

***Capítulo 1:***  
***INTRODUCCIÓN***  
***Y***  
***PRINCIPIOS BÁSICOS.***



## 1.1.- INTRODUCCIÓN.

En los últimos años, y dentro del conjunto de técnicas estadísticas empleadas por diferentes disciplinas, especialmente en las denominadas Ciencias Sociales, está cobrando un creciente interés un nuevo enfoque de análisis de datos denominado *Análisis Exploratorio de Datos (Exploratory Data Analysis)*, cuyo máximo impulsor y exponente es John W. Tukey tras la publicación de su libro pionero *Exploratory Data Analysis* en 1977<sup>1</sup>, en el que se presentan nuevos procedimientos estadísticos y se retoman otros, dentro de un marco globalizador.

A modo de introducción, se puede indicar que el Análisis Exploratorio de Datos es, por una parte, una perspectiva o actitud sobre el análisis de datos, en la que se exhorta a que el investigador adopte una actitud activa en y hacia el análisis de los mismos, como un medio para sugerir nuevas hipótesis de trabajo. Por otra parte, se compone de un renovado utillaje conceptual e instrumental respecto a lo que podríamos llamar Estadística Descriptiva "clásica", con el fin de optimizar la cantidad de información que los datos recogidos puedan ofrecer al investigador, bien a través de novedosas representaciones gráficas, bien a base de reducir la influencia de las puntuaciones extremas en los estadísticos a través del empleo de, los que por ello se ha convenido en llamar, "estadísticos resistentes", por citar sólo dos ejemplos.

De este modo, el Análisis Exploratorio de Datos (que en adelante será citado, como es habitual, por sus siglas, A.E.D.<sup>2</sup>) ha posibilitado la creación de una serie de representaciones gráficas de los datos (vg., los diagramas de "Tallo-y-Hojas" o de "Caja-y-Bigotes", que serán expuestos en el Capítulo 3), muy apropiadas para apreciar rápidamente la estructura y distribución de los datos, y que facilitan el poder optar por el empleo de los estadísticos descriptivos más representativos para los mismos. Además, se propugna el uso de nuevos conceptos, generalmente relacionados con los propuestos por la teoría estadística clásica (vg., uso del término "cuarto" de una manera

---

<sup>1</sup> Aunque ya circulaban ediciones previas al principio de la década.

<sup>2</sup> E.D.A. en la bibliografía anglosajona.

análoga al "clásico" cuartil, pero con ligeras diferencias de cálculo, como se verá).

Igualmente, se exhorta al empleo de los denominados estadísticos resistentes, es decir, estadísticos que son afectados sólo ligeramente por la existencia de algunas puntuaciones atípicas (*outliers*) en el conjunto de datos, a diferencia de los estadísticos más frecuentemente utilizados en los análisis estadísticos, como las medias o varianzas muestrales que sí se ven afectados por tales tipos de puntuaciones.

Si bien lo anteriormente señalado se hallaría dentro del marco de la denominada Estadística Descriptiva, los autores que propugnan el Análisis Exploratorio de Datos, estudian no sólo los estadísticos resistentes, sino que buscan asimismo estimadores adecuados para la inferencia a los parámetros. Consiguientemente, y ya en el área de la Estadística Inferencial, Tukey y otros autores del enfoque del Análisis Exploratorio de Datos exhortan a la utilización de los llamados "estimadores robustos", es decir, estadísticos que pueden servir para inferir parámetros (poblacionales) de diversos tipos de distribución con cierta eficiencia.

Aunque en un primer momento este enfoque suscitó algunas dudas acerca de las verdaderas novedades que aportaba, en cualquier caso, este enfoque ha hecho patente una fase de la investigación poco dada al conocimiento público, así como también ha intensificado el estudio de cierto tipo de gráficos, y de métodos resistentes y robustos. En estos momentos, aun careciendo de la necesaria perspectiva histórica, se puede predecir que las perspectivas de este enfoque del Análisis Exploratorio de Datos, tanto en su espíritu como en sus diversas técnicas, son prometedoras, como se podrá apreciar a lo largo de la lectura de las siguientes páginas.

## **1.1.1.- PERSPECTIVA GENERAL EN LOS ANÁLISIS DE DATOS.**

En el marco de la investigación científica empírica, desde una perspectiva clásica, cabe señalar tres fases de la investigación empírica que corresponden a la Estadística:

I) *La recogida de datos* (Muestreo), en la que se ha de lograr, a través de las diferentes técnicas existentes, una muestra adecuada de la población a la que se desee extrapolar las conclusiones que se obtengan en la muestra.

II) *La organización, descripción y representación gráfica de los datos* (fase Descriptiva), para hacerlos más comprensibles a nuestra mente.

III) *Los procesos de inferencia de parámetros* (fase Inferencial), con objeto de estimar, con una cierta probabilidad, características poblacionales (parámetros) a partir de las características de la muestra (estadísticos), así como diversas pruebas estadísticas para el contraste de las hipótesis de partida.

Debido a que este texto se ocupa básicamente del análisis de datos, se centrará principalmente en la segunda fase, en la que se incluirá el Análisis Exploratorio de Datos, y parcialmente en la tercera fase, inferencial, del análisis, incluyendo una referencia al concepto de los estimadores robustos.

### **1.1.1.1.- ANÁLISIS DE LA FASE DESCRIPTIVA EN LOS ANÁLISIS "CLÁSICOS" DE DATOS.**

Tras la recolección de una muestra de datos, el análisis de los mismos suele comenzar con la descripción de los datos a través de unos estadísticos, que tratarán de resumir la multiplicidad de los datos brutos. Estos estadísticos hacen referencia, especialmente, a la localización central de la

distribución (o índices de tendencia central, como la media aritmética) y a la dispersión de la misma (vg., la desviación típica), así como también índices relativos a la asimetría o a la curtosis de la distribución empírica de los datos.

Adicionalmente, junto al cálculo de los estadísticos se suelen utilizar algunas representaciones gráficas sobre los datos, tales como los Histogramas. Ambos tipos de procedimientos son realizados habitualmente por los investigadores usando cualquiera de los diferentes programas y paquetes estadísticos disponibles en la actualidad en el mercado informático (vg., SPSS, BMDP, SAS, SYSTAT, por citar los más conocidos).

Aunque el proceso anteriormente citado es correcto, el problema reside en que habitualmente, esta fase se realiza de forma rígida, es decir, se atiende casi "ciegamente" a los resúmenes de datos proporcionados por el ordenador, sin apenas atención a la adecuación de la aplicabilidad de tal o cual estadístico para la muestra de datos, lo cual puede oscurecer más que aclarar las propiedades de los datos (HARTWIG Y DEARING, 1979).

De este modo, los estadísticos descriptivos más comúnmente utilizados por los investigadores, clasificados en el enfoque descriptivo "clásico" han sido la media aritmética y la desviación típica. Sin embargo, el uso "indiscriminado" de estos índices no es aconsejable bajo ningún concepto. La media y la desviación típica son índices descriptivos convenientes sólo cuando la distribución de los datos es aproximadamente normal o, al menos, simétrica y unimodal. Sin embargo, muchas variables dependientes que son objeto de estudio en Psicología no suelen cumplir con estos requisitos (vg., el tiempo de reacción ante estímulos, que es bien conocido porque muestra una asimetría positiva). Por tanto, se precisa un examen riguroso de la forma y estructura de los datos previo al tipo de elección de los índices representativos del conjunto de datos.

