Qüestió - V. 9, n.º 2 (juny 1985) pp. 105-119

NUEVAS TÉCNICAS DE ANÁLISIS ESTADÍSTICO DE DATOS: TABULACIÓN Y SÍNTESIS NUMÉRICAS (ANÁLISIS EXPLORATIVO DE DATOS)

JOAN MANUEL BATISTA I FOGUET, MOISES VALLS I COLOM
UNIVERSITAT POLITÈCNICA DE CATALUNYA

Considerar por primera vez un conjunto de datos para analizarlos estadisticamente, requiere adoptar una cierta actitud frente a la información recogida, más que someterla automáticamente a un conjunto de técnicas de análisis. Este artículo, es el primero de una serie que se denomina genéricamente Análisis Exploratorio de Datos, cuyo objetivo global es proporcionar alternativas a la aproximación descriptiva que la estadística clásica propone. Se establece a continuación un marco adecuado para proceder a tabular y calcular resúmenes numéricos de datos. Próximamente, se afrontarán los problemas de la representación gráfica y de las adecuadas transformaciones de los datos originales, ilustrando cada uno de los procedimientos con aplicaciones a casos reales.

Keywords: VARIABLE ALEATORIA, DISTRIBUCION DE FRECUENCIAS, DISTRIBUCION DE PROBABILIDAD, HISTOGRAMA, ESPERANZA MATEMATICA, VARIANZA, ESTADISTICOS MUESTRALES, ANALISIS EXPLORATORIO DE DATOS, ROBUSTEZ, RESISTENCIA.

1. INTRODUCCION.

El gran número de datos que inicialmente se consideran en un estudio o investigación, supone a menudo que éstos sean virtualmente inútiles, a menos que no puedan recogerse o describirse de forma concisa.

La estadística habitualmente sugiere, en primer lugar, organizar estos datos en forma de tabla, agrupándolos según un número razonable de categorías. Esta disposición se conoce como distribución de frecuencias. No obstante, la arbitrariedad al determinar la amplitud de las clases, la pérdida de información que se produce al substituir por intérvalos los valores originales, la rigidez del método, y la necesidad ineludible de complementarlo con representaciones gráficas, cuestionan la eficiencia de sus aplicaciones.

En la actualidad, la mayor flexibilidad de la técnica de tabulación en "tronco y hojas", permite llevar a cabo aproximaciones más sistemáticas del tratamiento previo de los datos, revelando con nitidez las características más sobresalientes de una distribución de frecuencias.

Dado que solo la distribución de probabilidad de la variable considerada contiene toda la información con respecto a ésta, la estadística tradicional, recomienda en segundo lugar, sintetizarla, mediante el valor de un número limitado de estadísticos que determinen la tendencia central, la dispersión y en ocaciones su forma.

Hasta finales de los 70, la mayor parte de los procedimientos estadísticos, íntimamente vinculados con la hipótesis de normalidad, proporcionaban óptimos resultados siempre que este supuesto no fuese violado por los datos a los que se aplicaban. La experiencia sim embargo, demuestra que aún hoy, resulta una práctica inusual el excrutinio previo de los datos para corroborar hipótesis, y que cuando este se lleva a cabo, solo excepcionalmente la situación real se ajusta a la ideal esbozada por los presupuestos establecidos.

Las técnicas recién desarrolladas en el contexto del análisis exploratorio de datos, permiten descubrir patrones de comportamien-

- Joan Manuel Batista i Foguet, Moisés Valls i Colom - Universitat Politecnica de Catalunya - Escola Tecnica Superior d'Enginyers Industrials de Barcelona - Dep. Tecniques Quantitatives de Gestió - Av. Diagonal, 647 Barcelona 08028.

- Article rebut el maig de 1985.

to en los datos, sin presuponer modelo subyacente alguno, y detectar las posibles anomalías o errores que el conjunto inicial incluye "casi seguramente".

En particular, en este artículo insistiremos en la utilización de varios estadísticos en lugar de unos pocos, poniendo especial énfasis en dotarlos de "robustez", respecto de las desviaciones de los supuestos básicos y de "resistencia", frente a la presencia de valores anómalos en el conjunto de datos original.

Por último comentaremos los dos tipos de estimadores robustos más utilizados: Los L y los M estimadores.

A pesar de las precauciones sugeridas en éste, en posteriores artículos, fomentaremos la necesidad de transformar los datos originales cuando la exploración previa así lo aconseje y el escepticismo frente a la mera descripción numérica de las distribuciones en favor de descripciones de tipo gráfico.

2 REPRESENTACION EN TRONCO Y HOJAS.

Si el número de valores distintos de una cierta variable es considerable, tanto en las distribuciones de frecuencias como en las representaciones gráficas clásicas, se pierde cierta información al reagrupar los datos en un reducido número de intérvalos. En cambio, en la representación en tronco y hojas debida a J.W. Tukey /8/ puede aprovecharse toda la información contenida en la variable considerada. En realidad, esta disposición de los datos supone un híbrido entre tabulación y representación gráfica, sirviéndose de valores numéricos para configurar un perfil que permitirá caracterizar la distribución de la variable como en un histograma.

Un antecedente del método en su aspecto gráfico fue el método del Lot-plot ideado por Dorian Shainin /7/ para controlar la calidad en la recepción de piezas en base a un perfil numérico que reproduce la distribución de frecuencias muestral.

