

C.N.E. Biblioteca	
ARCHIVO PUBLICACIONES	
Nº 1	AÑO 1970

01.70.08

METODO GENERAL PARA AJUSTAR PARAMETROS DE FUNCIONES,

CUANDO TODAS SUS VARIABLES TIENEN ERROR.

Oscar José Bressan



Noviembre 1970

Bariloche - Argentina

METODO GENERAL PARA AJUSTAR PARAMETROS DE
FUNCIONES, CUANDO TODAS SUS VARIABLES
TIENEN ERROR

Oscar José Bressan

Centro Atómico Bariloche e Instituto de Física Balseiro,
Comisión Nacional de Energía Atómica y
Universidad Nacional de Cuyo - Bariloche

RESUMEN.-

Se generaliza el problema de cuadrados mínimos, para el caso de que todas las variables tengan error, llegando a un método general para ajustar parámetros de cualquier tipo de función. Se particulariza al caso lineal.

Se demuestra que el método es asintóticamente consistente y eficiente. Se encuentran intervalos de confianza de los parámetros y se sugiere una alternativa práctica para simplificar este problema.

Luego se estudian métodos estadísticos para verificar la bondad de la elección de la función, y se propone una dócima para detectar correlaciones locales en los errores.

Finalmente se discute como resolver los casos particulares donde las hipótesis hechas no se satisfacen.

Noviembre 1970

Bariloche - Argentina

I N D I C E

CAP. I	-	INTRODUCCION		
A	-	Inferencia Estadística.	Pág.	1
B	-	Método de cuadrados mínimos	"	1
CAP. II	-	METODO DE LAS DISTANCIAS MINIMAS		
A	-	Restricción del método de cuadrados mínimos	"	3
B	-	Planteo del método de distancias mínimas.	"	3
C	-	Un caso particular: La recta.	"	6
D	-	Otro caso particular: El hiperplano	"	9
CAP. III	-	CONSISTENCIA Y EFICIENCIA DEL METODO		
A	-	Función de máxima verosimilitud de Fisher	"	11
B	-	Estadística de la variable aleatoria "Distancia"	"	12
C	-	Aplicación de la función de máxima verosimilitud.	"	15
CAP. IV	-	INTERVALOS DE CONFIANZA		
A	-	Definición.	"	17
B	-	Varianza de los parámetros.	"	17
C	-	Intervalo de confianza de los parámetros.	"	19
D	-	Un método alternativo aproximado.	"	20
CAP. V	-	VERIFICACION DE HIPOTESIS		
A	-	Método de JI cuadrado.	"	22
B	-	Aplicación a la distancia mínima.	"	23
C	-	Docina para correlaciones locales	"	23
CAP. VI	-	NOTAS		
A	-	Discusión sobre la varianza de las coordenadas.	"	25
B	-	Función lineal: Diferencia entre M.C.M. y M.D.M.	"	26
C	-	Caso especial cuando en vez de una superficie se debe ajustar una curva	"	26
D	-	Requisitos de la función a ajustar	"	27

I - INTRODUCCIONA - INFERENCIA ESTADISTICA.

En lo que sigue, trataremos de problemas de interpretación de observaciones y mediciones que tienen una determinada estadística. Este tema se lo identifica como INFERENCIA ESTADISTICA. En esta pueden distinguirse tres etapas:

- 1º) Especificación: Consiste en la elección de la forma matemática de la ley estadística a seguir. Solo en pocos casos puede conocerse exactamente. Pero formulando ciertas hipótesis, con verificación a posteriori, a veces puede llegarse a un buen conocimiento de esa ley. Por ejemplo, la hipótesis básica de que el error cometido con una medición es el resultado de un número muy grande de pequeños errores independientes entre sí, tales que cada uno de ellos no contribuye más que a una pequeña parte del resultado, permite especificar como normal (gaussiana) a la distribución de errores cuando el número de observaciones es muy grande. Y el problema de la inferencia estadística es, en este caso, determinar a partir de los mismos errores, a los dos coeficientes (m , σ) que caracterizan totalmente a esa distribución.
- 2º) Estimación: Comprende la elección de los métodos de cálculo para estimar los parámetros de la ley estadística. El método de los cuadrados mínimos de Gauss, la regla χ^2 de Pearson, la función de máxima verosimilitud de Fisher, son ejemplos de estimadores. Según Fisher, el método que deberá emplearse para obtener los parámetros deseados debe satisfacer los siguientes requisitos:
 - a) - Consistencia: para un número elevado de observaciones, la ley de probabilidades del estimador debe concentrarse alrededor del verdadero valor del parámetro.
 - b) - Eficiencia: De todos los estimadores posibles corresponde elegir el de varianza mínima.
- 3º) Distribución: Incluye la discusión de cualquier función de cantidades cuya distribución es conocida, y en particular la verificación de hipótesis, que consiste en aceptarla o rechazarla en función de la misma distribución. Este es un problema muy delicado, aún hoy materia de controversia, y que en general puede llevar a alguno de los siguientes errores:
 - a) - Rechazar la hipótesis cuando esta es cierta (error de primera clase).
 - b) - Aceptar la hipótesis cuando esta es falsa (error de segunda clase).
 Como es obvio, estos dos problemas están intimamente ligados, y para su separación existen varios métodos, pero uno de los más efectivo es el empleo de χ^2 .

B - METODO DE CUADRADOS MINIMOS

El método de cuadrados mínimos (M.C.M.), para el caso más general, puede enunciarse como:

- a) Se tiene un conjunto de n puntos experimentales de un espacio de $R + 1$ dimensiones

$$(y_i, x_{1i}, x_{2i}, \dots, x_{Ri}) \quad I-1$$

donde el subíndice i va de 1 a n .

- b) Por razones teóricas o empíricas se supone que los puntos debieran satisfacer a la función de P variables independientes y P constantes indeterminadas:

$$y_i = f(x_{1i}, x_{2i}, \dots, x_{Pi}, k_1, k_2, \dots, k_p) \quad I-2$$

donde P es menor que n .

Generalmente se encuentra que el sistema es incompatible (lo que interesa es conocer las P constantes k a través del resultado muestral) debido a errores de dos tipos. Unos son los errores en la determinación de los puntos. Generalmente estos errores se toman como normales aunque si n no es muy grande lo riguroso sería considerarlos con distribución t -Student. Pero podrían eventualmente tener cualquier otra distribución. Es necesario destacar que si la determinación de los puntos se hace por métodos experimentales, cada una de sus coordenadas se verá afectada por un cierto error, pero el M.C.M. exige la siguiente hipótesis restrictiva:

- c) Todo el error está en la determinación de la coordenada dependiente, o por lo menos el error de las demás coordenadas es despreciable frente a aquel.

Además están los errores debidos a una mala elección de la función I-2. Su discusión es delicada y lo dejaremos para un capítulo posterior, que será la verificación de hipótesis. De cualquier modo el M.C.M. estipula:

- d) No existen errores de este segundo tipo.

Por la hipótesis c) la igualdad I-2 deberá expresarse en rigor como

$$\epsilon_i = y_i - f(x_{1i}, \dots, k_1, \dots) \quad I-3$$

y el método consiste en HACER MÍNIMA A LA SUMATORIA DE LOS CUADRADOS DE LOS ERRORES, CON RESPECTO A CADA UNA DE LAS CONSTANTES A DETERMINAR:

$$\left\{ \begin{array}{l} \frac{\partial \sum_{i=1}^n \epsilon_i^2}{\partial k_j} = 0 \quad j = 1, \dots, P \end{array} \right. \quad I-4$$

Esto significa resolver un sistema de P ecuaciones con P incógnitas (no siempre lineal) y eventualmente queda por estudiar el hessiano a fin de discutir la existencia de mínimo. También puede darse que existan varios mínimos relativos y cualquier clase de patología, originada en la forma particular de la función elegida I-2.

Es de destacar que I-4 es equivalente a hacer mínima a la varianza del error, lo que implica que es, por definición, el estimador más eficiente.

Cap. II - METODO DE LAS DISTANCIAS MINIMAS.

A - RESTRICCION DEL METODO DE CUADRADOS MINIMOS.

Como ya señalamos, la hipótesis c) del M.C.M. es desde el punto de vista teórico, una limitación importante. Desde el punto de vista físico, es fundamental, ya que siempre se tiene certeza que en mayor o menor grado, todas las determinaciones experimentales siempre tienen error, que afectan a todas las coordenadas de cada punto.

