Índice

Capítulo 1

Introducción a los Métodos Robustos

1.1. Introducción

Aunque puede decirse que la Estadística tuvo su origen en los censos romanos[1], sus métodos, tal y como los conocemos hoy en día, se deben fundamentalmente a Sir Ronald Fisher, quien en su trabajo de 1922 (*Sobre los fundamentos matemáticos de la estadística teórica*), estableció los principios a partir de los cuales se fueron desarrollando las técnicas y métodos que actualmente utilizamos y a los que ya denominamos *Métodos Clásicos*.

No obstante, su correcta aplicación requiere de condiciones muy rígidas, tales como un modelo probabilístico fijo (habitualmente la distribución normal) en el que sólo queden indeterminados uno o dos parámetros (su media y/o su varianza). Pero tal restricción es un problema, ya que los modelos probabilísticos habitualmente utilizados rara vez se ajustan bien al fenómeno aleatorio observado, razón por la cual, los resultados obtenidos bajo tales supuestos dejan de ser válidos, incluso en situaciones muy cercanas a la modelizada bajo la cual se obtuvieron.

Por estas razones surgieron los denominados *Métodos Robustos* que aquí estudiaremos. Su origen también es remoto. Rey (1978) lo sitúa en la antigua Grecia, en donde los sitiadores contaban las capas de ladrillos de algunos muros de la ciudad sitiada y tomaban la moda de los recuentos con objeto de determinar la longitud de las escalas a utilizar en el asalto; de esta forma, la estimación realizada no se veía afectada por valores extremos procedentes de muros muy altos o muy bajos y que hubieran conducido a escalas demasiado

[1] entre ellos, el más famoso, el ordenado por César Augusto y que obligó a trasladarse a José y María de Nazaret a Belén donde nació Jesús

largas o demasiado cortas. No obstante, fue en 1964 cuando, de la misma manera que los trabajos de R.A. Fisher dotaron a la estadística del rigor matemático del que hasta entonces carecía, el artículo de Peter Huber, *Estimación robusta de un parámetro de localización*, abrió las puertas del rigor matemático en robustez y, por ende, las del reconocimiento científico.

Posteriores trabajos suyos, así como las aportaciones fundamentales de Frank Hampel en los años 1971 y 1974, en donde definió la *robustez cualitativa* y la *curva de influencia*, terminaron de poner los cimientos de los Métodos Robustos tal y como son conocidos hoy en día.

1.2. Estimación de la media de una población

Con objeto de mostrar que, incluso en las situaciones más simples y que creíamos tener perfectamente claras, surgen serios problemas de sensibilidad —falta de robustez— de los estimadores habitualmente utilizados, consideremos el problema de la estimación de la media μ de una población, supuesto que contamos con diez observaciones de dicha población, dadas por la tabla 1.1.

Datos									
790	750	910	650	990	630	2510	820	860	710

Tabla 1.1

La idea de utilizar como estimador de μ la *media muestral*, definida como la media aritmética de las observaciones,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad [1.1]$$

hoy en día de uso tan extendido, se remonta a finales del siglo diecisiete (véase Plackett, 1958). Pero además de por ser el estimador *clásico* para el problema considerado, la razón fundamental de su extensa utilización es la de que $\overline{x}$ es el valor de a para el cual la suma de los cuadrados de las diferencias $\sum_{i=1}^{n}(x_i - a)^2$ es mínima; es decir, que $\overline{x}$ es el estimador de *mínimos cuadrados* de μ.

A pesar de esta propiedad de óptimo —en realidad a causa de ella—, $\overline{x}$ es un estimador sumamente sensible a valores extremos. Estos diez valores tienen como representación gráfica la figura 1.1.

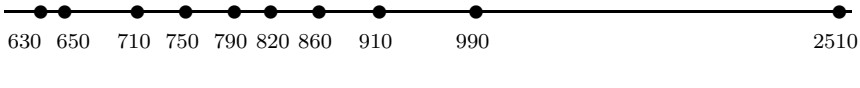

Figura 1.1

Como se ve, el valor 2510 está muy alejado del resto de los datos. Esto hace que la media muestral quede desplazada a la derecha, a causa del peso de este dato en [1.1], siendo $\overline{x} = 962$, (figura 1.2).