La organización de los datos que se propone se basa esencialmente en la idea de que todo número se puede dividir en dos partes, una de ellas más significativa que la otra. La limitada capacidad humana de procesamiento de la información según Miller, /6/, justifica esta separación.

Supongamos a título ilustrativo que nuestro interés se centra en el desempleo del estado español contabilizado en diciembre de 1983, /1/. Las tablas T1 y T2 proporcionan respectivamente los datos relativos al número y al porcentaje de parados de cada provincia. Si consideramos en primer lugar los dos primeros dígitos de esta última variable, la cifra de las decenas (tronco) será evidentemente más significativa que la de las unidades (hojas) y por tanto cada observación nos vendrá representada por sólo dos dígitos.

Por ejemplo, el porcentaje de parados correspondiente a Baleares es del 16,6. Este valor puede seguir uno de los dos procesos siguientes dependiendo de si preferimos aproximar o truncar la observación

			TRONCO	HOJA
16,6	APROXIMACION →	17	1	7
16,6	TRUNCADO →	16	1	6

El diagrama de los datos de la tabla T2 obtenido por aproximación será pues el de la figura F1.

Esta representación puede, sin embargo, mejorarse notablemente al incrementar el número de líneas y disminuir, por tanto, la amplitud de los intérvalos dividiendo cada tronco en dos líneas o ramas: Una que incluyera las hojas del 0 al 4 (*) y la otra, las comprendidas entre 5 y 9 (·), con lo cual se llegaría a la representación de la figura F2.

Se podría de nuevo aumentar el número de ramas subdividiendo cada tronco en 5, según la siguiente partición (Figura F3):

NOMENCLATURA	HOJAS
*	0;1
T	2;3
F	4;5
S	6;7
.	8;9

La codificación de las ramas proviene de la terminología anglosajona y corresponde a la

inicial de los números two, three, four, five, six, seven.

Si aun esta representación pareciese poco ilustrativa, convendría plantearse la selección de una nueva unidad (hoja) que permitiese obtener un mayor número de troncos (Figura F4).

En ocasiones como la presente, los datos incluyen valores excepcionalmente altos (A) o bajos (B) que distorsionarían la brevedad requerida en toda representación. En estos casos, los datos anómalos deben figurar al margen de la tabla junto a la unidad que convenga identificar con las hojas.(Figura F4).

El número total de ramas más adecuado dependerá evidentemente de la aplicación específica, existiendo, sin embargo, criterios objetivos para acotar el máximo N_R.

El método más conocido es quizás el que utilizaban Dixon y Kronmal /2/ para definir el número máximo de clases en la tabulación clásica:

$$N_R = [10 \log_{10} N] \qquad (1)$$

donde N representa el número total de datos y el corchete la función truncado.

En el ejemplo $N_R=[10 \log_{10} 50]=[16,99]=16$

Conociendo la amplitud total y el número máximo de ramas se puede determinar la amplitud mínima de estas I_m

$$I_m = \frac{\text{AMPLITUD TOTAL}}{N_R}$$

En nuestro caso, omitiendo el valor 37,3 correspondiente a Sevilla por ser excesivamente alto, se obtiene:

$$I_m = \frac{26,6 - 7,9}{16} = 1,17$$

Por tanto, cada línea contendrá 2 valores distintos y cada tronco 5 ramas según la representación F3. En resumen, el diagrama en tronco y hojas es una técnica flexible que ofrece tres posibles subdivisiones de la escala decimal, de las que se derivan ramas de distinta amplitud (2,5 ó 10).

Cuando el número de observaciones es reducido, Velleman /9/ sugiere tomar como máximo número de ramas

$$N_R = [2\sqrt{N}] \qquad (2)$$

En el ejemplo $N_R = [2\sqrt{50}] = 14$ y $I_m=1,34$

La propia experiencia nos ha demostrado que en general, este segundo criterio proporciona representaciones mucho más razonables -- que el logaritmico.

Sin embargo, cuando el número de observaciones es muy elevado (superior a 1000 por ejemplo), se podrían obtener valores de N_R excesivamente altos, por lo que recomendamos - utilizar

$$N_R = [2\sqrt{N}] \qquad N \leq 1000$$

$$N_R = [21 \log_{10} N] \qquad N \geq 1000$$

La razón de multiplicar el logaritmo precisamente por 21 reside en conseguir una función N_R contínua en N=1000.

Este criterio equivale a utilizar prácticamente siempre (2) ya que cuando el número de observaciones es tan elevado suele ser preferible reagrupar los datos y representarlos en histogramas.

Una vez representado el diagrama de tronco y hojas se le podrían añadir dos columnas adicionales referentes a la frecuencia absoluta de cada rama y a la localización. Mientras la frecuencia absoluta proporciona el número de observaciones por línea, en la localización se van acumulando en direcciones opuestas las frecuencias absolutas hasta -- llegar a la línea que contiene la mediana u observación que divide a la muestra en dos subconjuntos del mismo tamaño (ver figura F5). Como se verá, esta columna facilita en gran medida la obtención de estadísticos.

