Qüestiió - V. 9, n.º 3 (setembre 1985) pp. 163-176

TÉCNICAS GRÁFICAS EN ANÁLISIS EXPLORATORIO DE DATOS

JOAN MANUEL BATISTA I FOGUET, MOISÉS VALLS I COLOM
UNIVERSIDAD POLITÉCNICA DE CATALUNYA

En este artículo se plantea la necesidad de utilizar representaciones gráficas, antes, durante y después del análisis numérico de los datos. En particular, se sugieren dos gráficos: el de cuantilas y el de caja, señalando, en aplicaciones con datos reales, sus ventajas respecto a los procedimientos clásicos. Además, se propone una posible metodología, que permite generalizar el uso exploratorio del gráfico de caja para tratar datos procedentes de distribuciones no normales.

Keywords: ANALISIS EXPLORATORIO DE DATOS, ESTADISTICOS MUESTRALES, HISTOGRAMA, TABULACION EN TRONCO Y HOJAS, RESISTENCIA, ROBUSTEZ.

1. INTRODUCCION.

Desde que Playfair /6/, empezó a utilizar -- gráficos para observar datos, muchos han sido sus detractores /3/, /4/, /5/. En la actualidad, sin embargo, este tipo de representaciones resultan imprescindibles para analizar o comunicar información en todas las disciplinas.

La percepción en gráficos "decodificación visual de información codificada en gráficos" /2/, optimiza la capacidad de nuestro sofisticado sistema de procesamiento de la información. Ya advertimos /1/ que muchos procedimientos estadísticos para resumir información, en lugar de revelar, obscurecían cualquier estructura para cuyo tratamiento no estuviesen específicamente diseñados. Por el contrario, la información que se deriva de un gráfico es siempre de tipo cualitativo. De hecho, algunos de los gráficos que veremos representan un retrato más que un resumen de la distribución de los datos, por lo que en general se obtiene de ellos mayor conocimiento del que se derivaría de un estudio exclusivamente numérico. Usualmente, los gráficos mejorarán significativamente el posterior análisis estadístico.

La finalidad de un gráfico puede residir en registrar, almacenar y comunicar información, o en analizar datos para aprender más sobre ellos. Este artículo hace referencia al último de estos objetivos, poniendo el énfasis - en dos aspectos fundamentales de la utilización de los métodos gráficos en Estadística.

En primer lugar, para contrarrestar la tradicional identificación que se hace del análisis de datos con el cálculo de estadísticos, hacemos especial hincapié en que la representación visual debe preceder inevitablemente al análisis numérico.

En segundo lugar, ya que implícita o explícitamente, la mayor parte de los procedimientos estadísticos habituales se basan en supuestos acerca de la distribución de los datos de cuya veracidad depende la validez del análisis, los gráficos constituirán poderosas herramientas de diálogo para verificar asunciones previas y sugerir las oportunas correcciones cuando éstas hayan sido violadas.

Ya que suponemos conocidas las ventajas relativas de que gozan los diagramas de barras, los de puntos, los histogramas y otros métodos tradicionales para representar la distri

- Joan Manuel Batista Foguet - Moisés Valls Colom - Universitat Politècnica de Catalunya - E.T.S.E.I.B. Dep. de Tècniques Quantitatives de Gestió.

- Article rebut el setembre de 1985.

bución de las observaciones, referiremos aquí exclusivamente dos herramientas específicas del Análisis Exploratorio de Datos Univariante: El Gráfico de Cuantilas y el de caja (Box and Whiskers plot /7/).

2. GRAFICO DE CUANTILAS.

2.1. INTRODUCCION.

Las deficiencias que padecen los procedimientos clásicos de representación gráfica son de diversa índole. Al histograma, principalmente se le ha criticado la pérdida de información que supone substituir el verdadero valor de las observaciones por los extremos de un intérvalo y la selección arbitraria que se efectúa en la amplitud de las clases que constituyen la partición del espacio muestral. Por otro lado, los diagramas de puntos que no tienen esos inconvenientes, dejan de ser útiles cuando la concentración de los datos es elevada y sólo con dificultad se logra descongestionarlos.

El gráfico de cuantilas es un instrumento de fácil construcción que posibilita la solución de los tres problemas mencionados. Básicamente consiste en dibujar en unos ejes coordenados los pares:

$$(p_i, Q(p_i))$$

siendo $Q(p_i)$ la cuantila correspondiente a la fracción p_i /1/. En realidad, salvo un cambio de escala del eje de abcisas coincide con el gráfico

$$(i, x(i))$$

siendo x(i), la iésima observación de la muestra ordenada.

Ya que esta representación no presupone modelo alguno para los datos, ni se efectuan arbitrarias agrupaciones de las observaciones, representándose, por ende, todos los puntos, -- aunque éstos se repitiesen, el resultado configura un retrato de los datos más que un resumen de los mismos.

El PNB per cápita de los 180 paises relacionados en la tabla T1, se utiliza a continuación para destacar alguna de las ventajas que se derivan de este tipo de representación, ya que en esencia el método consiste en dibujar observaciones ordenadas resulta sencilla su programación en ordenador aunque sea de reducida capacidad. Sin embargo, la confección manual del gráfico no requiere demasiado tiempo y proporciona una conveniente visión - preliminar de los datos.

A partir de las fracciones, p_i, pueden identificarse algunas cuantilas típicas: (ver fig. F1).