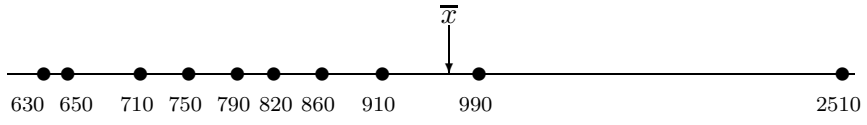

Figura 1.2

Es decir, la media muestral es un estimador sobre el que tienen gran *influencia* los datos extremos. *Cuanto más extremo sea el dato, mayor será su influencia sobre* $\overline{x}$.

Por la misma razón, la *cuasivarianza muestral*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2$$

o su análogo la *varianza muestral* $s^2 = (n-1)S^2/n$, también son muy sensibles a la presencia de datos extremos.

Un estimador sobre el que, por construcción, tendrán menos influencia los valores extremos será la *Mediana muestral* M_e definida (CB-sección 2.3.2) como el valor central de los datos, es decir, como aquel valor tal que, supuestos ordenados todos los datos en orden creciente, la mitad sean menores que M_e y la otra mitad sean mayores; en el caso de un número par de observaciones se define como el promedio de los dos valores centrales. Para los datos de nuestro ejemplo toma el valor $M_e = 805$, (figura 1.3).

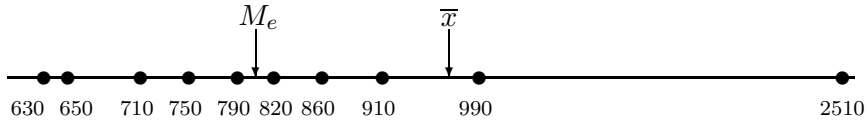

Figura 1.3

Claramente, M_e no variará si desplazamos aún más a la derecha el valor 2510; es decir, M_e parece *menos sensible* que $\overline{x}$ a observaciones extremas.

Este análisis realizado para el parámetro de localización también podía haberse seguido para el de escala. De esta forma, al igual que la varianza es la suma (normalizada) de los cuadrados de las desviaciones a la media (valor para el que se hacía mínima dicha suma de cuadrados), un estimador del

parámetro de escala, el cual veremos en el capítulo próximo que es *robusto*, es la *desviación absoluta mediana*, *MAD*, la cual se define (por analogía con la varianza) como la mediana de las desviaciones absolutas a la mediana,

$$MAD = Mediana\left\{|X_i - M_e|\right\}_{i=1}^{n}.$$

Por último, observemos que si un dato puede llegar a influir considerablemente en el valor tomado por un estimador, imaginemos el efecto que puede producir en un test de hipótesis. En un estimador puede tener un *efecto continuo*, es decir, a medida que el dato va influyendo, el estimador se va desplazando, pero en un test de hipótesis en donde sólo hay dos decisiones posibles —aceptar o rechazar la hipótesis nula—, la presencia de un dato extremo puede hacernos tomar la decisión errónea.

1.3. Métodos robustos y datos anómalos

La introducción de Métodos Robustos en estadística fue motivada, fundamentalmente aunque no de forma exclusiva, por la gran sensibilidad a los *datos anómalos* (*outliers* en la terminología anglosajona) de los estimadores habitualmente utilizados.

No obstante, a pesar de la relación existente entre el análisis de outliers y los Métodos Robustos, la cual veremos en esta sección, ambos campos han seguido desarrollos y caminos independientes.

1.3.1. Opciones ante la presencia de datos anómalos

Una de las primeras ideas que sugiere la presencia de datos anómalos en una muestra, entendidos éstos como datos sorprendentemente alejados del grupo principal de observaciones, es la de su *rechazo* o eliminación de la muestra, con objeto de *repararla* o *limpiarla*, antes de realizar inferencias con ella.

Esta idea se encuentra en numerosas publicaciones sobre el tema. Así por ejemplo, puede leerse en el trabajo de Ferguson (1961) "... el problema que se plantea en el tratamiento de los datos anómalos es el de introducir algún grado de objetividad en su rechazo...", dando por supuesto que los datos anómalos son necesariamente *erróneos* y que, por tanto, deben de ser eliminados.