Además de la "gran economía de tinta" que supone, el diagrama en tronco y hojas, suministra información referente a la localización del centro de la distribución, al grado de dispersión y simetría del conjunto, a la posible división de los datos en grupos, a la existencia de regiones con gran concentración y zonas intermedias sin observacio-

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

UNIDAD 1 = 1% parados/provincia
0|8 representa 8%

Fig. F1: Diagrama en tronco y hojas del porcentaje de desempleo por provincias (una rama por tronco)

```
0*|
 · | 8 8 9
1*| 0 0 0 1 1 1 2 2 2 2 4
 · | 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 7 7 8 8 8 9 9 9
2*| 0 0 1 2 3 3 4
 · | 6 7
3*|
 · | 7
```

UNIDAD 1 = 1% parados/provincia
0|8 representa 8%

Fig. F2: Diagrama en tronco y hojas del porcentaje de desempleo por provincias (dos ramas por tronco)

```
0*|
 T|
 F|
 S|
 · | 8 8 9
1*| 0 0 0 1 1 1
 T| 2 2 2 2
 F| 4 5 5 5 5 5 5 5 5 5
 S| 6 6 6 6 6 6 6 6 6 7 7
 · | 8 8 8 9 9 9
2*| 0 0 1
 T| 2 3 3
 F| 4
 S| 6 7
 ·|
3*|
 T|
 F|
 S| 7
 ·|
```

UNIDAD 1 = 1% parados/provincia
0|8 representa 8%

Fig. F3: Diagrama en tronco y hojas delporcentaje por provincias (cinco ramas por tronco)

```
07 | 9
08 | 3
09 | 3 9
10 | 3 3 7
11 | 0 3 6 8
12 | 1 4
13 | 5
14 | 6 6 6 6 7
15 | 0 1 2 4 7 8 9
16 | 0 2 3 3 3 4 6 7
17 | 8
18 | 0 3 9
19 | 2 4
20 | 2 3 7
21 | 6
22 | 9
23 | 1 5
24 |
25 | 9
26 | 6
```

UNIDAD 1 = 0,1% parados/provincia
07|9 representa 7,9%
A: 37,3

Fig. 4: Diagrama en tronco y hojas del porcentaje de desempleo por provincias (Una rama por tronco y cambio de unidad).

		u_i	L
0*			
T			
F			
S			
·	8 8 9	3	3
1*	0 0 0 1 1 1	6	9
T	2 2 2 2	4	13
F	4 5 5 5 5 5 5 5 5 5	10	23
S	6 6 6 6 6 6 6 6 6 7 7	11	(11)
·	8 8 8 9 9 9	6	16
2*	0 0 1	3	10
T	2 3 3	3	7
F	4	1	4
S	6 7	2	3
·			1
3*			1
T			1
F			1
S	7	1	1
·			

Fig. F5: Diagrama de la figura F3 conteniendo las frecuencias absolutas y la localización

1.		
2*	0 0	2
T	2 3 3 3	6
F	4 4	8
S	6 7	10
·	8 9 9	13
3*	0 0 0 0 1 1	19
T	3	20
F	4 4 5 5 5	25
S	6 6 6 6 7	30
·	8 8 8 8 8 8 9 9 9 9 9	41
4*	0 0 0 1 1 1 1 1 1 1 1	52
T	2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3	71
F	4 4 4 4 4 5 5 5 5 5 5 5 .5	84
S	6 6 6 6 6 6 7 7 7 7 7	95
·	8 8	(2)
5*	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1	94
T	2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3	66
F	4 4 4 4 5 5 5 5 5	50
S	6 6 6 6 7 7 7	41
·	8 8 8 8 8 8 9 9 9 9	34
6*	1 1 1 1	24
T	2 2 3 3 3 3	20
F	4 4 4 5 5	14
S	6 6 6	9
·	7	6
7*	0	5
T	2 3	4
F		
S		
·		

UNIDAD 1 = 0,1 puntos
2 0 REPRESENTA 2,0
A: 8,3 ; 8,1

Fig. F9: Diagrama en tronco y hojas de las notas finales de estadística (promedio de los dos parciales y "prácticas"). Curso 1983/84. E.T.S.E.I.B.

nes y a la aparición de observaciones excesivamente alejadas del resto o anomalías.

En el ejemplo se observa que el centro se localiza aproximadamente sobre el valor del 16%, que la distribución parece bastante simétrica exceptuando la posible anomalía correspondiente a Sevilla con un porcentaje de parados del 37,3% muy superior al resto.

Un ejemplo muy ilustrativo de la utilidad de la representación en tronco y hojas es el del conjunto de notas obtenidas en junio de 1984 por los alumnos de la E.T.S.E.I.B. en la asignatura de Estadística, después de promediar los dos exámenes parciales e incrementar la parte alicuota de las prácticas (Fig. F9).

La distribución de las notas es simétrica en gran medida si exceptuamos el agujero central que se detecta en la rama del "4". Esto se debe fundamentalmente al incremento adicional conseguido por la realización de las prácticas voluntarias que al añadirse con posterioridad a la nota promedio de los exámenes desvirtúa la forma aproximadamente normal que cabría esperar.

El tronco más frecuente corresponde al que incluye las hojas 50 y 51, lo cual corrobora la anterior interpretación y sugiere que "muchos estudian lo suficiente para solo -- aprobar", de hecho entre 39 y 59 se encuentran el 65% de los alumnos presentados a ambos exámenes, como se desprende de la columna de localizaciones.

Aunque puede estimarse cuantos erraron al evaluar cuanto era "lo suficiente", estos no han ocasionado sesgo alguno en la distribución. Seguramente, ello obedece a que la nota es un promedio de dos pruebas que además están espaciadas en el tiempo, y por tanto el alumno ha podido optar por retirarse, trabajar más o reducir el número de horas que venía dedicando a la asignatura.

