Tema 6 . Índices estadísticos de variables cuantitativas. Parámetros de tendencia central, dispersión, posición y forma.

Los parámetros o índices (ya vimos en el tema 3 que consideramos ambos conceptos como equivalentes) son otra forma de presentar resumidos los datos estadísticos.
Hay que distinguir:

- parámetros de tendencia central, que informan del centro de la distribución
- parámetros de dispersión, que informan de la dispersión de los datos
- parámetros de posición, que sitúan a los datos en el conjunto ce la distribución ordenada. Los más utilizados en Bioestadística son los percentiles. Algunos de ellos pueden ser considerados también como parámetros de tendencia central y otros como de dispersión.
- parámetros de forma, que precisan la forma de la distribución. Podría decirse que expresan numéricamente la forma del histograma.

Parámetros de tendencia central

Los más importantes son:

- la media aritmética, o simplemente la media
- la mediana
- la moda
- los percentiles "centrales" (p 25 a p75)

En la explicación de los parámetros se utilizarán tres grupos de datos en los ejemplos:
Supuesto A): 8 , 1 , 4 , 8 , 8 , 5 , 1
Supuesto B): los "70 DATOS" originales del tema 4
Supuesto C): la tabla que agrupa a esos 70 datos

--La **MEDIA** es la suma de todos los valores dividida por el número de ellos.
<u>Símbolo</u>: $\bar{x}$
<u>Cálculo</u>:
1) datos aislados, originales:

$$\bar{x} = \frac{\sum x}{N} \quad \text{; para el ejemplo A:} \quad \bar{x} = \frac{8+1+4+8+8+5+1}{7} = 5$$

$$\text{para el ejemplo B:} \quad \bar{x} = 39{,}6$$

2) datos agrupados en clases:

$$\bar{x} = \frac{\sum fc}{N} \quad \text{; en el ejemplo C:}$$

$$\bar{x} = \frac{(6*13)+(6*20)+(8*27)+(11*34)+(11*41)+(11*48)+(9*55)+8*62)}{70} = 39{,}4$$

<u>Propiedades de la media</u>

1- si a cada valor de x le sumamos, restamos, multiplicamos o dividimos por una constante, la media queda sumada, restada, multiplicada o dividida por esa constante
2- la media es sensible a la variación de cada valor de x
3- la media se expresa en la misma unidad de medida que los datos originales
4- si la media tiene decimales es habitual expresarla con uno más que los datos originales

PDF processed with CutePDF evaluation edition www.CutePDF.com

Media aritmética ponderada

Se usa cuando se quiere o se debe dar una fuerza distinta a determinados valores.

$\bar{x}_{pond} = \frac{\sum xF}{\sum F}$ **, siendo x el valor original y F el factor de ponderación**

Ejemplos:

1) Al introducirse los estudios de Diplomatura en esta Escuela, el Área de Ciencias de la Enfermería englobaba diversas asignaturas, de cuyas notas salía la nota del Área. Como eran de extensión e importancia dispares, se decidió que Microbiología (que para abreviar llamaremos A) participaría con el 33%, la Bioestadística (B) con el 28%, las Prácticas (C) con un 23% y el resto, la media de Salud Pública, Organización e Historia de la Profesión ((D1+D2+D3)/3) conjuntamente con un 16%.
   Si las notas de las asignaturas fueron : 6 en A, 5 en B, 8 en C, 6 en D1, 8 en D2 y 10 en D3 , la nota del Área fué 6,5 y no la media aritmética 7,2
   $\bar{x}_{pond}$ = (6*33 + 5*28 + 8*23 + 8*16)/(33+28+23+16) = 6,5
2) la media de una distribución calculada a partir de una tabla es realmente una media ponderada en la que x es el punto medio de clase y f (frecuencia) el factor de ponderación F.

Otras medias

En circunstancias especiales (distribución con sesgo muy intenso) hay autores que prefieren otras medias como la media geométrica o la trimedia , en las que no vamos a entrar.
En los concursos varios jueces dan una nota al actuante. Para disminuir favoritismos e inquinas se utiliza la media recortada, que se obtiene prescindiendo del valor más alto y del más bajo. Este sistema se puede aplicar también para evitar errores, cuando se manejan grandes cantidades de datos y aparecen valores marginales "anómalos". Así se puede decidir no tener en cuenta un pequeño porcentaje (no más allá de un 3%) de los valores más altos y más bajos.

