EL PROBLEMA DE LA INTERPOLACION EN ESTADISTICA

POR CLOTILDE A. BULA

1. Dada una sucesión discreta de valores de una variable independiente en correspondencia con los de una función desconocida, el problema de la interpolación consiste en encontrar una función contínua que tomando, para aquellos valores de la variable independiente, los valores conocidos, de la función, permita calcular valores intermedios.

La experiencia nos da, pues, correspondencias entre un número discreto de pares de valores (x_i, y_i), pero como conocidos n puntos hay infinitas funciones que pasan por ellos, es necesario elegir una.

La elección, de dicha función, tiene por fundamento: 1º) facilitar los cálculos, 2º) simplificar la expresión analítica, 3º) su presumible coincidencia con la ley que rige el hecho experimental.

En los hechos estadísticos, sobre todo en los económicos y sociales, en que la experiencia no puede repetirse a voluntad, a diferencia de lo que ocurre en otros campos, por ej. en la Física teórica, con la expresión analítica se trata en general, de tener una simple fórmula descriptiva del fenómeno y en la elección de la función prevalece, entonces, el criterio que determinan las condiciones 1ª y 2ª: facilidad en los cálculos y simplicidad de la expresión

2. Obtenida la función de interpolación, ella permite no solo calcular valores intermedios, sino también extrapolarlos, es decir hallar valores para la función fuera del campo en que la observación se realizó.

¿Qué sentido tiene la extrapolación? Indudablemente envuelve una concepción determinista: cuando la variable es el tiempo, los hechos acaecidos, dan la ley a que obedecerán los hechos futuros y también la que obedecieron los hechos pasados. En otras palabras: admitir la validez de la extrapolación es admitir que hay herencia en el fenómeno.

Pero es conveniente limitar aquellas extensiones a valores de la variable independiente cercanos a los valores conocidos, puesto que, aunque se aceptara la validez de la extrapolación, posiblemente, el concluir que los hechos pasados y futuros dependen, exclusivamente, de los hechos presentes, de la misma naturaleza, llevaría con facilidad a conclusiones erróneas.

3. Si abandonando toda posición explicativa, de la génesis y dinámica del fenómeno, se trata de obtener una expresión analítica sencilla, se está conducido, naturalmente, a escoger el polinomio.

Dentro de esta concepción, el método llamado de las ecuaciones normales o de Gauss da una solución inmediata. Si se tiene el cuadro de las observaciones:

	x	y
	x_0	y_0
	x_1	y_1
	,	.
$n+1)$	.	.
	.	.
	x_n	y_n

la fórmula de interpolación es

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

($n + 1$ datos se interpolan con una función racional y entera de grado n).

El problema se reduce a la determinación de las constantes $a_0, a_1, a_2 \ldots a_n$, que resultan de la solución del sistema lineal inhomogéneo

$$\left.\begin{aligned} \Sigma a_0 + a_1 \Sigma x_i + \ldots + a_n \Sigma x_i^n &= \Sigma y_i \\ a_0 \Sigma x_i + a_1 \Sigma x_i + \ldots + a_n \Sigma x_i^{n+1} &= \Sigma y_i x \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & \\ a_0 \Sigma x_i^n + a_1 \Sigma x_i^{n+1} + \ldots + a_n \Sigma x_i^{2n} &= \Sigma_i y_i x_i \end{aligned}\right\}$$

El cálculo se simplifica trasladando el origen a la media aritmética.

Podría plantearse un sistema de ecuaciones más sencillo en que el determinante de los coeficientes es del tipo de Vandermonde, como es sabido.

Otras fórmulas de interpolación, de uso corriente en Estadística, son las de Lagrange, Newton y de las diferencias finitas divididas.

La fórmula de Newton es de aplicación sólo cuando hay equiespaciamiento, es decir cuando el atributo es monomodular, siendo por lo tanto, las otras dos fórmulas, más generales.