3. SINTESIS NUMERICA DE UN CONJUNTO DE DATOS

3.1 INTRODUCCION

Es sabido que el conocimiento exacto acerca de una variable aleatoria requiere disponer de la función de distribución de probabilidad. Esta, incluye parámetros que suelen coincidir con la esperanza matemática, la variancia u otros momentos de la distribución.

Al introducir la tabulación en tronco y hojas, hemos resaltado que la distribución de frecuencias permite describir íntegramente el conjunto de datos muestrales. Sin embargo, en la práctica, unos pocos estadísticos - contrapartidas muestrales de aquellos parámetros - definidos convenientemente son quienes caracterizan numéricamente la posición, dispersión y a veces la forma de la distribución muestral.

Los estadísticos clásicos y en general toda la estadística "normal" ha sido diseñada para comportarse lo mejor posible siempre que se cumplan ciertos supuestos. No obstante, las situaciones en las que nos encontramos habitualmente difieren de la ideal que las asunciones establecen, por lo que los estadísticos tradicionales deben substituirse por otros que muestren mayor insensibilidad frente a datos anómalos (RESISTENCIA) y respecto a desviaciones de la normalidad (ROBUSTEZ).

i) En primer lugar, la síntesis que se efectúa de la distribución, considera la localización de su centro o valor típico alrededor del cual se agrupan los restantes. -- Tradicionalmente ello se ha indicado mediante valores promedio que caracterizan los datos. Sin embargo, esta descripción concisa del conjunto padece ciertas deficiencias -- que deterioran e incluso pueden invalidar el papel descriptivo de estos resúmenes numéricos.

Por un lado, al referirnos al centro se piensa en el centro de simetría de la distribución. Consecuentemente las distribuciones asimétricas carecen de localización natural en un punto determinado.

Por otro lado, aun cuando la distribución sea simétrica, la media aritmética ($\bar{x}$) se distorsiona fuertemente con la presencia de de datos anómalos o errores, ambos muy frecuentes en estudios exploratorios.

ii) Una vez identificado el valor más característico alrededor del cual se agrupa

el resto, se acostumbra a determinar la dispersión de los datos evaluando numéricamente la fluctuación de los mismos respecto del centro.

La medida más ampliamente utilizada por la estadística descriptiva clásica para caracterizar la variabilidad es sin duda la variancia muestral (s^2). Sin embargo, al ser ésta básicamente una media aritmética de desviaciones cuadráticas con respecto al valor central, estará aun más afectada por los problemas anteriores que dichos promedios.

iii) Además de los estimadores de los parámetros de localización y de dispersión utilizados para resumir la información incluida en la distribución de frecuencias de una variable, es posible definir otro tipo de medidas cuyo objetivo es sintetizar la forma de dicha distribución. El hecho de que hayan sido virtualmente omitidas de la mayor parte de los estudios estadísticos es debido al papel preponderante que en la inferencia tiene la distribución normal. (Es sabido que la función matemática de esta ley de probabilidad depende solo de dos parámetros que precisamente caracterizan su centro y su dispersión, pues tanto la simetría como el grado de apuntamiento son conocidos a priori).

3.2 INDICADORES DE LOCALIZACION.

3.2.1. Estadísticos de orden. L-estimadores

Para sintetizar la información que la distribución de frecuencias incluye, el análisis exploratorio de datos insiste en la utilización de varios estadísticos poniendo especial énfasis en su resistencia y robustez. Para ello se propone la utilización de estadísticos de orden insensibles a la presencia de valores anómalos.

Supongamos a partir de ahora que la reordenación ascendente según la magnitud de los valores muestrales viene caracterizada por subíndices entre paréntesis:

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(N)}$$

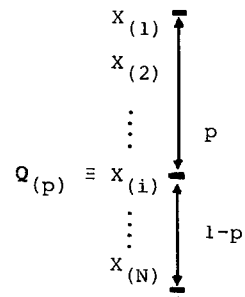
En estas condiciones, se define el i-ésimo estadístico de orden como el valor $x_{(i)}$. Esto permite caracterizar el centro refiriéndolo al valor de la mediana M que parte a la muestra en dos mitades. Obviamente, su cálculo es función de la paridad del número total de observaciones N.

$$N \text{ impar} \longrightarrow M = X_{\left(\frac{N+1}{2}\right)}$$

$$N \text{ par} \longrightarrow M = \frac{1}{2}\left(X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}\right)$$

En esta línea, del mismo modo que la mediana divide a la muestra en dos subconjuntos, las cuartilas la dividen en cuatro a los que corresponde una frecuencia del 25%, las quintilas en cinco del 20% y así sucesivamente, decilas, centilas, etc. En general, la división en un número de partes cualquiera determina las cuantilas.

Concretamente, definimos la cuantila $Q_{(p)}$, correspondiente a la fracción p, de una muestra ordenada, como aquel valor de la variable $X_{(i)}$ que divide al conjunto de datos en dos grupos, de forma que una fracción p se encuentre por debajo y la otra fracción (1-p) por encima.