De hecho, muchos analistas de datos, especialmente los más experimentados, siempre se han fijado con detalle en los datos antes de generar los estadísticos e intentar probar las hipótesis subyacentes a través de la inferencia y la comparación con algún modelo probabilístico. De este modo, aunque sin utilizar las técnicas propias del A.E.D., mantenían la misma actitud hacia los datos sintetizada y propugnada, en época reciente, por este enfoque. Sin embargo, dichas actitudes hacia

la fase descriptiva o de exploración de datos, a pesar de haberse venido tomando desde mucho tiempo atrás por muchos analistas de datos expertos, no han sido, o no han sabido ser, expuestas a la luz pública. En estos casos, el A.E.D. lo que proporciona es un amplio repertorio de métodos para el estudio detallado de los datos, que facilitan la tarea del investigador, especialmente para aquellos con sólo un conocimiento medio de las técnicas estadísticas.

### **1.1.1.2.- ANÁLISIS DE LA FASE INFERENCIAL EN LOS ANÁLISIS "CLASICOS" DE DATOS.**

El análisis inferencial trata de estimar a partir de los estadísticos obtenidos de la muestra, los parámetros de la población, según un determinado modelo o distribución supuesto para ella. Habitualmente, este proceso inferencial es de tipo confirmatorio, es decir, se trata de evaluar una hipótesis de partida más que meramente estimar el valor de unos determinados parámetros. En resumen, se intenta comprobar si se cumple nuestra hipótesis según el modelo supuesto.

Por otra parte, en el campo de la Estadística Inferencial un concepto de uso esencial es el de estimador, que va referido a los estadísticos empleados para inferir los parámetros de la población de la cual son origen.

He aquí un primer ejemplo de análisis inferencial. Pensemos que se desea estimar la altura media de los estudiantes universitarios valencianos. Para ello, se selecciona una muestra de estudiantes, se hallan los estadísticos descriptivos adecuados, y se emplea el estimador adecuado, supongamos la Media muestral, para estimar la Media poblacional. Tal estimación se realiza indicando un intervalo alrededor del cual se hallará la Media poblacional, siempre, al ser un procedimiento inferencial o inductivo, con una cierta probabilidad.

He aquí un segundo ejemplo de análisis a través de un contraste de hipótesis. Sea un estudio en el que se pretende

probar que un método de enseñanza novedoso, que será etiquetado por "A", es mejor que el método tradicional de enseñanza "B". Para ello se cuenta con una muestra homogénea de 30 personas, las cuales son asignadas a uno y otro método de enseñanza aleatoriamente, habiendo 15 personas en cada grupo. Una vez completado un curso con cada uno de los dos métodos, se pasa a probar el rendimiento (en este caso, la variable dependiente) a través de un examen de corrección objetiva. Pensemos que los sujetos del método "A" han conseguido una media de 8'5, mientras que los sujetos del método "B" consiguen una media de 6. A la hora de hacer inferencias lo que se ha de comprobar, siempre probabilísticamente, es que la media poblacional del rendimiento en el grupo "A" difiere de la del "B". Si hay diferencias entre los parámetros de "A" y de "B", entonces se considera que los dos métodos dan lugar a un rendimiento diferente, mientras que si no las hay, se suponen ambos métodos con igual rendimiento. En este último caso, en términos estadísticos se dice que no podemos rechazar la hipótesis Nula (lo que supondría que no podríamos concluir que existen diferencias), mientras que en el primer caso se dice que rechazamos la hipótesis Nula (por lo que admitiremos la existencia de diferencias entre las medias poblacionales entre los grupos).

## **1.2.- EL ENFOQUE DEL ANÁLISIS EXPLORATORIO DE DATOS.**

### **1.2.1.- APLICABILIDAD DEL ANÁLISIS EXPLORATORIO DE DATOS EN PSICOLOGÍA.**

En primer lugar, cabe indicar que, debido a que tales técnicas del Análisis Exploratorio de Datos son apropiadas no sólo para las variables de corte cuantitativo sino también las de índole cuasi-cuantitativo, son aplicables a muchas de las

variables de las que se ocupan las disciplinas como las Ciencias del Comportamiento o la Sociología.

De hecho, en muchos aspectos de la investigación psicológica no se suele avanzar más allá de las variables de índole ordinal, por lo que, en rigor, no se debería calcular ni siquiera la media aritmética, pese a que normalmente los investigadores sí lo hagan. Tales cálculos se efectúan en aras de una mayor potencia de cálculo bajo el supuesto de que se están "midiendo" variables cuantitativas, debido a lo fructífero de los resultados que así es posible obtener.

No obstante, quienes este texto editan (y que estas líneas escriben), coincidiendo con el filósofo pragmático Pierce al considerar que son disputas filosóficas absurdas las que consideran la verdad aparte de sus consecuencias prácticas en el ámbito de la investigación humana, son de la opinión (aunque en distinto grado..., y con las lógicas excepciones) de que "si una cosa sirve, es útil; y si es útil, es legítimo su uso"... El problema radicarán en saber situar bien los límites de ese "uso", y, consecuentemente, de la interpretación del resultado obtenido de él<sup>3</sup>.

En cualquier caso, debido al empleo de los llamados estadísticos de posición por este enfoque (vg., mediana, cuartos,...), muchas de estas técnicas se pueden usar sin necesidad de ningún supuesto en los casos en que haya variables que no sean, *de iure* <sup>4</sup>, cuantitativas, tales como por ejemplo, muchas de las puntuaciones obtenidas en los Tests.

Por otra parte, la utilización de estas técnicas se puede realizar con suma facilidad, dado que su popularidad es ya lo suficientemente considerable como para que hayan sido incluidas en diferentes paquetes estadísticos (SPSS, BMDP, SYSTAT, entre otros).

---

<sup>3</sup> Como indican Glass y Stanley (1986), diversas medidas en Psicología están a medio camino entre las ordinales y las cuantitativas, ya que pese a no tener unidad de medida (no hay unidad constante), sí que indican más información que la meramente ordinal. Por ejemplo, las puntuaciones de tres sujetos en un Test de inteligencia: 79, 123 y 118, no sólo indican que el segundo sujeto tiene la puntuación más alta y el primero la puntuación más baja, sino que, de algún modo, también indican una mayor diferencia entre el primero y los otros dos que la existente entre los sujetos segundo y tercero. En tales casos, los citados autores emplean incluso el término de "escalas de *cuasi-intervalo*".

<sup>4</sup> O "de derecho".

## **1.2.2.- ASPECTOS BÁSICOS DEL ANÁLISIS EXPLORATORIO DE DATOS.**

Las técnicas propuestas por el A.E.D. nos ayudan para abordar un conjunto de datos de un modo informal, guiándonos hacia una estructura de modo rápido y sencillo. Pero como se ha indicado, el A.E.D. no sólo es un conjunto de métodos para el análisis de datos, sino también una perspectiva de análisis de datos que enfatiza la utilización de representaciones gráficas de datos y de métodos resistentes y robustos. Consiguientemente, el A.E.D. es "un estado de ánimo, una forma de pensar acerca del análisis de datos, y también una forma de efectuarlo" (HARTWIG Y DEARING, 1979, p. 9).

