Apéndice F - Método de cuadrados mínimos análisis avanzado

Método de cuadrados mínimos. Regresión lineal. Función χ^2. Obtención de los parámetros de un modelo. Incertidumbre de los parámetros de un ajuste. Bondad de ajuste. Intervalos de confianza. Muestras pequeñas. Simulaciones: método de Montecarlo

Introducción

En la unidad 5 se presentaron algunos útiles para la de determinación de los parámetros de una relación lineal que mejor ajusta un conjunto de datos. También se discutió brevemente algunos criterios para evaluar la calidad del ajuste logrado. En este apéndice nos proponemos generalizar dichos procedimientos para el caso en que el modelo propuesto no sea lineal y los datos que estamos analizando estén afectado de errores. Asimismo se presentan criterios de análisis de bondad de un ajuste. Finalmente se introducen los rudimentos para simular un conjunto de datos experimentales a través del método de Monte Carlo, y su implementación usando una planilla de cálculo.

1 – 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 general definimos el valor de Chi-cuadrado como:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - y(x_i))^2}{\sigma_i^2} = \sum_{i=1}^{N} w_i \cdot (y_i - y(x_i))^2 \qquad \text{(F.1)}$$

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

$$v = N - n_{par}\,. \qquad \text{(F.2)}$$

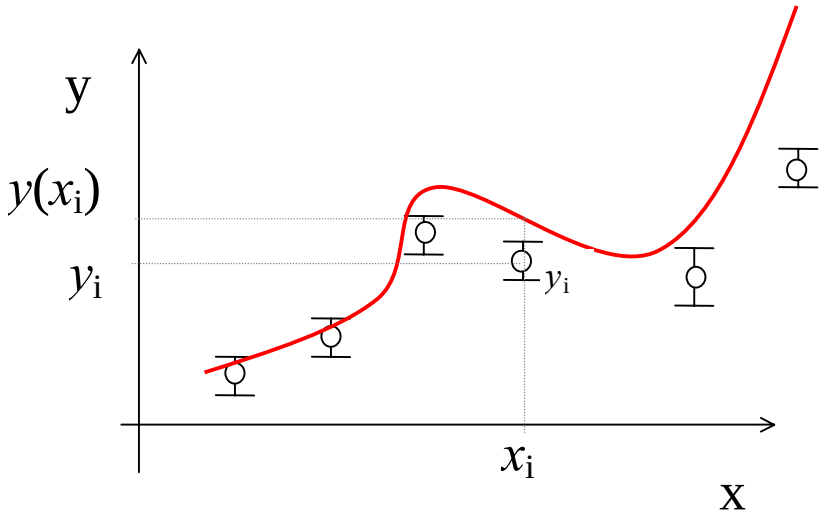

Figura F.1. Diagrama esquemático de un ejemplo de modelo no lineal representado por la función $f(x_i)$. σ_i representa el error absoluto asociado a cada observación y_i.

Introducimos la definición del error medio:

$$\sigma^2 = \frac{1}{\frac{1}{N}\cdot\sum_{i=1}^{N}\frac{1}{\sigma_i^2}} = \frac{1}{\frac{1}{N}\cdot\sum_{i=1}^{N} w_i} \qquad \text{(F.3)}$$

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

$$w_i = \frac{1}{y_i}, \quad \text{o} \quad w_i = \frac{1}{y_i^2}, \quad \text{etc.} \qquad \text{(F.4)}$$

El valor medio $\bar{y}$ de y_i se define como:

$$y = \frac{\sum_{i=1}^{N} w_i \cdot y_i}{\sum_{i=1}^{N} w_i}\,. \qquad \text{(F.5)}$$

Si todos los valores y_i tienen el mismo peso (los errores σ_i son iguales), esta expresión se reduce a la (2.7). También definimos la *variancia total* como:

$$S_t^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} w_i \cdot (y_i - \bar{y})^2 \qquad \text{(F.6)}$$

$\boldsymbol{S_t}$ 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 o el valor de chi-cuadrado por grado de libertad,* χ_v^2 *,* como:

$$S_f^2 = \frac{1}{N - n_{par}} \cdot \sum_{i=1}^{N} w_i \cdot (y_i - y(x_i))^2 = \frac{1}{v} \cdot \chi^2 = \chi_v^2 \qquad \text{(F.7)}$$

La varianza del ajuste, S_f, al igual que χ^2 o χ_v^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 propuesto *f(x)* 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, que también da una idea de la calidad del ajuste o bondad del modelo, como:

$$R^2 = \left(\frac{S_t^{\,2} - S_f^{\,2}}{S_t^{\,2}} \right) \qquad \text{(F.8)}$$

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 $R^2 \approx 0$.