En el caso que la función que se propone sea lineal, esta limitación se puede salvar en parte, tomando como variable dependiente a la que tenga más error. Por supuesto que el método falla si son más de una las variables que tienen error grande y además equivalente. Cuando la función que se propone no es lineal, y los mayores errores están en la variable independiente, no siempre puede ser despejada esta para invertir la función. En casos tan simples y comunes como un polinomio, o una suma de exponenciales, la inversión es imposible.

Es evidente que en primera aproximación el error de usar la hipótesis c) depende de la derivada de la función elegida (I-2) en el entorno del punto experimental. Si la derivada es pequeña con respecto a la variable x_1 , la determinación del error según el eje paralelo a "v" (ϵ_y) será casi independiente del error x_1 . Pero por el contrario, si la derivada es grande o tiende a infinito, un pequeño error en la determinación de x_1 influirá grandemente en ϵ_y , que al considerar su cuadrado, agravará más aún la importancia de este punto en la sumatoria que debe minimizarse.

Puede llegarse también a que ni siquiera quede definido ese error, o pese infinitamente. Por ejemplo, si la función es:

$$y = k_1 \cdot \arccos(k_2 \cdot x) \quad \text{II-1}$$

puede calcularse k_1 y k_2 con n puntos experimentales.

Supongamos que los valores calculados de las constantes sean A y B respectivamente. Entonces incluimos solo un punto más,

$$(x_{n+1}, y_{n+1}) \quad \text{donde} \quad x_{n+1} = (1/B) + g^2 \quad \text{y} \quad y_{n+1} = 0$$

Al calcular nuevamente por cuadrados mínimos k_1 y k_2 el último punto o no queda definido o pesa infinitamente, situación salvable obligando a que el nuevo k_2 valga ahora

$$1/((1/B) + g^2).$$

En consecuencia, por grande que sea n , el último punto altera las constantes.

B - PLANTEO DEL METODO DE DISTANCIAS MINIMAS.

Los inconvenientes planteados se solucionan si consideramos los errores en todas las coordenadas, como es obvio. Y una forma de hacerlo es considerar la distancia del punto a la función y trabajar con esa distancia.

El planteo del método (que para abreviar llamaremos M.D.M.) en su forma más general es el siguiente:

- $$(y_i, x_{1i}, x_{2i}, \dots, x_{Ri}) \quad \text{II-2}$$

b) Por razones teóricas o empíricas se supone que los puntos de a) satisfacen la función:

$$y = F(x_1, x_2, \dots, x_R, k_1, k_2, \dots, k_p) \quad \text{II-3}$$

c) $y_i, x_{1i}, x_{2i}, \dots, x_{Ri}$ son $R + 1$ variables aleatorias normales, todas con igual varianza σ . Con un sencillo cambio de variables, esto se satisface siempre, a menos que alguna variable no tuviera error, pero en este caso también es salvable. Ver nota A sobre una discusión de esta hipótesis c).

- d) No existe error debido a una mala elección de la función II-3. Esta hipótesis elimina la posibilidad de un error sistemático.

Dentro de la idea general del M.C.M., usaremos como estimador la condición de hacer mínima con respecto a los P parámetros k , a la sumatoria de los cuadrados de la distancia (d) de cada punto experimental a la función. Dicho en otras palabras: La distancia real entre un punto experimental y el correspondiente sin error (o sea el "real") es:

$$(d_r)^2 = (y_i - F(X_{1i}, \dots, k_1, \dots))^2 + \sum_{j=1}^R (x_{ji} - X_{ji})^2 \quad \text{II-4}$$

donde la variable con minúscula indica la coordenada medida y con mayúscula son los valores "reales". Pero como estas no pueden conocerse, la II-4 es solo una definición teórica.

Parece razonable sustituir las coordenadas "reales" por las que correspondan al punto que estando sobre la función II-3, tengan distancia mínima con respecto al punto experimental. Es decir, definir la distancia mínima:

$$(d_i)^2 = (y_i - F(x'_{1i}, \dots, k_1, \dots))^2 + \sum_{j=1}^R (x_{ji} - x'_{ji})^2 \quad \text{II-5}$$

donde los x'_{ji} se obtienen resolviendo el sistema de R ecuaciones con R incógnitas:

$$\left\{ \begin{array}{l} \frac{\partial d_i}{\partial x_1'} = \frac{\partial (d_i^2)}{\partial x_1'} = 0 \\ \dots\dots\dots \\ \frac{\partial d_i}{\partial x_R'} = \frac{\partial (d_i^2)}{\partial x_R'} = 0 \end{array} \right.$$

y lo que se busca es que la sumatoria:

$$\sum_{i=1}^n (d_i)^2 = \sum_{i=1}^n \left[(y_i - F(x'_{1i}, \dots, k_1, \dots))^2 + \sum_{j=1}^R (x_{ji} - x'_{ji})^2 \right]$$

II-7

sea mínima con respecto a los parámetros k , condición que implica resolver el sistema de P ecuaciones con P incógnitas:

$$\left\{ \begin{array}{l} \frac{\partial \sum_{i=1}^n (d_i)^2}{\partial k_1} = 0 \\ \dots \dots \dots \\ \frac{\partial \sum_{i=1}^n (d_i)^2}{\partial k_p} = 0 \end{array} \right. \quad \text{II-8}$$

Es evidente que el método puede llegar a ser bastante complicado, según sea II-3, y enseguida puede caerse, al tratar de calcularlo, en un círculo vicioso: El sistema II-6 no puede resolverse antes de conocer los coeficientes k_j , y para conocer los k_j habría que resolver el sistema II-8, para lo cual hay que conocer a las x'_{ji} que son el resultado de II-6.

Siempre queda el procedimiento heroico de resolverlo en forma iterativa, y así pueden obtenerse los parámetros con la precisión que se desea. Con las facilidades de una computadora, este es el procedimiento más expeditivo.

Hay que acotar, que tanto II-6 como II-8 en general van a dar como solución múltiples extremos, que no tienen porque ser mínimos. Debe estudiarse en cada caso el Hessiano, y elegir entre los distintos mínimos relativos al menor.

En un caso particular (que por supuesto es el lineal), el método toma una forma sencilla, haciendo uso del cálculo vectorial. Lo veremos seguidamente.

Para introducir la distancia II-5, usamos la expresión "parece razonable". Más adelante, daremos una justificación rigurosa del método, al demostrar su consistencia y eficiencia.

C - UN CASO PARTICULAR: LA RECTA

Para el caso particular en que II-3 es:

$$y = m x + b \quad \text{II-9}$$

(o sea la clásica recta) la distancia entre un punto y II-9 se expresa como:

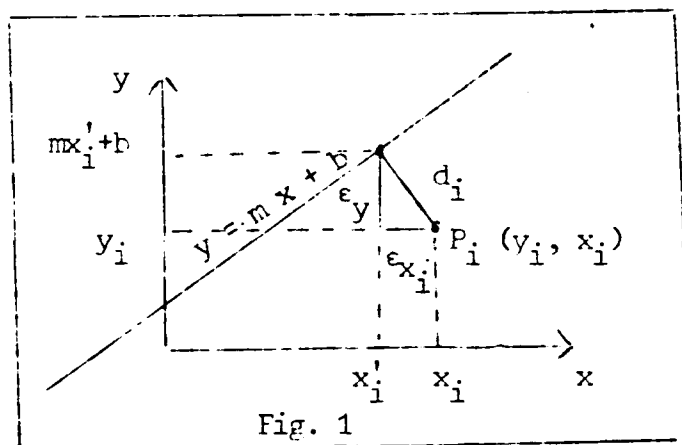
$$d_i^2 = (y_i - m x_i' - b)^2 + (x_i - x_i')^2 \quad \text{II-10}$$

y si hacemos el cambio de variables:

$$\begin{aligned} y_i - m x_i' - b &= \epsilon_{y_i} \\ x_i - x_i' &= \epsilon_{x_i} \end{aligned} \quad \text{II.11}$$

podemos escribir (ver fig. 1)

$$d_i^2 = \epsilon_{y_i}^2 + \epsilon_{x_i}^2 \quad \text{II-12}$$



Ahora imponemos que los d_i de II-12 sean mínimas con respecto a las x_i' . Por esta condición todos los d_i serán normales a la recta y paralelos entre si. O sea:

$$\frac{\epsilon_{y_i}}{\epsilon_{y_p}} = \frac{\epsilon_{x_i}}{\epsilon_{x_p}} \quad \therefore \quad \epsilon_{y_i} = K \quad \epsilon_{x_i} \quad \text{II.13}$$

y es trivial que esta constante K vale:

$$K = - \frac{1}{m} \quad \epsilon_{x_i} = - m \quad \epsilon_{y_i} \quad \text{II.14}$$

Y la II-12 puede escribirse ahora como:

$$d_i^2 = \epsilon_{y_i} + \epsilon_{x_i} = \epsilon_{y_j} + m^2 \epsilon_{y_j} = (1+m^2) \epsilon_{y_i} \quad \text{II-15}$$

donde en II-15 ya está implícito que es mínimo con respecto a x_i' .