El enfoque del A.E.D. acentúa el uso de técnicas alternativas para tratar el mismo conjunto de datos. Cuanto más se sepa de los datos, mejor se podrán utilizar para efectuar pruebas inferenciales y desarrollar las teorías bajo estudio. En resumen, el A.E.D. busca maximizar toda la información que ofrezcan los datos.

De hecho, muchas de las propuestas son paralelas a las de otras técnicas en términos de la teoría estadística clásica. En esta línea, el A.E.D. propugna la utilización de estadísticos como la Mediana como un estadístico resistente de localización, para evitar los problemas de puntuaciones extremas o de la asimetría de la distribución (empírica) de los datos. Mientras, los textos de Estadística Descriptiva "clásicos" suelen señalar, al comparar cuándo utilizar la Media y cuándo la Mediana, que ésta es preferible a aquélla sólo cuando en la distribución de datos de la muestra haya valores extremos, que podrían perjudicar la representatividad de la Media, o bien cuando la distribución sea marcadamente asimétrica, pero no se vislumbra la importancia de un uso masivo de tal índice de tendencia central de modo más generalizado, o bien de otros tipos de estadísticos, como señala el enfoque del Análisis Exploratorio de Datos (como se verá en el capítulo relativo a los estadísticos resistentes).

Otro aspecto de interés es que se pone énfasis en que no se mecanice el empleo de las técnicas del Análisis Exploratorio de Datos, porque se volvería al error de los análisis "clásicos". Así, aunque ciertas técnicas facilitan la exploración de datos, su uso por sí solo no convierte al investigador en "analista exploratorio de datos". En cambio, lo que se requiere es una cierta aproximación al análisis de datos, una cierta perspectiva (HARTWIG Y DEARING, 1979).

### **1.2.2.1.- FASES DE ANÁLISIS ESTADÍSTICO EN EL ANÁLISIS EXPLORATORIO DE DATOS.**

Los autores del enfoque del A.E.D. suelen denominar las fases de análisis estadístico no como *Fase Descriptiva* y *Fase Inferencial* del enfoque clásico (véase Apartado 1.1.1.), sino que indican la existencia de dos amplias fases, hasta cierto punto paralelas a las anteriores, aunque con matices diferentes (HOAGLIN, MOSTELLER Y TUKEY, 1983): la **Fase Exploratoria** y la **Fase Confirmatoria**.

#### **1.2.2.1.1.- ANÁLISIS DE LA FASE EXPLORATORIA DEL ANÁLISIS EXPLORATORIO DE DATOS.**

Así como los análisis descriptivos "clásicos" se han centrado más en el estudio de la localización y dispersión de las distribuciones, al estudiar la Media y la Desviación Típica respectivamente, mientras que una tercera característica de las distribuciones, como es la forma de la misma, no le ha otorgado la atención debida, la Fase Exploratoria del enfoque del Análisis Exploratorio de Datos señala que (HARTWIG Y DEARING, 1979, p. 15):

- 1) La forma de una distribución es como mínimo tan importante como su localización (tendencia central) y dispersión.

- II) Las representaciones visuales son superiores a las representaciones numéricas para descubrir la forma característica de una distribución.
- III) La elección de qué estadístico descriptivo es más apropiado para describir unos datos debería ser dependiente de lo apropiado de tal estadístico de acuerdo con la distribución que presentan dichos datos. Es decir, si en la distribución de los datos se observa una marcada asimetría no es correcto emplear la Media como índice de tendencia central, o si en la distribución se observan puntuaciones extremas en ambos lados de la distribución no es adecuado emplear la Desviación Típica como índice de dispersión.

Ante un conjunto de datos, **unidad básica de análisis para este enfoque**, por tanto, el enfoque del A.E.D. recomienda iniciar su análisis con la realización de gráficos que permitan visualizar su estructura. Por ejemplo, para datos cuantitativos, se aconseja comenzar con el denominado diagrama de "Tallo-y-Hojas", que es una técnica que preserva en lo posible los detalles de los datos, constituyendo un primer nivel de comprensión, o con el diagrama de "Caja-y-Bigotes". Con tales técnicas se pretende detectar su estructura subyacente, así como la existencia de posibles anomalías o patrones no previstos en los mismos, lo que proporciona al investigador una mejor comprensión del fenómeno a estudiar.

La Fase Exploratoria, pues, aísla patrones y rasgos de los datos, y los revela al analista/investigador. Normalmente, proporciona un primer contacto con los datos, precediendo cualquier elección de los modelos probabilísticos a emplear, así como para descubrir patrones o resultados no previstos por el modelo teórico bajo estudio.

De hecho, un investigador debería obtener la máxima información posible de una variable antes de utilizar los datos para poner a prueba las hipótesis, en la inferencia estadística, y así poder enriquecer la teoría subyacente bajo estudio.

En consecuencia, los resultados obtenidos en la Fase Exploratoria determinarán de modo importante qué otras técnicas podrán ser empleadas para examinar un conjunto de datos, a la hora de realizar una posterior fases confirmatoria.

### **1.2.2.1.2.- ANÁLISIS DE LA FASE CONFIRMATORIA DEL ANÁLISIS EXPLORATORIO DE DATOS.**

Esta fase trata de analizar los datos de manera que evalúen la reproductibilidad de los efectos observados (HOAGLIN, MOSTELLER Y TUKEY, 1983). Es paralela, con algunas diferencias, a la fase de inferencia estadística "clásica" ya que posibilita el contraste de hipótesis. Pero el análisis confirmatorio también cubre, habitualmente, otros pasos, que no serán analizados en el presente texto, como, por ejemplo (HOAGLIN, MOSTELLER Y TUKEY, 1983): (I) la incorporación de información obtenida del análisis de otro conjunto (o cuerpo) de datos, relacionado fuertemente con el conjunto de datos a estudiar, o (II) la validación de los resultados al recoger y analizar nuevos datos.

Lo primero que hay que observar, previo a la realización de las diversas pruebas de la carácter inferencial, es cuál es el estimador adecuado para extrapolar las características de la muestra a la población. En los casos prácticos de la Psicología y de otras Ciencias, ocurre que la distribución de los datos en muchas ocasiones no se parece a la distribución Normal, mientras que muchas pruebas precisan de una distribución normal de los datos o, al menos, de una distribución simétrica y unimodal. Ante esto, básicamente son dos las estrategias a seguir propugnadas por el A.E.D. para solucionar el problema del tipo de distribución de datos:

- I) Utilizar estadísticos resistentes ante la existencia de sesgo en la distribución o la presencia de datos atípicos, que sirvan, a su vez, de estimadores robustos de los parámetros.
- II) Modificar la forma de la distribución para conseguir la simetría mediante alguna transformación sencilla de los datos (vg., calcular el logaritmo de las puntuaciones en el caso de una distribución asimétrica positiva). Así, si obtenemos una distribución simétrica, conseguiremos una descripción más simple de sus características, y podremos utilizar la Media y Desviación Típica muestrales, es decir

los estadísticos más utilizados por la inferencia estadística.

