Por último planteamos la propuesta de Brown y Mood (1951) los cuales dividen los valores de X en dos grupos, por encima y por debajo de la mediana de X. Los valores de "a" y "b" se estiman ajustando a 0 los residuales de ambos grupos, es decir:

$$\text{med}_{x<Md(x)} \quad \{y - a - bx\} = 0 \tag{3.53}$$

$$\text{med}_{x>Md(x)} \quad \{y - a - bx\} = 0 \tag{3.54}$$

Debe añadirse que el establecimiento de una línea ajustada a la nube de puntos mediante estos procedimientos no es la única posibilidad en el tratamiento de situaciones que no soportan una estimación mínimo cuadrática. Como simple ejemplo de distintas propuestas de solución podemos citar las que se definen en Theil (1950) y Sen (1968) (Median of Pairwise Slopes); Hampel (1971) (Breakdown bound); Gentle (1977) (Least Absolute Residuals —LAR—); Siegel (1982) (Repeated Median Line) o más reciente, las estimaciones no paramétricas presentadas por Härdle (1990).

3.4. ANÁLISIS COMPARATIVO ENTRE LA EXPLORACIÓN Y LA CONFIRMACIÓN

Visto lo anterior, es necesario esbozar el comportamiento distinto que en unos mismos datos se observa con la línea resistente y con el modelo lineal de la regresión. No se trata aqui de establecer un desarrollo exhaustivo del modelo de la regresión, puesto que no es este el objetivo de este texto; sino simplemente facilitar al lector una cierta evidencia empírica

del distinto ajuste que se consigue con ambas técnicas bajo determinadas condiciones. Para ello utilizaremos los datos presentados en la tabla 3.1, en la que hemos efectuado los cálculos de la línea resistente con toda la muestra (18 sujetos) y los hemos repetido eliminando de la muestra los dos sujetos (13 y 17) con residuales altos. En ambas ocasiones se consiguieron las siguientes lineas resistentes (ofrecemos los datos sin corregir):

$$\text{Muestra entera} \quad Y = 31.94 + 0.49(X) \qquad (3.55.1)$$

$$\text{Muestra reducida} \quad Y = 30.31 + 0.54(X) \qquad (3.55.2)$$

Esos mismos datos tratados mediante la estimación mínimo cuadrática de la regresión, muestran los siguientes valores resúmen:

$$\text{Muestra entera} \quad Y = 30.21 + 0.51(X) \qquad (3.56.1)$$
$$R^2 = .378(p = .0066)$$

$$\text{Muestra reducida} \quad Y = 31.75 + 0.45(X) \qquad (3.56.2)$$
$$R^2 = .461(p = .0038)$$

No abordaremos el estudio de los residuales de los modelos de regresión en busca de anomalías, puesto que, como se ha dicho, ello escapa a nuestros propósitos. Baste con señalar que, si bien se obtienen ecuaciones de regresión muy parecidas a las lineas resistentes de ambas situaciones, debe tenerse en cuenta el elemento principal que radica en el incremento hallado en los coeficientes de determinación de la segunda ecuación de regresión con respecto a la primera (0.083) lo que conlleva una importante reducción en el grado de significación del ajuste de la recta de regresión. Con solo la reducción de dos sujetos, se consigue reducir drásticamente la probabilidad de cometer un error de tipo I en el rechazo de la hipótesis nula.

No es necesario abundar en este tema, puesto que cuanto mayor sea el desajuste aparente de la nube de puntos inicial, o cuanto mayor sea el número de sujetos "outliers", más claro será el sesgo en la estimación mínimo cuadrática y en consecuencia, más fácil establecer modelos lineales incorrectos.

3.5. EXPLORANDO LA RELACIÓN BIVARIABLE CON LA LÍNEA RESISTENTE

Ya se ha planteado en apartados anteriores la utilidad de la línea resistente para la exploración de la forma de relación entre las variables. Ello es especialmente necesario cuando se pretende modelizar esa relación mediante estrategias lineales y no se dispone de suficiente conocimiento previo de que tal relación es factible. En este punto la línea resistente ofrece un sencillo mecanismo para establecer la forma de la relación y unos indicadores, igualmente resistentes, para ofrecer una estrategia que permita la mejor transformación posible de la nube de puntos original para establecer la necesaria linealidad de esa nube. Lógicamente, los valores de "a" y de "b" son los responsables de tal exploración. A este efecto definiremos tres grandes tipos de relación entre variables cuantitativas:

a) Relación lineal monotónica:

   Se trata de una relación en la que la velocidad de incremento o decremento en los valores de Y se mantiene constante a lo largo de los valores de X. La siguiente figura muestra una relación de estas características:

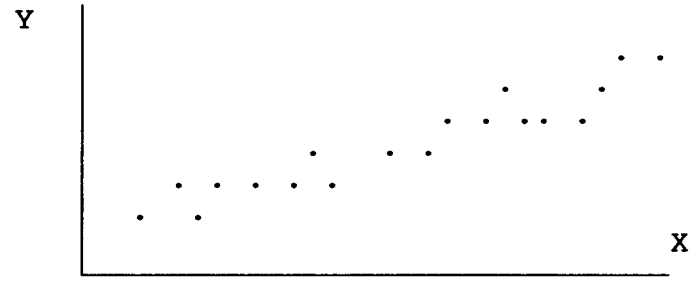


Fig.3.15. Relación lineal monotónica entre variables.

b) Relación No lineal Monotónica:

Al contrario de la anterior, la velocidad de incremento o decremento no permanece constante a lo largo de todo el dominio. Ello implica que en estas situaciones, al menos una vez se cambia esa velocidad. Mantiene, con respecto a la anterior relación, la similitud en el comportamiento general de la función sin cambio de dirección. La siguiente gráfica propone un ejemplo de este estilo:

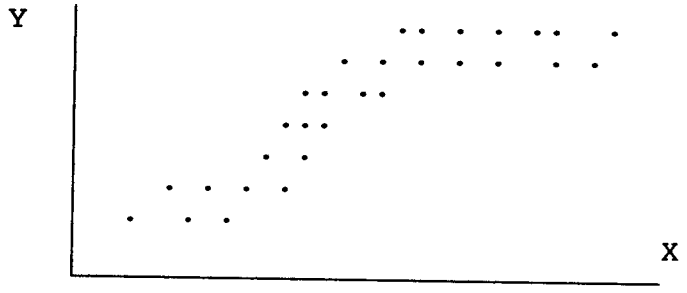


Fig.3.16. Relación no lineal monotónica entre variables.

c) Relación no monotónica:

En este caso, se plantean las mismas características de la relación no lineal monotónica pero con la salvedad que en este caso se presenta al menos un cambio de dirección en la nube de puntos original. Esta situación, obviamente, no permite incorporar el término lineal en su definición puesto que jamás lo es. La siguiente figura muestra un ejemplo de esta relación:

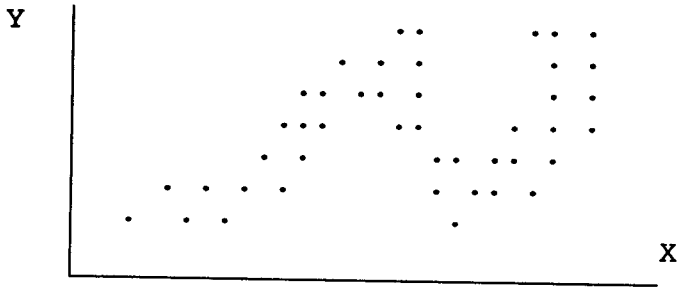


Fig.3.17. Relación no monotónica

Explorar la forma de la nube original pasa por la evaluación del valor de "*b*". Como ya se ha dicho, del comportamiento del signo de la pendiente podemos extraer la información de si la nube de puntos presenta una tendencia creciente (signo +) o decreciente (signo -). Sin embargo, aqui no nos estamos planteando el comportamiento de la función de una forma puntual, sino que deseamos explorar si es factible adoptar la linealidad como forma de la relación exhibida por la nube.

A partir de lo que se ha dicho, parece claro que las relaciones de carácter lineal monotónico son las que son susceptibles de ser modelizadas linealmente, y que en caso de obtener una relación de carácter no lineal monotónico, debemos establecer algún mecanismo para transformar esa nube en una relación lineal. Obviamente en el caso de las relaciones no monotónicas, tal empeño parece inútil, puesto que el cambio de dirección que las caracteriza impide su linealidad, aún sometiéndolas a transformaciones. Es necesario, pues, definir un tratamiento especial en las relaciones no lineales monotónicas, puesto que son recuperables para la linealidad. Este tratamiento pasa por dos puntos esenciales:

a) Identificar el carácter no lineal de la nube original.

b) En base a la forma original, establecer la mejor transformación posible de la nube para tender a la linealidad.

Con respecto al primer punto, debemos señalar que las diferentes posibilidades de obtener relaciones no lineales de las que aqui nos ocupan (monotónicas), no son excesivamente grandes. En concreto podemos establecer cuatro grandes tipos de nubes, las cuales se esquematizan en la siguiente figura:

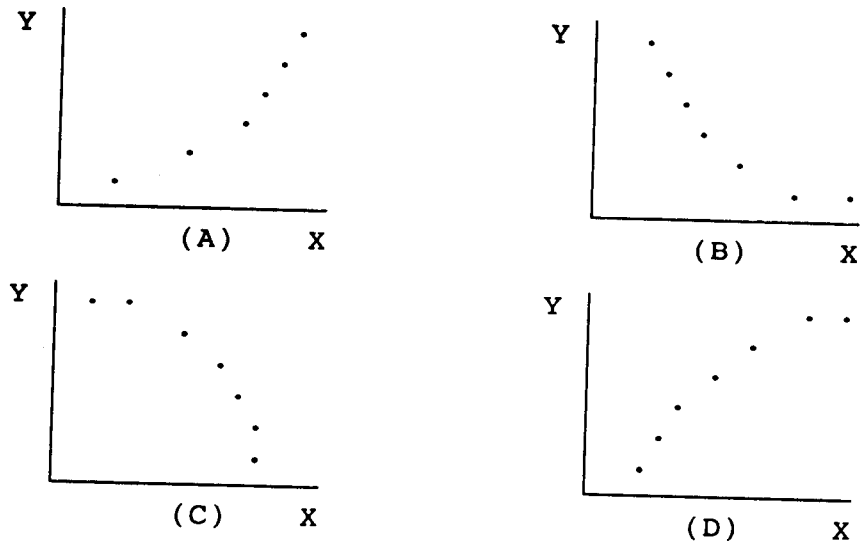


Fig. 3.18. Cuatro tipos básicos de funciones no lineales monotónicas.

Así, es importante poder determinar que tipo de situación es la que caracteriza a la nube de puntos inicial. Al margen del análisis gráfico (veáse capítulo 2 de este texto), la línea resistente, como decíamos, nos ofrece un sencillo mecanismo para la evaluación que nos proponemos. Se trata de establecer la pendiente de la línea resistente entre el tercio inferior y superior por una parte y, por otra, la misma pendiente pero entre el tercio medio y el superior. Con esas dos pendientes parciales, se establece una ratio entre ellas, denominada semipendiente y representada por $\frac{1}{2}(b)$, cuyo resultado nos informa de la existencia o no de linealidad, y en este último caso, permite decidir la conveniencia de emplear alguna transformación de la nube para llegar a la linealidad. En virtud de lo expuesto, recuperando para ello los puntos resumen de cada tercio de la nube original, podemos definir:

$$\text{Pendiente parcial inferior } b(inf) = (Y_m - Y_i)/(X_m - X_i) \quad (3.57.1)$$

$$\text{Pendiente parcial superior } b(sup) = (Y_s - Y_m)/(X_s - X_m) \quad (3.57.2)$$

$$\text{Semipendiente}\frac{1}{2}(b) = b(\text{inf})/b(\text{sup}) \quad (3.58)$$

Debe tenerse en cuenta que en la expresión de $\frac{1}{2}(b)$, la pendiente parcial mayor (en valor absoluto) debe figurar en el denominador de esa razón. Con ello se obtiene un valor que, como máximo, puede adoptar el valor 1.

Obviamente, aquellos valores de $\frac{1}{2}(b)$ menores de 0, es decir, que las pendientes parciales son de signo contrario, indican que en la nube de puntos original se da, como mínimo, un cambio de dirección, lo que supone que la inflexión de la función impide llegar a la linealidad. Un resultado en $\frac{1}{2}(b)$ cercano a 1, se relaciona con una linealidad de la nube inicial y con la posibilidad de ajustar este tipo de modelos por lo que a la forma de la nube se refire. En consecuecia, el rango de valores de $\frac{1}{2}(b)$ comprendido entre 0 y 1 corresponde a las nubes de puntos bivariables que no son lineales pero si son monotónas crecientes o decrecientes, y hacen factible pensar en alguna transformación de la nube. Las nubes que se muestran en la figura 3.18 son las prototípicas de esta situación, o sea, son las especialmente indicadas para estudiar las transformaciones de esa nube.

El siguiente cuadro presenta un breve resumen de las distintas posibilidades a partir del valor de $\frac{1}{2}(b)$.

Cuadro 3.2. Criterios generales de interpretación de la semipendiente $\frac{1}{2}(b)$.

VALORES DE $\frac{1}{2}(b)$	TRATAMIENTO DE LA NUBE ORIGINAL
$0.9 \leq \frac{1}{2}(b) \leq 1$	La relación es lineal, permite el empleo de tales modelos
$0.5 \leq \frac{1}{2}(b) < 0.9$	Una transformación adecuada permitirá obtener linealidad.
$0 \leq \frac{1}{2}(b) < 0.5$	Es factible plantear una transformación en X o en Y para conseguir linealidad. Si $\frac{1}{2}(b)$ es muy cercano a 0 o la nube es muy curva, probablemente no consiga la linealidad ni transformando
$\frac{1}{2}(b) < 0$	No es factible ninguna transformación puesto que esta situación indica un cambio de dirección en la función.

Con ello llegamos al segundo punto de los anteriormente planteados, el que se refiere al establecimiento de la mejor transformación de la nube. Como primer aspecto, debemos hacer hincapié en la necesidad de la inspección visual de la nube, una vez la semipendiente pone de manifiesto la no linealidad de la misma. El proceso de transformación debe articularse en base al movimiento necesario de la nube para "enderezarla", es decir reducir o aumentar el rango de una o de ambas variables (encoger o extender en términos más descriptivos), para llegar a la linealidad. Este es un proceso más artesanal que tecnológico, puesto que es importante adoptar la estrategia precisa, y lo más simple posible (son preferibles las transformaciones menos complejas a las más sofisticadas) que conduzca a una nube lineal. El cuadro que a continuación exhibimos presenta de forma esquemática una simple guía para elegir el mecanismo de "extensión" o "encojimiento" de las variables en base a las nubes de la figura 3.18.

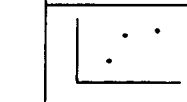
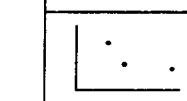
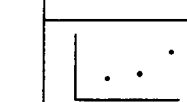
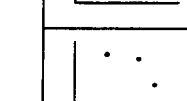
Cuadro 3.3. Estrategia de transformación para funciones no lineales.

TIPO DE RELACION	PENDIENTES PARCIALES	ACTUACION SOBRE EL RANGO	
		X	Y
	b(inf)>b(sup) ambas +	Reducir	Aumentar
	b(inf)>b(sup) ambas -	Reducir	Reducir
	b(inf)<b(sup) ambas +	Aumentar	Reducir
	b(inf)<b(sup) ambas -	Aumentar	Aumentar

No se trata aquí de efectuar un análisis exhaustivo de los diferentes procedimientos de transformación, sino la de mostrar la utilidad de la línea resistente en este ámbito. Por otra parte, son conocidas la mayoría de estrategias para aumentar o reducir (extender o encojer, respectivamente)

el rango de las variables originales. Sin embargo, remitiendo al lector a la exposición de este tema en el segundo capítulo de este texto, podemos presentar el siguiente cuadro resumen de algunas de las transformaciones clásicas en este tipo de situaciones:

Cuadro 3.4. Posibles transformaciones en función del tipo de relación original.

TIPO DE RELACION	PENDIENTES PARCIALES	TRANSFORMACIONES	
		X	Y
	b(inf)>b(sup) ambas +	Log (X) (-1/X) etc.	Y^2 Y^3 etc...
	b(inf)>b(sup) ambas -	Log (X) etc...	Log (Y) etc...
	b(inf)<b(sup) ambas +	X^2 X^3 etc...	Log (Y) etc...
	b(inf)<b(sup) ambas -	X^2 etc...	Y^2 etc...

Como se ha dicho, el cuadro anterior sólo persigue ofrecer un esquema inicial de actuación para la transformación de los datos originales y conseguir, así, la linealidad de la nube. El papel de la línea resistente es el de aportar, en base al estudio de las pendientes parciales y de la semipendiente, evidencia empírica para poder establecer la mejor actuación posible en el tratamiento estadístico de los datos. A continuación, veremos este proceso en un rápido ejemplo.

3.6. EXPLORACIÓN DE UNA NUBE DE PUNTOS MEDIANTE EL ANÁLISIS DE LA SEMIPENDIENTE (CASO PRÁCTICO)

Mostraremos la utilización de la pendiente con unos datos recogidos en una muestra de 15 sujetos. Se desea estudiar la relación existente entre el lenguaje escrito natural y la edad de los sujetos, pensando que el número de estructuras lingüísticas complejas que pone en práctica el sujeto en la redacción de un texto libre de duración controlada, depende en gran parte de la maduración del sujeto. En esa muestra se obtuvieron los datos que se reflejan en la tabla número 3.4:

Tabla 3.4. Matriz inicial de datos.

SUJETO	EDAD	NÚMERO ESTRUCTURAS
1	8	6
2	9	7
3	9	6
4	10	7
5	12	8
.....		
6	13	10
7	13	12
8	14	14
9	15	17
10	16	24
.....		
11	17	40
12	18	42
13	19	48
14	19	52
15	19	51

En la tabla anterior se han destacado los tres tercios de valores, según el rango de X, para facilitar el seguimiento de los diferentes pasos.

En primer lugar, podría pensarse en el ajuste de un modelo lineal confirmatorio en el que el número de estructuras se defina como variable criterio y la edad como variable regresora. Un análisis de este intento muestra un coeficiente de determinación de $R^2 = 0,95(P = 0.0000)$, lo cual podría llevar (dejándo al margen el habitual y clásico análisis de los residuales) a considerar a la ya obtenida ecuación de regresión [Y = -37.71 + (4.31 $\cdot X$)] como un modelo adecuado para, como mínimo, describir la relación entre ambas variables. Si estos datos se abordan desde una perspectiva exploratória, la línea resistente que se obtiene presenta la siguiente expresión:

$$Y = -34.4 + 4.1(X) \tag{3.59}$$

Se observan evidentes semejanzas entre la ecuación de regresión y la línea resistente establecida a partir de nuestros datos. Como dato complementario, se obtiene en esta línea resistente un valor de ajuste de δ = 0.51 (veáse la definición de δ en este mísmo capítulo). Sin embargo, una inspección visual de la nube de puntos original muestra que la relación lineal es más que dudosa. La siguiente figura muestra tal situación:

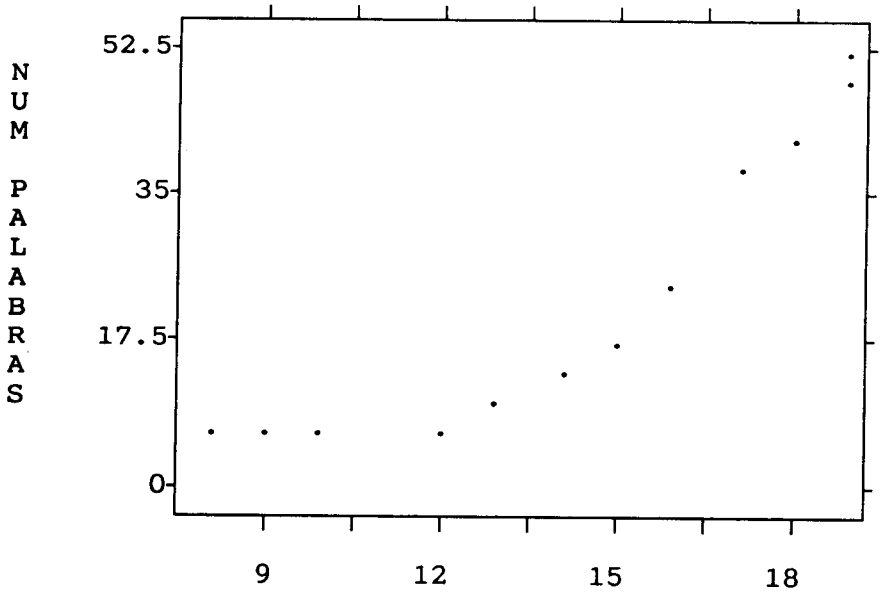


Fig. 3.19. Nube de puntos original ente las dos variables de la que se desprende una relación no lineal monotónica.

En la figura señalada se advierte que la relación entre ambas variables es de carácter monotónico (no cambia de dirección) pero no lineal.

Esta sospecha puede ser confirmada o desmentida mediante el análisis de la "semipendiente" generada a partir de los datos de la nube. Así, las medianas correspondientes a cada tercio muestran los siguientes resultados:

Tercio Inferior	Tercio Medio	Tercio Superior
$X_i = 9$	$X_m = 14$	$X_s = 19$
$Y_i = 7$	$Y_m = 14$	$Y_s = 48$

En consecuencia, podemos definir los siguiente valores:

$$b(\text{inf}) = (14 - 7)/(14 - 9) = 1.4 \qquad (3.60.1)$$

$$b(\text{sup}) = (48 - 14)/(19 - 14) = 6.8 \qquad (3.60.2)$$

$$\frac{1}{2}(b) = 1.4/6.8 = 0.206 \qquad (3.61)$$

El resultado obtenido en $\frac{1}{2}(b) = 0.206$ muestra claramente la idea de la inexistencia de linealidad en la nube original (veáse cuadro número 3.3). A la vista de esta circunstáncia, abordaremos el resultado de una simple transformación en la nube original para obtener una relación monotónica y lineal. En base al cuadro número 3.4, y por no complejizar esta cuestión, transformamos los valores de X elevándolos al cuadrado y los valores de Y obteniéndo su logaritmo neperiano. Con ello conseguimos el aumento (extensión) de X y la reducción (encojimiento) de Y, de forma que disponemos de unas nuevas variables, a las que nos referiremos como X' e Y', y que se definen del siguiente modo:

$$X' = X^2 \qquad Y' = \log(Y) \qquad (3.62)$$

Una vez efectuada esta transformación, la nube de puntos resultante se muestra en la siguiente figura:

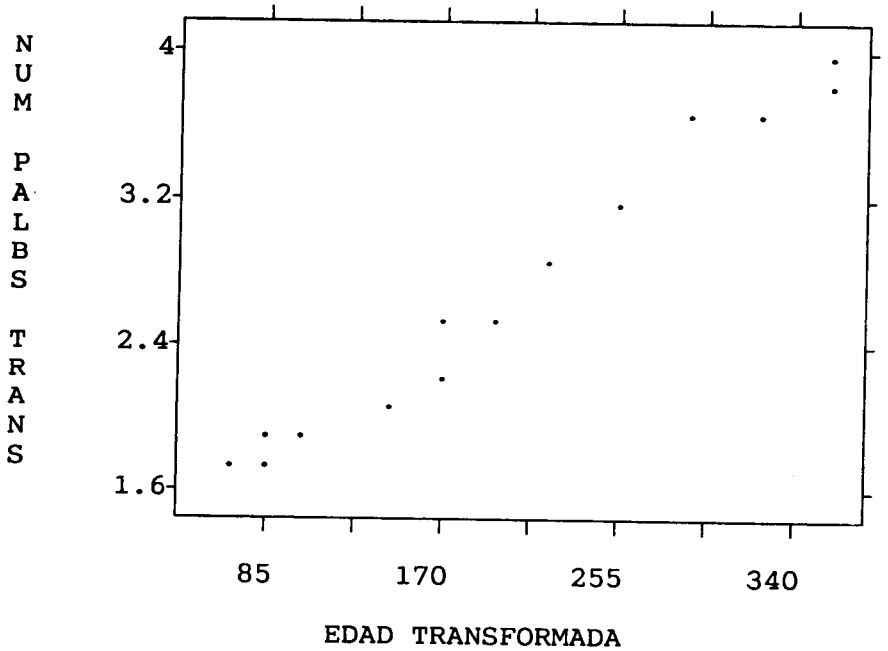


Fig. 3.20. Nube de puntos obtenida después de la transformación de X y de $Y(X', Y')$

En esta última figura, se observa una mayor linealidad que en la original, de forma y manera que podemos estar en condiciones de la exploración y posterior confirmación, de estrategias de análisis lineal. Si reproducimos los cálculos efectuados anteriormente, pero con la nube transformada, se obtienen los siguientes resultados:

Ecuación de regresión:	$Y = 1.180 + 0.00769(X)$	(3.63.1)
	$R^2 = 0.98(p = 0.0000)$	(3.63.2)
Línea resistente:	$Y = 1.356 + 0.06875(X)$	(3.63.3)
Pendiente parcial inferior	$b(\text{inf}) = 0.006026$	(3.63.4)
Pendiente parcial superior	$b(\text{sup}) = 0.007464$	(3.63.5)
Semipendiente	$\frac{1}{2}(b) = 0.81$	(3.63.6)
Valor del índice	$\delta = 0.53$	

A la vista de estos resultados, y de la inspección visual de la figura 3.20, podemos analizar las consecuencias de la transformación de la nube. La forma lineal de la relación parece clara, como lo asevera el valor de la

semipendiente (muy cercana al nivel ideal), siendo también adecuados los resultados de la línea resistente y del ajuste de la ecuación de regresión. El único índice que se mantiene constante antes y después de la transformación es "δ", lo que indica que ni antes ni ahora se dan valores "outliers" en los residuales.

Ya hemos mencionado que este caso práctico debe completarse mediante un exhaustivo análisis de los residuales, tanto en la versión exploratória como confirmatoria. Sin embargo, sirvan estos datos mínimos para poner de manifiesto la utilidad de la línea resistente y de las estrategias estadísticas que de ella se derivan en el estudio de nubes de puntos bivariables.

3.7. ASPECTOS MATEMÁTICOS DE LA TRANSFORMACIÓN DE NUBES DE PUNTOS BIVARIABLES

En este apartado queremos presentar una propuesta relativa a la transformación de nubes de puntos no lineales que, a diferencia del apartado anterior, emplea para su desarrollo algunos aspectos matemáticos más complejos. Su utilidad estriba en la mezcla entre un enfoque estadístico más clásico con las consideraciones gráficas y resistentes que propugnan las técnicas E.D.A.

El planteamiento al que nos referimos está desarrollado por Emerson (1982), el cual define un proceso matemático de transformación para cada una de las situaciones básicas del análisis estadístico (transformaciones univariables para establecer simetría, o reducción de la variación, o una determinada curtosis, transformaciones de tablas de contingencia, etc...) basadas en estadísticos resistentes. De entre todas las situaciones, aquí presentamos la que se refiere a las nubes de puntos no lineales. El esquema de esta propuesta se encuentra en Emerson y Stoto (1982) y, como decíamos, se desarrolla en Emerson (1982). En concreto, partimos

de una serie de "n" puntos (X_i, Y_i) originales que no se ajustan a una relación lineal. Emerson propone como mejor transformación para obtener linealidad, transformar los valores de Y con el establecimiento de una transformación de potencia con un exponente igual a "p". Se trata, pues, de identificar el valor adecuado de "p" para alcanzar el objetivo deseado. Partimos pues, de una serie de "n" puntos como los siguientes:

$$\{(X_1, Y_1)(X_2, Y_2) \ldots (X_n, Y_n)\} \tag{3.64}$$

Apliquemos una transformación de potencia (Φ) a la variable Y con un exponente (p). De este modo los puntos anteriores quedarían definidos como:

$$\{[X_1, \Phi(Y_1)][(X_2, \Phi(Y_2)] \ldots [(X_n, \Phi(Y_n)]\} \tag{3.65}$$

Definamos dos valores representativos en cada variable. Como es propio de la técnicas E.D.A. seleccionamos Med(X) y Med(Y) como puntos representativos y, por supuesto, Med$\{\Phi(Y)\}$ para la variable transformada. Si la transformación propuesta nos debe llevar a la linealidad, el modelo general debe adoptar la siguiente expresión:

$$\Phi(Y) - \text{Med}\{\Phi(Y)\} = K[X - \text{Med}(X)] \tag{3.66}$$

Obsérvese que si en este modelo anterior, la transformación se efectua con una potencia "$P = 0$", obtendríamos el modelo general ya conocido, puesto que se generaría el logaritmo de las transformaciones o, lo que es lo mismo, los valores iniciales, de modo que la expresión anterior se reformularía como:

$$Y - \text{Med}(Y) = b[X - \text{Med}(X)] \tag{3.67}$$

Esta operación requiere, pues, en primer lugar disponer de la evidencia de que la nube original no es lineal. Emerson, ante la no linealidad de los datos iniciales señala que una transformación que se basara solamente en el centrado de la variable Y con respecto a su mediana no resuelve la cuestión de modo satisfactorio, toda vez que tal tratamiento genera valores negativos y positivos, lo que hace más complejo el análisis. Por contra, establecer la transformación de Y en base a una potencia puede no dar lugar a esa situación. Así definiremos la variable transformada, que por comodidad denominamos:

$$\Phi(Y) = Y^p = Z \tag{3.68}$$

Con Z representamos los nuevos valores de Y. Se mantiene una expresión importante, puesto que si p es distinto de 0, se cumple que:

$$Y = Z^{(1/p)} \qquad \text{Med}(Y) \approx \text{Med}[Z^{(1/p)}] \tag{3.69}$$

Desarrollando la serie $Z^{(1/p)}$ en una serie de Taylor con respecto a Med(Z) se obtiene finalmente (no incorporamos todos los pasos de Emerson) la siguiente expresión:

$$Y - \text{Med}(Y) - b[b - \text{Med}(X)] \approx (1-p) \cdot [b^2(X - \text{Med}(x))^2/2\text{Med}(Y)] \tag{3.70}$$

siendo "b" la pendiente de la línea resistente que se ajusta a la nube inicial y siendo "p" el valor desconocido de la potencia que se desea obtener. Para obtenerlo, se siguen los siguientes pasos:

a) Establecimiento del valor de la pendiente (b) de la línea resistente original, tal y como se ha mostrado en los apartados correspondientes de este mismo capítulo.

b) Transformar los valores de X en X' con la siguiente expresión:

$$X' = [b^2(X - \text{Med}(X)^2]/2\text{Med}(Y) \tag{3.71}$$

que como se aprecia corresponde a la derecha del $\approx$ de la expresión anterior.

c) Transformar Y en Y' de la siguiente forma:

$$Y' = Y - \text{Med}(Y) - b[X - \text{Med}(X)] \tag{3.72}$$

que corresponde a la izquierda del $\approx$ de la expresión general.

d) Establecer la nube de puntos entre X' (ordenadas) e Y' (abcisas) de forma que, lógicamente, deben definir una nube claramente lineal con un valor de pendiente (m) cercano a 1.

e) Establecer el valor de $p = (1 - m)$, de forma que la transformación ideal a la que someter a Y sea la de elevar sus valores a ese exponente $(1 - m)$.

La evidencia empírica que presenta Emerson avala esta propuesta de transformación, que si bien es un tanto compleja de realización, facilita obtener la linealidad deseada de una forma casi segura.

Para finalizar estos apartados dedicados a las transformaciones de nubes de puntos bivariables, debemos remitir al lector a algunos trabajos que desarrollan este punto en toda su dimensión. Por ejemplo, si nos interesamos por transformaciones en análisis confirmatórios, Box y Tidwell (1962) y Box y Cox (1964) trata ampliamente esta cuestión. Si empleamos un enfoque exploratório, como el aqui presentado, Tukey (1977) y Tukey y Mosteller (1977) y Leinhardt y Wasserman (1979) son excelentes desarrollos de toda esta temática.

Como ejemplo de la propuesta de Emerson, podemos retormar los datos de la tabla número 3.4, los cuales están representados en la figura número 3.19. Veíamos en esa nube original la existencia de una relación no lineal. Intentaremos llegar a la linealidad a través del planteamiento que nos ocupa en este apartado, aunque sólo lo haremos a nivel gráfico para no reitarar datos ya calculados.

En consecuencia con lo anterior, los datos necesarios para llevar a cabo este objetivo son los siguientes:

$$\text{Med}(X) = 13.5 \quad \text{Med}(Y) = 13.0 \quad \text{“}b\text{”} = 4.1$$

recordando que el valor de la pendiente se halla a través de la línea resistente. Como primer paso, sometemos a los valores de X y de Y a la siguiente transformación:

$$Y' = Y - \text{Med}(Y) - b(X - \text{Med}(X)) = y - 13 - 4.1(X - 13.5) \tag{3.73.1}$$

$$X' = [b^2(X - \text{Med}(x))^2]/2 \cdot \text{Med}(Y) = 16.81(X - 13.5)^2/26 \tag{3.73.2}$$

Con estas transformaciones, la nube de puntos resultante sigue la siguiente figura:

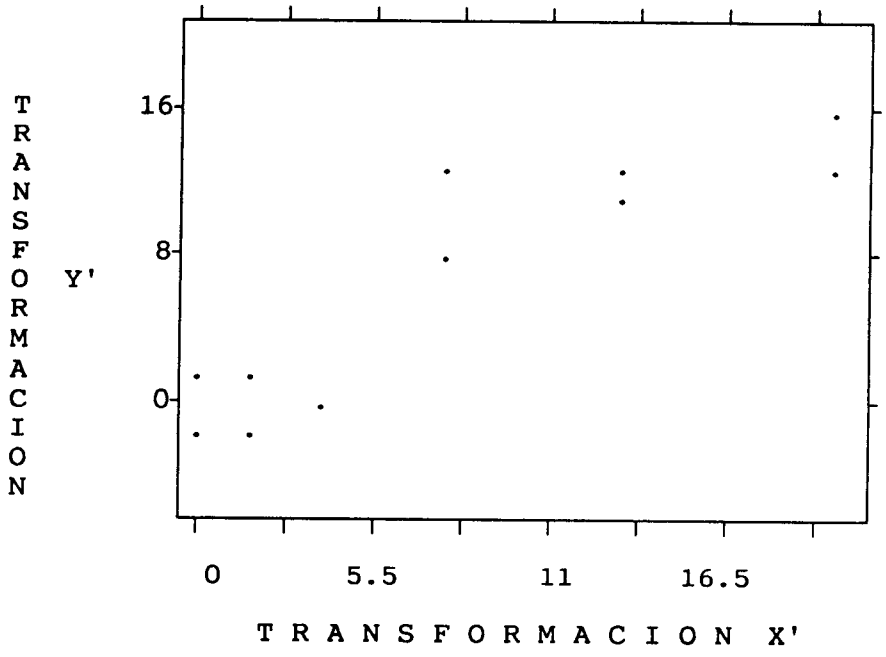


Fig. 3.21. Nube de puntos consecuencia de la transformación de X e Y propuesta por Emerson.

