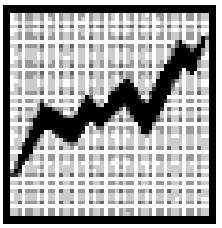
Ejemplos
Ejercicios
Misceláneas
Evaluación

Método cuantitativos de análisis gráfico



Método de cuadrados mínimos. Regresión lineal. Función χ^2. Obtención de los parámetros de un modelo. Correlación lineal. Incertidumbre de los parámetros de un ajuste.

3.1 – Método de cuadrados mínimos – Regresión lineal

En el capítulo anterior hemos enfatizado sobre la importancia de las representaciones gráficas. Asimismo hemos visto la utilidad de las versiones *linealizadas* de los gráficos de pares (X, Y) y las distintas maneras de llevar a cabo la linealización, puesto que nos encontraremos a menudo con tales situaciones en el laboratorio. En este capítulo formalizaremos algunos métodos analíticos para estimar los parámetros de un modelo que se confronta a los datos experimentales. De nuevo, el punto de partida será una representación gráfica de los datos experimentales, a la que queremos "superponer" la predicción de un modelo.

Por ejemplo, imaginemos que deseamos determinar la constante k de un resorte que sigue la ley de Hooke:

$$F = -k \cdot x \qquad (3.1)$$

donde F es la fuerza elástica y x la elongación del resorte. Para determinar k se procede a cargar al resorte con diferentos pesos P y medir la elongación que producen. Es fácil reconocer a estas cargas como el estímulo externo, el cual provoca como respuesta del "sistema resorte" la elongación observada x. Sin embargo, no es necesario que en la representación gráfica de los pares de datos (P, x), la carga P ocupe el lugar de la variable independiente sobre el eje de las abscisas. Por comodidad en el manejo de los datos, es más adecuado representar P en función de x, y de la pendiente de la recta obtener directamente la constante k buscada. En el presente caso, y como regla casi general, si representamos gráficamente los datos experimentales, éstos no caerán exactamente sobre una recta, sino que presentarán cierta dispersión como se ilustra en la Figura 3.1.

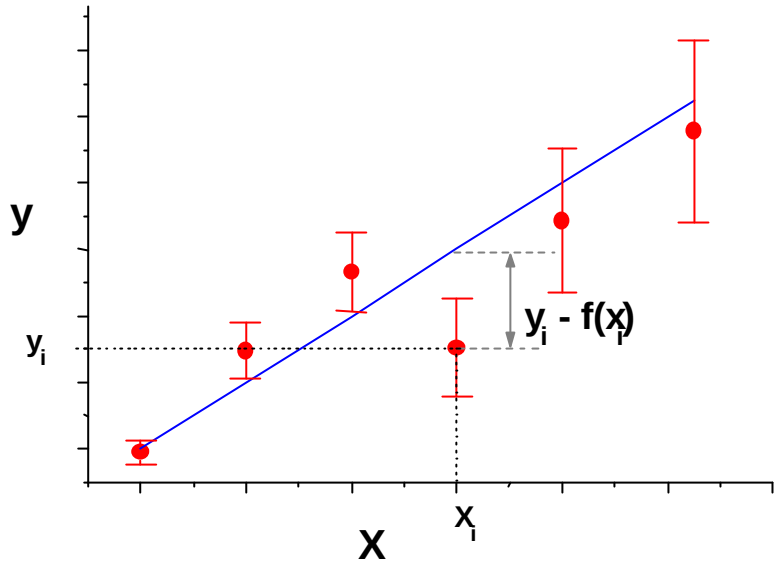


Figura 3.1. Gráfico de datos asociados a un modelo lineal. La cantidad y_I - $f(x_i)$ representa la desviación de cada observación de y_i respecto del valor predicho por el modelo $f(x_i)$.