Por otra parte, en relación a los análisis de corte confirmatorio, el A.E.D. mantiene que ha de haber una postura abierta respecto a los resultados, y no tener la mente prefijada en una determinada hipótesis que, incluso en el caso de que no se cumpla, se añadan modificaciones *post hoc* para que se "cumpla". Hay que fijarse más en la exploración de los datos, lo que no es óbice, por supuesto, para que haya fase confirmatoria, sino para que ésta sea suficientemente flexible (HARTWIG Y DEARING, 1979).

En resumen, la Fase Exploratoria enfatiza la búsqueda flexible de evidencia, mientras que la Fase Confirmatoria acentúa la evaluación/contraste de la evidencia disponible. El hecho de alternar ambos usos de técnicas de análisis, tanto exploratorias como confirmatorias es, a menudo, muy deseable (HOAGLIN, MOSTELLER Y TUKEY, 1983).

### **1.2.3.- PRINCIPIOS BÁSICOS DEL ANÁLISIS EXPLORATORIO DE DATOS.**

De acuerdo con Hartwig y Dearing (1979), cabe desarrollar dos principios básicos que guían a todo investigador que siga la perspectiva del Análisis Exploratorio de Datos, son el escepticismo y una actitud abierta, que serán tratados a continuación.

### **1.2.3.1.- ESCEPTICISMO.**

Un primer principio del A.E.D., según los citados autores, es que se ha de ser escéptico sobre los "resúmenes numéricos" del conjunto de datos (vg., ofrecer la Media, la Desviación Típica, la Mediana, el Número de Casos, etc., para describir un conjunto de datos), puesto que a veces pueden oscurecer o no revelar lo que podrían ser los aspectos más informativos de los datos. Habitualmente, por contra, suele tenerse demasiada fe en los resúmenes numéricos de los datos (HARTWIG Y DEARING, 1979).

La concepción de la Estadística como análisis de datos parece que acentúa la importancia de los números, es decir, los estadísticos, por sí solos (resúmenes numéricos de los datos), con la exclusión de otros métodos de análisis. Así, la estadística "clásica" tiende a rebajar la importancia de las representaciones gráficas de los datos, aspectos que en cambio, son ampliamente divulgados por el A.E.D.

El punto de vista habitual y clásico mantiene que un estadístico es más "sólido" que una representación gráfica de los datos. Sin embargo, estos estadísticos pueden oscurecer e incluso ignorar información que pueda ser de gran importancia. En el A.E.D., el análisis visual, como se ha indicado anteriormente, debiera preceder a los análisis estadísticos "numéricos", si bien estos últimos siguen siendo el producto final deseado.

### **1.2.3.2.- ACTITUD ABIERTA.**

El analista ha de estar abierto a patrones no previstos en los datos, ya que éstos podrían ser los aspectos más reveladores del análisis. Muchas veces, en las diversas Ciencias Sociales, pese a la habitual debilidad de sus teorías, los investigadores suelen ignorar patrones alternativos a las hipótesis de partida.

Como ya se ha indicado, los análisis estadísticos "clásicos" suelen ser de tipo confirmatorio, es decir, se parte de unas hipótesis de trabajo previas. En esta línea, en muchas

ocasiones, los datos no son explorados en busca de la existencia de patrones alternativos de explicación de los mismos (HARTWIG Y DEARING, 1979). En cambio, un "modo exploratorio" de análisis se halla abierto a un amplio número de explicaciones alternativas. Hay que recordar que los datos provenientes de las Ciencias Sociales suelen ser obtenidos bajo condiciones no repetibles, o con una relación vaga de los constructos con la teoría.

## **1.2.4.- TEMAS BÁSICOS EN EL ANÁLISIS EXPLORATORIO DE DATOS.**

Siguiendo a Hoaglin, Mosteller y Tukey (1983), se puede hablar de cuatro grandes temas que siempre se hayan presentes en el Análisis Exploratorio de Datos (Estadísticos Resistentes, Datos Residuales, Transformación de datos y "Revelación"), que serán analizados en los siguientes cuatro subapartados.

### **1.2.4.1.- ESTADÍSTICOS RESISTENTES.**

Primeramente, se procederá a definir lo que es una estadístico resistente: Se dice que un estadístico es resistente, en su acepción más general, cuando el valor de éste no se halla apenas afectado por cambios arbitrarios en una pequeña parte del conjunto de datos. De este modo, si cambiamos unos pocos datos en un conjunto de datos, y los sustituimos por otros que sean bastante diferentes a los anteriores, un estadístico resistente no se hallará apenas afectado. En otras palabras, los métodos resistentes prestan mucha atención al cuerpo de los datos y poco a las puntuaciones extremas, de modo que puedan "resistir" pequeñas alteraciones de los datos.

Desgraciadamente, la media y varianza muestrales no pueden ser consideradas en absoluto como estadísticos resistentes, ya

que de acuerdo con su fórmula de cálculo, una sola puntuación extrema puede afectar de modo importante a ambas. Sea el siguiente conjunto de datos 8,9,9,8,9,7,1 y 9; en el cual se aprecia que el dato "1" afecta de modo importante tanto a la media como a la varianza. En cambio, los denominados estadísticos de orden no se verían afectados en tal medida por esas puntuaciones. Ejemplos de ello son la Mediana para la tendencia central y Amplitud Inter-Cuartil para la dispersión.

Sin embargo, una matización a la definición indicada es la siguiente: **un estadístico es resistente cuando el valor de éste no se halla apenas afectado por un pequeño número de puntuaciones extremas<sup>5</sup> o por cierto número de errores de redondeo, truncado o agrupamiento.** La primera parte se refiere a la existencia de puntuaciones extremas, tal y como ocurría en el ejemplo anterior. La segunda parte se refiere a la resistencia no ya a las puntuaciones atípicas, sino a la resistencia a errores de redondeo o agrupamiento. Consiguientemente, un estadístico resistente no se debe ver tampoco apenas influido por errores de redondeo, truncado o agrupamiento de una pequeña fracción de los datos. En esta línea, o segunda acepción, la Mediana no sería un estadístico particularmente resistente, ya que, sabiendo que el valor de la mediana únicamente depende de unas pocas puntuaciones centrales de la distribución, un pequeño error de redondeo o agrupamiento afectaría al valor obtenido. Como Goodall (1893b, p. 349) señala, «...ordinariamente, tememos más los efectos de las puntuaciones atípicas que los procedimientos de redondeo o agrupamiento, de manera que se utiliza habitualmente la Mediana como estadístico resistente. Sin embargo, para análisis más refinados, preferimos un estadístico menos sensible al redondeo y agrupamiento, así como que se mantenga resistente ante las puntuaciones atípicas». Por ello, en el capítulo segundo se hablará de la Mediana y de otros estadísticos resistentes de mayor interés.

---

<sup>5</sup> O atípicas (que se desvían de la "norma").