Si ahora le sacamos la raíz a II-15 y hacemos la convención de tomar el signo de d_i igual al de ϵ_{y_i}

$$d_i = \sqrt{1+m^2} \epsilon_{y_i} \quad \text{II-16}$$

podremos expresar exactamente la ec. II-9 como:

$$\begin{aligned} y_i - \epsilon_{y_i} &= m x_i' + b \\ &= m (x_i - \epsilon_{x_i}) + b \\ &= m x_i - m \epsilon_{x_i} + b \\ &= m x_i + m^2 \epsilon_{y_i} + b \end{aligned}$$

$$\boxed{y_i - d_i / \sqrt{1+m^2} = m x_i + b + m^2 \cdot d_i / \sqrt{1+m^2}} \quad \text{II-17}$$

Y como i vale de 1 a n , son n ecuaciones II-17 independientes, que pueden ser interpretadas como las n componentes de un vector de n dimensiones:

$$y - \frac{\vec{d}}{\sqrt{1+m^2}} = m \left(x + m \frac{\vec{d}}{\sqrt{1+m^2}} \right) + b (1) \quad \text{II-18}$$

donde el vector (1) es el que tiene las n componentes iguales a 1. Desde el punto de vista geométrico, los dos vectores del miembro de la derecha de II-18 nos determinan un plano, de dos dimensiones en un espacio de n , que es justamente el II-9 cuando $d \rightarrow 0$; y otro vector de ese mismo plano es el del miembro de la izquierda. Y por supuesto, nos interesa conocer a los coeficientes m y b , que nos permiten expresar la combinación lineal II-18.

De la ecuación II-18 podemos despejar a la distancia:

$$\vec{y} - m \vec{x} - b (1) = \sqrt{1+m^2} \vec{d} \quad \text{II-19}$$

y esta será mínima, cuando $\vec{d}$ (como vector en el espacio de n dimensiones) sea normal al plano definido por la ec. II-18. Esto es totalmente equivalente a imponer la condición II-8. En cálculo vectorial, lo que se necesita es que el producto escalar de cada vector de II-18 por $\vec{d}$, o lo que es lo mismo, por el miembro de la izquierda de II-19, sea nulo. De donde se obtiene el sistema:

$$\langle y, y \rangle - m \langle y, x \rangle - b \langle y, 1 \rangle - \langle d, d \rangle = 0$$

$$\langle y, x \rangle - m \langle x, x \rangle - b \langle x, 1 \rangle + m \langle d, d \rangle = 0 \quad \text{II-20}$$

$$\langle y, 1 \rangle - m \langle x, 1 \rangle - b \langle 1, 1 \rangle = 0$$

que son tres ecuaciones con tres incógnitas: m , b y $\langle d, d \rangle = \sum_{i=1}^n d_i^2$ y aunque el sistema II-20 no es lineal, igualmente se resuelve fácilmente, como veremos a continuación en este ejemplo sencillo:

a) Sean los puntos sobre el plano: (ver fig. 2)

i	y_i	x_i
1	0	-1
2	1	1
3	2	0
4	3	2

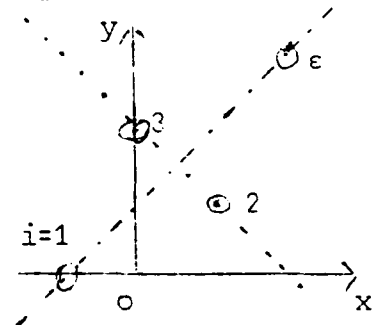


Fig. 2

- b) Suponemos que esos puntos corresponden a la observación experimental de una recta teórica: $y = m x + b$
- c) Tanto x como y tienen igual dispersión experimental (es decir, tienen igual varianza). Esto es un dato adicional que se da al problema, porque no puede sacarse como consecuencia de los mismos puntos. Ver discusión en VI- Nota A.
- d) Suponemos que la función elegida en b) es exacta (y no, por ejemplo, una aproximación de una forma más complicada).

Entonces ahora calculamos los productos escalares:

$$\langle y, y \rangle = \sum_{i=1}^n y_i \cdot y_i = 14$$

$$\langle y, x \rangle = 7$$

$$\langle y, 1 \rangle = 6$$

$$\langle x, x \rangle = 6$$

$$\langle x, 1 \rangle = 2$$

$$\langle 1, 1 \rangle = 4$$

con los cuales, reemplazando, resolveremos II-20. El sistema nos queda

$$14 - 7.m - 6.b - \langle d, d \rangle = 0 \quad (A)$$

$$7 - 6.m - 2.b + m. \langle d, d \rangle = 0 \quad (B)$$

$$6 - 2.m - 4.b = 0 \quad (C)$$

Si hacemos 2. (A) - 3. (C) y 2. (B) - (C) nos queda:

$$10 - 8.m - 2 \langle d, d \rangle = 0 \quad (A')$$

$$8 - 10.m + 2.m \langle d, d \rangle = 0 \quad (B')$$

y si ahora sumamos m. (A') + (B'):

$$- 8. m^2 + 8 = 0 \quad m = \pm 1$$

que generan dos conjuntos de soluciones:

$$m_1 = 1 \quad b_1 = 21 \quad \langle d, d \rangle = \sum_{i=1}^n d_i^2 = 1$$

$$m_2 = -1 \quad b_2 = 2 \quad \langle d, d \rangle = 9$$

y que corresponden a dos extremos que toma $\sum d_i^2$. La 1a. es el mínimo (graficada con - . - . en fig. 2) y es la solución que nos interesa.

D - OTRO CASO PARTICULAR: EL HIPEERPLANO

Es el caso general al tener el problema lineal. Como es totalmente similar al anterior, haremos referencia a las ecuaciones halladas para el caso de la recta, y pondremos sin discusión las que corresponden en este caso.

$$(II-9) \quad y = b + \sum_j^R a_j x_j \quad II-21$$

$$(II-10) \quad d_i^2 = (y_i - b - \sum_{j=1}^R a_j x_{ji})^2 + \sum_{j=1}^P (x_{ji} - x'_{ji})^2 \quad II-22$$

$$(II-11) \quad \left. \begin{aligned} y_i - b - \sum_{j=1}^R a_j x_{ji} &= \epsilon_{y_i} \\ x_{ji} - x'_{ji} &= \epsilon_{x_{ji}} \end{aligned} \right\} \quad II-23$$

$$(II-12) \quad d_i^2 = \epsilon_{y_i}^2 + \sum_{j=1}^R \epsilon_{x_{ji}}^2 \quad II-24$$

$$(II-13) \quad \frac{\epsilon_{y_i}}{\epsilon_{y_m}} = \frac{\epsilon_{x_{ji}}}{\epsilon_{x_{jm}}} \dots \epsilon_{y_i} = K_j \cdot \epsilon_{x_{ji}} \quad II-25$$

$$(II-14) \quad K_j = -1/a_j \quad \epsilon_{x_{ji}} = -a_j \cdot \epsilon_{y_i} \quad II-26$$

$$(II-15) \quad d_i^2 = \epsilon_{y_i}^2 + \sum_{j=1}^R \epsilon_{x_{ji}}^2 = \epsilon_{y_i}^2 + \epsilon_{y_i}^2 \sum_{j=1}^R a_j^2 = (1 + \sum_{j=1}^R a_j^2) \epsilon_{y_i}^2 \quad II-27$$

$$(II-16) \quad d_i = \sqrt{1 + \sum_{j=1}^R a_j^2} \cdot \epsilon_{y_i} \quad II-28$$

$$(II-17) \quad y_i - \frac{d_i}{\sqrt{1 + \sum_{j=1}^R a_j^2}} = b + \sum_{j=1}^R a_j x_{ji} + \frac{d_i \sum_{i=1}^R a_j^2}{\sqrt{1 + \sum_{j=1}^R a_j^2}} \quad II-29$$