Es útil definir la función χ^2 (Chi-cuadrado) como:

$$c^2 = \sum_i \left(P_i - k \cdot x_i\right)^2 \qquad (3.2)$$

que es una medida de la desviación total de los valores observados P_i respecto de los predichos por el modelo kx_i. El mejor valor de k es aquel que minimiza esta desviación total, o sea, el valor que puesto en (3.1) minimiza la función Chi-cuadrado. Por lo tanto, el mejor valor de k será el que se obtiene de resolver la siguiente ecuación:

$$\frac{dc^2}{dk} = \frac{d}{dk}\sum_i \left(P_i - k \cdot x_i\right)^2 = 2 \cdot \sum_i [\left(P_i \cdot x_i\right) - k \cdot \left(x_i\right)^2] = 0\,,$$

o sea:

$$k = \frac{\sum_i P_i \cdot x_i}{\sum_i x_i} \qquad (1.27)$$

En los programas como Excel, Origin, etc., este cálculo se realiza usando la herramienta "regresión lineal" o "ajuste lineal" que se aplica cuando la relación esperada entre las magnitudes medidas es lineal y todos los datos tienen la misma incertidumbre absoluta.

El método descripto aquí se aplica de manera análoga para un modelo lineal que incluya una ordenada al origen:

$$y = a \cdot x + b \qquad (3.3)$$

En este caso la función Chi-cuadrado es

$$\chi^2 = \sum_i \left(y_i - a \cdot x_i - b\right)^2 \qquad (3.4)$$

y para obtener los parámetros a y b se requiere minimizar la función repecto de ambos parámetros, es decir:

$$\frac{d\chi^2}{da} = 0$$

$$\frac{d\chi^2}{db} = 0$$

3.2 – Método de cuadrados mínimos incluyendo errores - Regresión no lineal

Supongamos que tomamos una serie de mediciones de dos magnitudes cuya relación deseamos determinar. El resultado de nuestras N mediciones dará lugar a un conjunto de N ternas de la forma (x_i, y_i, σ_i), donde σ_i es la incertidumbre asociada a la determinación de y_i. Aquí suponemos que la incertidumbre de x_i es despreciable. Supongamos que el modelo que ajusta los datos viene dado por la función $f(x;a,b,c,...)$, donde a, b, c, etc., son los n_{par} parámetros del modelo. Al estimador del valor de y dado por el modelo lo designamos por $y(x_i)=f(x_i\ ;a,b,c,...)$. Decimos que $y(x_i)$ representa la variación determinista de y con x.

En este caso más general definimos el valor de Chi-cuadrado como:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - y(x_i)\right)^2}{\sigma_i^2} = \sum_{i=1}^{N} w_i \times \left(y_i - y(x_i)\right)^2 \qquad (3.7)$$

donde los valores w_i son los factores de peso de cada triada de datos (x_i, y_i, σ_i); en este caso $w_i=1/\sigma_i^2$. Definimos el número de grados de libertad, ν, del modelo como:

$$\nu = N - n_{par}. \tag{3.8}$$

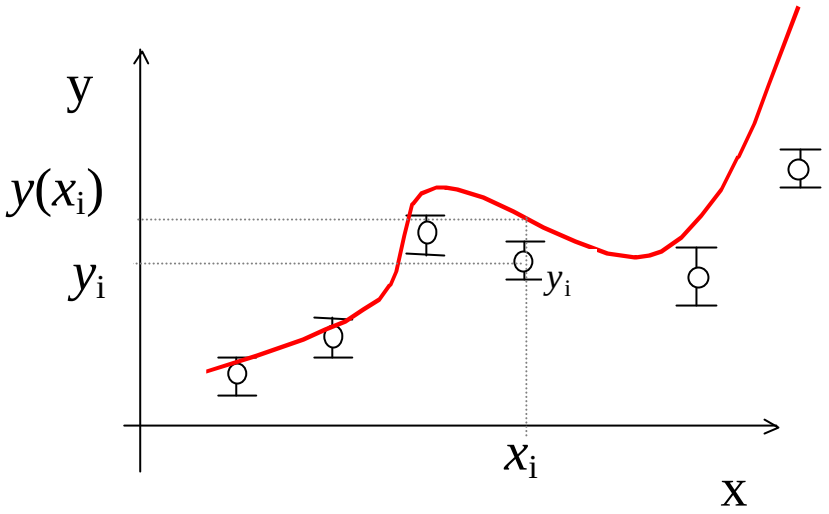


Figura 3.2. Diagrama esquemático de un ejemplo de modelo no lineal representado por la función $f(x_i)$. s_i representa el error absoluto asociado a cada observación y_i.