### 1.2.4.2.- DATOS RESIDUALES.

Pensemos, siguiendo el análisis estadístico clásico, que pretendemos observar la relación lineal existente entre las variables X e Y, para lo cual efectuamos un ajuste por el método de Mínimos Cuadrados. Pues bien, los datos residuales son los datos que tras la aplicación del modelo no encajan en el mismo. En el ejemplo, los residuales serían las desviaciones verticales respecto a la recta ideal (el modelo ajustado) entre las variables X e Y. Es decir,

$$\boxed{\text{DATOS BRUTOS}} = \boxed{\text{DATOS AJUSTADOS}} + \boxed{\text{DATOS RESIDUALES}}$$

o bien,

$$\boxed{\text{DATOS}} = \boxed{\text{PARTE AJUSTADA}^6} + \boxed{\text{PARTE RESIDUAL}^7} .$$

Esta ecuación, si se considera exclusivamente el aspecto descriptivo univariado, podría también expresarse como

$$\boxed{\text{CONJUNTO DE DATOS}} = \boxed{\text{TENDENCIA}} + \boxed{\text{EXCLUSIVIDAD}} ,$$

aspectos en los que podría descomponerse el conjunto de los datos:

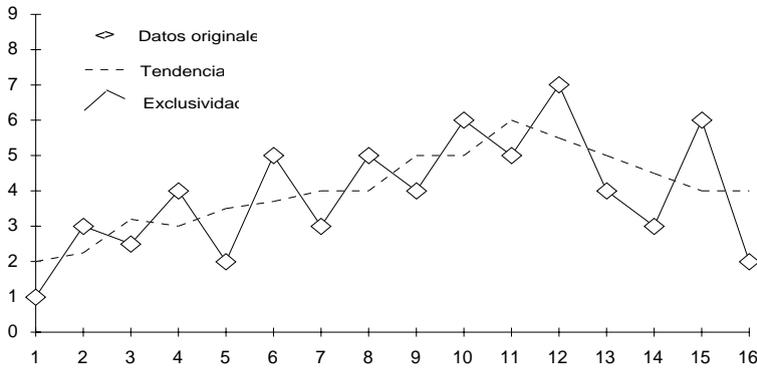
$$\text{CONJUNTO DE DATOS} \left\{ \begin{array}{l} \text{TENDENCIA} \\ \text{EXCLUSIVIDAD.} \end{array} \right.$$

El gráfico siguiente, referido a una serie temporal, ayudará a comprender mejor estos conceptos, en el que los "rombos" representan los datos obtenidos, la línea discontinua, no pasando por ninguno de los datos, indica la tendencia que sigue el conjunto de datos. Mientras, los segmentos de líneas quebradas que existen entre cada "rombo" y la línea discontinua muestran la parte exclusiva del conjunto de datos, lo que cada dato o caso tiene de específico (es decir, la desviación vertical de cada rombo respecto a la línea discontinua).

---

<sup>6</sup> "Smooth".

<sup>7</sup> "Rough".



El enfoque del A.E.D., además de ofrecer procedimientos más resistentes, al analizar la relación entre variables, que el de Mínimos Cuadrados, tal como la Línea de Tukey (que se verá en el Capítulo 4), mantiene que los datos residuales pueden avisar de aspectos sistemáticos importantes de los datos que pueden necesitar de atención, tal como la curvatura, no-aditividad y no-constancia de la variabilidad. Es decir, el análisis de residuales se emplea para detectar inadecuaciones de los modelos teóricos puestos a prueba, en los que habitualmente se supone que los residuales son meramente error aleatorio, sin ningún tipo de sistematismo.

En conclusión, un aspecto clave del A.E.D. es que un análisis de un conjunto de datos no se halla completo sin un cuidadoso examen de los datos residuales, aspectos que se verán ampliados en el Capítulo 4.

### **1.2.4.3.- RE-EXPRESIÓN O TRANSFORMACIÓN DE LOS DATOS.**

La transformación de los datos se refiere al cambio en la escala de medida, de la variable objeto de estudio, para adecuar los datos a los supuestos de los modelos paramétricos. Por ejemplo, un supuesto previo a la hora de realizar una prueba de inferencia estadística tan habitual como es el Análisis de Varianza (ANVAR o ANOVA) es que los datos deben seguir la distribución normal. En el caso de que se desee aplicar un Análisis de Varianza a unos datos que muestren una distribución positivamente asimétrica se podría, siguiendo el caso, no tomar los valores en la escala original, sino realizar el logaritmo de cada uno de ellos y observar si tras tal transformación la distribución de los mismos es normal, y así cumplir con el supuesto de la normalidad de puntuaciones del Análisis de Varianza.

La transformación de los datos va referida a que el investigador pueda analizar los datos a través de la utilización de otra escala (logarítmica, mediante raíz cuadrada, mediante el arco-seno de las puntuaciones, etc.), que pueda simplificar el análisis de datos. Es decir, se trata de considerar si la escala original de medida es satisfactoria. En el caso de que no lo sea, la transformación de datos en otra escala puede dar lugar a la deseada simetría de la distribución, a la constancia en su variabilidad, a la fuerza de una relación entre variables, a la aditividad de un efecto, o, en general, a los supuestos subyacentes a las distribuciones probabilísticas que se deseen a continuación aplicar.

### **1.2.4.4.- REVELACIÓN.**

Este aspecto va referido a la importancia de las representaciones gráficas en el A.E.D., que proporcionan la posibilidad de observar ciertos rasgos inesperados, o ciertas regularidades, que podrán favorecer la "intuición" (*serendipity*)

<sup>8</sup>) del investigador. En otras palabras, el empleo de las diversas técnicas gráficas disponibles aparece como una estrategia básica para revelar/desvelar la estructura de los datos. En diversos casos, tales como el descubrimiento de "agujeros" en las distribuciones o de "picos múltiples" (conceptos que serán tratados en los próximos apartados), las representaciones gráficas, tales como las propuestas por el A.E.D. resultan imprescindibles.

Además, las representaciones gráficas empleadas pueden resultar de utilidad no sólo para la búsqueda y exploración de características, sino también como una vía de comunicar información a audiencias no técnicas.

Pero más importante aún es su énfasis en el uso de los gráficos no sólo en la fase previa al análisis de datos sino también durante y después del mismo. Es decir, a pesar de disponer de resúmenes del conjunto de datos (estadísticos), no hay que olvidarse del origen de la información, que está mejor reflejado en una representación gráfica que preserve suficiente información tal y como ofrece, por ejemplo, el diagrama de "Tallo-y-Hojas", que conserva los datos originales.

## **1.2.5.- AFRONTAMIENTO DE PROBLEMAS BÁSICOS EN EL ANÁLISIS EXPLORATORIO DE DATOS.**

En los próximos cuatro subapartados se analizará el modo que propone el enfoque del Análisis Exploratorio de Datos de afrontar diversos problemas básicos que puede poseer la estructura de las distribuciones de un conjunto de datos, como son la asimetría de la distribución, la existencia de puntuaciones atípicas (o extremas), de "agujeros" o de múltiples "picos" (o modas).

---

<sup>8</sup> Algo así como: Idea brillante surgida "casualmente" a un hombre brillante que estaba brillantemente en el camino adecuado (vg., manzana de Newton).