Desafortunadamente esta definición plantea ciertos problemas operacionales cuando tratamos de calcular las cuantilas de un conjunto de datos. Así por ejemplo, si tenemos 20 observaciones, podemos hallar el valor que representa la cuantila de fracción $p_1 = 0,1$ ó $p_2 = 0,2$ pero no los correspondientes a fracciones comprendidas entre p_1 y p_2. Por ello, asignaremos el valor $x_{(i)}$ a la cuantila $Q_{(p)}$ siempre que p sea una de las fracciones

$$p_i = \frac{i - 0,5}{N} \qquad i = 1,2,\ldots,N \qquad (3)$$

pues definir p_i como el cociente i/N causa algunos inconvenientes: Por un lado, ambas colas de la distribución recibirán un tratamiento asimétrico y por otro lado, al va-

lor extremo $X_{(N)}$ le corresponderá una fracción p=1, lo cual no es deseable pues significará que la distribución de la variable finaliza en este valor muestral.

En cambio, mediante la expresión (3), una mitad de cada observación pertenece al grupo inferior ($< x_i$) i la otra mitad al superior ($> x_{(i)}$).

Así, si nos interesa saber entre que valores del porcentaje de paro se encontraría el 80% de las provincias, deberíamos determinar la primera (D_1) y novena (D_q) decilas:

$$P_i = 0,10 = \frac{i - 0,5}{50} \Rightarrow i = 5,5$$

En este caso, corresponden a las observaciones situadas en el punto medio entre la quinta y la sexta contadas a partir de cada uno de los extremos de la muestra ordenada:

$$D_1 = Q(0,10) = 10\%$$

$$D_9 = Q(0,90) = 23\%$$

La definición de cuantila (3) es susceptible de extenderse por simple interpolación a cualquier otro valor de p: Si θ representa una fracción del recorrido entre p_i y p_{i+1}, entonces

$$Q(p) = 1-\theta)\ Q(p_i) + \theta\, Q(p_{i+1})$$

En el ejemplo de las notas, las decilas anteriores se obtendrían de la siguiente forma:

$$P_i = 0,10 = \frac{i-0,5}{191} \Rightarrow i = 19,6 \Rightarrow \theta = 0,6$$

$$D_1 = (1-0,6)Q(p_{19}) + 0,6\ Q(P_{20}) = 3,22$$

$$D_9 = (1-0,6)Q(p_{172}) + 0,6\ Q(p_{173}) = 6,2$$

La generalidad de esta definición permite considerar como casos particulares gran parte de las medidas de localización mencionadas. Entre ellas, las cuartilas son sin duda las que se utilizan con mayor asiduidad:

Q_1, la primera cuartila se define como el valor de la variable para el cual el 25% de las observaciones se sitúan a su izquierda y el 75% a su derecha.

Q_2, la segunda cuartila coincide con el valor de la mediana M.

Q_3, la tercera cuartila se define como el valor por encima del cual se encuentran el 25% de las observaciones.

Como la frecuencia relativa toma sucesivos valores discretos que se diferencian en 1/N, las definiciones anteriores no siempre determinan un valor único, utilizándose varias estrategias para establecerlo.

J.W. Tukey /8/ propuso un método para calcular de forma expeditiva y aproximada las observaciones que dividen a la muestra en K partes, siendo K un múltiplo de 2. En esencia el procedimiento es una extensión del utilizado para la obtención de la mediana M. Concretamente, para obtener las cuartilas Q_1 y Q_3, puntos medios del recorrido entre la mediana y los respectivos extremos, procedemos según la expresión

$$L(Q_i) = \frac{[L(M)] + 1}{2} \quad \text{con} \quad i=1 \text{ ó } 3$$

donde la función L nos indica la localización o lugar que ocupa una observación en la muestra ordenada a partir del extremo más próximo, y el corchete, como antes, la función truncado.

El concepto genérico de cuartila nos sugiere definir otros estadísticos tales como el promedio de cuartilas ($\bar{Q}$) o la trimedia (TRI) que constituyen estimadores resistentes y robustos de la tendencia central.

$$\bar{Q} = \frac{1}{2}\ (Q_1 + Q_3)$$

$$TRI = \frac{1}{2}\ (M + \bar{Q}) = \frac{1}{4}\ (Q_1 + 2M + Q_3)$$

El hecho de utilizar otros estadísticos distintos de la mediana se debe a que ésta es muy sensible a errores de truncado redondeo o agrupamiento de las observaciones.

Generalizando, se definen los L-estimadores como combinaciones lineales de estadísticos de orden:

$$T = \sum_{i=1}^{N} c_i x_{(i)} \qquad (4)$$

El lector puede identificar sin dificultad los coeficientes c_i correspondientes a los estadísticos introducidos anteriormente.

Como quiera que gran parte de los promedios considerados en este apartado responden al concepto de media recortada (trimmed mean) convendría por último referirnos a este concepto genérico. Las medias recortadas $T(\alpha)$, representan un caso particular de (4) y consisten en eliminar una proporción α de observaciones extremas por ambos lados y promediar las restantes.

Si $g = \alpha N$ es un entero, $2g$ representará el número de valores eliminados y por tanto

$$T(\alpha) = \frac{1}{N-g} \sum_{i=g+1}^{N-g} x_{(i)}$$

Si αN no es entero, la fórmula anterior se puede generalizar

$$T(\alpha) = \frac{1}{N(1-2\alpha)}\left((1-\gamma)(X_{(g+1)}+X_{(N-g)}) + \sum_{i=g+2}^{N-g-1} X_{(i)}\right)$$

donde $g = [\alpha.N]$ y $\gamma = \alpha.N-g$

Como casos particulares de medias recortadas podemos citar

Media aritmética ($\bar{X}$)	$T(0)$
Mediana (M)	$T(\alpha)$ con $\alpha \simeq 0,5$
Midmean	$T(0,25)$

Utilizando una amplia gama de distribuciones Hoaglin /4/ compara todos los estadísticos de localización citados, llegando a la conclusión de que el que mejores resultados proporciona es precisamente el Midmean.