La fórmula de Lagrange es:

$$y_i = a_0 (x_i - x_1)(x_i - x_2) \ldots (x_i - x_n) + a_1 (x_i - x_0)(x_i - x_2) \ldots (x_i - x_n)$$
$$+ \ldots + a_n (x_i - x_0)(x_i - x_1) \ldots (x_i - x_{n-1}) \quad (i = 0, 1, 2, \ldots, n)$$

Sustituyendo las x_i y las y_i, se tiene un sistema de ecuaciones en $a_0, a_1, \ldots, a_n$, que, una vez calculadas, permiten escribir la fórmula de interpolación que es

$$f(x) = a_0 (x - x_1)(x - x_2) \ldots (x - x_n) + a_1 (x - x_0)(x - x_2) \ldots (x - x_n) + \ldots + a_n$$
$$(x - x_0) \ldots (x - x_{m-1})$$

La fórmula de Newton nos da la función buscada en función del valor inicial y_0, y de las diferencias $\Delta y_0, \Delta^2 y_0, \ldots, \Delta^n y_0$ donde

$$\Delta y_0 = y_1 - y_0 \ , \ldots\ldots, \Delta^n y_0 = \Delta^{n-1} y_1 - \Delta^{n-1} y_0$$

La fórmula de Newton es:

$$f(x) = y_0 + \frac{y_0}{0! \, h^0} (x - x_0) + \frac{\Delta^2 y_0}{2! \, h^2} (x - x_0)(x - x_1) + \ldots$$
$$+ \frac{\Delta^n y_0}{n! \, h^n} (x - x_0)(x - x_1) \ldots (x - x_{n-1})$$

siendo h el valor del módulo, es decir, de la amplitud de los subintervalos, que debe ser constante.

Finalmente, otra fórmula de carácter general, como la de Lagrange, es la de las diferencias divididas:

$$f(x) = y_0 + (x - x_0) \mathbf{D} y_0 + (x - x_0)(x - x_1) \mathbf{D}^2 y_0 + \ldots + (x - x_0) \ldots (x - x_{n-1}) \mathbf{D}^n y_0$$

donde el valor de la función, para una x cualquiera, está dado en función del primer valor conocido y_0 y de las diferencias divididas sucesivas, en ese punto.

Las diferencias divididas se definen así:

$$\mathbf{D}\, y_0 = \frac{y_1 - y_0}{x_1 - x_0}$$

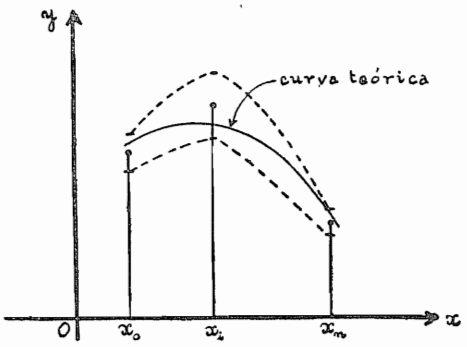
$$\mathbf{D}\, y_n = \frac{y_{n+1} - y_n}{x_{n+1} - x_n}$$

y, en general, es

$$\mathbf{D}^m y_n = \frac{\mathbf{D}^{m-1} y_{n+1} - \mathbf{D}^{m-1} y_n}{x_{m+n} - x_n}$$

4. Los hechos de la observación están afectados, en general, por errores (sistemáticos, por ej. instrumentos defectuosos; personales, por ej. defectos de percepción; accidentales, por ej. influencias atmosféricas). La función que interpola, es decir que pasa por los puntos conocidos, ¿puede decirse por eso que da el *verdadero valor?* Si se admite la existencia de errores —y hay que admitirla— es indudable que no.

Suficiente es, pues, que la función teórica pase por puntos próximos, en forma tal que llamando δ_i a los desvíos será: $\delta_i = f\ (x_i) - y_i$ y ellos deberán estar contenidos en un cierto entorno de los puntos conocidos, es decir, la curva teórica pasará dentro de una franja de tolerancia.



Esto permite simplificar los cálculos sin perder exactitud y el problema de la interpolación se convierte, de esta manera, en el del ajustamiento o perecuación como dicen los italianos.

5. Si en el problema del ajustamiento, abandonando todo propósito explicativo, nos proponemos dar la expresión analítica, más sencilla, del hecho experimental, nos encontramos, igual que en el caso de la interpolación, con que el polinomio es la función contínua que satisface tal propósito.