### **1.2.5.1.- ASIMETRÍA DE LA DISTRIBUCIÓN DE DATOS.**

Los textos de Estadística Descriptiva suelen indicar que si los datos se distribuyen más o menos simétricamente, la Media es el estadístico más adecuado, pero si la asimetría es apreciable, es más correcto el empleo de la Mediana. Ello se debe a que la Media se ve influida por todos los valores que, en el caso de una distribución asimétrica, se verá influida por los valores de la cola más larga. Análogamente, también la Desviación Típica se ve afectada por la asimetría, por lo que un estadístico como la Amplitud Inter-Cuartil, ofrecerá mayor resistencia. Mientras, el enfoque del A.E.D. señala un amplio número de estadísticos resistentes que permiten el análisis descriptivo de las distribuciones asimétricas.

Sin embargo, hay que advertir que lo anterior se refiere básicamente a las distribuciones ligera o moderadamente asimétricas. En el caso de las distribuciones marcadamente asimétricas no sería ni siquiera correcto hablar de medidas de tendencia central (HEATHCOTE, POPIEL Y NEWHORT, 1991), con lo que la única posibilidad que le queda al analista será, en este caso, observar y obtener conclusiones de una representación gráfica, tal y como un diagrama de "Caja-y-Bigotes" o de "Tallo-y-Hojas", o efectuar una transformación de los datos.

### **1.2.5.2.- PUNTUACIONES ATÍPICAS.**

Un problema que ha preocupado a los analistas de datos desde tiempo atrás es el de cómo abordar los estadísticos representativos cuando en el conjunto de datos recogidos hay los denominados datos atípicos (o lo que es lo mismo, puntuaciones

extremadas o "aberrantes")<sup>9</sup>, es decir, datos que se hallan separados del grueso de las puntuaciones. Este hecho puede ser debido a diversos factores, desde errores en la recogida de los datos, hasta datos que por cualquier circunstancia toman esos valores, pero no nos vamos a detener en las consideraciones de cuándo una puntuación llega a ser considerada como "atípica", debido a lo extenso y problemático que puede resultar el asunto, en sus matizaciones. Al igual que el tema de la asimetría es un tema de interés especial para los analistas de datos.

Debido a las características de estadísticos como la Media o Varianza muestrales<sup>10</sup> que, para su cálculo dependen de todas y cada una de las puntuaciones, la existencia de que una o varias de ellas se alejen de modo importante del resto hace que los valores que nos ofrezcan sean, de algún modo, un tanto irreales. Por ejemplo, piénsese en los datos 3,1,1,2,3,1,40,1; y en su media teniendo en cuenta ó no el dato atípico "40". Ante este problema, han sido propuestas diversas alternativas de actuación:

- a) Ateniéndonos a los datos provenientes de los estudios experimentales en Psicología que suelen tener como variable dependiente el Tiempo de Reacción, un procedimiento muy usual es el siguiente. Bajo la suposición de que la distribución de los tiempos de reacción sigue la curva normal (*sic*<sup>11</sup>), se procede a eliminar, de cada sujeto, bien aquellas puntuaciones que se alejen más allá de la media más/menos dos Desviaciones Típicas, o bien se indican unas puntuaciones fijas arbitrarias (por exceso y por defecto) más allá de las cuales los datos existentes serán suprimidos de los ulteriores análisis. Una vez realizada tal operación de "limpieza", y "normalizada" la distribución de los datos, se efectúa el cálculo de la Media muestral, proceso que equivale a una manera de recortar la Media, de forma hasta cierto punto análoga a los procedimientos de "medias recortadas" que se analizarán en el Capítulo segundo.

---

<sup>9</sup> *Outliers* en la literatura anglosajona.

<sup>10</sup> Especialmente la Varianza, que al calcularse con empleo de potencias (diferencias al cuadrado) maximiza el efecto de las puntuaciones extremas.

<sup>11</sup> Abreviatura que utilizaremos en su segundo sentido de advertir del uso de una afirmación o razonamiento (que se sabe) falso pero que se sigue bajo el supuesto de si fuera verdadero, por conveniencia para la argumentación.

b) Adoptar la estrategia propuesta por el enfoque del A.E.D. de utilizar, como se ha indicado con anterioridad, los diversos estadísticos resistentes, es decir, de índices que son poco afectados por las variaciones provocadas por la existencia de unos pocos (o presencia de algunos) datos atípicos. Tales estadísticos resistentes, los más empleados (vg., Mediana, Centrimea, Trimedia, entre otros) son analizados en el Capítulo 2, con un análisis comparativo de las ventajas de unos respecto a otros, en términos de robustez en la estimación.

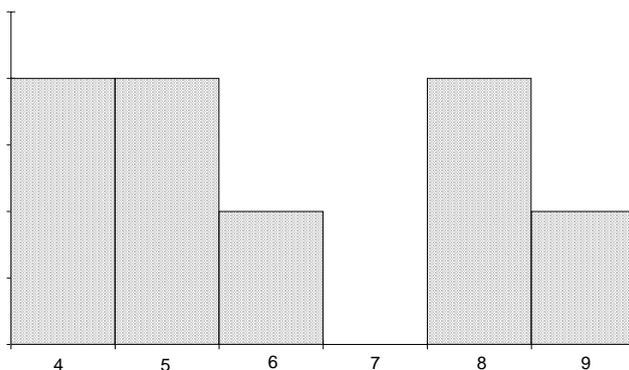
### 1.2.5.3.- "AGUJEROS".

Los "agujeros"<sup>12</sup> en la distribución de un conjunto de datos se refieren a la existencia, en la distribución de los mismos, de intervalos o zonas en las que no se han hallado datos o muy pocos en relación a las puntuaciones circundantes. Por ejemplo, en la secuencia: 4, 5, 6, 4, 5, 4, 5, 8, 6, 9, 8, 9, 8, 5, 4 y 8; hay un agujero en el "7", valor, por cierto, que se halla relativamente cercano a la media de los datos (6'125), lo que, sin embargo, da idea de lo poco adecuado que es, al menos en este caso, hablar de un valor de tendencia central o representativo del promedio, al recaer éste cercano a un "agujero". La mejor manera de apreciar la existencia de agujeros es a través de los métodos gráficos desarrollados por este enfoque (vg., los llamados diagramas de "Tallo-y-Hojas"), que serán descritos en el Capítulo 3, no obstante, puede servir un Histograma "clásico" para ilustrar la idea.

---

<sup>12</sup> *Gaps* en la literatura anglosajona.