3.2.2. INTRODUCCION A LOS M-ESTIMADORES.

Los L-estimadores son una herramienta muy útil ya que son estadísticos resistentes y robustos fáciles de calcular. Sin embargo, suelen sacrificar resistencia para dar un cierto peso a observaciones anómalas o bien por el contrario, pierden eficiencia dando poco peso a observaciones extremas pero no anómalas. Los M-estimadores en cambio, están basados en la idea de los estimadores máximo-verosímiles.

En efecto, sea $X_1, X_2, \ldots, X_N$ una muestra aleatoria de una variable aleatoria cuya familia de densidades es $\{f(X-T) \mid T \in \mathbb{R}\}$ donde T es un parámetro de la tendencia central y f es contínua. La densidad de probabilidad conjunta de

$$\underline{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \quad \text{dado T sería } f(\underline{x}|T) = \prod_{i=1}^{N} f(x_i - t) \text{ que}$$

considerada como función de T, dados $x_1, x_2, \ldots x_N$ será la función de verosimilitud

$$L(T|\underline{x}) = \prod_{i=1}^{N} f(x_i - T)$$ cuyo logaritmo neperiano será

$$\ell n\, L(T|\underline{x}) = \sum_{i=1}^{N} \ell n f(x_i - T) = -\sum_{i=1}^{N} P(x_i - T)$$

siendo $P(x) = -\ell n f(x)$

La estimación máximo-verosimil de T consiste en maximizar $\ell n L(T|\underline{x})$ o equivalentemente en minimizar

$$K(T) = \sum_{i=1}^{N} P(x_i - T)$$

Si $P(x)$ es diferenciable, la solución se obtendría haciendo

$K'(T) = 0$ es decir,

$$\sum_{i=1}^{N} P'(x_i - T) = \sum_{i=1}^{N} \frac{-f'(x_i - T)}{f(x_i - T)} = \sum_{i=1}^{N} \psi(x_i - T) = 0 \quad (5)$$

donde $\psi(x) = P'(x) = \frac{-f'(x)}{f(x)}$

El valor de T que cumple (5) es el estimador máximo-verosímil de T y se nota por T_N:

$$\sum_{i=1}^{N} \psi(x_i - T_N) = 0$$

Si se define $w_i = \frac{\psi(x_i - T_N)}{x_i - T_N}$ se cumplirá

$$\sum_{i=1}^{N} w_i (x_i - T_N) = 0 \text{ y por tanto } T_N = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$

será una combinación lineal de las observaciones con pesos dependientes de la muestra.

Para obtener los pesos w_i hemos de escoger una cierta función $\psi(x)$ que parezca razonable, para lo cual vamos en primer lugar a observar las funciones $\psi(x)$ asociadas a distintas leyes de probabilidad

-Ley normal $f(x|T) =$

$$= \frac{1}{\sqrt{2\Pi}\sigma} \exp\left(\frac{-1}{2\sigma^2}(x-T)^2\right) \quad -\infty<x<\infty$$

$$\psi(x)=x \;,\; \sum_{i=1}^{N} \psi(x_i-T_N)=0 \Rightarrow \sum_{i=1}^{N}(x_i-T_N)=0$$

(Fig.F6)

$$T_N = \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Ley exponencial doble $f(x|T) =$

$$= \frac{1}{2}\exp(-|x-T|) \qquad -\infty<x<\infty$$

$$\psi(x) = \begin{cases} -1 & x<0 \\ 1 & x>0 \end{cases} \qquad \sum_{i=1}^{N}\psi(x_i-T_N) = 0$$

(Fig.F7) $\qquad T_N = M$ (mediana)

Para obtener estimadores independientes de las unidades es conveniente modificar la definición de M-estimador en el sentido

$$\sum_{i=1}^{N} \psi\left(\frac{x_i-T_N}{s}\right) = 0$$

donde s es un estimador robusto y resistente de la dispersión.

Basándose en la idea de obtener una función $\psi(x)$ asociada a una distribución normal en el centro pero con distribución más alargada en las colas (doble exponencial), Huber /5/ propone la función (Fig. F8).

$$\psi(x)' = \begin{cases} -a & x<-a \\ x & |x|\leq a \\ a & x>a \end{cases}$$

La determinación de a depende de la eficiencia que se desee para el caso de normalidad:

Si $a = 0 \longrightarrow T_N = M$

Si $a\to\infty \longrightarrow T_N = \bar{x}$

En general, la solución de $\sum_{i=1}^{N} \psi\left(\frac{x_i-T_N}{s}\right) = 0$ suele hallarse iterativamente.