Si las observaciones son en número de $(n + 1)$, con un polinomio de grado n se interpola y con uno de grado menor se ajusta. Otra vez, es de aplicación el método de las ecuaciones normales o de Gauss y, en este caso, los coeficientes del polinomio de ajustamiento, hacen mínimo el error complexivo cuadrático.

Si el polinomio de ajustamiento fuera

$$f(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m \qquad (m \leq n)$$

en que los coeficientes a_j $(j = 0, 1, 2, \ldots, m)$, han sido determinados mediante la solución de aquel sistema de ecuaciones, es

$$\sigma^2 (a_j) := \sum_i [f(x_i) - y_i]^2 = \text{mínimo}$$

y como esta medida de la dispersión es función de aquellos coeficientes, si la condición de mínimo se cumple, es

$$\frac{\partial \sigma}{\partial a_0} = \frac{\partial \sigma}{\partial a_1} = \ldots = \frac{\partial \sigma}{\partial a_m} = 0$$

como es fácil demostrarlo.

El cálculo se simplifica trasladando el origen a la media aritmética.

Tiene este método el inconveniente de que si la aproximación no se juzgase satisfactoria, para mejorarla, es decir, para aumentar el grado de la función de aproximación, hay que rehacer íntegramente los cálculos. Lo hecho no se aprovecha.

6. Para obviar aquellos inconvenientes se introducen, en Estadística, los desarrollos en serie de funciones ortogonales.

La ortogonalidad es propiedad de los sistemas de funciones ordenadas y un sistema se llama ortogonal, cuando la integral del producto de dos de ellas (la suma en el caso discontínuo) es nula si son de distinto rango y no lo es en el caso contrario. Entonces si un sistema $[f_n(x_i)]$ de la variable discontínua x, es ortogonal, esta propiedad se suele indicar así:

$$\sum f_n(x_i) f_m(x_i) = E_{m|n}$$

Por lo tanto, con el símbolo $E_{m|n}$ se expresa $E_{m|n} = 0$, si es m distinto de n y un número positivo si es $m = n$.

Las series trigonométricas proporcionan un primer ejemplo de tales funciones, pues es conocida la propiedad de ortogonalidad que, en el intervalo $(0, 2\pi)$, tienen los senos y cosenos múltiplos (*).

En Estadística, los polinomios trigonométricos se prestan admirablemente, para ajustar algunos fenómenos periódicos. Así, por ej., observadas durante 30 años las alturas del río Paraná, hechos los premios mensuales se tiene.

(*) Euler en una memoria "Sur les inégalités du mouvement de Jupiter et de Saturne" (1748) utilizó, parece que por vez primera, desarrollos en serie de cosenos. También lo hizo Daniel Bernoulli en otra memoria, "Sur les cordes vibrantes" (1753), pero, como lo dice Riemann, (Oeuvres Mathématiques de Riemann-Trad. Laugel-pág. 233) fué Fourier (Théorie analytique de la Chaleur-1807) el primero que comprendió de una manera exacta y completa, la naturaleza de las series trigonométricas.

MESES	PROMEDIOS MENSUALES	MESES	PROMEDIOS MENSUALES
Enero	2.98	Julio	2.94
Febrero	3.63	Agosto	2.20
Marzo	3.92	Sbre.	1.85
Abril	4.00	Obre.	2.16
Mayo	3.63	Nbre.	2.60
Junio	3.23	Dbre.	2.63

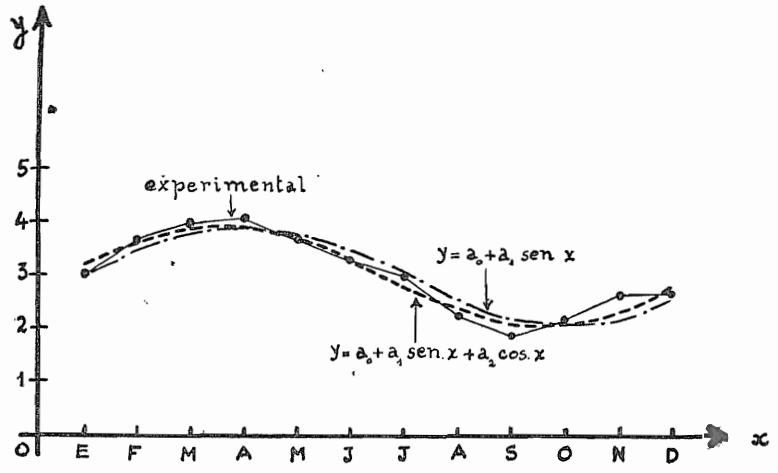
Ajustando con funciones de las formas:

$$y = a_0 + a_1, \operatorname{sen} x$$

e

$$y = a_0 + a_1 \operatorname{sen} x + b_1 \cos x$$

resulta



que como se ve, interpretan muy bien el hecho experimental

Pero es con los polinomios ortogonales, cómo se aborda, con toda generalidad, el problema del ajustamiento, en Estadística, dentro de la concepción descriptiva, utilizando funciones sencillas.

Entonces, si $[P_j\ (x_i)]$ es un sistema de polinomios ortogonales de la variable discontínua x, se verifica.

$$\Sigma\, P_j\ (x_i)\ P_k\ (x_i) = E_{m|n}$$

donde el subíndice indica el grado del polinomio, que, en el caso de una variable, coincide con su rango, siendo

$$P_k\ (x) = \alpha_{k|0} + \alpha_{k|1}\, x + \alpha_{k|2}\, x^2 + \ldots + \alpha_{k|k}\, x^k$$

Para una aproximación de grado m, la función teórica adopta esta forma

$$f(x) = a_0 P_0(x) + a_1 P_1(x) + \ldots + a_m P_m(x)$$

Debido a la propiedad de ortogonalidad, el cálculo de los coficientes a_j se hace por el clásico procedimiento de Fourier, siendo

$$a_j = \frac{\sum_i f(x_i)\, P_j(x_i)}{\sum_i P_i^2(x_i)}$$

expresión que se simplifica normalizando el sistema, en cuyo caso el denominador vale la unidad.

Los coeficientes de tal manera calculados, hacen mínimo el error complexivo cuadrático, es decir que

$$\sigma_m^2 = \sum_i [y_i - f(x_i)]^2 = \text{mínimo}$$

por lo cual es

$$\frac{\partial \sigma}{\partial a_0} = \frac{\partial \sigma}{\partial a_1} = \ldots = \frac{\partial \sigma}{\partial a_m} = 0$$

como se demuestra fácilmente.

Si ahora la aproximación no fuera satisfactoria, se pueden agregar términos, pero todo lo hecho se conserva y el trabajo se reduce a calcular, sólo, los nuevos coeficientes.

A estos polinomios se los puede generar, siempre, por determinantes y si se impone la condición de que el coeficiente del término máximo sea la unidad positiva, es

$$P_n(x) = \frac{(-1)^n}{\Delta_n} \begin{vmatrix} 1 & x & x^2 & \ldots & x^n \\ \mu_0 & \mu_1 & \mu_2 & \ldots & \mu_n \\ \mu_1 & \mu_2 & \mu_3 & \ldots & \mu_{n+1} \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \\ \mu_{n-1} & \mu_n & \mu_{n+1} & \ldots & \mu_{2n-1} \end{vmatrix}$$

donde es

$$\mu_n = \sum_i x_i^n$$

y Δ_n es el menor complementario de x^n.

Un método más breve, para generarlos, se tiene aplicando la fórmula de Gram-Romanovsky, que es la que se utiliza en las aplicaciones estadísticas.

Este procedimiento se basa en la propiedad que tienen los polinomios ortogonales de poder ser expresados como una combinación lineal de todos los anteriores más una potencia de la variable del grado que indica el subíndice del polinomio, es decir que

$$P_s(x) = \beta_{s|0} P_0(x) + \beta_{s|1} P_1(x) + \ldots + x^s \qquad [3]$$

El problema que se plantea ahora es el de determinar los coeficientes β,

lo que se hace aprovechando la propiedad de ortogonalidad de dichos polinomios y resulta

$$\beta_{s|j} = \frac{\sum_i P_j x_i^s}{\sum_i P_j^2}$$

y reemplazando en [3] estos valores se tiene que

$$P_s(x) = -\sum_{j=0}^{s-1} \frac{\sum_{i=0}^{n} P_j(\xi_i)\,\xi^s}{\sum_{i=0}^{n} P_j^2(\xi_i)} \quad P_j(\xi_i) + x^s$$

donde se ha cambiado la variable de sumación interna para evitar confusiones.