Si en la figura anterior, asumimos su linealidad (lo cual no es excesivamente difícil), obtenemos un valor de pendiente igual a 0.82728, lo cual, dado su valor cercano a 1, pone de manifiesto esa linealidad de la que hablábamos. Siguiendo con el esquema de Emerson, la mejor transformación de la nube original pasa por una transformación de potencia aplicada a Y con un exponente $(p = 1 - m)$ siendo "m" el valor de la pendiente hallada en la nube de la figura 3.21. Asi, el valor del exponente sería:

$$p = 1 - m = 1 - 0.82728 = 0.17672. \tag{3.74}$$

Transformando Y en el sentido descrito, es decir generando una Y'' con la siguiente expresión:

$$Y'' = Y^{0.17672} \tag{3.75}$$

se obtienen una nube de puntos (sin modificar los valores de X) claramente lineal. La figura 3.22 muestra esa nube transformada.

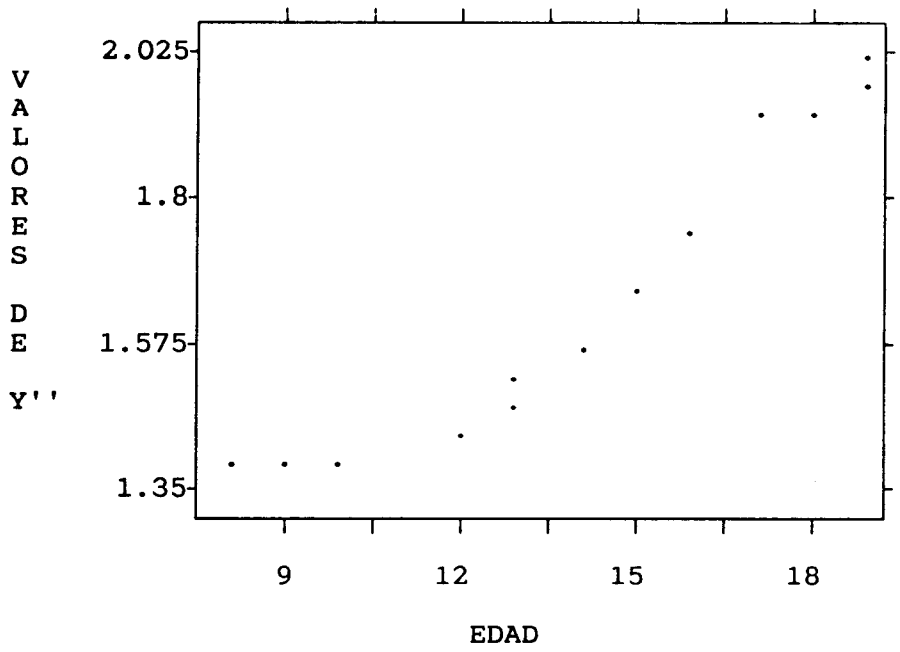


Fig.3.22. Nube de puntos una vez aplicada la transformación de potencia a Y con exponente igual a 0.17672.

Como hemos dicho, no entraremos en el análisis concreto de cada modelo, pues con la evidencia gráfica pensamos que es suficiente para hacer patente las características de esta posibilidad de transformación de Y.

3.8. UNA APROXIMACIÓN EXPLORATORIA A LA CORRELACIÓN PARCIAL

Como último apartado de este capítulo mostramos un breve comentario a la concepción del coeficiente de correlación parcial desde una perspectiva exploratória de las técnicas E.D.A.

Con ello queremos introducir, aunque de forma breve, la concepción de los modelos lineales múltiples, es decir con más de dos dimensiones.

De hecho, esta cuestión se enmarca en una temática más amplia, puesto que no es desdeñable la repercusión de la correlación parcial en la construcción de modelos lineales confirmatórios; ya sea por su vinculación a la tolerancia (Searle, 1971; Doménech y Riba, 1985), o por su relevancia en la evaluación causal de efectos complejos (Cliff, 1983; Guàrdia y Arnau, 1991).

Es ampliamente conocida la estructura del coeficiente de correlación parcial, la cual se puede esquematizar, en forma de estructura lineal, a partir de la concepción general que indica que este tipo de correlación pretende evaluar la relación entre dos variables controlando el efecto de una tercera. Así, si hemos definido la expresión resistente

$$Y = a + b(X_1) \tag{3.76}$$

deseamos ahora, evaluar la relación

$$Y = a' + b'(X_2) \tag{3.77}$$

asumiendo que X_1 y X_2 no son independientes entre si. A través de la estimación mínimo cuadrática típica de los modelos de regresión confirmatórios, podemos llegar a la evaluación de esa relación parcial de una forma lógica. Por ejemplo, definimos los residuales (Y') que se originan en el modelo que vincula a Y con X_1:

$$Y' = Y - (b_1 X_1 + a_1) \tag{3.78}$$

y por lo que se refiere a los residuales (X_2') relativos a la relación entre X_2 y X_1:

$$X_2' = X_2 - (b_2 X_1 + a_2) \tag{3.79}$$

Evaluar la aportación que X_2 efectua a Y, controlando X_1, pasa simplemente por analizar el modelo confirmatório que puede determinarse del siguiente modo:

$$Y' = b_3(X_2') + a_3 \tag{3.80}$$

Como se ve, intentamos analizar si existe relación entre los residuales de $Y = f(X_1)$ y los residuales de $X_2 = f(X_1)$; de manera que se obtiene una igualdad evidente:

$$r(Y'X_2') = r(YX_2 \cdot X_1) \tag{3.81}$$

No incoporamos aquí expresiones puntuales para la determinación de un coeficiente de correlación parcial, por otra parte ampliamente conocidas, ni acerca de la significación del mismo, puesto que es igualmente conocido; sino que se propone efectuar la misma evaluación a través de los indicadores de las técnicas E.D.A. derivadas de la exploración lineal o línea resistente bivariable.

Si se trata de evaluar si existe relación entre los residuales de $Y = f(X_1)$ y de $X_2 = f(X_1)$ mediante la expresión lineal de Y', tal relación puede estudiarse a través de los índices exploratórios de la línea resistente. En concreto podemos recuperar la propuesta de Jonhstone y Velleman (1982) para evaluar el ajuste de una línea resistente. Es factible definir los residuales que se derivan de $Y' = f(X_2')$ de la siguiente manera:

$$Y'' = Y' - (b_3 X_2' + a_3) \tag{3.82}$$

recordando que Y' es el residual de $Y = b_1 X_1 + a_1$
X_2' es el residual de $X_2 = b_2 X_1 + a_2$

De esta forma el valor de variación de Y' no explicado por X_2', que de forma confirmatória se establece en base al coeficiente de correlación parcial, se define por

$$1 - [r(X_2 Y \cdot X_1)]^2 \tag{3.83}$$

y aplicando las estrategias exploratórias podría estimarse con el índice δ de los mencionados autores, obteniéndose:

$$\delta = (dqY'')/(dqY') \tag{3.84}$$

siendo dq la distáncia entre cuartos de esas dos series de residuales.

Ya se ha comentado la necesaria incorporación de este tipo de análisis parciales en los modelos que nos ocupan, puesto que su valor dirige tanto la evaluación de efectos entre las variables regresoras (tolerancia, colinealidad, sesgo en las estimaciones, ...) como por lo que se refiere al contenido causal, si el diseño lo permite, de la interpretación de los efectos hallados.

Para mostrar la utilidad del indice "δ" aplicado a la evaluación del coeficiente de correlación parcial, utilizaremos unos datos propuestos por Berry y Lewis-Beck (1986), en los que se evalua a 13 estados de U.S.A.

en las siguientes variables: Índice de consistencia cognitiva, porcentaje de sujetos con estudios medios en cada estado e índice de competitividad. Se propone establecer un modelo de regresión múltiple que establezca el índice de consistencia cognitiva en función de las otras dos variables. Como dato inicial ofrecemos la matriz de correlaciones que se genera:

Con. Cognitiva	(Y)	1.0000		
(%) Est. Medios	(X_1)	.6921	1.0000	
Competitividad	(X_2)	.6499	.4660	1.0000

Como primer análisis hemos obtenido el coeficiente de correlación parcial $r(YX_2 \cdot X_1) = 0.51256$. Intentaremos llegar a un dato aproximado a este a través del índice "δ" que hemos propuesto anteriormente:

En primer lugar, obtendremos la ecuación de regresión (con estimación OLS) que se genera a partir de $Y = f(X_1)$:

$$Y = -0.31766 + 0.01213(X_1) \tag{3.85}$$

A continuación reproducimos este análisis, pero ajustando la función $X_2 = f(X_1)$, de donde:

$$X_2 = -15.23873 + 0.81304(X_1) \tag{3.86}$$

Generamos los valores residuales de cada una de esas dos ecuaciones de regresión de modo, que disponemos de dos nuevas variables:

$$Y' = Y - (-0.31766 + 0.01213 \cdot X_1) \tag{3.87.1}$$

$$X_2' = X_2 - (-15.23873 + 0.81304 \cdot X1) \tag{3.87.2}$$

A continuación, siguiendo nuestro esquema de actuación, obtenemos la estimación (OLS) de la función lineal $Y' = f(X_2')$, obteniendo:

$$Y' = 0.0000626 + 0.00420(X_2') \tag{3.88}$$

y, del mismo modo que en las ecuaciones anteriores, se puede definir los residuales de Y', de la siguiente forma:

$$Y'' = Y' - (0.0000626 + 0.00420 \cdot X_2') \tag{3.89}$$

Con las distribuciones de Y' y de Y'' podemos establecer el índice "δ" a traves de la distancias entre cuartos de Y' e Y''

$$\delta = dqY''/dqY' = 0.66 \tag{3.90}$$

Según el planteamiento efectuado, la relación entre el coeficiente de correlación parcial $[r(YX_2 \cdot X_1)]$ y el índice "δ" es la siguiente:

$$1 - [r(YX_2 \cdot X_1)]^2 \approx \delta \tag{3.91}$$

de donde puede calcularse que:

$$1 - (0.51256)^2 \approx 0.66 \tag{3.92}$$

Se observan unas ligeras diferencias en el valor de ambos índices, pero en ningún modo suficientes como para llegar a conclusiones distintas con cada uno de ellos.

En lineas generales, no proponemos la utilización del índice "δ" como substituto del coeficiente de correlación parcial, toda vez que este último es más rápido de cálculo que el primero. Simplemente, mostramos con este análisis comparativo, la igualdad en la sensibilidad de los índices clásicos con los que proponen las técnicas E.D.A; en especial en lo que se refiere a la conjunción del enfoque confirmatório y exploratório como es el caso de los modelos lineales múltiples que han ocupado nuestra atención.

4. TÉCNICAS DE SUAVIZADO

4.1. INTRODUCCIÓN

En el presente capitulo abordaremos la temática del **alisamiento** (—suavizado— "smoothing"). Diversos autores como C. Goodall (1.990), reconocen que probablemente la tarea fundamental del análisis de datos resida en descubrir y resaltar los patrones que están presentes en ellos. En el capítulo anterior ya se ha abordado esta perspectiva desde la utilización de la técnica de la línea resistente en el proceso de modelado, pero tal y como sugieren Velleman y Hoaglin, 1.981 cuando se trabaja con dos variables que presentan una estructura de relación, a veces será más interesante la busqueda de patrones mas generales que el proporcionado por una linea recta. Es decir interpretando a los mencionados autores, si en el capitulo anterior se promocionaba la utilización de transformaciones para obtener la linealidad, en esta técnica del suavizado la intención reside en descubrir cual es el patrón que mejor ajusta a los datos y que no necesariamente ha de ser el lineal.

Generalmente este tipo de proceder se circunscribe a las situaciones en que se dispone de una secuencia de datos identificable como una "**serie temporal**", aunque desde nuestro punto de vista no debe ser necesariamente la variable tiempo la que marque el orden de los datos dentro de la serie. Desde una perspectiva más amplia consideraremos que nos ha-

llamos ante una secuencia de datos que presentan una forma especial de relación, donde una de sus variables es importante especialmente por el orden que esta especifica. De todas formas cabe recoger la sugerencia de Tukey, 1.977 respecto a que los valores que proporciona la variable que nos permite la ordenación deben ser equidistantes entre si.

En conexión con el capitulo de línea resistente podriamos establecer que la intención de la técnica que se plantea, consiste en conseguir una descripción simple de la variable "Y" (dependiente) en función de la variable "X" (independiente) descomponiéndose cada dato a partir de la siguiente expresión general:

Dato = Ajuste + Residual (Data = Fit + Residual)

Evidentemente en el modelo planteado se pretende que la parte de ajuste intente recoger la mayor parte del patrón subyacente en los datos, de forma que se minimice la parte correspondiente al residual del modelo.

Pero desde la perspectiva del alisado se utiliza un tipo de descomposición que representa un caso especial de la anterior ecuación, separando cada dato en una parte alisada y una parte rugosa:

Dato = Parte Lisa + Parte rugosa (Data = Smooth + Rough)

donde la parte alisada no pretende ser una descripción mediante una formula sino simplemente una curva alisada que recoja a gran escala la estructura o conducta de la secuencia de datos, y por consiguiente la parte rugosa contenga la menor parte de estructura que sea posible, es decir sea un proceso de características de ruido blanco. Las técnicas de alisado son una aproximación simple para descubrir estas estructuras con el mínimo de presuposiciones "a priori" de como debe ser este patrón. La única suposición tal y como recoge C. Goodall (1.990) reside en que la relación entre las variables es alisada, la mayoria de las veces una curva continua, que no cambia rápidamente y que puede incluir un pequeño número de pasos o transiciones.

Una cuestión importante a remarcar se halla en que no necesariamente nuestro interés debe residir únicamente en la parte alisada sino que debemos resaltar la necesidad de analizar el componente residual o rugoso tal y como hemos mencionado en capítulos anteriores, y que desafortunadamente la tradición en la utilización de este tipo de técnicas no ha

contemplado. Quizás como señalan Velleman y Hoaglin, 1.981, a causa del nombre con que se conocen (Técnicas de Suavizado o Alisamiento).

Considerando por tanto que el proceso de establecimiento de la estructura puede ser llevado a cabo mediante el ajuste de una curva, de forma "casi visual" (mediante inspección visual) a la nube de puntos formada por la representación de los individuos en el espacio de las variables, es evidente que el ruido presente en los datos será el que nos impedirá o dificultará la visualización de cual es el patrón del proceso que los ha generado, por ello cabe plantearse la utilización de las técnicas de alisado como un proceso más de filtrado de los datos, encaminadas a eliminar o aislar este componente de la parte alisada (smooth).

Observese a modo de ejemplo la figura 4.1[1] que recoge el número de licencias de viviendas de nueva construcción concedidos mensualmente desde Enero de 1.965 hasta Diciembre de 1.975 en los EEUU, obviamente se hace dificil interpretar la existencia de ningún tipo de patron en los mencionados datos. Por otra parte observese la figura 4.2 correspondiente a los mismos datos que la anterior sometidos a un proceso de suavización, se puede constatar un patron cíclico aproximadamente anual que recoge un ascenso progresivo de las licencias durante los 8 primeros meses de cada año, seguido de un descenso gradual durante los cuatro últimos.

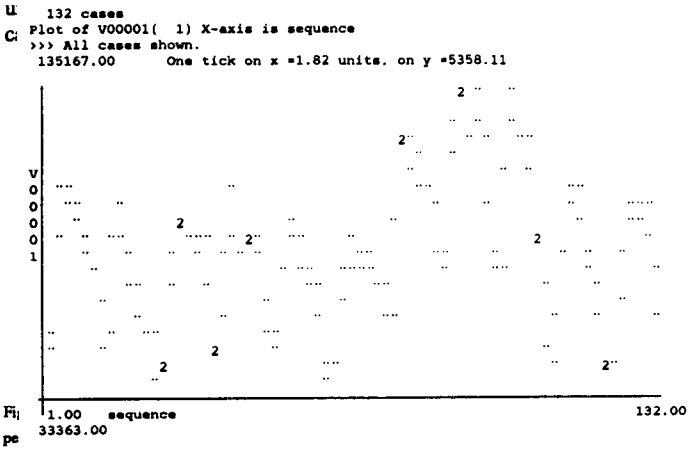


Figura 4.1. Gráfica correspondiente al número de licencias de viviendas de nueva construcción en el período 1.965 1.975 en los EEUU.

1. Todos los gráficos presentados en este capítulo han sido realizados mediante la versión 2.2 del paquete estadístico EDA de E. Horber.

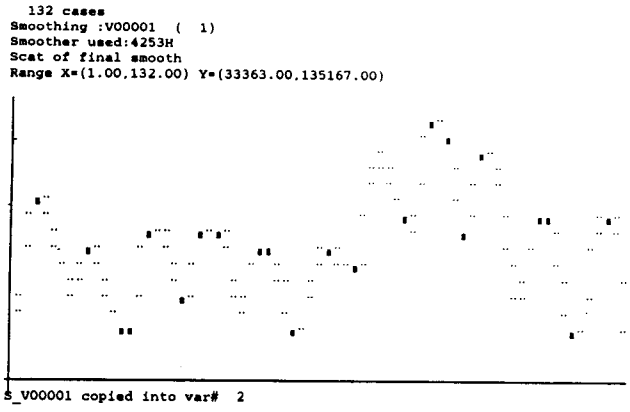


Figura 4.2. Gráfica correspondiente a los valores de la figura 4.1 suavizados.

4.2. PROCEDIMIENTOS BÁSICOS DE ALISADO

La propiedad fundamental de una secuencia alisada enunciada por Tukey, 1.977 y recogida por todos los autores que han aplicado este tipo de técnicas es que cada valor es mas similar a sus vecinos dado que los cambios en la secuencia no suelen tener lugar de forma repentina o inesperada. En este principio será el que se inspiren las diferentes técnicas de suavizado.

4.2.1. MEDIANAS MÓVILES

La justificación del nombre utilizado, **Mediana Móvil** reside por una parte en que el indicador capital del proceso de suavizado que se propone desde las técnicas de Análisis Exploratorio de Datos es el de la Mediana (índice suficientemente ponderado en cuanto a sus ventajas en el presente texto) y por otra parte en que el proceso de suavizado se concibe como un proceso dinámico que se efectua a lo largo de todo el recorrido de la variable.

4.2.1.1. MEDIANAS MÓVILES DE AMPLITUD IMPAR

El procedimiento que de forma más sencilla puede conseguir el objetivo planteado en la afirmación formulada al principio del punto 4.2, consiste en dada una secuencia de observaciones de la variable "Y" desde y_1 hasta y_n, ordenadas segun el criterio de la variable "X", se reemplaza cada valor de y_i por la Mediana de tres valores calculada a partir de el mismo valor, el siguiente y el anterior (y_{i-1}, y_i e y_{i+1}); donde "i" define el orden proporcionado por la variable "X".

Evidentemente la justificación de la utilización de la mediana reside en el hecho de la necesidad de plantearnos la consecución de una curva alisada que posea la característica de ser resistente ante la aparición de **valores anómalos** ("outliers"), característica que no poseería en el caso de plantearse las utilización de promedios para efectuar el alisado.

Desde el planteamiento que se realiza en las técnicas exploratorias es evidente que la utilización rígida de la **amplitud** ("span") tres no sería adecuada puesto que una mediana de tres valores conseguiria alisar el punto de la secuencia solamente en el caso de que no hubiera entre ellos ningún elemento anómalo o como máximo uno solo. Por ello se plantea la utilización de Medianas móviles de distintas amplitudes, por ejemplo una de amplitud cinco, conseguiría el objetivo de suavizar incluso en el caso de que se presentarán dos valores anómalos en la secuencia de cinco valores. Consecuentemente se hace patente que a mayor amplitud de la

secuencia utilizada mayor número de valores anómalos podrá soportar el proceso de suavizado.

Generalmente la notación de la operación de alisado se identifica con el dígito de su amplitud o número de valores sumarizados por la mediana (En los casos comentados "3" y "5" que son dos de los alisadores más utilizados).

El problema que se plantea es que hacer con los valores extremos de la serie, puesto que segun los criterios anteriormente planteados, no seria posible alisar el primer y último valor con el alisador "3" ni los dos primeros y los dos últimos con el alisador "5". Un primer criterio generalizado propuesto por Tukey, 1.977 consistiria en mantener inalterable en cualquiera de los dos casos el primer y último valor y con el segundo y el penúltimo proceder a la utilización de una mediana móvil "3".

Veamos en el siguiente ejemplo con datos simulados como se realizaria el proceso:

Tabla 4.1. Valores simulados de Y ordenados según el criterio de la variable X, y los correspondientes valores alisados.

Valores Originales	Mediana Móvil "3"	Mediana Móvil "5"
42	42	42
58	42	42
35	46	42
46	41	46
41	46	46
56	48	48
48	50	48
50	48	50
39	50	50
52	52	52

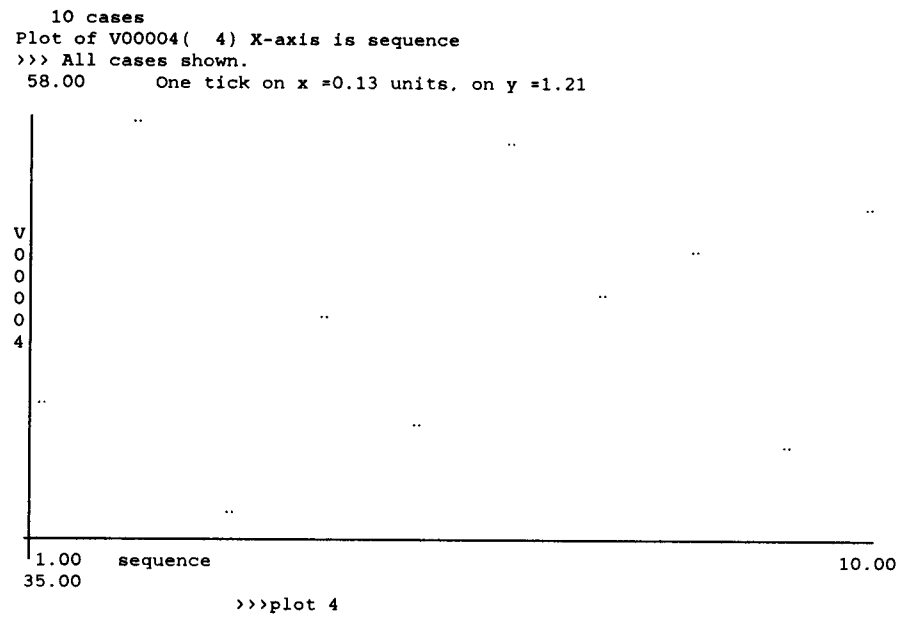


Figura 4.3. Nube de puntos correspondiente a la Variable Y ordenada según los valores de la variable X.

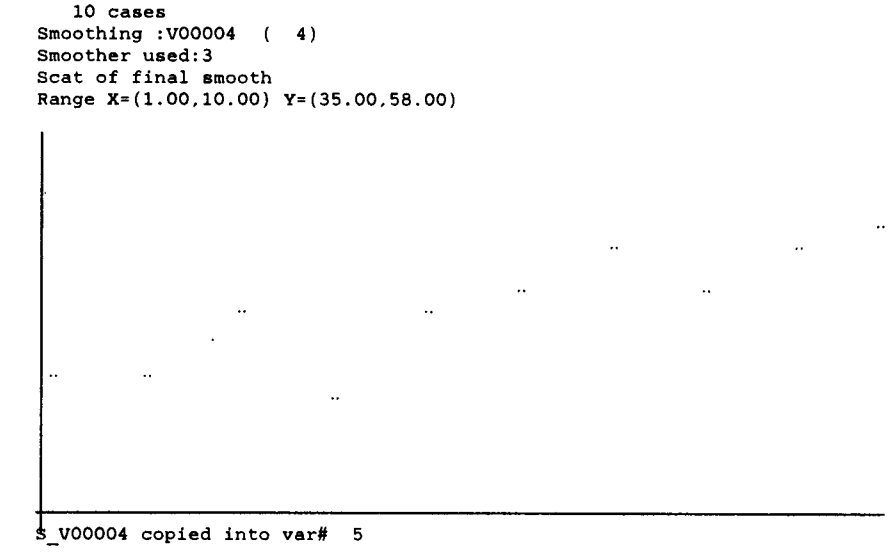
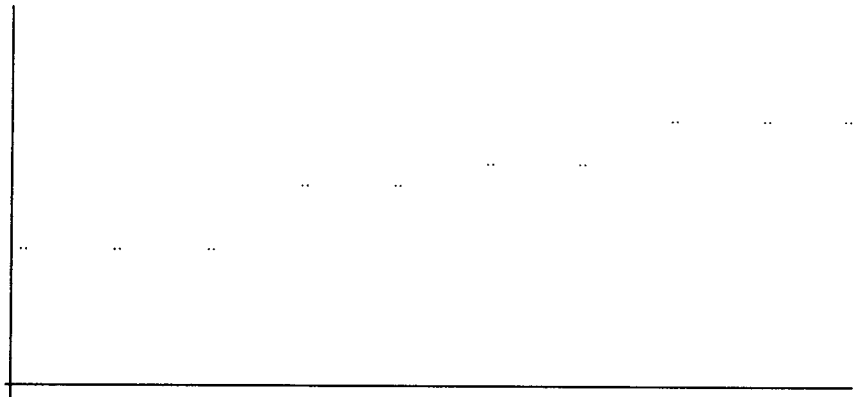


Figura 4.4. Variable Y suavizada mediante el alisador "3".

```
   10 cases
Smoothing :V00004   (  4)
Smoother used:5
Scat of final smooth
Range X=(1.00,10.00) Y=(35.00,58.00)
```



```
S_V00004 copied into var#  6
```

Figura 4.5. Variable Y suavizada mediante el alisador "5".

Logicamente puede cuestionarse el empleo de una alisador de amplitud "5" en una serie de datos tan poco numerosa como la presente, entiendase que su utilización tiene una finalidad únicamente didáctica.

4.2.1.2. MEDIANAS MÓVILES DE AMPLITUD PAR

Cabe plantearse la utilización de alisadores de amplitud de secuencia par, pero en este caso la actividad que realiza el alisador consiste en hallar el promedio de los dos valores centrales de la secuencia. Lógicamente si bien el centro natural de un segmento impar se halla situado en un valor de la variable "Y", en cambio en el caso de un segmento par se hallará en el centro de dos valores, no alineándose con ninguno de los valores de Y ordenados segun la variable X. Por lo que un par de medianas flanqueará los valores originales de la variable Y.

En esta situación se debe realizar un proceso complementario de alineamiento de los valores de la variable suavizada con los valores originales

de la variable según el órden que proporcionaba la variable X. Este es el proceso que se conoce con el nombre de **recentrado** ("recentering") que puede ser llevado a cabo aplicando a la serie suavizada un alisador de mediana móvil 2 (consistente en el promedio de cada par de puntuaciones alisadas).

Uno de los alisadores de estas características más utilizado es el que se conoce como "42",[2] en el que tras la aplicación de un alisador de mediana móvil "4" se procede a realizar el proceso de recentrado. Velleman y Hoaglin, 1.981 proporcionan una expresión algebraica para el presente alisador, que permite reemplazar cada valor de la serie original a partir de:

$$\text{Si} = 1/2\left(\text{Med}\left\{y_{i-2}, y_{i-1}, y_i, y_{i+1}\right\} + \text{Med}\left\{y_{i-1}, y_i, y_{i+1}, y_{i+2}\right\}\right) \tag{4.1}$$

Logicamente la utilización de este tipo de alisadores puede verse más afectado por la presencia de valores anómalos que los de amplitud impar, por lo que es conveniente utilizarlos cuando no existan dichos valores, o tal y como se sugiere más adelante emplearlos de forma combinada con algún alisador de amplitud impar.

Como ejemplo planteamos la variable utilizada en la tabla numero 4.1 puesto que como se observa en el diagrama de caja de esta variable en la figura 4.6 no se constata la presencia de ningún valor que se pueda calificar de anómalo. Observese en la figura 4.7. como de la aplicación de este alisador parece desprenderse de los datos un modelo de relación que se ajustaría a una tendencia cúbica.

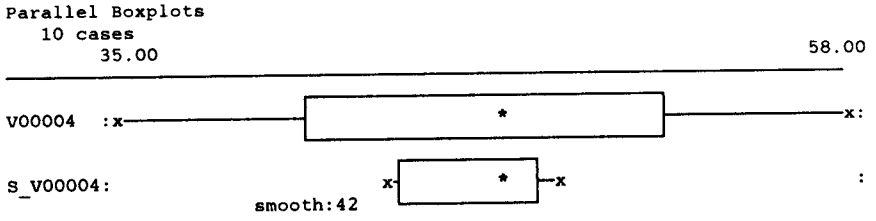


Figura 4.6. Diagrama de Caja de la variable Y y de su correspondiente versión alisada mediante el alisador "42".

2. Los dígitos utilizados para designar este alisador, no pueden inducir a confusión, puesto que las amplitudes máximas que suelen ser utilizadas son de 9 o 12.

Tabla 4.2. Alisado de la Variable Y mediante el alisador "42".

Valor Original	Valor Alisado "4"	Valor final "42"
42	42	42
58	50	47
35	44	43.75
46	43.5	43.5
41	43.5	45.25
56	47	48
48	49	49
50	49	49
39	49	47.25
52	45.5	52
	52	

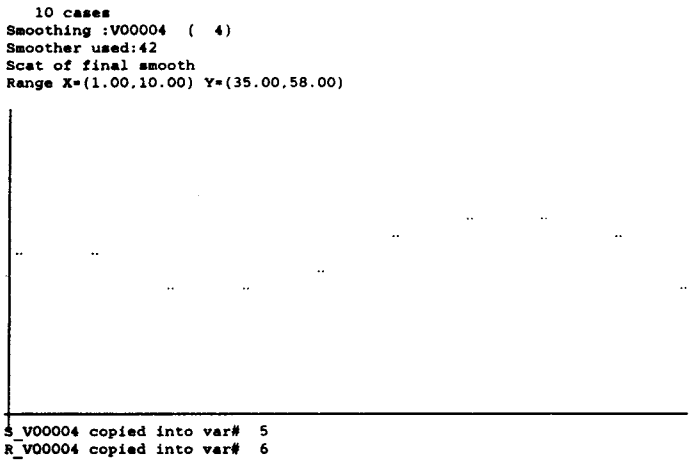


Figura 4.7. Variable Y suavizada mediante el alisador "42".

Se puede observar de los datos de la tabla 4.2 como el proceso de media móvil "4" presenta un problema adicional, puesto que genera un dato más en la secuencia, de forma que el primero y el último simplemente son copiados siguiendo la regla anteriormente propuesta, el segundo y el tercero son la mediana de la primera/segunda y penultima/última observaciones y las siete restantes son la mediana de las sucesivas agrupaciones de 4 puntuaciones, produciendo el proceso de recentrado una restauración de la secuencia a la longitud original, siendo nuevamente los valores primero y último copiados y los restantes datos, promedios de valores adyacentes.

4.2.2. ALISADO DE LOS PUNTOS FINALES

Tal y como se ha planteado anteriormente el alisado de los puntos extremos de la serie puede constituir un problema puesto que no existen suficientes valores en su entorno para que el proceso pueda ser llevado a cabo de forma correcta. La simple copia de los valores finales no puede ser considerada como una solución satisfactoria, por lo que Tukey, 1.977 propuso la **regla del alisado de los valores extremos.**

El proceso se inicia con la estimación del valor que precederia al primero de la secuencia o el que seguiria al último. En este proceso de estimación no pueden ser utilizados el primer y último valores puesto que todavía no han sido suavizados; por lo que una aproximación simple consiste en hallar la linea recta que pasaría por el segundo y tercer valor ya alisado y situar nuestra estimación en dicha linea en el lugar que hubiera ocupado de existir originalmente. Con el valor extrapolado, el valor extremo y el primer o último valor suavizado se realizaria la estimación de los valores extremos de la secuencia alisada a través de la mediana de los tres valores mencionados. Veamos la formalización de este proceso, recogida por diversos autores:

- La pendiente de la linea recta para una secuencia de datos en que los valores de "Y" fueran equidistantes según los valores de "X", se estableceria a través de:

$$(y_3 - y_2)/\triangle x \text{ para el principio de la serie} \qquad (4.2)$$

y

$$(y_{n-2} - y_{n-1})/\triangle x \text{ para el final de la serie.} \qquad (4.3)$$

- Los valores extrapolados se obtendrian a partir de:

$$y_0 = y_2 - 2\triangle x(y_3 - y_2)/\triangle x = 3y_2 - 2y_3 \qquad (4.4)$$

y de forma analoga para el final de la serie

$$y_{n+1} = 3y_{n-1} - 2y_{n-2} \qquad (4.5)$$

– Estableciéndose los valores extremos alisados a través de:

$$S_1 = \text{Med}\{y_0, y_1, S_2\} \quad (4.6)$$

$$S_n = \text{Med}\{y_{n+1}, y_n, S_{n-1}\} \quad (4.7)$$

Para los datos planteados en la tabla 4.1, utilizando un alisador "3":

$y_0 = 3(42) - 2(46) = 34 \quad y_{n+1} = 3(50) - 2(48) = 54$
$S_1 = \text{Med}\{34, 42, 42\} = 42 \quad S_n = \text{Med}\{54, 52, 50\} = 50$

4.2.3. HANNING ("h")

Con los alisadores planteados anteriormente se podria realizar una escala de dureza en cuanto a la fuerza con que actuan, siendo en este caso los "42" los más suaves. Pero se puede estar interesado en conseguir todavia un efecto más suave sobre los datos. En este caso se plantearia la utilización de un **Promedio Móvil Ponderado**. Este es ya un proceso que se puede clasificar como tradicional, y su objetivo es reemplazar cada dato observado por un promedio con diversos coeficientes de ponderación para los valores de la amplitud deseada. Este proceso se conoce como Hanning puesto que es debido a un metereólogo austríaco del siglo pasado llamado Julius Von Hann. Que propuso la siguiente expresión:

$$S_i(H) = (1/4)y_{i-1} + (1/2)y_i + (1/4)y_{i+1} \quad (4.8)$$

Se han desarrollado diferentes extensiones de este tipo de alisadores que implican diferentes tipos de ponderamiento y o bien diferente número de puntos que intervienen en el ponderamiento, mereciendose destacar los mencionados por Rappachi, 1.991 y E. Horber,1.991.