$$(II-18) \quad \overrightarrow{y} - \frac{\overrightarrow{d}}{\sqrt{1 + \sum_{j=1}^R a_j^2}} = b(1) + \sum_{j=1}^R a_j \left(\overrightarrow{x_j} + \frac{a_j \cdot \overrightarrow{d}}{\sqrt{1 + \sum_{j=1}^R a_j^2}} \right) \quad II-30$$

$$(II-19) \quad \overrightarrow{y} - b(1) - \sum_{j=1}^R a_j \overrightarrow{x_j} = \sqrt{1 + \sum_{j=1}^R a_j^2} \overrightarrow{d} \quad II-31$$

$$(II-20) \quad \left\{ \begin{array}{l} \langle y, y \rangle - b \langle 1, y \rangle - \sum_j a_j \langle x_j, y \rangle - \langle d, d \rangle = 0 \\ \langle y, 1 \rangle - b \langle 1, 1 \rangle - \sum_j a_j \langle x_j, 1 \rangle = 0 \\ \langle y, x_1 \rangle - b \langle 1, x_1 \rangle - \sum_j a_j \langle x_j, x_1 \rangle + a_1 \langle d, d \rangle = 0 \\ \langle y, x_2 \rangle - b \langle 1, x_2 \rangle - \sum_j a_j \langle x_j, x_2 \rangle + a_2 \langle d, d \rangle = 0 \\ \dots \dots \dots \\ \langle y, x_R \rangle - b \langle 1, x_R \rangle - \sum_j a_j \langle x_j, x_R \rangle + a_R \langle d, d \rangle = 0 \end{array} \right. \quad II-32$$

(el sist. II-32 son $R + 2$ ecuaciones con $R + 2$ incógnitas) (El sistema es no lineal) Incógnitas:

$$a_1, \dots, a_R, b \text{ y } \langle d, d \rangle$$

Teóricamente, resolviendo el sistema II-32 se solucionara totalmente el problema de ajuste de un hiperplano. Pero la resolución requiere técnicas especiales debido a que no es lineal. Uno de los inconvenientes (y no es el mayor), es que generalmente se obtienen 2^R conjuntos de soluciones para $a_1, a_2, \dots, a_R, b$ y $\langle d, d \rangle$. De todas debe elegirse la que corresponde al menor $\langle d, d \rangle$ (Notese que cada conjunto es un extremo, pero solo uno es mínimo absoluto).

CAP. III - CONSISTENCIA Y EFICIENCIA DEL MÉTODO.

A FUNCIÓN DE MÁXIMA VEROSIMILITUD DE FISHER.

Hemos estudiado en el Capítulo II, un método posible de considerar el error en todas las variables. Pero para ser rigurosos, debemos mirarlo desde el punto de vista que lo obtenido es simplemente un estimador y como tal debe satisfacer las condiciones de CONSISTENCIA Y EFICIENCIA definidas en el primer capítulo.

Una de las formas más rigurosas de hacerlo, es estudiar la distribución estadística de lo que hemos llamado distancia, y aplicarle a esa distribución la FUNCIÓN DE MÁXIMA VEROSIMILITUD (F.M.V.) de Fisher.

En realidad la F.M.V. ya fue utilizada por Gauss (Werke, vol. 4; Göttinga, 1880), pero como método general fue introducido por R.A. Fisher ("On an absolute criterion for fitting frequency curves", *Mess. of Math.*, 41, (1912), pag. 155). Lo importante de este método, y que es lo que nosotros necesitamos, es que la F.M.V. es un estimador, que bajo ciertas condiciones generales, es asintóticamente consistente y asintóticamente eficiente. Si podemos demostrar que los resultados que se obtienen con la F.M.V. son similares a los que se obtienen con el M.D.M., entonces habremos demostrado esas dos propiedades para el M.D.M.

El estimador de la F.M.V. consiste en lo siguiente:

Se define como función de verosimilitud "L" de una muestra de n valores de una población de tipo continua a:

$$L(y_1, y_2, \dots, y_n, a) = f(y_1, a) \cdot f(y_2, a) \dots f(y_n, a) \quad \text{III-1}$$

donde:

$y_1, y_2, \dots, y_n$ son los valores que toma la variable aleatoria y .

a - - - - - es un parámetro que no se conoce a priori y desea estimarse a partir de la misma muestra.

$f(y_i, a)$ - - es la función de densidad de probabilidad de encontrar el valor y_i , cuando el parámetro desconocido vale " a ".

Y el método consiste en elegir como parámetro " a ", a aquel valor que hace máxima a "L", o lo que es lo mismo, al logaritmo de L:

$$\frac{\partial \ln L}{\partial a} = 0 \quad \text{III-2}$$

donde, como siempre, es necesario estudiar el comportamiento de la derivada segunda.

El método se generaliza inmediatamente cuando el número de parámetros son dos o más, obteniéndose un sistema de dos o más ecuaciones (y un Hessiano que estudiar).

B - ESTADÍSTICA DE LA VARIABLE ALEATORIA "DISTANCIA"

Para poder aplicar la F.M.V. debemos conocer la función densidad de probabilidad que satisface la distancia. A continuación la discutiremos para diferentes casos.

a) Distancia "real": Si todas las variables son normales, con igual varianza σ^2 , entonces $(x_{ji} - X_{ji})$ es una variable normal $(0, \sigma)$ y tendremos que la distancia "real" definida por ec. II-4, por ser la suma del cuadrado de $R + 1$ variables aleatorias normales $(0, \sigma)$ es una variable aleatoria con distribución χ^2 (exactamente) con $R + 1$ grados de libertad, cuya función densidad de probabilidad es:

$$f. de d.p. (d_i^2) = \frac{1}{2^{\frac{R+1}{2}} \sigma^{R+1} \Gamma\left(\frac{R+1}{2}\right)} \cdot d_i^{R-1} \cdot e^{-\frac{d_i^2}{2\sigma^2}} \quad \text{III-3}$$

b) Distancia mínima al cuadrado: R.A. Fisher (Journal of the Royal Statistical Society 85 (1922) pag. 87 y 87 (1924) pag. 442) demostró que es necesario reducir el número de grados de libertad de la distribución en una unidad por cada uno de los parámetros estimados a partir de la muestra. Esto no es el caso, porque los R parámetros x_{ji} que hay que determinar para cada j , a fin de obtener la distancia mínima, caracterizan a cada variable aleatoria y no a la distribución de la variable aleatoria. Las condiciones de mínimo II-6, al afectar a las variables aleatorias, en una forma no predecible (dependen de la función II-3 y del punto), afectan a su distribución, también en una forma no predecible.

En primera aproximación, si la perturbación es pequeña, podemos suponer que a los $R + 1$ grados de libertad originales, deben descontarse las R condiciones de mínima distancia (sistema II-6) con lo que quedan $R + 1 - R = 1$ grado de libertad, y entonces la distribución de la distancia mínima, será una χ^2 con un grado de libertad. Y la función densidad de probabilidad de la distancia mínima al cuadrado será:

$$f.d.p. (d_i^2) = \frac{1}{\sqrt{2 \cdot \sigma^2 \cdot \pi} \cdot d_i^2} \cdot e^{-d_i^2 / (2 \cdot \sigma^2)} \quad \text{III-4}$$

Es interesante destacar que la varianza que aparece en III-4 es la misma que la de III-3. Pero la varianza de las variables, a priori no pueden calcularse, ya que la distancia real jamás puede conocerse. Entonces es evidente la importancia que tiene encontrar un método de calcularla a través de III-4. Lo haremos en este mismo capítulo, más adelante.

c) Distancia mínima: La ec. III-4 es la función densidad de probabilidad de la distancia mínima al cuadrado. Si a d_i^2 le extraemos la raíz y tomamos como signo el positivo, su distribución estadística en una χ^2 modificada, de un grado de libertad. Y la función densidad de probabilidad de la distancia mínima, tomada con signo positivo será:

$$f.d.p. (d_i^2) = \frac{2}{\sqrt{2 \cdot \sigma^2 \cdot \pi}} \cdot e^{-d_i^2 / (2 \cdot \sigma^2)} \quad \text{III-5}$$

La ec. III-5 es solo aproximada, ya que se deduce de la III-4, que también es aproximada.

d) - Distancia mínima con superficie orientada: Bajo ciertas condiciones, las fórmulas III-4 y III-5 pueden llegar a ser tan exactas como se quiera. Esas condiciones son las mismas que se necesitan para demostrar lo siguiente:

La distancia mínima tiene una distribución que tiende a la normal, cuando se cumple que:

- H1) Todas las coordenadas de cada punto experimental, son variables aleatorias con distribución normal, y todas con igual varianza: σ . No es necesario que se conozca σ .
- H2) Alrededor de cada punto x'_{1i} , x'_{2i} , ..., x'_{Ri} de coordenada mínima, en un entorno definido por

$$x'_{ji} - \lambda \sigma < x_j < x'_{ji} + \lambda \sigma, \quad \lambda > 1 \quad \text{III-6}$$

la función II-3 difiere en menos de un ϵ del plano tangente que pasa por dicho punto (se puede demostrar que el plano tangente tiene como vector normal a la distancia mínima). λ y ϵ son independientes de i y de j . $\epsilon < \sigma$.