Introducimos la definición del error medio:

$$s^{-2} = \frac{1}{\frac{1}{N} \cdot \sum_{i=1}^{N} \frac{1}{s_i^2}} = \frac{1}{\overline{w}} = \frac{1}{\frac{1}{N} \cdot \sum_{i=1}^{N} w_i} \tag{3.9}$$

En nuestro caso hemos definido los factores de peso de cada triada de datos como la inversa del cuadrado de la incerteza s_i, aunque a veces es útil emplear otros factores de peso de los datos, como por ejemplo:

$$w_i = \frac{1}{y_i}, \quad \text{o} \quad w_i = \frac{1}{y_i^2}, \quad \text{etc.} \tag{3.10}$$

También definimos la *variancia total* como:

$$S_t^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} w_i \cdot (y_i - \bar{y})^2 \tag{3.11}$$

S_t es una medida de la dispersión de los datos alrededor del valor medio de $\bar{y}$. Este valor no depende del modelo (función $f(x)$), o sea que S_t ignora toda variación determinista de y con x.

También definimos la v*arianza del ajuste*:

$$S_f^2 = \frac{1}{N - n_{par}} \cdot \sum_{i=1}^{N} w_i \cdot (y_i - y(x_i))^2 = \frac{1}{\nu} \cdot \chi^2 = \chi_\nu^2 \qquad (3.12)$$

La varianza del ajuste, S_f, al igual que χ^2 o χ_ν^2 (Chi-cuadrado por grado de libertad), miden la dispersión residual de los datos alrededor del valor determinista, o sea son medidas de la bondad del ajuste de $y(x_i)$ a los valores medido y_i. Si el modelo determinista fuese el adecuado, su valor estaría asociado a las fluctuaciones estadísticas de y_i respecto de su valor $y(x_i)$.

A veces es útil definir el coeficiente de regresión:

$$R^2 = \left(\frac{S_t^2 - S_f^2}{S_t^2} \right) \qquad (3.13)$$

Si el modelo $y(x_i)$ es una *buena* representación de los datos, es de esperar que tanto S_f como χ^2 sean pequeños y que $S_t >> S_f$, de donde se deduce que $R^2 \approx 1$. En caso contrario, tanto S_f como χ^2 serán grandes y $S_t \approx S_f$ por lo tanto $R^2 \approx 0$.

Una "receta" para la determinación de las incertidumbres de los parámetros del modelo

Al igual que en el caso del modelo lineal discutido anteriormente, los mejores valores de los parámetros del modelo se obtienen de la minimización de la función Chi-cuadrado:

$$a^* \Leftrightarrow \left. \frac{\partial \chi^2(a,b,c,...)}{\partial a} \right|_{a=a^*} = 0 . \qquad (3.14)$$

Esto es, $\chi_{min}^2 = \chi^2(a^*, b^*, \ldots)$.

La determinación de las incertidumbres en los parámetros $(a^*, b^*, c^*,...)$ es un procedimiento sofisticado sobre el que existen diversas teorías y opiniones[1,3]. Un método aproximado para calcular estas incertidumbres en forma gráfica se indica en la figura 1.5.

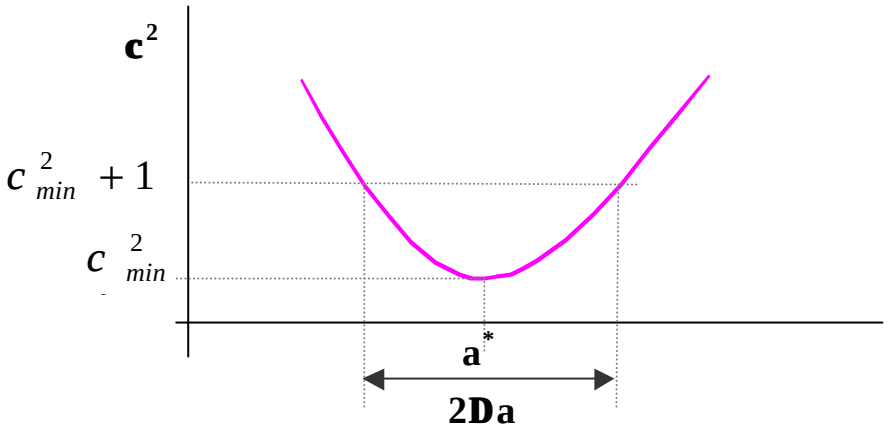


Figura 3.3. Esquema gráfico que ilustra un procedimiento aproximado para obtener las incertidumbres de los parámetros de un modelo no lineal.

