DESIGUALDAD, DIVERSIDAD Y CONVERGENCIA:

(MÁS) INSTRUMENTOS DE MEDIDA

-ESTADÍSTICA DESCRIPTIVA-.[*]

Francisco J. Goerlich

Correspondencia: Francisco J. Goerlich Gisbert
Departamento de Análisis Económico
Universidad de Valencia
Campus de los Naranjos, (Edificio Departamental Oriental)
46022 - Valencia

Tel.- 96-3828246, Fax.- 96-3828249
e-mail: Francisco.J.Goerlich@uv.es, web: http://www.uv.es/~goerlich

Editor: Instituto Valenciano de Investigaciones Económicas

Primera Edición Abril 2000

Depósito Legal: V-1426-2000

[*] Este trabajo recoge parte de los aspectos instrumentales de un informe más amplio titulado "Dinámica de la distribución provincial de la renta. II: La forma externa de la distribución -Evolución histórica-" realizado para el Instituto Valenciano de Investigaciones Económicas (I.V.I.E). Se agradece la financiación recibida de la DGICYT, proyecto SEC98-0895, y del Instituto Valenciano de Investigaciones Económicas.

DESIGUALDAD, DIVERSIDAD Y CONVERGENCIA:
(MÁS) INSTRUMENTOS DE MEDIDA
-ESTADÍSTICA DESCRIPTIVA-

Francisco J. Goerlich

RESUMEN

Este trabajo es una continuación de Goerlich (1998) y presenta un conjunto amplio de resultados cuyo objetivo es caracterizar la distribución *cross-section* de la renta *per capita* para un conjunto de individuos o unidades geográficas que agrupen a varios individuos. Comienza describiendo un conjunto de estadísticos útiles para ello, así como diversas formas de resumir la información proporcionada por los mismos, para continuar describiendo métodos de estimación directa de la función de densidad de la variable objeto de estudio.

PALABRAS CLAVE: Desigualdad, diversidad y convergencia. Estadísticos descriptivos, muestras ponderadas, estimación de funciones de densidad.

ABSTRACT

This work is a continuation of Goerlich (1998) and offers a wide set of results with the aim of characterize the cross-section distribution of per capita income for individuals or for a group of geographical units, such as regions or countries. We begin by describing a set of descriptive statistics and ways in which the information they provide can be sumarized, and continues by considering direct ways of estimating the probability density function of a variable mainly nonparametrically by means of kernel smoothing.

KEY WORDS: Inequality, divergence and convergence. Descriptive statistics, weighted samples, (kernel) density estimation.

1. Introducción y nomenclatura

Este trabajo es continuación de Goerlich (1998) y simplemente expone un conjunto amplio de estadísticos descriptivos con el ánimo de proporcionar un marco de referencia para una mejor comprensión de la evolución dinámica de determinadas variables económicas. Al igual que en el trabajo mencionado el análisis se realiza a partir de la exposición de una serie de técnicas con diversos grados de sofisticación comenzando con estadísticos totalmente elementales.

Aunque tomaremos como punto de referencia una variable clave en el proceso de crecimiento económico, como es la **renta *per capita***, los instrumentos que expondremos a continuación son aplicables con total generalidad. Si bien en Goerlich (1998) el análisis se realizó de forma exclusiva a partir de la utilización de conceptos tomados de la literatura de la desigualdad, que ha concentrado gran parte de sus esfuerzos en la elaboración de índices que posean determinadas propiedades (Atkinson (1970), Sen (1973), Chakravarty (1990), Cowell (1995)), este **segundo trabajo** toma prestados conceptos de la literatura aplicada sobre convergencia económica y busca básicamente estadísticos que nos permitan caracterizar la distribución *cross-section* de la renta *per capita* para un conjunto de individuos o unidades geográficas, como países o regiones, que engloben a varios individuos. Por tanto sea x la renta *per capita* objeto de estudio la finalidad es caracterizar $\phi(x)$, siendo $\phi(\bullet)$ una medida de la función *cross-section* de densidad de probabilidad de x. Hay dos características interesantes susceptibles de estudio en la evolución temporal de $\phi(x)$: (i) la forma cambiante en el tiempo de dicha función, y (ii) la dinámica intra-distribucional, es decir como una parte dada de la distribución en t transita a otra parte de dicha distribución en $t+j$. Las dos características sobre las que incidiremos son pues "forma externa" y "movilidad". El presente trabajo y su complementario (Goerlich (2000)) se centran básicamente en el estudio de la evolución dinámica de la forma externa de la distribución (*the external shape of the distribution*), aquí se examinarán estadísticos útiles para caracterizar $\phi(x)$, con especial hincapié en el concepto de σ-convergencia, así como los métodos que nos permiten inferir la forma de dicha función, mientras que Goerlich (2000) expondrá diversas formas de caracterización de $\phi(x)$ en el contexto de modelos de

regresión, centrándose fundamentalmente en el concepto de β-convergencia. El estudio de lo que sucede dentro de la distribución, es decir la movilidad, se abordará posteriormente.

Dos corrientes de literatura que han permanecido separadas, pero que hasta cierto punto son complementarias y cuyas técnicas de análisis pueden combinarse adecuadamente son: (1) la literatura tradicional sobre la desigualdad (Atkinson (1970), Sen (1973), Shorrocks (1980, 1982, 1984), Chakravarty (1990), Esteban y Ray (1993, 1994), Cowell (1995)), centrada fundamentalmente en el estudio de la distribución personal de la renta, y (2) la reciente literatura sobre la convergencia económica (Barro (1991), Barro y Sala-i-Martín (1991,1992,1995), Quah (1993a,b), Sala-i-Martin (1994)), preocupada por la convergencia o divergencia de la renta *per capita* o productividad de diversas unidades geográficas, ya sean regiones o países. Aunque ambas literaturas han tendido a permanecer separadas es evidente que tienen importantes puntos de contacto. Basta para ello ojear los trabajos de Esteban y Ray (1993) o Esteban (1996) sobre la polarización o los de Baumol (1986), DeLong (1988) o Quah (1996a,b,1997) sobre la existencia de clubes de convergencia para darse cuenta de que, a grandes rasgos, se está hablando de conceptos similares, grupos de individuos o regiones que presentan peculiaridades distintas del resto. Así pues aunque la literatura sobre la desigualdad parte del individuo y la del crecimiento de una unidad espacial considerablemente más amplia, las dos tratan de estudiar la evolución en el tiempo de la distribución de una variable económica considerada de especial relevancia desde el punto de vista del bienestar o de la actividad económica. Debe ser obvio entonces que las técnicas de análisis en un tipo de literatura pueden utilizarse satisfactoriamente en el otro. De hecho algunos autores (Rabadán y Salas (1996)) han propuesto medir directamente la convergencia mediante índices de desigualdad; este enfoque, llevado hasta su extremo, podría sufrir de algunas de las críticas de Quah (1993a,b) y Esteban (1996), ya que como veremos no parece adecuado reducir el concepto de convergencia a unos pocos estadísticos.

Si bien en Goerlich (1998) se examinaron conceptos procedentes de la literatura de la desigualdad, este trabajo y su complementario (Goerlich (2000)) utilizan fundamentalmente técnicas de análisis de la literatura aplicada sobre convergencia económica con la finalidad de examinar si la distribución de corte transversal de la renta *per capita* tiende en el tiempo hacia la igualdad en dicha renta o hacia una distribución

estacionaria, así como la forma de dicha distribución. El trabajo se centra en aspectos metodológicos y prácticos, no se ofrecen aplicaciones, muy numerosas por otra parte, si bien cuando requiramos de algún ejemplo este utilizará los datos de la renta *per capita* provincial de la Base de Conocimiento Económico Regional, **Sophinet**, de la Fundación BBV[1]. Dado que nuestra unidad de referencia no es necesariamente el individuo, introduciremos explícitamente la dimensión poblacional en el análisis, de forma que prestaremos una especial atención al tema de las ponderaciones en la construcción de los estadísticos, donde las ponderaciones vendrán dadas por las frecuencias relativas de la variable objeto de estudio, población en el caso de la renta *per capita*, aunque los resultados son aplicables con mayor generalidad. Esta dimensión poblacional es normalmente recogida por la literatura de la desigualdad pero, sin embargo y sin causa aparente, parece haber sido olvidada por la reciente literatura sobre la convergencia económica.[2]

Nomenclatura

Nuestro conjunto de observaciones de referencia se mueve en dos direcciones, el ámbito espacial y el ámbito temporal y constituye lo que la literatura reciente (Quah (1990)) ha dado en llamar un **campo de datos** (*data field*) en el que tanto n, el número de grupos, como T, el número de periodos, son razonablemente grandes o al menos de una de una dimensión similar. Sin embargo puesto que a lo largo del trabajo los estadísticos se calculan para cada *cross-section* podemos olvidarnos, de momento, de la dimensión temporal omitiendo el subíndice t, tal y como hicimos en Goerlich (1998). Así pues supongamos que disponemos de n agrupaciones de individuos para un determinado periodo temporal, $t = 1,...,T$, cuya **renta *per capita*** designamos por $\boldsymbol{x_i}$, $x_i = Y_i/N_i$,[3] siendo Y_i la renta y N_i la población de la agrupación $i = 1,2,...,n$. Sea además $\boldsymbol{p_i}$ la **frecuencia relativa**, esto

[1] Cuya dirección electrónica es http://bancoreg.fbbv.es/. Los datos de población proceden del *Anuario Estadístico* del INE.

[2] No obstante algunos autores si habían observado este olvido, Rabadan y Salas (1996, p.-15) o Jones (1997, p.-23).

[3] x_i es la **renta real equivalente *per capita***, es decir ha sido adecuadamente deflactada y ajustada por las diferentes necesidades de las agrupaciones, familias o individuos. (Deaton y Muellbauer (1980)).

es, el porcentaje de población por agrupación, $p_i = N_i/N$, $N = \Sigma_{i=1}^{n} N_i$, entonces la **renta *per capita* media** para el agregado puede expresarse como una media aritmética ponderada,

$$\mu = \frac{Y}{N} = \frac{\Sigma_{i=1}^{n} Y_i}{N} = \Sigma_{i=1}^{n} \frac{Y_i/N_i}{N/N_i} = \Sigma_{i=1}^{n} p_i x_i \qquad (1)$$

Nuestra **variable** de referencia es por tanto la **renta *per capita***, x_i, de forma que realizaremos la exposición en términos de esta variable y sus **pesos asociados**, p_i; sin embargo no deberemos perder de vista que si queremos construir estadísticos independientes de la escala, de forma que los estadísticos permanezcan inalterados si la renta de cada individuo en la población (o la renta *per capita* de cada agrupación) es alterada en la misma proporción (**homogeneidad de grado cero en rentas**), entonces deberemos normalizar x_i por su valor medio, μ, de forma que en la práctica muchas veces estaremos más interesados en la variable $z_i = \frac{x_i}{\mu}$; esta es la normalización adoptada por los índices de desigualdad relativos (Goerlich (1998)). En este caso los estadísticos son insensibles al nivel medio de renta y no consideran cuestiones de posición de la variable en cuestión.

Finalmente **dos breves reflexiones**, en primer lugar palabras como desigualdad, diversidad, diferenciación y convergencia son utilizadas como sinónimos en muchas partes del trabajo, lo que constituye un cierto abuso del lenguaje. Si la diversidad, o alternativamente la convergencia, es buena o mala, si debe aumentarse o disminuirse mediante políticas adecuadas, es algo que depende de juicios de valor y sobre lo que no nos pronunciaremos.

En segundo lugar la desigualdad y el crecimiento de las economías es un fenómeno complejo y multidimensional. Por ello, todo intento de resumir el proceso de convergencia en un único estadístico está abocado al fracaso. Quah (1993a,b) ha enfatizado satisfactoriamente este punto y a propuesto una serie de instrumentos metodológicos complementarios para analizar la evolución dinámica de distribuciones en el corte transversal (*model of explicit distribution dynamics*), parte de estos instrumentos serán presentados en este trabajo y su complementario (Goerlich (2000)). El trabajo se estructura

en dos grandes secciones, la **sección 2** presenta un conjunto amplio de estadísticos conocidos y diversas formas de resumir la información proporcionada por los mismos y la **sección 3** examina las diversas posibilidades de estimación directa de la función de densidad de una variable, es decir la función $\phi(x)$.

2. Estadísticos descriptivos: Posición, dispersión (convergencia-σ) y otras características interesantes de la distribución

Esta sección ofrece una descripción de estadísticos relevantes para una variable, teniendo como referencia la renta *per capita* de regiones o países, x_i; el objetivo es ir caracterizando la evolución en la distribución de dicha variable, $\phi(x)$, mediante una exposición exhaustiva de estadísticos descriptivos.

Uno de los **conceptos** fundamentales al que hace referencia esta sección es el denominado **σ-convergencia**, entendido en un sentido amplio como la **dispersión en toda la distribución** y no en el sentido más restringido acuñado por Barro y Sala-i-Martín (1995, Cap.-11.1, p.-383-387) como la varianza del logaritmo de la renta *per capita*; no obstante las medidas de posición también serán relevantes en la caracterización de la distribución, así como la simetría de la misma alrededor de un valor central y la identificación de los valores atípicos o *outliers*.

2.1. Estadísticos simples versus estadísticos ponderados: Una digresión[4]

Comenzamos esta sección con una digresión no trivial que ha sido largamente ignorada en la literatura; como ya hemos mencionado nuestra variable de referencia es la renta *per capita* de áreas geográficas que engloban a varios individuos, o más concretamente la función *cross-section* de densidad de probabilidad de dicha renta *per capita,* $\phi(x)$; la cuestión que se suscita inmediatamente es si el comportamiento de la renta *per capita* debe ser analizado en términos de dichas áreas geográficas o en términos de individuos; dicho con otras palabras la cuestión es si cuando trabajamos con rentas *per capita* medias de diferentes áreas geográficas la dimensión económica de dichas áreas debe contar para algo o no.

[4] Agradezco a Jose María Esteban algunas de las reflexiones contenidas en este epígrafe.

Los dos bloques de literatura que se analizan en este y en el anterior trabajo (Goerlich (1998)) han dado soluciones prácticas diferentes a esta cuestión, por una parte la literatura económica sobre la desigualdad, preocupada fundamentalmente por el bienestar individual, ha utilizado siempre **estadísticos ponderados**, donde la ponderación trata de reflejar la dimensión económica de las diferentes áreas geográficas analizadas, en este sentido todos los estadísticos analizados en Goerlich (1998) son estadísticos ponderados, razón por la cual no planteamos allí esta cuestión.[5] Por otra parte la literatura sobre crecimiento y convergencia económica, preocupada por los países o las regiones, ha utilizado de forma prácticamente exclusiva **estadísticos simples**, en el sentido de que la renta *per capita* de cada área geográfica era considerada como una observación individual, independientemente del tamaño o la importancia relativa de dicha área dentro del agregado.[6]

Aunque la **ponderación razonable** en este contexto parecen ser las **proporciones de población**, p_i, de cada región, así lo entienden los índices de Gini, G, Desviación Absoluta Media, M, Theil para $\beta = 0$, $T(0)$, y Atkinson, $A(\varepsilon)$, ya que en un contexto puramente estadístico estas proporciones representan las frecuencias relativas de las correspondientes rentas *per capita* y de hecho la media del agregado, que si es observable, es una media aritmética ponderada por proporciones de población, $\mu = \Sigma_{i=1}^{n} p_i x_i$; otras

[5] Sería posible de hecho calcular todos los estadísticos de desigualdad ofrecidos en Goerlich (1998) de forma no ponderada, estadísticos de desigualdad simples, simplemente considerando cada x_i como una sola observación; sin embargo nadie parece plantearse esta cuestión.

[6] Una cuestión similar, pero no idéntica, aparece cuando trabajamos con **datos de encuesta** y cada observación lleva asignada una ponderación muestral derivada del proceso de muestreo y relacionada con la probabilidad de que esa observación haya sido selecionada en la muestra, los denominados **factores de elevación**; tenemos de esta forma lo que se denomina una muestra ponderada, para la que es posible mantener el supuesto de independencia pero no el de idéntica distribución (Beach y Kaliski (1986), Bishop, Chakraborti y Thistle (1994)). La utilización de dichos factores en el cálculo de estadísticos descriptivos y medidas de desigualdad es estándar (Bosch, Escribano y Sánchez (1989), Atkinson, Rainwater y Smeeding (1995), Martín-Guzmán, Toledo, Bellido, López y Jano (1996), Goerlich y Mas (1999)) y su utilización en modelos de regresión o inferencia estadística ha sido objeto de atención diversa por parte de la literatura estadística que trabaja con datos de encuesta. (Nathan y Holt (1980), Hausman y Wise (1981), DuMouchel y Duncan (1983), Jewell (1985), Kott (1991), Cosslett (1993), Pfefferman (1993), Selden (1994), Korn y Graubard (1995a,b), Imbens y Lancaster (1996), Magee, Robb y Burbidge (1998), Wooldridge (1999)).

Nuestra muestra, por el contrario, no es propiamente dicha una muestra ponderada, pero si nos permite examinar la cuestión de si, dadas las rentas *per capita* medias observadas para las distintas regiones, nuestro interés debe dirigirse hacia estudio del comportamiento de dichas rentas en términos de las regiones mismas o en términos de los individuos que las habitan. Por supuesto, el ejercicio cuando la unidad de referencia es el individuo no dice nada acerca de la distribución de la renta dentro de cada región concreta, ya que para ello necesitaríamos datos de los individuos mismos, es decir datos microeconómicos (Goerlich y Mas (1999)).

ponderaciones son posibles, por ejemplo el índice de Theil para $\beta = 1$, $T(1)$, utiliza ponderaciones según proporciones de renta, e incluso en principio sería posible ponderar por superficie o cualquier magnitud que represente en alguna medida el tamaño económico. No obstante en este trabajo utilizaremos siempre ponderaciones por **proporciones de población**, ya que son las más fácilmente interpretables en el contexto de nuestra variable de referencia, si bien los resultados que presentaremos son aplicables a cualquier muestra en la que las observaciones lleven asociado un peso determinado.

Un ejemplo simple ayudará a transmitir la idea en la que estamos pensando. Considérese la distribución *cross-section* de rentas *per capita* en dos momentos del tiempo, t y t+1, para 3 regiones diferentes. El tamaño de la población, N, es constante e igual a 100 individuos desigualmente repartidos entre las regiones. Dicha distribución puede observarse en el cuadro 1.

Cuadro 1: Dos distribuciones hipotéticas de la renta *per capita*

Región	t		t+1	
	x_i	p_i	x_i	p_i
A	1	0.25	1	0.40
B	2	0.50	2	0.20
C	3	0.25	3	0.40

Obsérvese que dado que x_i es idéntica para cada región tanto en t como en t+1 los estadísticos simples no varían, i.e. la distribución de x_i en términos de regiones permanece constante, la media es igual a 2 y la varianza es igual a 2/3, tanto en t como en t+1. La desigualdad, sin embargo, en el sentido en el que se entiende tradicionalmente en la literatura, es decir la dispersión en la distribución de x_i en términos de individuos, ha aumentado, ya que una gran proporción de población se ha desplazado desde el centro, región B, hacia los extremos de la distribución; en concreto un 15% pasa al extremo inferior, región A, y otro 15% al extremo superior, región C. Los índices de desigualdad así lo reflejarían y también al cálculo de **estadísticos ponderados** por las proporciones de

población. Así aunque la media ponderada tanto en t como en t+1 sigue siendo igual a 2,[7] la varianza ponderada en t es 0.5 mientras que la varianza ponderada en t+1 es 0.8, es decir, se produce un aumento de la dispersión, dicho con otras palabras la distribución de x_i en términos de individuos muestra un aumento de la desigualdad. Por supuesto esto no dice nada acerca de la distribución de la renta dentro de cada región.

En términos de la debatida cuestión de la **convergencia** ¿que debemos concluir?, ¿se ha producido un proceso de divergencia o por el contrario la distribución se ha replicado a sí misma?. La literatura macroeconómica sobre la convergencia, utilizando estadísticos simples, concluiría que no ha habido ni convergencia ni divergencia. En efecto si como unidad de análisis consideramos las regiones entonces lo que nos interesa es la distribución no ponderada de x_i y por tanto no obtendríamos convergencia ni divergencia, sino una réplica de la distribución. En términos estadísticos, si lo que consideramos es que disponemos de una muestra aleatoria de regiones entonces la ponderación no es importante. Sin embargo si pensamos en que las rentas *per capita* que estamos analizando tienen detrás diferentes tamaños de población parece razonable que la dispersión de la distribución la midamos desde el punto de vista individual y por tanto alteraciones en las proporciones de población que las diversas regiones representan dentro del agregado pueden afectar al proceso de convergencia o divergencia, aún en casos extremos como los de nuestro ejemplo en el que ni la renta *per capita* media de cada región ni la agregada (simple o ponderada) se alteran. Dicho en términos estadísticos, si nuestra población de referencia son las personas, entonces deberemos otorgar más peso a aquellas regiones más densamente pobladas, no hacerlo así distorsionará las características de la distribución que tratamos de estudiar. Este enfoque nos llevaría a concluir, en nuestro ejemplo, que se ha producido un proceso de divergencia. ¿Tiene esto sentido desde el punto de vista de la literatura del crecimiento económico?, ciertamente lo tiene; al fin y al cabo el modelo de Solow (1956) y Swan (1956), que ha inspirado gran parte del debate teórico y aplicado sobre la convergencia económica, es un modelo que se aplica al comportamiento esperado de un país individual y que hace referencia al proceso de convergencia de dicho país a un

[7] El hecho de que la media simple y ponderada en t y t+1 sea siempre la misma se deriva del hecho de que la distribución, tanto simple como ponderada, es siempre simétrica en ambos periodos. Obviamente esto no es una característica general pero dado que lo que nos interesa es examinar el fenómeno de las ponderaciones en el cálculo de los estadísticos hemos simplificado al máximo el ejemplo.

estado estacionario; sin embargo el modelo ha sido aplicado a diferentes países y regiones y a distintos niveles de desagregación, por lo que extendiendo el argumento hasta el extremo podría ser aplicado a individuos, i.e. convergencia de las rentas individuales; de hecho Cass (1965) y Koopmans (1965), recuperando el análisis de agentes optimizadores de Ramsey (1928), desarrollaron el modelo de Solow (1956)-Swan (1956) en términos de un consumidor representativo y por tanto, estrictamente hablando, en términos de individuos y esta es la tendencia actual en la moderna teoría del crecimiento económico (Barro y Sala-i-Martín (1995))[8].

Los argumentos que acabamos de esgrimir hacen presagiar que la cuestión con la que iniciamos este epígrafe, si el comportamiento de la renta *per capita* debe ser analizado en términos de áreas geográficas o en términos de individuos, no va a ser nada fácil de discutir, al menos desde una actitud de principios. Probablemente lo que ha motivado la utilización de estadísticos simples (no ponderados) por parte de los investigadores dedicados a estudiar el tema de la convergencia económica son preguntas como la siguiente: ¿Cuanta diferencia hay entre las distintas especies de animales?. Esta es una cuestión que parece razonable contestar por medio de un índice de diferenciación (desigualdad) sin ponderaciones que recojan el total de población de cada especie dentro del mundo animal. La consideración de diferentes ponderaciones nos llevaría, entre otras cosas, a tener que sumar la población de elefantes con la de hormigas para calcular la población total del reino animal. El enfoque adoptado por la literatura aplicada del tema de la convergencia internacional parece ser este, donde cada país es una 'especie'. Hay, sin embargo, una cuestión importante que hace que el símil que acabamos de utilizar no sea del todo adecuado. En nuestro ejemplo la distancia, dentro del índice de diferenciación, entre dos especies es nula si y sólo si las dos especies son una misma. En el caso de la convergencia entre países, sin embargo, aunque la distancia en renta *per capita* entre dos países sea nula, siguen considerándose dos observaciones separadas y no una sola. El problema es, al menos parcialmente, semántico. Usamos una misma palabra, desigualdad, para referirnos a cuestiones muy distintas. Si de lo que se trata es de medir la diferenciación entre países, entonces podríamos utilizar el enfoque de la diversidad

[8] Otros autores han estudiado la convergencia económica entre regiones pero a partir de datos individuales sobre rentas y no ha partir de los valores medios de las rentas *per capita* regionales, Bishop, Formby & Thistle (1992).

biológica para analizar la cuestión (Weitzman (1992)). Esta sería una línea interesante desde la que examinar la opción tomada por los especialistas en el tema de la convergencia económica. Aunque implícito en este enfoque es que el preservar la diversidad es un valor positivo, evidentemente no hay problema en invertir los términos y presuponer que la diversidad es un valor negativo. En cualquier caso esta divagación pone de manifiesto que, desde una actitud de principios, nos introducimos rápidamente por caminos cuyo destino no parece obvio, al menos a primera vista.

Las reflexiones que hemos realizado aquí parecen apuntar hacia el hecho de que la dispersión, o en general el estudio de la evolución dinámica de la distribución de la renta *per capita*, en terminología de Quah (1996c,d), debe realizarse en términos de las distribuciones ponderadas por la población, o en otras palabras el comportamiento de la renta *per capita* debe ser analizado en términos de individuos; y ello por varias razones, en la práctica es razonable preguntarse cuestiones tales como: **¿es indiferente que países como España o Francia converjan al nivel medio de la renta *per capita* Europea a que lo haga Luxemburgo?**. Todo parece indicar que **no**; no sólo desde el punto de vista individual la convergencia es mayor si convergen países grandes en lugar de países pequeños, y ello independientemente de lo que suceda con la distribución de la renta dentro de cada país, sino que otras cuestiones relevantes, como los procesos de transferencias de regiones ricas a regiones pobres, dependen sustancialmente de los tamaños de población que hay detrás de una renta *per capita* concreta. Sería fácil construir ejemplos en los que un resultado de divergencia económica, obtenido a partir de estadísticos no ponderados, se debe al sistemático alejamiento respecto a la media de uno o dos países de tamaño insignificante, como Luxemburgo o Irlanda; mientras que una adecuada consideración de sus tamaños relativos dentro del agregado podría arrojar el resultado contrario de convergencia económica[9]. Por tanto al margen de una actitud de principios existe una actitud empírica, ¿proporcionan los estadísticos simples y los ponderados visiones diferentes, cuando no contradictorias, de un mismo fenómeno económico?; disponemos de algunos ejemplos indica que si, Korn y Graubard (1995b), mientras que otros indican que

[9] Comunicación personal del autor con L. Magee (Magee, Robb y Burbidge (1998)) indica que existe cierta evidencia de que muchos investigadores parecen estar en contra de las ponderaciones debido al hecho de que en contextos internacionales o regionales los paises o las regiones pequeñas, que suelen ser la mayoría, no tendrían prácticamente impacto sobre los resultados, que estarían dominados por unas pocas observaciones.

no, al menos no siempre, Goerlich y Mas (1998a,b). En consecuencia en este trabajo adoptaremos una aproximación práctica al problema y todos los estadísticos de esta sección serán presentados tanto de forma ponderada como de forma simple (no ponderada), al objeto de examinar como la consideración de las frecuencias relativas pueden alterar el cálculo de ciertos estadísticos y en consecuencia las conclusiones que podemos extraer de ellos.

Para hacernos una idea de las potenciales diferencias de la posible discrepancia entre los estadísticos ponderados respecto a los simples basta con examinar la estructura demográfica de las diferentes provincias españolas. Por ejemplo, Barcelona y Madrid, que se encuentran en el extremo superior de la distribución de la renta *per capita*, representaban más del 24% del total de la población española en 1995, sin embargo los estadísticos simples les asignan un peso conjunto del 4% en cualquier año, lo contrario sucede en provincias como Soria que, aunque representaba un porcentaje del 0.24% de la población en 1995, los estadísticos simples le asignan un peso fijo del 2% en cualquier año. Nos encontramos pues con dos efectos: (*i*) no todas las provincias tienen el mismo peso, y (*ii*) dicho peso varía en el tiempo. En este contexto la moda, entendida como un estadístico puntual asociado a una sola observación, vendría dada por la provincia con más población relativa. Además piénsese que si en lugar de realizar el análisis a nivel de provincias se realizara a nivel de Comunidades Autónomas, entonces una comunidad uniprovincial como es La Rioja, que representa porcentajes de población inferiores al 1% del total nacional, pasaría, en los estadísticos simples, de pesar un 2% a pesar un 5.88%, simplemente porque ha variado el número de unidades en el análisis; por lo tanto, en este caso, la división administrativa si importa, sin embargo la importancia relativa de La Rioja, ni en términos de renta ni en términos de población, ha variado por este motivo; por el contrario su peso en los estadísticos ponderados no se vería alterado. De todo ello se desprende que en principio si existe base para una dispar evolución entre los estadísticos simples y ponderados.