H3) La función II-3 es orientable en todo el dominio donde se han tomado puntos experimentales. (Esto se pide para darle un signo a la raíz cuadrada que define la distancia mínima. Ya se aplicó implícitamente al calcular la recta cuando se tomó como signo de la distancia mínima, el signo que tenía el error de la coordenada 'y' - Can. II/C-).

H4) Se supone que el punto "real" Y_i , X_{1i} , X_{2i} , ..., X_{Ri} se encuentra sobre la función.

Demostración: Podemos definir la distancia casi real, como la distancia entre el punto experimental y la proyección del punto real sobre el plano tangente.

Si las coordenadas de la distancia real están en el entorno III-6, la distancia casi real diferirá de la real en menos de ϵ y como $\epsilon < \sigma$, entonces estos puntos tendrán aproximadamente la misma distribución que la distancia real, o sea que su f.d.p. tenderá a la III-3, y será igual cuando $\epsilon = 0$.

Queda por considerar los puntos "reales" que salen fuera del entorno III-6. La cantidad de ellos será función de λ . Si la II-3 fuera función de una sola variable independiente, y si $\lambda = 1$, el 31,7% estarían fuera del entorno. Si $\lambda = 2$, estarían afuera el 4,5%. Si $\lambda = 3$ solo el 0,27%. Para el caso de que fueran dos las variables independientes y $\lambda = 1$, el 53,4 % estarían afuera. En tabla 1 se han calculado el número de puntos "reales" que quedarían excluidos del intervalo para distintos λ y para distinto número de variables independientes.

T A B L A 1

Nº de variab. indep.	$\lambda = 0$	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$
1	100 %	31,73 %	4,55 %	0,27 %	0,006%
2	100	53,39	8,89	0,54	0,01
3	100	68,18	13,04	0,81	0,02
4	100	78,28	17,00	1,08	0,02
5	100	85,17	20,77	1,34	0,03
6	100	89,88	24,33	1,61	0,04
7	100	93,09	27,82	1,87	0,04
8	100	95,28	31,10	2,14	0,05
9	100	96,78	34,24	2,40	0,05
10	100	97,8	37,23	2,67	0,06

Porcentaje de puntos "reales" fuera del intervalo definido en III-6 en función de λ para distinto número de variables independientes.

De la tabla se deduce que si se satisface H2) para $\lambda = 3$, aunque el número de variables independientes sea 10, puede afirmarse con error menor que el 3%, que la distancia casi real tiene una distribución χ^2 con $R + 1$ grados de libertad.

La proyección de la distancia casi real sobre la recta normal al hiperplano tangente es la distancia mínima (ya que la distancia mínima es normal al hiperplano tangente). Esta proyección es una combinación lineal de los errores sobre todas las coordenadas del punto, y es igual a la combinación lineal que resulta de hacer una rotación del sistema de coordenadas (diferente para cada punto experimental), tal que una de las nuevas coordenadas coincida con la distancia mínima, y las otras, ortogonales, están sobre el hiperplano tangente. Para hacer esta rotación, tenemos que hacer una transformación ortogonal, obteniendo (en la misma notación usada en II-23)

$$d_{i(\min.)} = \beta_{y_i} \epsilon_{y_i} + \beta_{x_{1i}} \epsilon_{x_{1i}} + \beta_{x_{2i}} \epsilon_{x_{2i}} + \dots + \beta_{x_{Ri}} \epsilon_{x_{Ri}}$$

Pero la combinación lineal de variables aleatorias normales, es una variable aleatoria normal, cumpliéndose además que

$$\begin{aligned} \sigma^2_{d(\min)} &= \beta^2_{y_i} \sigma^2_y + \beta^2_{x_1} \sigma^2_{x_1} + \beta^2_{x_2} \sigma^2_{x_2} + \dots + \beta^2_{x_R} \sigma^2_{x_R} \\ &= (\beta^2_{y_i} + \beta^2_{x_{1i}} + \beta^2_{x_{2i}} + \dots + \beta^2_{x_{Ri}}) \sigma^2 \end{aligned}$$

y como la transformación es ortogonal

$$g_{y_i}^2 + g_{x_{1i}}^2 + g_{x_{2i}}^2 + \dots + g_{x_{Ri}}^2 = 1$$

$$\boxed{\sigma \text{ (d min.)} = \sigma} \quad \text{III-7}$$

En consecuencia, queda demostrado, que bajo las hipótesis hechas, la distancia mínima tiene una distribución que tiende a la normal, con la misma varianza que la de las coordenadas:

$$f.d.p. \text{ (d min)} = \frac{1}{\sqrt{2 \pi \sigma^2}} e^{-d^2 / (2 \cdot \sigma^2)} \quad \text{III-8}$$

La orientabilidad de la función se requirió para darle un signo a d_i , con lo cual el campo de definición de III-8 va de $-\infty$ a $+\infty$. Nótese que si no se pudiera orientar a II-3, entonces puede considerarse a

$$+ \sqrt{d^2}$$

que tiene la distribución III-5, con campo de definición de 0 a $+\infty$ (esta es la razón porque tiene un factor 2) y que coincide exactamente con III-8.

La H3) será una útil propiedad cuando intentemos hacer verificación de hipótesis.

C - APLICACION DE LA FUNCION DE MAXIMA VEROSIMILITUD.

Ya conocida la función de probabilidad de la distancia, es muy fácil aplicar la F.M.V. Sin embargo pueden obtenerse sorpresas. Una de las mayores, es comprobar que si se usa este procedimiento a la f.d.p. de la distancia mínima al cuadrado, el resultado es totalmente diferente al método de la distancia mínimas (y es también diferente a lo que se obtiene si se aplica la F.M.V. a la distancia mínima con ó sin superficie orientada). La razón es que las derivadas de III-4 son divergentes, y la F.M.V. se escapa de las "ciertas condiciones generales", dentro de las cuales se prueban sus propiedades. Resumiendo, no es aplicable la F.M.V. a la f.d.p. III-4, siendo este caso uno de esos buenos ejemplos patológicos donde hay que tener sumamente cuidado.

Ahora aplicaremos este estimador, a la distancia mínima con superficie orientada (f.d.p. III-8). La función de verosimilitud (III-1) es para este caso:

$$L = \prod_{i=1}^n (2 \pi \sigma^2)^{-\frac{1}{2}} e^{-\frac{d_i^2}{2 \sigma^2}} = \left[(2 \pi \sigma^2)^{-\frac{1}{2}} \right]^n \cdot e^{-\frac{1}{2 \sigma^2} \sum_{i=1}^n d_i^2} \quad \text{III-9}$$

y su logaritmo:

$$\ln L = -\frac{n}{2} \ln (2 \pi \sigma^2) - \frac{1}{2 \sigma^2} \sum_{i=1}^n d_i^2 \quad \text{III-10}$$

y ahora debemos hacer máxima a III-10 con respecto a las constantes a determinar, es decir σ^2 , k_1 , k_2 , ..., k_p :

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2} - \frac{2}{2\pi\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n d_i^2 = 0$$

$$\boxed{\sigma^2 = \frac{1}{n} \sum_{i=1}^n d_i^2}$$

III-11

$$\frac{\partial \ln L}{\partial k_m} = -\frac{1}{2\sigma^2} \frac{\partial \sum_{i=1}^n d_i^2}{\partial k_m} = 0$$

o sea:

$$\left\{ \begin{array}{l} \frac{\partial \sum_{i=1}^n d_i^2}{\partial k_1} = 0 \\ \dots \dots \dots \\ \frac{\partial \sum_{i=1}^n d_i^2}{\partial k_p} = 0 \end{array} \right.$$

III.12

que es exactamente el sistema II-8. (La condición II-6 ya está implícita al deducir III-8 para la distancia mínima).