$Q_1 = Q(.25) \simeq 400$ dólares per capita
$M = Q(.50) \simeq 1400$ dólares per capita

y determinar rápidamente otros estadísticos de interés, calculados a partir de las cuantilas como por ejemplo la amplitud intercuartilica.

$$IQR = Q_3 - Q_1$$

En nuestro ejemplo una simple mirada al gráfico nos revela que la mayor concentración - se encuentra por debajo de M = 1400 dólares, y que por encima de 15000 dólares encontramos 6 paises excepcionalmente ricos cuyas causas no son difíciles de descubrir (QATAR, -- EMIRATOS ARABES UNIDOS, KUWAIT, SUIZA, BRUNEI, LUXEMBURGO).

La densidad local o concentración de datos viene indicada por la pendiente local del gráfico de cuantilas. Así, cuanto menos pronunciada sea esta pendiente, mayor es la concentración de puntos alrededor del valor correspondiente a la cuantila. En el límite, una serie horizontal de puntos representaría aquellos individuos con idéntico valor en la variable que se analiza.

En el ejemplo, debido a que apreciamos diferencias de un dólar en el PNB per cápita, difícilmente observaremos secuencias horizontales. No obstante, se hace patente la homogeneidad característica del 25% de países pauperrimos, que contrasta con la heterogeneidad conseutudinaria al 25% de los países más ricos. La impresión general es que la concentración de países decrece al aumentar su PNB per cápita.

El gráfico de cuantilas, no es un concepto novedoso. Si se intercambian sus ejes coordena-

Tabla T1: PNB per capita de 180 paises

	Países	dólares		Países	dólares
1	Qatar	27720	53	Samoa (EE.UU)	4170
2	Em Arab Un	24660	54	Bulgaria	4150
3	Kwait	20900	55	Polonia	3900
4	Suiza	17430	56	Reunion (FR)	3840
5	Bruner	17380	57	Gabon	3810
6	Luxemburgo	15910	58	Chipre	3740
7	Suecia	14870	59	Bahamas	3620
8	Noruega	14060	60	Malta	3600
9	Alemania R F	13450	61	Barbados	3500
10	Dinamarca	13120	62	Guyana franc.	3430
11	Bermudas	12910	63	Puerto Rico	3350
12	Islandia	12860	64	Sunnam	3030
13	Estados Unidos	12820	65	Irak	3020
14	Arabia Saudí	12600	66	Uruguay	2820
15	Islas Canal (RU)	12430	67	Yugoslavia	2790
16	Francia	12190	68	Surafrica	2770
17	Bélgica	11920	69	Macao (PORT)	2630
18	Holanda	11790	70	Argentina	2560
19	Canadá	11400	71	Chile	2560
20	I.Perces (DIN)	11100	72	Rumania	2540
21	Australia	11090	73	Portugal	2520
22	Groent (DIN)	10850	74	México	2250
23	Finlandia	10680	75	Brasil	2220
24	Austria	10210	76	Argelia	2140
25	Japón	10080	77	Hungria	2100
26	Reino Unido	9110	78	Fiyi	2000
27	Bahrem	8960	79	Namibia	1960
28	Libia	8450	80	Panama	1910
29	Nueva Zelanda	7700	81	Malasia	1840
30	Alemania R D	7180	82	Seychelles	1800
31	N Caledonia	7100	83	Corea del Sur	1700
32	Islas Vírgenes	7010	84	Montserrat	1640
33	Polinesia (FR)	6980	85	Paraguay	1630
34	Italia	6969	86	Jordania	1620
35	Guam (EE.UU)	6840	87	Siria	1570
36	Oman	5920	88	Antigua	1550
37	Checoslov.	5820	89	Turquia	1540
38	Trin y Tobago	5670	90	Costa Rica	1430
39	España	5640	91	Tunicia	1420
40	Isla Man (RU)	5390	92	Colombia	1380
41	Singapur	5240	93	Mauricí-	1270
42	Irlanda	5230	94	Rep. Dominc.	1260
43	Israel	5160	95	Costa de Marfil	1200
44	Hong Kong (RU)	5100	96	Ecuador	1180
45	Martinica (FR)	4820	97	Jamaica	1180
46	Gibraltar	4690	98	Islas Cook (NZ)	1170
47	URSS	4550	99	Perú	1170
48	Antillas Hol.	4540	100	Guatemala	1140
49	Grecia	4420	101	Congo	1110
50	Guadalupe	4340	102	Belice	1080
51	Venezuela	4220	103	Nive (NZ)	1080
52	Hungría	4180	104	S. Cristóbal	1040

	Países	dólares
105	Wallis Fortuna	1020
106	Botsuana	1010
107	Isl. Pacífico	1000
108	Santa Lucía	970
109	Camerun	880
110	Nigeria	870
111	Zinbabue	870
112	Marruecos	860
113	Nicaragua	860
114	Granada	850
115	Papua N.G.	840
116	Filipinas	790
117	Tailandia	770
118	Sauzilandia	760
119	Dominica	750
120	Guyana	720
121	Tuvaru	680
122	Tokelau	670
123	Egipto	650
124	El Salvador	650
125	Islas Salomon	640
126	San Vicente	630
127	Bolivia	600
128	Honduras	600
129	Zambia	600
130	Lesoto	540
131	Indonesia	530
132	Tonga	530
133	Liberia	520
134	Yibuti	480
135	Angola	470
136	Mauritania	460
137	Yemen Norte	460
138	Yemen del Sur	460
139	Senegal	430
140	Kenia	420
141	Kirbuti	420
142	Ghana	400
143	Sudan	380
144	Togo	380
145	Gambia	370
146	S Tome Princ.	370
147	Pakistan	350
148	Vanatu	350
149	Cabo Verde	340
150	Niger	330
151	Rep Malgache	330
152	Benin	320
153	Centroafrica	320
154	Còmores	320
155	Sierra Leona	320
156	China	300
157	Guinea	300
158	Haiti	300
159	Sn Lanka	300
160	Somalia	280
161	Tanzania	280
162	Mozambique	270
163	India	260
164	Maldivas	260
165	Ruanda	250
166	Alto Volta	240
167	Burundi	230
168	Uganda	220
169	Zaire	210
170	Malavi	200
171	Birmania	190
172	Guinea Bissau	190
173	Mali	190
174	Guinea Ecuat.	180
175	NEPAL	150
176	Bangladesh	140
177	Etiopia	140
178	Chad	110
179	Butan	80
180	Laos	80