ILUSTRACIÓN DE UN "AGUJERO" CON LOS DATOS DEL EJEMPLO EXPUESTO:

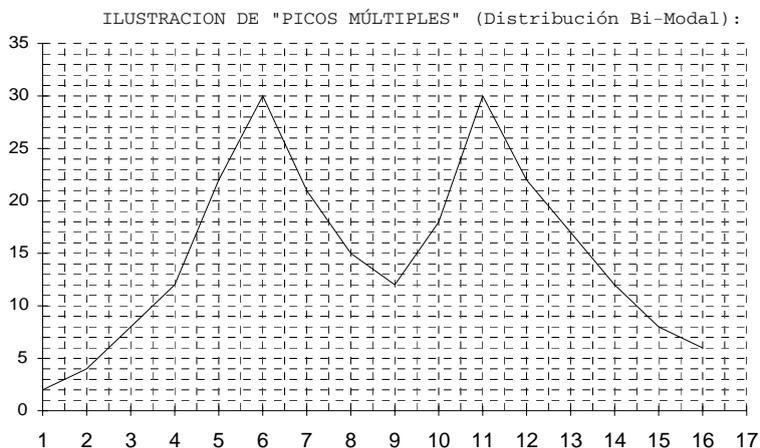


#### 1.2.5.4.- "PICOS MÚLTIPLES".

La existencia de múltiples "picos"<sup>13</sup>, o de múltiples modas<sup>14</sup>, se refiere a que hay diversos intervalos de puntuaciones en una distribución que tienen un parecido número de frecuencias, es decir, que es una distribución bi- o multimodal. Al igual que en el caso de los "agujeros", es a través de un procedimiento gráfico como mejor puede ser apreciado por el analista (óptimamente por medio de un diagrama de "Tallo-y-Hojas", no obstante, dado que aún no han sido expuestos los métodos de representación gráfica propios de este enfoque, un Polígono de Frecuencias "clásico" servirá de acercamiento para ilustrar la idea).

<sup>13</sup> Peaks en la literatura anglosajona.

<sup>14</sup> Nosotros preferimos hablar de "picos" para referirnos a este fenómeno, pues ateniéndonos a la definición usual de la Moda como el valor que más se repite en una distribución, podría darse el caso de que uno de los "picos" tuviera mayor altura que los demás, siendo, por tanto, la distribución, en rigor, unimodal; y sin embargo, estar afectada del fenómeno al que nos referimos. Por ello, algunos autores hablan de "moda mayor" para referir a la Moda en sentido estricto, y de "moda/s menor/es" para referirse a los valores cuya frecuencia destaca del grueso de los datos pero no cumplen esta definición de Moda.



Como podrá deducirse del ejemplo, en estos casos resultará muy difícil hablar de un índice de tendencia central adecuado, ya que posiblemente no tenga sentido indicar un índice representativo de la misma. Más bien, resulta más esclarecedor señalar los intervalos con mayor frecuencia y analizar posibles causas de tal resultado, como por ejemplo, la superposición en el conjunto de datos de dos distribuciones, supónganse normales, pero con diferente media.

## 1.3.- ESTADÍSTICA ROBUSTA.

### 1.3.1.- PROBLEMAS DE LA TEORÍA ESTADÍSTICA CLÁSICA.

El problema que se plantea a la hora de abordar los distintos procedimientos estadísticos de estimación o verificación de hipótesis es que se ha de partir de cierto conjunto de condiciones, o supuestos, entre las cuales se hallan

especialmente aspectos como la independencia de las observaciones entre sí, a la suposición de que siguen una misma distribución, así como a las asunciones respecto del carácter de la distribución de los elementos de la muestra. El incumplimiento de tales condiciones da lugar a que las afirmaciones que se logren al aplicar las pruebas inferenciales respectivas serán, habitualmente, inciertas.

Sin embargo, como se ha indicado con anterioridad, en la investigación empírica, tales condiciones son más la excepción que la regla. Muchos de los supuestos matemáticos son más bien el resultado de una inevitable idealización. Por consiguiente, tales condiciones no se suelen cumplir de manera exacta, surgiendo dudas sobre los resultados que se puedan conseguir con los procedimientos al uso una vez que se incumpla alguna de las asunciones subyacentes.

La teoría estadística clásica ha asumido distintas asunciones respecto a las muestras, tales como la noción de muestra de gran tamaño, así como la no violación de los supuestos subyacentes a las distribuciones estadísticas, temas sobre los que se reflexionará en los siguientes apartados.

### **1.3.1.1.- MUESTRAS PEQUEÑAS.**

La teoría estadística clásica enfatiza nociones como la de muestra que cuenta con un número de datos relativamente elevado. De este modo han surgido la "*ley (débil) de los grandes números*" que viene a indicar que conforme la muestra sea mayor la probabilidad de que un estadístico estime correctamente el parámetro correspondiente a su población tiende a uno. Análogo razonamiento sigue otra premisa básica de la Estadística Inferencial, como es el "*teorema central del límite*".

En resumen, conocer el comportamiento de un estadístico a medida que la muestra es de mayor tamaño tiene utilidad cuando, como frecuentemente ocurre, el comportamiento de las muestras grandes es más simple que el de las pequeñas.

Sin embargo, hay que reconocer que los conjuntos de datos que manejan los investigadores son, muy a menudo, pequeños, y carecen de la simplicidad de las grandes muestras, con lo que están sujetos a problemas a la hora de inferir a los parámetros (vg., el caso del estudio de muestras que provengan de poblaciones de alta variabilidad, lo que da lugar a grandes variaciones en las muestras pequeñas, es decir, a notables diferencias en los valores de los estimadores en muestras diferentes de una misma población).

El ejemplo más claro es que en las series de datos que se recogen, incluso las provenientes de una distribución normal, suele estar presente una pequeña porción de puntuaciones atípicas, es decir, generadas por errores de medición, o causadas por cualquier otro tipo de circunstancia. Habitualmente, la separación de este tipo de puntuaciones del resto es imposible. En lugar de ello, lo que se busca son procedimientos que sean poco sensibles a tales puntuaciones, como son los estadísticos resistentes, que, a su vez, serán empleados como estimadores.

### **1.3.1.2.- SUPUESTOS NO REALISTAS DE LAS DISTRIBUCIONES Y PRUEBAS TEÓRICAS.**

Las técnicas provenientes de la Estadística clásica están diseñadas para ser las mejores siempre que se les apliquen severas constricciones. Sin embargo, la experiencia y la investigación nos han forzado a reconocer que las técnicas clásicas pueden resultar equívocas cuando la situación práctica se aleja del modelo ideal descrito por tales asunciones o supuestos. Por ello, los recientes estudios que han desarrollado métodos robustos y exploratorios están ampliando la efectividad de los análisis estadísticos.

Así, la teoría estadística clásica responde a situaciones que podríamos denominar "ideales", que ocurren raramente en las investigaciones empíricas reales. De este modo, en lugar de buscar los estadísticos más adecuados en unas situaciones ideales, podría resultar más adecuado el empleo de estadísticos

que puedan afrontar gran cantidad de situaciones, tal como propone el A.E.D.

Veamos un caso habitual que ilustra el procedimiento a seguir. Si se sabe que los datos que hemos recogido realmente provienen de una población que sigue una distribución normal, entonces la Media muestral será un buen estimador de la Media poblacional. Sin embargo, incluso en distribuciones simétricas unimodales, en el caso de que tengan colas más gruesas en los extremos que la distribución normal, es decir, que sean leptocúrticas (vg., distribución de Cauchy<sup>15</sup>), y por ende, en distribuciones en las que posiblemente en sus muestras haya más puntuaciones extremas, la Media muestral resulta ser un estimador poco fiable de la Media poblacional (DEGROOT, 1988). En este caso, un estimador razonable y relativamente simple sería la Mediana muestral.

De forma esquemática, al respecto de lo que se acaba de exponer, cabe hablar de diversas situaciones (DEGROOT, 1988):

- I) Si se puede suponer que la distribución poblacional es una distribución normal, o aproximadamente normal, y la muestra es relativamente grande entonces se podría usar la Media muestral como un estimador aceptable de la Media poblacional.
- II) Si se cree que la distribución descrita por la población es simétrica y unimodal, pero con unas colas más gruesas en los extremos (leptocúrtica) que la distribución normal se podría utilizar la Mediana muestral como estimador de la Media poblacional.
- III) Si como ocurre muy comúnmente, en la investigación empírica, en disciplinas como la Psicología, no se sabe si la distribución de la población es normal, se debería tratar de encontrar un estimador de la Media poblacional adecuado para varios tipos posibles de distribuciones, es decir, se busca un *estimador robusto*. Pero eso será materia de otro apartado...

---

<sup>15</sup> Es un modelo de distribución muy especial: No se puede calcular su Media o Esperanza Matemática (da  $\infty$ ), aunque es una distribución simétrica en torno a 0, ya que su Función de Densidad de Probabilidad, que es:

$$f(x) = \frac{1}{\pi(1+x^2)} ; \quad -\infty < x < +\infty, \text{ se aproxima a } 0, \text{ en los extremos, mucho más}$$

lentamente que la distribución Normal, resultando muy leptocúrtica.

### 1.3.2.- ESTIMADORES ROBUSTOS.

El concepto de estimador robusto, previo a la introducción del Análisis Exploratorio de Datos, es un término que fue acuñado por Box en 1953. Un estimador se dice que es robusto cuando funciona bien para varios tipos distintos de distribuciones teóricas, aunque pueda no ser el mejor estimador disponible para ningún tipo concreto de distribución. Los estimadores robustos, en lugar de ser los mejores de una situación estrechamente definida, son los "mejores compromisos" para un amplio abanico de situaciones y, sorprendentemente, están cercanos a los "mejores" en cada situación por sí sola.

El uso de estimadores robustos es particularmente ventajoso en las Ciencias Sociales ya que a menudo no suelen haber unas teorías bien definidas, con lo que no se sabe con certeza el tipo de distribución poblacional de los datos bajo estudio.

Por tanto, **el concepto de robustez se refiere a la insensibilidad de un estimador a violar los supuestos subyacentes bajo un modelo probabilístico o distribución teórica.** Desde esta perspectiva, cabe indicar que la Media y Varianza muestrales no son estimadores robustos, mientras que un estadístico como la Mediana muestral aun siendo más robusto que los anteriores, no lo es particularmente, ya que otros estimadores tienen una *eficiencia*<sup>16</sup> mayor ante un amplio rango de distribuciones (como intentaremos mostrar en el Capítulo 2).

Una nota final a señalar en este punto es que tanto los métodos, llamémoslos tradicionales, como los métodos robustos son métodos paramétricos, es decir, pretenden inferir aspectos

---

<sup>16</sup> Las Distribuciones Muestrales de la Media y de la Mediana tienen la misma media. Sin embargo la varianza de la Distribución Muestral de Medias es menor que la de la Distribución Muestral de Medianas. Como podrá deducirse de las

fórmulas respectivas de sus desviaciones típicas (Errores Típicos):  $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$

y  $\sigma_{Md} = \frac{1,2533\sigma}{\sqrt{N}}$ , el numerador es mayor en Md, luego el valor resultante también será mayor.

provenientes de distribuciones o modelos probabilísticos, con la diferencia de que los estimadores robustos seguirán siendo relativamente eficientes aunque se incumplan las restricciones impuestas por los modelos paramétricos clásicos. A diferencia de los dos métodos, ambos paramétricos, los denominados métodos no paramétricos son aquellos que no tienen en cuenta el tipo de distribución poblacional para sus procedimientos, por ello es frecuente asociar a ellos el concepto de "distribución libre".

También es conveniente indicar que hay una cierta relación entre los métodos resistentes y los robustos. De este modo, diversos autores indican que la resistencia no es más que una "robustez cualitativa" (HOAGLIN, MOSTELLER Y TUKEY, 1983, p. 2).

En este texto serán analizados los estadísticos resistentes más conocidos, junto a su eficiencia en la estimación de parámetros de diferentes tipos de distribución (robustez).



Capítulo 1: INTRODUCCIÓN Y PRINCIPIOS BÁSICOS. ....	9
1.1.- INTRODUCCIÓN. ....	11
1.1.1.- PERSPECTIVA GENERAL EN LOS ANÁLISIS DE DATOS. ....	12
1.1.1.1.- ANÁLISIS DE LA FASE DESCRIPTIVA EN LOS ANÁLISIS "CLÁSICOS" DE DATOS. ...	13
1.1.1.2.- ANÁLISIS DE LA FASE INFERENCIAL EN LOS ANÁLISIS "CLÁSICOS" DE DATOS. ...	14
1.2.- EL ENFOQUE DEL ANÁLISIS EXPLORATORIO DE DATOS. ....	15
1.2.1.- APLICABILIDAD DEL ANÁLISIS EXPLORATORIO DE DATOS EN PSICOLOGÍA. ....	15
1.2.2.- ASPECTOS BÁSICOS DEL ANÁLISIS EXPLORATORIO DE DATOS. ....	16
1.2.2.1.- FASES DE ANÁLISIS ESTADÍSTICO EN EL ANÁLISIS EXPLORATORIO DE DATOS. ...	17
1.2.2.1.1.- ANÁLISIS DE LA FASE EXPLORATORIA DEL ANÁLISIS EXPLORATORIO DE DATOS. ....	18
1.2.2.1.2.- ANÁLISIS DE LA FASE CONFIRMATORIA DEL ANÁLISIS EXPLORATORIO DE DATOS. ....	19
1.2.3.- PRINCIPIOS BÁSICOS DEL ANÁLISIS EXPLORATORIO DE DATOS. ....	20
1.2.3.1.- ESCEPTICISMO. ....	20
1.2.3.2.- ACTITUD ABIERTA. ....	21
1.2.4.- TEMAS BÁSICOS EN EL ANÁLISIS EXPLORATORIO DE DATOS. ....	21
1.2.4.1.- ESTADÍSTICOS RESISTENTES. ....	22
1.2.4.2.- DATOS RESIDUALES. ....	23
1.2.4.3.- RE-EXPRESIÓN O TRANSFORMACIÓN DE LOS DATOS. ....	24
1.2.4.4.- REVELACIÓN. ....	25
1.2.5.- AFRONTAMIENTO DE PROBLEMAS BÁSICOS EN EL ANÁLISIS EXPLORATORIO DE DATOS. ....	26
1.2.5.1.- ASIMETRÍA DE LA DISTRIBUCIÓN DE DATOS. ....	26
1.2.5.2.- PUNTUACIONES ATÍPICAS. ....	27
1.2.5.3.- "AGUJEROS". ....	28
1.2.5.4.- "PICOS MÚLTIPLES". ....	29
1.3.- ESTADÍSTICA ROBUSTA. ....	30
1.3.1.- PROBLEMAS DE LA TEORÍA ESTADÍSTICA CLÁSICA. ....	30
1.3.1.1.- MUESTRAS PEQUEÑAS. ....	31
1.3.1.2.- SUPUESTOS NO REALISTAS DE LAS DISTRIBUCIONES Y PRUEBAS TEÓRICAS. ....	32
1.3.2.- ESTIMADORES ROBUSTOS. ....	33