No obstante, su eliminación no es más que una de las posibles opciones a considerar en el tratamiento de los datos anómalos. Y ello es porque no siempre los datos anómalos son datos erróneos, sino que, en ocasiones, son la indicación de algún inesperado y útil tratamiento industrial, o de la existencia de alguna nueva variedad agrícola, etc., siendo el objetivo, en estos casos, la *identificación*, para un posterior estudio más detallado, de tales datos anómalos.

Algo así ocurrió en el invierno de 1893-94, según recoge Tukey (1977), cuando Lord Rayleigh estaba investigando la densidad del nitrógeno procedente de varias fuentes. Con tal propósito realizó varios experimentos obteniendo los resultados de la figura 1.4, los cuales, proporcionaron el diagrama de hojas y ramas (véase CB-sección 14.2) representado en la figura 1.5, en cuya parte inferior se observa la presencia de valores anómalos.

	Fuente de Nitrógeno		
Fecha	*Procedencia*	*Agente purificador*	*Peso*
29-11-93	NO	Acero caliente	2′30143
5-12-93	NO	Acero caliente	2′29816
6-12-93	NO	Acero caliente	2′30182
8-12-93	NO	Acero caliente	2′29890
12-12-93	*Aire*	Acero caliente	2′31017
14-12-93	*Aire*	Acero caliente	2′30986
19-12-93	*Aire*	Acero caliente	2′31010
22-12-93	*Aire*	Acero caliente	2′31001
26-12-93	N_2O	Acero caliente	2′29889
28-12-93	N_2O	Acero caliente	2′29940
9- 1-94	NH_4NO_2	Acero caliente	2′29849
13- 1-94	NH_4NO_2	Acero caliente	2′29889
27- 1-94	*Aire*	Hidrato Ferroso	2′31024
30- 1-94	*Aire*	Hidrato Ferroso	2′31030
1- 2-94	*Aire*	Hidrato Ferroso	2′31028

Figura 1.4

RAMAS	HOJAS	
2298	25999	(Procedencia: NO, NH_4NO_2, NO, N_2O, NH_4NO_2)
2299	4	(Procedencia: N_2O)
2300		
2301	48	(Procedencia: NO, NO)
2302		
2303		
2304		
2305		
2306		
2307		
2308		
2309	9	(Procedencia: *Aire*)
2310	012233	(Procedencia: *Aire*, todos)

Figura 1.5

El estudio más detallado de la procedencia de éstos permitió a Lord Rayleigh observar que los pesos del nitrógeno obtenido a partir del aire son mayores.

Esto le hizo investigar más a fondo la composición del aire químicamente libre de oxígeno, descubriendo allí un nuevo elemento: el argón.

Barnett y Lewis (1994) recogen una serie de curiosos ejemplos judiciales en los que el propósito es, precisamente, la identificación de outliers.

Así, un ejemplo de como *dato anómalo* no es sinónimo de *dato a rechazar* lo encontramos en un proceso judicial inglés de 1949: el caso de *Hadlum* contra *Hadlum*. Mr Hadlum apeló contra el fallo de su anterior petición de divorcio alegando adulterio de Mrs Hadlum. La evidencia de tal adulterio —alegaba Mr Hadlum— era que el 12 de Agosto de 1945 Mrs Haldum había tenido un bebé, 349 días después de que Mr Hadlum abandonara el país para combatir en la segunda guerra mundial. Como el tiempo medio de gestación humana es de 280 días, Mr Hadlum consideró que 349 días estaba demasiado alejado del tiempo medio de gestación y que *algo debía haber ocurrido*. Es decir, pensó que 349 era un dato anómalo. Pero él no quería rechazarlo. Todo lo contrario, quería *identificarlo* para obtener las subsiguientes consecuencias.

Como anécdota diremos que Mr Hadlum no tuvo éxito en su petición de divorcio. El juez consideró que aunque 349 días era un valor poco probable, no se encontraba más allá de los límites considerados como posibles por la ciencia.