--La **MEDIANA** es el valor que ocupa el centro de la distribución una vez ordenados los datos. El símbolo es M

Cálculo:

1 – datos aislados, originales (¡que deben estar ordenados!)

a) N es impar: es el valor que ocupa el lugar (N+1)/2

b) N es par: es la media de los valores que ocupan los lugares N/2 y siguiente.

2 – datos agrupados

--de forma simplificada se toma como M el punto medio de la clase que contenga la mediana (el lugar se calcula como en los datos aislados) y se identifica la clase por la columna de frecuencias acumuladas.

--de forma un poco más exacta se utiliza la fórmula

$$M = L_i + i\left(\frac{N/2 - \sum f_M}{f_M}\right)$$

siendo L_i el límite inferior de la clase mediana, i su amplitud, N el nº total de datos, Σf_M las frecuencias acumuladas por debajo de la clase mediana y f_M la frecuencia de la clase mediana.

Ejemplos:

--supuesto A: se ordenan los 7 datos: 1 , 1 , 4 , 5 , 8 , 8 , 8 ; como N es impar la mediana será el valor que ocupe el lugar (7+1)/2 = 4 ; el 4º lugar es el 5

--supuesto B: se ordenan los 70 datos, número par. La mediana es la media de los valores que ocupen el lugar 70/2 = 35 y el siguiente, 36 . El 35º vale 40 y el 36º 41 , por tanto M = 40,5

--supuesto C: ***la clase mediana es la que contiene los valores 35º y 36º. En la columna de Σf se ve que pertenecen a la clase 38-44, que es la clase mediana. Por tanto M= c = 41

***aplicando la fórmula: $\mathbf{M = 37,5 + 7\left[\frac{\frac{70}{2} - 31}{11}\right] = 40}$

Propiedades de la mediana
Son las mismas que las de la media excepto la 2ª: la mediana sólo es sensible a la variación de los datos originales si se altera el orden en el centro de la distribución.

--La **MODA** es el valor más frecuente. Puede ocurrir que no haya moda o que haya más de una (empates en el máximo). El símbolo es Mo.
Cálculo:
-*en datos originales* se hace el recuento y se busca el valor más frecuente. Si hay empate, la moda es múltiple.
-*en datos agrupados* en tabla: la Mo será el punto medio de la clase modal, es decir, la más frecuente. En caso de empate se dan los puntos medios de las clases correspondientes.
Propiedades: como la mediana.
Ejemplos:
supuesto A: Mo = 8 ; supuesto B: Mo = 59 ; supuesto C: hay tres clases con frecuencia de 11; Mo = 34 , 41 y 48

De estos tres parámetros de tendencia central el mejor es sin duda alguna la media, pero hay algunos casos concretos (clases abiertas, valores muy discordantes) en que la mediana o incluso la moda son mejores. Cuando N≥30 la media suele ser un buen parámetro. En todo caso si el CV (coeficiente de variación), que luego veremos, supera el 50% la media no es buen representante del centro de la distribución.

Parámetros de dispersión

Informan de la dispersión de los datos, de la amplitud del conjunto. Los más importantes son:
-El RECORRIDO, que ya vimos en el tema 4 , o simplemente citar el máximo y el mínimo.
-La VARIANZA, que se basa en las diferencias entre cada valor y la media de la distribución.
-La DESVIACION ESTANDAR, que es la raíz cuadrada de la varianza.
-El COEFICIENTE DE VARIACIÓN, que relaciona la desviación estándar y la media.

--Varianza
Símbolo : s^2 (σ^2 , en la nomenclatura con caracteres griegos)
Cálculo: hay fórmulas distintas según los datos pertenezcan a una población o a una muestra.

– – población

– datos aislados: $$s^2 = \frac{N\sum x^2 - (\sum x)^2}{N^2}$$

– datos agrupados: $$s^2 = \frac{N\sum (fc^2) - (\sum fc)^2}{N^2}$$

– – muestra

– datos aislados: $$s^2 = \frac{N\sum x^2 - (\sum x)^2}{N(N-1)}$$

– datos agrupados: $$s^2 = \frac{N\sum (fc^2) - (\sum fc)^2}{N(N-1)}$$

Propiedades de la varianza

1- si a cada valor de x le sumamos o restamos una constante k, la varianza queda igual
2- si cada valor de x lo multiplicamos o dividimos por una constante k, la varianza queda multiplicada o dividida por k^2
3- la varianza es sensible a la variación de cada valor de x
4- la varianza se expresa en el cuadrado de la unidad de medida utilizada en la variable.
5- si la varianza tiene decimales, es habitual expresarla con dos decimales más que los datos originales