4.3. PROCEDIMIENTOS SOFISTICADOS DE SUAVIZADO

4.3.1. MEDIANAS MÓVILES REPETIDAS, PROCEDIMIENTO DE CORTADO Y ALISADORES COMPUESTOS

Para evitar subsecuencias llanas muy largas generalmente se sugiere evitar la utilización de alisadores de amplitud muy grande, prefiriéndose para conseguir una secuencia suave la utilización de un alisador "3" o incluso "5" de forma repetida, iterando hasta que la secuencia alisada se mantenga sin cambios; conociéndose estos alisadores como "$3R$" y "$5R$". Con este tipo de alisador se produce sin embargo el problema de que existe una tendencia a cortar los picos y valles de forma que presenten un aspecto de un pequeño alisamiento de dos puntos. Para evitarlo Tukey, 1.977 introduce el procedimiento de **cortado** ("splitting") denominada "S", en que dicho par de valores son tratados de forma similar a la regla de los valores extremos, intentando recoger en la secuencia final alisada su aspecto de pico o valle.

Evidentemente no solo cabe plantearse el encadenamiento de la misma amplitud de alisamiento, en muchos casos será más efectiva la construcción de un **alisador compuesto** en el que se combinen alisadores de diferente amplitud, combinándolos incluso con los promedios móviles ponderados anteriormente mencionados. De entre los más utilizados cabe mencionar "$4253h$", "$3RSSh$" en que se van encadenando los diferentes alisadores indicados por sus dígitos.[3]

Ciertamente las combinaciones de alisadores compuestos que se pueden realizar son infinitas, y dado que nos hallamos en un contexto exploratorio, para cada caso práctico es recomendable la prueba tentativa de más de uno para hallar el más adecuado. Por otra parte en los alisadores compuestos generalmente para el tratamiento de los valores extremos no se sigue la regla planteada en el punto 4.2.2 hata la última secuencia, utilizandose anteriormente la simple copia de estos.

3. Quizás sea conveniente explicitar el segundo de los mencionados en que el proceso sería $3R, S, 3R, S, 3R, h$.

4.3.2. REAPROXIMANDO ("REROUGHING")

Otro proceso complementario introducido por Tukey, 1.977 es el del **Reaproximado** "reroughing". Su misión consiste en intentar recapturar de los residuales algún patrón que un proceso de alisado demasiado fuerte hubiera hecho desaparecer de la parte suavizada, para añadirla a esta de forma que la parte suavizada sea mas similar a la secuencia original.

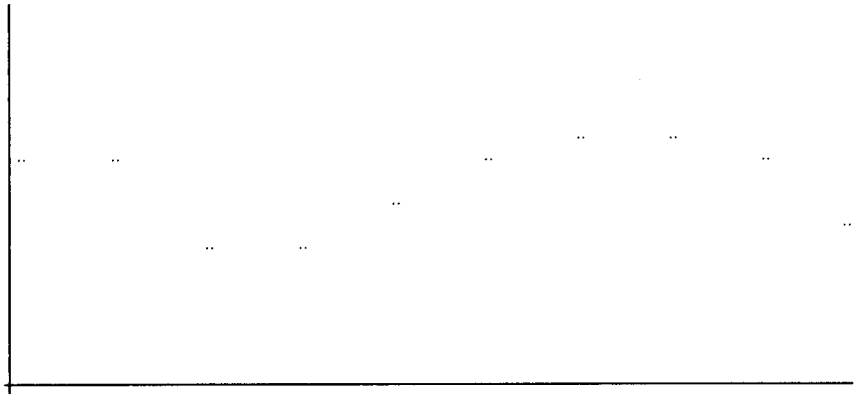
El proceso consiste en efectuar sobre el residual obtenido al final del proceso de suavizado de la secuencia original, el mismo alisador que con esta. Generándose la secuencia alisada definitiva al sumarse la secuencia alisada de residuales generada con este segundo proceso a la secuencia alisada que habia generado los primeros residuales. La notación empleada para representar este proceso de reaproximado es la palabra "twice" o la letra "t" siguiendo detrás de una coma a la secuencia del alisadr utilizado. Este tipo de proceso es bastante generalizado en las técnicas EDA, presentándose procesos similares en las técnicas de la Linea Resistente y el Análisis de Medianas.

En la tabla 4.3 se presenta como se realizaría este proceso a partir de la variable "Y"; la segunda columna presenta los datos de dicha variable después de haberlos alisado mediante un alisador de amplitud cuatro con el correspondiente proceso de recentrado; siendo la tercera columna la diferencia entre el valcr original y la secuencia alisada, a partir de la cual se genera la cuarta columna utilizando el mismo alisador; la quinta columna, resultado de sumar la cuarta a la segunda sería la secuencia resultante después del proceso de reaproximado. Observese asimismo el resultado de este proceso en la figura 4.8, y evalúese comparativamente esta con la figura 4.7 en la que se había efectuado el mismo proceso de alisado pero sin la operación de reaproximado.

Tabla 4.3. Proceso de Reaproximado de la Variable Y.

Valor Original	Valor Alisado "42"	Residual	Residual Alisado "42"	Valor Alisado "42, twice"	Residual Final
42	47.00	-5.00	0.88	47.88	-5.88
58	47.00	11.00	0.88	47.88	10.13
35	43.75	-8.75	-1.06	42.69	-7.69
46	43.50	2.50	-0.88	42.63	3.38
41	45.25	-4.25	-0.06	45.19	-4.19
56	48.00	8.00	0.38	48.38	7.63
48	49.00	-1.00	0.00	49.00	-1.00
50	49.00	1.00	0.00	49.00	1.00
39	47.25	-8.25	0.00	47.25	-8.25
52	43.75	8.25	0.00	43.75	8.25

```
   10 cases
Smoothing :V00004   (  4)
Smoother used:42
Scat of final smooth
Range X=(1.00,10.00) Y=(35.00,58.00)
```



```
S_V00004 copied into var#  9
R_V00004 copied into var# 10
                  >>>smooth 4 "42" twice smooth=9 rough=10
```

Fig.4.8. Variable Y suavizada mediante "42,*t*".

4.4. ANÁLISIS DE LOS RESIDUALES

Ya se ha comentado largamente, en otros capitulos y en este la importancia del termino residual (o rugoso tal y como lo hemos denominado aquí) puesto que en muchos casos será más interesante su analisis que el del propio ajuste o alisado. Así este componente nos puede revelar la presencia de valores anómalos, porciones de la secuencia que parecen estar sujetas a largas fluctuaciones, etc.

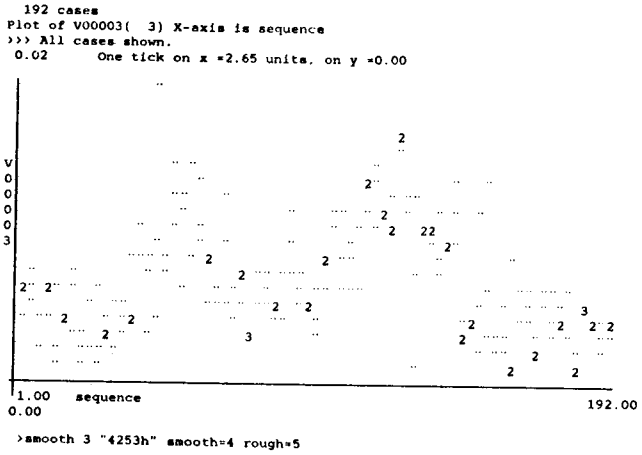


Figura 4.9. Tasa de inflación mensual en los EEUU 1.970-1.985.

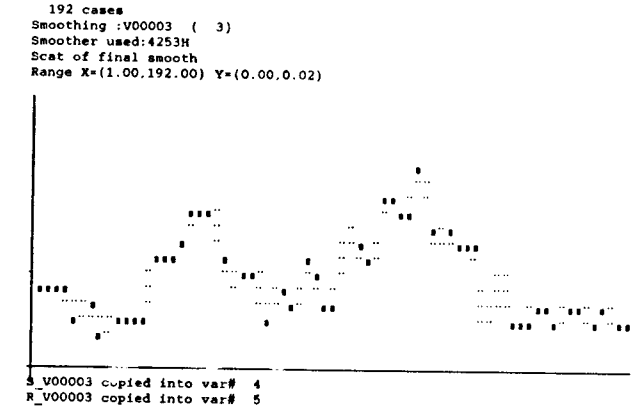


Figura 4.10. Suavizado de la tasa de Inflación en los EEUU "4253h".

```
  >plot 5

  192 cases
Plot of R_V00003(  5) X-axis is sequence
>>> All cases shown.
 0.01       One tick on x =2.65 units, on y =0.00
```

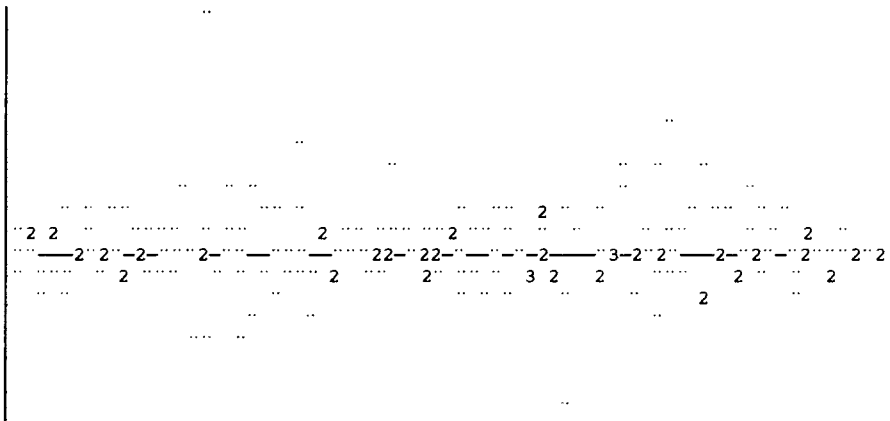


Figura 4.11. Gráfico de residuales de la tasa de inflación de los EEUU después de la aplicación del alisador "4253h".

```
  >boxplot 5 7 parallel

Parallel Boxplots
  192 cases
Boxplot :R_V00003(  5)                          rough :4253H
        -0.01                                                    0.01

@          @@ @  oox-------|  *  |-------xooo@ @@@       @              @
                  22                      2   2
Extreme values (LO,HI): 127  44   adjacent(LO,HI): 142  169
Lo outliers:127  52   43   45   55   147  68   8
Hi outliers:56   39   139  50   141  160  86   149  67   150  44

  >boxplot 7

  192 cases
Boxplot :R_V00003(  7)                          rough :4253H,t
        -0.01                                                    0.01

@        @   @ @@ooo x-------|  *  |-------oo @ @ @    @                @
               2                          32 2 2
Extreme values (LO,HI): 127  44   adjacent(LO,HI): 104  156
Lo outliers:127  52   43   45   68   55   147  8    142
Hi outliers:169  26   134  139  50   141  149  160  86   67   150  44
```

Figura 4.12. Diagramas de caja de los residuales de la tasa de inflación de los EEUU después de la aplicación de "4253h" y "4253h, Twice".

Observese por ejemplo la serie compuesta por los valores de inflación mensual desde Enero de 1.970 hasta Diciembre de 1.985 en los EEUU. Las figuras 4.9 y 4.10 representan las gráficas de la serie original y la secuencia de los datos alisados mediante "4253h", pero más interesante es la contemplación de las figuras 4.11 que recoge la distribución de los residuales y la 4.12 que presenta el diagrama de caja de los mencionados residuales, así como el de los residuales resultantes de aplicar el alisador "4253h,twice"; constatándose la inestabilidad de la inflación en los años 80 y la aparición de diversos valores anómalos.

Por otra parte el estudio de los residuales, generados mediante cualquier otra técnica de modelado mediante el proceso del alisado, también puede ser interesante puesto que nos proporcionará una ayuda para el estudio de las características de este componente (por ejemplo para el estudio de condiciones de aplicación). Así, y utilizando el residual obtenido mediante la técnica de la linea resistente en el capitulo anterior (ver tabla 3.2) se han generado las figuras 4.13 y 4.14 donde se observa respectivamente la distribución de residuales y los mismos alisados mediante "4253h", constatándose que efectivamente parecen ajustarse a un proceso estacionario con media cero.

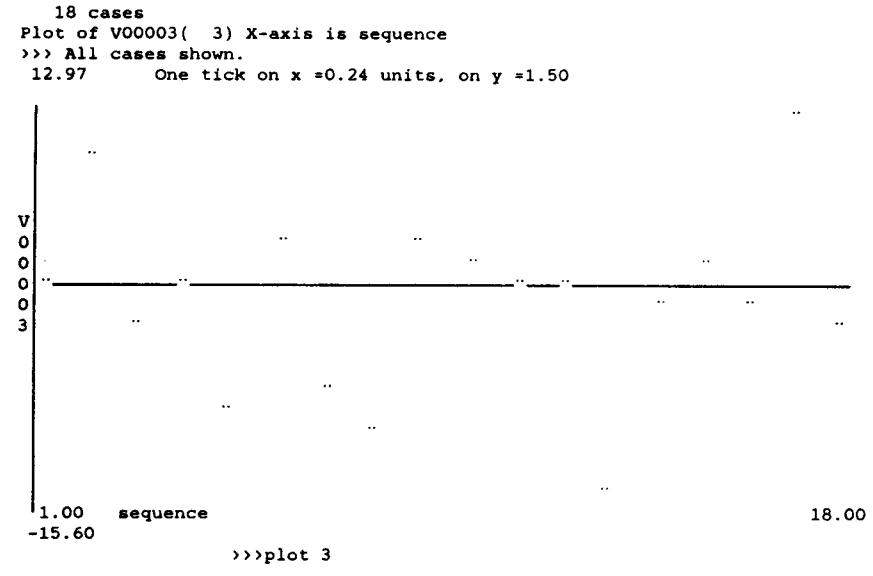
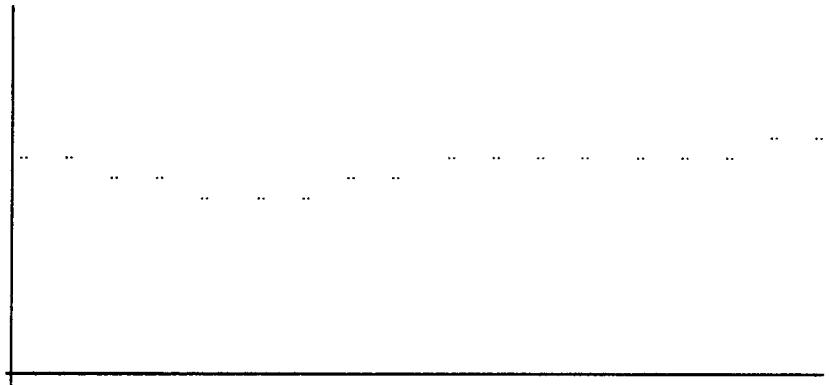


Figura 4.13. Gráfico de los residuales correspondientes a la tabla 3.2.

```
   18 cases
Smoothing :V00003   (  3)
Smoother used:4253H
Scat of final smooth
Range X=(1.00,18.00) Y=(-15.60,12.97)
```



```
S_V00003 copied into var#  4
››› Warning: var#  4 overwritten
R_V00003 copied into var#  4
                    ›››smooth 3 "4253h" rough=4
```

Figura 4.14. Residuales suavizados mediante el alisador "4253h".

4.5. ANÁLISIS DE SERIES TEMPORALES MEDIANTE LA TÉCNICA DEL SUAVIZADO

En este apartado pretendemos presentar algunos ejemplos de datos reales que puedan ser manipulados por el lector para estudiar el patrón subyacente en ellos mediante las técnicas de suavizado y empleando el Paquete estadístico EDA de E. Horber. Dado que las técnicas de alisado son de "las más tediosas" de realizar a mano y teniendo en cuenta que generalmente que la utilización de alisadores compuestos mejora la consecución de resultados, nos unimos a la consideración de Velleman y Hoaglin, 1.981 entorno a que para esta técnica es especialmente adecuada su implementación y resolución mediante ordenador. Se presentan aparte de las tablas de datos, algunas gráficas y comentarios que en absoluto pretenden ser exhaustivos ni definitivos, animándose al lector a que realice sus propios análisis y comentarios.

CASO PRIMERO: Consumo de Energía eléctrica en el sector industrial del País Valenciano en millones de Kilowatios desde Enero de 1.976 hasta Marzo de 1.984.

Observese en la Fig. 4.15 como el alisador utilizado muestra en los datos un patron cíclico, probablemente de carácter anual, pero asimismo también se puede determinar que el proceso no es estacionario sino que presenta una tendencia creciente.

Tabla 4.4. Datos correspondientes al Caso Primero.

	1.976	1.977	1.978	1.979	1.980
ENERO	292	366	334	336	405
FEBRERO	304	309	362	349	375
MARZO	298	353	321	344	355
ABRIL	316	338	354	355	370
MAYO	292	346	354	380	379
JUNIO	282	330	349	392	357
JULIO	370	377	382	378	369
AGOSTO	317	327	373	363	394
SEPTIEMBRE	323	321	348	327	344
OCTUBRE	325	348	368	389	387
NOVIEMBRE	363	355	398	384	403
DICIEMBRE	346	356	361	390	376
	1.981	1.982	1.983	1.984	
ENERO	384	391	368	383	
FEBRERO	370	364	362	384	
MARZO	384	358	381	421	
ABRIL	376	389	392		
MAYO	367	371	376		
JUNIO	375	391	410		
JULIO	417	391	411		
AGOSTO	399	439	423		
SEPTIEMBRE	363	353	367		
OCTUBRE	398	421	431		
NOVIEMBRE	397	363	421		
DICIEMBRE	396	315	392		

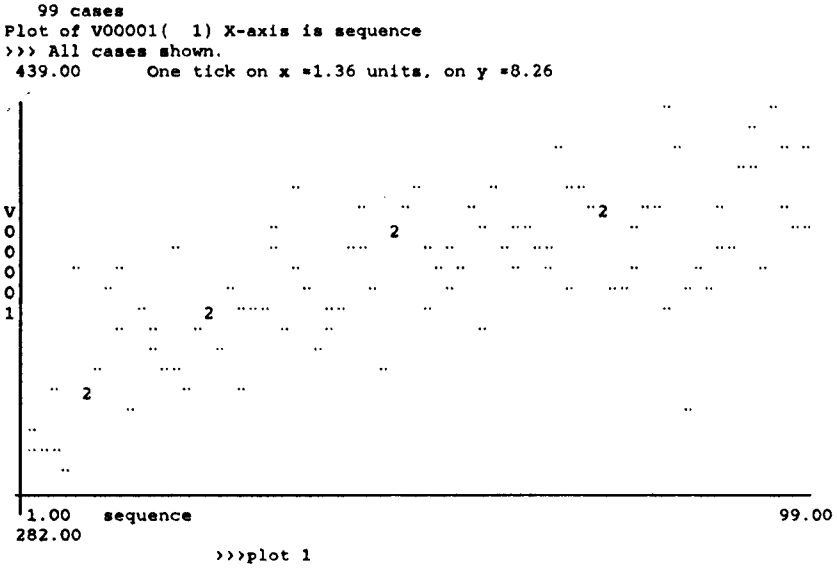


Fig.4.15. Gráfica del consumo de energía eléctrica en el sector industrial de la comunidad Valenciana.

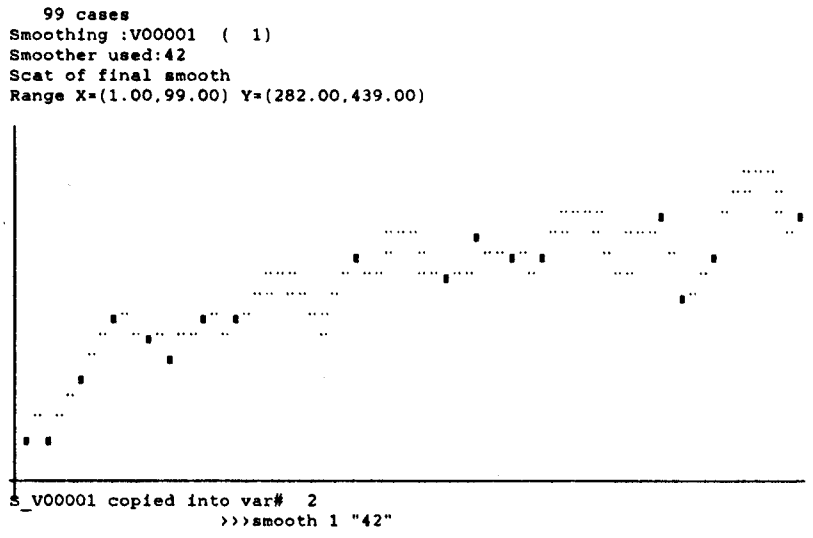


Fig. 4.16. Variable Consumo de Energía eléctrica suavizada mediante el alisador "42".

CASO SEGUNDO: Evolución del inventario diario de los distribuidores de potencia utilizados en impresoras de ordenador.

En este caso comparando las figuras 4.18 y 4.19 se puede constatar como en muchos casos la utilización de diferentes alisadores cada vez más complicados no modifican sustancialmente la gráfica obtenida.[4]

Tabla 4.5. Datos correspondientes al Caso Segundo.

001	1008	031	964	061	984	091	997	121	1015
002	1015	032	964	062	992	092	990	122	1014
003	1006	033	976	063	967	093	993	123	1016
004	1017	034	985	064	966	094	988	124	995
005	1015	035	997	065	970	095	1004	125	1002
006	1006	036	1008	066	982	096	995	126	993
007	1013	037	999	067	995	097	995	127	1018
008	1009	038	1003	068	981	098	1011	128	1008
009	1011	039	1024	069	990	099	991	129	1017
010	1009	040	1029	070	1003	100	994	130	996
011	995	041	1036	071	1005	101	991	131	1005
012	1024	042	1049	072	1016	102	999	132	1014
013	1008	043	1039	073	1028	103	995	133	1003
014	999	044	1040	074	1003	104	996	134	1000
015	997	045	1019	075	993	105	1002	135	1018
016	999	046	1007	076	995	106	1010	136	1002
017	1004	047	1008	077	1003	107	1016	137	997
018	1001	048	1020	078	1003	108	1017	138	1002
019	1003	049	1022	079	1013	109	1029	139	1007
020	1018	050	1023	080	1020	110	1036	140	990
021	1026	051	1031	081	1019	111	1024	141	994
022	1015	052	1011	082	1014	112	1022	142	1005
023	1019	053	1019	083	1014	113	1021	143	1012
024	1034	054	1009	084	1018	114	1019	144	1012
025	1033	055	1005	085	998	115	1009	145	1028
026	1021	056	1012	086	996	116	1021	146	1002
027	1017	057	1017	087	994	117	1025	147	998
028	1008	058	1000	088	983	118	1020	148	1029
029	1009	059	997	089	994	119	1021	149	1026
030	984	060	983	090	992	120	1030		

4. Gráficas muy similares a estas se han obtenido mediante el alisador "3RSSh" o empleando la técnica complementaria del reaproximado.

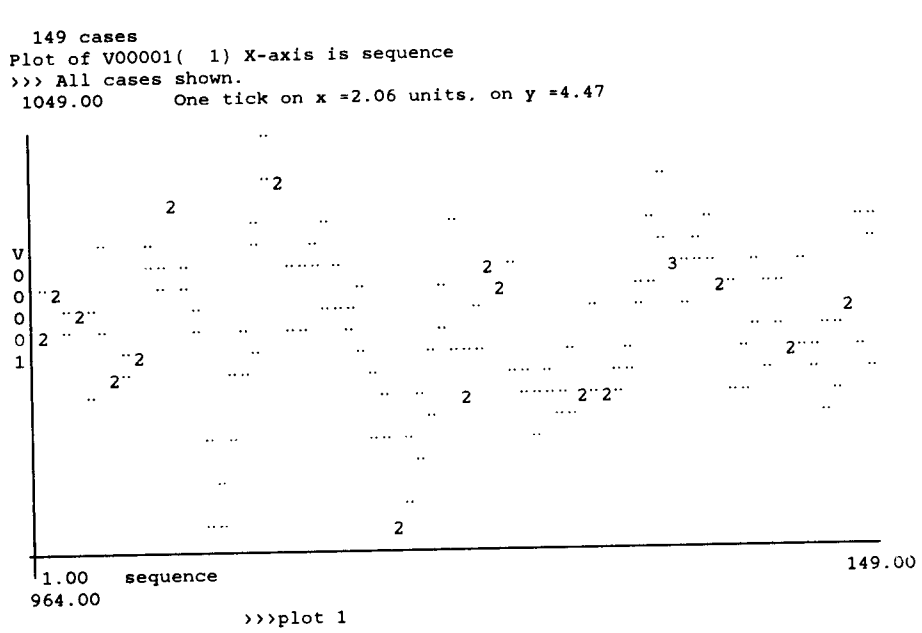


Fig.4.17. Gráfica de los datos del inventario tomados durante 149 dias.

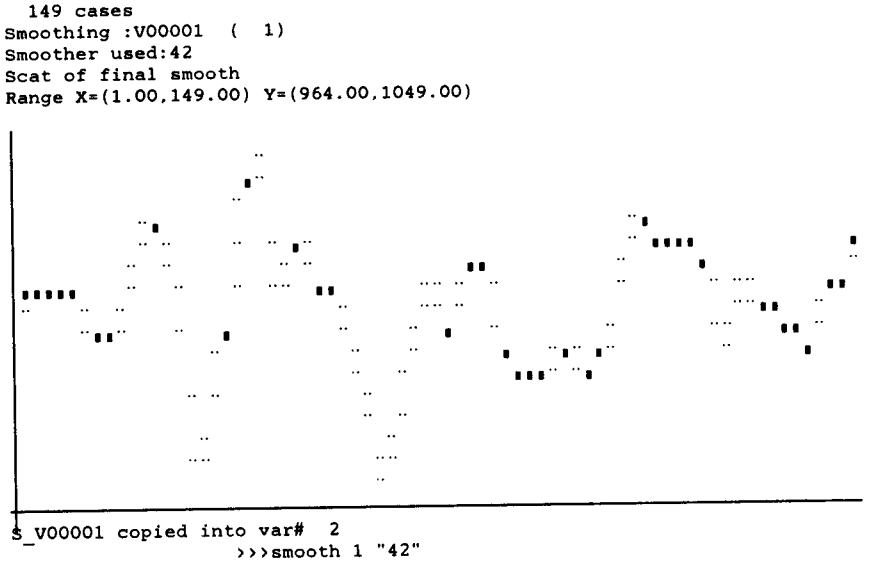
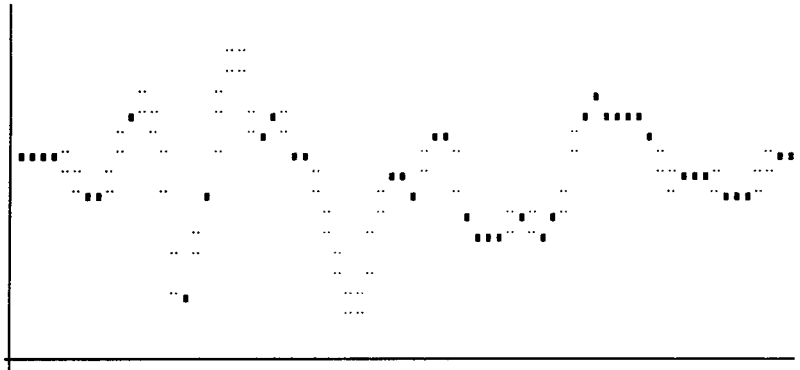


Fig. 4.18. Aplicación del alisador "42" a los datos del caso segundo.

```
  149 cases
Smoothing :V00001   (  1)
Smoother used:4253H
Scat of final smooth
Range X=(1.00,149.00) Y=(964.00,1049.00)
```



```
S_V00001 copied into var#  3
                >>>smooth 1 "4253h"
```

Fig. 4.19. Aplicación del alisador "4253*h*" a los datos del caso segundo.

CASO TERCERO: Medida del nivel mensual de ozono desde Octubre de 1.973 hasta Septiembre de 1.983 en la Bahia de Hudson, EEUU.

Este ejemplo puede ser especialmente interesante para observar como también se puede aplicar la técnica en el caso de "**missings**", y viendo en concreto como los trata el paquete estadístico uilizado.

Tabla 4.6. Datos correspondientes al Caso Tercero.

	1.973	1.974	1.975	1.976	1.977	1.978
ENERO		43.3	43.7			34.4
FEBRERO		42.9	22.7		34.4	
MARZO		64.0	44.7		58.3	50.6
ABRIL		40.1	49.7	39.6	28.6	39.1
MAYO		34.4	36.0	28.1	25.2	28.3
JUNIO		24.0	29.9	27.5	34.3	28.4
JULIO		24.4	28.7	23.3	27.9	25.0

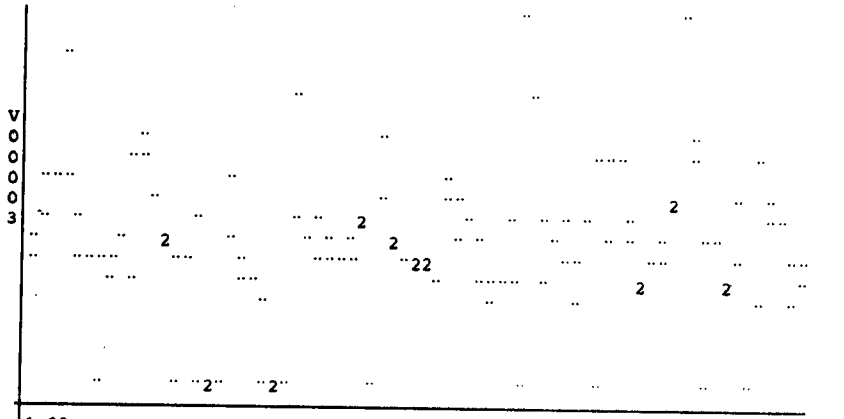
.../...

.../...

AGOSTO			24.9	20.6	27.0	25.1
SEPTIEMBRE		25.6		19.7	25.1	24.9
OCTUBRE	28.2	24.8	25.9		27.6	25.6
NOVIEMBRE	26.2	22.0			30.1	26.9
DICIEMBRE	33.2	30.4	31.8		31.9	22.8

	1.979	1.980	1.981	1.982	1.983
ENERO	38.5			36.6	18.8
FEBRERO	41.4	75.2	31.5	72.5	43.9
MARZO	31.1	59.2	44.1	47.9	37.5
ABRIL	37.8	34.1	44.1	45.3	32.6
MAYO	32.7	23.7	34.3		32.1
JUNI	28.2	31.1	29.6	29.2	17.5
JULIO	22.5	31.9	21.9	27.8	23.9
AGOSTO	19.5	25.2	22.7	22.0	23.7
SEPTIEMBRE	21.4	26.0	23.9	19.9	24.0
OCTUBRE	23.2	17.0	25.9	27.0	
NOVIEMBRE	20.1	34.2	31.4	38.3	
DICIEMBRE	33.9	45.1	37.1		

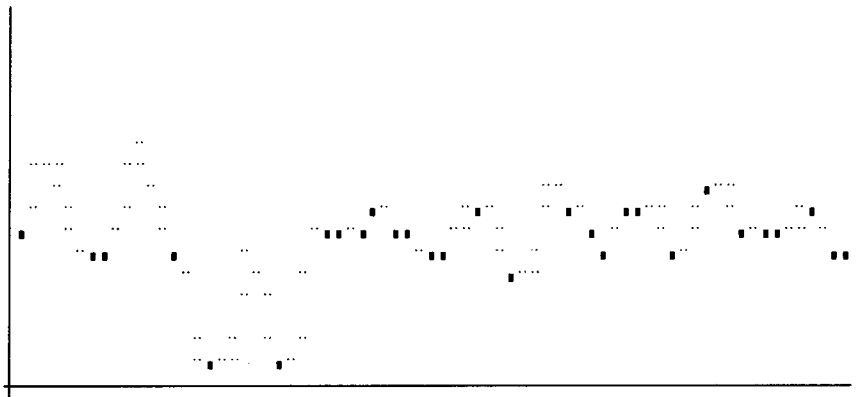
```
  120 cases
Plot of V00003(  3) X-axis is sequence
››› All cases shown.
 75.20       One tick on x =1.65 units, on y =3.96
```




```
›››plot 3
```

Fig. 4.20. Gráfico del nivel mensual de ozono en la atmósfera.

```
  120 cases
Smoothing :V00003   (  3)
Smoother used:3RSSH
Scat of final smooth
Range X=(1.00,120.00) Y=(0.00,75.20)
```



```
S_V00003 copied into var#  6
R_V00003 copied into var#  7
                >>>smooth 3 "3rssh" smooth=6 rough=7
```

Fig. 4.21. Aplicación del alisador "3RSS*h*" a los datos del caso tercero.