3.3 – Regresión lineal considerando las incertidumbres de medición

Un caso especial importante del esquema estadístico discutido precedentemente, es el de la regresión lineal, ya que en este caso es posible resolver las expresiones generales en forma analítica, lo que facilita su uso y programación en muchas aplicaciones prácticas. Igual que antes supondremos que se tienen una serie de mediciones de dos magnitudes x e y cuya relación se supone lineal, es decir:

$$y = a \cdot x + b$$

donde a y b son los parámetros del modelo que deseamos determinar y evaluar. El resultado de nuestras N mediciones dará lugar a un conjunto de N ternas de la forma (x_i, y_i, s_i), donde s_i es la incertidumbre asociada a la determinación de y_i. También aquí suponemos que la incertidumbre de x_i es despreciable. Al igual que antes definimos:

$$w_i = \frac{1}{s_i^2} . \qquad (3.15)$$

Ya vimos que este modo de definir el peso de los datos puede variarse según sea el caso. En particular si no se dispone de las incertidumbres s_I, los w_i pueden tomarse iguales a 1. Usando las siguientes definiciones:

$$SXn = \sum_{i=1}^{N} w_i \cdot x_i^n \,, \qquad SYn = \sum_{i=1}^{N} w_i \cdot y_i^n \,. \tag{3.16}$$

$$SXY = \sum_{i=1}^{N} w_i * x_i * y_i \,, \qquad Sum = \sum_{i=1}^{N} w_i \,. \tag{3.17}$$

$$< x > \equiv \overline{X} = \frac{SX}{Sum}, \qquad < y > \equiv \overline{Y} = \frac{SY}{Sum}$$

$$Var\,(x) = \frac{SX2}{Sum} - \overline{X}^2 \qquad \text{y} \tag{3.18}$$

$$\Delta = Sum \cdot SX2 - (SX)^2 = sum^2 \cdot Var(x)$$

y usando (1-38) es posible demostrar que:

$$a = \frac{1}{\Delta} \cdot \left[SXY \cdot Sum - SX \cdot SY \right]. \tag{3.19}$$

$$b = \frac{1}{\Delta} \cdot \left[SX2 \cdot SY - SX \cdot SXY \right] = < y > - a \cdot < y > \tag{3.20}$$

siendo sus incertidumbres:

$$\begin{aligned} \sigma_a^2 \equiv Var(a) \equiv (\Delta a)^2 &= \frac{Sum}{\Delta} = \frac{\sum_i w_i \cdot (y_i - a \cdot x_i - b)^2}{(N-2) \cdot Sum \cdot Var(x)} \,, \\ \sigma_b^2 \equiv Var(b) \equiv (\Delta b)^2 &= \frac{SX2}{\Delta} = Var(a) \cdot \frac{SX2}{Sum} \end{aligned} \tag{3.21}$$

respectivamente. De modo análogo se demuestra que el coeficiente de correlación viene dado por:

$$r = \frac{SXY - Sum \cdot \bar{x} \cdot \bar{y}}{\left[SX2 - Sum \cdot \bar{x}^2 \right] \cdot \left[SY2 - Sum \cdot \bar{y}^2 \right]} = \frac{Cov\,(x,y)}{Var(x) \cdot Var(y)} \tag{3.22}$$

Este parámetro da una idea de la bondad del modelo lineal propuesto. Si r es próximo a 1, el modelo es adecuado, mientras que si $r \approx 0$ el modelo lineal no es el modelo adecuado. Si $r \approx 0$ esto no significa que no haya una vinculación o correlación entre x e y, sino que el modelo lineal no es el adecuado. Por ejemplo, si los pares de puntos (x,y) tiene una relación tal que caen sobre un círculo, tendríamos $r \approx 0$. Desde luego, si los pares (x,y) no tienen ninguna correlación entre ellos, también tendríamos que $r \approx 0$.

Las incertidumbres en los valores de los parámetros a y b también pueden escribirse en términos de r del siguiente modo:

$$Var(a) = \frac{a^2}{(N-2)} \cdot \left(\frac{1}{r^2} - 1 \right) ,$$
$$Var(b) = Var(a) \cdot < x^2 > . \qquad (3.23)$$

Las ecuaciones (3.23) son de verdadera importancia cuando se realiza un análisis profundo de los datos experimentales.