Ejemplos:
Con datos originales es conveniente construirse una tabla auxiliar con dos columnas: x y x^2.
--Así en el supuesto A (asumiendo que es una muestra):

x	x^2
8	64
1	1
4	16
8	64
8	64
5	25
1	1
---	---
35	235

$$s^2 = \frac{7 * 235 - 35^2}{7 * 6} = 10$$

--en el supuesto B : $s^2 = 207,58$
--en el supuesto C: la tabla auxiliar tendrá las columnas f , c , f*c , c^2 , fc^2 para que podamos tener los sumatorios necesarios para aplicar la fórmula.

$$s^2 = 218,96$$

--La **DESVIACION ESTANDAR** es la raíz cuadrada de la varianza y por tanto es un número más manejable y de utilización más frecuente.
Símbolo: s .También se usa mucho D.E. y la abreviatura inglesa S.D. Y la letra griega σ .

Fórmula: $s = \sqrt{s^2}$

Propiedades: como la media
Ejemplos:
-supuesto A: s = 3,2
-supuesto B : s = 14,4
-supuesto C: s = 14,8

--El **COEFICIENTE DE VARIACION** es un índice abstracto, que no tiene unidad de medida. Da igual que midamos la variable en cm , kg, sec., etc, , el coeficiente de variación se expresa siempre como %. (que puede ser mayor del 100%).
Símbolo: CV

Fórmula: $CV = \frac{100s}{\overline{X}}$

Aplicaciones:
1) comparar dispersiones de variables, incluso si están medidas en unidades distintas. La variable con el CV menor tiene la menor dispersión (y viceversa).
2) valorar la representatividad de una media. Es buena si no supera el 50%.

Ejemplos:
-supuesto A: 64%
-supuesto B: 36,4%
-supuesto C: 37,6%
-otro ejemplo: Los niños de 3 años de la ciudad C tienen una talla media de 93 cm con s = 3,8 . Los niños de 15 años de esa ciudad miden en media 162 cm con s = 6. ¿A que edad es la talla más variable?
Se calcula el CV: -a los 3 años: 4,09% -a los 15 años: 3,70%
Respuesta: La talla es más variable a los 3 años.

PARAMETROS DE FORMA

1) **SESGO** : es el grado de asimetría de una distribución, expresado por el coeficiente de sesgo o asimetría, cuyo valor ideal es 0 (entonces hay simetría). Cuando hay un Sesgo la parte más alta del histograma (o de la campana de Gauss) se desplaza hacia la derecha o la izquierda y la campana tiene una cola larga, donde estará la media, y otra más corta, en la que suelen estar la mediana y la moda. Si la media es menor que la M y/o la Mo, el sesgo es negativo y si es mayor, el sesgo es positivo.
Símbolo: Sg
Hay una fórmula, muy compleja, para calcular el coeficiente de sesgo, en la que no entramos.
Un cálculo aproximado es: $\mathbf{Sg = \frac{3(\bar{x} - M)}{s}}$, aunque lo mejor es observar la campana o el histograma. **Mirando la campana, si se desplaza a la derecha el sesgo es negativo; si lo hace a la izquierda, positivo**. Si nos ponemos en lugar de la campana, al revés.

Mirando el histograma de los "70 DATOS" (página 5.4) se ve que tiene un pequeño sesgo hacia la derecha, es decir, negativo. Con los datos originales el cálculo exacto da un sesgo de –0,196; la fórmula aproximada da -0,187. Con los parámetros calculados a partir de la tabla el sesgo vale según la fórmula aproximada –0,324.

2) **CURTOSIS**
es el grado de apuntamiento de una distribución, expresado por el coeficiente de curtosis, cuyo cálculo es complejo y no se ve aquí.
Símbolo: ct o k
Se toma como referencia a la campana de Gauss de la distribución normal, cuya k vale 0 y se dice que es mesocúrtica. Si la distribución es más alta y delgada, se dice que es leptocúrtica. y k es >0. Si es achatada y ancha se denomina platicúrtica y k es <0.
Los "70 DATOS" tienen una k = -1,105 y por tanto la distribución es algo platicúrtica.