Tabla 4.7. Datos Originales, alisados y residuales del Caso Tercero.

	Valor Original	Alisad. "3RSSh"	Residual		Valor Original	Alisad. "3RSSh"	Residual
1	28.20	28.20	0.00	61	25.60	25.48	0.13
2	26.20	29.45	-3.25	62	26.90	25.92	0.98
3	33.20	34.38	-1.17	63	22.80	27.22	-4.42
4	43.30	40.58	2.72	64	38.50	29.65	8.85
5	42.90	43.10	-0.20	65	41.40	31.90	9.50
6	64.00	42.30	21.70	66	31.10	32.70	-1.60
7	40.10	39.38	0.72	67	37.80	32.70	5.10
8	34.40	33.33	1.08	68	32.70	31.58	1.13
9	24.00	26.85	-2.85	69	28.20	27.90	0.30
10	24.40	24.30	0.10	70	22.50	23.65	-1.15
11	0.00	24.45	-24.45	71	19.50	21.67	-2.17
12	25.60	24.70	0.90	72	21.40	21.40	0.00
13	24.80	24.80	0.00	73	23.20	21.40	1.80

.../...

.../...

14	22.00	26.20	-4.20	74	20.10	21.85	-1.75
15	30.40	29.00	1.40	75	33.90	25.43	8.48
16	43.70	33.72	9.98	76	0.00	32.78	-32.78
17	22.70	40.63	-17.92	77	75.20	38.55	36.65
18	44.70	44.45	0.25	78	59.20	38.60	20.60
19	49.70	42.53	7.17	79	34.10	34.85	-0.75
20	36.00	36.65	-0.65	80	23.70	31.85	-8.15
21	29.90	31.13	-1.23	81	31.10	31.10	0.00
22	28.70	28.05	0.65	82	31.90	29.83	2.07
23	24.90	25.85	-0.95	83	25.20	27.08	-1.88
24	0.00	24.90	-24.90	84	26.00	25.60	0.40
25	25.90	18.67	7.23	85	17.00	26.20	-9.20
26	0.00	6.22	-6.22	86	34.20	28.17	6.03
27	31.80	0.00	31.80	87	45.10	31.20	13.90
28	0.00	0.00	0.00	88	0.00	33.53	-33.53
29	0.00	0.00	0.00	89	31.50	34.22	-2.72
30	0.00	0.00	0.00	90	44.10	34.28	9.82
31	39.60	0.00	39.60	91	44.10	34.30	9.80
32	28.10	5.82	22.28	92	34.30	33.13	1.17
33	27.50	17.47	10.03	93	29.60	29.05	0.55
34	23.30	22.63	0.67	94	21.90	24.43	-2.53
35	20.60	21.05	-0.45	95	22.70	23.00	-0.30
36	19.70	15.00	4.70	96	23.90	24.10	-0.20
37	0.00	4.93	-4.93	97	25.90	26.77	-0.88
38	0.00	0.00	0.00	98	31.40	31.32	0.08
39	0.00	0.00	0.00	99	37.10	35.42	1.67
40	0.00	0.00	0.00	100	36.60	37.22	-0.63
41	34.40	7.15	27.25	101	72.50	37.85	34.65
42	58.30	21.45	36.85	102	47.90	36.58	11.33
43	28.60	28.60	0.00	103	45.30	32.83	12.47
44	25.20	28.42	-3.22	104	0.00	29.55	-29.55
45	34.30	28.08	6.22	105	29.20	28.15	1.05

.../...

.../...

46	27.90	27.67	0.23	106	27.80	27.80	0.00
47	27.00	27.11	-0.11	107	22.00	27.60	-5.60
48	25.10	26.92	-1.82	108	19.90	27.20	-7.30
49	27.60	27.96	-0.36	109	27.00	27.00	0.00
50	30.10	29.92	0.18	110	38.30	27.00	11.30
51	31.90	31.45	0.45	111	0.00	27.00	-27.00
52	34.40	31.90	2.50	112	18.80	28.40	-9.60
53	0.00	31.17	-31.17	113	43.90	31.20	12.70
54	50.60	29.63	20.97	114	37.50	32.60	4.90
55	39.10	28.65	10.45	115	32.60	32.47	0.13
56	28.30	28.42	-0.13	116	32.10	30.17	1.92
57	28.40	27.52	0.88	117	17.50	25.94	-8.44
58	25.00	25.88	-0.88	118	23.90	23.88	0.02
59	25.10	25.05	0.05	119	23.70	23.91	-0.21
60	24.90	25.20	-0.30	120	24.00	23.99	0.01

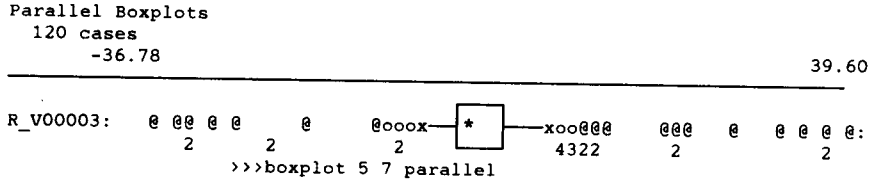


Fig. 4.22. Diagrama de caja de los residuales correspondientes al caso tercero.

CASO CUARTO: Índice de la producción Industrial desde Enero de 1.965 hasta Marzo de 1.984, en España.

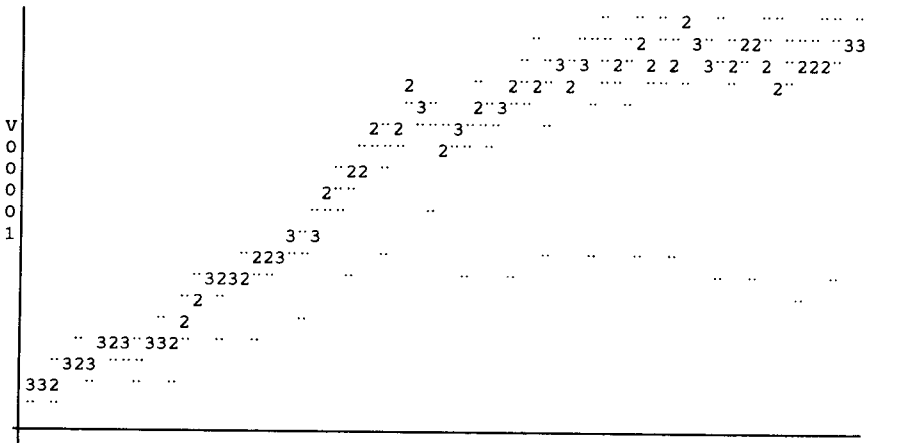
Tabla 4.8. Datos correspondientes al Caso Cuarto.

	1.965	1.966	1.967	1.968	1.969	1.970	1.971
ENERO	46.94	54.99	60.04	62.78	72.91	80.15	82.44
FEBRERO	49.26	56.14	57.85	63.86	71.11	80.85	83.84
MARZO	51.89	60.95	61.57	64.99	78.54	79.78	88.88
ABRIL	50.54	58.22	62.81	63.84	76.53	86.03	86.29
MAYO	52.11	59.45	61.92	67.08	78.43	81.90	87.54
JUNIO	50.26	58.94	62.92	64.86	76.56	83.02	89.03
JULIO	51.48	55.37	58.93	63.22	73.48	80.94	87.25
AGOSTO	43.24	49.17	50.32	53.10	61.88	64.37	69.68
SEPTIEMBRE	51.40	59.63	59.38	64.64	76.55	80.06	86.39
OCTUBRE	52.85	61.21	60.20	68.91	80.49	83.04	88.47
NOVIEMBRE	54.63	61.19	62.23	69.75	77.57	85.61	91.37
DICIEMBRE	55.21	63.51	62.25	71.85	78.55	82.31	91.03

	1.972	1.973	1.974	1.975	1.976	1.977	1.978
ENERO	94.33	110.28	126.10	111.78	112.15	122.97	129.98
FEBRERO	97.04	108.40	121.78	114.10	115.39	128.75	130.36
MARZO	103.08	115.37	126.32	119.80	124.20	140.77	133.69
ABRIL	99.06	109.39	121.18	118.44	122.69	128.47	131.67
MAYO	102.27	115.85	124.68	117.42	125.27	135.62	133.00
JUNIO	105.35	117.09	124.18	116.66	122.47	128.91	137.78
JULIO	97.70	108.13	115.35	111.19	124.11	119.52	123.85
AGOSTO	78.51	85.90	97.74	77.52	81.25	82.47	84.67
SEPTIEMBRE	102.77	111.32	119.13	120.13	128.08	133.90	136.75
OCTUBRE	104.77	116.73	124.74	129.47	130.57	134.00	141.72
NOVIEMBRE	108.71	119.58	118.35	121.57	132.91	133.17	142.81
DICIEMBRE	106.40	114.23	112.10	118.82	127.75	134.74	131.90

	1.979	1.980	1.980	1.981	1.982	1.983	1.984
ENERO	129.98	134.78	137.66	130.37	126.20	133.50	137.40
FEBRERO	130.36	125.91	137.24	132.74	129.30	133.00	137.80
MARZO	133.69	137.56	139.11	140.15	142.50	144.60	143.30
ABRIL	131.67	126.56	132.53	132.32	130.60	134.90	
MAYO	133.00	142.59	139.36	136.88	137.30	142.60	
JUNIO	137.78	136.94	133.49	136.76	133.30	138.70	
JULIO	123.85	131.02	133.02	139.54	134.50	131.60	
AGOSTO	84.67	86.25	79.88	76.94	75.80	79.70	
SEPTIEMBRE	136.75	133.68	135.84	133.49	139.10	140.10	
OCTUBRE	141.72	142.02	147.39	142.30	135.70	138.60	
NOVIEMBRE	142.81	144.82	140.84	139.40	140.10	141.70	
DICIEMBRE	131.90	128.34	133.40	132.90	131.60	138.30	

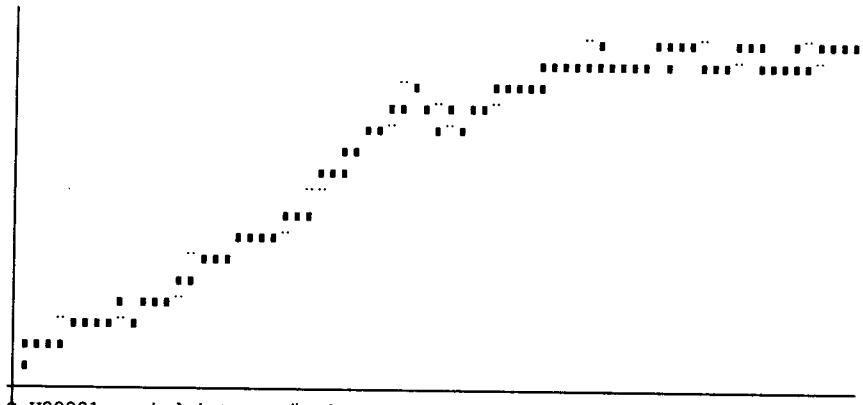
```
  243 cases
Plot of V00001(  1) X-axis is sequence
››› All cases shown.
 147.39       One tick on x =3.36 units, on y =5.48
```




```
                ›››plot 1
```

Fig. 4.23. Gráfico de los datos correspondientes al caso cuarto.

```
  243 cases
Smoothing :V00001   (  1)
Smoother used:3RSSH
Scat of final smooth
Range X=(1.00,243.00) Y=(43.24,147.39)
```



```
S_V00001 copied into var#  6
R_V00001 copied into var#  7
                ›››smooth 1 "3rssh" smooth=6 rough=7
```

Fig. 4.24. Secuencia alisada mediante "$3RSSh$" del caso cuarto.

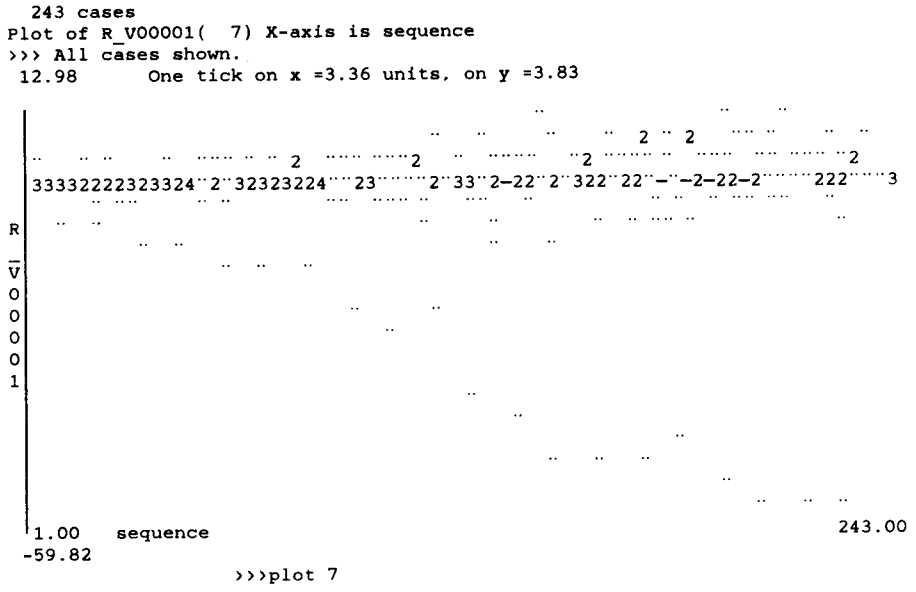


Fig.4.25. Residuales correspondientes a los datos del caso cuarto.

CASO QUINTO: Ensayo indiviual en un registro cerebral del Componente CNV (Variación Negativa Contingente) de un Potencial Evocado. Los datos corresponden al registro realizado durante un segundo, habiéndose obtenido un dato cada cuatro milisegundos.

En este tipo de datos (en general en todos los provinientes de registros electroencefalográficos) la técnica puede presentarse como especialmente útil, puesto que son procesos en los que se presenta una gran cantidad de ruido. Generalmente con este tipo de datos para hacer patentes los componentes que se dan es necesario realizar el registro de un gran número de ensayos para su posterior promediado para realzar los componentes de señal, haciéndose difícil, si no imposible, el análisis de los ensayos individuales. Un proceso de suavizado de los ensayos individuales puede permitir disminuir el ruido de cada ensayo, clarificándo probablemente el promedio, disminuyendo quizás el numero de ensayos necesarios para establecerlo, y tal vez posibilitando el análisis de algunos de los ensayos individuales.

Tabla 4.9. Datos correspondientes al caso Quinto.

001	-.0100	051	-.0100	101	-.0075	151	-.0050	201	.0000
002	-.0025	052	-.0225	102	-.0025	152	.0100	202	.0000
003	.0000	053	-.0150	103	.0000	153	.0100	203	.0050
004	.0050	054	-.0075	104	-.0025	154	-.0025	204	-.0050
005	.0025	055	-.0075	105	-.0025	155	-.0050	205	-.0025
006	-.0025	056	-.0150	106	-.0050	156	-.0075	206	.0025
007	.0075	057	-.0150	107	-.0050	157	.0025	207	-.0100
008	.0050	058	-.0200	108	-.0075	158	.0050	208	-.0025
009	-.0050	059	-.0125	109	-.0075	159	.0075	209	-.0025
010	-.0050	060	-.0175	110	-.0150	160	.0075	210	-.0150
011	-.0200	061	-.0125	111	-.0050	161	-.0050	211	-.0100
012	-.0100	062	-.0025	112	-.0125	162	-.0050	212	-.0025
013	-.0125	063	-.0150	113	-.0050	163	-.0075	213	-.0200
014	-.0125	064	-.0125	114	-.0100	164	-.0050	214	-.0050
015	-.0100	065	-.0150	115	-.0100	165	-.0025	215	-.0025
016	-.0100	066	-.0325	116	.0025	166	.0100	216	.0000
017	-.0125	067	-.0375	117	-.0025	167	.0025	217	-.0050
018	-.0025	068	-.0175	118	-.0075	168	.0075	218	-.0025
019	-.0075	069	-.0200	119	-.0050	169	.0125	219	-.0050
020	-.0075	070	-.0275	120	.0075	170	.0100	220	-.0100
021	-.0150	071	-.0250	121	-.0025	171	.0200	221	-.0050
022	-.0100	072	-.0300	122	-.0125	172	.0150	222	-.0100
023	-.0075	073	-.0175	123	-.0125	173	.0075	223	-.0050
024	-.0125	074	-.0250	124	-.0050	174	.0000	224	-.0075
025	-.0075	075	-.0275	125	.0000	175	.0025	225	-.0025
026	.0000	076	-.0150	126	-.0025	176	.0075	226	-.0125
027	.0000	077	-.0200	127	-.0050	177	.0050	227	-.0075
028	.0050	078	-.0075	128	.0050	178	.0050	228	-.0050
029	.0000	079	-.0050	129	.0000	179	.0100	229	-.0100
030	.0300	080	-.0100	130	.0100	180	.0025	230	-.0250
031	.0125	081	-.0200	131	.0050	181	.0050	231	-.0200
032	.0200	082	.0100	132	.0000	182	.0050	232	-.0150
033	.0075	083	.0150	133	.0050	183	-.0025	233	-.0150
034	.0075	084	.0075	134	.0025	184	-.0150	234	-.0225
035	.0100	085	.0100	135	-.0050	185	.0025	235	-.0150

.../...

.../...

036	.0025	086	-.0175	136	-.0100	186	-.0050	236	-.0150
037	.0025	087	.0050	137	-.0075	187	-.0100	237	-.0075
038	.0125	088	-.0075	138	-.0025	188	-.0075	238	-.0100
039	.0050	089	-.0075	139	-.0075	189	.0075	239	-.0175
040	-.0100	090	-.0075	140	-.0150	190	-.0025	240	-.0100
041	-.0125	091	-.0175	141	-.0100	191	.0075	241	-.0100
042	-.0125	092	-.0175	142	-.0150	192	.0075	242	-.0100
043	-.0050	093	-.0125	143	-.0100	193	-.0100	243	-.0150
044	-.0050	094	.0075	144	-.0025	194	-.0125	244	-.0100
045	-.0050	095	-.0050	145	.0075	195	-.0075	245	-.0050
046	-.0025	096	-.0050	146	.0075	196	-.0175	246	-.0075
047	-.0075	097	-.0025	147	-.0050	197	-.0050	247	-.0100
048	-.0025	098	-.0075	148	-.0050	198	-.0150	248	.0000
049	-.0150	099	-.0050	149	-.0075	199	-.0150	249	.0075
050	-.0075	100	-.0125	150	-.0050	200	-.0075	250	-.0050

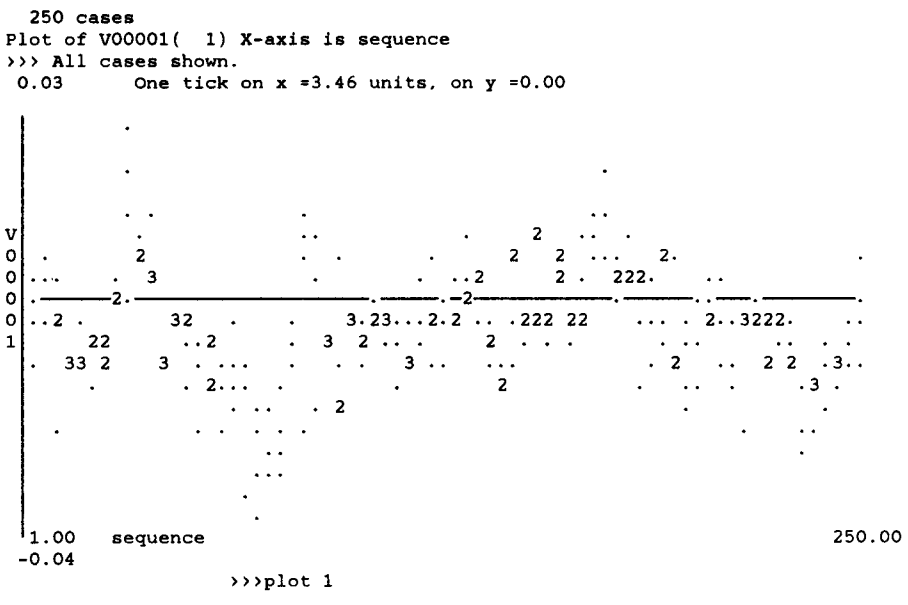


Fig. 4.26. Serie temporal de los datos originales correspondientes al caso quinto.

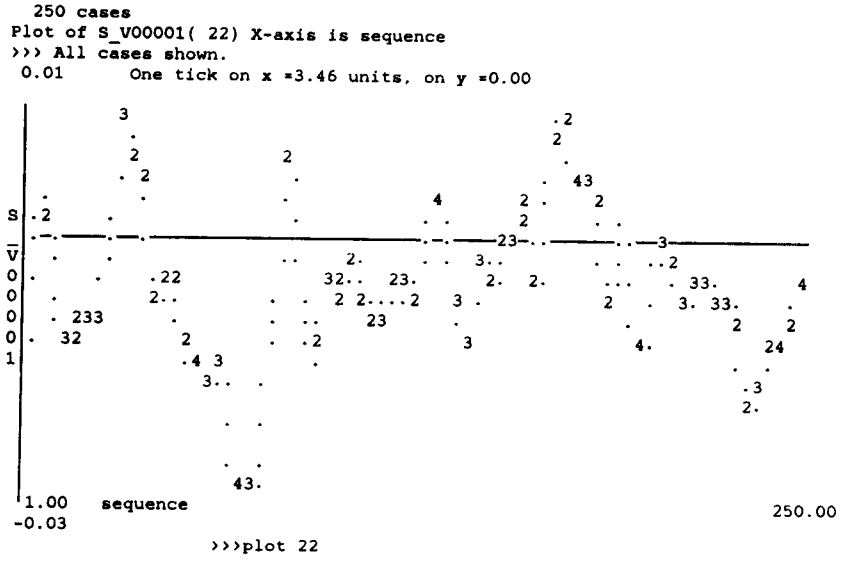


Fig. 4.27. Datos del caso quinto suavizados mediante el alisador "4253h".

5. AJUSTE DE MEDIANAS

5.1. INTRODUCCIÓN

El Análisis Exploratorio de Datos ofrece una serie de técnicas resistentes y robustas para examinar relaciones entre dos o más variables (independientes) cualitativas y una variable respuesta (dependiente) cuantitativa. Esta estructura de datos conocida como tablas de dos factores o diseño factorial se presenta mucho para para estudiar como cada uno de los factores varia regularmente y separadamente del otro y para observar los valores que va tomando la variable respuesta según las diferentes combinaciones de los niveles y de los factores. Por ejemplo, analizar el tiempo de reacción (respuesta) a distintos niveles de inteligencia y de intensidad de estimulo. Hay numerosos ejemplos en Psicología, Química, Medicina, etc.

Estas tablas son analizadas tradicionalmente en Estadística clásica con el analisis de la varianza de dos factores que :

> *"Este modelo constituye un caso especial de combinación lineal y en consecuencia, asume que cada uno de sus componentes debe ser independiente. Como todo modelo descriptivo los términos del mismo deben ser estimados a partir de los datos. El analisis de la varianza asume también*

que los errores siguen una distribución normal y son independientes con media cero y varianza constante".

(Arnau, 1987 pag. 175).

En Estadística Confirmatoria son frecuentes estos y otros supuestos. Este capitulo explora ajustes resistentes para varios modelos sin que sea preciso asumir ninguno de estos supuestos.

Entre ellas está el ajuste simple de medianas (Median Polish) que descompone los efectos de la variable dependiente:

Y = efecto común + efecto fila + efecto columna + residual

Un ajuste Y para tablas de dos factores describe los datos a través de la ecuación

$$Y = X\beta + \epsilon \tag{5.1}$$

Aunque en principio el ajuste de medianas usa un modelo aditivo similar al del Análisis de la Varianza (ANOVA), ajustando éste a partir de las medianas, a través de un proceso iterativo, pone mucho énfasis en mirar y analizar los residuos.

El análisis de los residuos revelan a menudo patrones que no son aparentes en los datos originales. El ajuste de medianas es un método simple resistente que nos ayuda a identificar una posible estructura aditiva de las variables a través de los residuos. Cobra especial interés el poder detectar una posible dependencia aditiva de la variable respuesta con los factores.

Si, por ejemplo, los análisis exploratorios revelan que los efectos factoriales se alejan substancialmente de una estructura aditiva y en consecuencia que no son independientes; entonces seria necesaria una transformación previa a un análisis confirmatorio basado en un modelo lineal, ya que no se cumplirian las condiciones de aplicación. Este capítulo le ayudará a escoger la mejor transformación.

Resumiendo, podemos decir que las técnicas que presentamos en este capítulo ofrecen, para explorar tablas de dos factores las siguientes ventajas:

- No es preciso asumir los rigidos supuestos que son necesarios para aplicar un modelo lineal.

- Puede analizarse con todo tipo de datos cuantitativos. (puntuaciones directas, porcentajes, proporciones, medias ,medianas, etc...).
- Puede realizarse el análisis con datos incompletos (casillas vacias).
- Es resistente.
- Explora la estructura aditiva entre las variables y busca la transformación más adecuada para conseguirla.
- Mediante la descomposición de los datos intenta detectar sus patrones de comportamiento complementando la búsqueda de estos patrones con el estudio de los residuales.
- Es en general más flexible y por tanto tiene gran diversidad y riqueza de análisis y aplicaciones.

Debe advertirse ,sin embargo, que el análisis de Medianas puede dar resultados un poco diferentes si el análisis se empieza por filas o por columnas. También puede dar resultados un poco distintos según el número de iteraciones que se hagan.

Aunque el análisis de medianas puede usarse como técnica alternariva al ANOVA, puede plantearse como estrategia exploratoria, aportando una visión distinta y previa al analisis confirmatorio.

Se podria decir que las informaciones provenientes del análisis exploratorio resistente ayudan a señalar indicios de estructura aditiva o de aproximación a patrones o modelos, siendo ello especialmente interesante en investigaciones en Ciencias Humanas, Sociales y de la Salud.

5.2. MODELO ADITIVO DE TABLAS DE DOS FACTORES

Para presentar la descomposición de los factores en un análisis de medianas, proponemos que cada observación se plantea de la siguiente forma:

$$Y_{ij} \qquad (i = 1 \ldots\ldots I;\ j = 1 \ldots\ldots J) \tag{5.2}$$

Estas tablas tienen tres componentes: el factor fila I ,el factor columna J y la respuesta (ij). Cuando observamos un dato Y_{ij} corresponde al modelo estadístico:

$$Y_{ij} = g_{(i,j)} + e \tag{5.3}$$

Donde, para probar el modelo aditivo más simple tenemos que

$$g_{(i,j)} = \alpha + \beta + \gamma \tag{5.4}$$

En un modelo aditivo, el dato observado en cada casilla[1] es la suma de tres componentes: un valor común constante que sumariza el nivel general de "Y", efectos fila que corresponden a los cambios de "Y" de fila a fila relativos al valor común y efectos columna que corresponden a los cambios de "Y" de columna a columna.

Se cumple que:

ajuste = término comun + efecto fila + efecto columna

Siempre que se ajusta un modelo a los datos se deben examinar las diferencias entre esos datos originales y los valores de la ecuación de ajuste.

residuo = datos - ajuste

de donde, substituyendo términos, se obtiene:

Datos = efect.comun + efect.fila + efect.columna + residual

El modelo aditivo simple, también llamado de efectos principales, tiene una interpretación fácil, ya que

$$Y'_{ij} = \alpha + \beta + \gamma_{ij} \tag{5.5}$$

de donde los residuales se plantean como:

$$\epsilon_{ij} = Y_{ij} - Y'_{ij} \tag{5.6}$$

que corresponden al modelo aditivo ya descrito:

$$Y_{ij} = \alpha + \beta + \gamma + \epsilon \tag{5.7}$$

1. Es importante no confundir este tipo de situaciones con una tabla de χ^2 en la que el valor de las casillas son siempre frecuencias observadas.

Con objeto de presentar las divergencias y similitudes se hará un ajuste de medias y uno de medianas al modelo simple con los mismos datos de partida.

5.2.1. AJUSTE DE UN MODELO ADITIVO A TRAVES DE LAS MEDIAS

Utilizaremos unos datos simulados, que reflejan el promedio mensual del número de accidentes mortales en un año, ocurridos en un cierto cruce de una carretera nacional.

Tabla 5.1. Promedio de accidentes según el momento del día y la edad del conductor.

	Edad Conductor				Medias de filas
	Más de 45	35-45	25-35	Menos de 25	
Hora del accidente					
Mañana	35	32.2	33.1	36.5	34.2
Tarde	33	33.5	33.5	36.0	34.0
Noche	35	34.2	34.5	35.5	34.8
Medias cols.	34.33	33.3	33.7	36.0	
Media de las medias de filas y columnas =				34.33	

El análisis aritmético es sencillo: se empieza con una de las variables "X" hallándose su media "por columnas o filas", se halla la media común y se resta de la primera media de "X" para encontrar los efectos, a continuación se buscan y se separan los efectos de la segunda variable. Por ejemplo, el valor de la casilla de la primera fila y primera columna se descompone del siguiente modo:

$$35 - 34,33 = 0,67$$

Tabla 5.2. Diferencias de cada casilla con respecto a las medias de las columnas.

	Edad Conductor			
	Más de 45	35-45	25-35	Menos de 25
Hora del accidente				
Mañana	0.67	-1.1	-0.6	0.5
Tarde	-1.33	0.2	-0.2	0.0
Noche	0.67	0.9	0.8	-0.5
Medias cols.	34.33	33.3	33.7	36.0
Efecto cols.	0	-1.03	-0.63	1.67

$$
\begin{aligned}
34.33 - 34.33 &= 0 \\
33.30 - 34.33 &= -1.03 \\
33.70 - 34.33 &= -0.63 \\
36.00 - 34.33 &= 1.67
\end{aligned}
$$

$$
\begin{aligned}
\text{Media residuos primera fila} &= (0,67 + (-1,1) + (-0,6) + 0,5)/4 = \\
&= -0.13
\end{aligned}
$$

El resto de medias de residuales por filas se calcularian del mismo modo a partir de los datos de la tabla 5.2. Obtener, con el mismo procedimiento, los efectos promedios por columna se puede observar en la siguiente tabla:

Tabla 5.3. Diferencias de cada casilla con respecto a las medias de las filas.

	Edad Conductor					
	Más de 45	35-45	25-35	Menos de 25	Fila	Efectos Fila
Hora del accidente						
Mañana	0.8	-0.97	-0.47	0.63	34.2	-0.13
Tarde	-1.0	0.53	0.13	0.33	34.0	-0.33
Noche	0.20	0.43	0.33	-0.97	34.8	0.47
Medias cols.	34.33	33.3	33.7	36.0		
Efecto cols.	0	-1.03	-0.63	1.67		

$$34.2 - 34.33 = -0,13$$

Podemos reproducir la tabla original. Por ejemplo, la casilla de la tercera fila y la cuarta columna (con un residual de -0.97) se puede expresar del siguiente modo:

$$0.97 + (-1.03) + (-0.13) + 34.33 = 32.2$$

Efecto columna =	La media de la columna menos la media común. El efecto en "Y" de pertenecer a una columna y no a otra.
Efecto fila =	La media de fila menos la media común. El efecto en "Y" de pertenecer a una fila y no a otra.
Residuos =	Lo que queda después del ajuste; los valores observados en los casilleros menos la media común, efecto de columna y de fila.

La interpretación de efectos de fila y columna es parecida a interpretar niveles de grupo cuando sólo se tiene una variable independiente.

Hasta ahora con un procedimiento simple hemos podido observar los efectos de dos variables categóricas "X" y una variable cuantitativa "Y". ¿Pero, es el análisis de medias resistente?. Nos remitimos al Capítulo primero de este texto, para poner de manifiesto que la media no es un estadístico resistente y, en consecuencia, será necesario plantear esta descomposición con un estadístico que si lo sea. El apartado siguiente aborda esta cuestión.

5.2.2. AJUSTE DE UN MODELO ADITIVO MEDIANTE EL ANÁLISIS DE MEDIANAS

En cualquier distribución en la que aparezcan valores anómalos, éstos afectarán fuertemente a la media. El uso de la media produce un efecto aplanador en los residuos que hace que sean menos aptos para generar nuevas averigüaciones.

Si partimos de la misma idea —descomponer "Y" en los componentes— pero queremos que los componentes sean niveles resistentes, utilizaremos medianas en vez de medias.

Obsérvese una vez más que el EDA utiliza preferentemente estadísticos de orden mientras que la Estadística Descriptiva Clásica utiliza estadísticos basados en la distancia.

La suavización de medianas requiere a menudo varias iteraciones hasta alcanzar la mediana cero. Téngase en cuenta que, a este nivel, se utiliza el error de redondeo (+0,1 -0,1).

La suavización por medias ajusta todos los datos mientras que la aproximación más resistente por medianas ajusta el núcleo de los datos más apretadamente (de residuos más pequeños). Todo ello demuestra mejor el carácter inusual de los datos que no se ajustan bien (residuos grandes). Evidentemente, cuanto mayor sea la tabla, más difícil será ver modelos en los residuos.

Debe observarse, por último, que el resultado será distinto si se empieza por filas que por columnas. Veamos con los datos de la tabla 5.1. como puede efectuarse un ajuste de medianas.

Empezamos la primera iteración por filas, obteniendo los siguientes valores:

a) Se halla la Mediana de cada fila.

b) Se halla la Mediana de las Medianas de las filas.

   De esta forma dispondremos de los valores que se presentan en la siguiente tabla.

Tabla 5.4. Promedio de accidentes según el momento del día y la edad del conductor, con las medianas de las filas y la mediana común.

	Edad Conductor				
Hora del accidente	Más de 45	35-45	25-35	Menos de 25	Mediana de las filas
Mañana	35	32.2	33.1	36.5	34.05
Tarde	33	33.5	33.5	36.0	33.50
Noche	35	34.2	34.5	35.5	34.75
				Mediana común	34.05

A continuación establecemos el siguiente paso:

c) A cada valor se le resta la mediana correspondiente a su fila.

Así se obtiene la primera tabla de residuos, que corresponden a los valores habiéndoles restado el efecto de la fila. El residuo, por ejemplo, de la primera fila y primera columna sería pues 0.95 (35 - 34.05). La tabla 5.5. muestra estos valores.

Tabla 5.5. Residuales de los efectos fila con respecto a sus medianas.

	Edad Conductor				
Hora del accidente	Más de 45	35-45	25-35	Menos de 25	Mediana de las filas
Mañana	0.95	-1.85	-0.95	2.45	34.05
Tarde	-0.50	0.0	0.0	2.45	33.50
Noche	0.25	-0.55	-0.25	-0.75	34.75
				Mediana común	34.05

d) Se hallan las medianas de los residuos de cada columna, como se aprecia en la siguiente tabla.

Tabla 5.6. Residuales de los efectos fila con respecto a sus medianas y medianas de cada columna.

	Edad Conductor					
	Más de 45	35-45	25-35	Menos de 25		Mediana de las filas
Hora del accidente						
Mañana	0.95	-1.85	-0.95	2.45		34.05
Tarde	-0.50	0.0	0.0	2.45		33.50
Noche	0.25	-0.55	-0.25	-0.75		34.75
Md. Columna	0.25	-0.55	-0.25	2.45	Md común	34.05

e) A cada residuo se le resta la mediana de los residuos correspondientes a su columna. Así se obtiene una segunda tabla de residuos después de haberles restado el efecto de la fila y columna, tal como aparece en la tabla 5.7.

Tabla 5.7. Residuales de los efectos fila y columna.

	Edad Conductor				
	Más de 45	35-45	25-35	Menos de 25	
Hora del accidente					
Mañana	0.70	-1.30	-0.70	0.00	
Tarde	-0.75	0.55	0.25	-0.05	
Noche	0.00	0.00	0.00	-1.70	
Md. Columna	0.25	-0.55	-0.25	2.45	Md común 34.05

f) A cada mediana de cada columna se le resta la mediana común tal como se presenta en la tabla 5.8.

Tabla 5.8. Resultados del paso (f) en el proceso de ajuste.

	Edad Conductor				
	Más de 45	35-45	25-35	Menos de 25	Medianas filas
Hora del accidente					
Mañana	0.70	-1.30	-0.70	0.00	-0.350
Tarde	-0.75	0.55	0.25	-0.05	-0.400
Noche	0.00	0.00	0.00	-1.70	0.700
Md. Columna	0.25	-0.55	-0.25	2.45	Md común 34.05

Con ello, se ha terminado una primera iteración. Se pueden hacer más iteraciones por el mismo procedimiento, repitiendo los pasos C y D hasta que el residuo sea despreciable o sea lo más cercano a cero posible.

Seguimos pudiendo reproducir los valores originales de la Tabla 5.1.. Por ejemplo, el valor de la tercera fila cuarta columna es igual a:

$$35.5 = 34.05 + 2.45 + 0.7 + (-1.65) + (-1.70)$$

De este modo, adquieren sentido los valores que componen el modelo aditivo general del que hemos partido (expresión 5.5), cuyos componentes se definen ahora del siguiente modo:

α = valor común
β = efecto principal debido a la fila
γ = efecto principal debido a al columna
ϵ = lo que queda en la tabla son los residuos después del ajuste.

Seguir este proceso puede ser complejo en los cálculos; por ello mostraremos (tabla 5.9) los ajustes conseguidos con nuestros datos tras dos iteraciones efectuadas con ordenador (Statgraphics).

Tabla 5.9. Ajuste de Medianas después de dos iteraciones.

	Edad Conductor				
	Más de 45	35-45	25-35	Menos de 25	Medianas filas
Hora del accidente					
Mañana	1.050	-0.950	-0.350	0.450	-0.350
Tarde	-0.900	0.400	0.100	0.000	-0.400
Noche	0.000	0.000	0.000	-1.600	0.700
Md. Columna	0.25	-0.55	-0.25	2.45	Md común 34.05

De forma esquemática, la figura 5.1. muestra los pasos seguidos hasta llegar a los resultados de la tabla anterior.

Se observa que los residuos más grandes están en -1.6 (conductores menores de 25 años y noche) y 1.05 (conductores de más de 45 años y mañana). Lo cual indica que hay una posible interación entre estos valores de las variables.

En los casos en que el residuo es 0 indica que toda la variabilidad es explicada por los efectos principales de la fila y de la columna.

En la figuras 5.2. y 5.3. se presentan los gráficos propios de E.D.A. (Tronco y Hojas y Caja) de los residuales presentados en la anterior tabla, para efectuar una evaluación de los mismos.

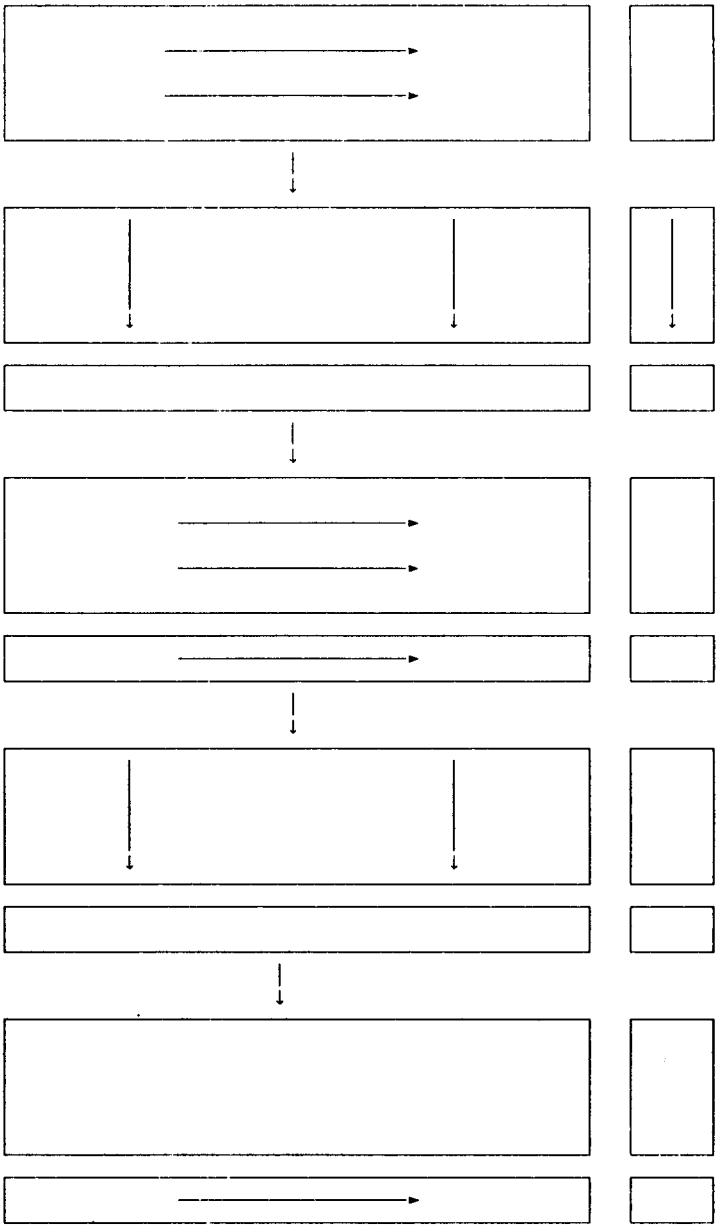


Fig.5.1. Esquema de iteraciones de la tabla 5.9.

Ajuste de Medias		Ajuste de Medianas	
-1	0	-1	6
-0	949	-0	399
0	82541363	0	0000144
1		1	0

Fig. 5.2. Gráficos de Tronco y Hoja de los residuales.

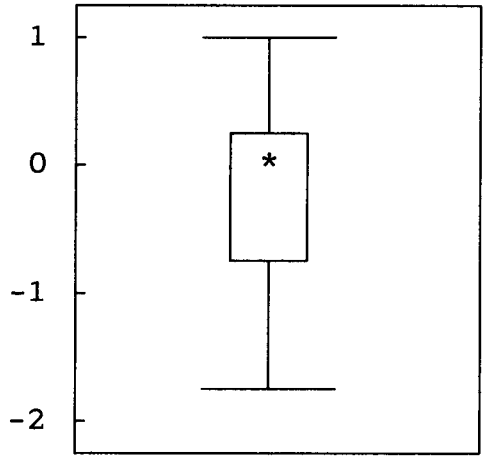


Fig. 5.3. Gráfico de Caja de los residuales del Ajuste de Medianas.

Resumiendo hemos intentado ajustar los datos a un modelo aditivo simple que corresponde a la ecuación (5.5) mediante un proceso iterativo de ajuste de medianas y hemos obtenido unos residuos. Tanto el ajuste de medias como el de medianas tienen ventajas e inconvenientes. Dependerá del tipo de datos y de la investigación el explorar una u otra, no obstante cabe resaltar que el análisis de medianas es siempre más resistente.

Con respecto al número de iteraciones necesarias, un criterio efectivo es el propuesto por Emerson y Hoaglin (1982; pp. 188)) que cifran en dos el número de iteraciones frecuentemente necesarias en la práctica.

Con objeto de mostrar con más detenimiento el proceso de descomposición del ajuste de medianas, proponemos un segundo ejemplo con los datos de la tabla 5.10.:

Tabla 5.10. Datos del número de respuestas acertadas en una prueba de memoria según la longitud de la lista y el intervalo de retención.

	Intervalo de Retención					
	0 seg.	15 seg.	30 seg.	60 seg.	5 min.	10 min.
Longitud serie						
8 Palabras	638	679	720	762	806	856
16 Palabras	696	740	781	829	875	926
32 Palabras	755	800	841	893	943	997
64 Palabras	785	828	876	925	976	999

Si hacemos el ajuste de medianas con dos iteraciones se obtiene la siguiente tabla de residuos

Tabla 5.11. Residuales después de dos iteraciones.

	Intervalo de Retención						
	0 seg.	15 seg.	30 seg.	60 seg.	5 min.	10 min.	Md. Común
8 pal.	7.875	4.375	2.125	-3.625	-7.625	-2.750	-95.125
16 pal.	1.375	0.875	-1.375	-1.125	-3.125	2.750	-30.625
32 pal.	-1.375	-0.875	-3.125	1.125	3.125	12.000	31.125
64 pal.	-1.875	-2.375	1.375	2.625	5.625	16.500	61.625
	-111	-66.5	-23.25	24.5	72.5	117.62	836.25

Igualmente que en el ejemplo anterior, la descomposición de cualquier casilla es factible. Por ejemplo, la primera fila y primera columna presenta la siguiente descomposición:

$$638 = 836.25 + (-111) + (-95.125) + 7.875$$

Ajuste de Medianas

-0.	7
-0*	0111122333
0*	011122234
0.	57

Extremos (12), (17)

Fig. 5.4. Gráfico de Tronco y Hoja de los residuales.

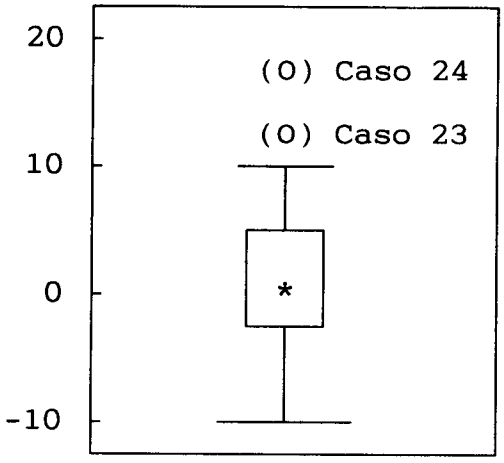


Fig. 5.5. Gráfico de Caja de los residuales de la tabla 5.11

Para efectuar un análisis de los residuales de una forma gráfica y rápida, Tukey sugiere simbolizar el valor de los residuales de cada una de las casillas mediante un código gráfico y así evaluar de forma global los residuales. Un criterio factible seria el siguiente:

Tabla 5.12. Códigos gráficos para la representación de los residuales.

Residual muy positivo (valores $> (Q_3 + 3IQR)$	P
Residual bastante positivo $(Q_3 + 1.5IQR) \div (Q_3 + 3IQR)$	#
Residual algo positivo $(Q_3) \div (Q_3 + 1.5IQR)$	+
Residual cercano a cero $Q_1 \div Q_3$	·
Residual algo negativo $(Q_1) \div (Q_1 - 1.5IQR)$	-
Residual bastante negativo $(Q_1 - 1.5IQR) \div (Q_1 - 3IQR)$	=
Residual muy negativo (valores $< (Q_1 - 3IQR)$	N

La tabla de residuales 5.11. representada mediante estos simbolos gráficos adopta la forma siguiente:

Tabla 5.13. Representación gráfica de la tabla de residuales 5.11.

+	+	·	-	-	-
·	·	·	·	-	·
·	·	-	·	+	#
·	-	·	·	+	#

Para finalizar con el proceso de descomposición y de ajuste a un modelo aditivo con el análisis de medianas, proponemos un tercer y último ejemplo, en el que se plantea una tabla (4×6) con los siguientes valores iniciales:

Tabla 5.14. Tabla de dos factores con datos simulados.

3.1911	3.2034	3.2157	3.2289	3.2421	3.2571
3.2091	3.2217	3.2346	3.2484	3.2628	3.2781
3.3268	3.2403	3.2538	3.2682	3.2832	3.2994
3.2358	3.2490	3.2631	3.2778	3.2931	3.3096

La aplicación del procedimiento ya analizado con anterioridad lleva a la siguiente tabla de residuales después de dos iteraciones.

Tabla 5.15. Tabla de residuales después de dos iteraciones.

0.0021	0.0013	0.0004	-.0004	-.0019	-.0027	-0.02895
0.0009	0.0004	0.0001	-.0001	-.0004	-.0009	-0.00975
-.0009	-.0004	-.0001	0.0001	0.0004	0.0009	-0.00975
-.0013	-.0012	-.0003	0.0003	0.0009	0.0016	0.01920
-.0333	-.0202	-.0070	0.0070	0.0217	0.0375	

Al igual que en los anteriores ejemplos, se puede especificar la descomposición de cualquier casilla de la tabla 5.14. De este modo, el valor de la casilla de la primera fila y primera columna se puede plantear como:

$$3.1911 = 3.25125 + (-0.02895) + (-0.0333) + 0.0021$$

5.3. AJUSTE A UN MODELO EXPANDIDO (CON INTERACCIÓN)

En este apartado estudiaremos la información sobre la aditividad que nos muestra la tabla de residuos procedente del ajuste de medianas, tal como hemos analizado en el apartado anterior (5.2.2.).

Presentaremos y construiremos un gráfico de gran utilidad llamado gráfico de diagnóstico, que nos detectará si hay indicios de no aditividad analizando los residuos encontrados y los valores comparativos. Siguiendo los contenidos del tercer capitulo de este texto, buscaremos la pendiente de la linea resistente entre los residuales y los valores comparativos. Según sea el valor de la pendiente de la linea resistente intentaremos descomponer

$$\epsilon = k\alpha\beta + r \tag{5.8}$$

es decir, intentaremos el ajuste a un modelo que describa la interacción entre los efectos fila y los efectos columna. Finalmente, buscaremos la mejor transformación para conseguir la aditividad.

5.3.1. INFORMACIÓN DE LA TABLA DE RESIDUALES PROCEDENTES DEL AJUSTE DE MEDIANAS

Los residuos procedentes del ajuste de Medianas informan de la estructura de los datos; siendo esta información esencial para explorar la aditividad del modelo. Esta estructura se pone de manifiesto si en la tabla de residuos, después de un ajuste de Medianas, se muestra un modelo de los llamados de silla (saddle), es decir los residuos opuestos diagonales tienen el mismo signo, como puede apreciarse en la tabla 5.15.

Cuando el efecto fila y el efecto columna tienen el mismo signo y el residuo correspondiente es pequeño o negativo y cuando el efecto fila y el efecto columna tienen signos opuestos y el residuo correspondiente es pequeño o positivo; hay evidencia de que no hay aditividad.

No obstante, existen tablas de residuales en los que será más difícil explorar la posible no aditividad del modelo. Las figuras siguientes (5.6) muestran las formas clásicas que adoptan las tablas de residuales para poner de manifiesto la existencia de no aditividad.

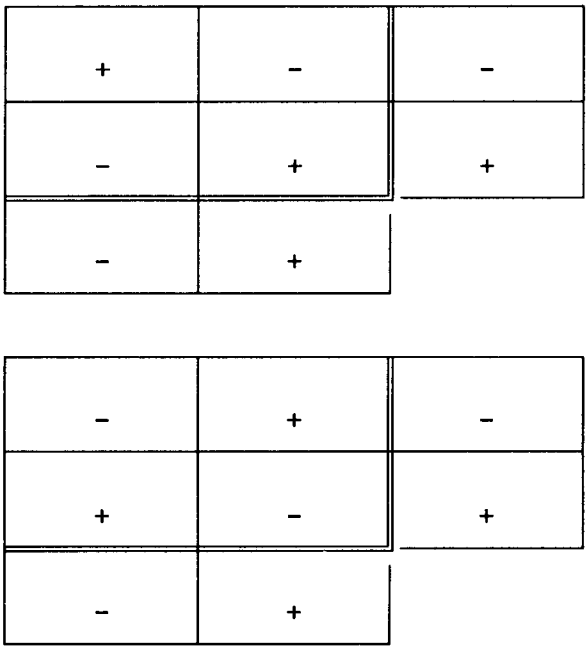


Fig. 5.6. Modelos de silla que pueden adoptar las tablas de residuales de los ajustes de Medianas.

En la tabla de residuales (5.15.) se observa un modelo clásico de silla. En situaciones tan claras no se requiere más evidencia para la aceptación de la no aditividad. Por contra, los datos analizados en el primer y segundo ejemplo (tablas 5.9. y 5.11) no dan información tan clara acerca de la aditividad.

De este modo, en la ecuación

$$Y = \alpha + \beta + \gamma + \epsilon \tag{5.9}$$

se asume que los datos siguen un modelo aditivo en su mayoria (quizás algun valor inusual). Si los datos se alejan sustancialmente y sistemáticamente de la estructura aditiva será necesario hacer una transformación antes de hacer los análisis que presuponen relación lineal entre las variables.

Si los residuales de un ajuste de medianas aditivo simple muestran separación sistemática de la aditividad vemos que el ajuste a la ecuación anterior no es el más adecuado.

El intento de ajuste de un modelo más complejo nos puede ayudar a comprender y suavizar la estructura no aditiva. Aunque podemos detectar la estructura mirando la tabla de residuales de un ajuste de medianas un gráfico de diagnóstico nos ayudará a escoger la mejor transformación para remover la no aditividad.

5.3.2. GRÁFICO DE DIAGNÓSTICO

Este gráfico sirve para detectar la no aditividad y para buscar la mejor transformación. Para hacer el gráfico de diagnostico, al que nos hemos referido, deberemos empezar calculando los valores comparativos (vc). Estos valores adoptan la siguiente expresión:

$$(vc) \text{ valor comparativo} = (a \cdot b)/m \tag{5.10}$$

siendo en esta fórmula "a" el efecto columna
"b" el efecto fila
"m" el efecto común

Se construye el Diagrama de dignóstico uniendo los puntos vc (valor comparativo) y ϵ (residual). El diagrama de Diagnóstico nos revela la tendencia o la estructura de los datos. Veamos su utilidad con la tabla de residuales 5.11.

Tabla 5.16. Tabla de valores comparativos a partir de los residuales de la tabla 5.11.

12.63	7.56	2.64	-2.78	-8.25	-13.38	-95.12
4.06	2.43	0.85	-0.89	-2.65	-4.30	-30.62
-4.13	-2.47	-0.86	0.91	2.69	4.37	31.12
-8.18	-4.90	-1.71	1.80	5.34	8.66	61.62
-111.0	-66.5	-23.5	24.5	72.5	117.62	836.25

El cálculo de (vc) en la primera fila y primera columna se efectuaría del siguiente modo:

$$vc = (-111.0 \cdot -95.12)/836.25 = 12.63$$

El diagrama de diagnóstico se construye, pues, como un gráfico bivariable con los residuales hallados y los valores comparativos (vc) establecidos como hemos visto. Si seguimos con nuestros datos, los resultados con que construir el gráfico son los siguientes:

Tabla 5.17. Residuales y valores comparativos de los datos de la tabla 5.10. (Residuales en la tabla 5.11. y valores comparativos en la tabla 5.16.).

RESIDUALES		VALORES COMPARATIVOS	
7.87		12.63	
4.37		7.56	
2.12		2.64	
-3.62		-2.78	
-7.62		-8.25	
-2.75		-13.68	
1.37		4.06	
0.87		2.43	
-1.37	*	0.85	*
-1.12		-0.89	
-3.12		-2.65	
2.75	*	-4.30	*
-1.37		-4.13	
-0.87		-2.47	
-3.12		-0.86	
1.12		0.91	
3.12		2.69	
12.00		4.37	
-1.87		-8.18	
-2.37		-4.90	
1.37	*	-1.71	*
2.62		1.80	
5.62		5.34	
-16.50	*	8.66	*

El gráfico de diagnóstico adopta la forma que se presenta en la figura 5.7.

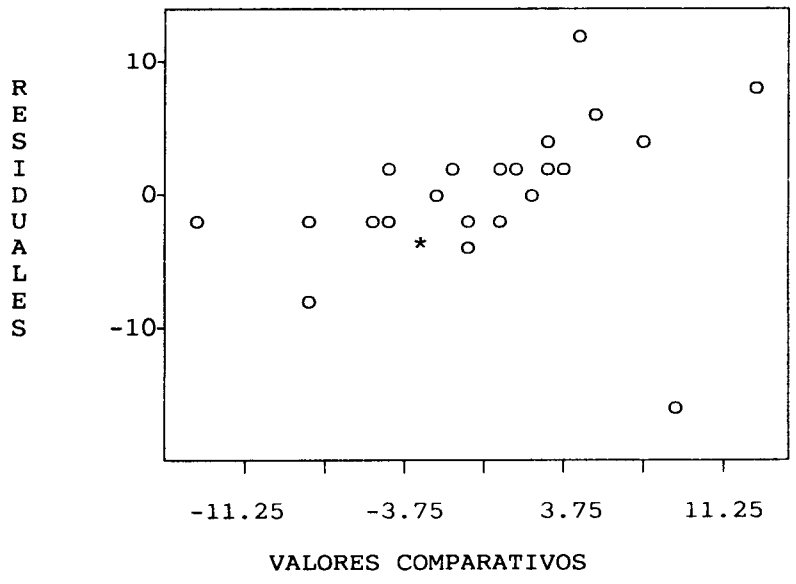


Fig.5.7. Gráfico de Diagnóstico.
(o un punto. * dos puntos).

Si entre los valores comparativos y los residuales no hay relación podremos concluir que los residuales son independientes entre si y en consecuencia que el modelo es aditivo, o lo que es lo mismo, que el residual es solo error producido por el azar y no hay interación entre las variables. En otras palabras, que la esperanza matematica de los residuales es 0 y que el modelo es aditivo.

Si por el contrario el diagrama de diagnóstico revela relación entre las residuales y los valores comparativos indicará que el modelo no es aditivo, que los errores no son independientes y que podemos intentar ajustar los datos a otro modelo para poder descomponer los residuales en que parte es interación y que parte es debido solo al azar.

Vale la pena resaltar que, en nuestra opinión, la importancia del test de aditividad que presentamos está más en el signo de los residuales y de los valores comparativos que propiamente en el valor de estos. En efecto, si

los residuales fueran realmente error producido por el azar los signos de los residuales y de los valores comparativos serian aleatorios. Observese, que en nuestro ejemplo (tabla 5.17) todos los pares de residuales y valores comparativos tienen el mismo signo menos cuatro (señalados con *).

Resumiendo, el gráfico de diagnostico es útil porque detecta la estructura interna de las tablas de dos factores es decir que patrones o tendencias o interacciones se hallan escondidos en los datos. Y esto será la mejor guia para

- Transformar.
- Para ajustar a un modelo extendido (con interación).

Esta nube de puntos que constituye el gráfico de diagnóstico puede ser objeto de tratamientos específicos para su alisamiento (aplicando un suavizador). El lector obtendrá información acerca de esta cuestión en el capítulo anterior de este texto. Baste señalar que un alisador frecuente a este respecto es el $3RSS$.

5.3.3. NUEVO AJUSTE UTILIZANDO LA LÍNEA RESISTENTE

La figura 5.7. pone en evidencia una relación lineal entre los residuales y los valores comparativos, lo cual nos lleva a determinar que no es factible plantear un modelo aditivo. No obstante, a veces esta tendencia no aparece tan clara, como puede observarse en la tabla 5.11. de residuales.

Si la tendencia lineal se observa en el gráfico de diagnostico, estamos en condiciones de pensar que muy probablemente los errores no son independientes entre sí y, en consecuencia, no son sólo debidos al azar.

De tal modo que, si queremos obtener un mejoramiento substancial sobre un ajuste simple probaremos un ajuste más amplio para estos datos. Este tipo de estrategia debe originar residuales más pequeños que los producidos en primera instáncia:

"...se espera que los residuales de un ajuste más amplio sean más pequeños en magnitud y tengan nenos patrón que los residuales procedentes a un ajuste simple".

(Emerson y Wong, 1990; pág. 75)

Si los datos no se ajustan al primer modelo podemos intentar ajustarlos a otro modelo:

$$Y = \alpha + \beta + \gamma + \beta\gamma + \epsilon \tag{5.11}$$

En realidad Tukey (1949) presenta este modelo en una forma más amplia:

$$Y = \alpha + \beta + \gamma + k(\beta\gamma/\alpha) + \epsilon \tag{5.12}$$

donde k es una constante que hay que determinar.

El valor de "k" más adecuado viene facilitado por la pendiente de la linea resistente, como demuestra el desarrollo presentado en el capítulo tres de este texto.

La linea resistente es una buena estrategia para explorar la relación entre los residuales y los valores comparativos sin que influyan en el resultado unos pocos valores alejados. En caso de que la relación lineal no este clara, podemos aplicar la transformación de potencia propuesta con anterioridad. El cálculo de la pendiente de la nube de puntos original, que constituye el gráfico de diagnóstico, de los resultados de los residuales de la tabla 5.11. ofrece un valor de "k" = 1.074. Este valor de "k" cumple el papel del valor de "m" planteado en el capítulo de la Linea Resistente.

Intentaremos, pues, ajustar el modelo anterior añadiendo esta constante ($k \cdot V.C.$), de tal manera que el modelo quedaria planteado como sigue:

$$r = Y - (\alpha + \beta + \gamma + k\beta\gamma) \tag{5.13}$$

Si los valores de la tabla 5.16. son multiplicados por el valor de "k" obtendremos una nueva tabla con la que evaluar el ajuste al modelo propuesto. La siguiente tabla ofrece estos valores:

Tabla 5.18. Términos iteractivos ($k\beta\gamma$), obtenidos multiplicando los valores comparativos de la tabla 5.16. por el valor de k = 1.074 hallado a través de la linea resistente.

13.56	8.11	2.83	-2.98	-8.86	-14.37
4.36	2.60	0.91	-0.95	-2.84	-4.61
-4.43	-2.65	-0.92	0.97	2.88	4.69
-8.78	-5.26	-1.83	1.93	5.73	9.30

Como ejemplo de esta operación, el valor de la casilla de la primera fila y primera columna se puede plantear como sigue:

$$13.56 = 12.63 \cdot 1.074$$

Si con la tabla anterior efectuamos un nuevo ajuste, sustrayendo de los residuales de la tabla 5.11. los términos interactivos, obtendremos unos residuales definitivos con respecto al modelo de interacción, evaluándolos con respecto a los residuales obtenidos en primera instáncia. La tabla siguiente muestra los residuales finales.

Tabla 5.19. Residuales del último ajuste de medianas efectuado con los datos de la tabla 5.18.

-5.69	-3.74	-0.71	-0.64	1.24	-17.12
-2.99	-1.73	-2.28	1.12	-0.28	7.36
3.06	1.77	-2.2	0.15	0.24	7.31
6.91	2.89	3.2	0.69	-0.11	-29.8

Análogamente a casos anteriores, el valor de la primera fila y primera columna se puede establecer:

$$-5.69 = 7.87 - 13.56$$

De esta forma, la descomposición de los datos originales debe variar, puesto que hemos incorporado un efecto de interacción en el modelo ajustado. En consecuencia, el valor original se puede expresar de tal manera que:

$$638 = 836.25 + (-111) + (-95.125) + 13.56 + (-5.69)$$

La comparación de las dos tablas de residuales (5.11. y 5.19.) confirman nuestras sospechas; los residuales procedentes de un ajuste más amplio (tabla 5.19) carecen de patrón y son más pequeños en magnitud.

Por último, debe destacarse que para una mayor comprensión del procedimiento de transformación de los residuales a partir del gráfico de diagnóstico, es necesario recurrir al capítulo tres, en el que se plantea con mucha más exhaustividad la estrategia de transformar los datos para el establecimiento de linealidad. Un desarrollo completo de esta temática puede obtenerse en Emerson y Stoto (1982). El tema de la transformación de las variables ya se ha abordado en este texto, de forma que si el lector desea obtener un aspecto más amplio con respecto a este tema, debe, asimismo, consultar en su totalidad el segundo capítulo de este volúmen.

Dentro del mismo ámbito, podemos señalar que se ha propuesto un indicador de variación explicada en el ajuste de este tipo de modelos. En concreto, Emerson y Wong (1985), presentan una expresión para el cálculo del porcentaje de variación absoluta:

$$P = [1 - (\Sigma\Sigma|e_{ij}|/\Sigma\Sigma\,|Y_{ij} - Md_{ij}(Y_{ij})|)] \cdot 100 \qquad (5.14)$$

Con los datos que estamos manejando en este apartado, antes de proponer el modelo expandido, podemos obtener los siguientes resultados:

$$Md_{ij}(Y_{ij}) = 828.5$$
$$\Sigma\Sigma|Y_{ij} - 828,5| = 1946.00$$
$$\Sigma\Sigma|e_{ij}| = 91.00$$

$$P = [1 - (91/1946)] \cdot 100 = 95.30\% \qquad (5.15)$$

En el caso del modelo expandido, el cual hemos analizado en este apartado, obtenemos el siguiente valor de P:

$$P = [1 - (75.130/1946)] \cdot 100 = 96,10\% \qquad (5.16)$$

Es fácil comprobar un ligero aumento en el porcentaje de variación absoluta entre los dos modelos.

5.4. APLICACIÓN DEL AJUSTE DE MEDIANAS A LAS MEDIDAS REPETIDAS

El diseño de medidas repetidas o diseño experimental intrasujeto es cada vez más usado en Ciencias Humanas. En muchas disciplinas, al investigar el efecto de un variable se hace necesario este tipo de estrategias, como por ejemplo en los estudios sobre el "cambio".

En el diseño de medidas repetidas, cada sujeto recibe todas las condiciones experimentales y sólo genera un dato para cada nivel o tratamiento. Los sujetos actuan como control propio, extrayendo de la variabilidad del error una de sus principales fuentes, siendo en realidad, el control de las diferencias individuales (Arnau, 1987). Una de las ventajas de este tipo de diseños es que se tiene una mayor precisión de la estimación de la acción de la variable tratamiento al disminuir el error experimental. Otra ventaja que los caracteriza es la reducción del coste del experimento, dado que se requieren menos sujetos.

Este diseño suele analizarse como el diseño factorial de dos factores independientes, en el que el primer factor es la variable de tratamiento y considera como segundo factor "la variable sujetos" (a pesar de que no sea propiamente un factor).

No obstante, estos diseños tienen también unos inconvenientes que suelen llamarse de **orden o secuenciales** y otros de **efectos residuales.**

Dado que en este tipo de diseños las observaciones son sucesivas sobre un mismo sujeto (Nambooridi, 1972), es frecuente que ocurra que los componentes de error no sean independientes entre si. En términos generales, si se pretende ajustar un diseño de este tipo a un modelo matemático aditivo, este adopta la expresión:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \qquad (5.17)$$

siendo ϵ_{ij} la desviación de Y_{ij} de μ que no es explicada ni por el efecto del tratamiento α_i ni por las diferencias individuales β_j.

Tukey (1949) elaboró una prueba que permite comprobar la aditividad con los datos de diseños de medidas repetidas que requiere la descomposición de la suma de cuadrados de la interacción en dos componentes:

$$SC_{\text{no-adit}} = \frac{[\Sigma_i(Y_{.j} - Y_{..})\Sigma_j Y_{ij}(Y_{i.} - Y_{..})]^2}{\Sigma_j(Y_{.j} - Y_{..})^2\Sigma_i(Y_{i.} - Y_{..})^2} \qquad (5.18)$$

$$F = CM_{\text{no-adit}} \,/\, \text{CMresidual} \qquad (5.19)$$

Un desarrollo exhaustivo de la prueba de Tukey puede encontrarse en Arnau (1987) y recomendamos al lector que se familiarice con la misma previamente a la lectura de este apartado. Sin embargo, como señalan Mayers (1972) y Arnau (1987), esta prueba no es sensible en aquellas situaciones en que las medias de las filas tienden a decrecer y la columna de los productos cruzados tiende a disminuir para posteriormente aumentar. Se pone de manifiesto, mediante la representación gráfica de la relación entre los productos cruzados y las medias de las filas, que se da interacción entre las fuentes de variación tratamiento y sujetos.