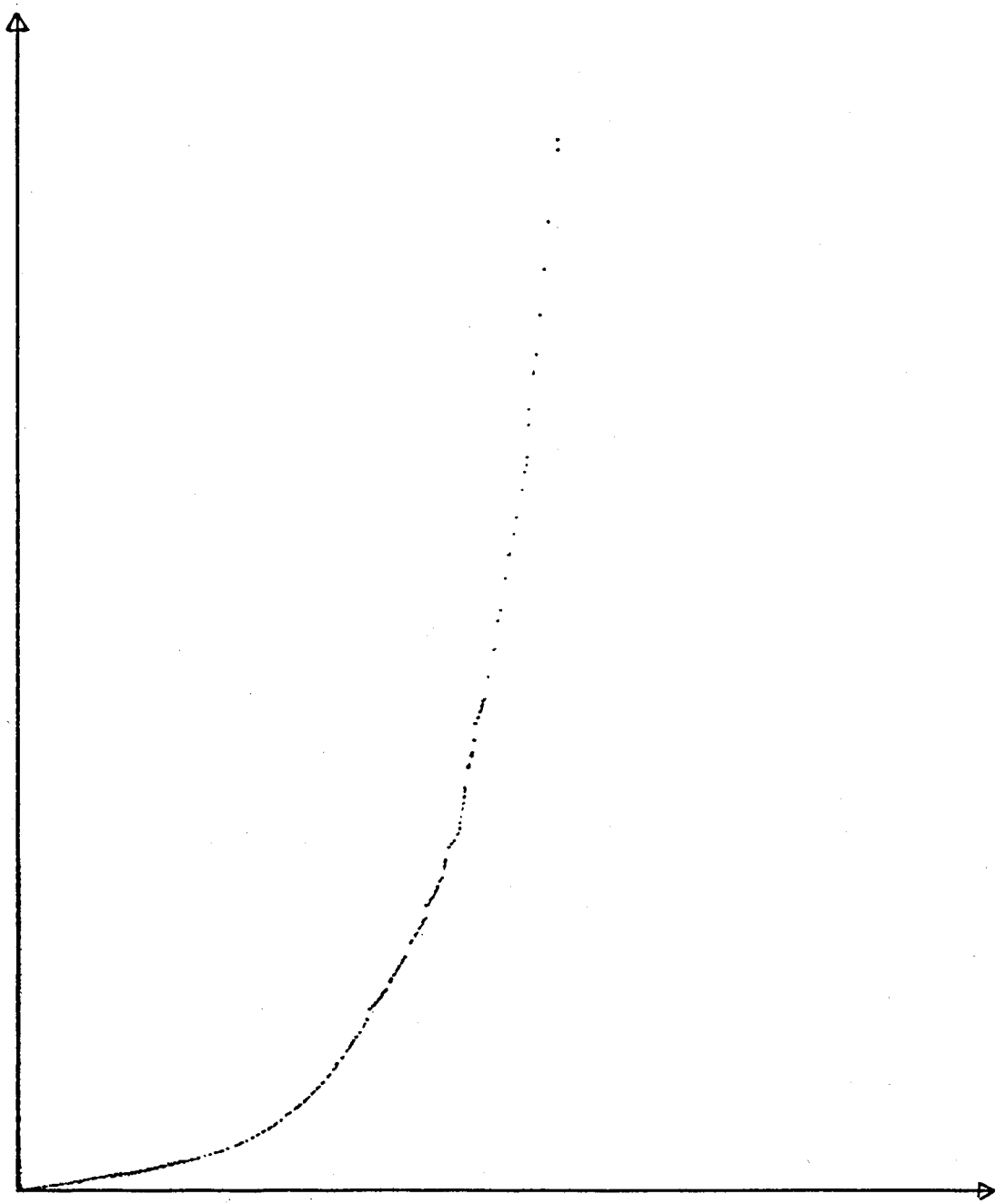


Figura F1: Gráfico de cuantilas correspondiente al PNB per cápita en Dólares, por paises (Tabla T1) Qatar (27720 $), Emiratos Arabes Unidos (24660 $) y Kuwait (20900 $) exceden los límites del gráfico.

dos, de manera que en ordenadas se situan las fracciones p_i y en abcisas las cuantilas -valores ordenados de la variable- es inmediato identificar el gráfico que aparecería, como el de la función de distribución empírica de la variable objeto de estudio. Sin embargo, para posteriores aplicaciones, es preferible mantener la configuración que el gráfico de cuantilas proporciona.