Se puede observar que se hubiera llegado al mismo resultado partiendo de la función densidad de probabilidad III-5.

El resultado obtenido, sirve como demostración rigurosa, de que el método de las distancias mínimas es un método asintóticamente consistente y asintóticamente eficiente, ya que la F.M.V. lo es, mientras sean aplicables las condiciones impuestas para demostrar la estadística de la distancia mínima.

CAP. IV - INTERVALOS DE CONFIANZA

A - DEFINICION

Hasta el presente, hemos desarrollado un método de evaluar constantes desconocidas de una función dada, a partir de una muestra de puntos experimentales, y demostramos que el método empleado es eficiente. Si se hiciera una nueva serie de medidas (otro muestreo), los nuevos valores de las constantes muy probablemente diferirán de los primeros, debido a que unos y otros tienen errores. Por eso, es de sumo interés dar a cada constante una cota de error, y primero definir como se interpreta el mismo.

Mucho se ha escrito sobre el tema, clasicamente basado en el teorema de Bayes, más recientemente con la teoría de las probabilidades fiduciales debida a R.A. Fisher y con la teoría de los intervalos confidenciales, debida a J. Neyman.

Supongamos que como resultado de un muestreo, hemos obtenido el valor particular k_m del parámetro exacto desconocido k_m^V . Siguiendo conceptos establecidos por J. Neyman, la precisión de k_m debe asociarse a dos números δ y ϵ ($\delta > 0$), ($1 < \epsilon < 0$) tales que la probabilidad de que el parámetro verdadero k_m^V quede incluido entre los límites $k_m \pm \delta$ sea igual a $1 - \epsilon$. En términos más generales, se trata de encontrar dos parámetros (funciones de los valores muestrales) k_{m1} y k_{m2} , tal que la probabilidad de que el intervalo (k_{m1}, k_{m2}) incluya el parámetro verdadero k_m^V tenga un valor dado $1 - \epsilon$. Cuando exista tal intervalo se llamará intervalo de confianza del parámetro k_m^V , y a la probabilidad $1 - \epsilon$ se llamará coeficiente de confianza del intervalo.

B - VARIANZA DE LOS PARÁMETROS

Estudiaremos primero la varianza de los parámetros, debido a que los intervalos de confianza pueden deducirse como una generalización de aquel.

Para simplificar, comencemos suponiendo que $V(x, y)$ es una medida indirecta de dos variables independientes x e y , que se miden directamente, siendo estas dos variables aleatorias normales, con varianza σ_x y σ_y respectivamente.

Podemos tomar el diferencial de V (en primer orden) alrededor del valor medio de las variables independientes: $\bar{x}$ e $\bar{y}$:

$$dV = \frac{\partial V}{\partial \bar{x}} dx + \frac{\partial V}{\partial \bar{y}} dy \quad \text{IV-1}$$

Para la iésima determinación de V , pasamos a incrementos finitos, sustituyendo dx_i por ϵ_{x_i} y dy por ϵ_{y_i} :

$$\delta V_i = \frac{\partial V}{\partial \bar{x}} \epsilon_{x_i} + \frac{\partial V}{\partial \bar{y}} \epsilon_{y_i} \quad \text{IV-2}$$

donde la aproximación será tanto mejor cuanto menores sean los errores. Si

se hicieron m determinaciones de V , el valor medio de la suma de cuadrados de IV-2 es:

$$\frac{\sum_{i=1}^m (\delta V_i)^2}{m} = \left(\frac{\partial V}{\partial x} \right)^2 \frac{\sum_{i=1}^m \epsilon_{xi}^2}{m} + \left(\frac{\partial V}{\partial y} \right)^2 \frac{\sum_{i=1}^m \epsilon_{yi}^2}{m} + 2 \frac{\partial V}{\partial x} \frac{\partial V}{\partial y} \frac{\sum_{i=1}^m (\epsilon_{yi} \epsilon_{xi})}{m} \quad \text{IV-3}$$

Si $n \rightarrow \infty$ las sumas se pueden poner en función de las varianzas. Además si los errores no están correlacionados, el último término de IV-3 es nulo, con lo que en el límite tendremos:

$$\sigma_V = \sqrt{\left(\frac{\partial V}{\partial x} \sigma_x \right)^2 + \left(\frac{\partial V}{\partial y} \sigma_y \right)^2} \quad \text{IV-4}$$

Y ahora pasemos a aplicar el resultado IV-4 al M.D.M. Cada parámetro k_m es expresable como una función g_m :

$$k_m = g_m(y_1, x_{11}, \dots, x_{R1}, \dots, y_n, x_{1n}, \dots, x_{Rn}) \quad \text{IV-5}$$

y como hemos supuesto que todas las coordenadas de todos los puntos son independientes, y que cada coordenada tiene varianza σ , tendremos:

$$\sigma_{k_m} = \sigma \sqrt{\left(\frac{\partial g_m}{\partial y_1} \right)^2 + \left(\frac{\partial g_m}{\partial x_{11}} \right)^2 + \dots + \left(\frac{\partial g_m}{\partial x_{R1}} \right)^2 + \dots} \quad \text{IV-6}$$

$$\dots + \left(\frac{\partial g_m}{\partial y_n} \right)^2 + \left(\frac{\partial g_m}{\partial x_{1n}} \right)^2 + \dots + \left(\frac{\partial g_m}{\partial x_{Rn}} \right)^2$$

que es la varianza del parámetro k_m en función de la varianza de las coordenadas. La misma debe interpretarse en el sentido de que si se hace un gran número de cálculos de k_m con puntos experimentales, el 68,3% estarán comprendidos dentro del intervalo cuyo ancho es dos veces IV-6, pero no implica que k_m sea una variable normal. Podría ocurrir que ni siquiera fuera aleatoria.

C - INTERVALO DE CONFIANZA DE LOS PARAMETROS.

Antes de continuar, haremos una observación para lo que sigue. Hasta aquí estuvimos llamando "varianza" indistintamente a

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{IV-7}$$

y a:

$$\sigma^2 = \lim_{n \rightarrow \infty} s^2 \quad \text{IV-8}$$

cuando con rigor, la primera (IV-7) es la "varianza muestral" mientras IV-8 es la verdadera definición de "varianza", sin que hasta el presente esta dualidad nos trajera inconvenientes. Pero, de ahora en adelante, distinguiremos un término del otro.

El paso de calcular la varianza muestral, a calcular los intervalos de confianza de los parámetros, es bastante fácil cuando las variables son normales. Para ello, tengamos en cuenta que la media de una distribución normal (m), para un coeficiente de confianza $1-p/100$ (donde p es el tanto por ciento de probabilidad de que el valor verdadero se encuentre f u e r a del intervalo de confianza), tiene un intervalo de confianza:

$$\bar{x} - \lambda_p \frac{s}{\sqrt{n-1}} < m < \bar{x} + \lambda_p \frac{s}{\sqrt{n-1}} \quad \text{IV-9}$$

donde, como siempre $\bar{x}$ es el valor medio calculado de una muestra de n determinaciones, s es la varianza muestral y λ_p es el valor del p por ciento de la distribución t-Student para $n-1$ grados de libertad (Ver tabla 4 de H. Cramer-Mathematical Methods of Statistics, p. ej.). Generalmente en física se usa hacer $\lambda_p = 1$, que corresponde a $p = 30$, con lo cual se estudia la probabilidad de que el verdadero parámetro se encuentre entre:

$$\bar{x} \pm s/(n-1)^{\frac{1}{2}} \quad \text{IV-10}$$

con una probabilidad del 70% aproximadamente ($\lambda_p \approx 1$ para $n = 30$, siendo casi independiente del número de grados de libertad, o sea del número de mediciones).