En este caso, pues, después de realizar la prueba de Tukey y sin necesidad de realizar el gráfico anteriormente descrito, consideraríamos que el modelo es aditivo. En realidad, estaríamos confundiendo el error "e" de un modelo aditivo cuando en realidad este se descompone en "$\alpha\beta$" (interacción tratamiento por sujetos) y "e" propiamente dicho. A la vista de todo ello, y con objeto de mejorar el diseño, es adecuado seguir estudiando los residuales en los diseños de medidas repetidas.

El ajuste de medianas es especialmente indicado en este tipo de diseños. Estudiar los residuales después de un ajuste de medianas es similar a la prueba de no aditividad de Tukey, ayudando a encontrar e identificar la forma con que cada sujeto reacciona a los diferentes tratamientos. Con el estudio exhaustivo de los residuales del ajuste de medianas, obtendremos evidencia de la existencia o no de interacción.

Veamos una aplicación de este aspecto con los datos que se muestran en la tabla 5.20.

Tabla 5.20. Diseño de medidas repetidas con seis sujetos y tres tratamientos.

	Tratamientos			
Sujetos	A	B	C	Medias
1	13.2	23.0	13.0	16.4
2	14.0	24.0	13.7	17.2
3	13.0	13.0	13.1	13.03
4	13.1	12.9	13.2	13.06
5	6.5	16.5	6.6	9.86
6	6.4	16	6.7	9.7

Con los datos de la tabla anterior se obtiene la tabla resúmen del ANOVA siguiente:

Tabla 5.21. Tabla resúmen del Análisis de la Varianza.

Fuentes de Variación	*S.C.*	g.l.	*C.M.*	*F*
Tratamiento	170.3037	2	85.15	9.52
Sujetos	149,7383	5	29,95	3.35
Residual (T×S)	89.4230	10	8.94	
Total	409.4751	17		

Si aplicamos la prueba de no aditividad de Tukey, a partir de los datos presentados, obtendremos:

$$SC_{\text{no-adit}} = 0.44 \qquad F = 0.04$$

Podemos concluir con la no significación estadística de la no aditividad y, en consecuencia, los datos presentados en este diseño se ajustan correctamente a un modelo aditivo. Si efectuamos el gráfico de la relación entre las medias de las filas y los valores de los productos cruzados, encontramos que se da interacción entre los tratamientos y los sujetos. La figura 5.8. presenta el gráfico al que aludimos.

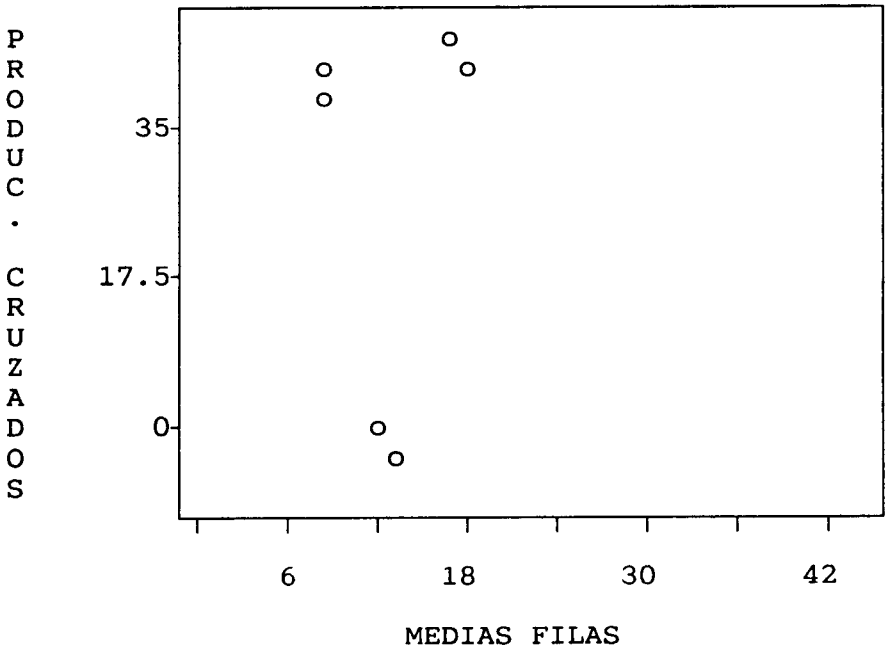


Fig. 5.8. Gráfico entre las medias de las filas y los productos cruzados.

A la vista del gráfico (5.8.) y del resultado en la prueba de Tukey, puede desprenderse que si bien la figura indica la presencia de interacción, el análisis empírico no lo ha detectado. Por su parte, si sometemos la tabla 5.20 a un ajuste de medianas, será factible poner de manifiesto este efecto. En la tabla 5.22. mostramos los residuales del ajuste de medianas trás dos iteraciones:

Tabla 5.22. Residuales del ajuste de medianas con dos iteraciones.

Sujetos	A	B	C	
1	0.000000	0.12500	-0.250000	0.15
2	0.000000	0.32500	-0.350000	0.95
3	0.000000	-9.65700	0.050000	-0.05
4	0.000000	-9.87500	0.050000	0.05
5	-0.100000	0.22500	-0.050000	-6.45
6	-0.050000	-0.12500	0.200000	-6.60
	0.0	9.675	0.05	13.05

Esta tabla refleja claramente que los sujetos 3 y 4 se comportan de forma distinta al resto en el tratamiento B. Como siempre, es factible plasmar la descomposición de cualquier casilla:

$$13.2 = 13.05 + 0.15 + 0.0 + 0.0$$

Será necesario, para poner de manifiesta el efecto de la no aditividad, establecer el gráfico de diagnóstico a partir de la tabla 5.22., calculando para ello los ya comentados valores compratívos (vc). Estos valores quedan reflejados en la siguiente tabla:

Tabla 5.23. Valores Comparativos a partir de la tabla 5.22.

Sujetos	A	B	C	
1	0.000000	0.11100	0.000570	0.15
2	0.000000	0.70000	0.003600	0.95
3	0.000000	-0.03000	-0.002500	-0.05
4	0.000000	0.03000	0.002500	0.05
5	0.000000	-4.78000	0.002400	-6.45
6	0.000000	-4.89000	-0.025000	-6.60
	0.0	9.675	0.05	13.05

Con los datos de las tablas 5.22. y 5.23. estamos en condiciones de establecer el correspondiente gráfico de diagnóstico, que puede analizarse en la siguiente figura:

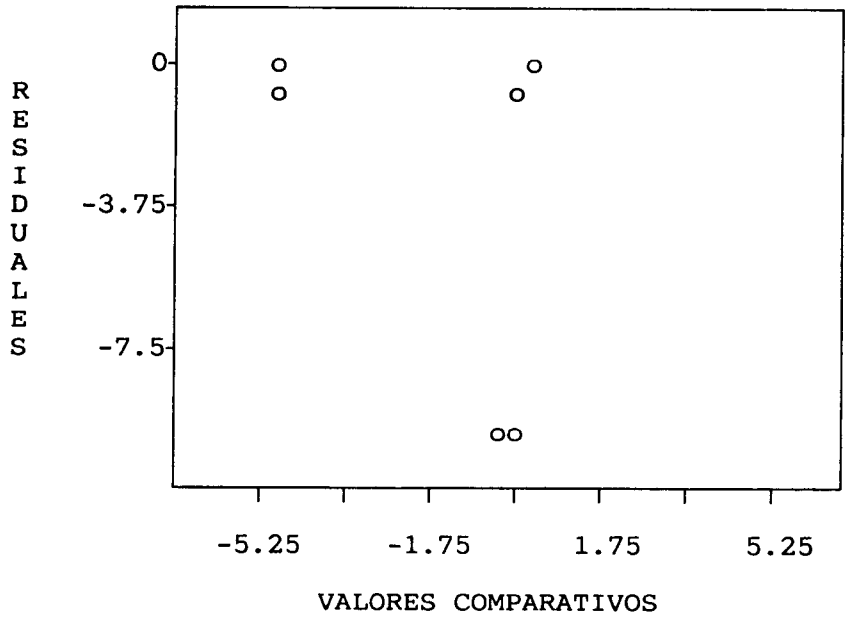


Fig. 5.9. Gráfico de diagnóstico entre los residuales y valores comparativos.

Esta representación evidencia la interacción (Tratamiento × Sujetos) que, de forma más precisa, se puede ver reflejada en los sujetos 3 y 4, que presentan un comportamiento peculiar con respecto, como decíamos, al tratamiento B. Tal comportamiento ya se ha puesto de manifiesta en la tabla de residuales (5.22.) con unos valores altamente distintivos del resto.

Añadiremos, por último, que con este procedimiento obtenemos, con suma rápidez y facilidad de cálculo, un análisis exhaustivo de los residuales, además de obtener información de la estructura aditiva del modelo que se propone evaluar y del comportamiento de los sujetos bajo los efectos de los distintos tratamientos.

6. INTRODUCCIÓN A LA ESTIMACIÓN ROBUSTA

La distribución de Gauss y la ley de los grandes números son algo que los matemáticos suponen que son producto de la observación y los experimentos, y que los naturalistas y los sociólogos creen que son teoremas matemáticos.

E. Poincaré. Demieres pensées.

6.1. INTRODUCCIÓN A LA PROBLEMÁTICA DE LA ESTIMACIÓN

Durante la decada de los 70 dos hechos revolucionaron el concepto de estimación en Estadística. El primer hecho es la invitación formulada por **John Tukey**, uno de los autores del término **estimación robusta**, a otros estadísticos estudiosos del tema a pasar el curso académico 1970-1971 en la Universidad de Princeton, Nueva Jersey. Los visitantes eran **Peter J. Huber** del Instituto de Tecnología de Zurich; **Peter Binkel** de la Universidad de California y **Frank Hampel** de la Universidad de Zurich. En este largo seminario participaron estudiantes, profesores, becarios y tambien científicos de los laboratorios Telefónicos Bell, entre los cuales cabe destacar a **David Andrews.**

En este largo seminario se clarificó el concepto de estimación robusta y se crearon varios nuevos estimadores robustos. Todo ello se recoge en la publicación "Robust Estimates of Location" de Andrews, Binkel, Hampel, Huber, Rogers y Tukey (1972).

El segundo hecho es la publicación en 1979 del artículo "Computers and the theory of statistics: thinking the unthinkable" de **Bradley Efron**, profesor de la Universidad de Stanford. En dicho artículo formaliza por primera vez la técnica **Bootstrap**.

Estos dos hechos, que más tarde se traducen en grandes avances estadísticos inferenciales, inviables sin métodos computerizados, tienen como objetivo común extraer **la máxima información de los datos de la muestra.**

No seria exagerado observar que este objetivo lo tienen también todas las técnicas expuestas en este libro sobre Análisis Exploratório de Datos, pues ésta debe ser la aspiración de cualquier técnica estadística.

En efecto, después de haber hecho un análisis exhaustivo de la muestra mediante técnicas E.D.A., a menudo hemos de dar un nuevo paso. Normalmente, toda investigación, simple o compleja, tiene como objetivo conocer **el valor verdadero en la población**. Ya que la mayoría de las veces, o casi siempre, se trabaja con una muestra, uno de los problemas estadísticos fundamentales reside en conseguir **la mejor estimación posible** de dicho valor o valores verdaderos en la población origen de la muestra. Expresado de otra forma, se trata de averiguar la aproximación mejor al "verdadero valor", llamado en estadística parámetro. En Estadística Inferencial Clásica un buen estimador puntual tiene que tener como mínimo cuatro propiedades:

1) Un estimador es **consistente** si tiende al valor verdadero a medida de que el número de datos tiende a infinito.

2) Un estimador es **no sesgado** si los valores esperados son iguales a los verdaderos.

3) Un estimador es **eficiente** si su variancia es pequeña.

4) Un estimador es **suficiente** si no es necesario añadir ningún otro estimador para mejorar la estimación del parámetro.

Los métodos estadísticos clásicos de inferencia nos permiten dar respuestas diversas a esta pregunta de naturaleza más o menos fiables. Es decir, nos dan los intervalos de confianza que el estimador merece.

La Estadística Inferencial Clásica basa casi todas las estimaciones en la hipótesis, generalmente imposible de verificar, de normalidad de la distribución o distribución de Gauss en la población origen, así llamada en honor del matemático alemán Karl Friedrich Gauss. Dicha hipótesis presupone además que los errores —las fluctuaciones aleatorias— de los valores empíricos se encuentran siempre simétricamente repartidos alrededor del valor verdadero.

La experiencia de años demuestra que la hipótesis "gaussiana" nos permite dar intervalos dignos de confianza con cálculos relativamente sencillos, aún en el caso de distribuciones que no se ajustan exactamente a la distribución normal. Sabido es, que la representatividad de la muestra depende principalmente de dos cosas: de como ha sido seleccionada y de su tamaño.

Así, cuando los datos empíricos no verifican las hipótesis "gaussianas", porque son pequeñas o por cualquier otro motivo, las prediciones del valor verdadero son mucho menos fiables, ya que, como se ha dicho, en la mayoría de los casos se han hecho predicciones fundándose en hipótesis de normalidad u otros supuestos.

Por este motivo es presumible que en algunos casos se esté en condiciones de dar resultados muy satisfactorios de "lo que sucede verdaderamente en los datos empíricos (muestra)" pero que sea muy difícil realizar inferencias a nivel poblacional.[1]

Quisieramos hacer hincapié en la idea de que los estimadores no son en si válidos o inválidos, buenos o equivocados, sino que lo que pretendemos es **buscar el estimador más adecuado a cada tamaño de muestra, a cada distrubución de la variable, etc.**

Como muestra de que este tema ha sido una gran preocupación de los estadísticos de todos los tiempos, baste recordar que ya Fisher (1935)

1. Sería deseable que el lector tuviera conocimientos de métodos clásicos de estimación de párametros, tales como el de los minimos cuadrados o el de la máxima verisimilitud.

propuso como una réplica o alternativa a las técnicas paramétricas, las denominadas "técnicas libres de función de distribución" también llamadas no paramétricas, que presentaron problemas de significación.

Toda teoría que logre hacer más precisas las investigaciones estadísticas inferenciales, ha de interesar a los investigadores en todas las areas y más especialmente a los de Ciencias Humanas, Sociales y de la Salud. En estas areas la mayoría de veces no se conoce la distribución exacta de las variables. En estos casos es mejor utilizar **estimadores robustos.**

Para que un estimador sea robusto necesita cumplir como mínimo dos propiedades: ha der ser resistente y eficiente. Las técnicas de estimación que proponemos en este capítulo son **resistentes y robustas.** Todas ellas requieren supuestos menos restrictivos, necesitan métodos intensivos de ordenador, y pueden aplicarse aún en el caso de que la muestra sea muy pequeña.

Siguiendo la filosofía del E.D.A., estos estimadores tendrán dos características fundamentales:

1) Calcularemos varios estimadores en vez de uno, simpre que sea posible, lo cual nos permitirá evaluar su constancia.

2) Evitaremos hacer hipótesis demasiado elaboradas sobre las funciones de distribución correspondientes a las variables aleatorias estudiadas.

Las técnicas que se exponen son, en principio, independientes de cualquier función de distribución; es preferible "explorar" el comportamiento de dicha función y sus posibles fluctuaciones.

Divideremos este capítulo en dos partes. En primer lugar se hará una revisión simple de estimadores robustos de posición y en segundo lugar, se introducirán las técnicas estadísticas basadas en la generación de nuevas muestras a partir de la original, tales como el Jaccknife y Bootstrap.

"Todas las distribuciones son normales en el centro".

Principio de Winsor. (Tukey 1960 pág. 457).

6.2. CARACTERÍSTICAS DE LOS ESTIMADORES ROBUSTOS DE POSICIÓN

Convendría, antes de iniciar el estudio de la estimación robusta, repasar algunos conceptos que caracterizan a tales estimadores; algunos de ellos ya planteados en el primer capítulo de este volumen.

6.2.1. RESISTENCIA

La resistencia (Tukey, 1977) permite alcanzar una cierta insensibilidad a los errores localizados, que pueden existir en los datos. Un método resistente debe dar un resultado tal que no cambiaría más que ligeramente si una pequeña parte de los datos se reemplazase por valores diferentes de los originales. Los métodos resitentes deben dar gran importancia a la parte central de los datos y poca a los posibles valores lejanos. Por ejemplo, la mediana es un valor resistente, la media, por contra, no presenta esta característica.

Un estimador es resistente si viene afectado sólo en un grado limitado, bien por un pequeño número de errores grandes o bien por cualquier número de pequeños errores de redondeo o de agrupamiento.

6.2.2. PUNTO DE COLAPSO

Se dice que un estimador alcanza su punto de colapso (breakdown point) (Hampel 1968) cuando limita la proporción máxima de observaciones anómalas que pueden presentarse sin que cambie el resultado del esti-

mador. Un estimador será más resistente en la medida que su punto de colapso sea superior a 0. Por ejemplo, el punto de colapso de la mediana es:

$$\begin{array}{ll} (1/2) - (1/n) & \text{si } n \text{ es par} \\ (1/2) - (1/2n) & \text{si } n \text{ es impar} \end{array}$$

El punto de colapso de un estimador que trate las observaciones uniformemente no puede ser mayor que 0.5. No obstante, hay muchos estimadores que, tratando uniformemente las observaciones, tienen como valor de punto de colapso un valor menor de 0.5. El punto de colapso de la media es 0. La media no es resistente. Un sólo valor puede hacer cambiar el resultado de la media.

Supongamos que cierto número de observaciones de una muestra se reemplazan en las colas; el punto de colapso de un estimador será la proporción más pequeña de la muestra que puede substituirse arbitrariamente sin que el valor del estimador se altere. Los estimadores con punto de colapso alto no cambiarán por el hecho de que haya valores alejados o se modifiquen unas pocas observaciones. El punto de colapso de los estimadores resistentes y robustos oscila normalmente entre 0.1 y 0.4. Veamos un sencillo ejemplo con los siguientes valores:

5 7 8 9 10 10 11 12 13 14

El punto de colapso de la mediana, siendo n=10 y por tanto par, seria:

$$(1/2) - (1/n) = (1/2) - (1/10) = 0.5 - 0.1 = 0.4$$

lo cual significa dejar el 40% de valores a cada uno de los dos lados:

5 7 8 9 |10 10| 11 12 13 14

6.2.3. ROBUSTEZ

La robustez (Andrews et als., 1972) implica una insensibilidad a las desviaciones de las hipótesis subyacentes a un modelo probabilístico. A

menudo queremos obtener métodos en que las hipótesis de aplicación no sean demasiado restrictivas. Como hemos dicho los métodos clásicos se justifican sólo si se supone una distribución "gaussiana" de los datos.

Un estimador robusto debería ser insensible como mínimo a dos tipos de anomalías encontradas en la muestra:

1) Unas pocas desviaciones grandes de los datos tomadas a a menudo como valores alejados (outliers).

2) Otras desviaciones más numerosas de los datos (pequeñas pero numerosas, suficientes para ser como mínimo comparables con la separación de observaciones ordenadas adyacentes), como, por ejemplo, agrupar o redondear.

Dado que queremos evitar hipótesis demasiado elaboradas sobre el comportamiento de las fluctuaciones de los datos, buscamos robustez, entendida como insensibilidad a los supuestos tácitos. La Robustez garantiza que un estimador es bueno para una serie de distribuciones sin que sea necesariamente mejor para una en particular.

Se dice que un estimador robusto es eficiente sobre una amplitud de distribuciones si su varianza es cercana al mínimo para cada distribución.

Los estimadores que dependen de grupos simples, no restrictivos acerca de la distribución subyacente y no son sensibles a tales hipótesis se llamam estimadores robustos. Rapacchi (1991) propone el ejemplo del análisis de la varianza como método no robusto, dado que necesita que los residuales sigan la ley normal con una misma desviación típica para cada uno de los grupos estudiados.

6.2.4. DISTRIBUCIÓN SIMÉTRICA

Rosember y Gasko (1989) definen una distribución simétrica del siguiente modo:

> *"La distribución de una variable aleatoria x es simétrica alrededor de un centro de simetria c si las variables aleatorias* $x - c$ *y* $-(x - c)$ *están idénticamente distribuidas, entonces* $c = x$".

Se sigue que una variable aleatoria con una distribución simétrica alrededor de c, de media $\overline{x}$ y de mediana $x_{0.5}$ necesariamente $c = x_{0.5}$ y si $\overline{x}$ es finita, entonces $c = \overline{x} = x_{0.5}$.

6.2.5. ESTADÍSTICOS DE ORDEN

Sea $[x_p, x_j, \ldots, x_k]$ una muestra de tamaño n, las observaciones reordenadas dentro de un orden de magnitud creciente, llamadas $[x_1, x_2, \ldots, x_n]$ son llamados estadísticos de orden de la muestra y x_i se llama $i-$ésimo estadístico de orden.

Una vez revisados estos conceptos y características, presentaremos, sucintamente, algunos de los estimadores robustos más empleados.

6.3. L – ESTIMADORES

Ante las posibles deficiencias de las estimaciones usuales de la media y la mediana debidas al no cumplimiento de los supuestos, como se ha explicado anteriormente, se presentan varios estimadores resistentes y robustos de posición como los más representativoss de los valores de tendencia central que habrá en la población origen de la muestra.

Introducimos varios procedimientos de estimación no sesgados para el centro de una distribución simétrica. A continuación, buscamos entre los calculados, el estimador más robusto.

Se llaman L–estimadores (Andrews et als., 1972) a los estimadores que son combinaciones lineales de estadísticos de orden.

Sea $[x_1 \leq x_2 \leq \cdots \leq x_n]$ los estadísticos de orden de una muestra de tamaño n. Sea $[a_1, a_2, \ldots, a_n]$ los números reales que verifican que

$0 \leq a_i \leq 1[i = 1, 2, \ldots, n]$, tal que $\sum_{(i=1)}^{n} a_i = 1$. Un L-estimador "T" con peso $[a_1, a_2, \ldots, a_n]$, es:

$$T = \sum_{(i=1)}^{n} a_i x_{(i)} \tag{6.1}$$

En este apartado estudiaremos estimadores simples de orden de una distribución simétrica. Normalmente este tipo de estimadores no se aplican a muestras grandes.

6.3.1. LA MEDIA Y LA MEDIANA

La Mcdia es el L-estimador en el que todos los pesos son iguales.

$$a_i = 1/n \quad [i = 1, \ldots, n]$$

La mediana, por su parte, es un L-estimador en el que sólo interviene el estadístico de orden central si n es impar, y que es la media de dos estadísticos de orden central si n es par.

$$\text{Si } n = (2p+1) \quad a_i \quad \begin{cases} 1 & \text{si } i = p \\ 0 & \text{si } i \neq p \end{cases}$$

$$\text{Si } n = 2p \qquad a_i \quad \begin{cases} 1/2 & \text{si } i = p \text{ ó } i = (p+1) \\ 0 & \text{si no lo es} \end{cases}$$

6.3.2. LAS MEDIAS RECORTADAS

Se llaman medias recortadas a las medias que se obtienen después de haber eliminado, a ambos lados de la distribución, una proporción α de valores. Asignamos al valor resultante el término media recortada con proporción α a cada trozo cortado por T_α.

Por ejemplo el 30% de media recortada de una muestra de tamaño 20 es una muestra simple de tamaño 14 en la que se han eliminado las tres puntuaciones más altas y las tres más bajas. Así, la media aritmética está considerada como 0% de media recortada y la mediana como, aproximadamente, el 50% de media recortada.

$$
\begin{array}{lll}
a_i = 0 & \text{si} & i \leq g \text{ ó } i \geq (n-g+1) \\
a_i = 1 - r/n(1-2\alpha) & \text{si} & i = g+1 \text{ ó} \\
 & & i = n-g \\
a_i = 1/n(1-2\alpha) & \text{si} & (g+2) \leq i \leq (n-g-l)
\end{array}
$$

siendo $g = \alpha n$
$r = \alpha n - g$
$\alpha =$ el corte
$n =$ número de casos de la muestra

Para reflejar el empleo de las medias recortadas proponemos unos datos que se relacionan con el consumo doméstico de energía eléctrica por 1000 habitantes (en miles de Kw/h) en el año 1985 en las poblaciones de la provincia de Castellón (Caja de Ahorros Valenciana, 1987. Indicadores Socio-económicos de la Comunidad Valenciana). En la tabla 6.1. se muestran 120 de estos datos.

Tabla 6.1. Consumo doméstico de energía eléctrica (en miles de Kw/h) en el año 1985 en 120 poblaciones de la provincia de Castellón.

	498.1	324.4	403.4
294.9	452.1	14.7	292.6
277.7	519.9	237.1	275.7
730.5	615.0	494.3	480.9
479.0	394.8	614.5	561.0
379.4	624.4	502.4	371.2
608.4	572.1	384.7	399.6
464.0	1040.1	558.1	455.5
439.6	588.8	332.7	649.3
482.6	348.8	308.4	316.9
695.5	519.7	375.0	477.0

.../...

.../...

485.5	1975.8	108.5	263.7
698.2	279.5	397.1	746.7
490.0	1343.7	662.6	210.3
808.6	276.2	186.7	537.0
424.4	310.3	517.2	590.4
227.2	273.5	615.0	271.8
424.4	266.7	1012.9	524.1
227.2	348.4	534.3	432.1
436.9	529.3	409.1	418.2
592.9	693.4	402.4	316.8
124.6	609.9	979.1	273.3
408.6	424.9	520.1	299.3
526.7	479.8	678.8	183.4
2190.0	478.7	443.0	460.2
260.0	325.4	380.0	517.0
550.4	539.2	473.7	713.6
602.5	125.1	424.3	153.6
620.5	552.6	604.9	438.7
430.5	487.4	398.8	372.0
409.3	504.5	415.6	373.9

Calcularemos las medias recortadas con diferentes valores de α: 0%, 0.05%, 0.10%, 0.20%, 0.30%, 0.40% y 0.50‰

$$T_{(0.00)} = (1/120)\sum_{i=1}^{120} x_{(i)} = (52.624/120) = 438.53$$

Este valor coincide, obviamente, con la media aritmética.

$$T_{(0.05)} = (1/108)\sum_{i=7}^{114} x_{(i)} = (49.114/108) = 454.76$$

$$T_{(0.10)} = (1/96)\sum_{i=13}^{108} x_{(i)} = (43.371/96) = 451.78$$

$$T_{(0.20)} = (1/72)\sum_{i=25}^{96} x_{(i)} = (32.427/72) = 450.38$$

$$T_{(0.30)} = (1/48)\sum_{i=37}^{84} x_{(i)} = (21.600/48) = 450.00$$

$$T_{(0.40)} = (1/24)\sum_{i=49}^{72} x_{(i)} = (10.754/24) = 448.08$$

$T_{(0.50)}$ coincide aproximadamente con la mediana.

6.3.3. CENTRIMEDIA Y TRIMEDIA

Es la media recortada para $\alpha = 0.25$. En realidad, es la media intercuartílica, ya que justamente es la media de los valores centrales, excluyendo el 25% de datos inferiores y el 25% de datos superiores. Para calcular la centrimedia se divide la suma de los valores comprendidos en el 50% central de la distribución por el número de estos (vease cap.1).

Se define la trimedia por :

$$TRI = 1/4(H_1 + 2Mdn + H_3) \qquad (6.2)$$

6.3.4. MEDIANA GENERALIZADA

Para "n" impar, la mediana generalizada es la media de los tres estadísticos de orden central $5 \leq n \leq 12$ o de 5 estadísticos de orden central para $n \geq 13$.

Para "n" par, la mediana generalizada es una ponderación de cuatro estadísticos de orden central para $5 \leq n \leq 12$ y con pesos 1/6, 1/3, 1/3, 1/6; para $n \geq 13$ es una media ponderada de seis estadísticos de orden central con pesos 1/10, 1/5, 1/5, 1/5, 1/5, 1/10.

En la bibliografía relativa a los L-estimadores se plantean estrategias más complejas de las aqui presentadas. Hogg (1974) ofrece algunas de estas posibilidades. Otra cuestión a considerar es la decisión que se debe adoptar con respecto a que recortar. Muchas veces, sólo haciendo el

diagrama de tallo y hojas ya vemos si la distribución es simétrica y con ello facilita la decisión de que és lo más conveniente. Si la muestra es muy pequeña Hoaglin, Mosteller y Tukey (1981) recomiendan lo siguiente:

1. Para $n < 7$ usar la mediana.
2. Para $n = 7$ recortar dos observaciones para cada cola.
3. Para $n > 8$ recorte 25% para cada cola.

6.4. M-ESTIMADORES DE POSICIÓN

Presentamos en este apartado unos métodos robustos de estimación un poco más complejos, pero que ofrecen grandes ventajas de ejecución, comodidad y flexibilidad.

Los M-estimadores minimizan las funciones objetivas[2] de una forma más general que la conocida suma de residuos al cuadrado asociada a la media de la muestra.

En lugar de elevar al cuadrado la desviación de cada observación x del estimador t, aplicamos una función $\rho(x;t)$ y hallamos la función objetiva extendiendo a toda la muestra la suma $\sum_{i=1}^{n} \rho(x_i;T)$; a menudo $\rho(x;t)$ depende de x y de t y también de $(x-t)$, con lo cual es más adecuado escribir $(x-T)$.

El término M, usado para estos estimadores, es el mismo que el estimador Máximo Verosímil. En los M-estimadores incluimos **todas las puntuaciones de la muestra**, pero ponderándolas de forma que las observaciones esten más cercanas al centro, es decir, modificando los valores para que sean más parecidos al valor central determinado. En consecuencia, pues, habrá tantos M-estimadores como pesos diferentes se utilicen.

2. Dado que el objetivo de la función del procedimiento mínimo cuadrático es minimizar la suma del cuadrado de los residuos, a la función que realiza esta misión se la denomina **función objetiva**.

Normalmente, todos los M-estimadores asignan pesos que decrecen cuando la distáncia al centro de distribución crece. Los cálculos para obtener los M-estimadores se hacen con la ayuda del ordenador, iterativamente, añadiendo a cada función objetiva una constante o peso.

La media aritmética puede considerarse un M-estimador al que asignamos el peso 1 a cada valor. Asimismo, los L-estimadores pueden considerarse M-estimadores cuando se calculan, por ejemplo, medias recortadas (se divide la distribución en dos grupos, uno se excluye y con el otro se calcula la media). Podemos considerar las medias recortadas como una media ponderada en la que en el caso de que se incluya un valor su peso vale 1 y en el caso de que se excluya, ese valor tiene por peso 0.

Los M-estimadores $T_n(x_1, \ldots\ldots, x_n)$ para la función ρ y la muestra $(x_1, \ldots\ldots, x_n)$ es el valor de t que minimiza la función objetiva $\sum_{i=1}^{n} \rho(x;t)$ de tal manera que T tiene que satisfacer que

$$\sum_{i=1}^{n} \Psi(x_i, t) = 0 \qquad (6.3)$$

Para que un M-estimador sea robusto es deseable que cumpla las siguientes propiedades:

1) Los buenos M-estimadores deben tener **un punto de colapso alto.**

2) Los buenos M-estimadores son relativamente **eficientes** para una amplia familia de distribuciones de probabilidad.

3) Los buenos M-estimadores deben ser **resistentes.**

4) Es claramente deseable que los buenos M-estimadores tengan las propiedades usuales de las muestras grandes, propias de los estimadores máximo verosímiles, las cuales son: **consistencia y normalidad asintótica.**

5) Finalmente, y dado que una de las filosofías E.D.A. en estimación es calcular varios estimadores en vez de uno, estos deben ser fáciles de calcular por ordenador.

Abordaremos a continuación, cuál son los aspectos generales del **procedimiento de cálculo de este tipo de estimadores.**

En general, todos los M-estimadores de posición (Tukey, Huber, Hampel y Andrews), excepto la media y la mediana, necesitan una transformación de los datos antes de empezar los cálculos.

Normalmente, los M-estimadores de posición usan unas escalas de índices auxiliares, que tienen como propiedad más relevante su resistencia, distintos a los clásicos. La mayoria centran y reducen la variable del siguiente modo:

$$u_i = (x_i - t)/cS_n \tag{6.4}$$

siendo x_i = el valor de cada observación.
c = es una constante, denominada **potenciómetro, que presenta la condición** $(o < c)$ **y que toma valores** $3 \leq c \leq 12$.
S_n = el índice que mide la dispersión de la distribución.

En general, los valores de T que minimizan la función son estimadores de posición. De este modo:

$$u = \frac{|\text{valor de cada caso} - \text{estimador de situación}|}{\text{estimador de dispersión}} \tag{6.5}$$

siendo los estimadores más usuales de dispersión:

- Mad (Mediana de las desviaciones absolutas).
- La amplitud intercuartílica.[3]
- La desviación típica.

Un M-estimador T_n será el valor que minimize $\sum_{i=1}^{n} \rho(u)$, siempre que se verifiquen las condiciones de igual varianza en posición y escala (Colin y Goodall, 1983).

En la mayoría de los casos no se selecciona $\rho(u)$, sino su derivada $\Psi(u)$, ya que, dado que el objetivo es minimizar la ecuación, es mejor tomar la derivada de esta. Veamos cuales son los M-estimadores de posición más comunes.

3. Para estos dos estimadores de dispersión consúltese el primer capítulo de este volumen.

6.4.1. LA MEDIA ARITMETICA Y LA MEDIANA

El más usual de los M-estimadores es el la media aritmética siendo la estimación mínima cuadrática el cuadrado del residual.

$$\rho(x;t) = (x-t)^2 \tag{6.6}$$

Otro ejemplo usual es el M-estimador de la mediana, donde la función objetiva es el valor absoluto del residual.

$$\rho(x;t) = |x-t| \tag{6.7}$$

que corresponde a la función

$$\Psi(x;t) = \text{sgn}(x-t) \tag{6.8}$$

$$\text{sgn}(u) = \begin{cases} +1 & \text{si } u > 0 \\ 0 & \text{si } u = 0 \\ -1 & \text{si} u < 0 \end{cases}$$

6.4.2. ESTIMADOR DE HUBER

El estimador de Huber es un M-Estimador que se define por:

$$T_n \text{ de } \sum_{i=1}^{n} \Psi(u_i) = 0 \tag{6.9}$$

o una vez centrada y reducida la variable

$$U_i = (x_i - T_n)/S_n \tag{6.10}$$

siendo x_i las observaciones y S_n el valor del MAD. De esta forma la expresión (6.9) se puede establecer como sigue:

$$\sum_{i=1}^{n} \Psi\left[(x_i - T_n)/S_n\right] = 0 \tag{6.11}$$

Aquí, el rango del valor absoluto de "U" se compara con una constante, denominada k, que tiene por condición $(0 < k)$, es decir, estrictamente positiva. Por ejemplo, el paquete estadístico SPSS/PC+ usa un valor de $k = 1.339$.

Si $|U| \leq k$ entonces $\Psi(U) = U$ siendo la correspondiente función objetiva ρ la siguiente:

$$\rho(U) = 1/2(U)^2 \tag{6.12}$$

Por otra parte, si $|U| > k$ entonces $\Psi(U) = k$ por signo (sgn) de (U), siendo la función objetiva ρ correspondiente:

$$\rho(U) = k|U| - 1/2k^2 \tag{6.13}$$

mostrando la figura 6.1. la derivada de la función que hemos definido en 6.13.

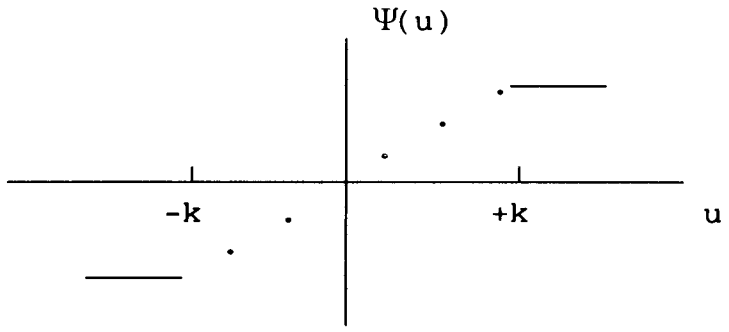


Fig. 6.1. Representación gráfica de la función $\Psi(u)$.

La función objetiva del estimador de Huber es cuadrática en el centro y lineal en los extremos. La curva de influencia del estimador de Huber es constante para todas las observaciones a partir de un cierto punto. La función Ψ es, como en la figura 6.1., monótona, el estimador de Hubber es un ejemplo de M-estimadores monótonos.

Este estimador es un buen estimador cuando la distribución se ajusta "razonablemente" a la normal, pero no es recomendable cuando hay valores extremos. Es, pues, un estimador no demasiado resistente y por lo tanto influenciado por los valores alejados (outliers).

6.4.3. ESTIMADOR "DOBLE PESO" DE TUKEY

El estimador "doble peso" (biweight) de Tukey es uno de los M-estimadores que se define por:

$$T_n \text{ de } \sum_{i=1}^{n} \Psi(u_i) = 0 \tag{6.14}$$

o una vez centrada y reducida la variable:

$$u_i = (x_i - T_n)/cS_n \tag{6.15}$$

donde x_i son las observaciones.
c es el valor del potenciómetro. El programa SPSS/PC+ define $c = 4.685$.
S_n es una medida de dispersión(MAD).

Normalmente, los valores T_n que minimizan la función son estimadores de posición. De este modo, la expresión T_n se puede plantear:

$$\sum_{i=1}^{n} \Psi\left[(x_i - T_n)/cS_n\right] = 0 \tag{6.16}$$

Si el valor absoluto del rango de "U" es $|U| \leq 1$ entonces $\Psi(U) = u(1-u^2)^2$, siendo la función objetiva ρ correspondiente

$$\rho(U) = 1/6[1 - (1-u^2)^3] \tag{6.17}$$

si el valor absoluto del rango de "U" es $|U| > 1$, entonces $\Psi(U) = 0$, siendo la función objetiva ρ la siguiente:

$$\rho(U) = 1/6 \tag{6.18}$$

6.4.4. ESTIMADOR DE ANDREWS

El estimador de Andrews (1972), normalmente llamado "onda de Andrews", es un M-estimador que se define por:

$$T_n \text{ de } \sum_{i=1}^{n} \Psi(U_i) = 0 \tag{6.19}$$

o, una vez centrada y reducida la variable:

$$U_i = (x_i - T_n)/cS_n \tag{6.20}$$

donde X_i son las observaciones.

c es el potenciómetro ($3 \leq c \leq 12$). Por ejemplo, el programa SPSS/PC+ emplea $c = 1.340$; Andrews (1972) $c = 2.1\pi$; Hogg (1979) $1.5\pi \leq c \leq 2\pi$ y Gross (1976) $1.8\pi \leq c \leq 2.4\pi$.

S_n en el valor del MAD.

De este modo la solución (6.19) se puede reordenadar de la siguiente forma:

$$\sum_{i=1}^{n} [\Psi(x_i - T_n)/cS_n] = 0 \tag{6.21}$$

Si el valor absoluto del rango de "U" es $|U| \leq 1$, entonces

$$\Psi(U) = 1/\pi \sin(\pi U) \tag{6.22}$$

siendo la función objetiva ρ correspondiente:

$$\rho(U) = (1/\pi^2)(1 - \cos \pi U) \tag{6.23}$$

pero, si por contra, el valor absoluto del rango de $|U| > 1$ entonces, $\Psi(U) = 0$, y la función objetiva ρ correspondiente:

$$\rho(U) = 2/\pi^2 \tag{6.24}$$

6.4.5. ESTIMADOR DE HAMPEL

Hemos visto como en el estimador de Huber la curva de influencia es constante para todas las observaciones a partir de un cierto punto. El estimador de Hampel pretende ser más resistente, haciéndolo, para ello, más compleja su formulación y cálculo. El estimador de Hampel es un M-estimador que se define por:

$$T_n \text{ de } \sum_{i=1}^{n} \Psi(u_i) = 0 \tag{6.25}$$

o una vez centrada y reducida la variable

$$u_i = (x_i - T_n)/S_n \tag{6.26}$$

donde x_i son las observaciones
S_n el valor del MAD.

De este modo, la expresión (6.25) se formula:

$$\sum_{i=1}^{n} \Psi\left[(x_i - T_n)/S_n\right] = 0 \tag{6.27}$$

Aquí, el rango del valor absoluto de "U" se compara con tres constances (a, b y c) con la única condición que $0 < a \leq b \leq c$. Si $|U| \leq a$, entonces $\Psi(U) = U$, siendo la función objetiva ρ la siguiente:

$$\rho(U) = 1/2(u^2) \tag{6.28}$$

En el caso que $a < |U| \leq b$, entonces $\Psi(U) = a$ por signo (sgn) de (U), siendo la función ρ:

$$\rho(U) = a|U| - 1/2a^2 \tag{6.29}$$

Por último, si $b < |U| \leq c$, entonces:

$$\Psi(U) = a\left[(c - |U|)/(c - b)\right] \operatorname{sgn}(U) \tag{6.30}$$

siendo la función ρ la siguiente:

$$\begin{aligned} \rho(U) &= ab - 1/2a^2 + (c - b)(a/2)\left\{1 - \left[(c - |u|)/(c - b)\right]^2\right\} \cdot \\ &\cdot \quad ab - 1/2a^2 + (c - b)(a/2) \end{aligned} \tag{6.31}$$

Señalar, a título informativo, que el programa SPSS/PC+ emplea los siguientes valores $a = 1.70$; $b = 3.4$ y $c = 8.5$.

Para finalizar este apartado dedicado a los M-estimadores, poner de manifiesto que el mismo sólo pretende exponer la formulación matemática de estos estimadores robustos que, en general, utilizan los programas estadísticos. En general, la utilidad máxima de este tipo de estimación se centra en el análisis que puede hacerse a partir de la comprobación de que los distintos estimadores ofrecen resultados parecidos, lo cual asegurará

una estimación exenta de sesgo. Para ejemplificar este aspecto, sirva la siguiente distribución simulada:

{28 26 33 24 34 27 16 40 29 02 29 22 24 21 25 30 23 29 31 18 03}

Los resultados que se obtienen con el paquete SPSS/PC+ (procedimiento Examine) a partir de estos datos se ofrecen en la figura 6.2. y en la tabla 6.2.

Tabla 6.2. Estadísticos descriptivos y estimaciones robustas de los datos simulados.

Media	24.4762	Error Est.	1.9888	Min.	2.00	Asim.	-1.181
Mediana	26.0000	Varianza	83.0619	Máx.	40.00	E.E.As.	.501
5% Trim	24.8757	Desv. Típ.	9.1138	Rango	38.00	Curtos.	1.883
				IQR	8.00	E.E.Cu.	.971

M-Estimadores

Huber	(1.339)	25.9177	Tukey	(4.685)	26.5397
Hampel	(1.700,3.400,8.500)	26.2620	Andrews	(1.340·π)	26.5371

```
Extremos  |  (2), (3)
      1.  |  68
      2*  |  12344
      2.  |  5678999
      3*  |  0134
Extremos  |  (40)
```

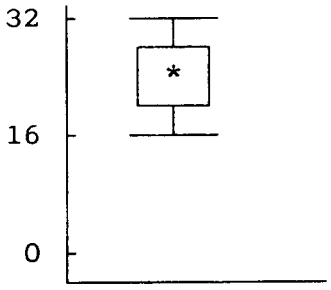


Fig. 6.2. Diagrama de Tronco y Hoja y Diagrama de Caja de los datos simulados.

Obsérvese que los estimadores de Hampel, Tukey y Andrew dan valores en torno de 26, cercanos a la mediana, mientras que el estimador de Huber ofrece un valor de 25.9, el cual está un poco influenciado por los valores alejados de la distribución analizada (2, 3 y 40).

Si reproducimos estos cálculos con los datos que aparecen en la tabla 6.1. obtendremos los resultados que a continuación se detallan

Tabla 6.3. Estadísticos descriptivos y estimaciones robustas de los datos de la tabla 6.1.

Mean	486.3408	Std Err	25.9255	Min	14.700	Skewness	3.3092
Median	441.3000	Varian.	80655.90	Max	2190.00	S E Skew	.2209
5%Trim	454.7167	Std Dev	283.9998	Range	2175.30	Kurt.	16.3965
				IQR	233.05	S E Kurt	.4383

M-Estimators

Huber	(1.339)	449.6460	Tukey	(4.685)	439.9513
Hampel	(1.700,3.400,8.500)	442.7948	Andrew	(1.340*pi)	439.9712

Las representaciones gráficas correspondientes a los datos que nos ocupan se encuentra en la siguientes figuras número 6.3.a y 6.3.b.

Frequency	Stem	&	Leaf
1.00	0	·	1
6.00	1	·	022588
17.00	2	·	12236667777777999
20.00	3	·	01112234477777889999
32.00	4	·	00000112222333334556677778888999
19.00	5	·	0011122233355567899
15.00	6	·	000011122467999
3.00	7	·	134
1.00	8	·	0
6.00	Extremes		(979),(1013),(1040),(1344),(1976),(2190)

Stem width: 100
Each leaf: 1 case(s)

Fig. 6.3.a. Diagrama de Tronco y Hojas de los datos de la tabla 6.1.

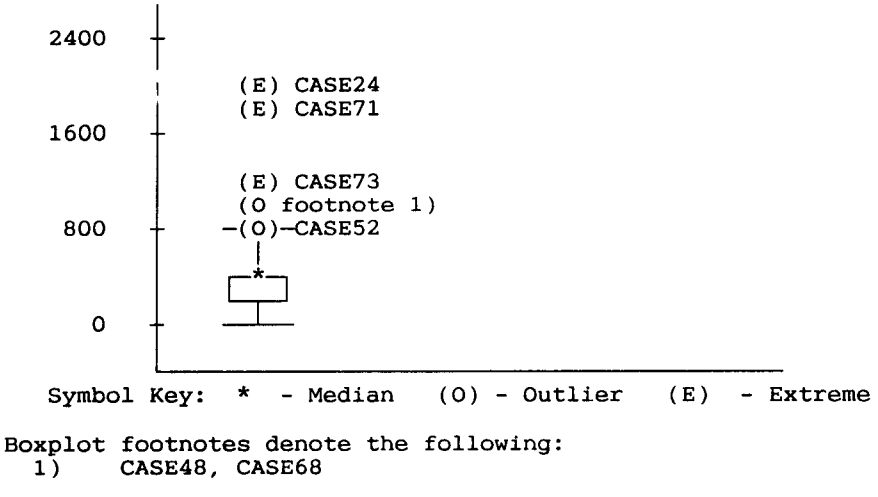


Fig. 6.3.b. Diagrama de Caja de los datos de la tabla 6.1.

Aquellos lectores que deseen profundizar en esta temática, buscando, por ejemplo, los correspondientes intervalos de confianza, encontrarán una magnífica revisión de esta cuestión en Iglewicz (1989).

6.5. MÉTODOS DE ESTIMACIÓN BASADOS EN EL REMUESTREO

Cualquier procedimiento que logre hacer más precisas las inferencias estadísticas ha de interesar a los investigadores en todas las áreas y, más especialmente, a los de Ciencias Sociales, Humanas y de la Salud.

Las técnicas que se exponen en este apartado pretenden resolver el problema de la fiabilidad de las estimaciones estadísticas sin necesidad de suponer que los datos gozan de distribución normal. También es útil para trabajar con muestras pequeñas y abaratar el coste de las investigaciones. Los métodos de investigación no paramétricos del error estándar se aprovechan de los adelantos técnicos de los cálculos por ordenador y serían complejos sin ellos.

6.5.1. JACKKNIFE

El **Jackknife** constituye una nueva propuesta sobre estimación estadística no paramétrica que fue introducida por Quennouille (1949) y Tukey (1958). Se puede hallar una excelente revisión técnica en el trabajo de Miller (1974) aparecido en el número 61 de Biometrika, titulado "The Jackknife, a review".

El nombre de "Jackknife", ideado por Tukey, por ser una técnica, no se acostumbra a traducir y es por ello que en este libro usaremos el nombre original.[4]

El Jackknife es un método para estimar el error muestral. La estimación Jackknife de σ, que se denomina normalmente $\sigma_{(j)}$ se obtiene normalmente de la siguiente forma:

1) Obtenemos una sola muestra de tamaño n.

2) Se obtienen k muestras de tamaño $n-1$ (por supresión de un valor cada vez distinto de la serie de datos independientes con reemplazamiento).[5]

3) Se calcula el estadístico deseado para cada serie de datos de tamaño $n-1$ (nueva muestra). Así obtendremos, a partir de una sola muestra de tamaño n, un estadístico $\theta_{(.)}$ estimado a partir de todos los

4. Propiamente, significa navaja o máquina de cortar cruel y rápida, que se puede usar en multitud de situaciones.

5. Vale la pena observar que no es necesario generar siempre las nuevas muestras con un tamaño $n-1$, sino que pueden ser definidas por un tamaño $n-g$, siendo g un número real menor que n.

estadísticos θ_i estimados con las muestras que hayamos generado a partir de la primera.

4) Se efectua la estimación del error muestral mediante la siguiente fórmula:

$$\hat{\theta}_{(.)} = (1/n)\sum_{i=1}^{n}\hat{\theta}_i \tag{6.32}$$

$$\hat{\sigma}_{(j)} = \left\{[(n-1)/n]\sum_{i=1}^{n}\left(\hat{\theta}_i - \hat{\theta}_{(.)}\right)^2\right\}^{1/2} \tag{6.33}$$

Se introduce el factor $[(n-1)/n]$ para conseguir que el Jackknife constituya una buena respuesta a la desviación estándar del estadístico.

El Jacknife, en cierta manera, genera nuevas muestras con la misma idea de las medias recortadas, en realidad cada nueva muestra no es más que un recorte de uno o más valores de la original.

Veamos un ejemplo de esta estrategia con una muestra simulada de ocho valores:

$$\{5,\ 6,\ 6,\ 6,\ 7,\ 8,\ 9,\ 10\}$$

presentando, estos datos, una mediana de 6.5 y una estimación del error estándar que sigue la siguiente expresión:

$$\hat{\sigma}_{(j)} = \left\{[(n-1)/n]\sum_{i=1}^{n}(Mdn_i - Mdn_{(.)})^2\right\}^{1/2} \tag{6.34}$$

donde Mdn_i es la mediana de cada muestra "Jacknife" y $Mdn_{(.)}$ es la media de esas medianas. Analizemos algunas de las posibles muestras "Jackknife", suprimiendo un solo caso en cada una de ellas:

Sin un valor 5	$\{6,6,6,7,8,9,10\}$	$Mdn = 7$
Sin un valor 6	$\{5,6,6,7,8,9,10\}$	$Mdn = 7$
Sin un valor 6	$\{5,6,6,7,8,9,10\}$	$Mdn = 7$
Sin un valor 6	$\{5,6,6,7,8,9,10\}$	$Mdn = 7$
Sin un valor 7	$\{5,6,6,6,8,9,10\}$	$Mdn = 6$
Sin un valor 8	$\{5,6,6,6,7,9,10\}$	$Mdn = 6$
Sin un valor 9	$\{5,6,6,6,7,8,10\}$	$Mdn = 6$
Sin un valor 10	$\{5,6,6,6,7,8,9\}$	$Mdn = 6$

Media de las medianas = 6.5

$$\hat{\sigma}_{(j)} = \{[(8-1)/8]\cdot(7-6.5)^2 + (7-6.5)^2 + \cdots + (6-6.5)^2\}^{1/2} = 1.32$$

6.5.2. BOOTSTRAP

Bradley Efron (1979), profesor de Estadística y Bioestadística en la Universidad de Stanford, propone un nuevo método sencillo de estimación. El método fue bautizado por Efron con el nombre de "bootstrap".[6] El nombre de "bootstrap" ha sido traducido por autodocimasia o docimasia por Enrique Cansado (citado en Diaconis y Efron, 1983), en la introducción del libro de Cramer "Métodos matemáticos en estadística" (Ed. Aguilar). Nosotros conservaremos el nombre original en inglés.

El bootstrap es en realidad una revisión y mejora del método Jackknife. Obsérvese que los dos nombres, en sentido figurado, reflejan el hecho de que a partir de una muestra, y sin más ayuda, se generan muchas más.

En realidad, tal como dice Hinkley (1986), con el nombre de métodos bootstrap, se conocen una variedad de técnicas basadas en la simulación y que se usan para unas tareas estadísticas particulares. Según Lunneborg (1987) podemos establecer la utilidad del bootstrap a tres niveles:

a) Valorar el sesgo y el error muestral de un estadístico o de una estimación de un parámetro poblacional calculado a partir de una muestra.

b) Establecer un intervalo de confianza para un parámetro estimado.

c) Realizar una prueba de hipótesis o significación en torno a uno o más parámetros poblacionales.

Para un análisis más exhaustivo, puede consultarse el trabajo de Efron y Tibshirani (1986), que hacen una excelente revisión de los métodos bootstrap.

6. Bootstrap significa "cordones para atar botas". Para el autor su sentido figurado significa que te ayudes a ti mismo y alude al viejo chiste de izarse uno mismo tirando hacia arriba de los cordones de las botas.

Supongamos que observamos

$$X_i = x_i \qquad [i = 1, 2, \ldots, n]$$

donde cada x_i son observaciones independientes e idénticamente distribuidos (i.i.d) de una variable de acuerdo con alguna distribución de probabilidad F de media μ y desviación típica σ. El problema principal es que F es desconocida, así como también lo son sus parámetros. En Estadística Inferencial Clásica y en el caso de la media,

$$\overline{x} = (1/n) \sum_{i=1}^{n} x_i \tag{6.35}$$

sabemos que la distribución muestral de medias es normal con media μ y $\sigma_{\overline{x}}$,

$$\text{siendo} \quad \sigma_{\overline{x}} = \sigma/(n)^{1/2} \tag{6.36}$$

dado que σ es desconocida, en la desviación tipica de la distribución sustituimos σ por un buen estimador:

$$\hat{\sigma}_{\overline{x}} = \hat{\sigma}/(n)^{1/2} \tag{6.37}$$

pero, esto no se puede hacer con otros estadisticos, dado que desconocemos su distribución muestral.

La idea intuitiva del bootstrap es sencilla tal como se detalla en los siguientes puntos, así como en la figura 6.4.

1) Se extrae de la población una sola muestra (generalmente pequeña) de tamaño n.

2) Se utiliza un generador de números al azar para extraer una nueva muestra del mismo tamaño "n" que la original de observaciones (i.i.d), con reemplazamiento a partir de la muestra original. De tal forma que cada nuevo valor sea una selección al azar de una de las observaciones originales. El nuevo conjunto de valores, llamado "muestra bootstrap", constituye una subserie de la muestra original. Algunas de estas observaciones originales pueden haber sido seleccionadas ninguna vez, otros una vez, otros dos veces, etc.

3) Se calcula el valor del estadístico $\hat{\theta}$ deseado para dicha nueva muestra generada de la original.

4) Se repiten los pasos 2 y 3 muchas veces (B veces). Es decir, generamos muchas muestras del tamaño de la original y a partir de esta.

5) Así obtenemos una nueva distribución muestral a partir del estadístico calculado en cada una de las muestras bootstrap. La desviación tipica de esta nueva distribución muestral de muchos valores del estadístico $\hat{\theta}_B$ se aproximaría adecuadamente al valor auténtico σ_θ del parámetro.

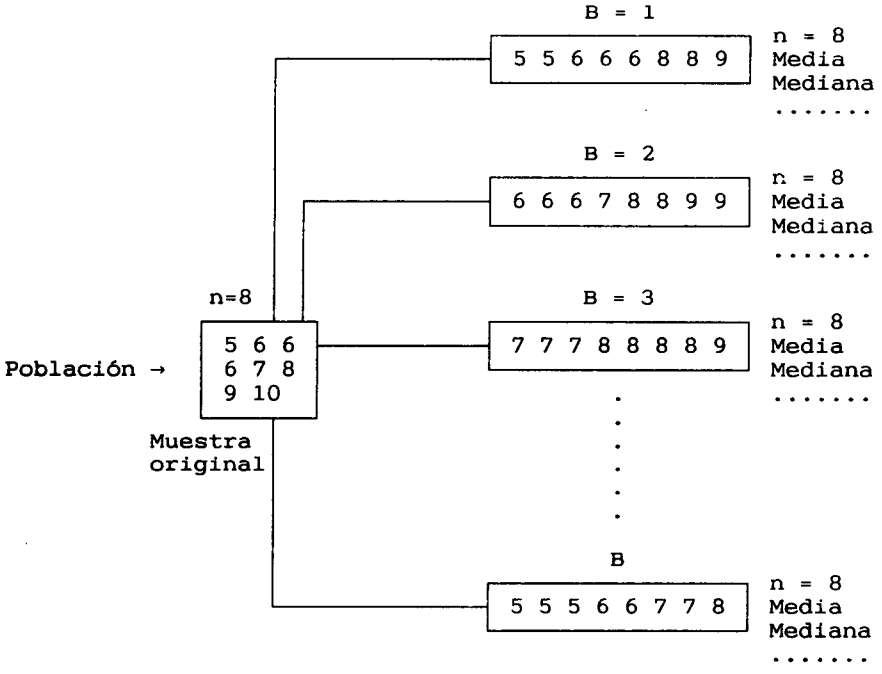


Fig.6.3. Representación esquemática de una estimación Bootstrap.

Al ser una técnica de carácter **no paramétrico** tiene la gran ventaja para el analista de poder liberarlo de la necesidad de realizar supuestos, a veces arriesgados, sobre la función de distribución de la variable en la población, así como en torno a la función de distribución muestral del

estadístico. Siendo asímismo aplicable a una gran cantidad de datos y diseños diferentes.

Para poder utilizar técnicas paramétricas hemos de conocer (al menos suponerla conocida) la función de distribución teórica. En caso de tomar un supuesto incorrecto respecto a la distribución de probabilidad, puede llevarnos a la obtención de estimaciones muy inexactas. Clásicamente, este problema se ha solventado mediante la utilización de técnicas no paramétricas, generalmente basadas en estadísticos de rango, con la sustancial pérdida de eficiencia que esto significa.

El "bootstrap", al ser una técnica no paramétrica, no precisa basarse en los supuestos anteriormente mencionados pero, contrariamente a las técnicas no paramétricas clásicas, no conlleva dicha pérdida de eficiencia al no requerir, entre otras cuestiones, un reescalamiento de la variable ordinal.

6.5.2.1. FORMULACIÓN DE LA TÉCNICA

El supuesto implícito en esta técnica, consiste en considerar que la función de distribución empírica de la muestra, que denominaremos F_n, nos proporciona un estimador máximo verosímil (MLE) no paramétrico de la función de distribución teórica.

Estableciéndose la igualdad $\hat{F}_\eta = F_n$; y siendo F_n la MLE no paramétrica de F_η y η el parámetro o vector de parámetros que cararterizan F, la verdadera función de distribución. Conociéndose a través del teorema de Glivenko-Cantelli, tal y como señala Borovkov (1984) la convergencia asintótica casi segura de la distribución F a medida que "n" tiende a infinito, tal como se recoge en la siguiente expresión:

$$\text{Si } n \to \infty \qquad \sup |F_n(x) - F(x)| \underset{\text{c.s.}}{\longrightarrow} 0 \qquad (6.38)$$

Obteniéndose, habitualmente, mediante remuestreo con reposición a partir de F_n, la función de distribución muestral "bootstrap" del estadístico (F_n^*). Una descripción formal de la técnica bootstrap, siguiendo a Efron (1979a), se puede establecer del siguiente modo.

Sea $\underset{\sim}{X}_n = (X_1, X_2, \ldots, X_n)$ una muestra de variables aleatorias independientes idénticamente distribuidas (i.i.d.) con una función de distribución F común y supongamos que $R_n(\underset{\sim}{X}_n; F)$ es alguna variable aleatoria especifica de interés (que en situación de una muestra puede tratarse de un parámetro θ), dependiente posiblemente de la función F desconocida. F_n denota la Función de Distribución Empírica de $\underset{\sim}{X}_n$ conseguida al dar masa n^{-1} a cada uno de los puntos $X_1, X_2, \ldots, X_n$.

La técnica Bootstrap consiste en aproximar la distribución muestral de $R_n(\underset{\sim}{X}_n; F)$ bajo F a través de la distribución Bootstrap de $R_n(\underset{\sim}{X}^*_n; F_n)$ bajo F_n, donde $\underset{\sim}{X}^*_n = (X^*_1, X^*_2, \ldots, X^*_n)$ denota una muestra aleatoria de tamaño "n" tomada de F_n.

Evidentemente la dificultad del procedimiento bootstrap reside en el cáculo de la distribución bootstrap. Efron (1979a), sugiere relizar la aproximación a través de la técnica de simulación Monte Carlo.

Se generan repetidas realizaciones de $\underset{\sim}{X}^*_n$ tomando nuestras aleatorias de tamaño "n" de F_n, es decir $\underset{\sim}{X}^*_n(1), \underset{\sim}{X}^*_n(2), \ldots\ldots, \underset{\sim}{X}^*_n(B)$, y el histograma de los valores ordenados correspondientes a $R_n(\underset{\sim}{X}^*_n(1); F_n), \ldots,$ $R_n(\underset{\sim}{X}^*_n(B); F_n)$ se toma como una aproximación de la distribución bootstrap de $R_n(\underset{\sim}{X}^*_n; F_n)$.

Nos referiremos, en esta introducción, únicamente a lo que Efron denomina bootstrap no paramétrico, o sea, remuestrear mediante simulación de Montecarlo, utilizando como estimación de F a F_n la función de distribución empírica. El mismo autor propone el denominado bootstrap paramétrico para aquellos casos en los que conocemos la verdadera función de distribución y por tanto no es necesario estimarla, simplemente se realizaría la estimación de los parámetros de la muestra obtenida y se remuestrearía sobre la función F conocida con los parámetros estimados. Es decir, en lugar de $\hat{F}_\eta = F_n$, nos centramos en $F_{\hat{\eta}}$.

Por otra parte, y en el caso en que estimamos F a partir de F_n, si podemos suponer que F es contínua, el propio Efron (1979b) y posteriormente Silverman y Young (1987), proponen el uso de una versión alisada de F_n, combinándo linealmente una función suave con la estimación de la función de distribución empírica, esto es lo que se ha venido a denominar **"bootstrap suavizado"**.

Uno de los aspectos que anteriormente se ha mencionado como utilidad del bootstrap por parte de Lunneborg (1987), reside en el establecimiento del error estándar. La estimación bootstrap del error estándar en situaciones simples de una muestra se puede esquematizar de la siguiente forma.

Supongamos que $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ es el verdadero valor de un estadístico, y sea $\sigma(F)$ el error estándar de $\hat{\theta}$, $\sigma(F) = [\text{VAR}_F\hat{\theta}]^{1/2}$, por supuesto $\sigma(F)$ también será función del tamaño de muestra "n" y de la forma del estadístico $\hat{\theta}$, que por conocido puede obviarse su indicación en la notación. La estimación bootstrap del error estándar sería:

$$\hat{\sigma} = \sigma(F_n) = [\text{VAR}_*\hat{\theta}^*]^{1/2} \tag{6.39}$$

donde F_n será la función de distribución empírica, dando masa n^{-1} a cada dato observado $X_i = x_i$, y $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)$ con $X_1^*, X_2^*, \ldots, X_n^*$ (i.i.d.) realizaciones de F_n. VAR_* denota la variancia bajo la ley condicional de $\underset{\sim}{X}_n^*$ dado $\underset{\sim}{X}$.

Como ejemplo clásico, que procionan la mayoría de autores, Efron (1982); Efron y Tibshirani (1986); Swanepoel (1989), supongamos que $\hat{\theta} = \overline{X}_n$, en cuyo caso se obtiene $\sigma(F) = [\mu_2(F)/n]^{1/2}$, donde $\mu_2(F) = \int_{-\infty}^{\infty}(x - E_F(X))^2\, dF(x)$, y obsérvese que $\hat{\sigma} = [S_n^2/n]^{1/2}$, expresión habitual utilizada para estimar la dispersión de la distribución muestral de medias, donde Sn^2 es el estimador no sesgado de $\mu^2(F)$.

En la mayoría de los casos, como ya hemos dicho, no existe una expresión simple para la función $\sigma(F)$ de estimadores como la mediana, las medias recortadas, el coeficiente de correlación o la pendiente de la regresión robusta, que como en el caso de la media, nos establezca una función simple de la distribución muestral. Sin embargo, es fácil evaluar $\hat{\sigma} = [\text{VAR}_*\hat{\theta}^*]^{1/2}$ a través del algoritmo de Monte Carlo que se concreta en los siguientes pasos:

a) Construcción de F_n.

b) Construcción de una muestra aleatoria $X_1^*, X_2^*, \ldots, X_n^*$ a partir de F_n y calcular $\hat{\theta}^* = \hat{\theta}(X_1^*, X_2^*, \ldots, X_n^*)$.

c) Repetir de forma independiente el paso número 2 "B" veces obteniendo replicaciones Bootstrap $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(B)$, y calculando:

$$\hat{\theta}^*(.) = B^{-1} \sum_{b=1}^{B} \hat{\theta}^*(b) \tag{6.40}$$

$$\hat{\sigma}_B = \left\{ (B-1)^{-1} \sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(.))^2 \right\}^{1/2} \tag{6.41}$$

Es fácil ver que en la medida que $B \to \infty, \hat{\sigma}_B \to \hat{\sigma} = \sigma(F_n)$, la estimación bootstrap del error estándar, es mejor conforme aumenta el número de réplicas bootstrap. Véase tabla 6.4. Según señalan algunos autores (por ejemplo Swanepoel, 1989), en la mayoría de los casos el que B oscile entre 50 y 200 es adecuado para la estimación de un error estándar; fijándolo Efron (1984), en la cifra intermedia de 100, aunque, en algunos estudios realizados, señala que se consiguen resultados razonables con solamente 25 replicaciones.

Tabla 6.4. Estimación del error estandar[7] de la media en una muestra $n = 16$.

5,5 8 4,5 5,5 4,5 3,25 7 7 0,25 6 1,25 5 4,5 4,25 5 9,25

	error estandar	$\sqrt{\epsilon(\hat{\sigma} - \sigma)^2}$ $M.S.E.$	Valor medio
Bootstrap $B = 100$	0,55	0,46	4,65
" $B = 200$	0,52	0,38	4,72
" $B = 300$	0,49	0,34	4,73
" $B = 400$	0,48	0,33	4,72
Jacknife	0,57		5,04
Error estandard de la muestra original $\frac{2,27}{\sqrt{16}} =$	0,568		5,04

7. Calculado con el programa elaborado por Alex Sánchez (Facultad de Biología). U.B.

Evidentemente, utilizando el error estándar bootstrap podemos construir intervalos de confianza para la estimación del parámetro poblacional a la manera clásica la cual, probablemente, mejorará la estimación realizada al emplear el error estándar habitual.

$$\hat{\theta}^*(.) \pm Z_{(\alpha/2)} \cdot \hat{\sigma}_B \tag{6.42}$$

Sin embargo, Efron (1980) propone también la construcción de diferentes intervalos de confianza no paramétricos, como el método del percentil, el método para la corrección del sesgo y el método acelerado para la corrección del sesgo. Asimismo, otros autores como Hall (1986) y Beran (1987) han propuesto, respectivamente, los métodos del percentil "t" y el método del bootstrap anidado. A excepción del método percentil, los restantes pretenden captar el sesgo del estimador. Los tres últimos incorporan la simetría de la función de distribución. En general todos ellos, suponen la construcción de un intervalo con una precisión de mayor orden, puesto que los dos primeros son asintóticamente correctos de primer orden, los dos segundos lo son de segundo orden y el último lo es de tercer orden (Huet y Jolivet 1989a; 1989b) (Huet, Jolivet y Messean, 1989) (Hall, 1988a; 1988b). Dado el carácter introductorio de este apartado no los expondremos, quedándonos simplemente en el primer nivel de los propuestos por Lunneborg (1987), remitiendo al lector a las fuentes originales citadas en la bibliografía.