PARAMETROS DE POSICION

1) **PERCENTILES**
Los percentiles (p) son parámetros de posición que nos indican la situación de cada valor en el conjunto de los datos ordenados, que se han dividido en 100 partes iguales. Se presentan como tabla o como gráfico.

Se expresan como pa siendo a el % de datos que queda por debajo del valor original al que corresponde ese percentil. Dicho de otra forma: a un valor le corresponde el percentil pa , cuando ordenados los datos el a% es menor que él y el (100-a)% es mayor.

Cálculo:

1- *en datos originales* : se ordenan los datos de menor a mayor y se calcula el lugar en el que estará el percentil (pa) buscado mediante la fórmula : lugar del pa = N*a/100. El valor que corresponda a es lugar o nº de orden será el pa
2- *en datos agrupados*: se utilizan la tabla o el gráfico de los porcentajes acumulados, interpolando, si es preciso. Hay una fórmula, parecida a la de la mediana, pero no suele ser necesaria.

Los percentiles se utilizan mucho en Pediatría en tablas y gráficos de crecimiento, pero en los últimos años su uso se ha extendido a muchos datos biológicos: colesterol, tensión arterial, densidad ósea... Han desplazado casi totalmente a otros parámetros de posición similares, como los deciles (el conjunto se divide en 10 partes iguales) y los cuartiles (el conjunto se divide en 4 partes).

Realmente hay100 percentiles, que van del p1 al p100, pero en la práctica se utilizan para mayor claridad sólo algunos de ellos. En Europa en las tablas y gráficos de crecimiento se utilizan el p3 , p10 , p25, p50, p75, p90, y p97.

El p50 se corresponde con el centro de la distribución: el 50% de los valores es mayor y el 50% es menor. Por tanto coincide con la mediana: p50 = M

En las variables biológicas los valores normales se obtienen a partir de muchas determinaciones en individuos sanos. Si un valor está por debajo del p3 se considera anormalmente bajo; si está por encima del p97, anormalmente alto; entre el p10 y el p90, totalmente normal. Entre el p3 y el p10, así como entre el p90 y el p97, aunque son aún normales, se consideran como en "zona de riesgo" o "sospecha", dada la proximidad de la zona anormal.

Los percentiles entre p25 y p75 pueden ser considerados también como parámetros de tendencia central y los mayores y menores como de dispersión.

Con los percentiles no pueden hacerse operaciones matemáticas, ya que son parámetros de posición . Así, pues, p50 ≠ (p25 + p75)/2

Al final de este tema puede verse un ejemplo de gráficos percentilados del peso y talla de niños de 2 a 18 años. Un niño de 5 ½ años que pesa 23 kg y mide 106 cm tiene una talla en el p10, un peso <p90 y una relación peso/talla >p97.

2) La **PUNTUACION TIPIFICADA O NOTA TIPIFICADA** puede ser también considerada como un parámetro de posición. Se verá con detalle en el tema 10. Adelanto:

Símbolos: se utilizan varios según las escuelas: c, z, SDS , SDE...

Fórmula: $$\mathbf{c = \frac{X - \bar{X}}{s}}$$

Equivalencias aproximadas entre percentiles y puntuaciones tipificadas:

p	**3**	**10**	**25**	**50**	**75**	**90**	**97**
c	**-2**	**-1,3**	**-0,7**	**0**	**0,7**	**1,3**	**2**

Dos observaciones finales

1) una distribución queda perfectamente definida conociendo todos los parámetros que hemos visto. Como el sesgo y la curtosis son de cálculo más difícil, el mínimo son la media y la desviación estándar, que suelen anotarse así : $\bar{\mathbf{x}} \pm s$ ó $\bar{\mathbf{x}} \pm DE$.
Que la media sola no es suficiente lo aclara el clásico ejemplo del pollo:" si una persona se come dos pollos y otra no come ninguno, la Estadística dirá que se comen un pollo cada uno". La media es ciertamente 1 . Pero si calculamos la desviación estándar la valoración puede ser distinta:

-uno come 2 pollos y el otro ninguno:

x	x^2
2	4
0	0
---	---
2	4

$$\mathbf{s} = \sqrt{\frac{(2*4)-2^2}{2*1}} = 1{,}4 \quad \text{y el CV} = 140\%$$

¡la media no es buena representante!

-cada uno come un pollo:

x	x^2
1	1
1	1
---	---
2	2

$$\mathbf{s} = \sqrt{\frac{(2*2)-2^2}{2*1}} = 0 \text{ y el CV será } 0\%$$

¡la media es buena representante!