2.2. APLICACION AL ESTUDIO DE LA SIMETRÍA.

En la mayor parte de estudios estadísticos, la simetría es la propiedad distribucional máx codiciada, por diversas razones: 1º) El centro de una distribución simétrica se define sin ambigüedad y cuando se trata de distribuciones unimodales, cualquiera de los estadísticos clásicos de tendencia central --- ($\bar{x}$, M, M_o) representa un resumen aceptable - del conjunto de observaciones, 2º) Razonar sobre distribuciones simétricas es mucho más sencillo que hacerlo cuando no lo son, 3º) Los indicadores tradicionales de centro, dispersión y forma pierden su utilidad al caracterizar distribuciones que no gozan de simetría /1/, 4º) Numerosos métodos estadísticos usuales son robustos frente a desviaciones de la normalidad, siempre y cuando los datos conserven su simetría.

La simetría de la distribución es susceptible de caracterizarse en términos de la función de cuantilas. En efecto, para que una distribución pueda considerarse simétrica, los puntos que se encuentran por debajo de la mediana (M) deben exhibir aproximadamente idéntico patrón de dispersión que los situados por encima de ella. Esto, se traduce en la igualdad de distancias:

$$M-X_{(i)}=X_{(n+1-i)}-M \quad i=1,\ldots,n/2 \ (\acute{o}\ n+1/2)$$

que en el gráfico de cuantilas sería:

$$Q(.5) - Q(p) = Q(1-p) - Q(.5)$$

Estas expresiones nos sugieren como construir un gráfico específicamente para evaluar la simetría. Así, si definimos las variables

$$v_i = M - X_{(i)}$$

$$u_i = X_{(m+1-i)} - M$$

siempre que la distribución sea simétrica la línea de puntos debe "ajustarse" a la recta v = u. El sesgo a la derecha o a la izquierda se revelará por medio de puntos que respectivamente se desviarán hacia arriba o hacia abajo de dicha recta.

La tabla en tronco y hojas /1/ de las notas de estadística del curso 83-84 revela una distribución aproximadamente simétrica, excep--- tuando el agujero central en los suspensos -- con nota próxima a 5.

El gráfico recién introducido para el estudio del grado de simetría, basado en las cuantilas, nos permitirá interpretar algunos rasgos más sobre la distribución de los estudiantes.

Si convenimos en representar por un punto cada par de distancias con respecto a la mediana, (v_i,u_i) y con las puntas de una estrella el número de observaciones coincidentes, se obtiene el gráfico de la figura F2.

Observamos en él, que la inmensa mayoría de los puntos se encuentran "en" o por debajo de la recta u=v, lo que indica un mayor alejamiento progresivo de las notas por debajo de la mediana (4,8) que por encima, es decir, si juzgásemos a los alumnos con el rasero M=4,8, interpretaríamos que los "malos" estudiantes están más alejados del centro que los "buenos". Sin embargo, las últimas cinco observaciones de la derecha ilustran que los cinco mejores estudiantes del curso eran "mas buenos" que "malos" fueron los cinco peores, lo cual no debe extrañarnos, pues el presentarse a todos los exámenes no es una obligación y permite abandonar a los alumnos con muy mala nota.

3. GRAFICO DE CAJA (BOX AND WHISKERS PLOT).

3.1 INTRODUCCION.

En el apartado 3 de /1/ nos centramos en la obtención de índices numéricos capaces de resumir el conjunto de observaciones. En este apartado, pretendemos convertir estos índices en un gráfico cuya interpretación visual sea útil y rápida.

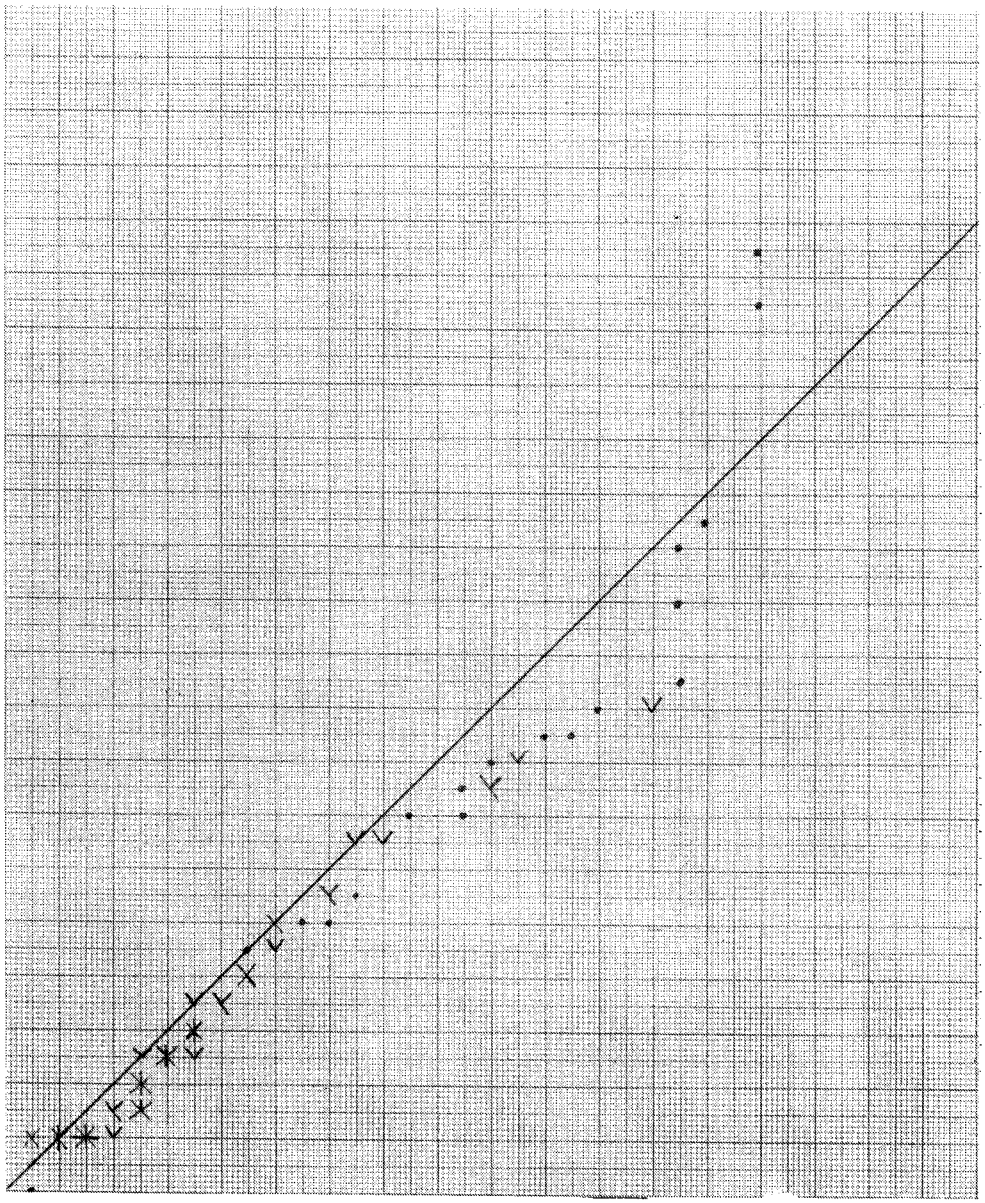
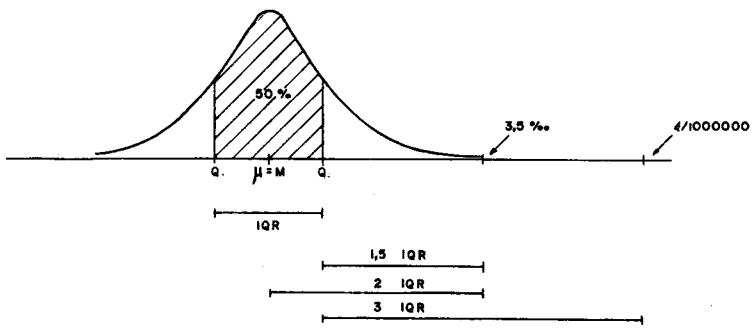