Evidentemente, la función $\psi(x)$ podría ser diferente de las anteriormente citadas y de hecho, numerosos autores tales como Hampel /3/, Tukey /8/, etc. han propuesto distintas funciones $\psi(x)$ que proporcionarán por tanto, distintos estimadores T_N. Así, por ejemplo, se podría pensar en funciones $\psi(x)$ asociadas a distribuciones con colas pesadas como es el caso de la distribución de Cauchy:

$$f(x) \frac{1}{\Pi(1+x^2)} \qquad -\infty<x<\infty$$

que presenta una función $\psi(x) = \frac{2x}{1+x^2}$

3.3 INDICADORES DE DISPERSION

Los estadísticos clásicos de dispersión tales como la desviación tipo s o la desviación media d_N definidos por

$$s = +\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{N-1}} \qquad d_N = \frac{1}{N}\sum_{i=1}^{N}|x_i-\bar{x}|$$

pierden eficiencia rápidamente al pasar de la distribución normal a distribuciones con colas más alargadas. En cambio, existen otros estadísticos robustos y resistentes que sin ser tan eficientes ante distribuciones normales consiguen ser bastante adecuados frente a una amplia gama de distribuciones. Destacamos entre ellos los dos siguientes:

AMPLITUD INTERCUARTILICA: $IQR = Q_3-Q_1$, siendo Q_1 y Q_3 la primera y tercera cuartila respectivamente.

MEDIANA DE LAS DESVIACIONES ABSOLUTAS: $MAD = \underset{i}{\text{Mediana}}\{|x_i-M|\}$, donde M es la mediana de las x_i.

Ambos pueden estandarizarse dividiendo por el valor correspondiente de la ley normal estandarizada pasando a denominarse pseudo-desviaciones tipo por su coincidencia con la desviación tipo en caso de distribución normal.

$$PSD_{IQR} = \frac{IQR}{1,349}$$

$$PSD_{MAD} = \frac{MAD}{0,6745}$$

En ocasiones es deseable disponer de alguna medida de dispersión relativa que mida las fluctuaciones de la variable, independientemente de la escala utilizada para ello. El denominado coeficiente de variación

$$cv = s/\bar{x}$$

es el estadístico más conocido de dispersión

relativa. Sin embargo, es un claro ejemplo de estadístico poco robusto por lo que proponemos sustituirlo en la etapa exploratoria por

$$cv_Q = \frac{1/2\ IQR}{\bar{Q}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

obtenido al sustituir numerador y denominador de cv por estadísticos más robustos y resistentes.

3.4. INDICADORES DE FORMA: SIMETRIA Y APUNTAMIENTO.

Tradicionalmente el grado de simetría y de apuntamiento de una distribución se han obtenido respectivamente a partir de los estadísticos

$$H = \frac{\sum_{i=1}^{N} ((x_i - \bar{x})/s)^3}{N}$$

$$K = \frac{\sum_{i=1}^{N} ((x_i - \bar{x})/s)^4}{N} - 3$$

que toman el valor H = K = 0 en caso de normalidad.

Sin embargo, la resistencia de estos coeficientes es prácticamente nula, por lo que no nos son útiles en el análisis exploratorio de los datos. Para solventar la falta de resistencia, Yule definió el coeficiente de simetría:

$$H_1 = \frac{Q_1 + Q_3 - 2M}{2M}$$

que toma valores mayores o menores que cero según el sesgo de la distribución sea a la derecha o a la izquierda. Basándose en cuantilas más extremas (decilas), Kelley propuso

$$H_2 = M - \frac{1}{2}(D_9 + D_1)$$

Nosotros proponemos utilizar ambos estadísticos, ya que proporcionan información complementaria. Mientras el primero considera la simetria en el centro de la distribución, el segundo examina las colas, lo cual permitiría distinguir ambos tipos de asimetría. Para homogenizar ambos coeficientes sería conveniente transformar el de Kelley en

$$H_3 = -H_2/M = \frac{D_1 + D_9 - 2M}{2M}$$

que será también adimensional como H_1

En cuanto al grado de apuntamiento sugerimos utilizar, empleando de nuevo cuartilas y decilas, el estadístico

$$D_2 = \frac{D_9 - D_1}{1{,}9(Q_3 - Q_1)}$$

que en caso de normalidad toma el valor -- $K_2 = 1$, con lo cual diremos que la distribución es más o menos apuntada respecto de la normal según K_2 tome un valor superior o inferior a 1.

3.4 APLICACION PRACTICA.

Para finalizar, hemos efectuado una síntesis numérica del ejemplo de las notas, obteniendo los siguientes resultados:

ESTADISTICOS DE LA TENDENCIA CENTRAL:

M = 4,8
TRI = 4,78
$\bar{Q}$ = 4,75
T(0,25)=4,76
T(0,1)= 4,76
$\bar{X}$ = 4,75

ESTADISTICOS DE DISPERSION:

IQR = 1,3 PSD_{IQR} = 0,96
MAD = 0,7 PSD_{MAD} = 1,04
S = 1,13

cv = 0,24
cv_Q = 0,14

ESTADISTICOS DE FORMA:

H_1 = 0,01
H_3 = 0,02

K_2 = 1,21

En resumen, los estadísticos de la tendencia central son todos muy similares y próximos a 4,75, los de dispersión son también coincidentes (desviación y pseudo desvia---