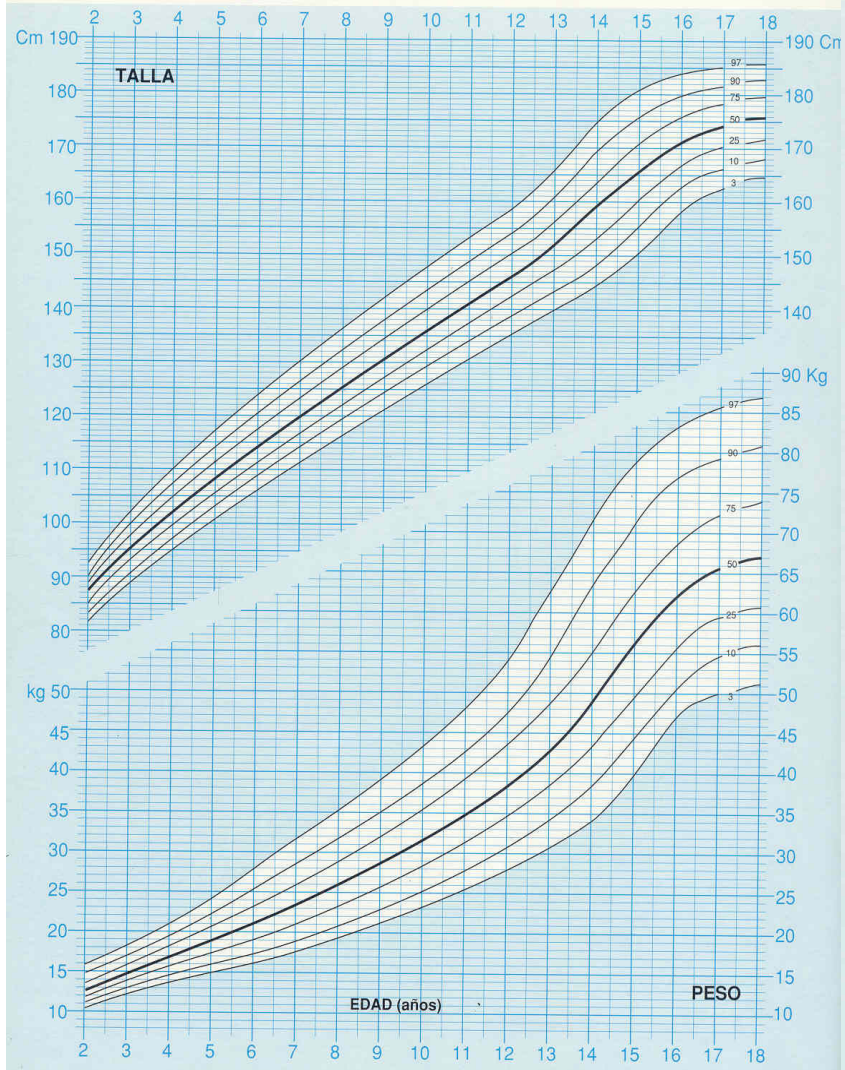
2) siempre que sea posible, los índices se calcularán a partir de los datos originales, ya que los cálculos a partir de la tabla conllevan algo de error. Como puede verse en este resumen con parámetros de algunos ejemplos que se han ofrecido en este tema:

"70 DATOS"	**Datos originales**	**Datos agrupados**
Media	**39,6**	**39,4**
Desviación estándar	**14,4**	**14,8**
Mediana	**40,5**	**40**
Moda	**59**	**34 , 41 , 48**
Coeficiente de variación	**36,4%**	**37,6%**