Estimación de las incertidumbres de los parámetros del modelo

Al igual que en el caso del modelo lineal discutido en la Unidad 5, los mejores valores de los parámetros del modelo se obtienen de la minimización de la función Chi-cuadrado, o sea el mejor valor de del parámetro *a* del modelo, *a**, vendrá dado por:

$$a^* \Leftrightarrow \left.\frac{\partial \chi^2(a,b,c,...)}{\partial a}\right|_{a=a^*} = 0. \qquad \text{(F.9)}$$

De modo que, $\chi_{min}^2 = \chi^2(a^*, b^*, \ldots)$, es el mínimo de χ^2.

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 y práctico para calcular estas incertidumbres en forma gráfica[3] se indica en la Fig. F.2.

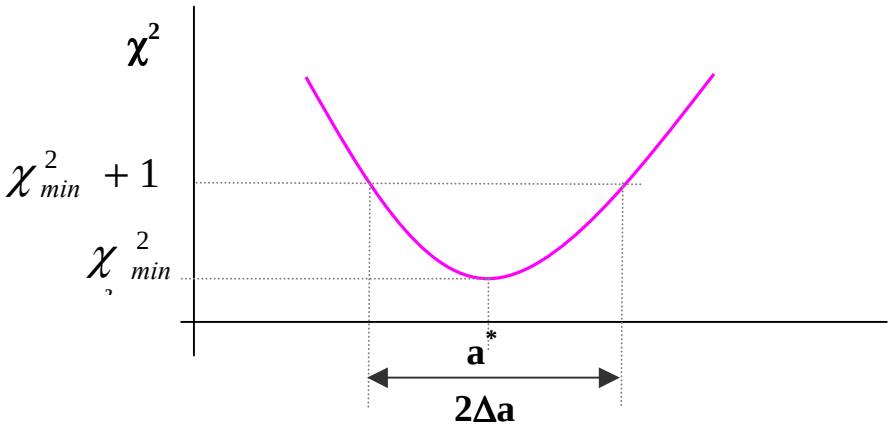

Figura F.2. Esquema gráfico que ilustra un procedimiento aproximado para obtener las incertidumbres de los parámetros de un modelo no lineal.

Para el caso de una variable, *a*, la técnica consiste en graficar χ^2 en función de *a*. χ^2 pasará por un mínimo (*a**) que determina el mejor valor del parámetro *a*. En este punto el valor de χ^2 será χ_{min}^2. Luego se determina el ancho del intervalo definido por las ordenadas que hacen $\chi^2 = \chi_{min}^2 + 1$. Este intervalo de ordenadas determina[3] el intervalo de incerteza Δa del mejor valor *a**.

2 – Regresión lineal considerando las incertidumbres de medición

Un caso especial de particular importancia es el de la regresión lineal. 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, σ_i), donde σ_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:

$$w_i = \frac{1}{\sigma_i^{\,2}} . \tag{F.10}$$

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 σ_I, los w_i pueden tomarse iguales a 1. Usando las siguientes definiciones:

$$SXn = \sum_{i=1}^{N} w_i \cdot x_i^n \,, \qquad SYn = \sum_{i=1}^{N} w_i \cdot y_i^n \,. \tag{F.11}$$

$$SXY = \sum_{i=1}^{N} w_i \cdot x_i \cdot y_i \,, \qquad Sum = \sum_{i=1}^{N} w_i \,. \tag{F.12}$$

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

$$Var\,(x) = \frac{SX2}{Sum} - \bar{X}^2 \qquad \text{y} \tag{F.13}$$

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

y usando (F.7) es posible demostrar[1,2,3] que:

$$a = \frac{1}{\Delta} \cdot \left[SXY \cdot Sum - SX \cdot SY \right]. \tag{F.14}$$

$$b = \frac{1}{\Delta} \cdot \left[SX2 \cdot SY - SX \cdot SXY \right] = < y > - a \cdot < y > \tag{F.15}$$

siendo sus incertidumbres:

$$\begin{aligned} \sigma_a^2 \equiv Var(a) \equiv (\Delta a)^2 &= \frac{Sum}{\Delta} = \frac{\sum_i w_i \cdot (y_i - a \cdot x_i - b)^2}{(N-2) \cdot Sum \cdot Var(x)} \,, \\ \sigma_b^2 \equiv Var(b) \equiv (\Delta b)^2 &= \frac{SX2}{\Delta} = Var(a) \cdot \frac{SX2}{Sum} \end{aligned} \tag{F.16}$$

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

$$\rho = \frac{SXY - Sum \cdot \bar{x} \cdot \bar{y}}{\left[SX2 - Sum \cdot \bar{x}^2\right] \cdot \left[SY2 - Sum \cdot \bar{y}^2\right]} = \frac{Cov\,(x,y)}{Var\,(x) \cdot Var\,(y)} \qquad \text{(F.17)}$$

Este parámetro da una idea de la bondad del modelo lineal propuesto. Si ρ es próximo a 1, el modelo es adecuado, mientras que si $\rho \approx 0$ el modelo lineal no es el modelo adecuado. Si $\rho \approx 0$ 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 $\rho \approx 0$. Desde luego, si los pares (*x,y*) no tienen ninguna correlación entre ellos, también tendríamos que $\rho \approx 0$. Ver Unidad 5.

Precausiones

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 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 es difícil es establecer si existe correlación lineal entre las variables, aun cuando los datos tengan 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 Fig. F.3. 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 ρ = 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 Fig. F.3.b están los logaritmos de los mismos datos, de donde se ve claramente la proporcionalidad $x \propto t^2$ que predice el modelo, difícilmente percibible a partir del gráfico de la Figura 3.5.a o del mero análisis de ρ.

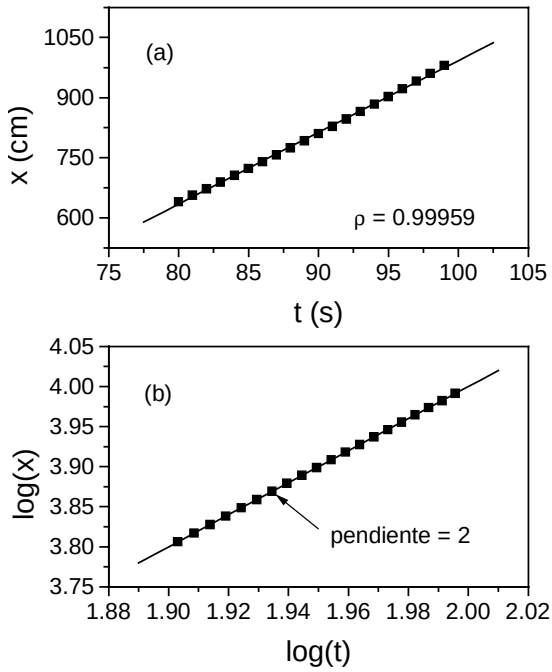

Figura F.3-Representación de x(t) para un cuerpo que se mueve con movimiento uniformemente acelerado. (a) No se aprecia la curvatura de los datos y bien podría suponerse que la correlación es lineal. El coeficiente de correlación lineal, en efecto, es muy alto. (b) *log(x)* en función de *log(t)*, de donde se ve que la relación es cuadrática.

4– Bondad del ajuste- Criterios cuantitativos

Volviendo al caso general de determinar los parámetros de un modelo que mejor ajustan un conjunto de datos experimentales, es útil disponer de un criterio cuantitativo de evaluación de la bondad o calidad del ajuste. Una medida de la bondad de un ajuste está dada por el valor de χ_v^2, definida por:

$$\chi_v^2 = \frac{1}{N - n_{par}} \cdot \chi^2 = \frac{1}{\nu} \cdot \sum_{i=1}^{N} \frac{\left(y_i - y(x_i)\right)^2}{\sigma_i^2}, \qquad \text{(F.18)}$$

donde $\nu = N - n_{par}$ es el *número de grados de libertad*. Es evidente que si todos nuestros datos experimentales (y_i) tienen desviaciones respecto del modelo ($y(x_i)$) que no

sobrepasan el error (σ_i), nuestro modelo es una descripción adecuada de nuestras observaciones. También es claro que en este caso, según (F.18), cada uno de los términos de la sumatoria será del orden de la unidad y por lo tanto χ_v^2 misma tendrá un valor cercano a uno. En otras palabras, si χ_v^2 es del orden de la unidad o menor decimos que el modelo propuesto para explicar los datos experimentales es adecuado y viceversa. Si χ_v^2 es mucho mayor que uno, el modelo no es una buena descripción de nuestros datos. Cuando $\chi_v^2 << 1$, se dice que el modelo es *demasiado bueno*, lo cual también es sospechoso o indicativo de que la distribución de los datos no es normal o que se sobre estimaron los errores. El criterio que acabamos de describir, aunque cualitativo, es un criterio práctico y útil en la mayoría de los casos.