Y el intervalo de confianza de los parámetros k_m se calcula simplemente usando la fórmula IV-4, pero reemplazando las varianzas muestrales por los intervalos IV-10. O sea que finalmente se tiene:

$$\delta_{k_m} = \delta \sqrt{\left(\frac{\partial g_m}{\partial y_1}\right)^2 + \left(\frac{\partial g_m}{\partial x_{11}}\right)^2 + \dots + \left(\frac{\partial g_m}{\partial x_{R1}}\right)^2 + \dots + \left(\frac{\partial g_m}{\partial y_n}\right)^2 + \left(\frac{\partial g_m}{\partial x_{1n}}\right)^2 + \dots + \left(\frac{\partial g_m}{\partial x_{Rn}}\right)^2} \quad \text{IV.11}$$

donde δ y δ_{k_m} son los intervalos de confianza para las coordenadas y para el parámetro k_m respectivamente, para un coeficiente de confianza del 0,7. δ se calcula como:

$$\delta = \frac{\sum_{i=1}^n (d_i)^2}{n \cdot (n - P)} \quad \text{IV-12}$$

donde se debe poner P , porque el cálculo de P constantes le quita a la distribución t-Student P grados de libertad.

D - UN METODO ALTERNATIVO APROXIMADO

Tanto la expresión IV-6 como la IV-9, que son proporcionales, son de muy difícil cálculo, y salvo casos particulares, no pasan de ser simples declaraciones de deseos, debido a que g_m es siempre una función complicada. Ya para la recta es el resultado de resolver un sistema no lineal. Muchas veces, ni siquiera podrá obtenerse la forma explícita IV-5.

Además, el cálculo efectivo de las k_m en la mayoría de los casos, se tendrá que hacer en forma iterativa, en donde no se calcula, por no hacer falta, g_m .

Todo esto indica la necesidad de buscar un método alternativo de calcular los intervalos de confianza.

La idea es conseguir un resultado aproximado, mediante el cálculo de cuanto habría que variar cada parámetro por separado para que la varianza varíe en la misma cantidad que su intervalo de confianza.

Para ello, recordemos primero el intervalo de confianza de la varianza. Para un coeficiente de confianza de $1-p/100$ (donde p es el tanto por ciento de la probabilidad de que el verdadero valor de la varianza quede fuera del intervalo), la varianza tiene el intervalo de confianza:

$$\frac{s \sqrt{n}}{\chi_{p''}^2} < \sigma^2 < \frac{s \sqrt{n}}{\chi_{p'}^2} \quad \text{IV-13}$$

donde: $p' = 100 - \frac{1}{2} P$

$$p'' = \frac{1}{2} p$$

$\chi_{p'}^2$ y $\chi_{p''}^2$ es el valor del p' y p'' por ciento de la distribución χ^2 para $n - 1$ grados de libertad. (Ver tabla 3 de H.

Cramer-Mathematical Methods of Statistic, p. ej.)

s es la varianza muestral.

n es el número de observaciones.

Después, notemos que la expresión IV-5 (y dado que los k_m se determinan a través de II-6 y II-8) podemos escribirla como:

$$k_m = g_m^0 (d_1, d_2, \dots, d_n) \quad \text{IV-14}$$

donde d_i es la distancia (II-5). Entonces podemos pensar en una expresión inversa a IV-14, que es justamente II-5.

Ahora fijamos un coeficiente de confianza (podría ser de 0,7 o sea el 70%). Buscamos en tablas

$$\sqrt{\frac{2}{\chi_p^2}}$$

y calculamos el valor superior del intervalo de confianza de la varianza (IV-13). Y finalmente vamos iterativamente variando k_m y calculando nuevos "s", hasta que llegue al límite superior fijado, que corresponderá a un dado k_m' . Y daremos la resta $k_m' - k_m$ como intervalo de confianza del parámetro k_m , para el coeficiente de confianza elegido.

El método usado es aproximado y semi-intuitivo. Merece un estudio más profundo, posible mediante las facilidades de una computadora.

En vez de estudiar el efecto sobre la varianza por perturbar algún k_m , podría hacerse sobre el valor medio de d_i . Esto trae el inconveniente que correlaciona a las distancias. Un ejemplo extremo de esto, es si se está calculando una recta y se le suma a todas las distancias una constante (y de esta forma, cambia su valor medio), esto modificará al término independiente, pero dejará inalterada a la pendiente.

CAP. V - VERIFICACION DE HIPOTESIS

A - METODO DE JI CUADRADO

Tanto la hipótesis d) del método de cuadrados mínimos, como la hipótesis d) del método de distancias mínimas, presuponen que no hay error en la elección de la función cuyos parámetros se van a ajustar.

Afortunadamente, podemos apelar en una forma indirecta a métodos estadísticos para verificar si la hipótesis hecha es o no correcta. Ya en la introducción, al referirnos a inferencia estadística, hablamos del problema de distribución y del error de primera y segunda clase a que puede llevarnos.

En general, en estadística se tiene una muestra que por una u otra razón se acepta que tiene una determinada distribución. Pero existen desviaciones entre los valores muestrales y la correspondiente distribución. Y se desea saber si estas desviaciones indican que se ha elegido mal a la distribución, o son fluctuaciones debido al carácter aleatorio de la variable.

Usaremos ahora el método de χ^2 para verificar hipótesis. Cuando la distribución es discreta, el método consiste en calcular las desviaciones entre el valor esperado y los valores observados, elevar al cuadrado estas desviaciones (para evitar la compensación de las desviaciones de distinto signo), dividir a cada uno de los cuadrados por el valor esperado, para tener una medida relativa de la importancia de la desviación, y sumar todas las medidas relativas así obtenidas. El resultado es el cuadrado de una variable asintóticamente normal, y por lo tanto tenderá asintóticamente a la distribución de Ji cuadrado (χ^2), y la llamaremos

$$\chi^2 \text{ (muestral)}$$

Se adopta luego un nivel de significación p, siendo p la probabilidad de rechazar la hipótesis cuando esta es cierta, y que generalmente se toma entre el 5 y el 0,1 %. Se busca en una tabla de χ^2 (p. ej. la citada de H. Cramer) el valor de

$$\chi^2 \text{ (teórico)}$$

para el dado p y para el número de grados de libertad del problema, y se compara el χ^2 (muestral) con el χ^2 (teórico). Si el primero es mayor que el segundo, se dice que la hipótesis hecha de que la muestra correspondía a la dada distribución, es falsa. (Siempre hay un riesgo p/100 de equivocarse)

Quando la distribución es continua, el valor observado se obtiene haciendo un agrupamiento de datos, resultante de dividir el dominio de la variable aleatoria en "z" clases contiguas (cuidando que el número de observaciones sea mucho mayor que z), contando cuantos sucesos se tiene en cada clase, y de allí se calcula su probabilidad clásica. El valor esperado se obtiene calculando la integral de la función densidad de probabilidad entre los límites de cada clase. Lo demás queda invariado.

B - APLICACION A LA DISTANCIA MINIMA

La aplicación al método de las distancias mínimas (o al M.C.M.) es un caso interesante. Debemos aplicar χ^2 a un dado muestreo de distancias mínimas, para verificar si el mismo cumple o no con la distribución III-8 (ó III-5 si la superficie no es orientable), y en el caso de que la χ^2 (muestral) sea mayor que la χ^2 (teórica), no significará que la distribución es falsa, sino que la función elegida (que por supuesto afecta a la distribución) está M A L elegida (a menos que no se cumpla alguna de las hipótesis que se necesitaron para demostrar III-8, con excepción de F3). Lamentablemente, por medios estadísticos no puede determinarse cual es la función que debe elegirse. Pero por lo menos puede saberse cual es la que NO DEBE ELEGIRSE.

El cálculo de χ^2 (muestral) es aquí fácil. Basta fijarse que dividiendo la sumatoria II-7 por σ^2 , ya tenemos:

$$\chi^2 \text{ (muestral)} = \frac{1}{\sigma^2} \sum_{i=1}^n (d_i)^2 \quad V-1$$

que es justamente lo que nos interesa, por ser una variable χ^2 , la suma del cuadrado de variables aleatorias normales con varianzas 1, al haber dividido por σ^2 , y que tiene $n - P - 1$ grados de libertad (pierde P grados de libertad por las P constantes k que deben determinarse de la misma muestra, más otro grado que pierde por determinar la varianza también de la misma muestra). Nótese que el hecho que la superficie sea o no orientable no importa para el cálculo de la χ^2 (muestral).

En principio podría esperarse que si se usa una función con más parámetros a determinar, al disminuir χ^2 (muestral) puede trabajarse con un mejor nivel de significación p . Pero (y esta es una de las propiedades más importantes de Ji cuadrado para nuestro M.D.M.) con mayor número de parámetros implica menor número de grados de libertad, con lo cual disminuye también el valor de χ^2 (teórico), y finalmente debe llegarse a una solución de compromiso. Esto es muy útil para el caso de tener que adoptar un criterio de cuantos términos conviene usar cuando se intenta un desarrollo en serie, o cualquier método similar.