7. En Estadística, cuando hablamos de frecuencias, nos referimos siempre a frecuencias relativas y si con η_i simbolizamos las que corresponden a x_i, se verifica que es

$$\eta > 0$$

$$\sum_i \eta_i = 1$$

La solución ideal de nuestro problema la daría una fórmula finita que respondiendo a un esquema de probabilidad, a la vez que describiera el hecho experimental lo explicara.

Es problema este que, no está resuelto, ni aún para este caso de una variable que es el más sencillo.

Pearson ha construido un utilísimo repertorio de funciones que definen las curvas que llevan su nombre y que se adaptan, admirablemente, para ajustar las observaciones experimentales.

El método se apoya en el esquema de las pruebas repetidas sin reposición que, como se sabe, define un polinomio hipergeométrico.

Partiendo de este esquema, se llega a plantear una ecuación, en diferencias finitas, dada en forma de función racional, con una expresión de primer grado en el numerador y otra de segundo en el denominador.

Como en el paso al límite, resulta una curva simétrica, la función pierde su característico contenido del esquema sin reposición que define curvas asimétricas. Pearson abandona el proceso analítico y utilizando la forma de la ecuación en diferencias finitas, plantea una ecuación diferencial de la misma forma, poniendo

$$\frac{y'}{y} = \frac{a - x}{b_0 + b_1 x + b_2 x^2} \qquad [4]$$

que resuelve por el llamado *método de los momentos*.

Se postula el anulamiento extremo, es decir que, si el intervalo es el (α, β), es

$$f(\alpha) = f(\beta) = 0 \qquad [5]$$

De la [4] resulta

$$y'(b_0 + b_1 x + b_2 x^2) = y(a - x)$$

multiplicando ambos miembros por x^n e integrando en (α, β)

$$\int_{\alpha}^{\beta} (b_0 + b_1 x + b_2 x^2)\ x^n\ y'\ d\,x = \int_{\beta}^{\alpha} (a - x)\ x^n\ y\ d\,x$$

Integrando por partes en el primer miembro, teniendo presente la condición [5] y poniendo

$$\mu' = \int_{\alpha}^{\beta} y\ x^n\ d\,x$$

que es lo que se llama un momento de orden n resulta

$$-n\ b_0 \mu'_{n-1} - (n+1)\ b_1\ \mu_n' - (n+2)\ b_2\ \mu'_{n+1} = a\ \mu'_n - \mu'_{n+1}$$

de donde, una sencilla fórmula de recurrencia que liga los momentos de tres órdenes consecutivos

$$[(n+2)\ b_2 + 1]\ \mu'_{n+1} = [a + (n+1)\ b_1]\ \mu'_n + n\ b_0\ \mu'_{n-1}$$

Variando n se plantea un sistema de ecuaciones y obtiene Pearson el repertorio de funciones a que hemos hecho referencia antes. En [4] a es el punto de máximo de la función. Se trata de curvas campanulares, excepto los tipos especiales que denomina "J-shaped" y "U-shaped".

El método se encuentra expuesto en la obra del actuario inglés W. P. Elderton "Frequency curves and Correlation", exposición, si se quiere, poco clara que, en un próximo trabajo, nos proponemos desarrollar en forma de hacerla accesible aún para los técnicos que, sin mayor bagaje matemático, deban efectuar aplicaciones.

BIBLIOGRAFIA

Bowley A.—Elements of Statistics. London, 1296-5ª ed.

Darmois Georges.—Statistique Mathématique. París, 1927.

Elderton W. Palin.—Frequency curves and Correlation. London, 1927.

Jackson Dunham.—The theory of approximation. (American Mathematical Society Colloquium Publications, Volume XI. New York, 1930).

Jordan Charles.—Statistique Mathématique. Paris, 1927.

Rietz H. L.—Handbook of Mathematical Statistics. New York, 1924.

Risser R. et Traynard C. E.—Statistique Mathématique. Paris, 1933.