Figura F2: Gráfico para evaluar la simetría de la distribución de las notas.

El método es particularmente adecuado para diagnosticar el grado de normalidad de las distribuciones. Puesto que por un lado, las distribuciones acostumbran a ser normales en su parte central, y por otro, la no normalidad de las colas distorsiona los estadísticos usuales $\bar{x}$,s, carentes de resistencia, ya que resultan gravemente afectados ante la existencia de anomalías, se propone tomar como referencia las cuartilas muestrales, Q_1, y Q_3, que incluyen el 50% de las observaciones centrales, y a partir de ellas construir un patrón para controlar el grado de normalidad de las colas de distintas distribuciones. Ya que la tabulación en tronco y hojas describe perfectamente la parte central de la muestra, es aconsejable utilizar ambos métodos secuencialmente para complementar el análisis de los datos.

3.2. DEFINICIONES DE ANOMALIA EN LA LEY NORMAL.

Antes de proceder a construir el gráfico y con el objeto de facilitar el estudio de las colas de las distribuciones, definiremos lo que se entiende por valor excepcional o anomalía en una curva normal.



En este tipo de distribuciones la amplitud intercuartílica cumple:

$$IQR = 1,35\sigma$$

Por tanto, si a partir de cada cuartila añadimos un segmento de longitud 1,5(IQR), obtendremos un intérvalo de probabilidad fuera del cual pueden observarse únicamente el 7‰ de las observaciones. En efecto

$$P\{|x-\mu|>2(IQR)\} = P\{|x-\mu|>2.70\sigma\}= 0,007$$

Tiene pues sentido, considerar como excepcionales o anómalos aquellos valores que observamos sólo en un 7‰ de los casos. Calificaremos de anomalías extremas, las observaciones que estén más alejadas de 3(IQR) a partir de las cuartilas. De hecho, la probabilidad de extraer aleatoriamente uno de estos valores en una población normal cuyas cuartilas muestrales fuesen las observadas, es:

$$P\{|x-\mu|>3,5IQR\} = 0,000002$$

3.3 SINTESIS NUMERICAS PROPIAS DEL GRÁFICO DE CAJA.

El porcentaje que las distintas provincias aportan al total de la producción nacional (año 1979, Fuente: Anuario del País 1982), ilustrará la construcción e interpretación del gráfico de caja. Partiremos de la tabulación en tronco y hojas que previamente deberíamos llevar a cabo para organizar nuestros datos.

L		
7	0*	2 3 3 4 4 4 4
17	.	5 7 7 7 7 8 8 8 9 9
(12)	1*	0 0 0 0 0 0 1 2 2 3 3 4
21	.	5 6 6 7 7 8 9
14	2*	0 0 1 1 2 3 4
7	.	9
6	3*	0 1 4

UNIDAD: 1 = 0, 1% 1,2% = 12

A: 16,23 ; 15,63 ; 5,74

La columna de localizaciones permite extraer inmediatamente el sumario de cinco números (7), de la tabla T2, cuya disposición facilita detectar algunos de los rasgos más sobresalientes de la distribución. Exceptuando los extremos (E), la característica principal de este resumen numérico es su resistencia.

	M (1,2)	
Q_1 (0,8)		Q_3 (2)
E_i (0,2)		E_s (16,2)

Tabla T2

Este resumen numérico, que usualmente acompaña al gráfico de caja, nos aclara que la mitad de las provincias en 1979 contribuian a la producción nacional con porcentajes inferiores al 1,2%, y el reducido valor de Q_3 (2%) frente a E_s(16,2%) detecta un posible sesgo hacia la derecha de la distribución. Esto es debido al denominado efecto suelo, provocado por la existencia de un origen -- (cero) natural, que constituye un límite infranqueable que determina el crecimiento de la variable en un sentido único.

3.4 CONSTRUCCION DEL GRÁFICO.

Para dibujar el gráfico de caja que nos permitirá visualizar la distribución en su conjunto procederemos en etapas:

a) En primer lugar seleccionaremos una escala conveniente que cubra el recorrido de la variable.

b) En segundo lugar trazaremos segmentos perpendiculares al eje que señalarán las posiciones de las cuartilas (Q_1, $Q_2 \equiv M$, Q_3).

c) La caja se completará uniendo con líneas paralelas al eje los extremos de las marcas del paso anterior.

d) A partir de ambas cuartilas se trazarán líneas a trazos denominadas patillas (whiskers), que se prolongan hasta el punto más alejado que no sea anomalía (adyacente, A)

e) Añadiremos, por último, anomalías moderadas como puntos vacíos (o) y las extremas como puntos sombreados (●).

Para facilitar la construcción e interpretación del gráfico es conveniente confeccionar un nuevo resumen numérico (ya que en la práctica se omiten todos los digitos del gráfico).

	PASO	
A_i		A_s
f_i		f_s
ANOMALIAS MODERADAS		ANOMALIAS MODERADAS
F_i		F_s
ANOMALIAS EXTREMAS		ANOMALIAS EXTREMAS

donde PASO = 1,5 IQR

$f_i = Q_1 - PASO$ límites interiores

$f_s = Q_3 + PASO$ de atención.

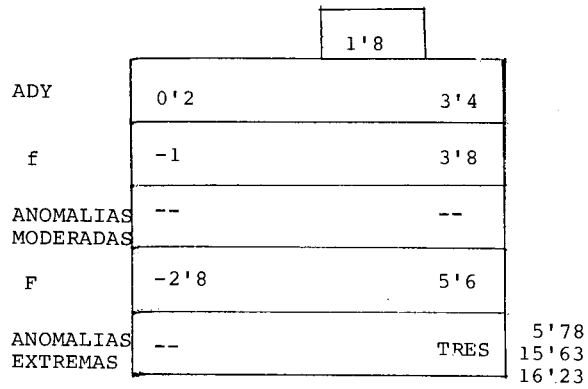
$F_i = Q_1 - 2PASO$ límites exteriores

$F_s = Q_3 + 2PASO$ de atención

Anomalías moderadas: Observaciones intermedias entre los límites interior y exterior de atención.

Anomalías extremas: Observaciones más alejadas que los límites exteriores de atención.

En el ejemplo (ver fig. F3)



3.5 INTERPRETACION DEL GRÁFICO DE CAJA

Como hemos visto, el gráfico suministra información análoga, en distribuciones unimodales , a la proporcionada por la tabulación en tronco y hojas respecto a la localización del núcleo de los datos, dispersión, sesgo y simetría de la distribución. Sin embargo, habrá ocaciones en las que se prefiera una impresión global, y menos detallada de la que ofrece el diagrama en tronco y hojas, para centrarnos más en el estudio de las colas de la distribución. Entonces, el diagrama de caja es el complemento adecuado al tronco y hojas, puesto que rapidamente sugiere:

i) la localización del centro de la distribución, definido por la situación de la mediana.

ii) la disposición de la parte central de la distribución, representada por la longitud de la caja y que coincidirá con la IQR.

iii) la forma del cuerpo central de la distribución, que establecería la posición relativa de la mediana dentro de la caja.

iv) el sesgo que sufren las colas de la distribución, en función de las longitudes relativas de las patillas.

v) las anomalías, que estarán situadas más allá de los valores frontera, f ó F.

El gráfico de la caja dificilmente detectaría estructuras bimodales o multimodales.

3.6. ALGUNAS MEJORAS EN EL GRÁFICO DE CAJA

Hemos visto antes que la probabilidad de que una observación no anómala sea considerada como tal es del 0,7% siempre que provenga de una ley normal. Ahora bien, si la población es substancialmente distinta de la normal, esta probabilidad varía considerablemente: desde el 16% en la distribución de Cauchy hasta el 0% en la uniforme.

Nuestros esfuerzos deben pues centrarse en intentar disminuir al máximo tanta variabilidad, por lo que nos parece conveniente cambiar la definición de los límites de atención f y F.

En primer lugar, existen distribuciones que por diversas razones (efecto, suelo, naturaleza de la variable, etc.) se extienden sólo en un sentido. En estos casos, demarcar fronteras simétricas a ambos lados de las medidas habituales de localización no parece ser lo más adecuado, por lo que proponemos dividir sistemáticamente la distribución en dos mitades a partir de la mediana y trabajar con ambas por separado.

Por lo tanto, podríamos redefinir los límites de atención de la siguiente manera (formulación 3.6.1.):

$$f'_s = Q_3 + 3(Q_3 - M) \qquad f'_i = Q_1 + 3(Q_1 - M)$$

$$F'_s = Q_3 + 6(Q_3 - M) \qquad F'_i = Q_1 + 6(Q_1 - M)$$

El hecho de trabajar con cuartilas como medidas de localización tiene el inconveniente de que están aún bastante alejadas de los extremos y por ello muchas veces nos definirán fronteras inadecuadas. Una rápida solución consiste en substituirlas por octilas, muy fáciles también de calcular y menos alejadas de los extremos. El único inconveniente es una

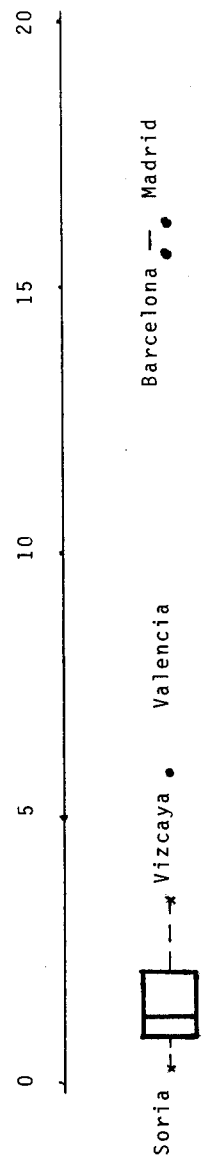


Figura F3.- Gráfico de caja correspondiente al porcentaje del PNB por provincia.

cierta pérdida de resistencia que, en general, no será relevante, pues difícilmente encontraremos datos con más de un 12,5% de observaciones anómalas en un mismo lado.

Por tanto, los nuevos límites interiores y exteriores de atención podrían venir definidos de la siguiente forma (formulación 3.6.2)

$$f_s''=0_7 + 1,35(0_7-M) \qquad f_i''=0_1 + 1,35(0_1-M)$$
$$F_s''=0_7 + 3,11(0_7-M) \qquad F_i''=0_1 + 3,11(0_1-M)$$

donde 0_1 y 0_7 son respectivamente la primera y séptima octila. Los valores de 1,35 y 3,11 se han tomado para que los nuevos límites coincidan con los anteriores en caso de normalidad.

Análogamente, si el número de observaciones es elevado y suponemos que no hay más del 6,25% de anomalías en ninguno de los dos lados, podemos proseguir en la línea anterior y definir (formulación 3.6.3.)

$$f_s'''=D_{15} + 0,76(D_{15}-M \qquad f_i'''=D_{15} + 0,76(D_1-M)$$
$$F_s'''=D_{15} + 2,08(D_{15}-M) \qquad F_i'''=D_1 + 2,08(D_1-M)$$

donde D_1 y D_{15} son las cuantilas correspondientes a 0,0625 y 0,9375 respectivamente:

$$D_1 = Q(0,0625) \qquad D_{15} = Q(0,9375)$$

Para simplificar el cálculo de 0_i y D_i podemos aproximar su localización (ver /1/) de la siguiente forma:

$$L(M) = (M + 1)/2$$
$$L(Q) = ([L(M)] + 1)/2$$
$$L(O) = ([L(Q)] + 1)/2$$
$$L(D) = ([L(O)] + 1)/2$$

donde el corchete indica la parte entera del valor considerado.

Evidentemente, al avanzar hacia los extremos, conseguimos, a base de disminuir la resistencia del método, aumentar su eficiencia. Por otro lado, si incrementamos el número de observaciones, la pérdida de resistencia se irá diluyendo con lo cual podríamos adoptar la siguiente solución de compromiso, que en base a nuestra experiencia, nos parece la más adecuada:

si $n \leq 25$ utilizar la formulación 3.6.1
si $25 < n \leq 100$ utilizar la formulación 3.6.2
si $n > 100$ utilizar la formulación 3.6.3

Las posibles mejoras del método no se acaban aquí, ya que aún puede reducirse sensiblemente la variabilidad entre las diferentes distribuciones. En efecto, supongamos que el número de observaciones es superior o igual a 100 y por tanto

$$f_s''' = D_{15} + 0,76(D_{15} - M)$$
donde $D_{15} = Q(0,9375)$ y $M = Q(0,5)$

En este caso D_{15} nos localiza en un punto bastante alejado del centro de la distribución mientras que al añadirle el término $0,76\ (D_{15}-M)$ nos distanciamos aún más dependiendo de la disposición de los datos. El grado de simetría no hace falta considerarlo de nuevo ya que al dividir la distribución en dos mitades solventamos ya este problema. Sin embargo, sí deberíamos tener en cuenta el grado de apuntamiento que presentan los datos pues dos distribuciones de igual dispersión tendrán las colas más o menos alargadas según el apuntamiento sea más o menos considerable respectivamente.

Si K_s representa un coeficiente de apuntamiento resistente de la mitad superior de la distribución y tal que $K_s=1$ en caso de normalidad, podríamos mejorar f_s''' de la forma (formulación 3.6.4):

$$f_{s,k}''' = D_{15} + 0,76\ (D_{15} - M)\ K_s^c$$

donde c es un exponente apropiado para K_s.

(si $c = 0$ $f_{s,k}^{III} = f_s^{III}$)

es decir, cuanto más apuntada sea la distribución, más largas serán sus colas y mayor será K_s, con lo cual nos acercaremos más al valor óptimo de f_s' .

Análogamente:

$$f_{i,k}^{III} = D_1 + 0,76(D_1 - M)\ K_i^c$$
$$F_{s,k}^{III} = D_{15} + 2,08(D_{15}-M)\ K_s^d$$
$$F_{i,k}^{III} = D_1 + 2,08(D_1 -M)\ K_i^d$$

En concreto se han probado coeficientes de

apuntamiento del tipo:

$$K_s = \frac{D_{15} - Q_3}{(D_{15}-M)\,0,56}$$

que consiguen aproximar extraordinariamente los límites de atención a los óptimos en un conjunto de distribuciones teóricas. Sin embargo, con datos reales la mejora no es sistemática debido a que entre otras razones, K_s no es un coeficiente demasiado eficiente.

Por tanto, aunque creemos conveniente considerar la formulación (3.6.4), pues nos sigue pareciendo válida para anomalías respecto de cualquier distribución con bastantes garantías de éxito, habría que hallar un coeficiente de apuntamiento que siquiera siendo resistente pero que además fuese suficientemente eficiente, para lo cual podríamos recurrir a los M - estimadores de apuntamiento como extensión de los de localización vistos en /1/.

Volviendo a la formulación (3.6.3.), podríamos generalizarla de otra forma definiendo:

$$f'''_{s,6} = D_{15} + C_1(D_{15}-M) + C_2(D_{15}-Q_3) + C_3(D_{15}-O_7)$$

donde C_1, C_2 y C_3 son valores cualesquiera con la única condición de que $f'''_{s,6} = 2,7$ si la variable se distribuye según una normal centrada y reducida.

Esta generalización nos parece más adecuada ya que por un lado tiene más en cuenta la forma de la distribución que en (3.6.3), es más resistente a los errores de redondeo o agrupamiento de las observaciones y todo ello sin mucha complejidad adicional.

En cuanto a los valores de C_1, C_2 y C_3, podríamos escogerlos de forma que los 3 términos de dispersión tuviesen el mismo paso en el caso de normalidad, con lo cual (formulación 3.6.5)

$$f'''_{i,6} = D_1 + 0,25(D_1-M) + 0,45(D_1-Q_1) + (D_1-O_1)$$
$$F_{s,6} = D_{15} + 0,69(D_{15}-M) + 1,24(D_{15}-Q_3) + 2,77(D_{15}-O_7)$$
$$F'''_i = D_1 + 0,69(D_1-M) + 1,24(D_1-Q_1) + 2,77(D_1-O_1)$$

Análogamente podríamos generalizar (3.6.2) de la siguiente manera (formulación 3.6.6)

$$f'''_{s,6} = O_7 + 0,67(O_7 - M) + 1,63(O_7-Q_3)$$
$$f'''_{i,6} = O_1 + 0,67(O_1 - M) + 1,63(O_1 - Q_1)$$
$$F'''_{s,6} = O_7 + 1,55(O_7 - M) + 3,76(O_7 - Q_3)$$
$$F'''_{i,6} = O_1 + 1,55(O_1 - M) + 3,76(O_1 - Q_3)$$

En el ejemplo de las notas, utilizando el método tradicional se obtiene:

$f_s = 7,35$ $f_i = 2,5$ $F_s = 9,3$ $F_i = 0,2$

mientras que con el método propuesto, al haber 19 casos emplearíamos la formulación 3.6.5.

$f'''_{s,6} = 7,65$ $f'''_{s,6} = 1,29$ $F'''_{s,6} = 9,85$ $F'''_{i,6} = 1,56$

Estos resultados nos detectan dos anomalías moderadas (8,1 y 8,3), mientras que con las fórmulas originales se hubiesen detectado como anómalas, además de las anteriores, las 2 observaciones correspondientes al valor "2", lo cual no parece lo más adecuado a la vista del diagrama en tronco y hojas.

En el caso del producto nacional bruto por países, pasaríamos de tener 16 anomalías moderadas y 3 extremas con la formulación original a tener únicamente 2 anomalías moderadas que corresponderían a los dos países más ricos (Qatar y Emiratos Arabes Unidos).

Por último, en el ejemplo de la contribución a la producción nacional, el único cambio remarcable es el correspondiente a Valencia que pasaría de ser una anomalía extrema a ser solamente moderada.

Aunque cualquier juicio sobre estos resultados será siempre subjetivo, creemos que en los 3 casos la mejoría respecto a la disposición original es evidente. Sin embargo, esto no es razón suficiente como para validar el método, por lo que además de experimentarlo repetidamente con diferentes conjuntos de datos reales, hemos comparado los resultados con un grupo bastante completo de distribuciones teóricas. Esta valoración más objetiva, nos ratifica la impresión satisfactoria que habíamos obtenido anteriormente.

Por tanto, a modo de resumen el método que se propone consiste en lo siguiente:

Utilizar la formulación 3.6.1. Si $n \leq 25$
" " " 3.6.6. Si $25 < n \leq 100$
" " " 3.6.5. Si $n > 100$

Esto supone que, complicando ligeramente la construcción del gráfico de caja, hemos aumentado sensiblemente su eficiencia para detectar anomalías sin disminuir excesivamente la resistencia del método.

4. REFERENCIAS.

/1/ BATISTA, J.M. y VALLS, M.: "Nuevas Técnicas de Análisis estadístico de datos. Tabulación y síntesis numéricas" - QUESTIIO, V. 9, N.2 (1985).

/2/ CLEVELAND, W.S. and McGILL, R.: "Graphical Perception: Theory, Experimentation and Application to the development of Graphical Methods". J.A.S.A. 79, 531-553 (1984).

/3/ CROXTON, F.E. and STRYLER, R.E.: "Bar Charts versus circle diagrams". J.A.S.A. 22, 473-482. (1927).

/4/ DARELL HUFF.: "How to lie with statistics". W.W. NORTON AND COMPANY (1954).

/5/ EELLS, W.C.: "The relative Merits of circles and bars for representing component parts". J.A.S.A. 21, 119-132.(1926).

/6/ PLAYFAIR, W.: "The Commercial and Political atlas". London (1986).

/7/ TUKEY, J.W.: "Exploratory Data Analysis". ADDISON-WESLEY. (1977).