No obstante, ese límite no siempre fue el mismo. En 1949, en el caso de M.T. contra M.T., la corte suprema de Inglaterra consideró que 340 días era imposible a la luz de la *moderna experiencia ginecológica*, concediendo el divorcio. Por contra, en 1951 el parlamento inglés consideró, en el caso de *Preston-Jones* contra *Preston-Jones* que el límite debía estar en 360 días. En 1921, Mr Gaskill no consiguió el divorcio basado en ausencia de 331 días del domicilio conyugal. (No obstante su mujer, Mrs Gaskill, lo consiguió en 1960 basándose en una ausencia de su marido de 39 años.)

Existe además un tercer aspecto a considerar en el tratamiento de los datos anómalos: la distribución modelo supuesta para la variable aleatoria en estudio.

Consideremos la siguiente muestra aleatoria

$$0'39\,,\ 1'40\,,\ -1'62\,,\ -0'05\,,\ -0'50\,,\ 0'02\,,\ 4'01\,,\ 1'23\,,\ -0'10\,,\ 1'69$$

extraída de una población de varianza 1, con objeto de contrastar la hipótesis nula, referente a su centro de simetría θ, $H_0 : \theta = 0$ frente a la alternativa $H_1 : \theta \neq 0$.

Si se quiere utilizar un test paramétrico —el de la t de Student—, deberemos admitir para la población una distribución normal —$N(0, 1)$ bajo la hipótesis nula— al ser el tamaño muestral pequeño (véase CB-sección 5.4).

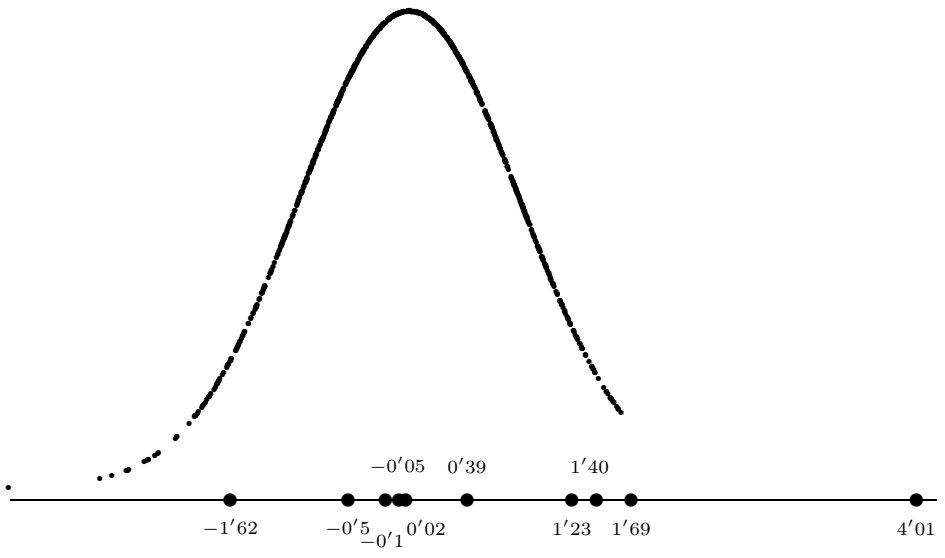

Figura 1.6

No obstante, en tal caso (figura 1.6), el dato $4'01$ debe ser considerado como un dato anómalo puesto que su probabilidad de obtención bajo la ley $N(0,1)$ (sobreimpresionada en dicha figura 1.6) parece muy pequeña. Sin embargo, su eliminación de la muestra sería un error, porque, en realidad, dicha muestra fue generada a partir de una distribución logística $L(0\,,\,\sqrt{\pi}/(2\sqrt{2}))$, que es una distribución con el mismo centro de simetría —cero— pero con colas más pesadas que la $N(0,1)$ (ver figura 1.7 en donde la distribución $N(0,1)$ es la de la línea de puntos y la $L(0\,,\,\sqrt{\pi}/(2\sqrt{2}))$ la de trazo continuo) y que, por tanto, da más probabilidad de ocurrencia a los casos extremos. Así pues, lo que deberíamos hacer en este caso es cambiar el modelo supuesto para la variable aleatoria en estudio; no eliminar el mencionado dato *aparentemente* anómalo.