8

NIÑOS: 2 a 18 años

TALLA
PESO

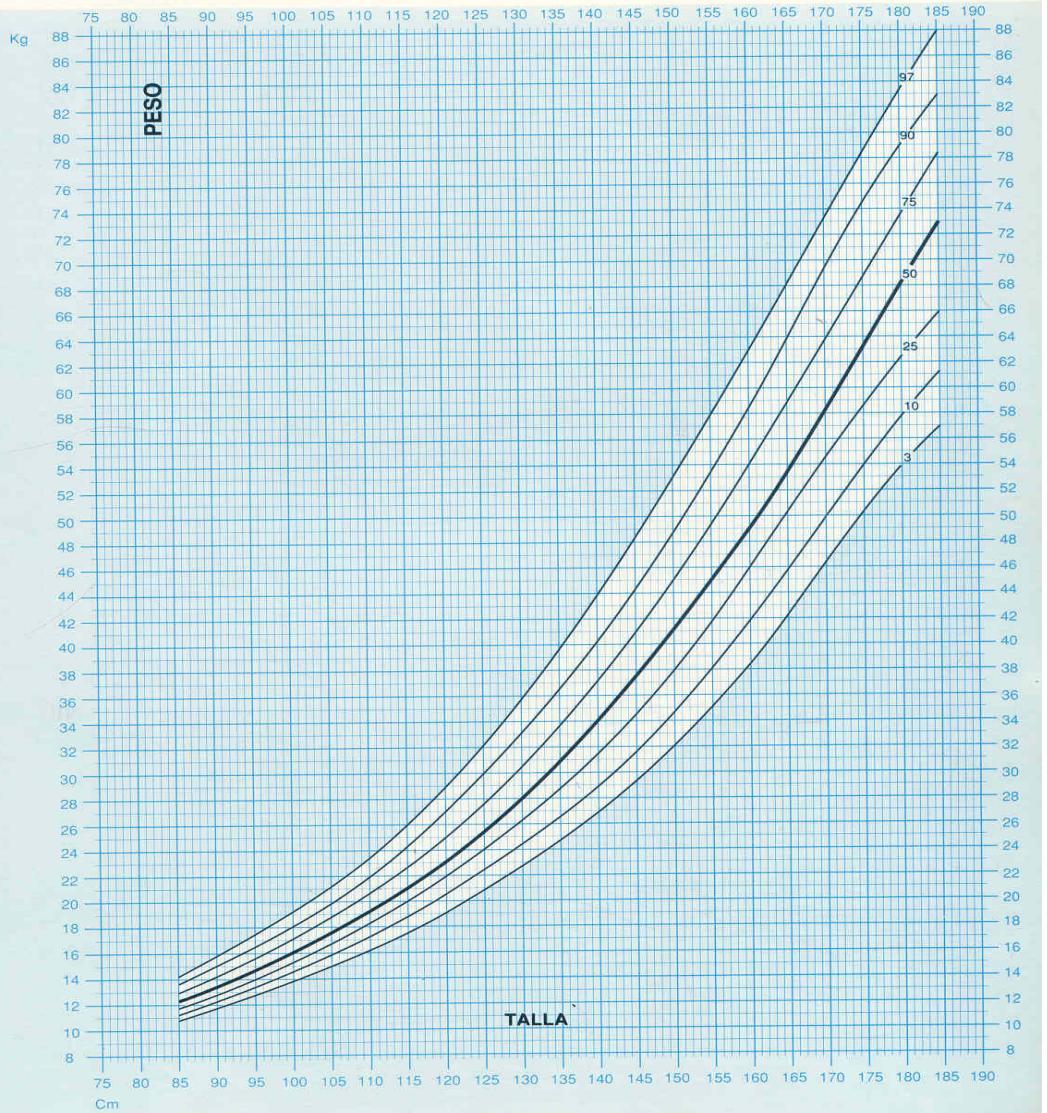


ESTUDIO LONGITUDINAL DE CRECIMIENTO, CURVAS DE 0 A 18 AÑOS.
M. Hernández, J. Castellet, J.L. Narvaiza, J. M. Rincón, I. Ruiz,
E. Sánchez, B. Sobradillo y A. Zurimendi.

INSTITUTO DE INVESTIGACION SOBRE CRECIMIENTO Y DESARROLLO.
FUNDACION F. ORBEGOZO. María Díaz de Haro, 10 bis, 48013 BILBAO
Realización Gráfica: Garsi Editorial - Londres, 17 - 28028 MADRID

NIÑOS

PESO-TALLA



ESTUDIO LONGITUDINAL DE CRECIMIENTO, CURVAS DE 0 A 18 AÑOS.
M. Hernández, J. Castellet, J. L. Narvaiza, J. M. Rincón, I. Ruiz,
E. Sánchez, B. Sobradillo y A. Zurimendi.

INSTITUTO DE INVESTIGACION SOBRE CRECIMIENTO Y DESARROLLO
FUNDACION F. ORBEGOZO. María Díaz de Haro, 10 bis, 48013 BILBAO
Realización Gráfica: Garsi Editorial - Londres, 17 - 28028 MADRID