TABLA T1

Nº de parados en millares

50 La Coruña
11 Lugo
13 Orense
38 Pontevedra
58 Asturias
23 Cantabria
14 Alava
70 Vizcaya
35 Guipúzcoa
26 Navarra
5 Avila
17 Burgos
18 Leon
8 Palencia
13 Salamanca
4 Segovia
2 Soria
21 Valladolid
2 Zamora
35 Zaragoza
6 Huesca
3 Teruel
7 La Rioja
301 Barcelona
24 Tarragona
10 Lérida
16 Gerona
28 Baleares
144 Valencia
17 Castellón
66 Alicante
42 Murcia
19 Cáceres
43 Badajoz
15 Albacete
26 Ciudad Real
4 Cuenca
6 Guadalajara
18 Toledo
207 Madrid
22 Almería
55 Cádiz
46 Córdoba
36 Granada
27 Huelva
23 Jaen
68 Málaga
130 Sevilla
50 Las Palmas
35 Sta Cruz de Tenerife

TABLA T2

Porcentaje de desempleo
(% de la población activa)

16,39 La Coruña
7,94 Lugo
9,91 Orense
14,64 Pontevedra
16,28 Asturias
14,56 Cantabria
15,84 Alava
16,2 Vizcaya
15,36 Guipúzcoa
19,45 Navarra
10,73 Avila
15,19 Burgos
12,11 Leon
18,3 Palencia
16,31 Salamanca
10,33 Segovia
11,58 Soria
14,63 Valladolid
13,51 Zamora
14,99 Zaragoza
11,34 Huesca
8,30 Teruel
12,35 La Rioja
21,63 Barcelona
15,9 Tarragona
9,25 Lérida
11,8 Gerona
16,62 Baleares
17,96 Valencia
11,0 Castellón
16,75 Alicante
15,7 Murcia
17,8 Cáceres
19,22 Badajoz
20,30 Albacete
20,22 Ciudad Real
10,29 Cuenca
16,04 Guadalajara
14,58 Toledo
15,16 Madrid
18,94 Almeria
23,13 Cádiz
25,85 Córdoba
21,93 Granada
26,63 Huelva
16,32 Jaén
23,55 Málaga
37,25 Sevilla
20,67 Las Palmas
14,75 Sta. Cruz Tenerife

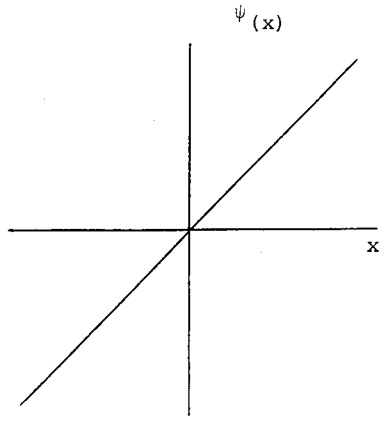


Fig. F6: Función $\psi_{(x)}$ asociada a la ley normal

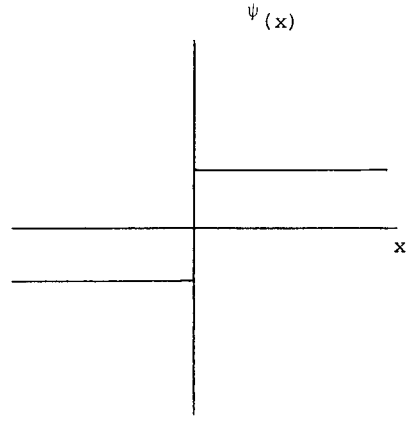


Fig. F7: Función $\psi_{(x)}$ asociada a la ley exponencial doble

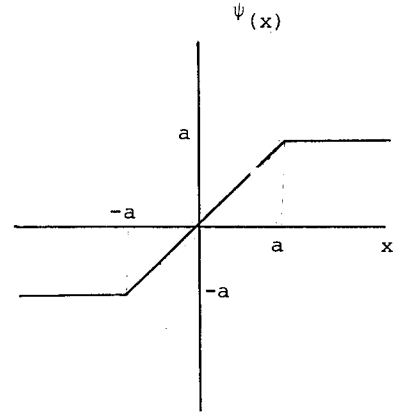


Fig. F8: Función $\psi_{(x)}$ propuesta por Huber.

ción tipo próximos a 1), mientras que los indicadores de forma muestran por un lado una leve asimetría con sesgo a la izquierda, tanto en el centro como en las colas, y por otro, un cierto apuntamiento de la distribución.

4. BIBLIOGRAFIA.

/1/ ANUARIO DEL PAIS - Promotora de Informaciones, S. A. (1984).

/2/ DIXON, W.J., KRONMAL, R.A.:"The choice of origin and scale graphs" . Journal of the Association for Computer Machinery, 12, 259-261. (1965).

/3/ HAMPEL, F.R.: "Contributions to the Theory of Robust Estimation", Ph.D. Dissertation Univ. of California. (1978).

/4/ HOAGLIN, D.C.; MOSTELLER, F.; TUKEY, J.W. "Understanding Robust and Exploratory Data Analysis", J. Wiley & Sons Inc., (1983).

/5/ HUBER, D.R.: "Robust Statistics: A Re view".Annals of Math. Stat. 43: 1041-1067. (1975).

/6/ MILLER, G.A.: "The Magical Number Seven Plus or Minus Two". Psychological Review, nº 63: 81-97.

/7/ SHAININ, D.: "The Hamilton Standard Lot Plot Method of Acceptance Sampling by Variables Industrial Quality Control". (1950).

/8/ TUKEY, J.W.: " Exploratory Data Analysis" Addison-Wesley. (1977).

/9/ VELLEMAN, P.F.: "Interactive Computing for Exploratory Data Analysis I: Display Algorithms". (1976).
Proceedings of the Statistical Computing Section. Washington DC: American Statistical Association. (1975).