Es decir, los datos pueden ser o parecer anómalos *en relación con el modelo supuesto*, por lo que una posible alternativa a su rechazo es la de su *incorporación*, *ampliando* el modelo, es decir, cambiarlo por uno con colas más pesadas.

Después de considerar este último aspecto, podemos dar ya una definición más precisa de lo que entenderemos por *dato anómalo* —o *outlier* en la terminología anglosajona.

Definición 1.1

Denominaremos *dato anómalo* a aquella observación que parece ser inconsistente con el resto de los valores muestrales, en relación con el modelo supuesto.

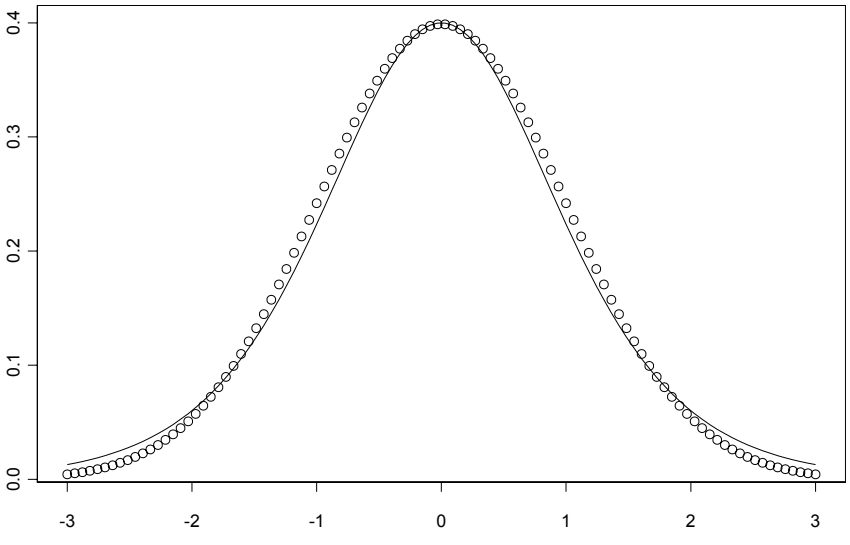

Figura 1.7

Tanto en los ejemplos judiciales antes considerados como en la definición 1.1, aparece una cierta componente subjetiva en la calificación de un dato como anómalo.

Existe una manera, ciertamente más objetiva, de poder llegar a tal conclusión.

Se trata de utilizar unos tests de hipótesis, denominados *tests de discordancia*, los cuales están basados en unos estadísticos para los que es posible determinar, o al menos tabular, su distribución en el muestreo.

Mediante dichos tests podemos calificar a uno o varios datos como *discordantes* —valores que resultan significativos en un test de discordancia—, y como consecuencia podemos, como hemos visto,

- *Rechazarlos*, eliminándolos del resto de la muestra.
- *Identificarlos*, resaltando algún aspecto que pudiera resultar interesante.
- *Incorporarlos*, ampliando la distribución modelo supuesta.

A pesar del esfuerzo por conseguir una calificación objetiva de los datos, el carácter subjetivo permanece, en cierta medida, en los tests de discordancia, tanto en su nivel de significación, como en la propia elección del contraste a considerar.

Además, como todo test de hipótesis, los tests de discordancia no son simétricos; es decir, no son tratadas de igual manera la hipótesis nula de ausencia de outliers en la muestra que la alternativa de, digamos, tres outliers a la

derecha. Y una vez concluido el test, deberían considerarse los dos tipos de error asociados al test.

Pero lo peor de considerar tests de discordancia, rechazando los outliers y luego utilizando métodos clásicos, es la pérdida de eficiencia con respecto a la utilización de Métodos Robustos como veremos en la siguiente sección.

Otro problema adicional relacionado con el tratamiento de outliers es que éstos no sólo se presentan en situaciones simples, sino que también aparecen en situaciones más estructuradas: modelos lineales, series cronológicas, datos circulares, etc.