GLOSARIO

ADYACENTE: denominación empleada por Tukey (1977) para aquellos valores que limitan las patillas o "whiskers" en un Diagrama de Caja.

AGUJERO ("gap"): Ausencia de valores en intervalos determinados de una distribución.

AJUSTE DE MEDIANAS ("Median Polish"): Proceso de búsqueda de patrones o relaciones entre variables a partir de los residuales. Estos se encuentran sustrayendo iterativamente las medianas, alternativamente de filas y columnas, si la tabla es de doble entrada.

ALISADOR ("Smoother"):

- **COMPUESTO**: Alisador en que se combina la acción secuencial de diferentes amplitudes para conseguir la secuencia alisada final.
- **IMPAR**: Mediana móvil en la que la amplitud de la subsecuencia utilizada es impar.
- **PAR**: Mediana móvil en la que la amplitud de la subsecuencia utilizada es par.
- **CORTADO**: Proceso que sigue generalmente a un alisador de amplitud 3 ó 5 para evitar la desaparición de picos o valles presentes en la serie original.
- **HANNING**: Promedio móvil ponderado de amplitud 3.
- **IANNING**: Promedio móvil ponderado de amplitud 5.
- **JANNING**: Promedio móvil ponderado de amplitud 7.
- **KANNING**: Promedio móvil ponderado de amplitud 9.

AMPLITUD INTER-CUARTÍLICA (IQR) ("interquartilic range"): dispersión media ("midspread"), diferencia de cuartiles (dq): medida de dispersión que proporciona la distancia entre el primer y tercer cuartiles. Puede aproximarse al valor que se obtendría en el cálculo de la S de una distribución que siguiera el modelo de Laplace-Gauss mediante estandarización (ver pseudo-iqr).

ANOMALÍA, VALOR ANÓMALO ("outlier", "far outlier"): (exteriores, remotos) valor muy alejado del conjunto central de los datos, que afecta especialmente los índices de tendencia central y dispersión clásicos.

Existen diversas categorizaciones según autores en función de su grado de alejamiento o afectación.

BOOTSTRAP: Método de estimación no paramétrico, ideado por B. Efron, basado en la reconstrucción de la distribución muestral, generalmente, mediante la simulación de Monte Carlo.

CENTIL: valor que divide a la distribución en una serie de puntos equidistantes. Se acostumbran a utilizar los percentiles (100), deciles (10) y cuartiles (4). Puede asimismo aproximarse esta subdivisión mediante el cálculo de cuantilas (Batista & Valls, 1989a; 1989b) o mediante sucesivas divisiones en 2 mitades de la distribución original.

CENTILES, GRÁFICO DE: función de distribución empírica de los datos originales, situando éstos en el eje de abcisas y sus correspondientes percentiles en el de ordenadas. Es de gran utilidad para efectuar aproximaciones visuales a las agrupaciones o disgregaciones en los datos, así como a su simetría.

CENTRIMEDIA o MEDIA INTERCUARTÍLICA ("Midmean", "interquartile mean"): índice de localización resultante del promedio de todos los valores entre el primer y tercer cuartil. En su cálculo acostumbran a eliminarse los valores repetidos, procurando que resten el mismo número de datos a un lado y otro de la Md.

COEFICIENTE DE CURTOSIS (K_2): coeficiente centrado sobre el valor 1 (indicador de distribución mesocúrtica) que pondera el primer y noveno decil con el IQR.

COEFICIENTES DE LA LÍNEA RESISTENTE: Valores estimados a partir de los tercios de la nube de puntos inicial, correspondientes a la constante y a la pendiente de una ecuación lineal.

COEFICIENTE DE VARIACIÓN CUARTÍLICO (CVc): índice de dispersión relativa, más robusto que el Coeficiente de Variación clásico, que posibilita asimismo la comparación del cociente dispersión-centro de una distribución, independientemente de sus unidades de medida.

CUARTO, BISAGRA ("hinge", "fourth"): subdivisión de las dos mitades de una distribución, definida a partir de la Md, a partir de la que puede calcularse el fourth-spread, para delimitar los valores anómalos en un diagrama de Caja.

DIAGNÓSTICO, GRÁFICO DE: Gráfico para detectar el componente no aditivo. Se realiza a partir de los residuos y de los valores comparativos después de un ajuste de medianas.

DIAGRAMA DE CAJA ("Box Plot", "box-and-dot plot", "box and whiskers plot"): Gráfico que permite evaluar el rango, simetría central y extrema así como las diferentes gradaciones de valores anómalos en una distribución de datos.

DIAGRAMA DE RAIZ CUADRADA ("Rootogram"): Gráfico similar al Histograma de frecuencias, donde se sustituyen las frecuencias por la raiz cuadrada de la densidad del intervalo.

DIAGRAMA SUSPENDIDO DE RAIZ CUADRADA ("Suspended Rootgram") : Figura en la que se superponen el modelo teórico de la distribución de frecuencias, con el histograma realmente observado, de forma que se pueden observar los residuales, o desajustes, que se producen entre las dos representaciones.

ESTANDARIZACIÓN DE UNA VARIABLE: Transformación lineal que se realiza sobre los valores de una variable, con objeto de centrarla y reducirla.

ESTIMADOR ROBUSTO: Estimador que presenta la característica de robustez (Ver robustez; L y M estimadores).

FUNCIÓN OBJETIVA: Función que minimiza la desviación de cada valor (x_i) con respecto al estimador "t".

H-PSEUDOSIGMA: índice de dispersión correspondiente al valor aproximado a la S muestral de una distribución ajustada a la curva normal, que incrementa la resistencia de ésta.

ÍNDICE DE SIMETRÍA DE KELLY (H_2): indica la simetría de la distribución en sus extremos o colas respecto de la Md, por lo que es muy útil en combinación con el $H1$. Al depender de las unidades de medida es aconsejable transformarlo en $H3$.

ÍNDICE DE SIMETRÍA DE YULE (H_1): indica la simetría del 50% central de la distribución. Es aconsejable interpretarlo en combinación con el índice de simetría de Kelly.

JACKNIFE: Método de estimación no paramétrico del error muestral basado en el remuestreo de los datos originales.

L-ESTIMADORES: Estimadores robustos que se obtienen mediante combinaciones lineales de estadísticos de orden.

LOTE ("batch"): conjunto de datos a analizar, que comparten una característica común. Datos generados por un mismo proceso. Equivalente a "muestra" en Estadística clásica.

MEDIANA MÓVIL: Técnica de suavizado consistente en substituir cada valor de la serie por la mediana del conjunto de datos compuestos por el propio valor y los que se hallan en su entorno (dependiendo el número de éstos de la amplitud del alisador escogido).

MEDIANA DE LAS DESVIACIONES ABSOLUTAS (MAD): índice de dispersión que indica la Md de las diferencias, en valor absoluto, respecto de la Md de la distribución original. Puede efectuarse una estandarización del índice, que aproximará su valor al de la S de una distribución que siga el modelo de Laplace-Gauss.

MEDIANA GENERALIZADA: Mediana que se obtiene con cualquiera de los tres o cuatro estadísticos de orden central.

M-ESTIMADORES: Estimadores robustos que se obtienen minimizando la función objetiva. Los más usuales son los de Tukey, Hempel, Huber y Andrews.

MEDIAS RECORTADAS ("Trimmed Mean"): Media obtenida después de haber eliminado a ambos lados de la distribución una proporción α de valores.

OCTAVOS ("eights"): valores que dividen en dos partes iguales a los cuartos, equidistantes entre Md y el límite superior o inferior de la muestra.

PROFUNDIDAD ("Depth"): distancia de un dato, equivalente a su valor ordinal, eligiendo el menor entre los que le correspondan en el orden establecido desde cada uno de los extremos de la muestra.

PROMEDIO DE CUARTILES (Q): índice de localización consistente en la suma promediada del primer y tercer cuartiles.

PROMEDIO MÓVIL PONDERADO: Promedio de "k" valores adyacentes en que no todos ellos tienen el mismo peso.

PSEUDO-AMPLITUD INTER-CUARTÍLICA (PSD_{IQR}) ("interquartilic range"): estandarización del IQR correspondiente al 50% central de los datos, comparable con la S que se obtendría si éstos siguieran una distribución normal, incrementando la resistencia del índice.

PSEUDO-MEDIANA DE LAS DESVIACIONES ABSOLUTAS (PSD_{MAD}): estandarización del IQR correspondiente al 50% central de los datos, comparable con la S que se obtendría si éstos siguieran una distribución normal, incrementando la resistencia del índice. TRIMEDIA; índice de localización que define la distancia media entre la Md y Q, eliminando un 25% de las observaciones de cada extremo. Puede asimismo definirse otro límite (5% bilateral, por ejemplo, etc.) que proporcione información comparable con otros estadísticos.

PUNTO DE COLAPSO ("Breakdown point"): Se dice que un estimador alcanza su punto de colapso cuando limita la proporción máxima α de observaciones anómalas a ambos lados de la distribución que pueden producirse sin que cambie el resultado del estimador.

REAPROXIMANDO ("Reroughing"): Proceso que pretende recuperar para la serie suavizada algún patrón que un alisado demasiado fuerte haya trasladado a la serie de residuales.

RECENTRADO: Proceso que sigue a un alisador de amplitud par para que los valores alisados se vuelvan a alinear según los valores marcados por la variable que ordena la secuencia original.

REEXPRESION: Veáse Transformación.

REGLA DE VALORES EXTREMOS: Regla utilizada para conseguir suavizar los valores extremos de cada serie temporal.

RESIDUAL: Diferencia que hay entre los datos reales y el resultado de su ajuste a un modelo previamente determinado o subyacente.

RESISTENCIA: característica que indica el grado de sensibilidad a los valores anómalos (alejados del intervalo central de la distribución).

ROBUSTEZ: grado de sensibilidad ante las desviaciones de las condiciones de ajuste o aplicación de modelos probabilísticos.

RESIDUALES DE DOBLE RAIZ ("Double-Root Residuals") : Estrategia gráfica que nos proporciona, a partir del estudio de los residuales, una prueba de bondad de ajuste, entre una distribución de frecuencias observada y una teórica.

SEMIPENDIENTE: Coeficiente para la evaluación de la linealidad de la nube de puntos inicial.

SERIE TEMPORAL: Conjunto secuencial de datos de una variable ordenados en función de otra variable, generalmente de carácter temporal.

TERCER ÍNDICE DE SIMETRÍA (H_3): transformación del índice de simetría de Kelly para hacerlo interpretable de forma independiente de las unidades de medida, al igual que el $H1$.

TRANSFORMACIÓN: Sustitución de una serie de datos por una nueva serie de valores, función de los originales.

- **DE POTENCIA** ("Power Trans".) : funciones contínuas aplicadas sobre el conjunto original, y que mantienen el orden presentado originalmente por los datos.
- **LINEALES**: subconjunto de las transformaciones de potencia, que implica una traslación del origen y un cambio en la escala de los datos originales.
- **NO LINEALES**: transformaciones de potencia que deforman la forma original de la distribución de la variable.
- **MONOTONAS NO LINEALES**: subconjunto de las anteriores, que producen ratios de crecimiento no constantes entre los diferentes valores, pero sin alterar la ordenación original de estos.
- **ESCALA DE TUKEY** ("Ladder of Power Trans.") : escala de transformaciones de potencia mas usuales.
- **PARA PROMOVER SIMETRÍA**: estrategia que ayudará a la elección de al mejor transformación que simetrice la distribución.
- **PARA CONSEGUIR DISPERSIÓN ESTABLE**: estrategia gráfica que permitirá la elección de la mejor transformación que homogeneice las dispersiones de varios subconjuntos de datos.

- **COMPARADAS** ("Matched") : transformación lineal que se realiza sobre los datos transformados, para poder comparar el efecto de la transformación no lineal sobre los datos directos.

TRIMEDIA: Índice de localización que define la distancia media entre la Mediana y el Promedio de Cuartiles, eliminando un 25% de las observaciones de cada extremo.

TRONCO-Y-HOJA, GRÁFICO DE: Diagrama de la distribución, a medio camino entre un histograma clásico y una distribución de frecuencias, en el que los datos son representados por sus mismo dígitos, subdivididos en una parte definidora genérica (tronco) y su precisión dentro de los anteriores intervalos (hojas). Su gran versatilidad le confiere una gran utilidad en una primera aproximación a la descripción de los datos en todas sus facetas de localización, tendencia central y forma.

VALOR COMPARATIVO: Se obtiene multiplicando el efecto columna por el efecto fila y dividido por el efecto común. (Veáse Ajuste de Medianas).

VALORES-LETRA, GRÁFICO DE ("Letter-value display"): conjunto de observaciones extraídas sistemáticamente de la muestra, que definen una serie de límites internos, externos y extremos, a partir de los cuáles pueden categorizarse los datos en función de su distancia respecto de la Md e identificar valores anómalos.

BIBLIOGRAFÍA

Andrews, D.F. (1.971). "A note on the selection of data transformations". *Biometrika*, 58 (2), 249–254.

Andrews, D.F.; Binckel, P.J.; Hampel, F.R.; Huber, P.J.; Rogers,W.H. y Tukey, J.W. (1972). *Robust Estimates of Location*. Princeton, New Jersey: Princeton University Press.

Arnau, J. (1986). *Diseños experimentales en Psicología y Educación*, Vol. I. México: Trillas.

Bartlett, M.S. (1949). "Fitting a straight line when both variables are subject to error". *Biometrics*, 5, 207–212.

Batista, J. M. & Valls, M. (1985a). "Nuevas técnicas de análisis estadístico de datos: Tabulación y síntesis numérica Análisis Exploratorio de Datos)". *Qüestiió*, 9, 2, 105–119.

Batista, J. M. & Valls, M. (1985b). "Técnicas gráficas en Análisis Exploratorio de Datos". *Qüestiió*, 9, 3, 163–176.

Behrens, J.T.; Stock, W.A. & Sedgwick, C. (1990). "Judgment errors in elementary Box-plot displays". *Commun. Statis. Simula.*, 19(1), 245–262.

Beran, R. (1987). "Prepivoting to reduce level error of confidence sets". *Biometrika*, 74, 457–468.

Berk, R.A. (1984). "A premier on Robust Regression". En J. Fox y J. Scott Long (Eds.). *Modern Methods of data analysis (pp. 292–324)*. Newssbury Park: Sage Pub.

Berry, W.D. (1984). *Non-recursive causal models*. Beverly Hills, California: Sage.

Berry, W.D. y Lewis-Beck, M.S. (1986). *New Tools for Social Scientist. Advances and Applications in Research Methods*. Beverly Hills, California: Sage.

Borovkov, A.A. (1984). *Estadística Matemática*. Moscú: MIR.

Box, G.E.P. (1.987). "Signal to noise ratios, performance criteria and transformations". *Technical Report 26*. Center for quality and productivity improvement. University of Wisconsin-Madison.

Box, G.E.P. y Cox, D.R. (1.964). "An Analysis of Transformations". *J. of Royal Statistic Soc. Ser. B, 26*, 211–243.

Box, G.E.P. y Tidwell, P.W. (1962). "Transformations of the independent variables". *Technometrics, 4*, 531–550.

Brown, G.W. y Mood, A.M. (1951). "On median test for linear hypotheses". En J. Neyman (Ed.), *Proceedings of the Second Berkeley Symposium on*

Mathematical Statistics and Probability. Berkeley y Los Angeles: University of California Press.

Chambers, J. M.; Cleveland, W. S.; Kleiner, B. & Tukey, P. A. (1983). *Graphical methods for data analysis*. Boston: Duxbury Press.

Cliff, N. (1983). "Some cautions concerning the applications of causal modelling methods". *Multivariate Behavioral Research, 18*, 115–126.

De la Rosa, A.; Sánchez, M.; Freixa, M. y Sierra, V. (1990). "Estudio de la tendencia de consumo de alcohol en función del sexo en adolescentes de la provincia de Barcelona". *Comunicación presentada al VII Congreso Nacional de Psicologia*, Barcelona (Noviembre).

Diaconis, P. y Efron, B. (1983). "Métodos estadísticos intensivos por ordenados". *Revista de Investigación y Ciencia* (Julio), 70–128.

Dixon, W.J. & Kronmal, R. A. (1965)."The choice of origin and scale for graphs". *J. Association for computing machinery, 12*, 259–261. (Citado por Hoaglin & cols. 1983).

Domenech, J.M. y Riba, M.D. (1985). *Métodos estadísticos. Modelo lineal de la regresión*. Barcelona: Herder.

Efron, B. (1979a). "Bootstrap Methods: Another look at the Jacknife". *The Annals of Statistics, 7* (1),1-26.

Efron, B. (1979b). "Computers and the theory of statistics: Thinking the unthinkable". *SIAM Review, 21* (4),460-319.

Efron, B. (1982). *The Jackknife ,the Boostrap and the other Resampling Plans*. Bristol: CBMS-NSE.

Efron, B. (1984). *Better Bootstrap Confidence Intervals*. LCS Thecnical Report 14. Dept of Statistics. Stanford. California.

Efron, B. y Tibshirani, R. (1986). "Bootstrap Methods for Standard Errors, Confidence Intervals, and other Measures of Statistical Accuracy". *Statistical Sciencie, 1* (1).54-77.

Emerson, J.D. (1.985). "Mathematical Aspects of Transformation". En D.C. Hoaglin, F. Mosteller y J.W. Tukey (Eds.). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley and Sons. Cap. 8, 247-281.

Emerson, J.D. y Hoaglin, D.C. (1983). "Analysis of two-way tables by medians". En E. Hoaglin; D.C. Mosteller; J.W. Tukey, (Eds.). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Emerson, J.D. y Hoaglin, D.C. (1985). "Resistant Multiple Regression, one variable at a time". En D.C. Hoaglin; F. Mosteller y J.W. Tukey (Eds.), *Exploring Data Tables, Trends, and Shapes*, pp. 241-280. New York: John Wiley & Sons.

Emerson, J.D. y Stoto, M.A. (1982). "Exploratory Methods for choosing power transformations to normality". *Journal of the Royal Statistical Society Association, 77*, 472-476.

Emerson, J.D. y Stoto, M.A. (1.983). "Transforming Data". En D.C. Hoaglin, F. Mosteller y J.W. Tukey (Eds.). *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley and Sons. Cap. 4, 97-127.

Emerson, J.D. y Wong, G (1985). "Resistant non additive fits for two-way tables". En E. Hoaglin; D.C. Mosteller; J.W. Tukey, (Eds.) *Exploring Data Tables, Trends and Shapes*, (pág. 67-119). New York: Wiley.

Erickson, B. & Nosanchuk, T. (1979). *Understanding Data*. Open Univ. Press: Milton Keynes.

Everitt, B. S. & Dunn, G. (1983). *Advanced methods of data exploration and modelling*. London: H.E.B.

Ferrer, R.; Freixa, M.; Guardia, J.; Salafranca, Ll.; Solanas, A. y Turbany, J. (1991). *Análisis de Datos en Ciencias del Comportamiento. Introducción al paquete estadístico SPSS/PC+*. Barcelona: Signo.

Fisher, R.A. (1935). *The design of experiments*. Londres: Oliver & Boyd.

Freixa, M. y Salafranca, Ll. (1989). *Análisis Exploratorio de Datos. Documento de circulación interna*. Dept. de Metodología de las Ciencias del Comportamiento. Universidad de Barcelona.

Fox, J. (1.990). "Describing univariate distributions". En J. Fox y J. Scott Long (Eds.). *Modern Methods of Data Analysis*, (pág. 58-176). California: Sage.

Fox, J. y Scott Long, J. (Eds.) (1990). *Modern Methods of Data Analysis*. California: Sage.

Gentle, J.E. (1977). "Least absolute values estimation: an introduction". *Communications on Statistics, B/*, 313-328.

Good, J. (1983). "The philosophy of Exploratory Data Analysis". *Philosophy of Science, 5* (2), 283-295.

Goodall, C. (1982). "M-Estimators of Location:an outline of the theory". En D.C. Hoaglin; F. Mosteller y J.W. Tukey,J.W. (Eds.) (pp. 339-400). *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

Goodall, C. (1982). "Examining Residuals". En E. Hoaglin; F. Mosteller y J.W. Tukey,J.W. *Understanding Robust and Exploratory Data Analysis* (pp. 211-243). New York: Wiley.

Goodall, C. (1990). "A survey of smoothing techniques". En J. Fox y J. Scott Long (Eds). *Modern Methods of Data Analysis* (pp. 126-176). California: Sage.

Goodfrey, K. (1989). "Fitting by organized comparisons: the square combining table". En E. Hoaglin; F. Mosteller y J.W. Tukey (Eds.) (pp. 37-62). *Exploring Data Tables, Trends and Shapes.* New York: Wiley.

Gross, A.M. (1976). "Confidence interval robustness with long-tailed symmetric distributions". *Journal of the American Statistical Association, 71*, 409-416.

Guàrdia, J. y Arnau, J. (1991). "Análisis evaluativo de las características teórico-empíricas de los sistemas de ecuaciones estructurales". *Anuario de Psicología, 48* (1), 5-16.

Gujarati, D.N. (1987). *Basic Econometrics.* New York: McGraw-Hill Int. Ed.

Hampel, F.R. (1968). *Contributions to the theory of robust estimation.* Tesis doctoral no publicada. University of Berkley, California.

Hampel, F.R. (1971). "A general qualitative definition of robustness". *Annals of Mathematical Statistics, 42*, 1887-1896.

Hall, P. (1986). "On the Bootstrap and Confidence intervals". *Annals of Statistics, 14*, 1431-1452.

Hall, P. (1988a). "On Symetric confidence intervals". *Journal of the Royal Statistical Society (B), 50* (1), 35-45.

Hall, P. (1988b). "Theoretical comparison of Bootstrap Confidence intervals". *The Annals of Statistics, 16* (3), 927-953.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Hinkley, D.V. (1975). "On power transformations to simetry". *Biometrika, 62* (1), 101-111.

Hinkley,D.U. (1986). "Bootstrap Methods: efficiency and validity". Comunicación presentada en el *2 Simposium Internacional Catálan de Estadística.* 18-19 Septiembre,Barcelona.

Hoaglin, D. Iglewicz, B. y Tukey, J.W. (1981). "Small-sample performance of a resistant rule for oulier detection". *1980 Proceedings of the Statistical Computing Section.* Washington DC: American Statistical Assoc. págs. 148-152.

Hoaglin, D. Mosteller, F. y Tukey, J.W. (Eds.) (1983) *Understanding Robust and Exploratory Data Analysis*. New York: John Wiley & Sons.

Hoaglin, D.; Mosteller, F. y Tukey, J.W. (1982). "Introduccion to more refined estimators". En D. Hoaglin; F. Mosteller y J.W. Tukey (Eds.). *Understanding Robust and Exploratory Data Analysis* (pp. 283-294). New York: John Wiley & Sons.

Hoaglin, D.; Mosteller, F. y Tukey, J.W. (1991). *Fundamentals of Exploratory Analysis of variance*. New York: J. Wiley.

Hogg, R.V. (1979). "An introduction to robust estimation". En R.L. Launer y G.N. Wilkinson (Eds.), *Robustness in Statistics*, (pp. 1-18). New York: Academic.

Honrubia, M.L.; Freixa, M.; Guillen, F.; Aznar, J.A. y Salafranca, Ll. (1990). "Estilo cognitivo DIC en juicios igual-diferente: Aplicación del análisis exploratorio de datos". *Comunicación presentada al VIII Congreso Nacional de Psicología*. Barcelona (Noviembre).

Horber, E. (1989). *EDA user's Manual*. Dept. des Science Politique. Univbersité de Géneve.

Horber, E. (1990). *Conceptes de base en Analyse Exploratoire des Donées*. Documento interno de la "IV Ecole de été en EDA", Carcasonne, Setembre 1991.

Horber, E. (1991). *Manual del paquete estadístico EDA*. Faculté des Sciences Politiques. Ginebra.

Hartwig, F. y Dearing, B. R. (1979). *Exploratory Data Analysis*. London: Sage.

Huet, S. y Jolivet, E. (1989a). "Exactitude au second ordre des intervalles de confiance bootstrap pour les paramètres d'un modele de régression non linéaire". *C.R. ACAD. SCI. Paris, t308*, 429-432.

Huet, S. y Jolivet, E. (1989b). "Bootstrap and Edgeworth expansion: the nonlinear regression as an example". *Technical report INRA, Dept. de Biometrie*.

Huet, S.; Jolivet, E. y Messean, A. (1989). "Some simulations results about confidence intervals and bootstrap methods in non linear regression". *Technical Report INRA, Dept de Biometrie*.

Iglewicz, B. (1982). "Robust scale estimators and confidence intervals for location". En D. Hoaglin; F. Mosteller y J.W. Tukey (Eds.). *Understanding Robust and Exploratory Data Analysis* (pp. 404-433). New York: Wiley.

Johnstone, I y Velleman, P.F. (1982). "Tukey's resistant line and related methods: asymptotics and algorithms". 1981 *Proceedings of the Statistical Computing Section.* Washington D.C.: American Statistical Association, pp. 218-223.

Leinhardt, S. y Wasserman, S.S. (1979). "Exploratory Data Analysis: an introduction to selected methods". En K.F. Schuessler (Ed.), *Sociological Methodology 1979.* San Francisco: Jossey-Bass.

Levey, A.B. (1.980). "Measurement units in psichophysiology". En I. Martin y P.H. Venables (Eds.). *Techniques in Psychopsysiology.* Chichester: John Wiley and Sons. Cap 12, 597-628.

Li, G. (1985). "Robust Regression". En E. Hoaglin; F. Mosteller y J.W. Tukey (Eds.), (pp. 281-341). *Exploring Data Tables, Trends and Shapes.* New York: Wiley.

Lunneborg,C.E. (1987). "Bootstrap Application for the Behavioral Sciencies". *Education and Psychological Measurement, 47,* 627.

Mayers, J.L. (1982). *Fundamentals of Experimental Design.* Boston: Allyn and Bacon Inc.

Maritz, J.S. (1981). *Distribution-free statistical methods.* London: Chapman and Hall.

Marsh, C. (1988). *Exploring Data. An introduction to Data Analysis for Social Scientist.* Cambridge: Polity Press.

Miller, R.G. (1974). "The Jackknife-A review". *Biometrika 61* (1), 1-15.

Mosteller, F.; Siegel, A.F.; Trapido, E. y Youtz, C. (1985). "Fitting straight lines by eye". En D.C. Hoaglin; F. Mosteller y J.W. Tukey (Eds.). *Exploring Data Tables, Trends, and Shapes,* pp. 225-240. New York: John Wiley & Sons.

Mosteller, F. y Tukey, J.W. (1977). *Data Analysis and Regression.* New York: Addison.

Nair, K.R. y Shrivastava, M.P. (1942). "On a simple method of curve fitting". *Sankhya, 6,* 121-132.

Namboodiri, N.K. (1972). "Experimental designs in which each subject is used repeatedly". *Psychological Bulletin, 77,* 54-64.

Ocaña, J.; Sánchez, A. y Solanas, A. (1991). *Seminari d'introducció als mètodes bootstrap.* Documento de circulación interna. Barcelona: Universidad Autónoma de Barcelona.

Pressey, A.W. & Smith, N.E. (1986). "The Effects of Location, orientation, and acumulation of boxes in the Baldwin illusion". *Percept. & Psychophysics, 40*, 344-350.

Qennnouille, M.H. (1949). "Aproximate test of correlation in time series". *Journal of the Royal Statistical Society Series B 11*, 68-83.

Rappachi, B. (1991). *Une introduction a la notion fde robustesse.* Documentación de la Universidad de Verano de EDA. Septiembre, Carcassonne.

Romanowiccz, C.N. (1989). "Three-Way Analysis". En E. Hoaglin; F. Mosteller y J.W. Tukey,J.W. (Eds.), *Exploring Data Tables, Trends and Shapes*, (pág. 125-185). New York: Wiley.

Rosember,J.L. y Gasko, M. (1982). "Comparing location estimators: Trimmed Means, medians and trimean". En E. Hoaglin; F. Mosteller y J.W. Tukey, (pp. 297-336). *Understanding Robust and Exploratory Data Analysis.* New York: Wiley.

Salafranca, Ll.; Freixa, M. y Guàrdia, J. (1990). "Aplicación del análisis exploratorio de datos en los sistemas de ecuaciones estructurales". *Comunicación presentada al VIII Congreso Nacional de Psicología.* Barcelona (Noviembre).

Salafranca, Ll. (1991). *Neurociencia Cognitiva: Problemática del análisis de datos.* Tesis Doctoral no publicada.Universidad de Barcelona.

Salafranca, Ll., Turbany, J. y Solanas, A. (1989). *Intervalos de Confianza bootstrap: un estudio comparativo de la precisión.* Comunicación presentada al II Congreso Español de Biometria. Segovia.

Sampson, P.D. y Guttorp, P. (1.991). "Power Transformations and Tests of Environmental Impact as Interaction Effects". *The American Statistician, 45* (2), 83-89.

Sánchez, A. (1990). "Efficient bootstrap simulation: an overview". *Qüestiió, 14*, 43-88.

Sarriá, A.; Guàrdia, J. y Freixa, M. (1986). *Introducción de la estadistica en Psicología.* Barcelona: Alamex.

Schmid, C.P. (1983). *Statistical graphics, design principles and graphics.* New York: John Wiley & Sons.

Schwartz, D. (1985). *Métodos estadísticos para médicos y biólogos.* Barcelona: Herder.

Searle, S.R. (1971). *Linear Models.* New York: John Wiley & Sons.

Sen, P.K. (1968). "Estimates of the regression coefficient based on Kendall's tau". *Journal of the American Statistical Association, 63*, 1379-1389.

Siegel, A.F. (1982). "Robust regression using repeated medians". *Biometrika, 69*, 242-244.

Silverman y Young (1987). "The bootstrap: to smooth or not to smooth?". *Biometrika, 74* (3), 469-479.

SPSS/PC (1988). *Getting started with SPSS/PC+ Trends V.2.0 for the IBM.*

Staudte, R.G. y Sheather, S.J. (1990). *Robust estimation and Testing.* New York: Johm Wiley & Sons.

Sturges, H. A. (1926). "The choice of a class interval". *J. American Statistical Association, 21*, 65-66. (Citado por Hoaglin & cols. 1983).

Swanepoel, J.W.H. (1989). "A Review of Bootstrap Methods". Comunicación presentada en el congreso de Matemáticas de Santiago.

Theil, H. (1950). "A rank-invariant method of linear and polynomial regression analysis, I,II, and III". *Proceedings of Koninklijke Nederlandse Akademie van Wetenschappen, 53*, 386-392, 521-525, 1397-1412.

Tukey, J.W. (1949). "One degree of freedom for nonadditivity". *Biometrics, 5*, 232-242.

Tukey, J.W. (1960). "A survey of sampling from contamined distributions". En I. Olkin, S.G. Ghurye, W. Hoeffding, W.G. Madow y H.B. Mann (Eds.), *Contributions to Probability and Statistics, Essays in Honor of Harold Hotelling*, (pp. 448-485). Standford: Standford University Press.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Reading, Massachussets: Addison-Wesley.

Uriel, E. (1985). *Análisis de Series Temporales. Modelos ARIMA.* Madrid: Paraninfo.

Velleman, P. F. (1976). "Interactive computing for exploratory data analysis I: display algorithms". *1975 Proceedings of the Statistical Computing Section.* Washington DC: American Statistical Assoc. págs. 148-152.

Velleman, P.F. y Hoaglin, D.C. (1981). *Applications, Basics and Computing of Exploratory Data Analysis.* Boston: Duxbury.

Wald, A. (1984). "The fitting of straight lines if both variables are subject to error". *Annals of Mathematical statistics, 11*, 284-300.

Yoav, B. (1988). "Openening the box plot". *The American Statistician 42* (4).

Análisis Exploratorio de datos: Nuevas técnicas estadísticas es una obra que puede considerarse de auténtica novedad dentro del campo estadístico. Recoge, por primera vez en lengua castellana, el conjunto de técnicas resistentes y robustas expuestas en Tukey, 1977. Estas técnicas intentan descubrir patrones o modelos en los datos y para ello se valen de importantes innovaciones, principalmente gráficas, como por ejemplo el diagrama de caja.

Estas nuevas técnicas no solo constituyen un complemento a las clásicas sino también una valiosa alternativa.

El libro termina con una introducción a la estimación robusta presentando brevemente los métodos basados en el remuestreo Jackknife y Bootstrap (Efron, 1979).

El lector encontrará la técnica adecuada a multitud de campos como la Economía, el Marketing, las Ciencias Humanas, Sociales y de la Salud. El libro está expuesto de una manera clara y sistemática, con numerosos ejemplos.

Los autores M. Freixa, L. Salafranca, J. Guàrdia, R. Ferrer y J. Turbany son Profesores de Estadística Aplicada a las Ciencias Humanas en la Facultad de Psicología (División de Ciencias de la Salud) de la Universidad de Barcelona.

Al margen de sus contribuciones científicas a diversos congresos, tanto de carácter nacional como internacional, destacan por sus trabajos vinculados con las distintas problemáticas surgidas de la aplicación de las diversas técnicas estadísticas en el campo de las Ciencias Sociales, en especial, obviamente, en el ámbito psicológico.

De ahí que, en conjunto, hayan contribuido a la realización de diversos trabajos subvencionados por Instituciones, tanto públicas como privadas, y al desarrollo de proyectos de investigación básica a los que han aportado el soporte estadístico y de análisis.

Son autores de varias publicaciones, algunas de las cuales figuran en la Bibliografía del presente volumen.

9 788476 651797