3.4 - Aplicaciones

Consideremos el estudio experimental de un péndulo simple, al que se mide el período T para distintas longitudes L. Supongamos que cada período $T_i(L_i)$ (al que consideraremos la variable dependiente del problema) está determinado con la misma incertidumbre $\sigma(T_i)$, como se muestra en la Figura 3.4 a.

En caso de linealizar la representación mediante el cambio de variables (L, T^2), la nueva variable dependiente T^2 tiene incertidumbre $\sigma(T_i^2)$ dada por las fórmulas de propagación:

$$\sigma(T^2) = \left| \frac{\partial T^2}{\partial T} \sigma(T) \right| = 2T\sigma(T)$$

de donde se ve inmediatamente que $\sigma(T_i^2)$ es función de T (ver también la Figura 3.4 b). Si procedemos a estimar los pará,etros del ajuste lineal a partir de los datos de la Figura 3.4 b debemos usar las fórmulas de la Sección 3.3.

Las mismas consideraciones son válidas en el caso de transformaciones usando la función logaritmo. Recordamos que, en caso de efectuar una linealización usando escalas logarítmicas, la incertidumbre propagada de una variable Y es (ver Capítulo 1):

$$\sigma[\log(Y)] = \left| \frac{\partial \log(Y)}{\partial Y} \sigma(Y) \right| = \frac{\sigma(Y)}{|Y|}$$

que es una función de la variable Y.

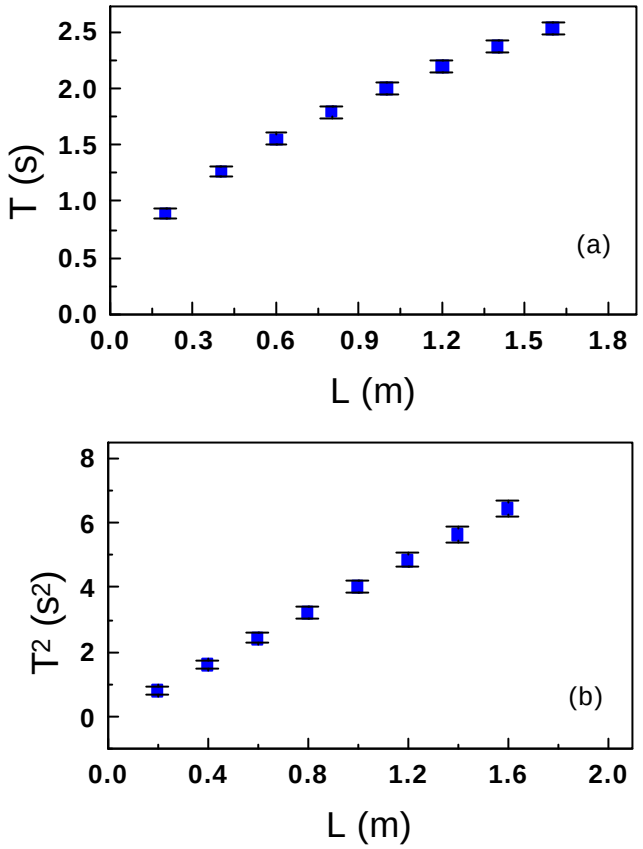


Figura 3.4. Al cambiar la representación hay cambios en las incertidumbres de las variables. a) la incertidumbre de la variable dependiente es uniforme. b) el cambio T^2 por T implica no-uniformidad de las incertidumbres. Este hecho debe considerarse cuando se evalúen los parámetros del ajuste.

3.4 – Comentarios finales

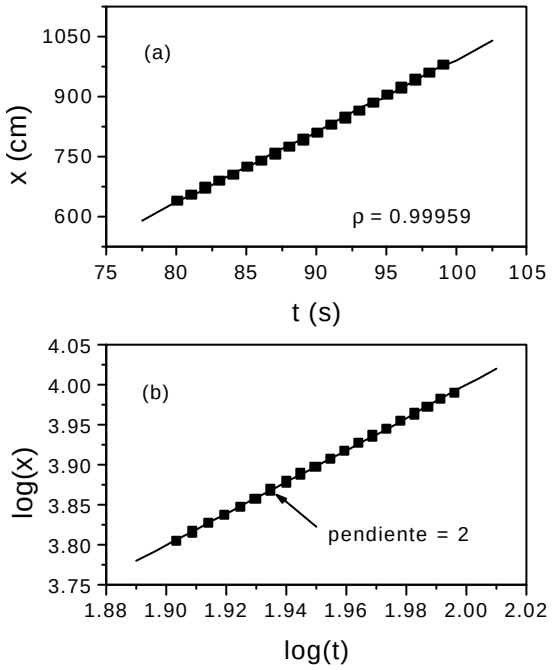
Vale la pena notar que no siempre es suficiente admitir que dos variables siguen una relación lineal guiándonos por lo que muestra un gráfico de los datos en escalas lineales. Menos aun si *sólo* evaluamos el coeficiente de correlación del ajuste lineal que propondríamos a partir de este gráfico. Un gráfico de $Y = X^{1.1}$ (variables sin correlación lineal) puede ajustarse por una recta y obtenerse a la vez un coeficiente de correlación lineal (inexistente) de, por ejemplo, 0.998. Un gráfico de datos experimentales de $Y=X$ con algo de dispersión fortuita de los puntos, podría devenir en un coeficiente de, por ejemplo, 0.995, menor que el anterior. Entre los coeficientes hay una diferencia, apenas, del 3 por mil. Pero en un gráfico log-log, la diferencia de pendientes será la que hay entre 1.1 y 1.0, lo que representa un 10% de discrepancia entre los exponentes de la variable X. Estos métodos de análisis nos enseñan que los efectos de correlación pueden estar enmascarados por el efecto del "ruido" de los datos. Muchas veces lo difícil es establecer si existe

correlación entre las variables, aun cuando los datos provengan de fuentes “limpias”, las que hayan producido datos con relativamente poca dispersión.

Imaginemos un experimento donde se mide la distancia que recorre un móvil sobre una línea recta mientras una fuerza constante actúa sobre él. Esperamos, por tanto, que el movimiento sea uniformemente acelerado. Supongamos que el cuerpo parte del reposo, que medimos *x(t)* y que los datos colectados son los de la Figura 3.5. Si los datos experimentales se analizan sobre este gráfico con escalas lineales, el ajuste por un modelo lineal es más que tentador. Hecho ésto, se obtiene la ecuación de la mejor recta y un coeficiente de correlación $\boldsymbol{r}$ = 0.9995. Sin embargo, un modelo basado en las ecuaciones de la dinámica dice que

$$x = \frac{1}{2} a t^2$$

donde *a* es la aceleración. En la Figura 3.5.b están los logaritmos de los mismos datos, de donde se ve claramente la proporcionalidad $x \propto t^2$ que predice el modelo, difícilmente demostrable a partir del gráfico de la Figura 3.5.a.



Representación de x(t) para un cuerpo que se mueve con movimiento uniformemente acelerado. (a) No se aprecia la curvatura de los datos y bien podría suponerse que la correlación es lineal. El coeficiente de correlación lineal, en efecto, es muy alto. (b) *log(x)* en función de *log(t)*, de donde se ve que la relación es cuadrática. Para el análisis de errores, ver los comentarios de la Sección 3.3.

Bibliografía

1. *Data reduction and error analisys for the physical sciences*, 2nd ed., P. Bevington and D. K. Robinson, McGraw Hill, New York (1993).

2. *Numerical recipies in Fortran*, 2nd. ed., W.,H. Press, S.A. Teukolsky, W.T. Veetterling and B.P. Flanner, Cambridge University Press, N.Y. (1992). ISBN 0-521-43064x.

3. *Data analysis for scientists and engineers*, Stuardt L. Meyer, John Willey & Sons, Inc., N.Y. (1975). ISBN 0-471-59995-6.

4. *Estadística*, Spiegel y Murray, 2da ed., McGraw Hill, Schaum, Madrid (1995). ISBN 84-7615-562-X.

5. *Uncertainty in the linear regression slope*, J. Higbie, Am. J. Phys. **59**, 184 (1991); *Least squares when both variables have uncertainties*, J. Orear, Am. J. Phys. **50**, 912 (1982).

6. *Probability, statistics and Montecarlo*, Review of Particle Properties, Phys. Rev. D **45**, III.32, Part II, June (1992).

7. *Teoría de probabilidades y aplicaciones*, H. Cramér, Aguilar, Madrid (1968); *Mathematical method of statistics*, H. Cramér, Princeton Univ. Press, New Jersey (1958).