C - DOCIÑA PARA CORRELACIONES LOCALES

Supongamos que los puntos obtenidos en un muestreo (o sea en un ciclo completo de experiencia) puedan ser ordenables, por ejemplo según una coordenada creciente, y que la superficie definida por la función II-3 sea orientable, lo que nos permite elegir un signo a la distancia mínima, según se encuentre a un lado u otro de la superficie. (Recordemos que la distancia mínima es perpendicular a la superficie).

Con estas armas, podremos subsanar un inconveniente que tiene el cálculo anterior. no distingue que varios grupos diferentes de puntos contiguos tengan una distancia del mismo signo, es decir que existe una correlación local del signo de las distancias. El cálculo del inciso B) solo toma en cuenta que las distancias, como un todo, tengan una distribución normal. Esto ocurre cuando se trata de ajustar puntos que están sobre una parábola por medio de una recta (o viceversa), con lo cual todos los puntos que están a la derecha del punto de intersección de la recta con la parábola

la tienen distinto signo que los que están a la izquierda (hasta que haya un nuevo cruce), y en general cuando se trata de ajustar puntos que tienen poco error con una función que no es la correcta.

Una forma contundente de verificar estas correlaciones, es usar la ordenación hecha y sumar algebraicamente la distancia de cada punto a la distancia del punto siguiente, con lo cual se tiene la nueva variable aleatoria:

$$D_1 = d_1 + d_2 ; \dots ; D_i = d_i + d_{i+1} ; \dots ; d_n = d_n + d_1$$

(la última rigurosamente no tiene sentido, pero no pesa mucho y nos permite tener n variables aleatorias independientes), que por ser la suma de dos variables aleatorias normales independientes, tiene una distribución normal, con una nueva media (m') que es la suma de las dos medias anteriores, y con una nueva varianza (σ') cuyo cuadrado es la suma de los cuadrados de las varianzas de cada una de las variables suma:

$$m' = m_1 + m_2 = 0 \quad \text{V-2}$$

$$\sigma'^2 = \sigma_1^2 + \sigma_2^2 = 2 \sigma^2$$

$$\boxed{\sigma'^2 = \sqrt{2} \cdot \sigma^2} \quad \text{V-3}$$

con lo cual calculamos la nueva χ^2 (muestral)

$$\chi^2 \text{ (muestral)} = \sum_{i=1}^n \frac{D_i^2}{2 \sigma^2} = \left(\sum_{i=1}^n \frac{d_i^2}{\sigma^2} \right) + \frac{1}{\sigma^2} \left(d_n d_1 + \sum_{i=1}^{n-1} d_i d_{i+1} \right) \quad \text{V-4}$$

y como puede verse de V-4, si no existe ninguna clase de correlaciones el último término es nulo y vuelve a obtenerse V-1 (que debe a su vez compararse con la misma χ^2 (teórica) del inc. A), con lo que sería solamente una repetición. Pero si existe correlación, el muestreo de la variable suma va a tener dispersión mayor que la calculada V-3, y el método de Ji cuadrado va a detectar si estas desviaciones son o no significativas.

Para detectar una correlación de mayor alcance, puede pensarse en generar nuevas variables sumas de más de dos distancias contiguas.

CAP. VI - NOTAS.-

A) - DISCUSION SOBRE LA VARIANZA DE LAS COORDENADAS

Discutiremos seguidamente tres casos que se apartan de la hipótesis b) del planteo del M.D.M., que dice que todas las coordenadas tienen igual varianza (e implícitamente el mismo peso).

1) Variables con distinta varianza: Hay que entender que la varianza muestral solo puede conocerse a posteriori del cálculo de los parámetros, que es recién cuando se puede calcular la expresión III-11. Sin embargo, a través de las mediciones de las mismas coordenadas, o como resultado de un estudio previo del método de medida, pueden establecerse con bastante precisión un juego de relaciones del tipo:

$$\frac{\sigma_{x_m}}{\sigma_{x_j}} = \gamma_{m,j} \quad \text{VI-1}$$

donde σ_{x_m} y σ_{x_j} son las varianzas de las coordenadas x_m y x_j y ambas se desconocen. Solo se conoce $\gamma_{m,j}$ que así resulta un dato más del problema, o mejor dicho un juego de datos. Entonces es fácil hacer el cambio de variable:

$$\begin{aligned} y_i \text{ (nuevo)} &= y_i \text{ (viejo)} / \gamma_{y,j} \\ x_{1i} \text{ (nuevo)} &= x_{1i} \text{ (viejo)} / \gamma_{1,j} \\ &\dots\dots\dots \\ x_{Ri} \text{ (nuevo)} &= x_{Ri} \text{ (viejo)} / \gamma_{R,j} \end{aligned} \quad \text{VI-2}$$

con lo cual seguimos sin saber la varianza, pero estamos seguros que todas nuestras coordenadas tienen igual varianza.

2) Una (o varias) variables independientes con error nulo: Si una (o varias) de las coordenadas tiene error nulo o despreciable frente a los demás (lo que implica que su varianza sea cero o tienda a cero), no es aplicable lo anterior. Sin embargo el caso es fácil de salvar.

Supongamos que sea la coordenada x_m la que tiene varianza nula. Entonces en vez de usar II-5 (y el sistema II-6) para definir la distancia mínima, usamos:

$$\begin{aligned} d_i &= \left(y_i - F(x'_{1i}, \dots, x'_{m-1,i}, x_{m,i}, x'_{m+1,i}, \dots, x'_{Ri}) \right)^2 + \\ &+ \sum_{j=1}^{m-1} (x_{ji} - x'_{ji})^2 + \sum_{j=m+1}^n (x_{ji} - x'_{ji})^2 \end{aligned} \quad \text{VI-3}$$

y el sistema II-6 serán las derivadas con respecto a

$$x'_1, x'_2, \dots, x'_{m-1}, x'_{m+1}, \dots, x'_R,$$

lo que será un sistema de $R-1$ ecuaciones con $R-1$ incógnitas. Y lo demás sigue invariado, recordando que al hacer las sumatorias sobre j (subíndice que hemos distinguido para sumar sobre variables independientes) debe omitirse el término de $j = m$.

3) La variable dependiente con error nulo: Para el caso particular en que la variable dependiente tenga error nulo, tendremos que hacer mínima a la distancia definida como:

$$d_i^2 = \sum_{j=1}^R (x_{ji} - x'_{ji})^2 \quad \text{VI-4}$$

con la condición que

$$y_i - F(x'_{1i}, \dots, x'_{Ri}, k_1, \dots, k_p) = 0$$

que es un caso de extremo de funciones con variables ligadas, y para su resolución puede apelarse al método de multiplicadores de Lagrange.

4) Puntos de distinto peso: Si los puntos tienen distinto peso, el problema se resuelve haciendo la suma II-7 con un peso determinado para cada d_i . Por ejemplo, si estos pesos fueran q_1 para el punto 1, ..., q_n para el punto n , habría que tratar de hacer mínima con respecto a los parámetros k_j a la sumatoria:

$$\sum_{i=1}^n q_i^2 \cdot d_i^2$$

y lo demás todo igual.

B - FUNCION LINEAL: DIFERENCIA ENTRE M.C.M. y M.D.M.

Si se compara el sistema que se tiene que resolver para ajustar una recta por el M.C.M. con el similar para el M.D.M. se encuentra que los mismos solo coinciden cuando la pendiente de la recta tiende a cero o el error tiende a cero. Esto es razonable, porque en ese caso las distancias mínimas son equivalente a errores en la coordenada "y". Caso contrario el sistema, y por ende los valores de los parámetros resultan diferentes, aunque asintóticamente consistentes. Sin embargo, los valores de la varianza, aún cuando n tienda a infinito, son intrínsecamente diferentes.

C - CASO ESPECIAL CUANDO EN VEZ DE UNA SUPERFICIE SE DEBE AJUSTAR UNA CURVA.

Supongamos que en vez de tener que ajustar una función del tipo de la II-3, pretendemos ajustar una hipercurva de M dimensiones en el espacio de $R+1$. Si $Q + M = R + 1$, la forma más simple (y la única que consideraremos) de la expresión analítica de la curva será:

