Ejemplos
Ejercicios
Misceláneas
Evaluación

Análisis avanzado
Bondad de ajuste Simulaciones

Bondad de ajuste. Intervalos de confianza. Muestras pequeñas. Simulaciones: método de Montecarlo.

3.1 – Bondad del ajuste

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

$$c_v^2 = \frac{1}{N - n_{par}} \cdot c^2 = \frac{1}{\nu} \cdot \sum_{i=1}^{N} \frac{\left(y_i - y(x_i)\right)^2}{s_i^2}, \qquad (4.1)$$

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 (s_i), nuestro modelo es una descripción adecuada de nuestras observaciones. También es claro que en este caso, según (4.1), cada uno de los términos de la sumatoria será del orden de la unidad y por lo tanto c_v^2 misma tendrá un valor cercano a uno. En otras palabras, si c_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 c_v^2 es mucho mayor que uno, el modelo no es una buena descripción de nuestros datos. Cuando $c_v^2 << 1$, 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 $c^2 = c_0^2$ haya ocurrido sólo por azar; esta probabilidad viene dada por:

$$P_{c_0^2,N} = P(c^2 \geq c_0^2) = Q\left(\frac{N - n_{par}}{2}, \frac{c_0^2}{2}\right) = \\ = CL(c_0^2) = Chi(v, c_0^2) = Dist.Chi(c_0^2, v) \quad , \tag{4.2}$$

con

$$Q(a,x) = \frac{1}{\Gamma(a)} \cdot \int_x^\infty e^{-t} \cdot t^{a-1} \cdot dt \quad , \tag{4.3}$$

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

donde ***Q(a,x)*** se conoce como la función Gama incompleta. Los ***límites de confianza*** CL(c_0^2) [usamos CL la denominaicón 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, c^2)*, su valor medio es *v* y su varianza es *2v*. *CL* representa la probabilidad de que una repetición al azar del experimento dará un valor de $c^2 \geq c_0^2$, suponiendo que 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 denota en (4.2). Por lo general se considera que un valor de $Q \geq 0.1$ es indicativo de un modelo creíble. Un valor de Q entre (0.1 y 0.001) es todavía marginalmente aceptable o indicativo de que los errores fueron subestimados. Si $Q < 0.001$ la validez del modelo debe ser revisada y su credibilidad es dudosa .

Nota: Los programas utilitarios realizan en su la mayoría 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 disposición (o puede requerirse vía e-mail a uno de los autores sgil@df.uba.ar) disponemos 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 bs pesos w_i:

$$MODE = \begin{cases} 1 & w_i = 1 \text{ no se consideran los errores} \\ 2 & w_i = 1/s_i^2 \text{ se consideran los errores} \\ 3 & w_i = 1/y_i^2 \text{ se consideran los errores } s_i^2 = y_i \\ 4 & w_i = 1/|y_i| \text{ se toma como peso la inversa de } |y_i| \end{cases} \tag{4.4}$$

3.2 – 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 (x_1, x_2,...x_N), con el objeto de determinar o estimar el valor de algún parámetro desconocido $a=a(x_1,x_2, ...x_N)$. Imaginemos que el *estimador* de este parámetro es $a^*=a^*(x_1,x_2, ...x_N)$, cuyo valor podría haberse obtenido, por ejemplo, por un proceso de minimización similar al descripto en las secciones anteriores. Muchas veces es deseable tener es la relación entre dos número positivos y pequeños d y e, tal que podamos afirmar que el mejor valor de α (o en algunos casos el verdadero valor de a) está incluido en el intervalo $\alpha^* \pm d$ con probabilidad 1-e, o sea:

$$P(a^* - d \le a \le a^* - d) = 1 - e \tag{4.5}$$

El intervalo (α^*- d, α^*+d) que con probabilidad $P=1-e$ contiene al *mejor valor* (o verdadero valor) de a se denomina *intervalo de confianza del parámetro* d. La probabilidad $P=1-e$ se denomina *coeficiente de confianza*. Cuando e se expresa en porcentaje ($e\%=100*e$), 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 s desconocidos y cuyos valores deseamos determinar a partir del análisis de la muestra. Imaginemos que el *valor medio muestral*, $<x>$, y la *desviación estándar muestral*, S_x, vienen dadas por las expresiones (1.12) y (1.13) 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 $<x>$ y 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:

$$t = \sqrt{N-1}\cdot\left(\frac{<x>-m}{S_x}\right) \tag{4.6}$$

tiene una distribución ***t-Student*** con $N-1$ grados de libertad. Por lo tanto si 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:

$$P\left(-t_p < \sqrt{N-1}\left(\frac{<x>-m}{S_x}\right) < t_p\right) = 1 - \frac{p}{100} \tag{4.7}$$

o lo que es equivalente,

$$P\left((< x > -t_p \cdot \boldsymbol{s}_{\bar{x}}) < m < (< x > -t_p \cdot \boldsymbol{s}_{\bar{x}})\right) = 1 - \frac{p}{100}. \qquad (4.8)$$

donde hemos hecho uso de la relación (1.14) entre S_x y $\boldsymbol{s}_x$. Los valores ($<x>$ - $t_p.\boldsymbol{s}_x$) y ($<x>$ + $t_p.\boldsymbol{s}_x$) 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 los que se expuso en la Sección (1.2), o sea que para $\boldsymbol{t_p}$=1, $\boldsymbol{p}$%=31.73, para $\boldsymbol{t_p}$=2, $\boldsymbol{p}$%=4.55, para $\boldsymbol{t_p}$=3, $\boldsymbol{p}$%=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 $\boldsymbol{s}_x$.

1.16 – 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, , Dy_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 Excel. 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 $\boldsymbol{y_{teor}}$ ***(=a.x+b)***.En la tercer y cuarta columna calculamos los valores que van a caracterizar la dispersión de los datos(Dy_{teor} y Dy_{err}) dados por:

$$\Delta y_{teor} = y_{teor} \cdot \frac{disp\%}{100},$$
$$\Delta y_{err} = y_{teor} \cdot \frac{err\%}{100}, \qquad (4.9)$$

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 Excel 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 Dy_{sint} definidos como:

$$y_{sint} = y_{teor} + \Delta y_{teor} \cdot rnd1,$$
$$\Delta y_{sint} = \frac{1}{2}\left(rnd2 + \sqrt{2 \cdot \Delta y_{teor} + 1}\right)^2 \approx \Delta y_{teor} \cdot \left(rnd2 + 1\right)^2 \qquad (4.10)$$

Los valores de *y* y *Dy* así obtenidos tienen las características de dispersión preestablecida (caracterizada por *disp%*) y errores caracterizados por *err%*. Esta última expresión para Dy_{sint} se basa en el hecho de que para el caso de muchos grados de libertad ($N > 30$) la distribución $\left(\sqrt{2}\,\chi^2 - \sqrt{n-1}\right)$ 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://home.ba.net/~sgil) 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 c^2 y los ajustes asociados al parámetro variando su valor por su incerteza obtenida del gráfico (1.5).

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. *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).