Más cuantitativamente, para el caso en que la distribución estadística de los valores y_i sea normal, podemos calcular la probabilidad que un dado valor de $\chi^2 = \chi_0^2$ haya ocurrido sólo por azar. Esta probabilidad viene dada por:

$$P_{\chi_0^2,N} = P(\chi^2 \geq \chi_0^2) = Q\left(\frac{N - n_{par}}{2}, \frac{\chi_0^2}{2}\right) = \\ = CL(\chi_0^2) = Chi(v, \chi_0^2) = Dist.Chi(\chi_0^2, v) \qquad \text{(F.19)}$$

con

$$Q(a,x) = \frac{1}{\Gamma(a)} \cdot \int_x^\infty e^{-t} \cdot t^{a-1} \cdot dt \qquad \text{(F.20)}$$

$$Q(a,0) = 1, \qquad Q(a,\infty) = 0 \quad y \quad a > 0$$

donde ***Q(a,x)*** se conoce como la función Gama incompleta. Los l***ímites de confianza*** CL(χ_0^2) [usamos CL de la denominación en inglés *Confidence Level*] están dados por la integral de la función de distribución Chi-Cuadrado. La función de distribución chi-cuadrado es la derivada $Q(v/2, \chi^2)$, su valor medio es v y su varianza es $2v$. *CL* representa la probabilidad de que una repetición del experimento dará un valor de $\chi^2 \geq \chi_0^2$, sólo por azar, aún cuando el modelo sea correcto. Estas funciones están incorporadas en diversos programas matemáticos (Mathematica, MatLab, etc.) y en planillas de cálculos como Excel. En particular en esta planilla de cálculo, la función se denota con la expresión *Dist.Chi(x,v)*, como se indica en (F.19). Por lo general se considera que un valor de $\boldsymbol{Q} \geq 0.1$ es indicativo de un modelo creíble. Un valor de $\boldsymbol{Q}$ entre (0.1 y 0.001) es todavía marginalmente aceptable o indicativo de que los errores fueron subestimados. Si $\boldsymbol{Q} < 0.001$ la validez del modelo debe ser revisada y su credibilidad es dudosa .

Nota: Prácticamente todas la planillas de cálculo y programas de ajuste realizan estos cálculos, pero sin incluir los errores de las variables. Para tener en cuenta a los mismos es necesario programar estas funciones en, por ejemplo, Excel. En la hoja de calculo de Excel, Errores_SG.xls, que está a disponible en www.fisicarecreativa.com se dispone de estas rutinas en VBA (*Visual Basic for Application*), que pueden usarse, precisamente,

desde una planilla Excel. En estas subrutinas, el parámetro MODE se usa para indicar el modo como se evalúan los pesos $\boldsymbol{w_i}$:

$$MODE = \begin{cases} 1 & w_i = 1 \quad \text{no se consideran los errores} \\ 2 & w_i = 1/\sigma_i^2 \quad \text{se consideran los errores} \\ 3 & w_i = 1/y_i^2 \quad \text{se consideran los errores } \sigma_i^2 = y_i \\ 4 & w_i = 1/|y_i| \quad \text{se toma como peso la inversa de } |y_i| \end{cases} \qquad \text{(F.21)}$$

6 – Intervalos de confianza y nivel de significación. Muestras pequeñas

En muchos casos prácticos, se realiza un conjunto de ***N*** mediciones o se extrae una muestra de ese tamaño, cuyos valores son ($\boldsymbol{x_1, x_2,...x_N}$), con el objeto de determinar o estimar el valor de algún parámetro desconocido $\alpha=\alpha(x_1,x_2, ...x_N)$. Imaginemos que el *estimador* de este parámetro es $\alpha^*=\alpha^*(x_1,x_2, ...x_N)$, cuyo valor podría haberse obtenido, por ejemplo, por un proceso de minimización similar al descripto en las secciones anteriores. Muchas veces es deseable tener una relación entre dos número positivos y pequeños δ y ε, tal que podamos afirmar que el mejor valor de α (o en algunos casos el verdadero valor de α) está incluido en el intervalo $\alpha^* \pm \delta$ con probabilidad 1-ε, o sea:

$$P(\alpha^* - \delta \le \alpha \le \alpha^* - \delta) = 1 - \varepsilon \qquad \text{(F.22)}$$

El intervalo (α^*- δ, α^*+δ) que con probabilidad $P=1-\varepsilon$ contiene al *mejor valor* (o verdadero valor) de α se denomina *intervalo de confianza del parámetro* α. La probabilidad $P=1-\varepsilon$ se denomina *coeficiente de confianza*. Cuando ε se expresa en porcentaje ($\varepsilon\%=100*\varepsilon$) , se lo denomina el *nivel de significación.* Es importante destacar que estas definiciones no tienen aún carácter universal, pero las definiciones presentadas aquí son las adoptadas por una fracción importante de científicos y tecnólogos. Para fijar ideas imaginemos el siguiente ejemplo: supongamos que extraemos una muestra de tamaño *N* de una población que suponemos tiene una distribución normal de parámetros *m* y σ desconocidos y cuyos valores deseamos determinar a partir del análisis de la muestra. Imaginemos que el *valor medio muestral*, $\boldsymbol{<x>}$, y la *desviación estándar muestral*, $\boldsymbol{S_x}$, vienen dadas por las expresiones (2.7) y (2.8) respectivamente y son desde luego conocibles a partir de los datos muestrales. Nuestro objetivo primero es estimar el valor medio poblacional ***m*** (en este caso sí podemos hablar de verdadero valor de ***m***). Si, como supusimos, la población madre tiene una distribución normal, los estimadores $\boldsymbol{<x>}$ y $\boldsymbol{S_x}$ tienen distribuciones normales (m, $S_x/\sqrt{N-1}$) y Chi-cuadrado con ***N-1*** grados de libertad respectivamente. Entonces se cumple que la variable:

$$t = \sqrt{N-1}\cdot\left(\frac{<x>-m}{S_x}\right) \qquad \text{(F.23)}$$

tiene una distribución ***t-Student*** con ***N-1*** grados de libertad. Por lo tanto si $\boldsymbol{t_p}$ es un parámetro del intervalo de confianza asociado al coeficiente de confianza *p*, a través de la distribución de probabilidad *t-Student*, tenemos:

$$P\left(-t_p < \sqrt{N-1}\left(\frac{<x>-m}{S_x}\right) < t_p\right) = 1-\frac{p}{100} \qquad \text{(F.24)}$$

o lo que es equivalente,

$$P\left((<x>-t_p\cdot\sigma_{\bar{x}}) < m < (<x>-t_p\cdot\sigma_{\bar{x}})\right) = 1-\frac{p}{100}. \qquad \text{(F.25)}$$

donde hemos hecho uso de la relación (2.10) entre S_x y σ_x. Los valores ($<x>$ - $t_p.\sigma_x$) y ($<x>$ + $t_p.\sigma_x$) definen un *intervalo de confianza* de *100-p* para *m*. Por lo tanto si afirmamos que el mejor valor de m está comprendido en este intervalo, la probabilidad de equivocarnos cuando hacemos esta afirmación es de *p*%. Este valor de *p* se conoce con el nombre de *nivel de significación*. Cuando se trabaja con muestras grandes ($N > 30$) o con muchos grados de libertad, es útil recordar que en este caso la distribución *t-Student* se aproxima muy bien con una curva normal y en este caso la relación entre los valores críticos t_p y *p* son para $\boldsymbol{t_p}$=1, ***p***%=31.73, para $\boldsymbol{t_p}$=2, ***p***%=4.55, para $\boldsymbol{t_p}$=3, ***p***%=0.27, etc. Este último ejemplo es similar al que comúnmente encontramos en el caso de determinar el mejor valor de un parámetro que se ha medido *N* veces. También es claro que un análisis similar podría hacerse para determinar los límites de confianza de σ_x.

7 – Simulación de resultados experimentales – Método de Montecarlo

A menudo es útil simular las características de un experimento antes de llevarlo a cabo. Esto no permite por ejemplo decidir el tamaño de los errores permitidos para observar un dado efecto. La técnica de Montecarlo es un formalismo probabilístico para generar números con una distribución de probabilidad prefijada y que simulen los resultados de una variable física. Dado que una familia muy amplia de programas comerciales ya posee generadores de números aleatorios con distribuciones de probabilidad preestablecida, la tares de realizar simulaciones de Montecarlo se ha

facilitado grandemente. Para fijar ideas imaginemos que deseamos generar datos sintéticos de un experimento en el que cada medición dará como resultado la terna $(x_i, y_i, , \Delta y_i)$. Supongamos además que la relación esperada entre ***x*** e ***y*** es lineal, de la forma $y = a.x + b$. Vamos a suponer que sólo los valores de *y* tienen error (o es el error dominante del problema) con una distribución normal cuya desviación estándar esta caracterizada por un parámetro de dispersión ***disp%*** prefijado. También supondremos que los errores experimentales tendrán una distribución estadística que puede ser bien descripta por una distribución Chi-cuadrado con un número grande de grados de libertad y cuya magnitud está caracterizada por un error relativo porcentual de ***err%***. Para hacer más claro el ejemplo en consideración vamos a suponer que trabajamos con una planilla de cálculo. En la primera columna de la planilla debemos definir el rango de valores de ***x*** en los que estamos interesados. En la segunda columna, introducimos los valores de ***y*** obtenidos a través de la expresión analítica, con los valores de ***a*** y ***b*** que suponemos representativos del problema en cuestión. A estos valores de *y* lo designamos como y_{teor} ***(=a.x+b)***.En la tercera y cuarta columna calculamos los valores que van a caracterizar la dispersión de los datos(Δy_{teor} y Δy_{err}) dados por:

$$\Delta y_{teor} = y_{teor} \cdot \frac{disp\%}{100},$$
$$\Delta y_{err} = y_{teor} \cdot \frac{err\%}{100}, \qquad \text{(F.26)}$$

Estas definiciones se proponen a modo de ejemplo y en cada caso particular se pueden considerar otras caracterizaciones de la dispersiones y los errores de los datos. Seguidamente procedemos a introducir el carácter aleatorio del experimento usando el método de Montecarlo. Para ello, en dos nuevas columnas, usando la función de generación de números aleatorios de la planilla introducimos en la primera columna los números ***rnd1*** que los elegimos de modo tal que se distribuyan normalmente con media ***0*** y desviación estándar ***1***, o sea *N(0,1)*. En la otra columna introducimos los valores de los números al azar ***rdn2*** que suponemos que también tienen una distribución *N(0,1)*. Con estos valores ahora estamos en condiciones de definir los valores de los datos sintéticos para la variables y_{sint} y sus errores correspondientes Δy_{sint} definidos como:

$$y_{s\,\mathrm{int}} = y_{teor} + \Delta y_{teor} \cdot rnd1,$$
$$\Delta y_{s\,\mathrm{int}} = \frac{1}{2}\left(rnd2 + \sqrt{2 \cdot \Delta y_{teor} + 1}\right)^2 \approx \Delta y_{teor} \cdot (rnd2 + 1)^2 \qquad \text{(F.27)}$$

Los valores de *y* y Δy así obtenidos tienen las características de dispersión preestablecida (caracterizada por *disp%*) y errores caracterizados por *err%*. Esta última expresión para Δy_{sint} se basa en el hecho de que para el caso de muchos grados de libertad ($N > 30$) la distribución $\left(\sqrt{2 \cdot \chi^2} - \sqrt{\nu - 1}\right)$ es normal *N(0,1)*. Esta técnica puede generalizarse para reproducir y simular situaciones reales en forma rápida y económica.

✍ Usando la hoja de cálculos Errores_SG.xls, (que puede obtenerse de http://www.fisicarecrativa.com/) definir una función simple, por ejemplo un polinomio de segundo grado ($y=ax^2+bx+c$) de coeficientes conocidos, estos parámetros definen el modelo original. Usando el modelo de Montecarlo indicado en la hola de cálculo, generar datos "sintéticos al azar y sus correspondientes errores", de modo tal que el tamaño de los errores y la magnitud de la dispersión de los datos pueda regularse de manera controlada, por ejemplo introduciendo un valor porcentual del error medio y la dispersión media. Seguidamente, con el programa de su preferencia:

- Determinar los parámetros que mejor ajustan los datos sintéticos y sus errores. Comparar con los valores de los datos originales.

- Para cada uno de los parámetros del modelo obtenidos, realizar un gráfico de χ^2 versus el mismo. Obtener de esta figura las incertezas de cada uno de estos parámetros. Comparar con los obtenidos con el programa de ajuste usado y los valores de los parámetros del modelo original. c) Para por lo menos dos parámetros de modelo, comparar gráficamente los datos sintéticos con sus correspondientes errores con los ajustes obtenidos usando los parámetros que minimizan χ^2 y los ajustes asociados al parámetro variando su valor por su incerteza obtenida según la Fig. F.2.

Bibliografía

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

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

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

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

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

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

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