El tipo de argumentos que hemos ofrecido en este epígrafe parece que han estado totalmente ausentes en la literatura empírica sobre convergencia. Por una parte los autores procedentes del análisis microeconómico y la desigualdad (Theil y Sorooshian (1979), Berrebi y Silber (1987), Esteban (1994), Duro y Esteban (1998)) no parecen cuestionarse

el problema y simplemente aplican el instrumental de índices de la literatura de la desigualdad al análisis regional de la convergencia, sin ningún tipo de mención al respecto. Por otra parte, y salvo por comparaciones de las distintas regiones respecto a la media del agregado, que es una media ponderada, los estudiosos del problema provenientes de la macroeconomía, utilizan de forma prácticamente exclusiva estadísticos simples. Así por ejemplo Decresin y Fatás (1995, p.-1630) reconocen el problema pero no hacen nada al respecto, salvo escoger la regiones de forma que sean comparables en tamaños de población. Ante esta situación es lícito preguntarse que hubiera pasado si hubiéramos trazado las fronteras de forma diferente, ¿habría ello supuesto una alteración substancial en los resultados?. Algunas excepciones a este respecto son el trabajo de Rabadán y Salas (1996), quienes por el motivo que hemos mencionado proponen medir la convergencia mediante índices de desigualdad; procedimiento lícito aunque no el único posible, ni siquiera tiene porque ser el más adecuado, ya que por ejemplo desde el punto de vista de la convergencia no hay porque asignar más importancia a las transferencias de renta en el extremo inferior de la distribución, lo que por el contrario si puede ser deseable en términos de la medición de la desigualdad individual, ni porque basar los estadísticos en conceptos normativos sobre el bienestar social; y el de Jones (1997, p.-22), quien argumenta que aunque el análisis de la renta *per capita* a nivel internacional se realiza normalmente en términos de los países esta puede ser una forma engañosa de examinar los datos ya que simplemente la alteración de las fronteras modificaría los resultados.

2.2. Inferencia con estadísticos ponderados: Un comentario

Aunque nuestro interés en este trabajo se centra en el cálculo de estadísticos descriptivos ponderados, la inferencia estadística en este caso, tan desarrollada con muestras simples, merece un comentario.

La cuestión de la inferencia con estadísticos ponderados ha sido objeto de atención desde hace tiempo por parte de la literatura estadística y econométrica que trabaja con datos de encuesta (Klein y Morgan (1951), Nathan y Holt (1980), Hausman y Wise (1981), DuMouchel y Duncan (1983), Jewell (1985), Beach y Kaliski (1986), Kott (1991),

Pfefferman (1993), Cosslett (1993), Kakwani (1993), Selden (1994), Bishop, Chakraborti y Thistle (1994),), Korn y Graubard (1995a,b), Imbens y Lancaster (1996), Magee, Robb y Burbidge (1998), Wooldridge (1999)), estos datos típicamente llevan asociado un peso relacionado en alguna medida con la probabilidad de que dicha observación sea incluida en la muestra y la cuestión de interés ha sido el tratamiento adecuado de estos pesos al objeto de lograr estimadores consistentes y eficientes con los que poder realizar inferencia acerca de los parámetros de la población. Esta literatura suele ser cuidadosa en la descripción de los procesos de muestreo que han dado lugar a las observaciones disponibles (Cosslett (1993), Selden (1994), Imbens y Lancaster (1996), Wooldridge (1999)), ya que las características de los datos y sus pesos dependen de dicho proceso y por tanto los estimadores propuestos, así como sus propiedades, varían en función de la información disponible acerca del muestreo utilizado.

Nuestra muestra de referencia, la renta *per capita* de regiones que engloban a varios individuos, no procede, sin embargo, de ninguna encuesta, no ha sido obtenida mediante ningún proceso de muestreo, simplemente disponemos de un conjunto de observaciones y pretendemos caracterizar la distribución de la variable que representan tales observaciones, $\phi(x)$. Si dicha distribución la consideramos en términos de las rentas *per capita* de las regiones individuales entonces podemos suponer que disponemos de una muestra de observaciones independientes e idénticamente distribuidas (*i.i.d.*), los estadísticos que debemos calcular son estadísticos simples y la inferencia puede proceder de forma estándar. Por otra parte si la distribución que deseamos analizar es la distribución de x en términos de los individuos que hay detrás de cada renta *per capita* regional entonces las cosas no son tan sencillas, puesto que las regiones difieren en población cada observación muestral tiene una diferente representatividad dentro de la población de forma que podemos seguir suponiendo que las observaciones son independientes pero no idénticamente distribuidas (*i.n.i.d.*). Fue esta observación la que motivó los comentarios del epígrafe anterior y aunque es razonable en este caso describir la población mediante el cálculo de estadísticos ponderados queda por resolver la cuestión de como realizar inferencia sobre la población con este tipo de muestras. Es decir tratamos a continuación de responder a preguntas tales como: ¿podemos realizar contrastes sobre la media de la distribución mediante los procedimientos estándar?, ¿podemos contrastar la simetría o la normalidad mediante los estadísticos habituales (Jarque y Bera (1980)) en los que cualquier

momento poblacional es sustituido por el correspondiente momento muestral ponderado?. Los argumentos que ofrecemos a continuación responden, bajo ciertas condiciones, **afirmativamente** a estas cuestiones, de forma que en ciertos casos la inferencia puede proceder de forma similar a situaciones estándar[10].

Con muestras ponderadas la correcta utilización de los pesos y las propiedades de los estimadores dependen crucialmente del proceso de muestreo y de lo que supongamos acerca de la población subyacente (DuMouchel y Duncan (1983), Cosslett (1993)), por lo tanto para responder a las preguntas anteriores deberemos ser específicos acerca de estas cuestiones para nuestra muestra de referencia. Sin embargo el tipo de muestra que utilizamos en este informe no parece haber sido analizado por la literatura estadística y/o econométrica una vez incorporamos el hecho de que cada observación tiene una representatividad diferente para la población, es por ello que la mejor forma de pensar en el problema es tratar de adecuar nuestra muestra a los resultados existentes en la literatura sobre datos de encuesta con muestreo no aleatorio, de forma que deberemos distinguir entre la distribución de la población y la distribución de acuerdo con la cual los datos han sido generados. Nuestra **muestra** está constituida por observaciones de áreas geográficas, donde cada observación lleva asociada una **frecuencia muestral,** que viene dada por **1/*n***, en el caso de las provincias españolas $n = 50$ con lo que la frecuencia muestral sería del 2% y además es constante en el tiempo; mientras que la **población** está constituida por los individuos que habitan las áreas geográficas, $N = \Sigma_{i=1}^{n} N_i$, cada observación lleva asociada una **frecuencia poblacional**, que refleja la importancia de dicha observación en la población, en nuestro caso la frecuencia poblacional viene dada por las proporciones de población, $p_i = N_i/N$, que son variables en el tiempo. Este esquema puede ser visto como un proceso de **muestreo estratificado estándar** (Cosslett (1993), Imbens y Lancaster (1996), Wooldridge (1999)) en el que hay tantos estratos como observaciones, disponemos de una sola observación por estrato y en el que las proporciones de población coinciden con la importancia del estrato (y de la observación) dentro de la población. Por tanto, en

[10] Disponemos de una sola variable de forma que estamos interesados en este epígrafe en los momentos que caracterizan a $\phi(x)$ y en inferencia estadística sobre dichos momentos al objeto de concluir algo acerca de la forma de $\phi(x)$. La cuestión de la utilización de las ponderaciones en modelos de regresión y la inferencia estadística asociada a dichos modelos es notablemente más compleja (Cosslett (1993), Imbens y Lancaster (1996) y Wooldridge (1999)).

nuestro caso pensamos en la observación *i*-ésima como extraída aleatoriamente de una subpoblación de tamaño N_i (Magee, Robb y Burbidge (1998)). El **resultado** es una muestra de **observaciones independientes pero no idénticamente distribuidas**.

En esta situación la densidad de probabilidad de[11] x_i, $\phi_i(x_i)$, no coincide con la densidad de probabilidad de la población subyacente, $\phi(x)$, pero los momentos de esta última distribución pueden ser estimados de forma consistente mediante momentos muestrales ponderados con ponderaciones p_i; es decir si θ es un parámetro de la distribución de x en la población y $g(x,\theta)$ es una función dependiente de x y de θ tal que $E[g(x,\theta)] = 0$ en la población[12], entonces en nuestro contexto el estimador ponderado de momentos obtenido al resolver la ecuación

$$\Sigma_{i=1}^{n} p_i g(x_i, \hat{\theta}) = 0$$

es un estimador consistente para θ, $\hat{\theta} \xrightarrow{p} \theta$, un parámetro de la distribución de la población. (Wooldridge (1999, p.-1401))[13]. Esta es una fundamentación estadística que justifica la utilización de momentos ponderados para caracterizar la distribución de x en términos de la población subyacente a nuestras observaciones, así para el caso de la **media** tomamos $g(x,\theta) = x - \theta$, de lo que resulta $\hat{\theta} = \mu = \Sigma_{i=1}^{n} p_i x_i$.

Puesto que consistencia es una propiedad de grandes muestras antes de resolver la cuestión de la inferencia estadística deberemos considerar una regla para extender nuestra muestra de forma indefinida. Este no es un problema que se plantee la literatura sobre datos de encuesta donde es fácil pensar en términos de muestreo a partir de una población infinita, sin embargo en nuestro caso, en el que disponemos de observaciones de un conjunto dado de regiones, es difícil imaginar cualquier regla que permita extender el

[11] x_i es ahora no una observación sino una variable aleatoria de la que sólo dispondremos de una realización.
[12] La esperanza debe ser entendida de acuerdo con la densidad de la población.
[13] En realidad las ponderaciones son el cociente entre la frecuencia poblacional y la frecuencia muestral, en nuestro caso $n.p_i$, pero dado que la frecuencia muestral es constante $\forall i$ desaparece en el proceso de estimación que iguala momentos muestrales ponderados a momentos poblacionales.

proceso generador de datos a muestras arbitrariamente grandes[14]. En la práctica haremos uso de la ficción de las **muestras repetidas** para nuestro proceso generador de datos (Davidson y MacKinnon (1993)), de forma que si la muestra observada fuera de tamaño m, consideraremos muestras de tamaño $n = k.m$, $k = 1,2,3,....$ Los resultados asintóticos que mencionaremos mantendrán la distribución de la población, $\phi(x)$, fija y permitiremos que k crezca de forma indefinida con lo que tanto el tamaño de la muestra, n, como el de la población, $N = \Sigma_{i=1}^{n} N_i$, crecerán de forma arbitrariamente grande, manteniendo constantes los pesos asignados a cada observación, $N_i = N_{i+(k-1)m}$, $i = 1,2,3,...,m$; en efecto este supuesto no hace mas que replicar nuestra población de referencia, manteniendo intactas las propiedades de las observaciones en relación a su representatividad respecto a la población. Una ficción conveniente para realizar análisis asintótico y justificar la inferencia estadística; por supuesto en una aplicación concreta n es fijo y dado, y por tanto $k = 1$.

Finalmente deberemos establecer la relación entre la densidad de probabilidad de cada variable individual, x_i, $\phi_i(x_i)$[15], y la densidad de probabilidad de la población subyacente, $\phi(x)$; si el muestreo fuera aleatorio (*i.i.d.*) estas dos distribuciones serían idénticas, $\phi_i(x_i) = \phi(x)\ \forall i$, y en consecuencia cualquier momento de $\phi_i(x_i)$ sería idéntico a cualquier momento de $\phi(x)$, con lo que momentos poblacionales pueden ser estimados mediante momentos muestrales y la inferencia puede proceder de forma estándar. En nuestro caso, en el que cada observación tiene un contenido informativo diferente acerca de la población, necesitamos dos requerimientos (Stigler (1974), Teorema 6 y ejemplo 5.5):

$$(i) \qquad \lim_{n\to\infty} \frac{\Sigma_{i=1}^{n} \phi_i(x_i)}{n} = \phi(x)$$

de forma que la densidad de la población subyacente, $\phi(x)$, tenga sentido, y

$$(ii) \qquad \phi_i\left(g(x_i)\right) = \phi\left(np_i g(x_i)\right) \qquad \forall i$$

[14] Dicho de otra forma, el número de regiones de un país es limitado y el número de paises de la Tierra es un número finito no muy grande, por no mencionar el conjunto de países de la OCDE, de la Unión Europea o de un continente. Por citar algunos ejemplos del tipo de muestras que estamos considerando.

[15] De la cual no hay forma de inferir nada, al menos sin una mayor desagregación en los datos, ya que sólo dispondremos de una sola realización proveneniente de dicha distribución.

es decir que las densidades de probabilidad de cualquier función de cada variable individual, $\phi_i(g(x_i))$, pertenezcan todas a la misma familia y que sean idénticas una vez la función $g(x_i)$ ha sido ajustada por un factor de proporcionalidad, siendo el factor de proporcionalidad la *ratio* entre la frecuencia poblacional, p_i, y la frecuencia muestral, $1/n$, con lo que obtenemos el factor $n.\ p_i$.

Estas condiciones son más fuertes de lo necesario pero son suficientes para garantizar la inferencia por los métodos habituales simplemente sustituyendo momentos poblacionales por momentos muestrales ponderados. Debe observarse que estas condiciones no son normalmente satisfechas por los procedimientos de muestreo estándar (Wooldridge (1999)) pero si pueden ser mantenidas en nuestro caso.

Una forma de entender la intuición de este factor de escala consiste en observar que puesto que suponemos que la observación *i*-ésima ha sido extraída aleatoriamente de una subpoblación de tamaño N_i es natural inflar la contribución de x_i por este factor en la población, pero puesto que sólo disponemos de n observaciones esta contribución debe ser escalada por la *ratio* entre muestra y población, n/N[16]. De esta forma si $p_i = 10\%$ y $n = 50$ la contribución de x_i en la población es escalada por 5. Obsérvese que no se trata de un caso de corrección por heterocedasticidad, como algunos autores sugieren (Beach y Kaliski (1986, p.-41)). Además si $N_i = 1,\ \forall i$, el muestreo puede ser considerado como aleatorio, en cuyo caso los requerimientos (*i*) y (*ii*) anteriores son superfluos, puesto que $n.p_i = 1,\ \forall i$.

Como hemos mencionado las condiciones anteriores son suficientes para que la inferencia pueda ser realizada de forma estándar. Por ejemplo, $\mu = \Sigma_{i=1}^{n} p_i x_i$ es un estimador consistente de la media de la población, digamos θ; si deseamos realizar inferencia acerca de la media de la distribución poblacional de x necesitamos derivar la distribución asintótica de μ, observando que $\mu = \frac{1}{n}\Sigma_{i=1}^{n} n p_i x_i$ y que los requerimientos anteriores implican

[16] Ver Imbens y Lancaster (1996) y Wooldridge (1999) para el caso de muestreo multinomial.

$$Var\left(\frac{\Sigma_{i=1}^{n} np_i x_i}{\sqrt{n}}\right) = \frac{\Sigma_{i=1}^{n} Var(np_i x_i)}{n} = \frac{\Sigma_{i=1}^{n} Var_i(x_i)}{n} \rightarrow Var(x) = \sigma^2$$

lo que nos permite derivar el resultado estándar ya conocido.

$$\sqrt{n}(\mu - \theta) \xrightarrow{d} N(0, \sigma^2)$$

Inferencia acerca de θ procede pues mediante los métodos habituales, sustituyendo σ^2 por un estimador consistente de este parámetro, la varianza ponderada de las observaciones. El mismo argumento funciona para momentos de orden más elevado de forma que simetría o normalidad podrían ser contrastadas con los estadísticos estándar y sus distribuciones derivadas bajo muestreo aleatorio (Jarque y Bera (1980)), simplemente sustituyendo momentos simples por momentos ponderados. En cualquier caso el énfasis en este trabajo radica más en la descripción de $\phi(x)$ que en inferencia acerca de esta distribución.

2.3. ¿Que estadísticos descriptivos constituyen nuestro objeto de interés?

Este epígrafe ofrece una descripción pormenorizada de estadísticos descriptivos, aunque en su mayor parte se trata de estadísticos habituales de posición, dispersión y orden, y cuya discusión puede encontrarse en los libros de estadística tradicionales (Mood, Graybill y Boes (1974)), la consideración simultánea de estadísticos ponderados y simples hace conveniente una exposición de los mismos con una nomenclatura unificada. Los momentos de nuestra variable, x, serán definidos en términos ponderados utilizando como frecuencias relativas los porcentajes de población[17], p_i; los correspondientes momentos simples se obtendrán dando el mismo peso a cada observación, es decir $N_i = 1, \ \forall i$, con lo que $N = n$ y $p_i = 1/n, \ \forall i$; en consecuencia los momentos simples utilizarán como divisor el número de observaciones, n, de forma que no se incorporan ajustes por grados de libertad.

[17] Otras ponderaciones son posibles y de hecho los resultados mencionados son válidos bajo un conjunto arbitrario de ponderaciones.

- **Media**: es la medida de **posición** por excelencia.

La media de *x* es una medida alrededor de la cual los valores de la variable están "centrados", si no conocemos nada acerca de la distribución de *x* la media nos da una idea de la posición de la variable en cuestión. Otras medidas de posición serán consideradas en relación a los estadísticos de orden.

En realidad la media ya ha sido definida al introducir la nomenclatura, obsérvese que sólo la media ponderada coincide con la media del agregado, que en la práctica es observable:

MEDIA:	**ponderada**	**simple**	
	$\mu = \Sigma_{i=1}^{n} p_i x_i$	$\bar{x} = \frac{\Sigma_{i=1}^{n} x_i}{n}$	(2)

- **Desviación típica**: es la medida de **dispersión absoluta** más habitual.

La desviación típica se define a partir de la **varianza**, que no es más que el segundo momento central alrededor de la media, como la raíz cuadrada positiva de la misma.

VARIANZA:	**ponderada**	**simple**[18]	
	$Var_{\omega}(x) = \Sigma_{i=1}^{n} p_i (x_i - \mu)^2$	$Var(x) = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}{n}$	(3)

Para dos distribuciones con la misma media una disminución de la varianza implica una mayor concentración de la masa de probabilidad entorno a la media, al menos para ciertos intervalos alrededor de dicha media, pero ello no nos dice necesariamente nada acerca lo que sucede en las colas de la distribución.

---

[18] En ocasiones la varianza simple incorpora un ajuste por grados de libertad, $Var_s(x) = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$.

Los ajustes por grados de libertad mencionados están pensados de forma que los estadísticos muestrales constituyan estimadores insesgados de los parámetros poblacionales subyacentes; por razones obvias estos ajustes solo se pueden realizar en el caso de estadísticos simples.

La **varianza** es una medida de dispersión de los valores de una variable alrededor de la media y puesto que su cálculo implica elevar al cuadrado las desviaciones respecto a la media la varianza posee como unidad de medida el cuadrado de las unidades de *x*, razón por la cual es normalmente más conveniente utilizar la **desviación típica** como medida de dispersión, puesto que este estadístico tendrá las mismas unidades de medida que *x*.

DESVIACIÓN	**ponderada**	**simple**[19]	
TÍPICA:	$SD_{\omega}(x) = +\sqrt{Var_{\omega}(x)}$	$SD(x) = +\sqrt{Var(x)}$	(4)

Otras medidas de dispersión serán consideradas en relación a los estadísticos de orden.

Para futuras referencias conviene definir los **momentos de orden *r***, que son simplemente la media de las potencias de los valores de la variable original, **momentos respecto al origen o simplemente momentos**, o la media de las potencias de los valores de la variable en desviaciones respecto a un determinado valor, **momentos centrales**.

MOMENTOS:	**ponderados**	**simples**	
	$\mu'_r = \Sigma_{i=1}^{n} p_i x_i^r$	$m'_r = \frac{\Sigma_{i=1}^{n} x_i^r}{n}$	(5)

Observamos que $\mu'_1 = \mu$ y $m'_1 = \bar{x}$, la media de *x*.

Las potencias de *x* pueden centrarse en un valor determinado y obtener de esta forma los denominados **momentos centrales**,

MOMENTOS	**ponderados**	**simples**	
CENTRALES:	$\mu_r(a) = \Sigma_{i=1}^{n} p_i (x_i - a)^r$	$m_r(b) = \frac{\Sigma_{i=1}^{n} (x_i - b)^r}{n}$	(6)

[19] Si la varianza simple incorpora un ajuste por grados de libertad entonces la desviación típica viene dada por $SD_s(x) = +\sqrt{Var_s(x)}$.

de donde observamos como los momentos respecto al origen se obtienen fijando $a = b = 0$, $\mu_r(0) = \mu'_r$ y $m_r(0) = \bar{x}'_r$.

Si $a = \mu$ y $b = \bar{x}$ obtenemos los **momentos centrales respecto a la media**, que son los más habituales,

MOMENTOS	**ponderados**	**simples**	
CENTRALES:	$\mu_r = \Sigma_{i=1}^{n} p_i (x_i - \mu)^r$	$m_r = \frac{\Sigma_{i=1}^{n} (x_i - \bar{x})^r}{n}$	(7)

Observamos que $\mu_1 = 0$ y $m_1 = 0$, y que $\mu_2 = Var_\omega(x)$ y $m_2 = Var(x)$, la varianza de x[20].

Es importante observar además que si las observaciones están distribuidas de forma simétrica en torno a la media entonces todos los momentos centrales respecto a la media de orden impar son nulos, para el caso de μ_r ello requiere no sólo que las observaciones estén distribuidas de forma simétrica sino que también lo estén sus frecuencias relativas.

- **Coeficiente de Variación**: es la medida de **dispersión relativa** más habitual.

Como mencionamos en Goerlich (1998) la desviación típica no es invariante respecto a la escala y una forma de solucionar esta cuestión es dividir este estadístico por la media, el resultado es el denominado **coeficiente de variación**.

COEFICIENTE	**ponderado**	**simple**	
DE VARIACIÓN:	$CV_\omega(x) = \frac{SD_\omega(x)}{\mu}$	$CV(x) = \frac{SD(x)}{\bar{x}}$	(8)

que no está definido cuando la media es cero y cuya significación no está del todo clara cuando la variable puede tomar **valores negativos**, ya que en este caso obtendríamos una

[20] En el caso de la varianza simple el ajuste por grados de libertad mencionado anteriormente puede ser escrito en función de los momentos como $Var_s(x) = \frac{n}{n-1} . m_2 = k_2$, donde k hace referencia a los llamados estadísticos-k (Fisher (1929), Kendall y Stuart (1977, Cap.-12), Spanos (1999, Cap.-13.2.1)).

medida de dispersión negativa. Aunque este no es nuestro caso si puede plantearse en general y en la práctica esto se obvia considerando el **valor absoluto** del coeficiente de variación.

El coeficiente de variación es uno de los estadísticos más habituales para medir el concepto de σ-convergencia, que como ya hemos mencionado se preocupa de la dispersión en la distribución, precisamente por ser invariante respecto a la escala; vale la pena observar, sin embargo, que la concentración de la distribución en un punto[21], lo que exige que $SD(x)\to 0$, es condición suficiente para que $CV(x)\to 0$, pero no es condición necesaria, ya que esto puede suceder si $\mu\to\infty$, aunque $SD(x)$ permanezca estable o incluso crezca pero a una tasa menor que μ. Este comentario, que se aplica a todas las medidas de desigualdad relativa examinadas en Goerlich (1998), debe ser tenido presente cuando se examinan resultados concretos ya que periodos de crecimiento generalizado pueden ser vistos como periodos de intensa convergencia y lo que puede estar sucediendo es simplemente que el nivel de vida agregado crezca sin cesar aunque las diferencias entre las unidades económicas se mantengan.

Como observamos en Goerlich (1998) el cuadrado del **coeficiente de variación** es **cardinalmente equivalente** al **índice de Theil** (1967) con **parámetro igual a 2**, $T(2)=\frac{1}{2}CV_{\omega}(x)^2$[22].

Hasta ahora nos hemos centrado en los dos primeros momentos de una variable que nos ofrecen una idea de la posición y dispersión de la misma, adicionalmente los momentos centrales respecto a la media de orden tres y cuatro son útiles para examinar diversas características de la densidad de probabilidad de x, $\phi(x)$, pero examinaremos primero los denominados **estadísticos de orden** y funciones de los mismos que nos permiten observar otras características interesantes de $\phi(x)$.

[21] Lo que en términos estadísticos llamaríamos convergencia puntual (*pointwise*) en probabilidad.
[22] La versión simple del coeficiente de variación, $CV(x)$, sería cardinalmente equivalente a la versión simple del índice de Theil, $T(2)$.

- **Estadísticos de orden**: Dadas nuestras observaciones de renta *per capita* regional, $\{x_i\}_{i=1}^n$, una **ordenación no decreciente de** dichas **observaciones**, $x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq ... \leq x_{(n-1)} \leq x_{(n)}$, constituyen los denominados **estadísticos de orden**, donde el paréntesis en los subíndices indica que las observaciones han sido ordenadas en la forma indicada.

Los estadísticos de orden no tienen en cuenta, en principio, las frecuencias relativas de cada observación, pero lógicamente si queremos examinar las características de $\phi(x)$ en términos de individuos, las ponderaciones, $\{p_i\}_{i=1}^n$, deberán ser introducidas en el análisis; de esta forma paralelamente a la ordenación de x consideraremos la ordenación de las frecuencias relativas, $\left(p_{(1)}, p_{(2)}, p_{(3)}, ..., p_{(n-1)}, p_{(n)}\right)$, donde dicha ordenación se corresponde con la derivada para x, es decir $p_{(i)}$ es la proporción de población de la región que ocupa la posición i-ésima en nuestra muestra ordenada de forma no decreciente.

Varios estadísticos de orden son útiles en la caracterización de $\phi(x)$, en primer lugar debemos mencionar los **estadísticos de valor extremo**, es decir el **mínimo**, $x_{(1)} = \min \{x_i\}_{i=1}^n$, y el **máximo**, $x_{(n)} = \max \{x_i\}_{i=1}^n$, de los valores observados, que además de ser útiles en sí mismos nos permiten definir una medida alternativa de dispersión, el **rango**[23],

RANGO: $$R(x) = x_{(n)} - x_{(1)} = \max\{x_i\}_{i=1}^n - \min\{x_i\}_{i=1}^n \quad (9)$$

y una medida alternativa de posición, el **medio-rango**,

MEDIO-RANGO: $$Mid - R(x) = \frac{x_{(1)} + x_{(n)}}{2} = \frac{\min\{x_i\}_{i=1}^n + \max\{x_i\}_{i=1}^n}{2} \quad (10)$$

[23] Tal y como está definido el rango no depende de las frecuencias relativas, además como se observa en Goerlich (1998) el **rango** podría ser normalizado respecto a varios estadísticos de interés para hacer su intervalo de variación más interpretable, siendo los más obvios la media o los propios valores máximo o mínimo. Obsérvese además que si en lugar de considerar la variable x consideramos z entonces obtenemos lo que en Goerlich (1998, p.-22) se denomina el **rango relativo**.

Ninguno de estos dos estadísticos, el **rango** y el **medio-rango**, dependen de las frecuencias relativas e igualmente ignoran todo lo que sucede entre los valores extremos.

Otra medida de posición alternativa a la media es la **mediana**[24], *Med*(*x*), que se define como el **estadístico de orden que divide la distribución de *x*, $\phi(x)$, en dos partes con igual probabilidad** en cada una de ellas, de forma que el 50% de la masa de probabilidad estará por debajo de la mediana y el 50% restante por encima. Para una distribución simétrica la mediana coincide con la media.

En el caso de una **muestra simple** la **mediana** es simplemente **el estadístico de orden que divide la muestra en dos partes iguales**, es decir la **observación central**; de forma que si *n* es impar la mediana viene dada por $x_{((n+1)/2)}$, ya que este valor deja a izquierda y derecha el mismo número de observaciones, mientras que si *n* es par la mediana se define convencionalmente como la media entre los dos valores centrales, $\frac{x_{(n/2)} + x_{((n/2)+1)}}{2}$ [25].

En el caso de muestras simples todas las observaciones tienen asignado el mismo peso y por tanto dividir la muestra en dos partes iguales es equivalente a distribuir la masa de probabilidad de forma simétrica. Sin embargo ello no es así si queremos obtener la **mediana** para una **muestra ponderada**, en este caso cada observación, x_i, lleva asociada una frecuencia relativa, p_i; el problema es por tanto ligeramente diferente, ahora no se trata de dividir las observaciones sino de dividir la masa de probabilidad que representan dichas observaciones, de forma que **la mediana no puede definirse directamente a partir de las observaciones**, es necesario invertir el proceso, en este caso debemos acumular los $p_{(i)}$, $F_s = \Sigma_{i=1}^{s} p_{(i)}$, $s = 1,2,...,n$, y buscar el valor *s* tal que $F_s = 0.50$, si dicho valor existe podría ser utilizado para definir la mediana, $x_{(s)}$.

---

[24] El término "mediana" fue utilizado por primera vez por Galton (1883).

[25] Si *n* es par entonces *n*/2 es un número entero y cualquier valor en el intervalo cerrado $\left[x_{(n/2)}, x_{((n/2)+1)}\right]$ puede ser utilizado para definir la mediana (Patel y Read (1982), p.-261), convencionalmente tomamos el valor medio (Kendall y Stuart (1977) p.-39) pero obsérvese que cualquier otro valor del intervalo (abierto) dividiría la muestra en dos partes iguales al contener cada una de ellas idéntico número de observaciones; por otra parte el valor medio es el valor natural ya que es lo que obtenemos si interpolamos linealmente entre ambas observaciones.

En la práctica sin embargo este no es un procedimiento totalmente adecuado para la obtención de la mediana ya que aunque existiera un valor exacto s tal que $F_s = 0.50$ encontraríamos un resultado diferente si empezamos a contar la probabilidad asociada a las observaciones por la parte inferior de la distribución, $x_{(1)}$, o por la parte superior, $x_{(n)}$; este no es por tanto un procedimiento simétrico. Además en la práctica un valor exacto s tal que $F_s = 0.50$ será la excepción y no la regla por lo que será necesario arbitrar **algún esquema de interpolación** para las observaciones en el entorno de $F_s = 0.50$. El procedimiento utilizado busca el valor s tal que $F_{s-1} < 0.50$ y $F_s \geq 0.50$ y distribuye linealmente $p_{(s)}$ a lo largo del intervalo comprendido entre los puntos medios entre la observación (s)-ésima y sus dos observaciones adyacentes, (s−1) y (s+1)[26], lo que es equivalente a asignar el valor de $p_{(s)}$ al final de dicho intervalo, para posteriormente obtener el valor de la **mediana** por interpolación lineal entre los puntos $\left(\frac{x_{(s-1)} + x_{(s)}}{2}, F_{s-1} \right)$ y $\left(\frac{x_{(s)} + x_{(s+1)}}{2}, F_s \right)$, dado el valor de s tal que $F_{s-1} < 0.50$ y $F_s \geq 0.50$.

Una tercera medida de posición es la **moda**, *Mode*(x), que se define como el **valor de x, si existe, para el cual $\phi(x)$ alcanza su valor máximo**. Como estadístico descriptivo calculado para variables continuas y a partir de observaciones simples carece de utilidad ya que en la práctica nunca observamos dos valores de x exactamente iguales, sin embargo si consideramos estadísticos ponderados la moda vendrá dada por el valor que alcance mayor frecuencia relativa, que en el caso de las provincias españolas sería Barcelona entre 1951 y 1977 y Madrid entre 1978 y 1998. Aún en este caso su utilidad es muy limitada; la moda será importante en la sección siguiente cuando estimemos $\phi(x)$ de forma no paramétrica.

Hemos visto que la **mediana** divide la distribución de x, $\phi(x)$, en dos partes con igual probabilidad en cada una de ellas, no hay motivo sin embargo para restringirse a que estas dos partes sean iguales, y podemos buscar estadísticos de orden que dividan la distribución de x de forma asimétrica. Esta idea la recogen los denominados **quantiles**, el **quantil de orden p, ξ_p, se define como el estadístico de orden, ξ, que divide la distribución de x, $\phi(x)$, en dos partes tal que $\Phi(\xi) = p$, $0 \leq p \leq 1$, siendo $\Phi(\bullet)$ la función**

[26] Este procedimiento es válido siempre y cuando $1 < s < n$, cuando $s = 1$ se toma como límite inferior del intervalo $x_{(1)}$ y cuando $s = n$ se toma como límite superior del intervalo $x_{(n)}$.

de distribución acumulativa de *x*, $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$; es decir el *p*-% de la masa de probabilidad estará por debajo del quantil de orden *p*, ξ_p, y el (1-*p*)-% restante por encima. Por tanto la **mediana** no es más que el **quantil de orden 0.5,** $\xi_{.5} = Med(x)$, el **mínimo** puede ser considerado como el **quantil de orden 0**, $\xi_{0.0} = x_{(1)} = \min \{x_i\}_{i=1}^{n}$, y el **máximo** como el **quantil de orden 1**, $\xi_{1.0} = x_{(n)} = \max \{x_i\}_{i=1}^{n}$.

Varios quantiles son habituales en la literatura estadística, los tres estadísticos de orden que dividen la distribución de *x*, $\phi(x)$, en **cuatro partes iguales** son los denominados **cuartiles**, correspondientes a $p = 0.25,\ 0.50$ y 0.75; cuatro estadísticos de orden que dividen la distribución de *x* en **cinco partes iguales** son los denominados **quintiles**, correspondientes a $p = 0.2,\ 0.4,\ 0.6$ y 0.8; **nueve estadísticos de orden** que dividen la distribución de *x* en diez partes iguales son los denominados **deciles**, correspondientes a $p = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,\ 0.6,\ 0.7,\ 0.8$ y 0.9; 19 estadísticos de orden que dividen la distribución de *x* en **20 partes iguales** son los denominados **veintiles**, correspondientes a valores de p en incrementos de 0.05; y finalmente 99 estadísticos de orden que dividen la distribución de *x* en **100 partes iguales** son los denominados **percentiles**[27], correspondientes a valores porcentuales de *p*. Obviamente el conocimiento de un número suficientemente elevado de quantiles proporciona una idea bastante buena de la forma de $\phi(x)$ razón por la cual estos estadísticos son importantes[28].

En el caso de una **muestra simple** la obtención de los **quantiles** se basa en buscar el **estadístico de orden que divide la muestra en las dos partes adecuadas**, obviamente para un conjunto de observaciones siempre hay una pequeña indeterminación que puede ser resuelta de forma similar al caso de la definición práctica de la mediana. El procedimiento

---

[27] Los percentiles fueron definidos por Galton (1885).

[28] Algunos autores (Mills (1990), p.-21-26) han propuesto extender el concepto de mediana a partir de ir dividiendo por la mitad sucesivamente los intérvalos de observaciones que quedan despues de calcular la mediana; es decir, una vez calculada la mediana se obtiene la mediana para las observaciones entre el mínimo y la mediana y otra mediana para las observaciones entre la mediana y el máximo, en la practica ello equivale apróximadamente al cálculo de los cuartiles $\xi_{.25}$ y $\xi_{.75}$. Este proceso de ir calculando sucesivas medianas puede hacerse recursivo y proporciona una caracterización de $\phi(x)$ idéntica a la ofrecida por los quantiles; un escalón más en el proceso de ir calculando sucesivas medianas sería aproximadamente equivalente a la obtención de $\xi_{.125}$ y $\xi_{.875}$, y procediendo recursivamente ello equivaldría aproximadamente a calcular $\xi_{.0625}$ y $\xi_{.9375}$ y a continuación $\xi_{.03125}$ y $\xi_{.96875}$, y así sucesivamente.

empleado primero determina $np = p_x(n-1) + 1$ y luego calcula el quantil correspondiente por interpolación lineal entre $x_{([np])}$ y $x_{([np]+1)}$, donde $[np]$ es el mayor entero menor o igual a np[29], es decir,

QUANTILES: $$\xi_p = (1-(np-[np])).x_{([np])} + (np-[np]).x_{([np]+1)} \qquad (11)$$

el procedimiento distribuye de forma uniforme la probabilidad teniendo en cuenta que para n observaciones sólo disponemos de $n-1$ huecos entre las mismas[30]. Obsérvese que $np-[np]$ no es más que la parte fraccional de np y que para $p = 0.5$ obtenemos la fórmula para la mediana mencionada anteriormente.

En el caso de una **muestra ponderada** la obtención de los **quantiles no puede proceder a partir de las observaciones** por la misma razón que la mediana no podía definirse directamente a partir de dichas observaciones, en este caso cada x_i lleva asociada una frecuencia relativa, p_i, por lo que deberemos proceder a obtener los quantiles a partir de la función de distribución acumulativa empírica, es decir a partir de la acumulación de $p_{(i)}$, $F_s = \Sigma_{i=1}^{s} p_{(i)}$, $s = 1,2,...,n$. Dado un valor $0 \le p \le 1$ podemos buscar el valor entero s tal que $F_s = p$, si dicho valor existe podría ser utilizado para definir el quantil de orden p, ξ_p.

Por las mismas razones que expusimos al hablar de la mediana este no es un procedimiento totalmente adecuado ya que no es simétrico y además no es de esperar que encontremos un valor exacto de s tal que $F_s = p$, por lo tanto será necesario arbitrar algún **esquema de interpolación** para las observaciones en el entorno de $F_s = p$. El procedimiento utilizado es idéntico al que mencionamos para la mediana y se basa en distribuir linealmente $p_{(s)}$ a lo largo del intervalo comprendido entre los puntos medios entre la observación (s)-ésima y sus dos observaciones adyacentes, ($s-1$) y ($s+1$), lo que equivale a asignar el valor de $p_{(s)}$ al final de dicho intervalo, buscar el valor entero s tal que

---

[29] [•] debe ser leido como la "parte entera de" y denota la operación de eliminar la parte fraccional.

[30] Este no es el único procedimiento práctico para calcular quantiles a partir de un conjunto de observaciones, aunque es el más lógico. Patel y Read (1982, p.-261) proponen un procedimiento alternativo pensado básicamente en distribuir observaciones a ambas partes del quantil más que en distribuir de forma continua la probabilidad a lo largo del rango de variación de x. Según esta regla $np = p_x n$ de forma que si np no es entero, entonces el estadístico de orden $x_{([np]+1)}$ es el quantil de orden p, mientras que si np es entero, entonces se toma como quantil de orden p la mitad entre $x_{([np])}$ y $x_{([np]+1)}$. Obsérvese que esta regla proporciona el mismo valor para la mediana que la regla mencionada en el texto.

$F_{s-1} < p$ y $F_s \geq p$ y finalmente obtener el **quantil de orden *p*** por interpolación lineal entre los puntos $\left(\frac{x_{(s-1)} + x_{(s)}}{2}, F_{s-1}\right)$ y $\left(\frac{x_{(s)} + x_{(s+1)}}{2}, F_s\right)$, dado s tal que $F_{s-1} < p$ y $F_s \geq p$[31]. Para $p = 0.5$ obtenemos la mediana tal y como fue definida anteriormente[32].

El esquema de interpolación que acabamos de mencionar no funciona para las observaciones extremas, $x_{(1)}$ y $x_{(n)}$, ya que en este caso no podemos distribuir la probabilidad por debajo de $x_{(1)}$ ni por encima de $x_{(n)}$, si queremos mantenernos dentro del rango de variación de x. Así pues para el primer hueco entre observaciones $p_{(1)}$ es distribuido entre $x_{(1)}$ y $\frac{x_{(1)} + x_{(2)}}{2}$, de forma que si $p_{(1)} = F_1 \geq p$ obtenemos el quantil correspondiente por interpolación lineal entre $\left(x_{(1)}, 0\right)$ y $\left(\frac{x_{(1)} + x_{(2)}}{2}, p_{(1)}\right)$. De forma simétrica para el último hueco entre observaciones $p_{(n)}$ es distribuido entre $\frac{x_{(n-1)} + x_{(n)}}{2}$ y $x_{(n)}$, de forma que si $F_{n-1} < p$ obtenemos el quantil correspondiente por interpolación lineal entre $\left(\frac{x_{(n-1)} + x_{(n)}}{2}, F_{n-1}\right)$ y $\left(x_{(n)}, 1\right)$.

Este es un procedimiento que debe proporcionar resultados razonables a menos que la muestra sea pequeña, los valores de $p_{(1)}$ o $p_{(n)}$ sean muy elevados y estemos interesados en los quantiles en las colas de la distribución. Su principal inconveniente es que si fijamos $N_i = 1,\ \forall i$, entonces no obtenemos los mismos resultados que la regla para la obtención de quantiles en el caso de muestras simples como consecuencia de la asimetría en el tratamiento de la probabilidad en los extremos de la distribución; sin embargo ambos

[31] Otros procedimientos de interpolación como el *kernel smoothing* (suavizado) analizado en la sección siguiente serían posibles.

[32] Este no es el único procedimiento práctico para calcular quantiles a partir de una muestra ponderada. Un procedimiento que imita la regla de Patel y Read (1982, p.-261) para observaciones mencionada anteriormente tomaría como quantil de orden p el estadístico de orden $x_{(s)}$ tal que $F_{s-1} < p$ y $F_s \geq p$. Con datos de encuesta en el que el número de observaciones es muy elevado los procedimientos de interpolación no deben afectar mucho a la obtención de los quantiles pero con datos regionales y/o de paises parece razonable utilizar reglas que interpolen entre observaciones y sus probabilidades asociadas.

procedimientos son asintóticamente equivalentes en el sentido de que si fijamos $N_i = 1$, $\forall i$, entonces ambas reglas proporcionarán los mismos resultados conforme $n \to \infty$.

Finalmente señalar que una forma útil de inspeccionar visualmente los quantiles consiste en dibujar la **función de distribución acumulativa empírica de probabilidad** (Mood, Graybill y Boes (1974), p.-264), es decir un gráfico-XY de $F_s = \Sigma_{i=1}^{s} p_{(i)}$ frente $x_{(s)}$, $s = 1,2,...,n$, en el caso ponderado, o de $\frac{s}{n}$ frente $x_{(s)}$, $s = 1,2,...,n$, en el caso simple. Volveremos sobre esta función en la sección siguiente, cuando consideremos explícitamente el procedimiento de inferir a partir de una muestra la forma de $\phi(x)$.

Asociados a los quantiles podemos definir **medidas adicionales de dispersión**, los **rangos inter-quantílicos, *cuasi*-rangos** o **rangos de orden *p***,

RANGO DE ORDEN p: $$R(\xi_p) = \xi_{1-p} - \xi_p, \qquad 0 \le p < 0.5 \tag{12}$$

y **medidas adicionales de posición**, los **medios-rangos de orden *p***,

MEDIO-RANGO DE ORDEN p: $$Mid - R(\xi_p) = \frac{\xi_p + \xi_{1-p}}{2}, \qquad 0 \le p < 0.5 \tag{13}$$

Obsérvese que para $p = 0$ obtenemos, $R(\xi_{0.0}) = R(x)$ y $Mid–R(\xi_{0.0}) = Mid–R(x)$.

***R*($\xi_{.25}$)** es conocido como el **rango inter-cuartílico**, una medida de dispersión muy popular como alternativa a la desviación típica y en la definición de observaciones atípicas (*outliers*). Para una distribución simétrica todos medios-rangos de orden *p* deben coincidir y ser igual a la mediana que a su vez debe ser igual a la media, de esta forma estos estadísticos pueden proporcionarnos información muy útil acerca de la simetría de la distribución y en caso de ser asimétrica sobre la forma de dicha asimetría[33].

[33] Idénticas medidas adicionales de posición y dispersión podrían ser definidas a partir del cálculo de sucesivas medianas de las observaciones (Mills (1990), p.-21-26).

- **Medidas de simetría**: Puesto que para una **distribución simétrica** la **media y la mediana coinciden**[34] parece natural medir el alejamiento de una distribución de la simetría a partir del estadístico

	ponderado	**simple**	
SIMETRÍA:	$S_{\omega}(x) = \dfrac{\mu - Med_{\omega}(x)}{SD_{\omega}(x)}$	$S(x) = \dfrac{\bar{x} - Med(x)}{SD(x)}$	(14)

cuyos límites de variación vienen dados por $-1 \leq S_{\omega}(x), S(x) \leq 1$ (Hotelling and Solomons (1932)).

Además puesto que todos los **momentos centrales respecto a la media de orden impar son nulos** para **distribuciones simétricas** parece natural utilizarlos para examinar la simetría de una distribución. En la práctica se suele utilizar solamente el tercer momento, μ_3, que para distribuciones simétricas es nulo, $\mu_3 = 0$. El gráfico 1 permite observar dos funciones de densidad simétricas, la de una distribución normal estándar y la de una distribución *t*-Student ("Student" (1908a,b)) con 5 grados de libertad, en ambos casos puede demostrarse que $\mu_3 = 0$.

El gráfico 2 (Mood, Graybill and Boes (1974), p.-76) ofrece una impresión visual de lo que esperamos cuando pensamos en distribuciones asimétricas, así la densidad $\phi_1(x)$ se dice que es asimétrica hacia la izquierda, la cola de la izquierda decae más lentamente que la de la derecha, y en este caso $\mu_3 < 0$; mientras que la densidad $\phi_2(x)$ es asimétrica hacia la derecha, la cola de la derecha decae más lentamente que la de la izquierda, y puede demostrarse que ahora $\mu_3 > 0$. Sin embargo las medidas de asimetría deben interpretarse con cautela ya que el conocimiento de las mismas no proporciona realmente una información fiable acerca de la forma de la distribución. De hecho para una distribución simétrica $\mu_3 = 0$ pero lo contrario no es cierto, $\mu_3 = 0$ no implica que la distribución sea simétrica (Ord (1968)); por ejemplo, la densidad $\phi_3(x)$ en el gráfico 2 tiene $\mu_3 = 0$, pero

[34] Tambien la **moda** en el caso de distribuciones unimodales.

Gráfico 1. Distribución Normal y t-Student

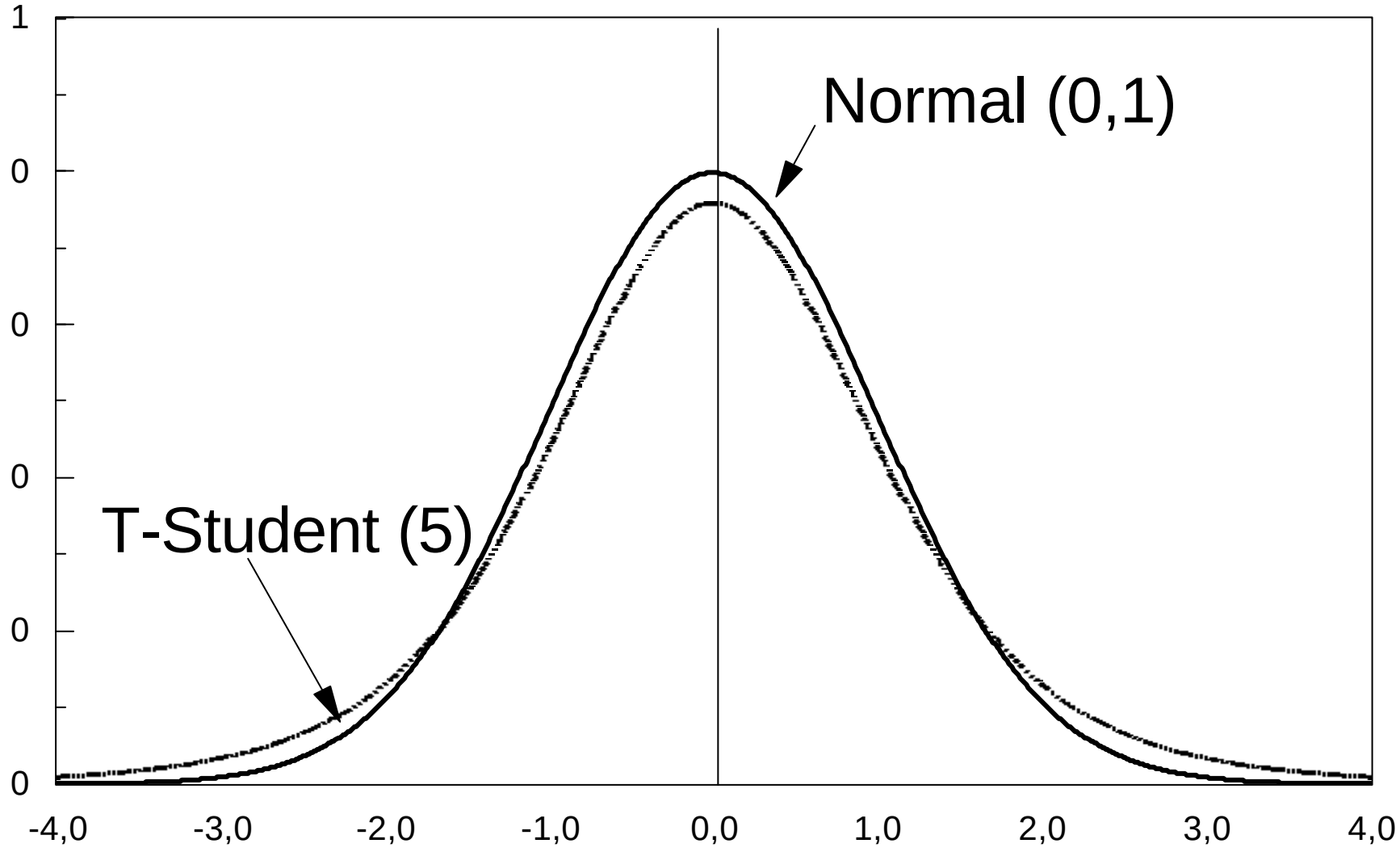


Gráfico 2. Simetría

Asimétrica hacia la izquierda

Asimétrica hacia la derecha

Asimétrica pero con el tercer momento igual a cero

obviamente su forma está lejos de ser simétrica, además pequeños cambios en la curvatura de $\phi_3(x)$ podrían proporcionar valores positivos o negativos de μ_3.

El la práctica se utiliza como **medida de simetría** el **tercer momento convenientemente estandarizado**, $\frac{\mu_3}{\mu_2^{3/2}}$, para librarlo de las unidades de medida, y que es lo que se conoce como el **coeficiente de simetría**.

	ponderado	**simple**[35]
COEFICIENTE DE SIMETRÍA:	$\gamma_1 = \sqrt{\beta_1} = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{SD_\omega(x)^3}$	$c_1 = \sqrt{b_1} = \frac{m_3}{m_2^{3/2}} = \frac{m_3}{SD(x)^3}$ (15)

El estadístico (15) puede ser utilizado para realizar un contraste sobre si nuestras observaciones proceden de una distribución simétrica, esto es si $\phi(x)$ es simétrica. Si denominamos θ_1 al coeficiente de simetría poblacional entonces H_0: $\theta_1 = 0$ puede ser contrastado a partir del resultado[36]

$$\sqrt{\frac{n}{6}}\gamma_1 \xrightarrow{d} N(0,1) \qquad \text{bajo} \quad H_0: \theta_1 = 0$$

- **Medidas de curtosis**: el **cuarto momento alrededor de la media**, μ_4, es utilizado con frecuencia como **medida del grado de curvatura de una distribución alrededor de su centro**.

[35] Al igual que sucede con la varianza simple el tercer momento simple, m_3, puede incorporar un ajuste por grados de libertad, $k_3 = \frac{n^2}{(n-1)(n-2)}.m_3$, con lo que el coeficiente de simetría podría ser calculado como $\frac{k_3}{k_2^{3/2}}$ (Kendall y Stuart (1977, p.-73, 88 y 300), Doan (1992, p.-14-238)).

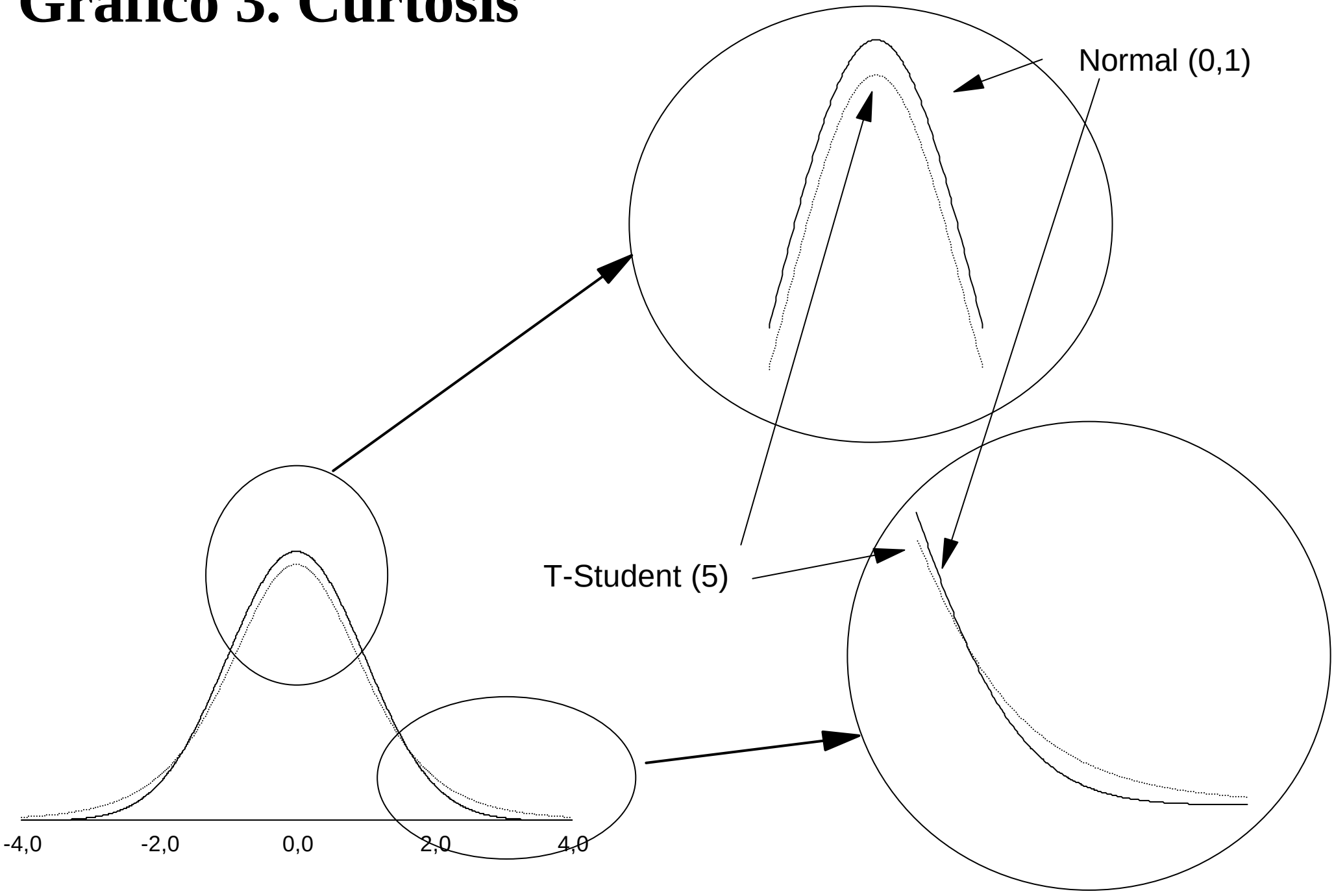
[36] En términos simples el contraste puede incluir un ajuste por grados de libertad (Kendall y Stuart (1977, p.-317), Patel y Read (1982, Cap.-5.7), Doan (1992, p.-14-238)).

En la práctica, puesto que μ_4 tiene unidades de medida, lo que se utiliza es el **coeficiente de curtosis**[37], que no es más que el **cuarto momento estandarizado**, $\frac{\mu_4}{\mu_2^2}$. Para una distribución normal estándar, gráfico 1, el valor de dicho coeficiente es 3, por lo que normalmente el coeficiente de curtosis se define respecto a la normal como el **coeficiente de exceso de curtosis**, $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$. Distribuciones para las que $\gamma_2 = 0$ se denominan **meso-cúrticas**, cuando $\gamma_2 > 0$ **lepto-cúrticas** y cuando $\gamma_2 < 0$ **plati-cúrticas** (Pearson (1906)). Aunque estos nombres se aplican en la práctica al valor del coeficiente (de exceso) de curtosis su origen se debe a que para ciertas distribuciones simétricas regulares (unimodales), la normal es la referencia más evidente, valores de $\gamma_2 > 0$ indican una densidad más puntiaguda alrededor de su centro que la distribución normal; mientras que valores de $\gamma_2 < 0$ indican una densidad más plana alrededor de su centro que la distribución normal. Esto no es sin embargo necesario para otras distribuciones simétricas o para distribuciones asimétricas, por lo que el coeficiente (de exceso) de curtosis sufre del mismo defecto que las medidas de simetría, es decir que no siempre mide lo que se supone que debe medir.

El gráfico 3 permite observar con más detalle la distribución *t*-Student con 5 grados de libertad en relación a la distribución normal estándar, observamos que la *t*-Student es ligeramente más puntiaguda que la normal, de hecho para esta distribución $\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = 6 > 0$. En general para la *t*-Student $\frac{\mu_4}{\mu_2^2} = 3 + \frac{6}{\nu - 4}$, siendo ν el número de grados de libertad. Observamos igualmente como esta distribución tiene más densidad en las colas que la normal.

[37] El coeficiente de curtosis fue introducido en estadística por Pearson (1895).

Gráfico 3. Curtosis



COEFICIENTE DE CURTOSIS: **ponderado** $\gamma_2 = \beta_2 - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{Var_\omega(x)^2} - 3$ **simple**[38] $c_2 = b_2 - 3 = \frac{m_4}{m_2^2} - 3 = \frac{m_4}{Var(x)^2} - 3$ (16)

Es posible demostrar que $\beta_2 \geq 1$ $(\gamma_2 \geq -2)$ siempre y que para distribuciones simétricas y unimodales $\beta_2 \geq 1.8$, además se cumple que $\beta_2 > 1 + \beta_1$ (Kendall y Stuart (1977), p.-88 y 95). Puesto que el cuarto momento eleva a la cuarta potencia la distancia de las observaciones respecto a la media **el coeficiente de curtosis es muy sensible a los *outliers***.

El estadístico (16) puede ser utilizado para realizar un contraste sobre si la distribución empírica de nuestros datos se asemeja a la forma de una distribución normal, lo que se denomina un contraste de curtosis. Si denominamos θ_2 al coeficiente de exceso de curtosis poblacional entonces H_0: $\theta_2 = 0$ puede ser contrastado a partir del resultado[39]

$$\sqrt{\frac{n}{24}}\gamma_2 \xrightarrow{d} N(0,1) \qquad \text{bajo } H_0: \ \theta_2 = 0$$

En la práctica es más útil un **contraste conjunto de simetría y curtosis**, lo que se interpreta como un **contraste de normalidad** (Jarque y Bera (1980)). Al ser los estadísticos (15) y (16) asintóticamente independientes la hipótesis nula H_0: $\theta_1 = \theta_2 = 0$ (*Normalidad*) puede ser contrastada a partir del resultado[40]

---

[38] En el caso del coeficiente de exceso de curtosis simple el numerador de $c_2 = \frac{m_4 - 3m_2^2}{m_2^2}$, $m_4 - 3m_2^2$, puede incorporar un ajuste por grados de libertad,

$$k_4 = \frac{n^2}{(n-1)(n-2)(n-3)}\{(n+1)m_4 - 3(n-1)m_2^2\}$$

con lo que el coeficiente de exceso de curtosis podría ser calculado como $\frac{k_4}{k_2^2}$ (Kendall y Stuart (1977, p.-73, 88 y 300), Doan (1992, p.-14-238)).

[39] En términos simples el contraste puede incluir un ajuste por grados de libertad (Kendall y Stuart (1977, p.-326), Patel y Read (1982, Cap.-5.7), Doan (1992, p.-14-238)).

[40] De nuevo en términos simples el contraste puede incluir un ajuste por grados de libertad.

$$n\left(\frac{\gamma_1^2}{6}+\frac{\gamma_2^2}{24}\right)\xrightarrow{d}\chi^2_{(2)} \qquad \text{bajo} \quad H_0: \theta_1 = \theta_2 = 0 \ (\textit{Normalidad})$$

Hasta aquí hemos descrito un conjunto de estadísticos que nos permitirán una primera caracterización de nuestro objeto de estudio, $\phi(x)$, sin embargo esta caracterización será necesariamente incompleta, los momentos y los quantiles sólo proporcionan visiones parciales de la forma de $\phi(x)$, es posible encontrar densidades con formas muy diferentes pero con sus cuatro primeros momentos idénticos (Joiner y Rosenblatt (1971)). En términos prácticos los dos primeros momentos son de gran importancia puesto que normalmente es necesario conocer la posición de nuestra variable y tener alguna idea acerca de su dispersión, los quantiles pueden proporcionarnos una idea del comportamiento de nuestra variable en las colas de la distribución pero los momentos de orden tercero y cuarto son de poca utilidad ya que normalmente es difícil concluir algo acerca de la forma de $\phi(x)$ a partir de ellos, momentos de orden más elevado son de relevancia práctica nula y por ello no han sido considerados. Una forma de resumir la información proporcionada por (casi) todos nuestros estadísticos será examinada en el epígrafe 2.6 y la cuestión de como inferir la forma de $\phi(x)$ a partir de nuestras observaciones será retomada en la sección siguiente.

2.4. "Outliers": Identificación

No existe un criterio universalmente aceptado para la definición de observaciones atípicas o *outliers*. Sin embargo la identificación de *outliers* es muy importante ya que pueden distorsionar gravemente los resultados de un estudio, las observaciones atípicas pueden deberse a errores en la construcción o publicación de estadísticas, en cuyo caso deberán subsanarse o de no ser posible eliminarse del análisis, pueden ser fenómenos puramente aleatorios o por el contrario pueden llevar consigo información genuina de interés acerca de determinados fenómenos que deben ser analizados con más cuidado o estudiados separadamente. Por ejemplo, Goerlich y Mas (1998c) muestran como sólo dos observaciones de una muestra de 24 son suficientes para generar los resultados de convergencia-σ observados en la muestra de países de la OCDE.

Una **regla** utilizada con frecuencia para la definición de *outliers* se basa en el **rango inter-cuartílico**, $R(\xi_{.25}) = \xi_{.75} - \xi_{.25}$, y considera **observaciones atípicas todas aquellas que caen fuera del intervalo definido por** $\xi_{.25} - 1.5 \text{x} R(\xi_{.25})$, **como límite inferior, y** $\xi_{.75} + 1.5 \text{x} R(\xi_{.25})$**, como límite superior** (Tukey (1977)); es decir x_i es considerado un *outlier* si

$$x_i < \xi_{.25} - 1.5 \text{x} R(\xi_{.25}) \qquad \text{o} \qquad x_i > \xi_{.75} + 1.5 \text{x} R(\xi_{.25})$$

Para una distribución normal estándar $\xi_{.25} = -\xi_{.75} = -0.674$ con lo que $R(\xi_{.25}) = 1.349$ y por tanto los límites del intervalo representan 2.698 desviaciones típicas a ambos lados de la media o mediana, lo que cubre una probabilidad del 99.30% y en consecuencia representa una probabilidad de observar *outliers* del 0.70%.

Esto sugiere una **regla alternativa** para la definición de observaciones atípicas **basada en la probabilidad de observación de las mismas a partir de la referencia a una normal**, por ejemplo para una normal estándar 2.5 desviaciones típicas a ambos lados de la media cubren una probabilidad del 98.76%, dejando una probabilidad de observación de *outliers* del 1.24%; 3.0 desviaciones típicas cubren una probabilidad del 99.73%, lo que deja una probabilidad de observación de *outliers* del 0.27%. Todo se reduce por tanto a fijar *a priori* nuestra probabilidad subjetiva asociada a la observación de un suceso muy raro, de la misma forma que fijamos el nivel de significación en un contraste de hipótesis, y obtener a partir de aquí los límites de un intervalo en términos de desviaciones típicas de una normal estándar. Así por ejemplo, si dicha probabilidad es fijada de forma arbitraria en un 0.1% entonces consideraríamos *outliers* a todas aquellas observaciones que cayeran fuera del intervalo definido por 3.29 desviaciones típicas a ambos lados de la media muestral, y si dicha probabilidad fueran fijada en el 1 por millón, 0.0001%, entonces consideraríamos observaciones atípicas todas aquellas cayeran fuera del intervalo definido por 4.89 desviaciones típicas a ambos lados de la media muestral. En términos prácticos consideraremos que estamos en presencia de un *outlier* cuando una observación caiga fuera del intervalo definido por 3.0 veces la desviación típica observada en los datos a ambos lados de la media muestral; es decir x_i es considerado un *outlier* si

$$x_i < \mu - 3.0 \text{x} SD_{\omega}(x) \qquad \text{o} \qquad x_i > \mu + 3.0 \text{x} SD_{\omega}(x)$$

en términos de estadísticos ponderados o

$$x_i < \bar{x} - 3.0 \text{x} SD(x) \qquad \text{o} \qquad x_i > \bar{x} + 3.0 \text{x} SD(x)$$

en términos de estadísticos simples, lo que como ya hemos dicho para una distribución normal representa una probabilidad de observación de *outliers* del 0.27%, es por tanto una regla algo más restrictiva que la basada en el rango inter-cuartílico.

2.5. Un comentario sobre la transformación logarítmica

Entre los estadísticos analizados en el epígrafe 3 destaca el hecho de que no hemos incluido la **varianza de los logaritmos** como medida de dispersión, o más concretamente la **desviación típica de los logaritmos**, a pesar de que este es el estadístico más frecuentemente utilizado por la literatura del crecimiento económico para medir el concepto de σ-convergencia[41] y que como ya señalamos en Goerlich (1998) constituye una medida habitual de desigualdad al ser independiente de la escala. Su definición es sencilla y simplemente consiste en aplicar el concepto de **varianza** o **desviación típica** al logaritmo de nuestra variable de referencia:

VARIANZA	**ponderada**	**simple**
DE LOS	$Var_{\omega}(\log x) = \Sigma_{i=1}^{n} p_i (\log x_i - \log \tilde{\mu})^2$	$Var(\log x) = \dfrac{\Sigma_{i=1}^{n} (\log x_i - \log \tilde{x})^2}{n}$ (17)
LOGARITMOS:		

[41] De hecho Barro y Sala-i-Martín (1995, Cap.-11.1,p.-383-387) identifican el concepto de σ-convergencia con el de la desviación típica del logaritmo de la renta *per capita*.

donde $\log\tilde{\mu} = \Sigma_{i=1}^{n} p_i \log x_i$ y $\log\tilde{x} = \frac{\Sigma_{i=1}^{n} \log x_i}{n}$ son el logaritmo de la media geométrica, ponderada o simple respectivamente. En consecuencia la desviación típica de los logaritmos se define como

	ponderada	**simple**	
DESVIACIÓN TÍPICA DE LOS LOGARITMOS:	$SD_{\omega}(\log x) = +\sqrt{Var_{\omega}(\log x)}$	$SD(\log x) = +\sqrt{Var(\log x)}$	(18)

La razón de tal omisión es deliberada y responde al hecho de que lo que pretendemos analizar es la distribución de x, $\phi(x)$, no la distribución del logaritmo de x, $\phi(\log x)$. Aunque evidentemente las dos distribuciones están relacionadas no constituyen el mismo objeto de estudio y no nos parece razonable tratar de caracterizar $\phi(x)$ por medio de la transformación logarítmica de x. Ciertamente la transformación logarítmica tiene propiedades útiles y muy deseables en ciertos contextos, por ejemplo,

(i) la transformación logarítmica es monótonamente creciente y por tanto mantiene el *ranking* entre observaciones,

(ii) los modelos teóricos son más fácilmente resolubles mediante aproximaciones logarítmico lineales en torno al estado estacionario (Barro y Sala-i-Martín (1995)) y en consecuencia $SD(\log x)$ puede tener un sentido concreto en un modelo particular,

(iii) si $\log x$ tuviera una distribución normal entonces la distribución de x sería lognormal (Aitchison y Brown (1957), Nelson (1973, Cap.-6.7), Hart (1995)) y esta es una distribución frecuentemente utilizada en el análisis de la distribución personal de la renta y la riqueza por algunas de sus especiales características (Cowell (1995))[42],

[42] Además es este caso $\log(1 + CV(x)^2) = Var(\log x)$, por lo que existe una relación uno a uno entre $CV(x)$ y $SD(\log x)$ como medidas de dispersión y en consecuencia ambos estadísticos proporcionan la misma información. Aitchison y Brown (1957) Tabla A.1, p.-154 tabulan para la distribución log-normal la relación entre $CV(x)$ y $SD(\log x)$.

(iv) si la dispersión en una variable es proporcional al nivel de la misma la transformación logarítmica estabiliza la varianza y reduce problemas de heterocedasticidad (Spanos (1986, p.-487)), esta es una de las razones por la que la transformación logarítmica es tan popular en econometría aplicada, y

(v) los logaritmos tienen una clara justificación en la literatura sobre índices de desigualdad, donde normalmente se desea dar más importancia a las transferencias de renta en el extremo inferior de la distribución, discriminando de esta forma positivamente hacia los pobres (Villar (sin fecha, p.-13)), sin embargo este no tiene por que ser el caso en el tema de la convergencia entre regiones económicas.

Es cierto que la transformación logarítmica libra a los estadísticos de las unidades de medida y los hace independientes de la escala, sin embargo no encontramos ninguna clara ventaja en esta transformación como forma de caracterizar $\phi(x)$, la reducción de los problemas de heterocedasticidad puede ser más un inconveniente que una ventaja al enmascarar características importantes en la evolución de $\phi(x)$ en el tiempo, especialmente en el extremo superior de la distribución; por otra parte no estamos interesados en discriminar a favor o en contra de la reducción en la dispersión en determinadas partes de la distribución y perdemos claramente intuición, así por ejemplo podemos examinar el rango de nuestra variable pero no está muy claro el significado que debemos otorgar a los logaritmos de las observaciones extremas de nuestra muestra. Sin embargo la razón más importante que encontramos para no utilizar la desviación típica de los logaritmos como medida de dispersión, al menos en un sentido único, es que como es bien conocido no verifica el principio de las transferencias de Pigou (1912)-Dalton (1920) (Cowell (1995, p.-149), por lo que tal y como han puntualizado acertadamente Foster y Ok (1999) es posible encontrar casos de relevancia práctica en los que una reducción en la dispersión global en la distribución, en el sentido de dominancia de Lorenz (1905), vayan acompañados de un incremento en $SD(\log x)$.

Debemos observar además algunas peculiaridades de interés asociadas a esta medida. En primer lugar, tal y como ha sido utilizada por la literatura del crecimiento, se utiliza siempre la versión no ponderada del estadístico, por lo que implícitamente esta literatura está interesada en la distribución de la renta *per capita* de los países o las

regiones y no de la población subyacente a los mismos. En segundo lugar el estadístico ponderado utiliza como ponderación para $\log x_i$ la misma que para x_i, lo que hace perder de nuevo intuición y sugiere una agregación por medias geométricas en lugar de por medias aritméticas (Attanasio y Weber (1993)), sin embargo la media del agregado, que es observable, es la media aritmética, μ, no la geométrica, $\tilde{\mu}$. Esta es la razón por la que en ocasiones la desviación de los logaritmos de x se realiza respecto del logaritmo de la media aritmética, $\log \mu$, en lugar de respecto del logaritmo de la media geométrica, $\log \tilde{\mu}$; generando de esta forma una medida alternativa de dispersión, la denominada **varianza logarítmica** (Cowell (1995), Goerlich (1998))[43]; sin embargo esta medida tampoco verifica el principio de las transferencias de Pigou (1912)-Dalton (1920) (Cowell (1995, p.-149)) y no será considerada.

Sin embargo la transformación logarítmica puede alterar algunas de las características importantes que podemos inferir acerca de $\phi(x)$ de los estadísticos calculados, por ejemplo la transformación logarítmica, al comprimir la escala de la variable tiene éxito en reducir o eliminar el número de observaciones atípicas, esto puede ser una clara ventaja en ciertos contextos, por ejemplo en el análisis de regresión, Goerlich (2000), pero es en realidad un inconveniente en términos de la caracterización de $\phi(x)$, ya que podemos perder algunas de las características importantes de la distribución de nuestra variable, y el enmascaramiento de los *outliers* puede ser una de ellas[44].

En este sentido **la desviación típica de los logaritmos no proporcionan ninguna ventaja adicional sobre el coeficiente de variación como medida de dispersión invariante respecto a la escala**, y aunque útil en ciertos contextos no parece presentar ventajas si lo que pretendemos es caracterizar la distribución de una variable.

---

[43] Obsérvese que centrar los logaritmos de x_i en $\log \mu$ en lugar de en $\log \tilde{\mu}$ no es equivalente a considerar $Var_\omega(\log z_i)$, ya que por definición $Var_\omega(\log z_i) = Var_\omega(\log x_i)$, independientemente de las ponderaciones.
La **varianza logarítmica** quedaría definida en términos de los momentos centrales como $\mu_2(\log x, \log \mu) = \Sigma_{i=1}^n p_i (\log x_i - \log \mu)^2$ en el caso ponderado, y como $m_2(\log x, \log \bar{x}) = \frac{\Sigma_{i=1}^n (\log x_i - \log \bar{x})^2}{n}$ en el caso simple.

[44] Esto indica que ciertos ejercicios (Gardeazabal (1996)) pueden no proporcionar los mismos resultados si no tomáramos logaritmos.

2.6. A modo de resumen: "Box-plots"

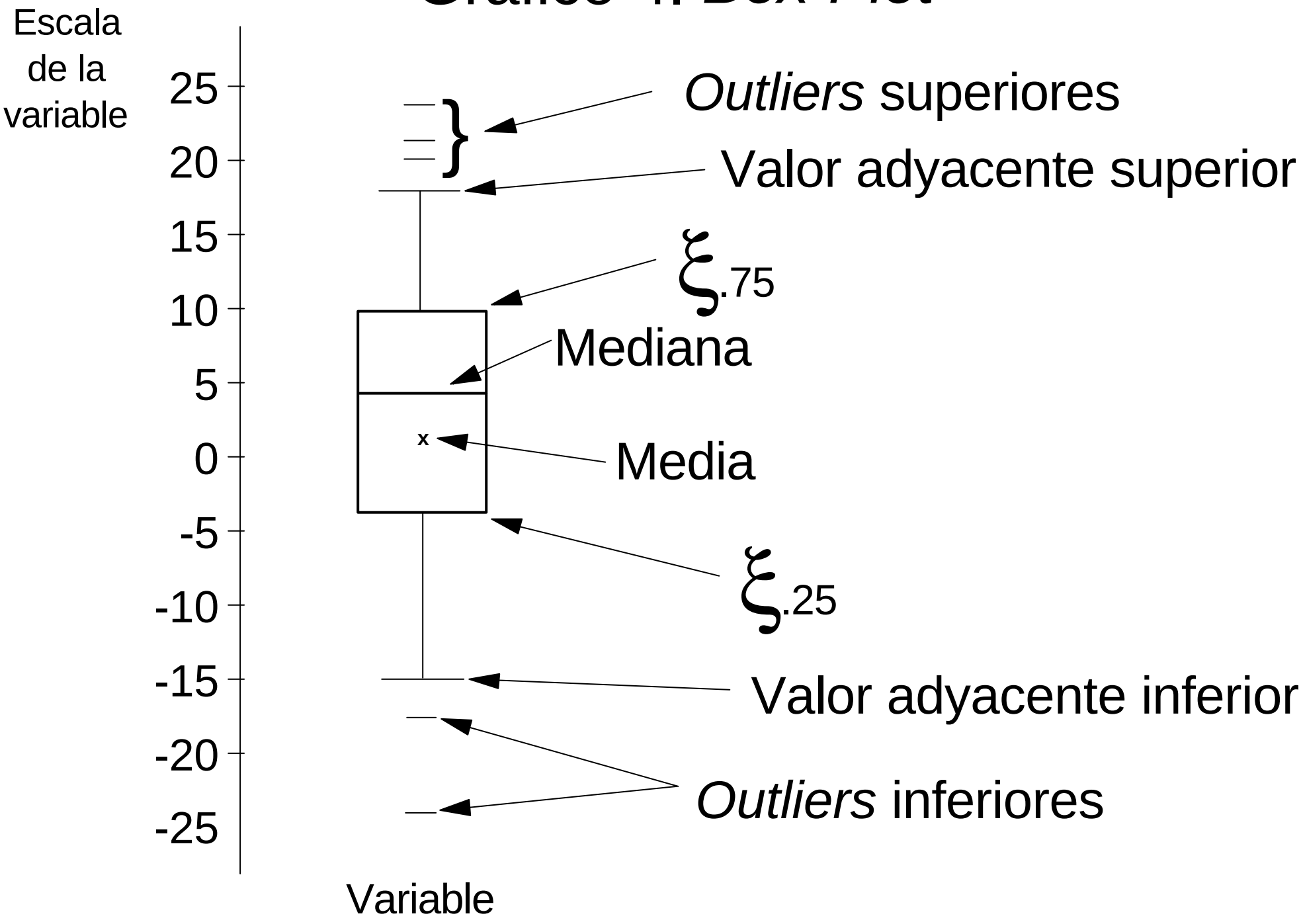
Ofrecemos en este epígrafe una forma gráfica y conveniente de resumir gran parte de la información suministrada por los estadísticos descriptivos que hemos descrito en esta sección, los denominados diagramas de caja o *box-plots* que proporcionan una forma rápida de examinar los datos.

Un *box-plot* no es más que una representación plana de algunas de las características más sobresalientes de un conjunto de datos. Proporciona información que está a medio camino entre los estadísticos descriptivos y una representación de un histograma, su principal ventaja es que dado que es una representación plana pueden observarse simultáneamente varios *box-plots* en un mismo gráfico lo que permite el estudio dinámico de la evolución de algunas características importantes de la distribución de la variable en cuestión, por ejemplo existencia, aparición o desaparición de *outliers*, dispersión o concentración de los datos, así como la simetría o asimetría de la distribución. De hecho una de las utilidades básicas de los *box-plots* es el análisis gráfico de *outliers*.

A continuación describimos un *box-plot* estándar, que adopta la definición de *outliers* basada en el rango inter-cuartílico y examinada en el epígrafe 2.4, existen otros tipos de *box-plots* más completos o que adoptan otra definición de las observaciones atípicas pero no serán mencionados en este trabajo (Tukey (1977), McGill, Tukey y Larsen (1978), Velleman y Hoaglin (1981), Mills (1990, Cap.-3, Sec.-3.4), Cleveland (1993), Everitt (1994)).

Un *box-plot*, con todos sus elementos, puede examinarse en el gráfico 4. El eje horizontal carece de sentido y simplemente representa cada variable en cuestión, mientras que el eje vertical representa la escala de la variable. El **cuadrado** o caja, ***box***, representa el **rango inter-cuartílico**, el cuartíl 0.75, $\xi_{.75}$, constituye la parte superior y el cuartíl 0.25, $\xi_{.25}$, constituye la parte inferior del cuadrado. Por construcción dentro del *box* está contenido el 50% de la masa de probabilidad de la distribución. La **altura del *box*** representa, por tanto, el **rango inter-cuartílico**, que como ya hemos indicado constituye una medida de dispersión habitual. Un rango inter-cuartílico mayor se visualizará mediante un *box* de mayor altura indicando que el 50% de la densidad de *x* está relativamente

Gráfico 4. *Box-Plot*



dispersa. Por el contrario, un rango inter-cuartílico menor se visualizará mediante un *box* más corto, e indica que el 50% de la densidad de *x* está relativamente concentrada.

La **línea horizontal dentro del *box***, es la **mediana** o **cuartíl 0.50**. Una medida de **posición** de la distribución de la variable. La localización de esta línea respecto a los límites superiores o inferiores del *box* proporciona información gráfica sobre la forma de la distribución, si la mediana no está en el centro del *box* la distribución es asimétrica. En el caso del gráfico 4 existe evidencia de asimetría hacia la izquierda, es decir hacia la parte inferior de la distribución. En ocasiones la línea que representa la mediana se complementa con una indicación de la **media**, una **x** en el gráfico 4; la relación entre la mediana y la media proporciona evidencia adicional sobre la simetría de la distribución, así en nuestro ejemplo del gráfico 4 la distancia entre la media y la mediana refuerza la evidencia sobre la asimetría mencionada anteriormente.

Dos **líneas verticales** aparecen en los límites superior e inferior del *box*, el final de estas líneas, dibujadas de forma horizontal, se conoce como **valor adyacente, superior** e **inferior** respectivamente. A partir del rango inter-cuartílico, $R(\xi_{.25})$, el **valor adyacente superior** se define como el **valor observado de la variable representada no mayor que** $\xi_{.75} + 1.5 \text{x} R(\xi_{.25})$**,** y el **valor adyacente inferior** como el **valor observado de la variable representada no menor que** $\xi_{.25} - 1.5 \text{x} R(\xi_{.25})$. La máxima longitud posible entre valores adyacentes vendrá dada por el intervalo $\left[\xi_{.25} - 1.5 \text{x} R(\xi_{.25}) , \xi_{.75} + 1.5 \text{x} R(\xi_{.25})\right]$ pero en general presentará un recorrido menor ya que dentro de este intervalo buscaremos las observaciones extremas para determinar dichos valores. Los **valores adyacentes** son, por tanto, **estadísticos de orden**, $x_{(s)}$, que se corresponden con observaciones actuales de la variable en cuestión y que cubren el rango de observaciones que no consideraremos como *outliers*.

Finalmente, **las observaciones más allá de los valores adyacentes son los *outliers***, **superiores si son mayores que el valor adyacente superior**, e **inferiores si son menores que el valor adyacente inferior.** Estos valores son representados de forma individual por pequeñas líneas horizontales, así en el ejemplo del gráfico 4 podemos observar 3 *outliers* superiores y 2 inferiores. Los valores adyacentes cumplen de esta forma

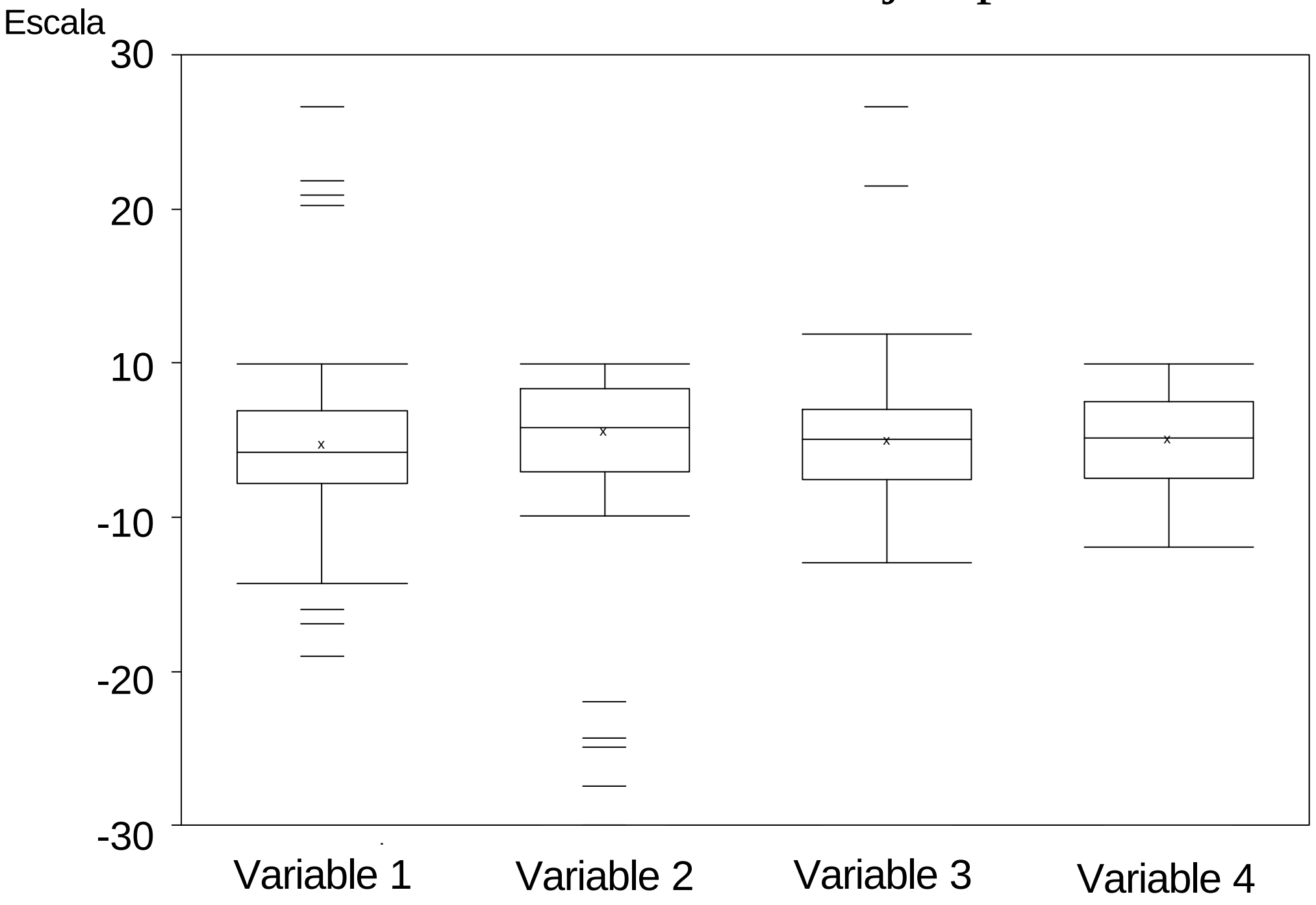
una doble misión, por una parte nos delimitan el rango de observaciones que no consideraremos como atípicas y por otra nos indican la distancia entre los valores extremos de dichas observaciones y los *outliers*, lo que permite observar la lejanía o proximidad de los mismos respecto a la mayor parte de la distribución. Es posible que no existan *outliers*, de forma que los valores adyacentes sean en realidad los valores extremos del conjunto de observaciones, el máximo y/o el mínimo de la distribución. Las diferentes posibilidades pueden examinarse en el gráfico 5.

Obviamente los *box-plots* pueden calcularse a partir de estadísticos simples o ponderados si bien en este último caso la existencia de observaciones atípicas no nos dice nada acerca de la masa de probabilidad asociada a dichas observaciones. En este caso se impone un tratamiento individualizado de los *outliers*.

En resumen, dado el $R(\xi_{.25})$, obtenido a partir de $\xi_{.25}$ y $\xi_{.75}$, calculamos el intervalo $\left[\xi_{.25} - 1.5 \text{x} R(\xi_{.25}) , \xi_{.75} + 1.5 \text{x} R(\xi_{.25})\right]$ y determinamos los valores observados máximo y mínimo dentro de dicho intervalo, estos valores constituyen los **valores adyacentes, superior e inferior** respectivamente. Todas las observaciones que caen fuera de dicho intervalo son consideradas *outliers*. Los *outliers* se definen pues como aquellos valores que caen fuera de 1.5 veces $R(\xi_{.25})$ por encima y por debajo del mismo. Por construcción **si no existen *outliers*** los valores adyacentes son los estadísticos de valor extremo de la distribución, y en este caso la **distancia entre valores adyacentes representa el rango de las observaciones**, $R(x)$, otra medida de dispersión. En consecuencia los *box-plots* resumen gran parte de la información ofrecida anteriormente y son útiles fundamentalmente por dos motivos; (*i*) para la determinación y evolución de los *outliers*, y (*ii*) en relación al estudio de la dispersión o concentración de la distribución, más exactamente del 50% de la densidad de probabilidad asociada al $R(\xi_{.25})$.

En la práctica se representan varios *box-plots* correspondientes a diferentes variables en un mismo gráfico de forma que podemos observar rápidamente las características principales de los datos, así como las diferencias entre variables. El gráfico 5 contiene *box-plots* para cuatro variables que cubren todos los casos posibles de relevancia práctica. Para la variable 1 observamos *outliers*, tanto superiores como inferiores, sin

Gráfico 5. *Box-Plots* - Ejemplos



embargo mientras los *outliers* inferiores se encuentran relativamente cerca de su valor adyacente los *outliers* superiores están mucho más distanciados del valor adyacente superior lo que indica una mayor singularidad en estas observaciones. Para la variable 2 sólo se observan *outliers* inferiores de forma que el valor adyacente superior se corresponde con el máximo de los valores observados, $x_{(n)}$; por el contrario para la variable 3 sólo se observan *outliers* superiores de forma que el valor adyacente inferior se corresponde con el mínimo de los valores observados, $x_{(1)}$. Finalmente la variable 4 no presenta observaciones atípicas por lo que los valores adyacentes son en realidad el valor máximo y mínimo de la variable, $x_{(n)}$ y $x_{(1)}$, de esta forma observamos el rango de la variable. Dejando al margen los *outliers* sólo la variable 1 parece presentar una cierta asimetría hacia la derecha.

3. Estimando la función de densidad de probabilidad de una variable (convergencia-δ)

La sección anterior ha descrito de forma exhaustiva una serie de estadísticos que nos permitieran estudiar diversas peculiaridades de $\phi(x)$ prestando especial atención a la dispersión en la distribución, es decir al concepto de σ-convergencia. Esta sección ampliará el análisis anterior considerando $\phi(x)$ en su totalidad en la línea sugerida repetidamente por Quah (1993a,b) y que ya vislumbramos al finalizar la sección anterior, donde tratamos de resumir la información proporcionada por los estadísticos calculados; los *box-plots* eran útiles pero una caracterización completa de $\phi(x)$ todavía quedaba lejos. Una vez $\phi(x)$ haya sido caracterizada en su totalidad será posible examinar la convergencia de toda la distribución, no sólo de algunas características parciales de la misma, lo que podríamos denominar **δ-convergencia**.

Buscamos ahora una estimación directa de la forma de $\phi(x)$, ya que ello nos proporcionará información relevante sobre las características de esta función, tales como la dispersión, la asimetría, la forma de la distribución en relación a la normal, o la posibilidad de la existencia de múltiples máximos locales (modas) y la formación de clubs diferenciados de regiones. La gran cantidad de estadísticos descriptivos calculados en la sección anterior ofrecían respuestas parciales a estas cuestiones y en ocasiones era difícil extraer conclusiones claras a partir de tanto estadístico, sin duda alguna la mejor forma de obtener una visión clara acerca de $\phi(x)$ es conocer su forma. La primera cuestión de interés es, probablemente, saber si es posible inferir la forma de $\phi(x)$ a partir de los estadísticos estudiados en la sección anterior. El problema teórico de si una función de densidad[45] es determinada o no por la secuencia de sus momentos es conocido en estadística teórica como el **problema de los momentos**, y la respuesta es en general negativa, aunque bajo ciertas condiciones poco restrictivas una secuencia de momentos si determina de forma única $\phi(x)$, la razón por la cual los momentos son muy útiles desde un punto de vista teórico y en consecuencia son ampliamente utilizados (Mood, Graybill y Boes (1974, p.-

[45] De distribución para ser rigurosamente exactos.

81), Kendall y Stuart (1977, p.-89)). Desde un punto de vista práctico un conjunto de estadísticos descriptivos, por muy numeroso que sea, no permite inferir la forma de $\phi(x)$, si bien como hemos observado podemos determinar algunas de las características relevantes de la densidad de x, por ello deberemos acometer directamente el problema de estimar $\phi(x)$.

La literatura estadística ha tomado **dos aproximaciones** diferentes a la hora de enfrentarse con el problema de **estimar $\phi(x)$**, en **primer lugar** la **aproximación paramétrica** postula una forma funcional para $\phi(x)$, tal como la distribución normal, la *t-Student* o la distribución log-normal (Aitchison y Brown (1957)), esta función depende de una serie de parámetros que caracterizan completamente la densidad de x por lo que el problema se reduce a estimar estos parámetros, una vez estos parámetros han sido estimados $\phi(x)$ está caracterizada completamente, su forma puede observarse simplemente dibujando la función $y = \phi(x)$ y todos los momentos, quantiles y estadísticos descriptivos como las medidas de desigualdad, incluyendo la curva de Lorenz (1905), analizadas en Goerlich (1998) pueden ser directamente calculadas a partir del conocimiento de $\phi(x)$. El ejemplo más sencillo lo proporciona la distribución normal, si la densidad subyacente a las observaciones, $\phi(\bullet)$, fuera la normal, entonces nuestro problema se reduciría a encontrar estimaciones de la media y la varianza, ya que estos dos parámetros caracterizan completamente la distribución normal, estimaciones de estos parámetros pueden ser obtenidos a partir de los datos como μ y $\mu_2 = \sigma^2$ y sustituyendo estas estimaciones en la fórmula correspondiente nuestro problema ha sido resuelto,

$$y = \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2} . \frac{(x-\mu)^2}{\sigma^2}\right\}$$

De esta forma hemos invertido el proceso, no buscamos características indirectas de $\phi(x)$, ahora conocemos la distribución de la variable de interés y a partir de ella podemos calcular todos los estadísticos y medidas de desigualdad que deseemos, si queremos comparar dos distribuciones en el tiempo no tenemos más que estimar los parámetros relevantes en esos dos momentos y comparar las funciones resultantes, o calcular a partir de ellas el estadístico en el que estemos interesados y observar su evolución.

La teoría de la estimación estadística está lo suficientemente desarrollada como para que esta sea una forma operativa de estimar $\phi(x)$ una vez la forma funcional se supone conocida, sin embargo esta es la **cuestión crucial** en el procedimiento, hemos de **conocer la forma funcional de $\phi(x)$** y no existe un procedimiento general para saber cual es la forma funcional más apropiada en cada caso concreto. De las características particulares de las distribuciones (Johnson y Kotz (1970), Hastings y Peacock (1974), Evans, Hastings y Peacock (1993)) es posible inferir algunas **distribuciones útiles** en casos particulares, por ejemplo en el análisis de la distribución personal de la renta las dos distribuciones más utilizadas (Cowell (1995, Cap.-4)) por sus peculiares características son **la distribución log-normal** (Aitchison y Brown (1954, 1957)) y la **distribución de Pareto** (1965, 1972) para el extremo superior de la distribución (Spanos (1986, p.-61)); sin embargo muchas otras son utilizadas, tales como la **distribución Beta** (Thurow (1979), Slottje (1984)), la **distribución Gamma** (Salem y Mount (1974), McDonald y Jensen (1979)) o la **distribución sech^2** (Fisk (1961)), con el argumento de que representan mejor conjuntos particulares de datos. En ocasiones, y dada la compleja estructura de las muestras reales, se trata de ajustar distribuciones mixtas (Titterington, Smith y Makov (1985), Hamilton (1994, Cap.-22.3)), donde diversas distribuciones se combinan para representar diferentes partes de la distribución.

Este inconveniente, la necesidad de conocer la forma funcional, convierte a la aproximación paramétrica para la estimación de $\phi(x)$ en poco operativa en la práctica como instrumento descriptivo, y ello a pesar de la existencia de *software* especializado que ofrece, para una muestra de datos concreta, un *ranking* de distribuciones probables entre un gran conjunto a partir de la comparación de las diversas estimaciones mediante estadísticos de bondad del ajuste (*BESTFIT*[46], Palisade (1997)). Cuando este procedimiento semi-automático es aplicado a los datos de la renta *per capita* provincial normalizada en 1955 y 1995 los resultados que obtenemos se muestran en el gráfico 6. En ambos casos se ofrece la "mejor" distribución obtenida a partir de una ordenación de todas las intentadas y donde la ordenación se ha efectuado de acuerdo con el criterio de bondad del ajuste de Anderson-Darling (1954), un criterio similar al de Kolmogorov-Smirnov (Mood, Graybill y Boes (1974, p.-508)) pero que pone más énfasis en las colas de la distribución y no depende del

[46] Este programa considera un total de 26 funciones de distribución posibles.

número de intervalos en los que se clasifican las observaciones[47] (Stephens (1974, 1977), Chandra, Singpurwalla y Stephens (1981)).

Gráfico 6 (*i*) - Renta per capita provincial normalizada: 1955 y "mejor" distribución ajustada: Pearson V (α,β)

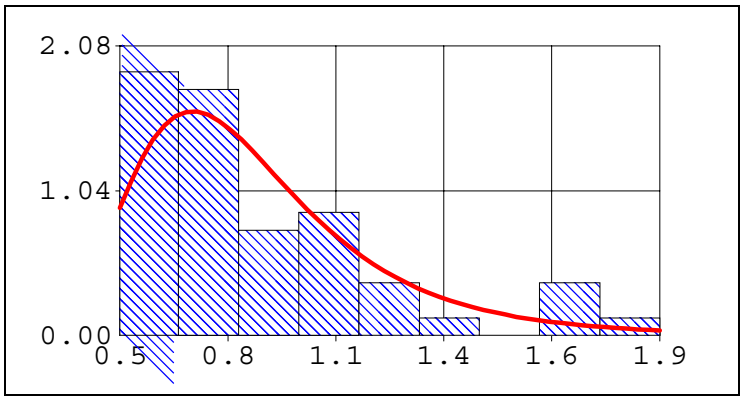
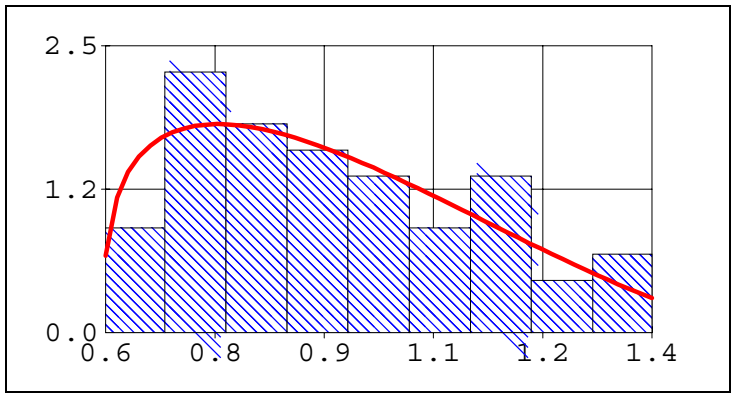


Gráfico 6 (*ii*) - Renta per capita provincial normalizada: 1995 y "mejor" distribución ajustada: Beta (α_1,α_2)



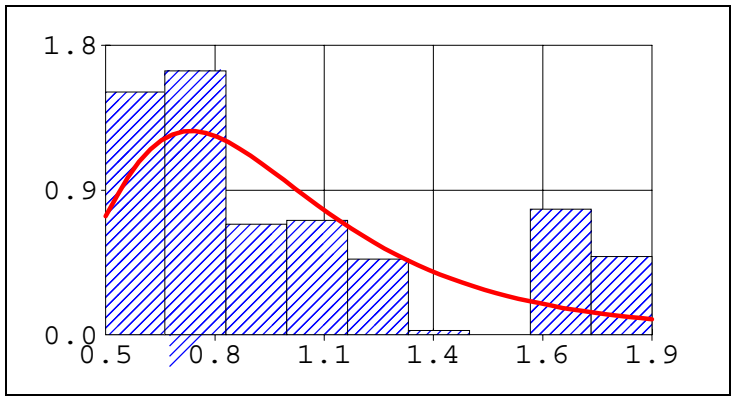
Observamos que para 1955 la "mejor" distribución es de tipo Pearson V[48], para este año el histograma de partida permite observar algunos *outliers* en la cola derecha de la distribución. En 1995 la distribución ha cambiado sustancialmente, en primer lugar el rango de los datos se ha reducido, nótese que la escala en ambos gráficos es ligeramente

[47] El procedimiento construye un histograma para las observaciones y ajusta la distribución a partir de él, Palisade (1997) ofrece información sobre los procedimientos de cálculo.
[48] Los valores estimados de los parámetros fueron $\alpha = 9.04$ y $\beta = 7.03$.

diferente, ahora la cola derecha de la distribución no aparece tan aislada del resto como en 1955 y finalmente la distribución también ha cambiado, la que mejor ajusta nuestros datos ahora en una distribución Beta[49].

Lamentablemente el *software* empleado no permite la utilización de datos ponderados[50] aunque estos pueden ser aproximados simplemente replicando las observaciones de acuerdo con su frecuencia relativa. En la práctica los datos de la renta *per capita* normalizada del gráfico 6 fueron ampliados hasta un total de 2.000 observaciones manteniendo la estructura poblacional del año correspondiente, de esta forma la muestra fue replicada 40 veces de acuerdo con la ponderación observada. Los resultados de la estimación de las funciones de densidad ponderadas se ofrecen en el gráfico 7 a partir de un histograma construido con los mismos intervalos que en el caso anterior.

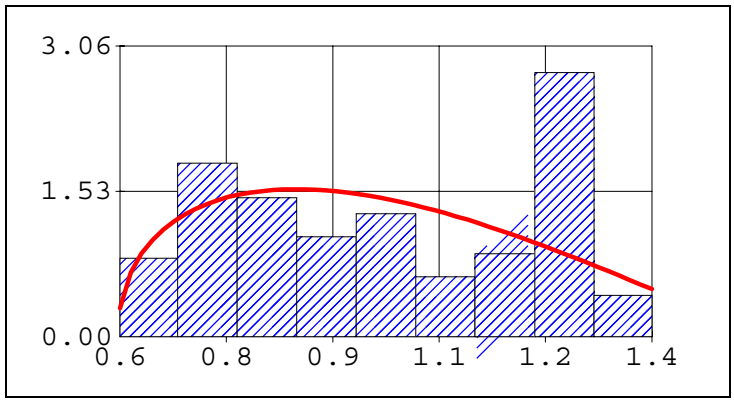
Gráfico 7 (*i*) - Renta per capita provincial normalizada: 1955
y "mejor" distribución ajustada: Pearson V (α,β) - Datos ponderados



[49] Los valores estimados de los parámetros fueron $\alpha_1 = 1.32$ y $\alpha_2 = 2.60$; puesto que el dominio de definición para la distribución Beta es el intervalo cerrado [0,1] los datos son transladados a este intervalo antes de proceder al ajuste.

[50] Si permite, sin embargo, la utilización de datos en términos de densidad, es decir un valor y su probabilidad asociada. Introduciendo los datos de renta *per capita* y población de cada provincia de esta forma el programa produjo resultados totalmente insatisfactorios y carentes de sentido.

Gráfico 7 (*ii*) - Renta per capita provincial normalizada: 1995 y segunda "mejor" distribución ajustada: Beta (α_1,α_2). Datos ponderados

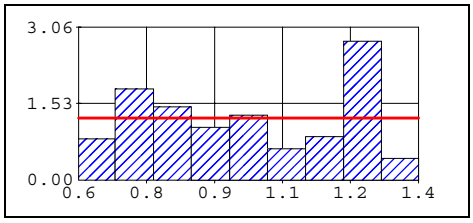


Para 1955 la "mejor" distribución sigue siendo de tipo Pearson V pero la distribución no es la misma puesto que los parámetros ahora son diferentes[51]. En 1995 la distribución ponderada ha cambiado sustancialmente, al igual que sucedía con la distribución simple, de hecho la "mejor" distribución estimada por *BESTFIT* se corresponde con una distribución uniforme[52], lo que es poco informativo, por ello el gráfico 7 (*ii*) ofrece la segunda distribución en el *ranking* que se corresponde, al igual que en el caso del gráfico 6 (*ii*), con una distribución Beta[53].

Estos ejemplos ponen de manifiesto que **la aproximación paramétrica al problema de estimar $\phi(x)$ carece de generalidad** y es **poco flexible**.

[51] Los valores estimados de los parámetros fueron $\alpha = 6.39$ y $\beta = 5.40$.
[52] La "mejor" distribución ajustada fue en realidad la siguiente distribución uniforme

Gráfico 8 - Renta per capita provincial normalizada: 1995 y "mejor" distribución ajustada: Uniforme (α_1,α_2) - Datos ponderados



[53] Los valores estimados de los parámetros fueron ahora $\alpha_1 = 1.45$ y $\alpha_2 = 2.24$. Los datos fueron transladados al intervalo cerrado [0,1] antes de proceder al ajuste.

En **segundo lugar**, la literatura estadística ha adoptado una **aproximación no-paramétrica** al problema de la estimación de $\phi(x)$ que es mucho más útil en nuestro contexto. Esta aproximación **no presupone ninguna forma funcional para la densidad de *x***, simplemente supone que cada observación, x_i, proporciona cierta información acerca de la densidad subyacente a las observaciones dentro de un intervalo ("ventana") alrededor de x_i. A continuación describiremos brevemente la intuición de esta forma de proceder y el método concreto seguido en nuestro caso, al igual que en el resto del trabajo enfatizaremos la intuición de los aspectos teóricos del método y describiremos con cierto detalle los aspectos prácticos más relevantes. Monografías útiles en este contexto, que constituye una verdadera rama de la estadística y el análisis de datos, son Silverman (1986), Scott (1992), Wand y Jones (1994) y Simonoff (1996).

Señalemos en primer lugar que lo que **tratamos de estimar** en esta sección es **una función**, $\phi(x)$, sin ofrecer una forma funcional para la misma; por el contrario en la sección anterior tratábamos de estimar características particulares de $\phi(x)$ que eran reflejadas en los correspondientes estadísticos y que en la práctica eran simples números reales, ahora obtendremos no un número real sino un conjunto de puntos (x,y) que corresponden a la función $y = \phi(x)$ y cuya representación gráfica nos proporcionará una impresión visual de la densidad que estamos buscando. Es este conjunto de puntos, (x,y), el que constituye nuestra estimación de $\phi(x)$, la función de densidad de probabilidad de la población subyacente, y a la que denominaremos $\hat{\phi}(x)$. En segundo lugar, la exposición se realizará ahora en términos de la variable x y una muestra de n observaciones, cuya función de densidad subyacente tratamos de estimar, de esta forma, para ganar intuición y claridad en los argumentos, expondremos el caso de la estimación simple mencionando a lo largo del texto las modificaciones correspondientes para la estimación de la densidad en términos ponderados.

La mejor forma de entender la construcción de una estimación no paramétrica de $\phi(x)$ es, probablemente, partir del **histograma** (Kendall y Stuart (1977, Cap.-1)), que constituye en realidad el estimador más antiguo y conocido de la función de densidad. Un histograma no es más que el conjunto de rectángulos que aparece en el gráfico 6 detrás de

la densidad estimada por *BESFIT* [54] (Palisade (1997)), su **construcción** se basa en elegir un **punto de origen**, digamos x_0, y una **longitud**, h, que permitan definir los **intervalos que dan base al histograma** y que vienen definidos por $[x_0 + m.h, x_0 + (m+1).h) m \in \mathbf{N}$, donde los intervalos se eligen cerrados por la izquierda y abiertos por la derecha para que la definición sea apropiada y no existan ambigüedades. Una vez determinados los intervalos el histograma, $H(x)$, se define como

$$H(x) = \frac{\Sigma_{i=1}^{n} I\left(x \& x_i \in [x_0 + m.h, x_0 + (m+1).h)\right)}{n.h} m \in \mathbf{N} \qquad (19)$$

donde $I(\bullet)$ es la función índice que toma el valor 1 si se cumple la condición y 0 en caso contrario y el divisor asegura que la suma de las áreas de los rectángulos es igual a la unidad[55], lo que permite dar al histograma una interpretación en términos de frecuencias relativas. El numerador de (29) simplemente determina el número de observaciones x_i que pertenecen al mismo intervalo que x.

En muchas ocasiones la longitud del histograma se limita al rango de variación de la variable de forma que el histograma se representa entre los valores mínimo, $x_{(1)}$, y máximo, $x_{(n)}$, de x; en este caso $x_0 = x_{(1)}$ y dado exógenamente el número de intervalos en los que clasificar las observaciones[56], $k \geq 1$, el valor de h se determina automáticamente como $h = \frac{x_{(n)} - x_{(1)}}{k}$, donde obviamente $x_{(n)}$ siempre se asigna al último intervalo, ahora m toma valores en el rango $0 \leq m < k$. En este caso la elección de k determina automáticamente la longitud del intervalo, h, y el rango de variación de m. Esta es la forma en la que los histogramas de los gráficos 6 y 7 han sido generados.

---

[54] De hecho este programa organiza la información mediante un histograma antes de proceder a ajustar las funciones de densidad paramétricas.

[55] En este sentido n en el divisor es opcional y en ocasiones no se incluye, h es opcional sólo si la longitud de todos los intervalos es la misma, en otro caso una impresión visual correcta requerirá el correspondiente ajuste según la longitud del intervalo (Kendall y Stuart (1977, Cap.-1)).

[56] Por razones que se harán evidentes a lo largo de esta sección el problema de la selección del número de intervalos en la construcción de histogramas es enteramente equivalente a la selección del tamaño de la "ventana" o parámetro de suavizado en la estimación no paramétrica de la función de densidad. En ausencia de información *a priori* podemos utilizar la aproximación normal de Scott (1979) y determinar k como $k = \left[(4n)^{2/5}\right]$, donde [•] indica la "parte entera de".

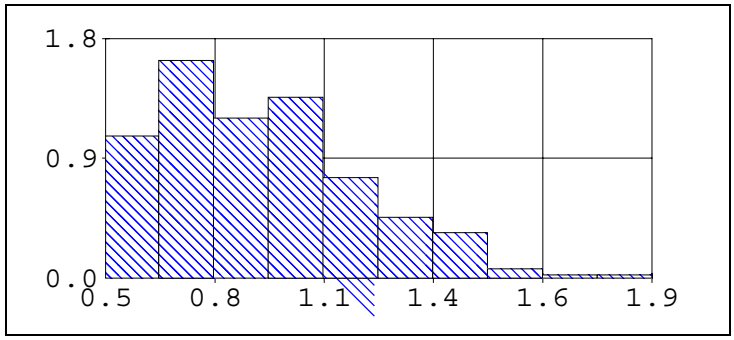
Obsérvese que la construcción de un histograma exige **dos elecciones**, un **punto de origen**, x_0, y una **longitud de intervalo**, h, y que es esta última elección la que fundamentalmente determina el grado de suavidad (*smoothing*) que impondremos sobre la representación gráfica, es decir sobre los datos[57]. Esta última afirmación puede ser vista claramente si consideramos los dos casos extremos entre los que puede variar h, por una parte si h cubre todo el rango de la variable entonces $H(x) = \frac{1}{h} \quad \forall x \in [x_{(1)}, x_{(n)}]$ con lo que obtendremos un histograma plano completamente inútil para describir los datos, en el otro extremo si h es tan pequeño que en cada intervalo sólo entra como máximo una observación entonces obtendremos tantos rectángulos como observaciones, todos ellos de la misma altura, $H(x) = \frac{1}{n.h}$, lo que tampoco nos permitirá ver nada interesante, como argumenta Scott (1992) tenemos ahora un problema de "demasiada tinta". Sin acudir a ejemplos tan extremos el efecto del valor de h sobre la impresión visual que transmite el histograma puede ser observada en el gráfico 9, en el que $x_0 = x_{(1)}$ y $h = \frac{x_{(n)} - x_{(1)}}{k}$, de forma que la representación gráfica abarca sólo el rango de la variable, x.

Podemos observar en el gráfico 9 como un valor de h demasiado grande, pocos intervalos considerados en (*i*), hace que perdamos características importantes de los datos, en otras palabras las observaciones han sido suavizadas en exceso, mientras que un valor de h demasiado pequeño, muchos intervalos considerados en (*iii*), da la impresión de una estructura demasiado errática en las observaciones, en ambos casos se hace difícil extraer conclusiones, en el primer caso, por ejemplo, la estructura de las observaciones en el entorno de 1 está prácticamente ausente del histograma, mientras que en el último caso aparecen demasiadas puntas de las que no parece extraerse un patrón excesivamente claro. Obviamente un valor de h entre estos dos extremos, (*ii*), proporciona no sólo una mejor impresión visual sino una estructura más interpretable.

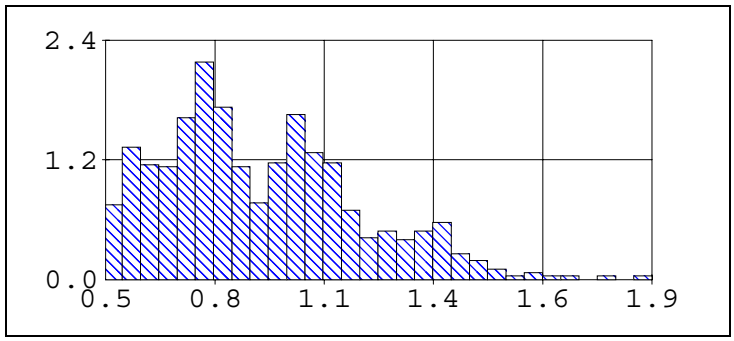
[57] La elección del punto de origen es menos problemática aunque no está exenta de problemas (Silverman (1986, Cap.-2.2)).

Gráfico 9

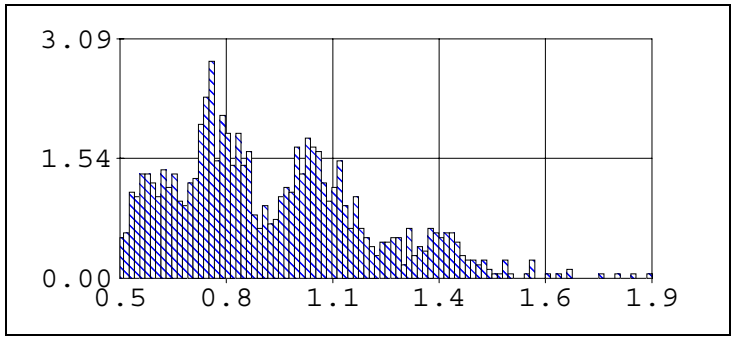
(*i*) $k = 10$, $h = 0.142557$



(*ii*) $k = 30$, $h = 0.047519$



(*iii*) $k = 100$, $h = 0.014256$



Dos observaciones son importantes antes de proseguir: **(*i*)** el histograma puede ser generalizado permitiendo **longitudes de los intervalos variables**; formalmente, definidos k intervalos de longitud variable, h_k, que cubran el rango de las observaciones, el histograma queda definido ahora como

$$H(x) = \frac{1}{n} \cdot \frac{\Sigma_{i=1}^{n} I(x_i \text{ en el mismo intervalo que } x)}{\text{longitud del intervalo que contiene } x} \qquad (20)$$

lo que obviamente añade complejidad a la elección de los diferentes *h* adecuados para cada intervalo. Esta es una situación corriente en los estudios sobre distribución de la renta con datos microeconómicos donde pocas observaciones en la cola derecha de la distribución hacen aconsejable la ampliación del los intervalos en los tramos de renta más elevados (Cowell (1995, Cap.-5)).

Y ***(ii)***, la introducción de **ponderaciones** en el análisis simplemente requiere sumar el peso N_i de la observación x_i cuando esta pertenece a un intervalo determinado, puesto que esta observación cuenta como N_i observaciones, y sustituir el tamaño muestral, *n*, por la suma de los pesos, $N = \Sigma_{i=1}^{n} N_i$, de esta forma **sumamos pesos en lugar de observaciones**; con ello (19) se redefine como

$$\begin{aligned} H_{\omega}(x) &= \frac{\Sigma_{i=1}^{n} N_i . I\left(x \,\&\, x_i \in [x_0 + m.h,\ x_0 + (m+1).h)\right)}{N.h} \\ &= \frac{\Sigma_{i=1}^{n} p_i . I\left(x \,\&\, x_i \in [x_0 + m.h,\ x_0 + (m+1).h)\right)}{h} \end{aligned} \qquad m \in \bigstar \qquad (21)$$

y (20), cuando los intervalos son variables, como

$$\begin{aligned} H_{\omega}(x) &= \frac{\Sigma_{i=1}^{n} N_i . I\left(x_i \text{ en el mismo intervalo que } x\right)}{N_x(\text{longitud del intervalo que contiene } x)} \\ &= \frac{\Sigma_{i=1}^{n} p_i . I\left(x_i \text{ en el mismo intervalo que } x\right)}{\text{longitud del intervalo que contiene } x} \end{aligned} \qquad (22)$$

Los histogramas son muy útiles como instrumentos para describir ciertas características de los datos pero son claramente insuficientes como estimaciones de $\phi(x)$. En primer lugar, puesto que nuestra variable es continua, desearíamos una estimación suficientemente suave de $\phi(x)$ como para que no presentara discontinuidades, en general podemos suponer que la densidad de *x* es diferenciable en todo el dominio de definición de nuestra variable y desearíamos que esta característica fuera reflejada por nuestra estimación, por tanto las discontinuidades y saltos asociados al histograma son un grave inconveniente para su utilización en ciertos contextos. En segundo lugar, es posible mejorar la precisión y la eficiencia con la que son utilizadas las observaciones por un

histograma en términos de varias descripciones matemáticas generalmente aceptadas de precisión o eficiencia, por ello parece natural tratar de mejorar nuestra estimación de $\phi(x)$ a partir de un histograma. Vale la pena señalar, sin embargo, que los histogramas fueron los únicos estimadores no-paramétricos posibles de $\phi(x)$ hasta mediados de los 50 en que Rosenblatt (1956) y Whittle (1958) propusieron el tipo de estimadores que consideraremos a continuación.

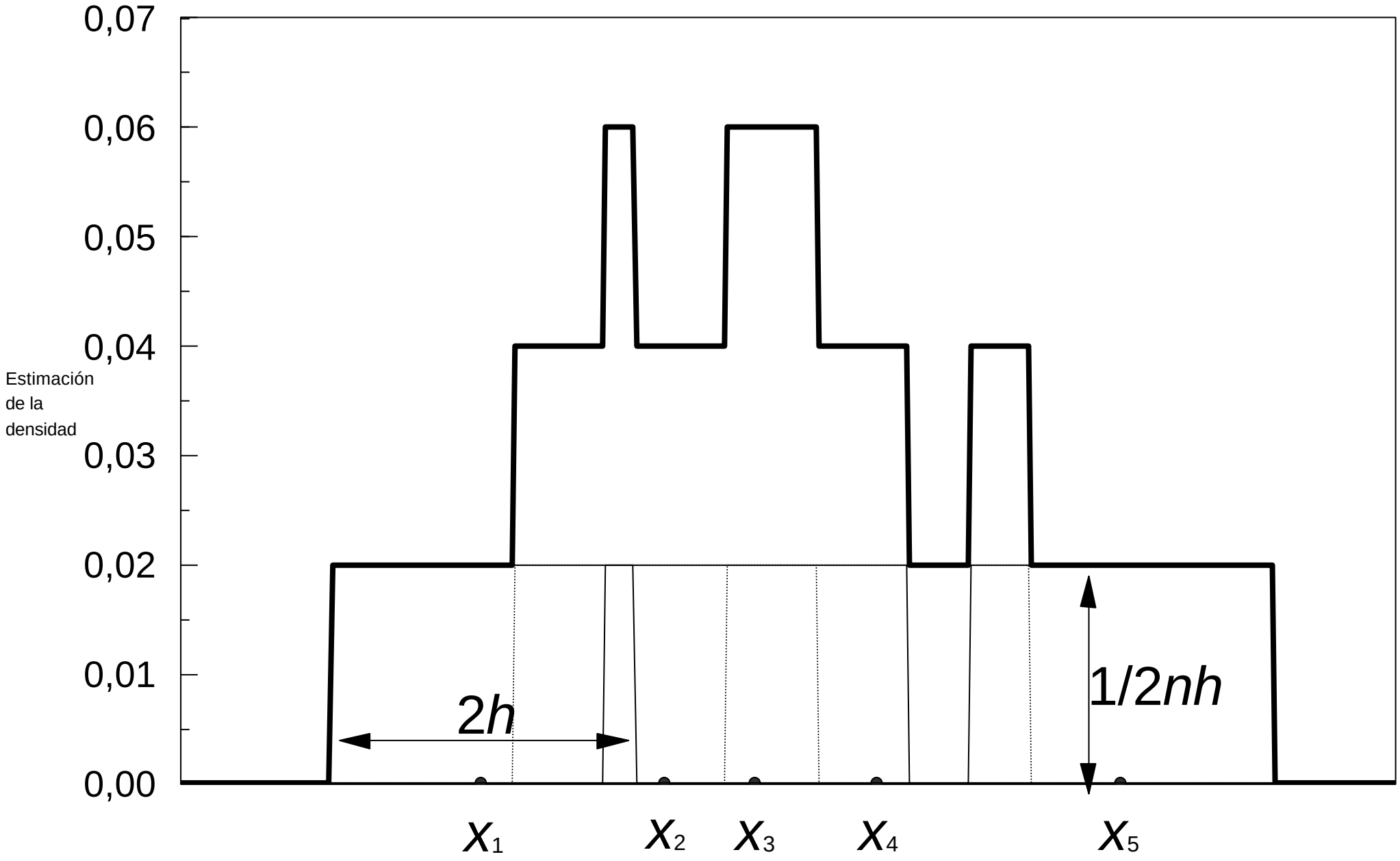
Supongamos que cada observación, x_i, de nuestra muestra proporciona cierta información acerca de la densidad subyacente a las observaciones dentro de un intervalo alrededor de x_i y dibujemos, con centro en cada x_i, un rectángulo de base $2h$ y altura $\frac{1}{2nh}$, obtengamos ahora la ordenada de nuestra función, $y = \phi(x)$, como la suma de las alturas de los rectángulos superpuestos, el gráfico 10 ilustra esta situación con 5 observaciones.

Este estimador, que fue inicialmente propuesto por Fix y Hodges (1951), puede ser visto como un intento de construir un histograma donde cada observación, x_i, sea el centro de un intervalo, liberando de esta forma al histograma de una elección particular del origen del mismo; queda por resolver, sin embargo la elección de la longitud de los intervalos, controlada por el parámetro h, y que determina el grado de suavidad que impondremos sobre los datos.

El estimador representado en el gráfico 10 mejora, respecto al histograma del gráfico 9, la eficiencia con la que los datos son utilizados pero no es todavía un estimador satisfactorio como estimador de nuestra función $\phi(x)$, al igual que sucedía con el histograma presenta saltos, ahora en los puntos $x_i \pm h$, por lo que la estimación no es suficientemente suave, esta naturaleza de escalones puede proporcionar una visión demasiado errática de las observaciones, además no recoge la diferenciabilidad de $\phi(x)$ y finalmente todavía es posible mejorar la precisión y la eficiencia con la que son utilizadas las observaciones. Por todo ello consideraremos una generalización de este estimador.

Sin embargo desde nuestro punto de vista lo importante es que el estimador que hemos descrito puede ser formulado algebraicamente como

Gráfico 10. *Kernel* rectangular



$$\hat{\phi}(x) = \frac{1}{nh}\sum_{i=1}^{n}\omega\left(\frac{x - x_i}{h}\right) \tag{23}$$

siendo $\omega(\bullet)$ una función de ponderación definida por

$$\omega(s) = \begin{cases} \frac{1}{2} & \text{si } |s| < 1 \\ 0 & \text{en otro caso} \end{cases} \tag{24}$$

A partir de la formulación (23) es fácil superar algunas de las dificultades asociadas al estimador definido por (23)-(24), simplemente remplazando la función de ponderación $\omega(\bullet)$ por una función continua y diferenciable, $K(\bullet)$, tal que

$$\int_{-\infty}^{+\infty} K(s)\mathrm{d}s = 1 \tag{25}$$

con lo que el estimador que estamos buscando queda definido como

$$\hat{\phi}(x) = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x - x_i}{h}\right) \tag{26}$$

este estimador de $\phi(x)$ es conocido en la literatura estadística sobre estimación no-paramétrica de funciones de densidad como **estimador *kernel*** con *kernel* $K(\bullet)$, del que el estimador (23)-(24) es un caso particular. Este será el único tipo de estimador de $\phi(x)$, además del histograma, que consideraremos en este trabajo, aunque no es el único existente (Silverman (1986), Scott (1992), Simonoff (1996)). La **construcción** de dicho estimador requiere, al igual que en el caso del histograma, **dos elecciones**, la **función *kernel***, $K(\bullet)$, y el **parámetro de suavizado**, h, también llamado **ancho de "ventana"** o **de banda**, por diversos autores. De estas dos elecciones nos ocuparemos a continuación pero primero debemos ganar algo de intuición acerca de nuestra estimación.

Las propiedades de $K(\bullet)$ hacen que **usualmente**, aunque no siempre, $\boldsymbol{K(\bullet)}$ sea una **función de densidad de probabilidad simétrica**, siendo la **densidad normal** la más

utilizada, *kernel* gaussiano; de hecho cualquier función de densidad es un candidato adecuado como elección de $K(\bullet)$, correspondiendo la función $\omega(\bullet)$ a la distribución uniforme.

La generalización llevada a cabo al sustituir (23)-(24) por (25)-(26) puede ser vista gráficamente en el gráfico 11, donde los rectángulos del gráfico 10 han sido sustituidos por funciones de densidad normales,

$$K(s) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{s^2}{2}\right\} \tag{27}$$

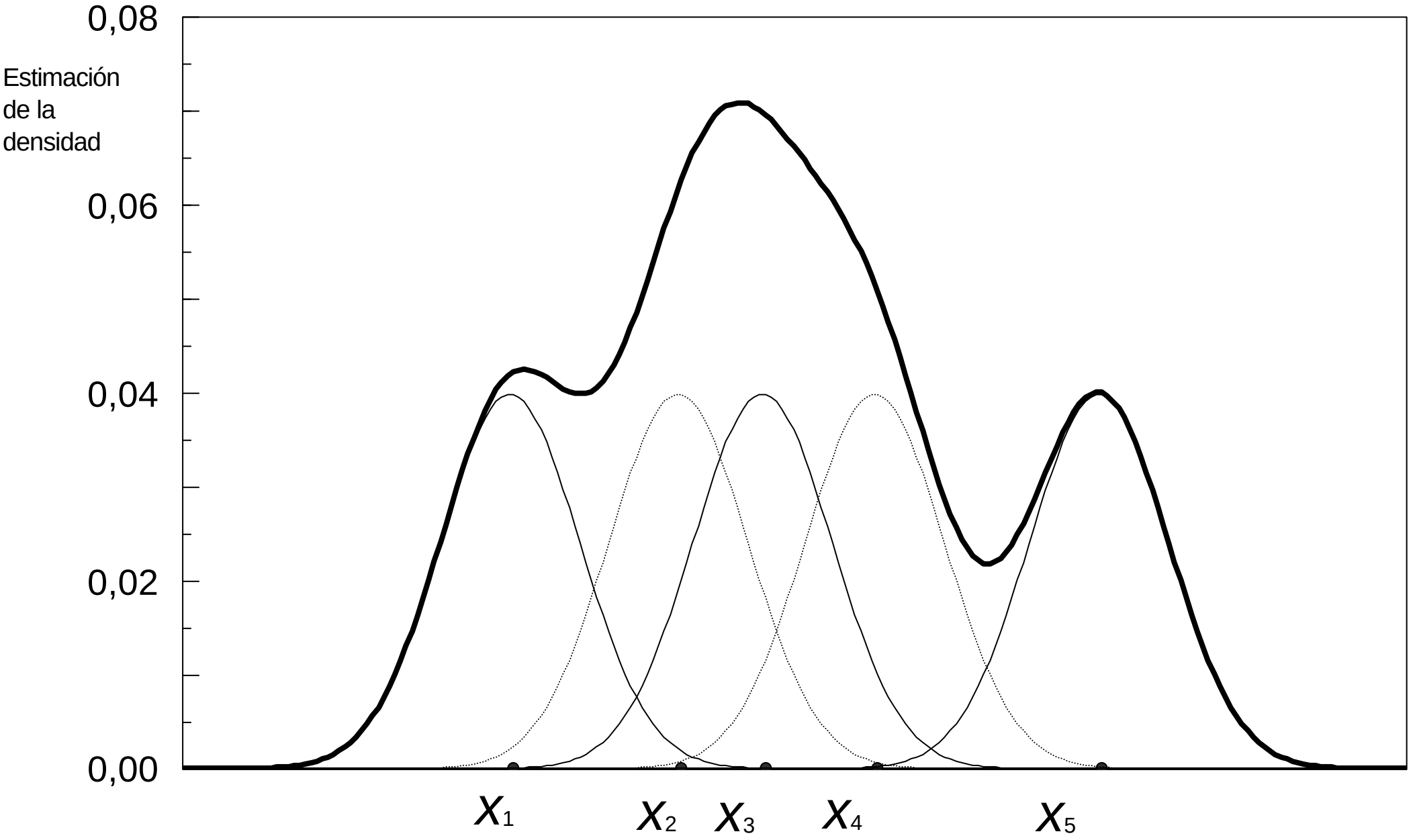
de forma que en dicho gráfico la ordenada y es obtenida como

$$y = \hat{\phi}(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \exp\left\{-\frac{1}{2}.\left(\frac{x - x_i}{h}\right)^2\right\} \tag{28}$$

Así pues las ordenadas de nuestra función, $y = \phi(x)$, se obtienen como la suma de las alturas de las densidades superpuestas.

El gráfico 11 y la fórmula (28) permiten observar el efecto que la elección del **parámetro de suavizado**, h, tiene sobre la forma estimada de $\phi(x)$. A partir de (28) observamos que este parámetro juega el papel de la desviación típica en una función de distribución normal, por lo tanto un h demasiado elevado aumentará la dispersión alrededor de cada x_i y todos los detalles existentes en las observaciones quedarán difuminados, por el contrario un h demasiado pequeño generará densidades muy poco dispersas alrededor de cada x_i, lo que tenderá a poner de manifiesto una estructura demasiado errática en las observaciones. No es difícil intuir que **la elección de *h* será crucial** en los resultados que obtengamos por lo que deberemos analizar esta cuestión más detalladamente. Observamos por tanto que los efectos en la elección de h son enteramente equivalentes a los ya ilustrados para el caso de la construcción de histogramas.

Gráfico 11. *Kernel* normal



Antes de considerar las dos elecciones básicas para la construcción de nuestro estimador (25)-(26), la **función *kernel***, $K(\bullet)$, y el **parámetro de suavizado**, h, deberemos introducir las **ponderaciones** en el análisis. Al igual que en el caso del histograma simplemente **sustituimos 1/*n* (36) por la frecuencia relativa, p_i, de forma que la ordenada en cada punto se obtiene como una suma ponderada en lugar de como una suma simple** (DiNardo, Fortin y Lemieux (1996)), esto da lugar al siguiente **estimador ponderado no-paramétrico de la función de densidad**

$$\hat{\phi}_{\omega}(x) = \frac{1}{h}\sum_{i=1}^{n} p_i K\left(\frac{x - x_i}{h}\right) \tag{29}$$

que cuando $K(\bullet)$ viene dado por la densidad normal queda especificado como

$$\hat{\phi}_{\omega}(x) = \frac{1}{h\sqrt{2\pi}}\sum_{i=1}^{n} p_i . \exp\left\{-\frac{1}{2}.\left(\frac{x - x_i}{h}\right)^2\right\} \tag{30}$$

El efecto de las ponderaciones sobre la estimación de la función de densidad puede observarse visualmente en los gráficos 12 para el caso de los rectángulos y 13 para el caso de una función *kernel* normal, donde a las observaciones de los gráficos 10 y 11 se les han asignado diferentes ponderaciones, en estos últimos gráficos todas las observaciones tenían el mismo peso en la estimación de $\phi(x)$, un 20% cada una, supongamos ahora que las ponderaciones asignadas a las observaciones son las siguientes 15% para x_1, 25% para x_2, 30% para x_3, 20% para x_4 y 10% para x_5, de esta forma una mayor masa de probabilidad es asignada a las observaciones centrales. Obsérvese como las diferentes ponderaciones alteran la altura de los rectángulos o las densidades asignadas a cada observación individual pero mantienen constante el tamaño de la "ventana", es decir el valor de h.

Los gráficos 12 y 13 muestran como lógicamente al asignar más peso a las observaciones centrales de la distribución la estimación de la densidad se hace más puntiaguda alrededor de su centro y más plana en las colas, lo que es particularmente evidente en el caso del *kernel* normal, obsérvese que la escala del eje de ordenadas en estos gráficos y su contrapartida no ponderada es diferente. De esta forma se vuelve a ilustrar

Gráfico 12. *Kernel* rectangular

Densidad ponderada

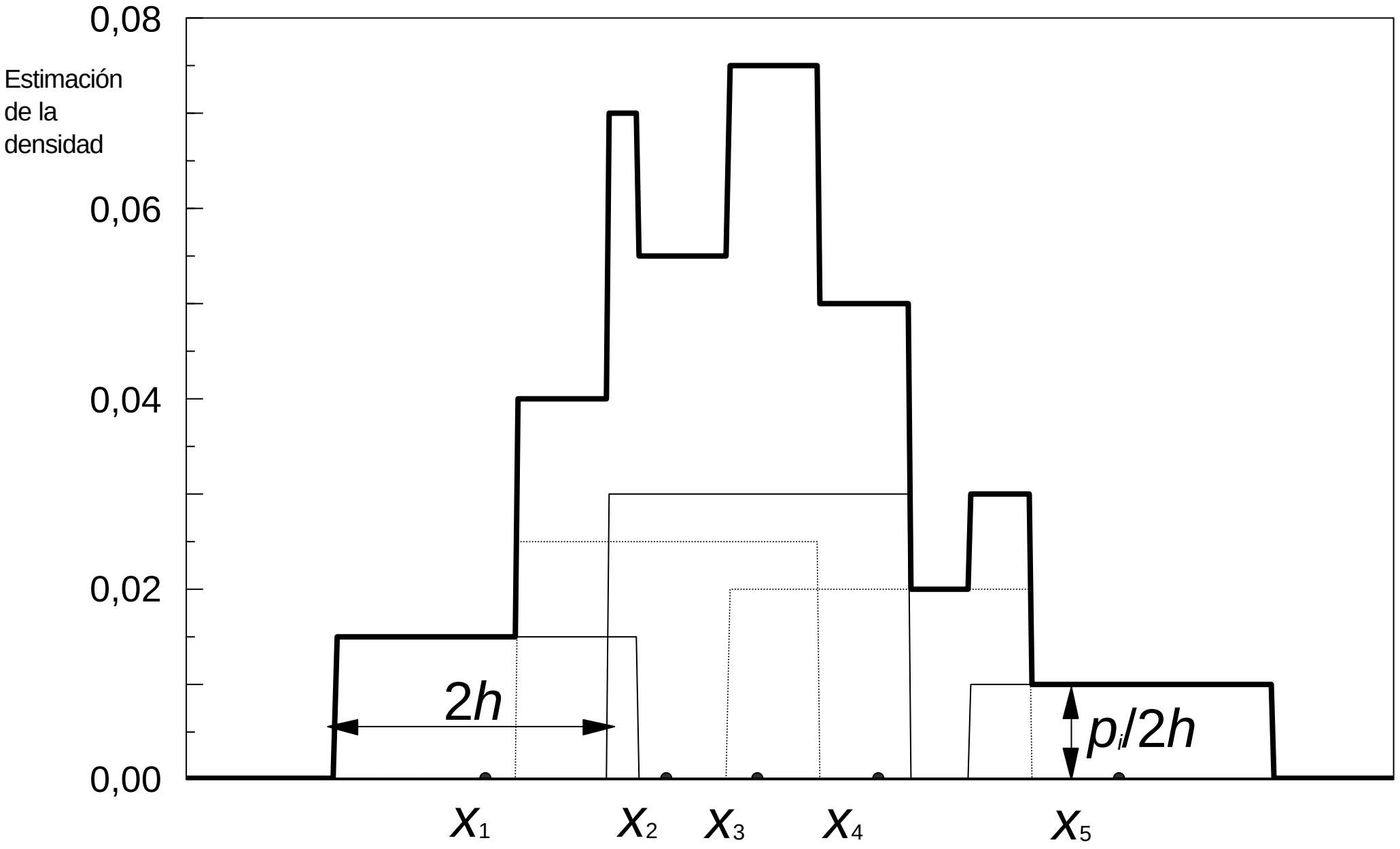
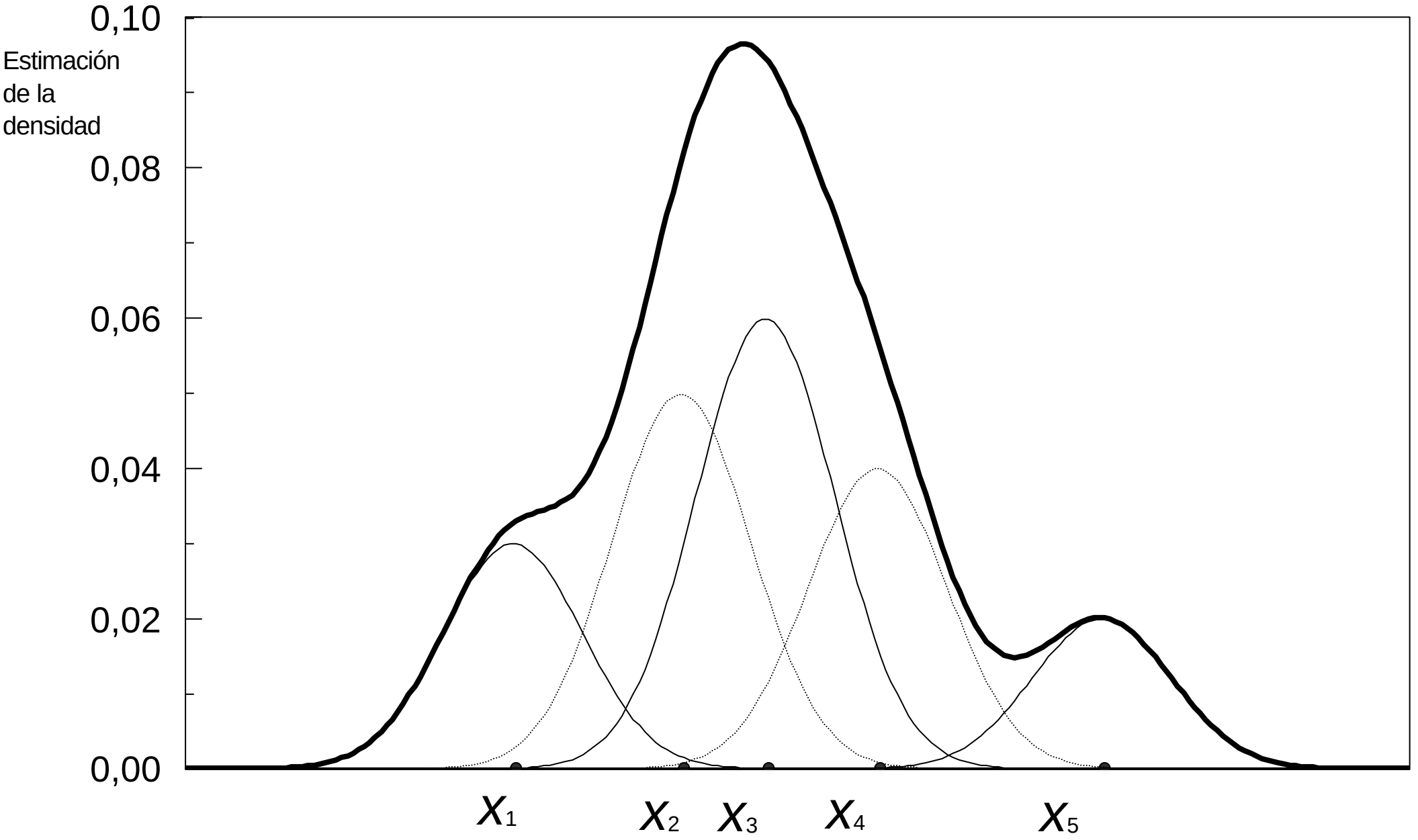


Gráfico 13. *Kernel* normal

Densidad ponderada



como la consideración de estadísticos ponderados puede proporcionar visiones muy diferentes de la que proporcionaría su contrapartida no ponderada, como ya mencionamos anteriormente.

Esta es la forma habitual, (29)-(30), en la que las ponderaciones son introducidas en el cálculo de densidades ponderadas si bien como veremos a continuación no es la única, ni quizá la forma más conveniente de hacerlo.

La función kernel, K*(•)*

Las propiedades de nuestro estimador *kernel* de $\phi(x)$ se derivan directamente de su definición, (25)-(26), y de las propiedades de la función *kernel*, *K*(•). De hecho es posible demostrar que nuestra estimación es de la forma (Silverman (1986, Cap.-3.2.1))

Densidad subyacente de la población suavizada + Perturbación aleatoria

donde *"densidad subyacente de la población suavizada"* depende de forma determinista de la elección particular de los parámetros del método de estimación, pero no directamente del tamaño muestral, *n*. De esta forma tanto **el sesgo** como **la varianza de $\hat{\phi}(x)$ dependen de** las dos elecciones básicas para la construcción del estimador, **la función *kernel*, *K*(•), y el parámetro de suavizado, *h***. Por supuesto sólo en la medida en que *h* sea elegido como función de *n* el sesgo y la varianza de $\hat{\phi}(x)$ dependerán indirectamente del tamaño muestral.

La elección de la función *kernel* es menos problemática por lo que comenzaremos comentando sobre esta primera. Supuesto que ***K*(•)** satisface (25) y no-negatividad, en otras palabras se trata de una **función de densidad de probabilidad**, entonces **la estimación de $\phi(x)$ será una función de densidad de probabilidad**, que es lo que queremos[58].

[58] Existen algunos argumentos en favor de las funciones *kernel* que toman valores negativos y positivos (Parzen (1962), Bartlett (1963), Müller (1984), Silverman (1986, Cap.-3.6)) ya que pueden reducir los sesgos

Adicionalmente $\hat{\phi}(x)$ **heredará todas las propiedades de continuidad y diferenciabilidad de la función *kernel*, $K(\bullet)$,** de forma que si $K(\bullet)$ es la distribución normal entonces la estimación de $\phi(x)$ será una curva suave, continua y con derivadas de todos los órdenes.

Estos argumentos apoyan la **utilización de la función de densidad normal**, (27), **como *kernel* en nuestra estimación** ya que sus propiedades son ampliamente conocidas, tiene derivadas de todos los órdenes y no impone requerimientos de cálculo excesivos. El *kernel* gaussiano no es sin embargo el único posible, el cuadro 2 ofrece algunos de los más utilizados, todos ellos funciones de densidad, en orden decreciente de eficiencia[59]. El *kernel* de Epanechnikov (1969) es el más eficiente (Hodges y Lehmann (1956)) y la eficiencia del resto se mide respecto a la de este último, sin embargo vale la pena mencionar que las eficiencias relativas de todos los *kernels* del cuadro 2 son superiores al 92.50%, siendo la eficiencia del *kernel* gaussiano del 95.12%, el **mensaje** es por tanto que **la elección importante no es la del *kernel***, sino como veremos a continuación la del parámetro de suavizado, h.

---

en la estimación de $\phi(x)$, sin embargo ellos no garantizan que dicha estimación sea una función de densidad de probabilidad y por tanto no serán tratados en este trabajo.

[59] Puesto que nuestra estimación es una función la eficiencia se mide respecto a una medida global de precisión, el llamado **error cuadrático medio integrado** (Rosenblatt (1956)), y supone que h ha sido elegido de forma óptima.

Cuadro 2: Funciones *kernel*

Kernel	$K(\bullet)$
Epanechnikov	$K(s)=\begin{cases}\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}s^2\right) & \text{si } \lvert s\rvert<\sqrt{5}\\ 0 & \text{en otro caso}\end{cases}$
Biponderado	$K(s)=\begin{cases}\frac{15}{16}\left(1-s^2\right)^2 & \text{si } \lvert s\rvert<1\\ 0 & \text{en otro caso}\end{cases}$
Triangular	$K(s)=\begin{cases}1-\lvert s\rvert & \text{si } \lvert s\rvert<1\\ 0 & \text{en otro caso}\end{cases}$
Normal (Gaussiano)	$K(s)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{s^2}{2}\right\}$
Rectangular	$K(s)=\begin{cases}\frac{1}{2} & \text{si } \lvert s\rvert<1\\ 0 & \text{en otro caso}\end{cases}$

Fuente: Silverman (1986), p.-43.

El parámetro de suavizado, **h**

Ya hemos mencionado como la elección de *h* afecta a nuestra estimación de $\phi(x)$, si el grado de suavizado es insuficiente entonces la densidad estimada resultante contendrá características espurias fruto de la variabilidad muestral en los datos, por el contrario si el grado de suavizado es excesivo, características importantes de la muestra se perderán en el proceso de construcción del estimador. Por tanto un *h* demasiado pequeño tendrá mostrar una fina estructura espuria en gran parte, mientras que un *h* demasiado grande oscurecerá la verdadera estructura de los datos, por ejemplo es posible en este último caso suavizar una estructura bimodal haciendo que esta no aparezca en nuestro estimador. El mismo fenómeno era observado en la construcción de histogramas.

La experiencia práctica ha demostrado que la elección de h es de crucial importancia en la impresión visual que obtengamos de nuestra estimación de $\phi(x)$ por lo que la literatura ha sugerido, y sigue haciéndolo, un sinfín de métodos sin que parezca existir un consenso definitivo sobre el tema (Silverman (1986, Cap.- 3.4), Park y Marron (1990), Sheather y Jones (1991), Scott (1992), Wand y Jones (1994), Simonoff (1996), Jones, Marron y Sheather (1996)). En los comentarios que siguen a continuación trataremos de ilustrar de forma general la problemática en la elección de h para acabar decantándonos por una elección concreta que aunque no tiene por que ser la óptima creemos que es suficientemente robusta.

Desde el punto de vista estadístico **la elección del parámetro de suavizado, *h*, implica el conocido *trade-off* entre sesgo y varianza**, el sesgo puede ser reducido eligiendo un h pequeño, pero ello tiene un coste en términos de incremento de la varianza; por otra parte la elección de un h elevado reducirá la varianza pero incrementará el sesgo en la estimación. Por lo tanto la elección de h implica siempre un *trade-off* entre el error sistemático (sesgo) y el aleatorio (varianza), este es un resultado general. En este contexto parece natural tratar de minimizar el error cuadrático medio, pero puesto que nuestra estimación es una función y no un parámetro puntual necesitamos una medida global de precisión, la medida generalmente utilizada en este contexto es el **Error Cuadrático Medio Integrado (ECMI)** (Rosenblatt (1956)), que se define como

$$\mathrm{ECMI}(\hat{\phi}) = E\int \left\{\hat{\phi}(x) - \phi(x)\right\}^2 \mathrm{d}x \qquad (31)$$

El valor óptimo de h, h_{opt}, será aquel que minimice (31); en la práctica dicha expresión es intratable analíticamente por lo que es sustituida por una aproximación asintótica, $n\to\infty$, que desafortunadamente muestra como h_{opt} depende a su vez de la densidad poblacional que queremos estimar, $\phi(x)$, y que obviamente es desconocida (Silverman (1986, Cap.-3.3)). Sin embargo una característica importante de h_{opt} es que $h_{\mathrm{opt}}\to 0$ conforme $n\to\infty$, dicho en palabras, conforme aumenta n el parámetro de suavizado, h, debe disminuir; la razón intuitiva es que el estimador de $\phi(x)$ debe ser más local cuanta más información tengamos.

La dependencia de h_{opt} de la densidad poblacional desconocida hace que una forma natural de elegir el valor de h en la práctica sea mediante la referencia a una familia paramétrica de distribuciones. Puesto que el parámetro de escala (dispersión) es muy importante para la elección de h, pero el de posición no lo es, una elección natural vuelve a ser la distribución normal, $N(0, \sigma^2)$. Silverman (1986, Cap.-3.4.2) muestra que en este caso

$$h_{opt} = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5} \tag{32}$$

siendo $\sigma = SD(x)$. Por tanto una forma rápida de obtener h es simplemente calcular la desviación típica de nuestra muestra e introducir dicha estimación en (32).

La experiencia práctica y los estudios de monte carlo (Jones, Marron y Sheather (1996)) muestran que el valor de h obtenido a partir de (32) funciona bien si la densidad de la población subyacente no se aleja mucho de la normal, pero que tiende a suavizar en exceso los datos en caso contrario, ya sea como resultado de la asimetría o la curtosis o como consecuencia de que la densidad de la población es multimodal. En general es posible obtener mejores resultados si $\sigma = SD(x)$ en (32) es sustituido por un estadístico de dispersión más robusto, por ejemplo (32) para una densidad normal puede ser escrito en función del rango inter-cuartílico, $R(\xi_{.25})$, como

$$h_{opt} = \left(\frac{4}{3}\right)^{1/5} \frac{R(\xi_{.25})}{1.349} n^{-1/5} \tag{33}$$

lo que proporciona un mejor ajuste de $\hat{\phi}(x)$ a $\phi(x)$ en el caso de distribuciones asimétricas o con colas relativamente pesadas (Silverman (1986), p.-47). Desafortunadamente (33) suaviza los datos en exceso en el caso de distribuciones multimodales y acaba funcionando peor como criterio para la elección de h que (32) si las modas están muy separadas unas de otras, por ello Silverman (1986, p.-47) recomienda utilizar lo mejor de ambas fórmulas y utilizar como criterio de selección de h

$$h_{opt} = \left(\frac{4}{3}\right)^{1/5} An^{-1/5} \tag{34}$$

siendo

$$A = Min\left\{\sigma, \frac{R(\xi_{.25})}{1.349}\right\} \tag{35}$$

Al parecer (34)-(35) funciona razonablemente bien con densidades unimodales con colas más pesadas que la normal y no demasiado mal si densidad poblacional es moderadamente bimodal. Para reducir el exceso de suavizado detectado en los estudios de monte carlo para este estimador (Marron (1989), Scott(1992), Jones, Marron y Sheather (1996)) Silverman (1986, p.-47) recomienda reducir el factor de proporcionalidad $\left(\frac{4}{3}\right)^{1/5} \approx 1.06$ en (34) a un factor de 0.9 para un *kernel* normal. En consecuencia la elección de *h* vendrá dada por (35) y

$$h_{opt} = 0.9An^{-1/5} \tag{36}$$

En resumen la elección de *h* a partir de (35)-(36) parece proporcionar resultados razonables en un buen número de situaciones y su obtención es muy sencilla; sin duda alguna otros métodos de elección del parámetro de suavizado pueden ser mejores en otros contextos, aunque también su obtención es más compleja y menos intuitiva y no será considerada (Park y Marron (1990), Jones, Marron y Sheather (1996)).

Obsérvese que en la discusión sobre la elección de ***h*** que hemos realizado suponemos que este es un parámetro **constante** lo que implica que utilizamos el **mismo grado de suavizado en todos los puntos** de la muestra. La cuestión es similar a la de la longitud de los intervalos en la construcción de histogramas, aunque inicialmente consideramos histogramas en los que la longitud de los intervalos era siempre la misma ya observamos como estos podían ser generalizados permitiendo longitudes de los intervalos variables de forma que si en un tramo de la distribución existían pocas observaciones un intervalo más ancho daba continuidad al histograma, esto es típico en estudios sobre la

distribución de la renta con datos microeconómicos donde la existencia de pocas observaciones en el extremo superior de la distribución hacen aconsejable la ampliación de los intervalos en los tramos de renta más elevados (Cowell (1995, Cap.-5)). De forma similar algunos autores (Silverman (1986, Cap.-5.1), Scott (1992, Cap.-6.6)) han observado como en ciertos contextos es posible mejorar notablemente la estimación de $\phi(x)$ por medio de la utilización de **parámetros de suavizado locales**, es decir que se adapten a la densidad local de los datos. Obviamente esto requiere no la elección de un parámetro *h* sino la elección de una función completa que nos proporcione un valor de *h* para cada observación, h_i.

Aunque este tipo de estimadores de $\phi(x)$ no serán considerados si mencionaremos con cierto detalle el **estimador *kernel* adaptativo** (Breiman, Meisel y Purcell (1977), Abramson (1982), Silverman (1986, Cap.-5.3)) al proporcionar una forma alternativa a (29) de introducir las ponderaciones en el análisis.

La idea básica del **estimador *kernel* adaptativo** es simplemente construir un estimador *kernel* permitiendo que *h* varíe de una observación a otra de la muestra. El procedimiento se basa en la **intuición** ya mencionada de que **una forma natural de proceder en distribuciones con colas relativamente largas es usar un parámetro de suavizado mayor en regiones con poca densidad**, de la misma forma que los intervalos de un histograma eran ensanchados cuando teníamos pocas observaciones, de esta forma una observación en la cola de la distribución tendrá su masa de probabilidad esparcida sobre un intervalo más amplio.

La primera cuestión práctica que deberemos resolver es como decidir si una observación está en una región de baja densidad o no[60], el estimador *kernel* adaptativo solventa este problema por medio de un procedimiento en dos etapas. Una estimación inicial, digamos $\tilde{\phi}(x)$, es utilizada para hacernos una idea de $\phi(x)$; esta estimación es utilizada para generar h_i correspondientes a cada observación y finalmente estos parámetros de suavizado son utilizados para construir el estimador adaptativo propiamente dicho. Sin

[60] Obsérvese que cuando mencionamos los intervalos variables en los histogramas no indicamos nada acerca de como determinarlos y no es una cuestión obvia (Scott (1979)).

entrar en detalles específicos los **pasos a seguir en la construcción de un estimador *kernel* adaptativo** son los siguientes:

1. Obtener una **estimación inicial**, $\tilde{\phi}(x)$, que satisfaga $\tilde{\phi}(x_i) > 0 \quad \forall i$.

2. Definir los **factores de suavizado locales**, λ_i, como

$$\lambda_i = \left\{ \frac{\tilde{\phi}(x_i)}{g} \right\}^{-\alpha} \tag{37}$$

donde g es la media geométrica[61] de $\tilde{\phi}(x_i)$, $\log g = \frac{\Sigma_{i=1}^{n} \log \tilde{\phi}(x_i)}{n}$, y α es un parámetro de sensibilidad tal que $0 \leq \alpha \leq 1$. Obsérvese que para (37) la media geométrica de λ_i es igual a la unidad.

3. Construir el **estimador *kernel* adaptativo** como

$$\hat{\phi}_a(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h\lambda_i} K\left(\frac{x - x_i}{h\lambda_i} \right) \tag{38}$$

donde $K(\bullet)$ es la función *kernel* y h es el parámetro de suavizado global que junto con λ_i determina el grado de suavizado local impuesto sobre los datos, puesto que en la estimación final el parámetro de suavizado para x_i viene dado por $h\lambda_i$.

No entraremos en más detalles sobre este estimador, simplemente mencionar que la literatura (Breiman, Meisel y Purcell (1977), Abramson (1982), Silverman (1986, Cap.-5.3)) ha señalado que el estimador final es bastante insensible a la elección del estimador inicial, por lo que estimador *kernel* estándar con h constante será adecuado, y que hay buenas razones teóricas para fijar $\alpha = ½$ en (37) (Abramson (1982)), obsérvese que $\alpha = 0$ reduce el método al estimador *kernel* con parámetro de suavizado, h, fijo, puesto que en este caso $\lambda_i = 1 \ \forall i$.

[61] Podríamos utilizar igualmente la media aritmética pero entonces ni la media geométrica ni la aritmética de λ_i sería la unidad.

La introducción de ponderaciones en el estimador *kernel* adaptativo, (38), es trivial; a partir de (29),

$$\hat{\phi}_{\omega a}(x) = \sum_{i=1}^{n} \frac{p_i}{h\lambda_i} K\left(\frac{x - x_i}{h\lambda_i}\right) \tag{39}$$

El estimador *kernel* adaptativo puede ser utilizado para racionalizar la introducción de ponderaciones en la estimación de funciones de densidad de una forma alternativa a la considerada en (29) mediante el siguiente argumento. Ya hemos indicado como una buena elección de h requiere que el estimador de $\phi(x)$ sea más local cuanta más información tengamos y que en este mismo sentido la razón fundamental para considerar parámetros de suavizado locales se basa en la intuición de adaptar la estimación de $\phi(x)$ a la densidad local de los datos de forma que en regiones de alta densidad h debe ser menor que en regiones de baja densidad, donde deberemos utilizar un parámetro de suavizado mayor.

De hecho a partir de (37) podemos observar como para $0 < \alpha \leq 1$ cuando la densidad es alta

$$\tilde{\phi}(x_i) > g \qquad \Rightarrow \qquad \lambda_i < 1 \qquad \Rightarrow \qquad h\lambda_i < h$$

mientras que si la densidad es baja

$$\tilde{\phi}(x_i) < g \qquad \Rightarrow \qquad \lambda_i > 1 \qquad \Rightarrow \qquad h\lambda_i > h$$

por tanto el suavizado es menor que en promedio[62] para las observaciones situadas en regiones de alta densidad y mayor que en promedio para las observaciones situadas en regiones de baja densidad.

Cuando disponemos de una **muestra ponderada**, x_i con su peso, p_i, asociado, entonces **no todas las observaciones tienen el mismo contenido informativo acerca de**

[62] Recuérdese que la media geométrica de λ_i es la unidad.

la densidad subyacente, y esta era precisamente la razón para que pesaran de forma diferente las observaciones en la estimación de la densidad ponderada $\hat{\phi}_{\omega}(x)$, (29). Como observamos entonces esto alteraba la altura de los rectángulos o las densidades asignadas a cada observación individual, pero mantenía intacto el intervalo sobre el que dichos rectángulos o densidades se esparcían, esto es, el tamaño de la "ventana" (gráficos 12 y 13). Dicho con otras palabras $\hat{\phi}_{\omega}(x)$ es un estimador ponderado, pero no adaptativo.

Puesto que las observaciones para las que p_i es alto en relación a la ponderación media, $p_i > \frac{1}{n}$, son indicativas de una mayor densidad de probabilidad en ese punto, adaptar la estimación de $\phi(x)$ a la densidad local de los datos equivale en este sentido a utilizar un parámetro de suavizado menor para dicha observación; de igual forma observaciones para las que p_i es bajo en relación a la ponderación media, $p_i < \frac{1}{n}$, son indicativas de una menor densidad de probabilidad en ese punto y adaptar la estimación de $\phi(x)$ a la densidad local de los datos equivaldrá a utilizar un parámetro de suavizado mayor para dicha observación. Este argumento sugiere tomar $\lambda_i = \left\{ \frac{p_i}{1/n} \right\}^{-\alpha}$ en (38), es decir utilizar como factor de suavizado local, λ_i, para cada observación la inversa de la *ratio* entre su frecuencia poblacional, p_i, y su frecuencia muestral[63], $1/n$, ajustado por un factor de sensibilidad α tal que $0 \leq \alpha \leq 1$. Obsérvese que ahora ni la media aritmética ni la geométrica de λ_i es igual a la unidad.

De esta forma las **ponderaciones son introducidas en la estimación de $\phi(x)$ como indicativas de la densidad de probabilidad asociada a cada observación** y alteran la longitud, y no solo la altura, del intervalo ("ventana") sobre el que se supone que cada observación x_i proporciona información acerca de la densidad subyacente.

Dos **comentarios finales** respecto a esta cuestión son de interés:

[63] Obsérvese que esta *ratio* ya apareció cuando hablamos de la inferencia con estadísticos ponderados.

(*i*) Adicionalmente a la introducción de las ponderaciones a través de los factores de suavizado locales, λ_i, , es posible introducir las ponderaciones en la forma habitual; es decir utilizar el estimador (39), $\hat{\phi}_{\omega a}(x)$, con $\lambda_i = \left\{\frac{p_i}{1/n}\right\}^{-\alpha}$.

La literatura estadística es totalmente silenciosa a este respecto por lo que sería interesante considerar las diferentes alternativas de introducir las ponderaciones en el análisis y examinar la bondad de los diferentes estimadores. En particular es necesario examinar si ahora hay razones teóricas o prácticas para fijar $\alpha = ½$ o a cualquier otro valor en $\lambda_i = \left\{\frac{p_i}{1/n}\right\}^{-\alpha}$.

(*ii*) La consideración de $\lambda_i = \left\{\frac{p_i}{1/n}\right\}^{-\alpha}$ elimina la necesidad de la estimación inicial de $\phi(x)$, $\widetilde{\phi}(x)$, en la construcción del estimador adaptativo puesto que una vez conocemos p_i podemos calcular λ_i para un valor dado de α. Sin embargo el concepto de densidad asociado a las ponderaciones es muy diferente del concepto de densidad que motivó la definición del estimador *kernel* adaptativo (37)-(38), en este último caso densidad hacía referencia a si había muchas o pocas observaciones en un determinado intervalo, mientras que densidad en términos de las ponderaciones asociadas a las observaciones hace referencia al contenido informativo de cada observación en función de su peso dentro de la muestra. En la práctica es posible que ambos conceptos de densidad no coincidan, un ejemplo claro sería la renta *per capita* de Madrid en 1955, esta observación representa en dicho año un 7.61% de la población española, lo que está muy por encima de la frecuencia muestral del 2%, constituyendo desde este punto de vista una observación en una región de alta densidad lo que sugeriría la utilización de un parámetro de suavizado menor que el promedio, sin embargo esta observación aparece como un *outlier* con un valor de 1.93 en relación a la media nacional, lo que representa 3.11 desviaciones típicas simples y 1.97 veces $R(\xi_{.25})$, y a la que le sigue a bastante distancia la renta *per capita* de Vizcaya con un valor de 1.71 en relación al promedio nacional; por tanto Madrid se sitúa desde otro punto de vista

en una región de baja densidad lo que sugeriría utilizar un parámetro de suavizado mayor que el promedio, existen dos fuerzas sobre el suavizado que operan en direcciones opuestas y no es evidente la forma adecuada de proceder ya que la consideración de uno solo de los conceptos de densidad puede empeorar la estimación de $\phi(x)$ en lugar de mejorarla.

Desde un punto de vista práctico es posible combinar ambos conceptos de densidades y a partir de una estimación inicial de $\phi(x)$, $\widetilde{\phi}(x)$, definir

$$\lambda_i = \left\{ \frac{\widetilde{\phi}(x_i)}{g} . \frac{p_i}{1/n} \right\}^{-\alpha}$$

o incluso utilizar parámetros de sensibilidad diferentes en cada caso,

$$\lambda_i = \left\{ \left(\frac{\widetilde{\phi}(x_i)}{g} \right)^{-\alpha_1} . \left(\frac{p_i}{1/n} \right)^{-\alpha_2} \right\}$$

para $0 \le \alpha_1 \le 1$ y $0 \le \alpha_2 \le 1$, y utilizar este valor de λ_i en el estimador *kernel* adaptativo (38) ó (39). Obsérvese que en ninguno de estos casos la media geométrica de λ_i es igual a la unidad.

Dominios restringidos

Hasta ahora hemos procedido como si el dominio de definición de nuestra variable fuera toda la recta real, **R**, de hecho la densidad normal está definida en todo **R**, sin embargo con frecuencia nos encontramos con situaciones en las que la variable bajo estudio sólo puede tomar valores en un determinado rango, así por ejemplo la renta *per capita* sólo puede tomar valores no negativos, $x \in \mathbf{R}^+$, pero si aplicamos los métodos que hemos descrito hasta ahora podríamos obtener una estimación de $\phi(x)$ cuyo soporte incluyera valores negativos y por tanto fuera una estimación inaceptable. En la práctica ello

significa que debemos enfrentarnos al problema de asegurar que $\hat{\phi}(x)$ sólo esté definida para el rango de definición de la variable *x*. En la exposición que sigue a continuación supondremos que $x > 0$, pero idénticos métodos se aplican cuando el rango de definición de *x* es diferente.

La literatura ha utilizado básicamente **tres aproximaciones al problema de limitar el soporte de** $\hat{\phi}(x)$. En **primer lugar** la forma más sencilla de asegurar que $\hat{\phi}(x) = 0$ para $x \leq 0$ consiste simplemente en estimar $\hat{\phi}(x)$ para $x > 0$ y simplemente fijar $\hat{\phi}(x) = 0$ para $x \leq 0$. En este caso nuestra estimación es por tanto

$$y = \begin{cases} \hat{\phi}(x) & \text{para } x > 0 \\ 0 & \text{para } x \leq 0 \end{cases} \tag{40}$$

El principal **inconveniente** de este método para la estimación de funciones de densidad de probabilidad es que la densidad estimada no integra (necesariamente) la unidad. La solución natural a este problema consiste en escalar (truncar) la densidad estimada para convertirla en una verdadera función de densidad, con lo que nuestra estimación quedaría

$$y = \begin{cases} \dfrac{\hat{\phi}(x)}{\hat{\text{P}}\text{rob}(x > 0)} & \text{para } x > 0 \\ 0 & \text{para } x \leq 0 \end{cases} \tag{41}$$

Esta no es sin embargo una solución enteramente aceptable ya que aunque ahora nuestra estimación integra la unidad, la contribución a dicha estimación de los puntos en el entorno de cero será mucho menor que la contribución a la estimación de las observaciones suficientemente alejadas de cero, por tanto la estimación de $\phi(x)$ en el entorno de cero estará infra-estimada.

En **segundo lugar** es posible **transformar los datos**, estimar la densidad para los datos transformados y deshacer finalmente la transformación para obtener nuestra estimación de $\phi(x)$. Obviamente no existe un criterio universalmente válido en cualquier

situación pero en nuestro contexto podríamos tomar logaritmos, $\log x$, estimar no-paramétricamente la densidad para $\log x$, $\hat{\varphi}(\log x)$, y finalmente deshacer la transformación para obtener la densidad de x,

$$\hat{\phi}(x) = \frac{1}{x}\hat{\varphi}(\log x) \qquad \text{para } x > 0 \tag{42}$$

A pesar de ser propugnada por algunos autores (Copas y Fryer (1980)) esta aproximación al problema no ha ganado mucha popularidad en la práctica.

En **tercer lugar** es posible adaptar los métodos originalmente desarrollados para todo **R** al caso de dominios restringidos. La idea básica es **replicar o reflejar los datos** fuera del rango de definición de x, estimar la densidad para esta muestra ampliada y finalmente reflejar la densidad estimada dentro del rango de definición de la variable objeto de estudio (Boneva, Kendall y Stefanov (1971)).

Por ejemplo en nuestro caso, $x > 0$, podemos reflejar nuestras observaciones en $\mathbf{R}^-$, de la siguiente forma $(x_1, -x_1, x_2, -x_2, x_3, -x_3, \ldots, x_n, -x_n)$, creando una muestra de tamaño $2n$ para la cual podemos estimar $\phi(x)$ mediante el método *kernel* descrito anteriormente. Una vez disponemos de la estimación para esta muestra reflejada la estimación para las observaciones originales viene dada por

$$y = \begin{cases} 2\hat{\phi}(x) & \text{para } x > 0 \\ 0 & \text{para } x \leq 0 \end{cases} \tag{43}$$

Para el estimador *kernel* que hemos venido utilizando (43) es equivalente a estimar $\phi(x)$ mediante la siguiente función

$$\hat{\phi}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{x - x_i}{h}\right) + K\left(\frac{x + x_i}{h}\right)\right] \tag{44}$$

Supuesto que la función *kernel*, $K(\bullet)$, es simétrica y diferenciable es fácil demostrar que la estimación tendrá derivada nula en el origen y al mismo tiempo si $K(\bullet)$ es una función de densidad la estimación resultante también lo será[64].

Los comentarios realizados pueden adaptarse a casos en los que el soporte del estimador es un intervalo finito [a, b], siendo de aplicación práctica en economía el caso del intervalo [0, 1] ya que muchas variables de interés están en forma de tasas o porcentajes (Tortosa-Ausina (1999)). Así por ejemplo los métodos de transformación pueden basarse en transformaciones del tipo $\Phi^{-1}\left(\frac{x_i - a}{b - a}\right)$, donde Φ^{-1} es cualquier función de distribución acumulativa estrictamente creciente en **R**. Y de igual forma los métodos de replicado pueden reflejar las observaciones fuera de dicho intervalo a ambos lados de los límites del mismo, así en el caso de que necesitemos que la estimación de $\phi(x)$ se encuentre acotada dentro del intervalo [0, 1] es posible replicar la muestra en los intervalos [−1, 0] y [1, 2] obteniendo una muestra de tamaño 3*n* dada por

$$\left(x_1 - 1, x_2 - 1, x_3 - 1, \ldots, x_n - 1, x_1, x_2, x_3, \ldots, x_n, x_1 + 1, x_2 + 1, x_3 + 1, \ldots, x_n + 1\right)$$

y proceder como hemos indicado anteriormente[65].

Finalmente señalar que una vez la densidad de interés ha sido estimada y disponemos de $\hat{\phi}(x)$ es posible obtener los momentos, quantiles, medidas de desigualdad, incluyendo la curva de Lorenz (1905), o prácticamente cualquier otro estadístico o funcional que dependa de la densidad desconocida de forma similar a cuando disponíamos de una estimación paramétrica de $\phi(x)$, si bien ahora deberemos emplear métodos de simulación (Silverman (1986) Cap.-6.4 y 6.5). En este contexto una estimación no-paramétrica de la función de distribución acumulativa de probabilidad, $\hat{\Phi}(s) = \int_0^s \hat{\phi}(x)\mathrm{d}x$,

[64] En la práctica no es necesario reflejar todas las observaciones ya que aquellas suficientemente alejadas de cero no presentarán ninguna contribución a la estimación de $\phi(x)$. Omitiremos, sin embargo, detalles computacionales, muy numerosos en este campo.

[65] Es posible continuar replicado las observaciones más allá de los intervalos [−1, 0] y [1, 2] pero en la práctica no suele ser necesario.

obtenida mediante integración numérica de $\hat{\phi}(x)$, y sus quantiles asociados, resulta una estimación alternativa a la función de distribución acumulativa empírica mencionada en el epígrafe 2.3[66].

Contrastes de multimodalidad

Aunque ya hemos mencionado que nuestro interés no se centra en aspectos relacionados con la inferencia estadística vale la pena mencionar brevemente una de las aplicaciones más interesantes de la estimación no-paramétrica de $\phi(x)$, el contraste estadístico sobre la presencia de varias **modas** (máximos) locales en una distribución.

La estimación de $\phi(x)$ suele generar en la práctica varios máximos locales o cimas (Quah (1996a)) que pueden indicar alguna característica subyacente en la población que valdría la pena investigar con detalle. Así por ejemplo la literatura sobre crecimiento y convergencia económica ha interpretado la presencia de dos modas en la distribución de la renta *per capita* como indicativa de la formación de clubes o grupos diferenciados y en consecuencia como ausencia de convergencia global (Quah (1993b, 1996a,d,e, 1997), Bianchi (1995)).

La primera cuestión que debemos plantearnos una vez considerado el concepto de moda como máximo local en una función de densidad es si una moda realmente observada en $\hat{\phi}(x)$ se corresponde con una moda en la densidad poblacional, $\phi(x)$, o por el contrario se trata de un fenómeno debido a la variabilidad muestral o es fruto de la elección particular del parámetro de suavizado *h*; en otras palabras debemos ir más allá de una mera descripción de los datos para preguntarnos sobre la significación estadística de las modas observadas.

[66] Obsérvese que a menos que la muestra sea muy grande la estimación de los quantiles en los extremos de la distribución obtenidos a partir de la estimación de $\phi(x)$ estarán ahora fuera del rango de *x*.

Los contrastes no-paramétricos sobre multimodalidad en funciones de densidad se basan en el concepto de **parámetro de suavizado crítico**, h_{crit}, introducido por Silverman (1981, 1983), y definido como el **valor más pequeño de *h* que genera una densidad estimada unimodal**[67]. Resulta ilustrativo recordar a estos efectos el comportamiento de $\hat{\phi}(x)$ en función del valor de *h* para una muestra dada. Para un valor de *h* muy grande el grado de suavizado sobre los datos será tan elevado que esperaremos que la estimación de $\phi(x)$ sea unimodal, conforme *h* disminuye imponemos un grado de suavizado sobre los datos cada vez menor y en consecuencia existirá un punto en el que $\hat{\phi}(x)$ se convertirá en bimodal; disminuyendo *h* todavía más podremos hacer que aparezcan todavía más modas en $\hat{\phi}(x)$, en el límite para un valor de *h* suficientemente pequeño aparecerán tantas modas como observaciones, de igual forma que éramos capaces de construir un histograma en el que en cada rectángulo entrara una sola observación. La conclusión es por tanto que **el número de modas en $\hat{\phi}(x)$ es una función decreciente de *h*** (Silverman (1981)[68]) y por tanto podemos buscar un valor de *h* donde la estimación de $\phi(x)$ cambie de unimodal a multimodal, este es el valor de *h* crítico, h_{crit}.

En consecuencia h_{crit} verifica que para

$$h \geq h_{crit} \qquad \Rightarrow \qquad \hat{\phi}(x) \quad \text{unimodal}$$

mientras que para

$$h < h_{crit} \qquad \Rightarrow \qquad \hat{\phi}(x) \quad \text{multimodal}$$

Un simple procedimiento de búsqueda puede ser utilizado para determinar h_{crit} en la práctica con el grado de precisión adecuado.

---

[67] En general $h_{crit}(m)$ se define como el valor más pequeño de *h* que genera una densidad estimada con *m* modas; en el texto consideramos el caso de unimodalidad *versus* bimodalidad que es el de más relevancia práctica y por tanto $m = 1$, aunque los razonamientos son fácilmente generalizables a valores superiores de *m*.

[68] Técnicamente esta afirmación requiere que la función *kernel* cumpla ciertas propiedades, que son satisfechas si $K(\bullet)$ es la densidad normal, Silverman (1981) ofrece los detalles técnicos sobre esta cuestión.

Para **poblaciones bimodales** debemos esperar un valor de h_{crit} **relativamente grande** puesto que en este caso necesitaremos suavizar mucho los datos para que $\hat{\phi}(x)$ sea unimodal, por el contrario para poblaciones unimodales el grado de suavizado a imponer sobre las observaciones para que $\hat{\phi}(x)$ sea unimodal será mucho menor y en consecuencia el valor de h_{crit} también será menor. De hecho es posible demostrar (Silverman (1983)) que conforme $n \to \infty$ entonces $h_{crit} \to 0$ si la distribución de la población subyacente, $\phi(x)$, es unimodal, pero $h_{crit} \to \delta > 0$ si $\phi(x)$ es multimodal. Este resultado puede ser utilizado para realizar un contraste de hipótesis sobre el número de modas en $\phi(x)$ donde valores de h_{crit} relativamente grandes tenderán a indicar la existencia de más de una moda.

La cuestión ahora es como decidir cuando un valor concreto observado de h_{crit} es "relativamente grande", en la terminología de los contrastes de hipótesis necesitamos la distribución muestral del estadístico h_{crit}, que sin embargo no es conocida. Existen de nuevo dos aproximaciones para solucionar este problema. Una primera aproximación es comparar h_{crit} frente a una familia paramétrica de distribuciones unimodales, por ejemplo la normal, de esta forma podemos preguntarnos si para una muestra dada el valor observado de h_{crit} es inusualmente elevado respecto al que observaríamos si los datos provinieran de una distribución normal con la misma desviación típica que la muestra. En este sentido Jones (1983) mediante un estudio de monte carlo estima que valores de h_{crit} superiores a $1.25\sigma n^{-1/5}$ se observarán sólo en aproximadamente un 5% de los casos si la muestra proviene de una distribución normal con varianza σ^2 (Silverman (1986), p.-140), por lo que este valor, $1.25\sigma n^{-1/5}$, podría ser utilizado como valor crítico en un contraste de hipótesis donde **H_0:** $\phi(x)$ **unimodal** *versus* **H_1:** $\phi(x)$ **multimodal**. Esta aproximación sin embargo no ha ganado popularidad por su excesiva dependencia respecto a la normal y la falta de generalidad frente a otras distribuciones paramétricas unimodales.

Una segunda aproximación en la dirección del análisis no paramétrico es posible mediante los métodos *bootstrap* (Silverman (1981, 1983, 1986, Cap.- 6.4), Izenman y Sommer (1988), Efron y Tibshirani (1993), Bianchi (1995)). Recordemos que el problema fundamental radica en determinar cuando h_{crit} es relativamente grande, en cuyo caso rechazaremos **H_0:** $\phi(x)$ **unimodal**. La idea básica es construir una estimación de $\phi(x)$ a

partir del valor observado de h_{crit}, $\hat{\phi}_{h_{crit}}(x)$, esta densidad será por definición una estimación unimodal, y considerar dicha densidad como la verdadera bajo la hipótesis nula de unimodalidad. A partir de aquí muestreamos repetidamente con remplazo de $\hat{\phi}_{h_{crit}}(x)$ muestras de tamaño *n* un número suficientemente elevado de veces y evaluamos el nivel de significación empírico del valor observado de h_{crit}. Para ello simplemente determinamos la proporción de muestras generadas de $\hat{\phi}_{h_{crit}}(x)$ para las que su valor de suavizado crítico, digamos $\hat{h}_{crit}$, es mayor que el valor de suavizado crítico obtenido a partir de los datos originales, es decir la proporción de veces para las que $\hat{h}_{crit} > h_{crit}$. De esta forma seremos incapaces de rechazar la hipótesis nula de unimodalidad si el nivel de significación empírico del valor observado de h_{crit} es superior a los niveles de significación estándares. Además puesto que para una muestra generada a partir de $\hat{\phi}_{h_{crit}}(x)$ tendremos que $\hat{h}_{crit} > h_{crit}$ si y sólo si la densidad estimada para esa muestra generada es multimodal si utilizamos h_{crit} como parámetro de suavizado, en la práctica sólo es necesario determinar el porcentaje de muestras generadas para las cuales la densidad estimada utilizando h_{crit} como valor de *h* es multimodal, no siendo necesario calcular $\hat{h}_{crit}$ para cada una de las muestras generadas. El procedimiento puede generalizarse con facilidad al contraste del número exacto de modas en $\phi(x)$ (Bianchi (1995)).

Referencias bibliográficas

Abramson, I.S. (1982) "On bandwidth variation in kernel estimates - a square root law", *Annals of Statistics*, 10, 1217-1223.

Aitchison, J. & Brown, J.A.C. (1954) "On criteria for descriptions of income distribution", *Metroeconomica*, 6, 88-107.

Aitchison, J. & Brown, J.A.C. (1957) *The lognormal distribution*, Cambridge University Press, Cambridge.

Anderson, T.W. & Darling, D.A. (1954) "A test of goodness of fit", *Journal of the American Statistical Association*, 765-769.

Attanasio, O.P. & Weber, G. (1993) "Consumption growth, the interest rate and aggregation", *Review of Economic Studies*, 60, 631-649.

Atkinson, A.B. (1970) "On the measurement of inequality", *Journal of Economic Theory*, 3, 244-263.

Atkinson, A.B.; Rainwater, L. & Smeeding, T.M. (1995) *Income Distribution in OECD Countries: Evidence from Luxembourg Income Study*, OCDE, París.

Bartlett, M.S. (1963) "Statistical estimation of density functions", *Sankhya*, Series A, 25, 245-254.

Barro, R.J. (1991) "Economic growth in a cross section of countries", *Quarterly Journal of Economics*, 106, (May), 407-443.

Barro, R.J. & Sala-i-Martín, X. (1991) "Convergence across states and regions", *Brookings Papers on Economic Activity*, 1, (April), 107-182.

Barro, R.J. & Sala-i-Martín, X. (1992) "Convergence", *Journal of Political Economy*, 100, 2, 223-251.

Barro, R.J. & Sala-i-Martín, X. (1995) *Economic Growth*, McGraw Hill, New York.

Baumol, W.J. (1986) "Productivity growth, convergence, and welfare", *American Economic Review*, 76, 5, (December), 1072-1085.

Beach, C.M. & Kaliski, S.F. (1986) "Lorenz curve inference with sample weights: An aplication to the distribution of unemployment experience", *Applied Statistics*, 35, 1, 38-45.

Berrebi, Z.M. & Silber, J. (1987) "Regional differences and the components of growth and inequality change", *Economics Letters*, 25, 295-298.

Bianchi, M. (1995) "Testing for convergence: Evidence from nonparametric multimodality tests", Bank of England, Working Paper Series 36, (June).

Bishop, J.A.; Chakraborti, S. & Thistle, P.D. (1994) "Relative inequality, absolute inequality, and welfare: Large sample tests for partial orders", *Bulletin of Economic Research*, 46, 1, 41-59.

Bishop, J.; Formby, J.P. & Thistle, P. (1992) "Convergence of the south and the non-south income distributions, 1969-79", *American Economic Review*, 82, 262-272.

Boneva, L.I.; Kendall, D.G. & Stefanov, I. (1971) "Spline transformations: Three new diagnostic aids for the statistical data-analyst" (with discussion), *Journal of the Royal Statistical Society*, Series B, 33, 1-70.

Bosch, A.; Escribano, C. & Sánchez, I. (1989) *Evolución de la Desigualdad y la Pobreza en España. Estudio basado en las Encuestas de Presupuestos Familiares 1973-74 y 1980-81*. Instituto Nacional de Estadística (INE), Madrid.

Breiman, L.; Meisel, W. & Purcell, E. (1977) "Variable kernel estimates of multivariate densities", *Technometrics*, 19, 135-144.

Cass, D. (1965) "Optimum growth in an agregative model of capital accumulation", *Review of Economic Studies*, 32, (July), 233-240.

Chakravarty, S.R. (1990) *Ethical Social Index Numbers*, Springer Verlag, Berlin.

Chandra, M.; Singpurwalla, N.D. & Stephens, M.A. (1981) "Kolmogorov statistics for tests of fit for the extreme value and Weibull distribution", *Journal of the American Statistical Association*, 76, 375, 729-731.

Cleveland, W.S. (1993) *Visualizing Data*, Hobart Press.

Copas, J.B. & Fryer, M.J. (1980) "Density estimation and suicide risks in psychiatric treatment", *Journal of the Royal Statistical Society*, Series A, 143, 167-176.

Cosslett, S.R. (1993) "Estimation from endogenously stratified samples", in G.S. Maddala, C. R. Rao & Vinod, H. D. (Eds.) *Handbook of Statistics*, Volume 11, Amsterdam, North-Holland, 1-43.

Cowell, F. (1995) *Measuring Inequality*, 2nd Edition, LSE Handbooks in Economics, Prentice Hall, London. (1st. Edition 1977, Phillip Allan Publishers Limited, London).

Dalton, H. (1920) "The measurement of inequality of income", *Economic Journal*, 30, 348-361.

Davidson, R. & MacKinnon, J.G. (1993) *Estimation and Inference in Econometrics*, Oxford University Press, Oxford.

Deaton, A. & Muellbauer, J. (1980) *Economics and Consumer Behavior*, Cambridge University Press, Cambridge.

Decresin, J. & Fatás, A. (1995) "Regional labor market dynamics in Europe", *European Economic Review*, 39, 1627-1655.

DeLong, J.B. (1988) "Productivity growth, convergence, and welfare: A comment", *American Economic Review*, 78, 5, (December), 1138-1155.

DiNardo, J.; Fortin, N.M. & Lemieux, T. (1996) "Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach", *Econometrica*, 64, 5, (September), 1001-1044.

Doan, T.A. (1992) *RATS, User's Manual*, Version 4.0, Estima. Evanston, IL.

DuMouchel, W.H. & Duncan, G.J. (1983) "Using sample survey weights in multiple regression analysis of stratified samples", *Journal of the American Satistical Association*, 78, 535-543.

Duro, J.A. & Esteban, J. (1998) "Factor decomposition of cross-country income inequality, 1960-1990", *Economics Letters*, 60, 269-275.

Efron, B. & Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, Monographs on Statistics and Applied Probability nº 57, London.

Epanechnikov, V.A. (1969) "Nonparametric estimation of a multidimensional probability density", *Theory of Probability and its Applications*, 14, 153-158.

Esteban, J.M. (1994) "La desigualdad interregional en Europa y en España: Descripción y análisis", en Esteban, J.M. & Vives, X. (Eds.) *Crecimiento y Convergencia Regional en España y en Europa*, 2 volúmenes, Vol 2, Cap.-1, 13-84.

Esteban, J.M. (1996) "Desigualdad y polarización. Una aplicación a la distribución interprovincial de la renta en España", *Revista de Economía Aplicada*, 4, 11, (Otoño), 5-26.

Esteban, J.M. & Ray, D. (1993) "El concepto de polarización y su medición", en *Igualdad y Distribución de la Renta y la Riqueza*, vol.-2, Fundación Argentaria, Madrid, 1-35.

Esteban, J.M. & Ray, D. (1994) "On the measurement of polarization", *Econometrica*, 62, 819-852.

Evans, M.; Hastings, N. & Peacock, B. (1993) *Statistical Distributions*, 2nd. edition, John Wiley & Sons, New York.

Everitt, B.S. (1994) "Exploring multivariate data graphically: A brief review with examples", *Journal of applied Statistics*, 21, 3, 63-94.

Fisher, R.A. (1929) "Moments and product moments of sampling distributions", *Proccedings of London Mathematical Society*, 2, 30, 199.

Fisk, P.R. (1961) "The graduation of income distribution", *Econometrica*, 22, 171-185.

Fix, E. & Hodges, J.L. (1951) "Discriminatory analysis, nonparametric estimation: Consistency properties", *Report 4, Project nº 21-49-004*, USAF School of Aviation Medicine, Randolph Field, Texas.

Foster, J.E. & Ok, E.A. (1999) "Lorenz dominance and the variance of logarithms", *Econometrica*, 67, 4, (July), 901-907

Gardeazabal, J. (1996) "Provincial income distribution dynamics: Spain 1967-1991", *Investigaciones Económicas*, 20, 263-269.

Galton, F. (1883) *Inquiries into Human Faculty and Its Development*, MacMillan, London.

Galton, F. (1885) "Regression towards mediocrity in hereditary stature", *Journal of the Anthropological Institute of Great Britain and Ireland*, 14, 246-263.

Goerlich, F.J. (1998) "Desigualdad, diversidad y convergencia: (Algunos) instrumentos de medida", *Monografía*, Instituto Valenciano de Investigaciones Económicas, (Diciembre).

Goerlich, F.J. (2000) "Desigualdad, diversidad y convergencia: (Más) instrumentos de medida -Modelos de regresión-", *Monografía* en elaboración, Instituto Valenciano de Investigaciones Económicas.

Goerlich, F.J. & Mas, M. (1998a) "Medición de la desigualdad: Variables, indicadores y resultados", *Moneda y Crédito*, 207, (Noviembre), 59-86.

Goerlich, F.J. & Mas, M. (1998b) "Inequality and convergence in the OECD area", Working Paper WP-EC 98-09, Instituto Valenciano de Investigaciones Económicas.

Goerlich, F.J. & Mas, M. (1998c) "Japan/USA: The (apparent) miracle of convergence", Working Paper WP-EC 98-20, Instituto Valenciano de Investigaciones Económicas.

Goerlich, F.J. & Mas, M. (1999) "Medición de la desigualdad: Contribución a una base de datos regional", Monografía, Instituto Valenciano de Investigaciones Económicas.

Hamilton, J.D. (1994) *Time Series Analysis*, Princeton University Press, Princeton, New Jersey.

Hart, P.E. (1995) "Galtonian regression across countries and the convergence of productivity", *Oxford Bulletin of Economics and Statistics*, 57, 3, (August), 287-293.

Hastings, N.A. J. & Peacock, J.B. (1974) *Statistical Distributions*, Butterworth, London.

Hausman, J.A. & Wise, D.A. (1981) "Stratification on an endogenous variable and estimation: The Gary income maintenance experiment", in C. F. Manski & D. McFadden (Eds.) *Structural Analysis of Discrete Data with Econometric Applications*, Cambridge, MA., MIT Press, 365-391.

Hodges, J.L. & Lehmann, E.L. (1956) "The efficiency of some nonparametric competitors of the *t*-test", *Annals of Mathematical Statistics*, 27, 324-335.

Hotelling, H. & Solomons, L.M. (1932) "The limits of a measure of skewness", *Annals of Mathematical Statistics*, 3, 141.

Hsiao, C. (1986) *Analysis of Panel Data*, Cambridge University Press, Cambridge.

Imbens, G.Y. & Lancaster, T. (1996) "Efficient estimation and stratified sampling", *Journal of Econometrics*, 74, 289-318.

INE (varios años) *Anuario Estadístico de España.* Instituto Nacional de Estadística, Madrid.

Izenman, A.J. & Sommer, C.J. (1988) "Philatelic mixtures and multimodal densities", *Journal of the American Statistical Association*, 83, 404, 941-953.

Jarque, C.M. & Bera, A.K. (1980) "Efficient tests of normality, homoscedasticity and serial independence of regression residuals", *Economic Letters*, 6, 255-259.

Jewell, N.P. (1985) "Least squares regression with data arising from stratified samples of the dependent variable", *Biometrika*, 72, 11-21.

Johnson, N.L & Kotz, S. (1970) *Distributions in Statistics: Continuous Univariate Distributions*, 2 vols., John Wiley & Sons, New York.

Jones, C.I. (1997) "On the evolution of the World income distribution", *Journal of Economic Perspectives*, 11, 3, (Summer), 19-36.

Jones, M.C. (1983) *The proyection pursuit algorithm for exploratory data analysis*, Ph. D. Thesis, University of Bath.

Jones, M.C.; Marron, J.S. & Sheather, S.J. (1996) "A brief survey of bandwidth selection for density estimation", *Journal of the American Statistical Association*, 91, 433, (March), 401-407.

Joiner, & Rosenblatt, M. (1971) "Some properties of the range in samples from Tukey's symmetric lambda distributions", *Journal of Amerian Statistical Association*, 66, 394-399.

Kakwani, N. (1993) "The coefficient of determination for a regression model based on group data", *Oxford Bulletin of Economics and Statistics*, 55, 2, (May), 245-251.

Kendall, M.G. & Stuart, A. (1977) *The Advanced Theory of Statistics. Volume 1: Distribution Theory*. 4th. Ed. Griffin, London.

Klein, L.R. & Morgan, J.N. (1951) "Results of alternative statistical treatments of sample survey data", *Journal of Amerian Statistical Association*, 46, 442-460.

Koopmans, T.C. (1965) "On the concept of optimal economic growth", in *The Econometric Approach to Development Planning*, Amsterdam, North Holland.

Korn, E.L. & Graubard, B.I. (1995a) "Analysis of large health surveys: Accounting for sample desing", *Journal of the Royal Statistical Society*, Series A, 158, 263-295.

Korn, E.L. & Graubard, B.I. (1995b) "Examples of differing weighted and unweighted estimates from a sample survey", *The American Statistician*, 49, 291-295.

Kott, P.S. (1991) "A model-based look at linear regression with survey data", *The American Statistician*, 45, 107-112.

Lorenz, M.C. (1905) "Methods of measuring the concentration of wealth", *Publications of the American Statistical Association*, 9, 209-219.

Magee, L.; Robb, A.L. & Burbidge, J.B. (1998) "On the use of sampling weights when estimating regression models with survey data", *Journal of Econometrics*, 84, 251-271.

Magnus, J.R. & Neudecker, H. (1988) *Matrix Differential Calculus. With Applications in Statistics and Econometrics*, John Wiley & Sons Ltd, New York.

Marron, J.S. (1989) "Automatic smoothing parameter selection: A survey", *Empirical Economics*, 13, 187-208.

Martín-Guzmán, P.; Toledo, M.I.; Bellido, N.; López, J. & Jano, N. (1996) *Encuesta de Presupuestos Familiares. Desigualdad y Pobreza en España. Estudio basado en las Encuestas de Presupuesto Familiares de 1973-74, 1980-81 y 1990-91*, Instituto Nacional de Estadística y Universidad Autónoma de Madrid.

McDonald, J.B. & Jensen, B. (1979) "An analysis of some properties of alternative measures of income inequality based on the Gamma distribution function", *Journal of the American Statistical Association*, 74, 856-860.

McGill, R.; Tukey, J.W. & Larsen, W.A. (1978) "Variations of Box Plots", *American Statistician*, 32, 12-16.

Mills, T.C. (1990) *Time Series Techniques for Economists*. Cambridge University Press, Cambridge.

Mood, A.M.; Graybill, F.A. & Boes, D.C. (1974) *Introduction to the Theory of Statistics*, 3rd. Edition, International Student Edition, McGraw-Hill Book Company, London.

Müller, H. G. (1984) "Smooth optimum kernel estimators of densities, regression curves and modes", *Annals of Statistics*, 12, 766-774.

Nathan, G. & Holt, D. (1980) "The effect of survey design on regression analysis", *Journal of the Royal Statistical Society*, Series B, 42, 377-386.

Nelson, C.R. (1973) *Applied Time Series Analysis for Managerial Forecasting*, Holden-Day, Inc., San Francisco.

Ord, J.K. (1968) "The discrete Student's *t* distribution", *Annals of Mathematical Statististics*, 39, 1513.

Palisade Corporation (1997) *BESTFIT. Probability Distribution Fitting for Windows*. User's Guide. (June), New York.

Pareto, V. (1965) *Écrits sur La Courbe de la Repartition de la Richesse*, G. Busino (Ed.) Vol 3, *Oeuvres Complètes*, Libraire Droz, Geneva.

Pareto, V. (1972) *Manual of Political Economy*, A. S. Schwier & A. N. Page (Eds.), MacMillan, London.

Park, B.U. & Marron, J.S. (1990) "Comparison of data-driven bandwidth selectors", *Journal of the American Statistical Association*, 85, 409, (March), 66-72.

Patel, J.K. & Read, C.B. (1982) *Handbook of the Normal Distribution*, Statistics: Textbooks and monographs, Volume 40, Marcel Dekker, Inc., New York and Basel.

Parzen, E. (1962) "On the estimation of a probability density function and mode", *Annals of Mathematical Statistics*, 33, 1065-1076.

Pearson, K. (1895) "Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material", *Philosophical Transactions of the Royal Society of London*, series A, 186, 343-414.

Pearson, K. (1906) "Skew frequency curves, a rejoinder to Professor Kapteyn", *Biometrika*, 5, 168-171.

Pfefferman, D. (1993) "The role of sampling weights when modeling survey data", *International Statistical Review*, 61, 317-337.

Pigou, A.C. (1912) *The Economic of Welfare*, London. (Editado por MacMillan, New York en 1952).

Quah, D. (1990) "International patterns of growth: I. Persistence in cross-country disparities", Mimeo. Economics Department, MIT. Cambridge, MA.

Quah, D. (1993a) "Galton's fallacy and test of the convergence hypothesis", *The Sandinavian Journal of Economics*, 95, 4, (December), 427-443.

Quah, D. (1993b) "Empirical cross-section dynamics in economic growth", *European Economic Review*, 37, 2/3, (April), 426-434.

Quah, D. (1996a) "Twin peaks: Growth and convergence in models of distribution dinamics", *Economic Journal*, 106, 437, (July), 1045-1055.

Quah, D. (1996b) "Ideas determining convergence clubs", Working Paper, Economics Department, LSE. (April).

Quah, D. (1996c) "Regional cohesion from local isolated actions: I. Historical outcomes." Working Paper, Economics Department, LSE. (December).

Quah, D. (1996d) "Regional convergence clusters across Europe", *European Economic Review*, 40, 3/5, (April), 951-958.

Quah, D. (1996e) "Empirics for economic growth and convergence", *European Economic Review*, 40, 1353-1375.

Quah, D. (1997) "Empirics for growth and distribution: Stratification, polarization, and convergence clubs", *Journal of Economic Growth*, 2, (March), 27-59.

Rabadan, I. & Salas, R. (1996) "Convergencia y redistribución intertemporal en España: Efecto de los impuestos directos, cotizaciones sociales y transferencias", *Economía Pública*, (Septiembre), Fundación BBV.

Ramsey, F.P. (1928) "A mathematical theory of saving", *Economic Journal*, 38, (December), 543-559.

Rosenblatt, M. (1956) "Remarks on some nonparametric estimates of a density function", *Annals of Mathematical Statistics*, 27, 832-837.

Sala-i-Martín, X. (1994) "Cross-sectional regressions and the empirics of economic growth", *European Economic Review*, 38, 739-747.

Salem, A.B.Z. & Mount, T.D. (1974) "A convenient descriptive model of income distribution: The gamma density", *Econometrica*, 42, 1115-1127.

Scott, D.W. (1979) "On optimal and data-based histograms", *Biometrika*, 66, 605-610.

Scott, D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*, John Wiley & Sons, New York.

Selden, T.M. (1994) "Weighted generalized least squares estimation for complex survey data", *Economics Letters*, 46, 1-6.

Sen, A. (1973) *On Economic Inequality*, Oxford University Press, Oxford.

Sheather, S.J. & Jones, M. C. (1991) "A reliable data-based bandwidth selection method for kernel density estimation", *Journal of the Royal Statistical Society*, Series B, 53, 683-690.

Shorrocks, A.F. (1980) "The class of additively descomposable inequality measures", *Econometrica*, 48, 613-625.

Shorrocks, A.F. (1982) "Inequality decomposition by factor components", *Econometrica*, 50, 193-211.

Shorrocks, A.F. (1984) "Inequality decomposition by population subgroups", *Econometrica*, 52, 1369-1386.

Silverman, B.W. (1981) "Using kernel density estimates to investigate multimodality", *Journal of the Royal Statistical Society*, Series B, 43, 97-99.

Silverman, B.W. (1983) "Some properties of a test for multimodality based on kernel density estimates", in J. F. C. Kingmand & G. E. H. Reuter (Eds.) *Probability, Statistics, and Analysis*, Cambridge University Press, 248-260.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, Monographs on Statistics and Applied Probability nº 26, London.

Simonoff, J.S. (1996) *Smoothing Methods in Statistics*, Springer-Verlag, Berlin

Slottje, D.J. (1984) "A measure of income inequality based upon the beta distribution of the second kind", *Economics Letters*, 15, 369-375.

Solow, R. (1956) "A contribution to the theory of economic growth", *Quarterly Journal of Economics*, 70, 1, (February), 65-94.

Spanos, A. (1986) *Statistical Foundations of Econometric Modeling*, Cambridge University Press, Cambridge.

Spanos, A. (1999) *Probability Theory and Statistical Inference. Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge.

Stephens, M.A. (1974) "EDF statistics for goodness of fit and some comparisons", *Journal of the American Statistical Association*, 69, 347, 730-737.

Stephens, M.A. (1977) "Goodness of fit for the extreme value distribution", *Biometrika*, 64, 3, 583-588.

Stigler, S.M. (1974) "Linear functions of order statistics with smooth weight functions", *The Annals of Statistics*, 2, 4, 676-693.

"Student" (1908a) "On the probable error of a mean", *Biometrika*, 6, 1.

"Student" (1908b) "On the probable error of a correlation coefficient", *Biometrika*, 6, 302.

Swan, T.W. (1956) "Economic growth and capital accumulation", *Economic Record*, 32, (November), 334-361.

Titterington, D.M.; Smith, A.F.M. & Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*, John Wiley & Sons, New York.

Tortosa-Ausina, E. (1999) "Especialización productiva, eficiencia y convergencia de las empresas bancarias españolas", Tesis Doctoral, Facultad de Ciències Econòmiques y Jurídiques, Departament d'Economía, Universitat Jaume I.

Theil, H. (1967) *Economics and Information Theory*, Amsterdam, North-Holland.

Theil, H. & Sorooshian, C. (1979) "Components of the change in regional inequality", *Economics Letters*, 4, 191-193.

Thurow, L.C. (1970) " Analyzing the American income distribution", *American Economic Review*, *Papers and Proceedings*, 60, 261-269.

Tukey, J.W. (1977) *Exploratory Data Analysis*, Reading, Massachusets, Addison-Wesley.

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

Villar, A. (sin fecha) "Medición de la desigualdad: Análisis de los índices de desigualdad económica", Mimeo, Universidad de Alicante.

Wand, M.P. & Jones, M.C. (1994) *Kernel Smoothing*, Chapman & Hall, London.

Weitzman, M.L. (1992) "On diversity", *The Quarterly Journal of Economics*, 107, 2, (May), 363-405.

Whittle, P. (1958) "On the smoothing of probability density functions", *Journal of the Royal Statistical Society*, Series B, 20, 334-343.

Wooldridge, J.M. (1999) "Asymptotic properties of weighted M-Estimators for variable probability samples", *Econometrica*, 67, 1385-1406.