En estas situaciones, los datos anómalos tenderán a ser menos aparentes, siendo en ocasiones la discrepancia con el modelo propuesto lo que hará anómalo a un dato.

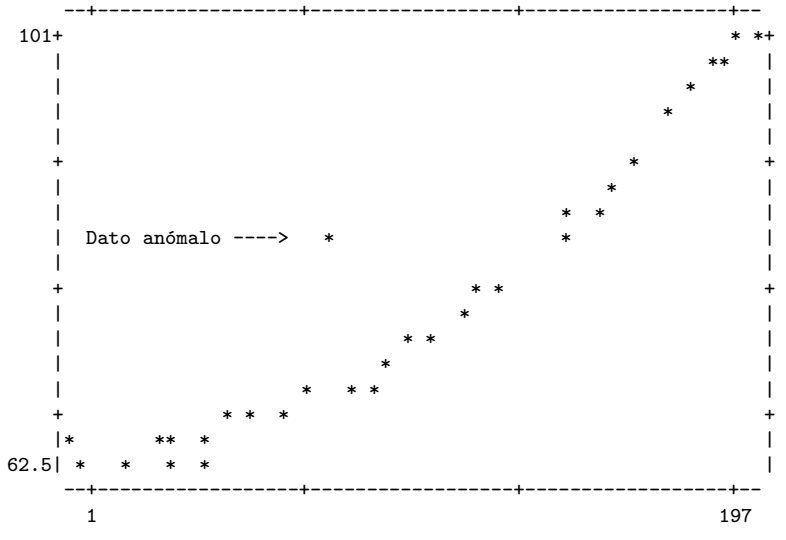

Figura 1.8

Así por ejemplo, en el caso de una regresión lineal simple, el ser anómalo un dato consiste en no estar alineado con el resto de las observaciones. Ahora, el ser anómalo no consiste en ser extremo; puede encontrarse en el grupo principal de observaciones y ser tratado como anómalo (figura 1.8).

Por tanto, el término *modelo*, en la definición de outlier dada más arriba, debe entenderse en un sentido amplio.

Como resumen, digamos que mientras los tests de discordancia tienen como objetivo el estudio de los outliers en sí mismos, proponiendo como acción ante un outlier alguno de los tres puntos antes reseñados, los Métodos Robustos están diseñados para realizar inferencias sobre el modelo, reduciendo la posible influencia que pudiera tener la presencia de datos anómalos. De hecho, los Métodos Robustos son denominados, en ocasiones, *Técnicas de acomodación de outliers*.

Es decir, en los tests de discordancia los outliers son el objetivo, mientras que en los Métodos Robustos los outliers son el mal a evitar.

En el presente texto se estudian los Métodos Robustos y no las técnicas para el tratamiento de outliers; un excelente texto para el estudio de estas técnicas es el de Barnett y Lewis (1994).

1.4. Distribuciones contaminadas

De la sección anterior se puede deducir una forma aparentemente razonable de proceder: Primero se realiza un *test de discordancia* mediante el cual se detectan los outliers y, después de eliminados éstos, se utilizan los métodos habituales de la Inferencia. Al final de esta sección veremos que este procedimiento no es el más adecuado.

Consideremos los datos representados en las figuras 1.1 ó 1.6. Se les puede suponer como procedentes de una distribución, digamos F, salvo uno de ellos, supuestamente procedente de una distribución desplazada a la derecha.

Por tanto, podíamos modificar el modelo de forma que con una determinada probabilidad conocida (generalmente grande), $1-\epsilon$, un dato venga de F y con una determinada probabilidad, (generalmente pequeña), ϵ, el dato venga de una distribución alternativa —*contaminante*— G. (En los ejemplos mencionados puede ser la misma que F pero desplazada a la derecha.)

Es decir, una forma razonable de prevenirse contra outliers (o de incorporarlos), es admitir que el modelo del que proceden los datos es de la forma $(1-\epsilon)F + \epsilon G$.

Este planteamiento, propuesto por Tukey (1960), consiste en suponer como alternativa a la distribución modelo $X_i \equiv F \quad , \; i = 1, ..., n$ el denominado *modelo de contaminación* $X_i \equiv (1-\epsilon)F + \epsilon G \quad , \; i = 1, ..., n$, con objeto de explicar o incorporar los outliers.

Es decir, suponer que en lugar de proceder los datos de una distribución F provienen de una *mixtura* (mezcla) de distribuciones, $(1-\epsilon)F + \epsilon G$, denominada *distribución contaminada*, de forma que ahora se pueda explicar la presencia de outliers en la muestra: Son los datos procedentes de G. El resto de las observaciones (la mayoría si ϵ es pequeño) procederá de F.

Tukey desarrolló su trabajo considerando que F era una distribución normal y G otra normal, con la misma media que F pero con mayor varianza.

En este modelo, ϵ representa la *proporción de contaminación* de la muestra y los datos procedentes de G, las *observaciones contaminadas*.

Admitiendo un modelo contaminado como generador de los datos, el número de observaciones contaminadas de una muestra de tamaño n será una variable aleatoria, R, con distribución binomial $B(n, \epsilon)$, por lo que (véase CB-sección 4.4.1), en una muestra de tamaño n, el número esperado de datos

contaminados (procedentes de G) será $n\epsilon$ y la proporción esperada de datos contaminados, ϵ.

Aunque, como antes dijimos, en su origen los modelos contaminados estaban compuestos por mixturas de dos normales, éstos han sido generalizados. G no tiene por qué ser necesariamente normal aunque F lo sea; ni siquiera tiene por qué ser simétrica.

En muchas ocasiones se tomará como distribución contaminante la distribución δ_x degenerada en un punto x, la cual es la distribución discreta (véase CB-sección 4.4) que toma con probabilidad 1 el valor x. Su función de distribución es una escalera pero que sólo da un salto (igual a 1) en el punto x; es decir, $\delta_x(y)$ es la función de y siguiente:

$$\delta_x(y) = \begin{cases} 0 & \text{si } \ y < x \\ \\ 1 & \text{si } \ y \geq x. \end{cases}$$

Utilizando esta distribución contaminante se considerará la distribución contaminada $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$ igual a

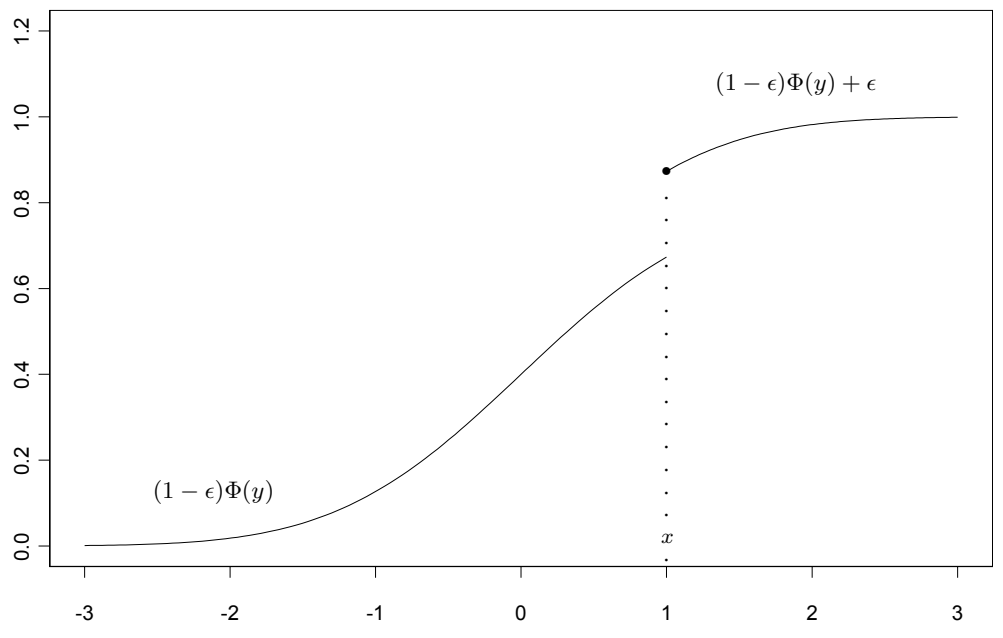

Figura 1.9