Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Profesionales Zaragoza

Diseño de un Programa para Computadora Digital para obtener Superficies de Respuesta (Bidimensionales), de Ecuaciones Lineales.

T E S I S

Que para obtener el título de:

INGENIERO QUIMICO

P r e s e n t a :

ROBERTO RODAS LECONA

México, D. F. 1984

I N D I C E

RESUMEN

En este trabajo, se resumen las bases teóricas y principales conceptos de los modelos lineales, los cuales son utilizados ampliamente en la metodología de superficie de respuesta, cuyo objetivo es encontrar "óptimos" de procesos de una o más variables independientes, en base a experimentación y análisis de resultados en forma estadística, además se dan ejemplos de uso del programa graficador de superficies desarrollado para complementar la mayoría de los paquetes estadísticos actuales.

1

PROYECTO INICIAL PARA
APROBACION DE TEMA DE TESIS

A) TITULO DEL PROYECTO

Diseño de un programa para computadora digital - para obtener superficies de respuesta (bidimensionales), de ecuaciones lineales.

B) AREA ESPECIFICA DEL PROYECTO

Estadística y Computación aplicados al análisis - de experimentos y optimización de procesos.

C) PERSONAS QUE PARTICIPAN

Alumno: Roberto Rodas Lecona

Asesor: Ing. Ruben Cariño Garay

D) FUNDAMENTACION DE LA ELECCION DEL TEMA

La superficie de respuesta es una técnica que se desarrolló principalmente en el área de Ingeniería Química; es - utilizada para optimizar procesos físicos, químicos, biológicos o sociológicos donde se mide una variable cuantitativa que depende - de uno o más factores cuantitativos. Esto se puede escribir como:

$$Y_u = F(x_{1u}, x_{2u}, x_{3u}, \ldots\ldots, x_{ku}) + \varepsilon_u$$

donde $u = 1,2,\ldots\ldots,N$ representa n-observaciones de un experimento, o diseño experimental, x_{iu} representa el valor del i-ésimo - factor en la u-ésima observación. A la función F se le llama superficie de respuesta, la cual sería k-dimensional.

Cuando no se conoce la forma matemática de F, la función puede aproximarse satisfactoriamente dentro de la región -

experimental por un polinomio de las variables x_{iu}. Si solo hay 2 factores, la ecuación puede graficarse en 3 dimensiones, o alternativamente en 2, con los ejes, en esté último caso, siendo los factores y graficando líneas o superficies de respuesta (de aquí el nombre de la técnica) constantes, algo análogo a las isotermas en un diagrama P-V de un gas. Este tipo de diagrama es muy útil ya que en él se puede ir localizando los puntos de operación óptima, además de observar el número de combinaciones de los factores (x_1, x_2) que nos pueden dar una misma respuesta.

El diagrama es una parte importante de la técnica, pero es demasiado laborioso elaborarlo; esto ha motivado a crear programas de computadora que realizen este trabajo. Sin embargo, ellos se han incluido como parte del software de 'paquetes' de programas que son rentados en diferentes partes del mundo. En México existen pocos paquetes de este tipo y solo unos cuantos, como los ubicados en la Escuela Nacional de Agricultura y el Instituto de Investigaciones Agrícolas, pueden graficar superficies de respuesta. Según el software y el equipo utilizado, el tiempo de CPU de cada grafica varía (afectando también la forma de la superficie), pero un promedio representativo sería de 1 minuto. El costo de un minuto de CPU anda alrededor de 2000 pesos, en paquetes rentados. Una amplia investigación en el área de Ingeniería Química, como puede ser la investigación de las condiciones óptimas de operación de un reactor, de una columna de destilación o de un sedimentador, puede fácilmente llevar de 30 a 60 minutos de CPU en la sola elaboración de las gráficas. Debido a este y a otros altos costos involucrados, este tipo de metodología no se ha im-

plementado con fines didácticos. Afortunadamente, en el caso de la ENEP Zaragoza, la existencia de una planta piloto facilita la realización de experimentos; estos experimentos podrían diseñarse en forma estadística, y luego, con ayuda de programas como el que se pretende desarrollar, analizar los resultados para lograr, en última instancia, optimizar las condiciones de operación existentes. De esta forma, el alumno podría tener una sólida formación práctica, ya que el tiempo y dinero ahorrados en la manipulación y análisis de un equipo, podrían utilizarse para operar otro; además, contaría con un método definido que le permitiría comprender y valorar la importancia de las variables involucradas en el fenómeno, y por ende, poder estimar las condiciones óptimas para llevar a cabo el proceso determinado, redundando en un mayor ahorro de recursos. De esta forma, el nivel de preparación de los egresados será de una mayor calidad en todos sentidos.

E) PLANTEAMIENTO DEL PROBLEMA

Se intenta desarrollar un programa que grafique una superficie de respuesta, con la sola información básica de la ecuación de interés, y de los rangos de valores de los factores X_1 y X_2 que se desea estén incluídos en la gráfica. El diagrama, además de esta gráfica, deberá imprimir en sus ejes respectivos los valores de los factores; asimismo se deberán incluir los rangos de la respuesta y su signo de impresión utilizado. La ecuación y los rangos de valores, serán leídos como datos, por lo que será necesario manipularlos antes de proceder a la graficación. Por lo mismo, se deberá elaborar un manual de uso claro y

conciso.

F) OBJETIVOS

a) Describir técnicas de diseño de experimentos - así como la metodología de superficie de respuesta, para que sea - utilizada por los alumnos como guía para:

1.- La definición de condiciones de operación para sus experimentos.
2.- El deliniamiento del análisis a seguir de los resultados obtenidos.
3.- La optimización de la operación del proceso de interes.

b) Proveer de la parte medular del material necesario para aplicar la metodología de superficie de respuesta, elaborando un programa que grafique dicha superficie, para el caso de 2 factores independientes cuantitativos (superficie de respuesta - bidimensional).

G) MATERIAL Y METODOS

Para la elaboración del programa será necesario contar con una clave de computadora que tenga un tiempo de CPU suficiente, para no agotarse en la etapa de verificación y corrección de errores, tomando como estimado que cada gráfica utilizara un minuto de CPU. El lenguaje a usar será el FORTRAN, para facilitar la ejecución de un tipo a otro de máquina, ya que así tendra leves modificaciones. El método que se piensa utilizar en el programa es básicamente modular.

H) BIBLIOGRAFIA QUE APOYA EL PROYECTO

En este aspecto, se podrá contar con una cantidad de información suficiente, si existen en el pais los volúmenes que se detallan, los cuales aunados a los siguientes ya obtenidos, conjuntan la bibliografía a utilizar en este tema.

Bibliografía obtenida:

1.- Bacon, W. David
Making the most of a 'one-shot' experiment
Ind. Eng. Chem.
Vol. 62, No 7 (Julio 1970), pp 27-34

2.- Burtis, C.A. et al.
Optimization of Kinetic Method by Response Surface Methodology and Centrifugal Analysis and Application to the Enzymatic Measurement of Ethanol.
Anal. Chem.
Vol. 53, 1981, pp 1154-1159

3.- Mustacchi, Carlos, y Moresi, Mauro
A strategy to obtain semi-empirical correlations for deterministic systems.
Chemical Engineering Science.
Vol. 35, 1980, pp 737-741.

4.- Murphy, Thomas D.
Design and Analysis of industrial experiments
Chemical Engineering
Junio 6, 1977, pp 168-182

5.- Atkinson, A.C.
Statistical designs for pilot plant and laboratory experiments-Part I
Chemical Engineering
Mayo 9, 1966, pp 149-154

6.- Hunter, W.G. y Atkinson, A.C.
Statistical designs for pilot plant and laboratory experiments-Part II
Chemical Engineering
Junio 6, 1966, pp 159-164

7.- Cochran y Cox
Experimental Design
2 ed., Wiley, New York 1957

Bibliografía a buscar:

1.- Box, G.E. y Wilson, K.B.
On the experimental atainment of optimum conditions
Jour. Roy. St. Soc. B.
13: 1-45

2.- Hill, W.J. y Hunter, W.G.
A review of response surface methodology a literature survey
Technometrics
Vol. 8, 1966, pp 571-579

3.- Box, G.E. y N.R. Draper
A basis for the selection of response surface design
Journal of American Statistical A.
Vol. 54, 1959, pp 622-654

4.- Karson, M. J. et al
Minimum bias estimation and experimental design for response surface
Technometrics
Vol. 11, 1969, pp 461-475

5.- Box, G.E. y J. S. Hunter
Multifactor experimental design for exploring response surfaces.
Annals of Math. Stat.
Vol. 28, 1957, pp 195-241

6.- Johnson, W.W.
A least-squares method of interpreting magnetic anomalies caused by two-dimensional structures
Geophysics
Vol. 34, 1969, pp 65-74

7.- Sefa, D. y Stanley D.
Cowpea Proteins. 1. Use of response methodology in predicting Cowpea (Vigna unguiculuta) protein extractability.
J. Agric. Food Chem.
Vol. 27, 1979, pp 1238-1243

8.- Hopkin, D. y Moss B.
Automata
MacMillan Press Ltd., 1976, London

9.- Lewis II, P. M. et al.
Compiler Design Theory
The Systems programming Series
Addison-Wesley, U.S.A., 1976

10.- Presser, L. et al.
Ciencias de la computación Vol. 1
Limusa-Wiley, México, 1972

11.- Cárdenas, A. F. et al.
Ciencias de la computación Vol. 2
Limusa-Wiley, México, 1972

12.- Hunter, S. J.
Plant experiment
Chemical Engineering
Marzo 28, 1966, pp 57-64

13.- Himmelblau, David M.
Process analysis by statistical methods
Wiley, New York, 1970

14.- Draper, N. R.
Applied Regresión Analisys
Wiley, New York, 1966

15.- Box, G. E. y N. R. Draper
Evolutionary operation
Wiley, New York, 1969

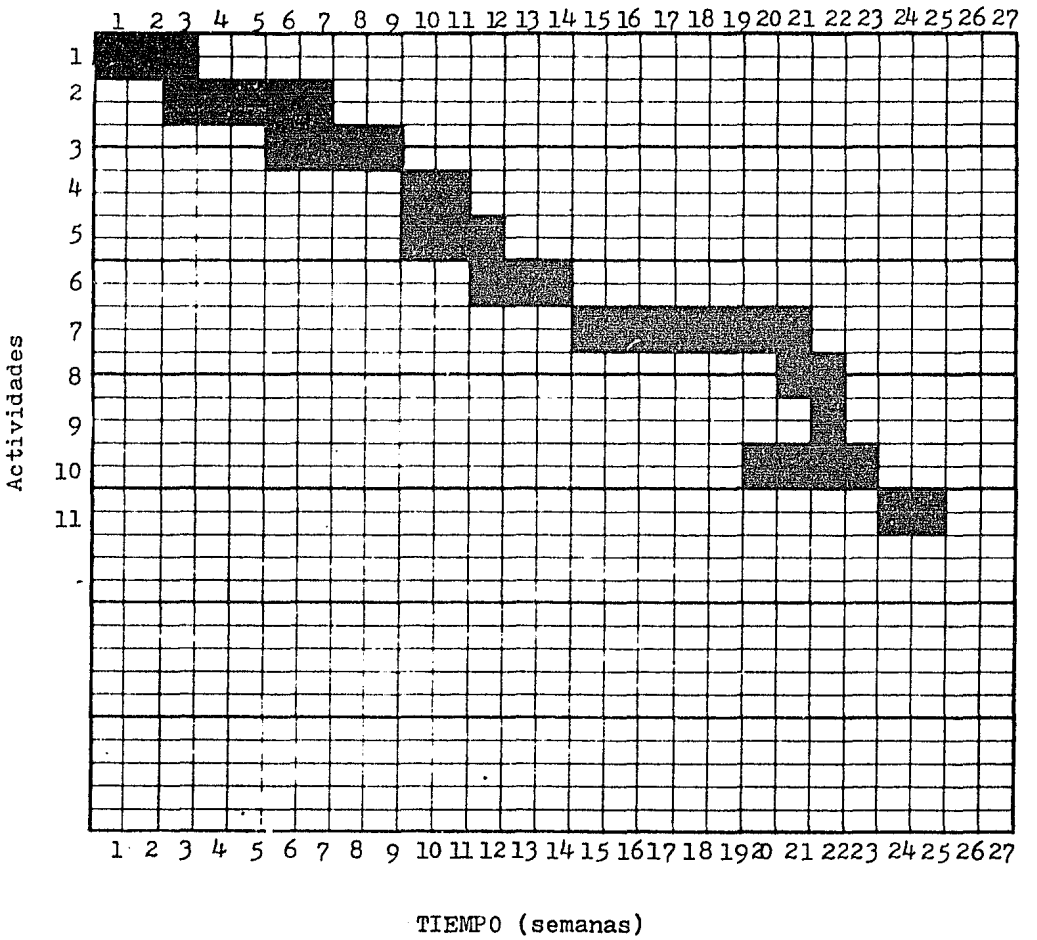
16.- Isaacson, W. B.
Statistical analyses for multivariable sistems.
Chemical Engineering
Junio 29, 1970, pp 69-75

I) CRONOGRAMA DE ACTIVIDADES

A continuación se describen las actividades a realizar y el tiempo estimado que llevara cada una de ellas. En el caso de las actividades que relacionan la elaboración escrita de la tesis, la primera parte se refiere a los antecedentes, así como la descripción de las técnicas de diseño experimental y análisis estadístico y la enmarcación dentro de ellas de la metodología de superficie de respuesta. La segunda parte tratara sobre el programa en sí, y la técnica utilizada en su elaboración, incluyendo los resultados de las corridas de prueba efectuadas. La parte final de la tesis incluirá el manual de uso.

	Actividad	Tiempo(semanas)
1	Obtención de clave de computadora, cinta magnética y manuales FORTRAN.	3
2	Obtención y comprensión de información bibliográfica.	5

	Actividad	Tiempo(semanas)
3	Elaboración de tesis 1ª - parte.	4
4	Revisión 1ª parte de tesis por parte del asesor, y corrección de errores.	2
5	Elaboración de diagrama de flujo y sintaxis del programa.	3
6	Codificación y perforación	3
7	Corridas de prueba, verificación y corrección de errores.	7
8	Obtención de gráficas a incluir como ejemplos.	2
9	Elaboración de manual de uso	1
10	Elaboración de tesis 2ª parte	4
11	Revisión y aprobación	2


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
1
2
3
4
5
6
7
8
9
10
11
Actividades
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
TIEMPO (semanas)

2

INTRODUCCION

Actualmente, la mayoria de las ramas de la ciencia estan usando diversas técnicas de optimización para hacer más eficientes los modelos matemáticos con los cuales trabajan; las principales causas de este fenómeno lo constituyen la intensa competencia existente en el área de procesos industriales y la creciente necesidad de poder reconocer y resolver en forma más rápida los problemas que surgen dentro del marco de la economía, la sociología y la ingeniería. En nuestro país, sin embargo, el uso de estos métodos se ha restringido a algunas carreras universitarias (Ingeniería, Estadística, Computación) y a Institutos de Investigación.

Este trabajo se elabora con el fin de introducir en él los conceptos fundamentales sobre los cuales se basan los métodos de optimización por medios estadísticos, específicamente, por modelos lineales de regresión y diseño de experimentos.

Estos métodos han probado ser muy eficaces en el área de Ingeniería Química[10,17,36,37] sin embargo es difícil manejarlos debido a la complejidad de los cálculos involucrados. Para subsanar esta limitante, se ha recurrido al uso de la computadora; los métodos basados en modelos lineales son altamente susceptibles de ser programados; sin embargo, obviamente, la creación de software* de este tipo requiere gran cantidad de recursos, como son tiempo de cómputo, conocimiento del método a implementar y programadores de alta calidad; debido a lo anterior, comúnmente

---

* El software de una computadora es el conjunto de programas que están escritos en un lenguaje apropiado a la estructura física de las máquinas, y con los cuales es posible hacer uso de ellas; este

se recurre al alquiler de paquetes elaborados en alguna otra parte del mundo (S.A.S.[28], G.P.S.S.[**]), lo cual genera una fuente importante de fuga de divisas. Entre los programas usados en una investigación de este tipo, uno de los que consume más tiempo y memoria es el que genera las llamadas "superficies de respuesta", debido al alto número de puntos a graficar, y a la sobreimpresión utilizada en ellos. La UNAM cuenta con un paquete estadístico (BASIS), y entre los pocos módulos que le faltan, se encuentra precisamente el que genera las "superficies de respuesta". De ahí, que uno de los resultados más notables de este trabajo, es precisamente haber creado dicho programa, con lo cual, los estudiantes de Ingeniería Química de la ENEP Zaragoza, a los que se dirige este trabajo, podrán implementar en la practica, los conocimientos aquí adquiridos. Esto no quiere decir, que solamente ellos podrán hacer uso de este programa; cualquier otra persona podrá usarlo, sin costo alguno, siempre y cuando su interés sea meramente académico y en beneficio de la UNAM.

---

* término no tiene sinónimo en el idioma español, por lo cual se utilizara con ese significado de aquí en adelante.

**
G.P.S.S.: General Purpose Systems Simulator.- Es un paquete de simulación de sistemas discretos; Se ha utilizado para simular desde modelos de guerra submarina hasta modelos de crecimiento demográfico.

3

FUNDAMENTACION DEL TEMA

La metodología de superficie de respuesta es una técnica que abarca todas las etapas necesarias para el desarrollo de modelos matemáticos (generalmente empíricos) con combinación - óptima de los niveles de los factores involucrados.

Esta técnica es ampliamente usada en el campo de la Ingeniería Química, aplicandose a la optimización de condiciones de operación de equipos como columnas de destilación, reactores, filtros y en general, de cualquier equipo relacionado con - las operaciones unitarias; sin embargo, la aplicación de la técnica requiere de conocimientos especializados, así como de tiempo de proceso en alguna computadora que tenga implementados uno o - vàrios paquetes estadísticos.

En México existen pocos paquetes de este tipo, y la mayoría no tienen implementado el software necesario para llevar a cabo este tipo de investigación. Además, el costo de CPU en paquetes rentados es altísimo, ya que de 2000 pesos/min. que costaba en promedio en 1982, actualmente se ha elevado a mas de - 15000/min.

Un amplio programa de investigación puede requerir de 2 a 3 horas de CPU*; esto en el caso de modelos lineales, - en donde los programas que más tiempo consumen son el graficador de superficies de respuesta, y los métodos especiales de análisis como el stepwise** y el forward***, representa un costo aproximado de 250 000 a 400 000 pesos, en el solo tiempo de máquina. Adicionalmente, habría que evaluar los costos de experimentación, de salarios del personal especializado, de depreciación del equipo experimental, etc., que pueden multiplicar por varias veces ese costo.

---

*El CPU es la Unidad de Procesamiento Central en una computadora, -

Lo anterior ha traído como consecuencia que estas técnicas no sean utilizadas dentro de la Industria Mexicana.

Obviamente, el implantar la infraestructura para la aplicación de estas técnicas, tal y como lo estan en paises desarrollados, no es la solución, sino adecuar los recursos disponibles para aprovechar al máximo estos métodos, sin efectuar grandes costos de implemetación.

En el caso de la UNAM, y particularmente, de la - ENEP Zaragoza, la existencia de una planta piloto facilita la realización de experimentos, los cuales serían reducidos, ya que se planearían estadísticamente; para la etapa de análisis, puede usarse - el paquete BASIS (que se complementa con el programa desarrollado - aquí) y de esa forma conocer los factores dominantes, para posteriromente, investigar la combinación óptima de sus niveles. Esto - pudiera conducir a varias etapas de experimentación-análisis antes de lograr dicho objetivo.

Dado que este trabajo no se limita a la elaboración del programa graficador de superficies de respuesta, sino que describe los lineamientos de la metodología, así como los funda- -

---

* y realiza las siguientes funciones: Entrada de datos a la memoria principal, operaciones lógico matematicas con dichos datos, salida de resultados a perifericos como la impresora o el disco, y el control implicito a cada una de esas acciones.

** Es un método en el cual se analizan los factores, introduciendolos uno por uno, de tal forma, que solo permanecen en el modelo los - factores que en el momento de introduccion, provocaron un cambio - significativo en el valor de F (capitulo 7), y ademas, en el momento de introducción de otra variable, su valor de F especifico, no tuvo disminución significativa.

*** Igual que el stepwise, pero sin efectuar el segundo paso.

mentos en los cuales se basa, los alumnos podrán contar con los recursos necesarios para aplicarla, proveendolos de un método definido para observar, comprender y valorar la importancia de las variables involucradas en un determinado proceso. Con este tipo de actividades, los alumnos podrán tener una sólida formación práctica, ya que podrían manipular varios equipos completamente, y no solo a nivel de demostración, ahorrando al mismo tiempo recursos en la planeación y experimentación, así como en la etapa de análisis, ya que el costo de usar el paquete BASIS es el más bajo de la República; - esto, a mediano plazo, redundaría en una mayor competitividad de - los productos elaborados con tecnología nacional, que es lo que se necesita en un país como el nuestro.

4

PLANTEAMIENTO DEL PROBLEMA

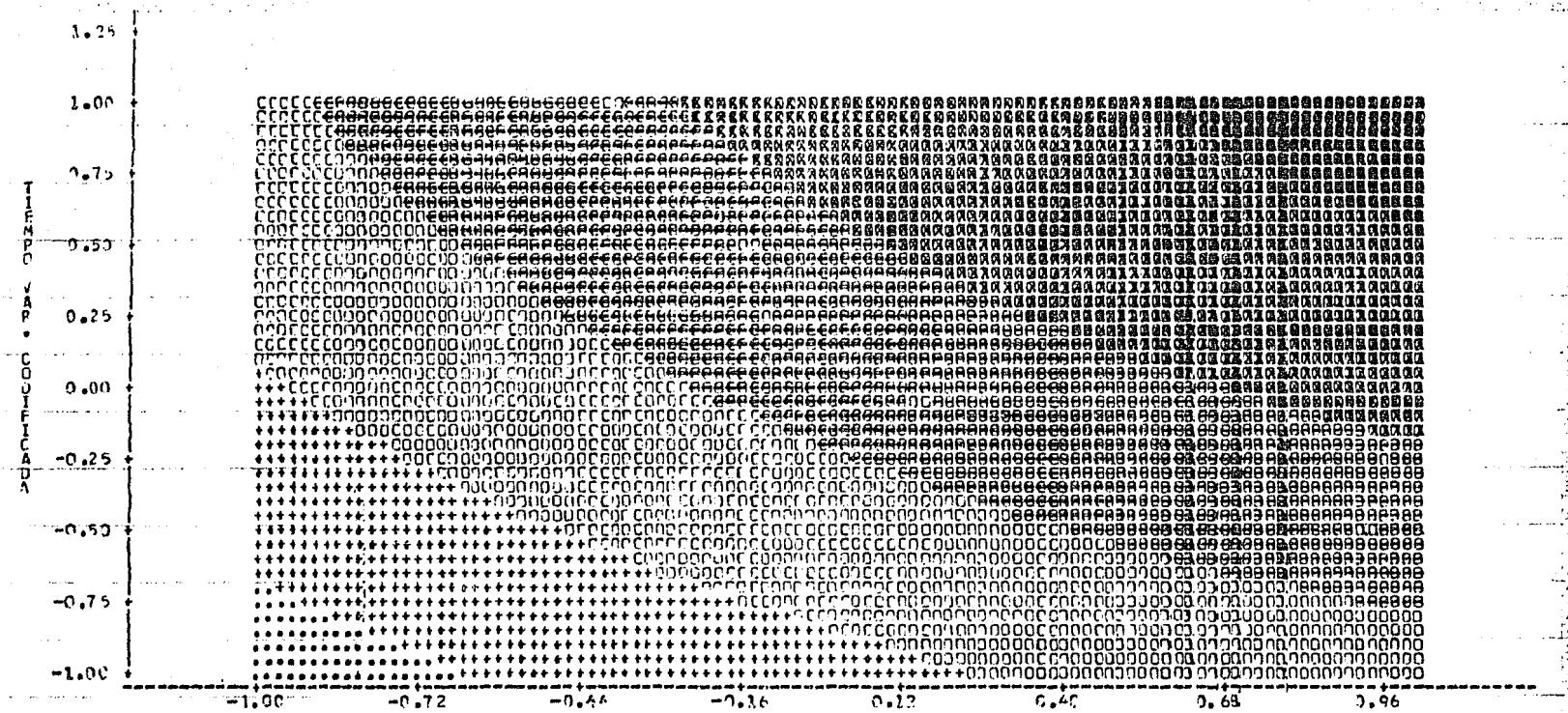
Fundamentalmente, se desea desarrollar un programa que grafique una superficie de respuesta. En la hoja siguiente, se muestra una superficie obtenida mediante el paquete S.A.S., y la cual consumió 1 min. 06 seg. en una IBM 370*. La siguiente hoja muestra el conjunto de instrucciones necesarias para obtener dicha gráfica. Puede verse, que tan solo usar este programa del paquete implica conocer sus instrucciones muy particulares. Para facilitar el uso del programa a desarrollar, la información de entrada se debe diseñar para que se lea en un archivo de datos, y no como líneas de otro, o aún del mismo, programa. Esta información la constituiría la ecuación de interes, y los rangos de valores de los factores X_1 y X_2 que se desea sean los ejes de la gráfica. Asimismo, el programa deberá imprimir los valores de estos factores, correspondientes a su ubicación geométrica, tal y como se muestra en la gráfica ejemplo.

El signo de impresión utilizado para representar la respuesta, consiste en un carácter compuesto; este y el rango de valores que incluye, debe escribirse abajo de la gráfica correspondiente.

Adicionalmente al programa, se deberá describir la metodología de superficie de respuesta, haciendo énfasis en los conceptos, más que en sus desarrollos matemáticos, con el proposito de que se reduzca el tiempo de comprensión de la técnica, y paralelamente, aplicarla eficientemente. Por último, se deberá elaborar un manual de uso claro y conciso.

---

*Modelo 150



SYMBOL		SMAXT	SYMBOL		SMAXT	SYMBOL		SMAXT
:::::::	2.3525680 =	2.6033248	CCCCCC	3.1124378 =	3.6183508	@@@@@@	4.1242638 =	4.6301768
++++++	2.6065248 =	3.1124378	@@@@@@	3.6183508 =	4.1242638	@@@@@@	4.6301768 =	4.8831333

NOTE: 6227 CAS HIDDEN

NOTE: THE JOB CONTORNO HAS BEEN RUN UNDER RELEASE 79.3B OF SAS AT COLEGIO DE POSTGRADUADOS.

```
1          DATA CONTORNO>
2          TEMP=-1>
3          PRES=-1.02>
4          PRES3:PRES+.02>
5          TI=-1.02>
6          TI2:TI+.02>
7          SMAXT=4.41961575+0.51131978*TEMP+.34827254*TI+.57411994*PRES-.12595940*TEMP*TEMP
8          -0.12707683*TI*TI-0.10609788*PRES*PRES-0.31941869*TEMP*TI-0.22288996*TEMP*PRES+0
9          .06110767*TI*PRES>
10         OUTPUT>
11         IF TI2< THEN GO TO TI3>
12         IF PRES2< THEN GO TO PRES3>
13         FORMAT SMAXT 12.7>
```

NOTE: DATA SET WORK.CONTORNO HAS 10404 OBSERVATIONS AND 4 VARIABLES. 232 OBS/TRK.
NOTE: THE DATA STATEMENT USED 20.13 SECONDS AND 136K.

```
14         PROC PLOT NOLEGEND>
15         PLOT TI*PRES=SMAXT/CONTOUR=6>
16         LABEL TI=TIEMPO VAR. CODIFICADA>
17         LABEL PRES=PRESION VAR. CODIFICADA>
18         TITLE SUPERFICIE CON LA TEMPERATURA A NIVEL BAJO>
```

NOTE: THE PROCEDURE PLOT USED 37.48 SECONDS AND 150K AND PRINTED PAGE 1.

NOTE: SAS USED 150K MEMORY.

NOTE: SAS INSTITUTE INC.
SAS CIRCLE
BOX 8000
CARY, N.C. 27511

5

OBJETIVOS

I.- Describir técnicas de diseño de experimentos así como su enmarcación en la metodología de superficie de respuesta, poniendo enfásis en el fundamento y el análisis a seguir en los modelos estadísticos lineales, para que dicha metodología sea utilizada como guía para:

a) La definición de condiciones de operación en experimentos.

b) El delineamiento del análisis a seguir de los resultados obtenidos.

c) La optimización de la operación del proceso de interés.

II.- Proveer de la parte medular del material necesario para aplicar la metodología de superficie de respuesta, elaborando un programa que grafique dicha superficie, para el caso de dos factores independientes cuantitativos (superficie de respuesta bidimensional).

6

MATERIAL Y METODO

Para la elaboración del programa, y con el fin de que el programa objeto pueda ser soportado por cualquier tipo de máquina que posea ese compilador, será utilizado el lenguaje FORTRAN, ya que es el único, con caracteristicas cientificas, que es uniforme en sus versiones. De igual manera, se cuidara que el conjunto de caracteres a usar en la impresión sea un subconjunto del código ASCII*, debido a que este es el código universalmente aceptado para la comunicación entre CPU y periféricos, o bien entre sistema-sistema de cómputo. El programa se desarrollará básicamente en modulos, auxiliandose en cada uno de ellos de la programación estructurada (en lo posible, ya que el lenguaje no es estructurado), y en asignación dinámica de memoria, con el objeto de que el programa y el subsecuente proceso pueda ser corrido en cualquier máquina con al menos 32 KB de memoria (la IBM 370 ocupa de 192 a 320 KB).

---

*American Standard Character Interchanged Information; utiliza 7 bits.

7

DESARROLLO

7.1 MODELOS ESTADISTICOS LINEALES

7.1.1 DESARROLLO HISTORICO DE LOS MODELOS LINEALES.

La Estadística forma parte importante de la Matemática en general. Su origen usualmente está ligado tanto a los juegos de azar como a la ahora llamada ciencia política. Los estudios en probabilidad dieron lugar al tratamiento matemático de los errores, y las leyes resultantes guiaron a la teoría que hoy forma la base de la Estadística Matemática o Inferencial; por otro lado el interés en el análisis de fenómenos políticos guió a la llamada Estadística Descriptiva.

La Estadística trabaja con modelos matemáticos que toman en cuenta los aspectos aleatorios de los fenómenos, y se conocen como modelos estocásticos, o estadísticos. Los modelos estadísticos lineales son los más sencillos de este tipo, y su fuerte base teórica los ha hecho muy útiles en gran número de situaciones; esa base, se ha moldeado a traves de varios siglos, y entre los avances más actuales, se encuentra la formulación de la llamada "teoría de decisiones"[31].

Gitolomo Cardano fue el primero en observar la regularidad de los fenómenos aleatorios[35], aplicando intuitivamente la probabilidad teórica, para llegar a ser un jugador de renombre. Tiempo después, el Caballero de Méré, jugador profesional, propuso al matemático Pascal que encontrara la solución del "problema de puntos", referente al reparto equitativo de las apuestas

en un juego cuando se interrumpe una partida antes de acabarse. El intercambio de correspondencia entre él y Fermat, sentó las primeras bases de la teoría de la probabilidad.

Intuitivamente, se puede considerar a la probabilidad* como una medida de la oportunidad que hay de que ocurra un determinado proceso, y a dicho proceso, se le conoce como fenómeno estocástico o estadístico.

Desde un punto de vista estricto, todos los modelos matemáticos son estadísticos, lo que sucede es que la probabilidad asociada a ellos puede ser muy grande, muy baja, 1 ó 0, por lo cual no se toma esta en cuenta explícitamente**.

A. M. Legendre fue el primero en proponer un modelo lineal, dentro de su obra sobre métodos nuevos para la determinación de las órbitas de los cometas[30]. El modelo, que no poseía propiedades distribucionales explícitas, era:

$$e_i = \sum_{j=1}^{q} B_j z_{ji} - x_i \quad (i = 1,2,.....,n;\ n \geq q)$$

donde

X_i mediciones en estudio

Z_{ji} coeficientes conocidos

B_j variables desconocidas

e_i errores

El principio propuesto por Legendre fue la minimi-

---

*Ver apéndice A

**Ver apéndice B

zación, por variaciones de B_j, de la suma de cuadrados de los errores, empleando inconcientemente, lo que en optimización se conoce como "función objetivo".

La primera discusión del modelo de Legendre fue hecha por C. F. Gauss, quien postuló que los errores e_i tenían una distribución normal, descrita ya por P. S. Laplace, quien fue el primero que consideró el concepto de distribución de errores. El propio Gauss determinó en 1825 las propiedades distributivas de los estimadores obtenidos al aplicar el método de mínimos cuadrados.

En 1837, el matemático alemán Hagen postuló que:

1.- Un error observado es la suma algebraica de un número muy grande de errores elementales infinitesimales de igual magnitud.
2.- Los errores elementales positivos y negativos se producen con igual frecuencia al considerar muchos casos.
3.- La contribución de los errores elementales es independiente una de otra.

Galton[16], al estudiar alturas de una población, observó que los hijos de padres altos, aunque con tendencia a ser altos, eran, en promedio, más bajos que sus padres; igualmente los hijos de padres bajos, con tendencia a ser bajos, en promedio, eran más altos que sus padres. En base a esto, en 1886, postuló la ley de regresión universal: "Toda peculiaridad de un hombre es compartida por su pariente, pero, en promedio, en grado menor".

K. Pearson demostró que la estimación de parámetros mediante mínimos cuadrados es la combinación lineal de las z_i que maximiza la correlación con la x_i. Ademas tuvo el mérito de extender el uso de los modelos lineales propuestos por Gauss, a una clase de problemas mucho más amplia que los problemas de medición de constantes físicas. Esto lo hizo entre 1897 y 1898.

En 1901, Liapunov demostró que los errores elementales, aunque no fuesen de igual magnitud, tenderían a distribuirse normalmente. De esta manera se observaba ya, incipientemente, la base del teorema de límite central.

Brunt, en 1917, definió los errores sistemáticos, constantes y accidentales, señalando que los dos primeros deben eliminarse al planear y corregir un experimento. Los errores accidentales los considera inevitables e irregulares, con las siguientes propiedades:

1.- Un gran número de errores accidentales muy pequeños están presentes en cualquier observación.
2.- Los errores positivos y negativos son igualmente presentes.
3.- El error total no puede exceder de una cantidad razonablemente pequeña.
4.- La probabilidad de un error pequeño es mayor que la de uno grande.

sin embargo, le faltó señalar que los errores accidentales se consideran independientes entre sí, lo cual Hagen ya había observado.

R. A. Fisher[15], matemático inglés, entre 1920 y - 1930, desarrolló trabajos que impulsaron principalmente las técnicas para el uso de los modelos lineales en Estadistica; entre sus contribuciones se encuentran:

1.- Desarrolló el concepto de prueba de hipótesis así como el método para efec- -- tuarlas; esto condujo a la distribución Z de Fisher, que con un ligero cambio, - fue nombrada F en honor a Fisher, por - G. W. Snedecor.

2.- Consideró la introducción de valores de las variables independientes x_i iguales a uno o cero para designar presencia o - ausencia de factores de tipo cualitativo que afecten a los datos.

3.- Introdujo el concepto de bloque*, para - reducir la variablidad de los experi- - mentos.

4.- Consideró que era necesario que las observaciones se hicieran en un orden determinado aleatoriamente, así como también que en forma aleatoria se asignaran las variantes en estudio a las diferen-- tes unidades experimentales. En experimentación, la aleatorización se puede - considerar como la única innovación verdaderamente moderna.

---

*Grupo de experimentos con caracteristicas comunes (Capitulo 9).

5.- Puso énfasis en el uso de varios factores simultáneamente e inicio las ideas de independencia u ortogonalidad entre los factores estudiados.

A partir de ese momento, se han aclarado innumerables conceptos y se han refinado las técnicas de los modelos lineales. Esas técnicas tienen una pequeña variación dependiendo del campo en que se apliquen. En este trabajo se hace énfasis sobre el campo industrial, y por lo tanto, puede haber resultados erróneos si se aplican indiscriminadamente tales métodos a otras áreas.

7.1.2 EL MODELO LINEAL $Y = \mu + \varepsilon_i$.

7.1.2.1 Regresión Lineal entre dos Variables.- El modelo lineal estadistico más sencillo, es aquel que involucra una variable dependiente y una independiente.

En muchos trabajos experimentales se presenta este caso. En electricidad, si la resistencia de un circuito es constante, la intensidad I varía directamente proporcional con el voltaje V aplicado, según la ley de Ohm. Igualmente, en un recipiente a presión constante, un gas contenido en él aumentara su volumen si aumenta su temperatura, según la ley de Charles.

Suponiendo que no se conoce la ley de Ohm, y que se deseara conocer la relación entre el voltaje y la intensidad de un circuito dado, se podrían cambiar los valores de V y observar I, con R cte., graficar estos valores en un diagrama V-I, y el

conjunto de puntos definiría "más o menos", una línea recta que pasaría por el origen (fig. 7.1). Se dice más o menos, porque aun-

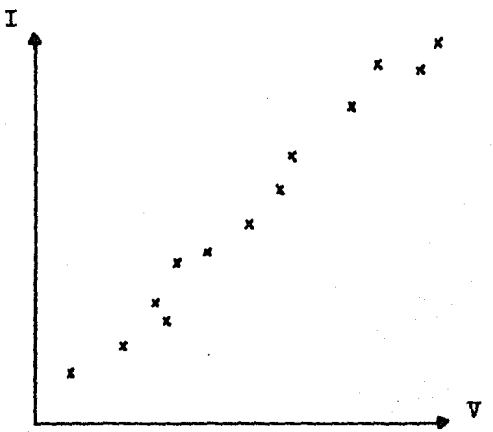


Fig. 7.1

que la relación verdadera es exactamente lineal, las medidas hechas estaran sujetas a pequeños errores aleatorios.

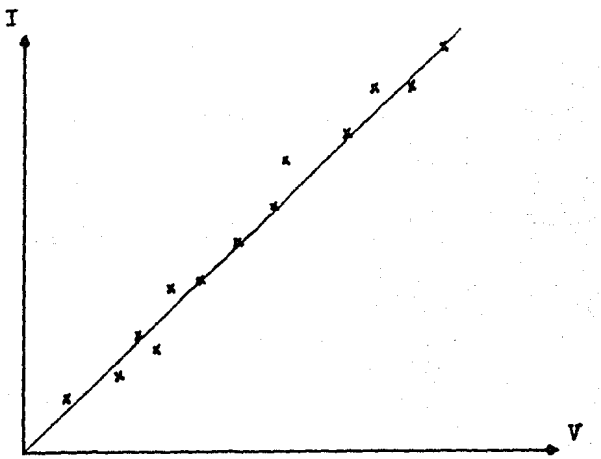


Fig. 7.2

Considerando esto último, se podría trazar una línea recta de tal manera que se intentara que la mayoría de los puntos quedaran dentro de ella (Fig. 7.2). A la distancia que existe entre un par ordenado (V_i, I_i) de la recta, y un punto experimental (x_i, y_i) colocado en $x_i = V_i$, se le llama error aleatorio y se representa por ε_i.

En ocasiones, no solo el error aleatorio es provocado por fallas en las mediciones, sino que ádemas se suma a él una variación característica de la respuesta observada. Así por ejemplo, al querer estimar una relación entre el peso y la altura de hombres adultos, se pudieron haber hecho observaciones y graficado estas en un diagrama altura-peso. Acto seguido, se procedería a trazar una recta que pase por la mayoría de los puntos posibles. En este caso, el trazar una determinada recta sería muy subjetivo

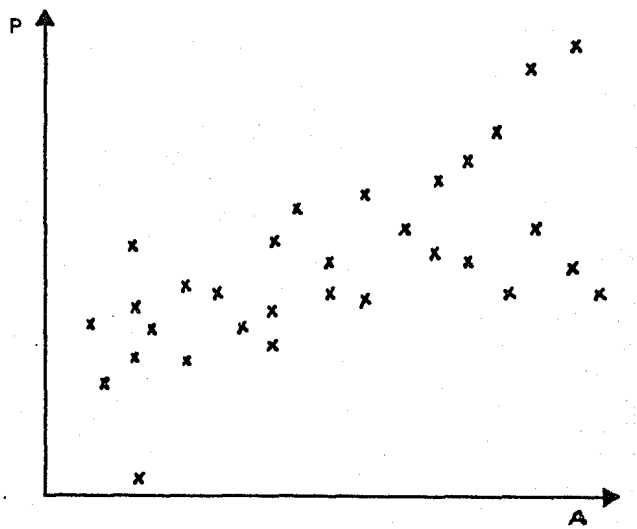


Fig. 7.3

(Fig. 7.3), ya que esta puede tener casi cualquier inclinación; esto es, una recta no da una relación satisfactoria entre altura y peso. Ante esto, se puede tomar el promedio de los pesos de una altura determinada, y generar la Fig. 7.4. Aquí el trazo de una rec-

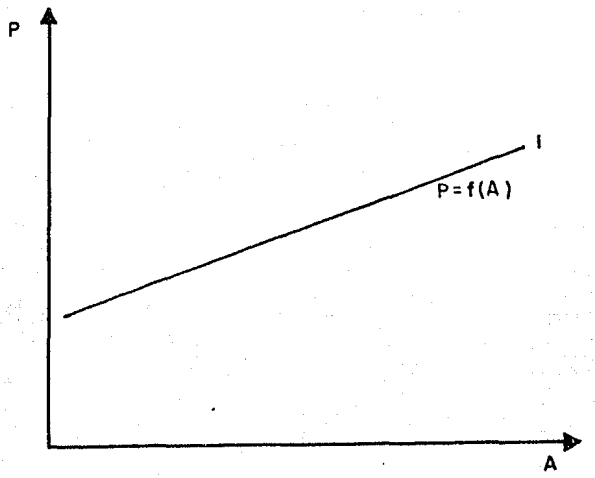


Fig. 7.4

ta es más objetivo, y la curva generada recibe el nombre de "curva de regresión" del peso sobre la altura. Se puede observar, sin embargo, que también se puede generar una curva de regresión de altura-peso (Fig. 7.5). Ambas curvas no son iguales, y sin embargo proveen información útil.

Supongase que se tienen medidas de la altura de algunos individuos, pero no su peso, y que deseamos estimar este. Lo que procedería sería tomar la recta 1 y hallar el peso promedio observado para cada altura determinada, sirviendo esto como una estimación del peso que no se midió.

Igualmente, si se tienen pesos registrados, se po-

dría hallar la altura promedio y usarla como estimación del peso - registrado.

Para poder obtener una y solo una ecuación de regresión, es necesario que la respuesta Y sea una variable aleatoria y la cantidad X, sea variable, pero no aleatoria. En este caso

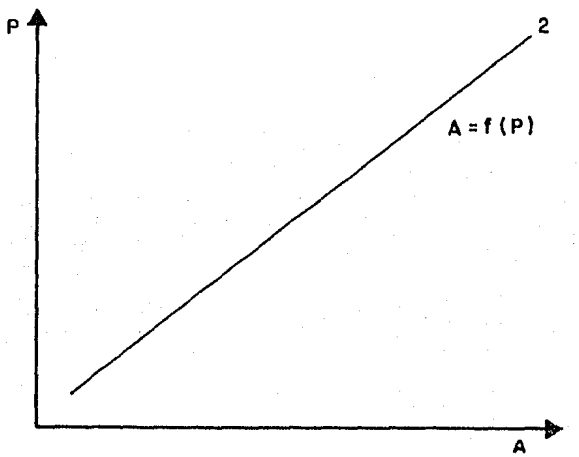


Fig. 7.5

ambas variables, peso y altura, son aleatorias y siguen algún ti--

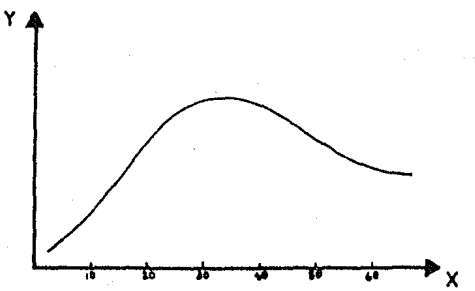


Fig. 7.6

po de distribución de probabilidad bivariada; por esta razón, se tu

vo que recurrir al uso de pesos promedios y alturas promedio.

Como se ha observado, una relación lineal es útil en muchos casos, y puede serlo igualmente en casos en que se sabe que la relación no es lineal. Las funciones de crecimiento obviamente no son lineales (Fig. 7.6). Sin embargo, si consideramos el rango $0 \leq X \leq 30$, una línea recta puede proveer una adecuada representación del fenómeno. Esta relación, por otro lado, no puede usarse con fines de predicción, porque se corre el riesgo de tomar conclusiones incorrectas.

Una recta se puede representar matemáticamente por:

$$\hat{Y} = b_o + b_1 X \qquad (7.1)$$

b_o y b_1 son llamados los parámetros del modelo y corresponden a la ordenada al origen y a la pendiente, respectivamente. $\hat{Y}$ es el valor predicho o estimado de Y para una X dada. El modelo es lineal y de primer orden. En regresión, lineal se refiere a los parámetros, y el orden es la potencia más alta a la que una variable independiente esta elevada dentro del modelo, así:

$$\hat{Y} = b_o + b_1 X + b_{11} X^2 \qquad (7.2)$$

es un modelo lineal de segundo orden.

Obviamente, los valores de b_o y b_1 se desea que sean tales, que el valor de $\hat{Y}_i$ sea lo más aproximado posible al valor de Y_i, para $i = 1,\ldots,n$. Esto se puede representar como:

$$Y_i = b_o + b_1 X_i + \varepsilon_i \qquad \varepsilon_i \text{ sea el mínimo para } i \qquad (7.3)$$

sustituyendo la Ec. 7.1 en la Ec. 7.3:

$$Y_i - \hat{Y}_i = \xi_i = Y_i - (b_o - b_1 X_i) \qquad (7.4)$$

la suma de los cuadrados de los errores para la recta verdadera - sería:

$$S = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} (Y_i - \beta_o - \beta_1 X_i)^2 \qquad (7.5)$$

β_o y β_1 serían los coeficientes verdaderos, y se obtendrían si se tomara en cuenta la población total específica.

Dado que comunmente se trabaja con muestras, se usan los estimadores b_o y b_1 respectivamente:

$$S = \sum_{i=1}^{n} \xi_i^2 = \sum_{i=1}^{n} (Y_i - b_o - b_1 X_i)^2 \qquad (7.6)$$

para que S sea mínimo, diferenciamos parcialmente con respecto a - b_o y b_1:

$$\frac{\partial S}{\partial b_o} = -2 \sum_{i=1}^{n} (Y_i - b_o - b_1 X_i) \qquad (7.7a)$$

$$\frac{\partial S}{\partial b_1} = -2 \sum_{i=1}^{n} (X_i (Y_i - b_o - b_1 X_i)) \qquad (7.7b)$$

igualando a cero y efectuando sumatorias

$$b_o n + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \qquad (7.8a)$$

$$b_o \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \qquad (7.8b)$$

Las ecuaciones 7.8 se conocen como ecuaciones -

normales, y su solución es:

$$b_1 = \frac{\sum X_i Y_i - \left[(\sum X_i)(\sum Y_i)\right]/n}{\sum X_i^2 - (\sum X_i)^2/n} \qquad (7.9a)$$

$$b_o = \sum Y_i/n - b_1 \sum X_i/n \qquad (7.9b)$$

donde todas las sumatorias van desde i = 1,...,n.

Si se definen $\bar{X} = \sum X_i/n$, $\bar{Y} = \sum Y_i/n$, y ádemas se hacen algunos rearreglos, las Ecs. 7.9 pueden escribirse como:

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \qquad (7.10a)$$

$$b_o = \bar{Y} - b_1 \bar{X} \qquad (7.10b)$$

La cantidad $\sum X_i^2$ es llamada la suma incorrecta - de cuadrados de las X's y $(\sum X_i)^2/n$ es la correción para la media de las X's. La diferencia $\sum X_i^2 - (\sum X_i)^2/n$ es la suma de cua- - drados correcta de las X's. Similarmente, $\sum X_i Y_i$ es llamada la suma no corregida de productos, $(\sum X_i)(\sum Y_i)/n$ es la corrección - para las medias y la diferencia es llamada la suma correcta de - productos de X y Y.

Substituyendo la Ec. 7.10b en la Ec. 7.4, resulta:

$$\hat{Y} = \bar{Y} + b_1(X - \bar{X}) \qquad (7.11)$$

o bien

$$\hat{Y} = \bar{Y} + b_1 X - b_1 \bar{X} \qquad (7.12)$$

$$\hat{Y} = (\bar{Y} - b_1 \bar{X}) + b_1 X \qquad (7.13)$$

Si se toman como datos las columnas 1 (Y) y 2 - (X) de la tabla C.1, y efectuando las operaciones indicadas, se obtiene:

$$n = 25$$

$$\sum Y_i = 235.6$$

$$\bar{Y} = 9.42$$

$$\sum X_i = 1315.$$

$$\sum X_i Y_i = 11821.43$$

$$\sum X_i^2 = 76323.42$$

$$b_1 = -0.079892$$

$$Y = 9.424 + 0.079829(52.60) - 0.079829\ X$$

$$Y = 13.623005 - 0.079829\ X$$

con esta ecuación y los valores correspondientes reales se puede - generar la tabla de residuos 7.1.

Note que la ecuación 7.11 se puede escribir:

$$Y_i - \hat{Y}_i = (Y_i - \bar{Y}) + b_1(X_i - \bar{X})$$

o

$$\sum (Y_i - \hat{Y}_i) = \sum (Y_i - \bar{Y}) - b_1 \sum (X_i - \bar{X}) = 0$$

o sea que

$$\sum \xi_i = 0$$

este resultado se obtendrá siempre en problemas de regresión.

Con la tabla de residuos es posible observar si la ecuación ajustada es satisfactoria. Nuevamente, sin embargo, se

Número de Observación	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	10.98	10.81	0.17
2	11.13	11.25	-0.12
3	12.51	11.17	1.34
4	8.40	8.93	-0.53
5	9.27	8.72	0.55
6	8.73	7.93	0.80
7	6.36	7.68	-1.32
8	8.50	7.50	1.00
9	7.82	7.98	-0.16
10	9.14	9.03	0.11
11	8.24	9.92	-1.68
12	12.19	11.32	0.87
13	11.88	11.38	0.50
14	9.57	10.50	-0.93
15	10.94	9.89	1.05
16	9.58	9.75	-0.17
17	10.09	8.89	1.20
18	8.11	8.04	0.07
19	6.83	8.04	-1.21
20	8.88	7.68	1.20
21	7.68	7.87	-0.19
22	8.47	8.98	-0.51
23	8.86	10.06	-1.20
24	10.36	10.96	-0.60
25	11.08	11.34	-0.26

Tabla 7.1

busca ser lo más objetivo posible, y se utiliza un método que se conoce como Análisis de Varianza. Considere la siguiente identidad:

$$Y_i - \hat{Y}_i = Y_i - \bar{Y} - (\hat{Y}_i - \bar{Y}) \qquad (7.14)$$

$$\sum (Y_i - \hat{Y}_i)^2 = \sum \left[(Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y}) \right]^2$$

$$\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 + \sum (\hat{Y}_i - \bar{Y})^2 -$$

$$-2 \sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})$$

el tercer término se puede desarrollar; sustituyendo la Ec. 7.11:

$$-2\sum(Y_i - \bar{Y})b_1(X_i - \bar{X}) = -2b_1\sum(Y_i - \bar{Y})(X_i - \bar{X})$$

con la Ec. 7.10a

$$= -2b_1^2\sum(X_i - \bar{X})^2$$

y con la Ec. 7.11

$$= -2\sum(\hat{Y}_i - \bar{Y})^2$$

así, la ecuación 7.14 queda como

$$\sum(Y_i - \hat{Y}_i)^2 = \sum(Y_i - \bar{Y})^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

o bien

$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 \qquad (7.15)$$

donde

$Y_i - \bar{Y}$ es la desviación de la i-ésima observación con respecto a la media.

$Y_i - \hat{Y}_i$ es la desviación de la i-ésima observación <u>real</u> con respecto a su valor predicho o ajustado $\hat{Y}_i$.

$\hat{Y}_i - \bar{Y}$ es la desviación del i-ésimo valor predicho con respecto a la media.

Así, la Ec. 7.15 se puede expresar como:

Suma de cuadrados con respecto a la media.	=	Suma de cuadrados con respecto a la regresión	+	Suma de cuadrados debida a la regresión
SS_{rm}		SS_{rr}		SS_{dr}

Esto significa que la variación de las Y's con respecto a su media, se compone de una parte debida a la línea de regresión, y otra parte debido a que las observaciones no caen en su totalidad sobre la línea de regresión; esto graficamente sería representado en la Fig. 7.7:

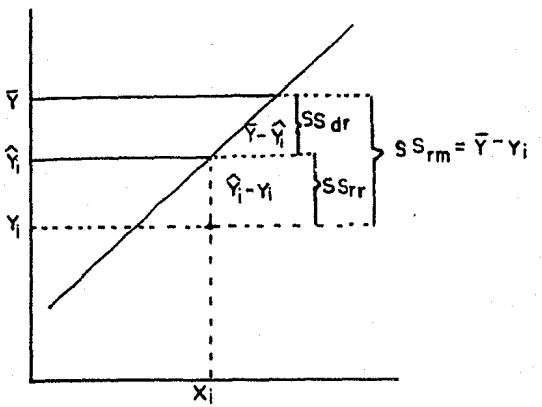


Fig. 7.7

Obviamente, es deseable que las observaciones se encuentren lo más cercanas posible de la recta, o sea, que SS_{dr} sea mucho mayor que SS_{rr}; si se define un parámetro R^2 tal que:

$$R^2 = \frac{SS_{dr}}{SS_{rm}}$$

deseamos que este valor sea lo más cercano posible a la unidad, con lo cual aseguramos que SS_{rr} es muy pequeño:

$$R^2 = \frac{SS_{dr}}{SS_{rr}} = \frac{SS_{rm} - SS_{rr}}{SS_{rm}}$$

Las sumas de cuadrados (SS) siempre tienen asociado un número llamado grados de libertad. Al igual que en Termodinámica, los grados de libertad dan el número de observaciones cuyo valor no esta restringido por alguna función matemática; es decir, su valor no se ve afectado por los valores que tomen las otras observaciones. Por ejemplo, si suponemos que en un recipiente cerrado se encuentran 3 compuestos, al especificar la fracción mol de dos de ellos, el 3° automaticamente depende de los otros 2, debido a la ecuación:

$$X_1 + X_2 + X_3 = 1$$

$$X_3 = 1 - X_1 - X_2$$

aquí el número de grados de libertad es precisamente dos; en general, para n-componentes, el número de grados de libertad a especificar con respecto a la fracción mol es n-1.

Para la SS_{rm} se tienen n-1 grados de libertad. Esto es debido a que se pueden fijar n-1 relaciones del tipo:

$$Y_i - \bar{Y}$$

pero la última relación debe ser tal, que se cumpla que:

$$\sum Y_i/n = \bar{Y}$$

de esta forma, si tenemos 5 datos, y la media es 10, solo (5-1) valores seran independientes; si estos fueran 3,5,9,21, el 5° necesariamente debe ser 12.

Para la SS_{dr}, existe un grado de libertad ya que con el solo valor de b_1, queda fija la relación:

$$(Y_i - \bar{Y}) = b_1(X_i - \bar{X})$$

Los grados de libertad para la SS_{rr} se obtienen por sustracción:

$$DF_{SS_{rm}} = DF_{SS_{rm}} - DF_{SS_{dr}} \qquad (7.16)$$

donde DF son grados de libertad (Degrees of Freedom).

De esta forma, se pueden construir tablas de ANOVA (Analysis of Variance):

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	MEDIA CUADRADA	F
Regresión	$b_1 \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$	1	MS_R	$\frac{MS_R}{s^2}$
Residual	Por diferencia	$n - 2$	$s^2 = \frac{SS_{rr}}{(n-2)}$	
Total, con respecto a la media	$\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}$	$n - 1$		

Tabla 7.2

Otro tipo de tabla se presenta en la 7.3; lo útil de este tipo de tabla, es que los coeficientes se van validando uno a uno, para de esa forma conocer si son significativos o no. Esto se explica más detalladamente en la Sec. 7.1.2.3.

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	MEDIA CUADRADA	F
Regresión (b_o)	$SS(b_o) = \frac{(\sum Y_i)^2}{n}$	1		$\frac{SS(b_o)}{s^2}$
Regresión ($b_1 \mid b_o$)	$SS(b_1 \mid b_o) = b_1 \sum X_i Y_i - \frac{(\sum X_i)(\sum Y_i)}{n}$	1	MS_R	$\frac{SS(b_1 \mid b_o)}{s^2}$
Residual	Por diferencia	$n - 2$	$s^2 = \frac{SS_{rr}}{(n-2)}$	
Total, con respecto a la media (no corregida)	$\sum Y_i^2$	n		

Tabla 7.3

Aquí, $SS(b_o)$ es la corrección para la media de las Y's; $SS(b_1 \mid b_o)$ es la suma de cuadrados del coeficiente b_1 estando ya presente b_o; s^2 es un estimador basado, en n-2 grados de libertad, de la varianza de la regresión σ^2. El valor de F es el cociente MS_R/s^2. Las tablas para el ejemplo serían las 7.4a y la 7.4b.

Así como el valor de R^2 nos indica la precisión de la recta ajustada, el valor de F nos proporciona un parámetro para verificar si las conclusiones obtenidas con él modelo son confiables. Para aceptar un modelo o coeficiente como significativo, el valor de F obtenido deberá ser mayor al valor de F que se encuentra en tablas, para cierto nivel de significancia.

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	MEDIA CUADRADA	F
Regresión	45.5924	1	45.5924	57.54
Residual	18.2234	23	0.7923	
Total (corregida)	63.8158	24		

Tabla 7.4a

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	MEDIA CUADRADA	F
Regresión	2220.2944	1		2802.34
Regresión $(b_1 \mid b_o)$	45.5924	1	45.5924	57.54
Residual	18.2234	23	0.7923	
Total (no corregida)	2284.1102	25		

Tabla 7.4b

Hasta este punto, se ha desarrollado la base del método de mínimos cuadrados y se ha hecho una introducción al Análisis de Varianza. La forma presentada es la más correcta para un modelo lineal de primer orden con una sola variable independiente. No obstante, esto se ha hecho solamente con manipulaciones algebraicas sin tomar en cuenta los conceptos teóricos que nos permiten efectuar dichos cálculos. Sobre este tema se tratará en la sección siguiente.

7.1.2.2 Población y Distribución de Frecuencia.- Se llama población al conjunto de mediciones que se pueden efectuar sobre una característica común de un grupo de seres u objetos[30]. La característica común se obtiene al fijar un cierto número de factores como constantes, dejando a las demás variables involucradas variar entre individuo e individuo de la población. Ejemplos de poblaciones son:

1) Conjunto de mediciones en kilos, del peso de hombres adultos que trabajan como obreros en la planta A.
2) Conjunto de bombas defectuosas/mes producidas por una compañía en un lapso Z, al utilizar el material X para la carcaza y el material Y para el rotor.
3) Conjunto de valores, de la producción diaria de café granulado, en Kg, del Estado de Veracruz, durante la temporada Noviembre-Febrero de los últimos 10 años.

4) Conjunto de valores, de la temperatura de salida de las torres de enfriamiento ubicadas en el Valle de México, tomando datos cada 5 - minutos, cuando la potencia utilizada es de - 2000 KJ por día y la temperatura ambiente 30 grados centígrados.

Las poblaciones se clasifican de acuerdo al grado de generalidad que poseen. Una población tendra mayor grado de generalidad con respecto a otra población , si los factores que se mantienen constantes en la primera población son menos que los de la 2^a población.

Ejemplos de poblaciones con mayor grado de generalidad que las anteriores son:

1) Conjunto de mediciones, en Kg, del peso de hombres adultos que trabajan en la fábrica A.
2) Conjunto de bombas defectuosas/mes producidas en una compañia en un lapso Z.
3) Conjunto de valores de la producción diaria de café granulado, en Kg, de la República Mexicana en los últimos 10 años.
4) Conjunto de valores, de la temperatura de salida de las torres de enfriamiento ubicadas en el D.F., tomando datos cada minuto.

El último ejemplo es especial; frecuentemente se ignora la variación ocasionada por errores de medida, considerando que la mayor contribución a la variación entre individuos u -- objetos de la población se debe a las características mismas de -

ellos. Así, en ese ejemplo, si se considera esto último, la diferencia entre lapsos de medida dara mayor precisión, pero no crearа por sí sola una población con diferente grado de generalidad. Lo mismo ocurre con otras condiciones e instrumentos de medición. Muy rara vez, este tipo de aspectos se incluyen para definir una población.

Teóricamente, una población dada debe consistir de un número infinito de observaciones; en la práctica esto se logra con solamente asegurarse que el número de individuos es muy grande. Debido a esto, se recurre al uso de gráficas para hacer más sencilla la presentación de la información. Dichas gráficas se generan por medio de una tabla de frecuencia y la tabla 7.5 es un típico ejemplo; para construirla, se establece un número arbitrario de rangos de valores a los que se les denomina clases, registrandose el número de individuos u objetos (frecuencia) que

ANALISIS DE 182 MUESTRAS DE AGUA

Concentración de sales % peso.	Número de muestras	Porcentaje del total
3 - 7	3	1.648
7 - 11	10	5.495
11 - 15	19	10.440
15 - 19	29	15.934
19 - 23	36	19.780
23 - 27	34	18.687
27 - 31	25	13.736
31 - 35	16	8.791
35 - 39	8	4.396
39 - 43	2	1.099
SUMA	182	100.000

Tabla 7.5

caen dentro de un intervalo específico; frecuentemente, es usada - la frecuencia relativa de la clase que se calcula como

$$F_r = \frac{\text{frecuencia de la clase}}{n} \times 100$$

donde n es el número total de individuos de la población.

Con la tabla 7.5, pueden generarse las Figs. 7.8, 7.9 y 7.10.

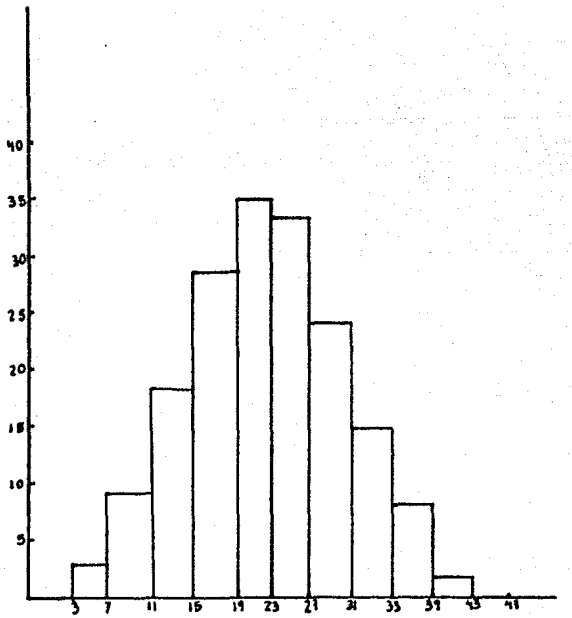


Fig. 7.8-HISTOGRAMA

Si consideramos intervalos muy pequeños, la figura 7.8 se transforma (Fig. 7.11). El modelo matemático surge al - considerar el caso límite*, en donde tanto el número de clases co-

*Apéndice B

mo el de individuos u objetos es infinito, lo cual genera la Fig. 7.12. Esta curva recibe el nombre de normal y puede representarse

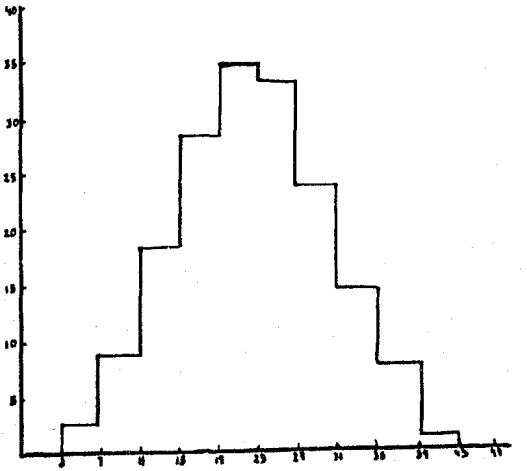


Fig. 7.9-PERFIL

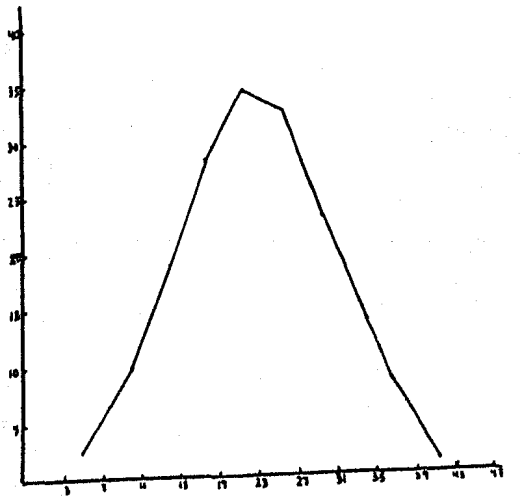


Fig. 7.10-POLIGONO

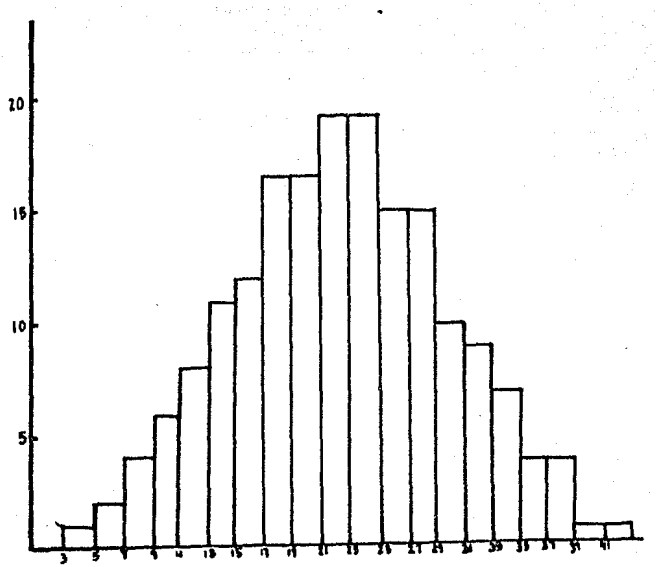


Fig. 7.11

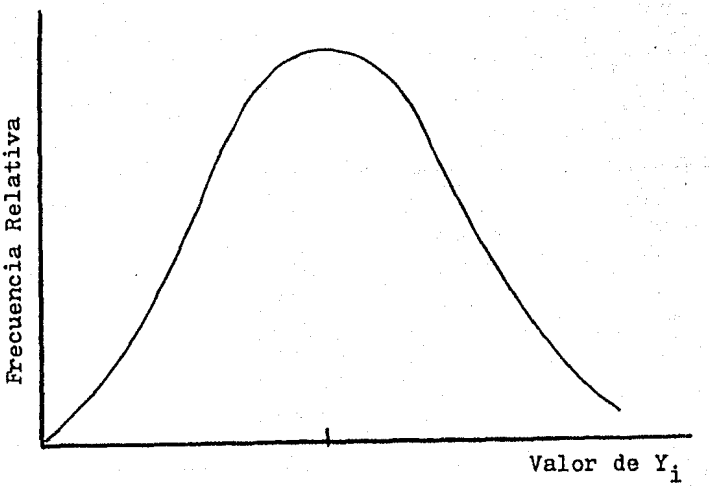


Fig. 7.12

matemáticamente por el modelo:

$$Y = \mu + \varepsilon_i \qquad (7.17)$$

En esta curva, la media aritmética de la población constituye la parte más alta de la curva, y se representa por μ; la amplitud de la curva puede representarse por la varianza poblacional σ^2, la cual rigurosamente es el promedio aritmético de los cuadrados de las discrepancias ε_i entre cada valor de la población y la media poblacional, y que se estima como s^2 en poblaciones finitas así:

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1}$$

Se dice entonces, que Y se distribuye normalmente con media y varianza poblacionales μ y σ^2 $(Y \sim N(\mu, \sigma^2))$, en donde Y son los valores de la medición de interés de los individuos u objetos de la población.

El hecho de que Y tenga una distribución de frecuencias es consecuencia lógica del modelo; μ es un valor característico que depende de los factores que se mantienen constantes, sin embargo, ε_i es una variable <u>aleatoria</u> cuyo valor no puede predecirse en un momento dado; los ε_i se distribuyen normalmente, con media cero y varianza poblacional σ^2, cuando el número de observaciones (n) es muy grande (Fig. 7.13). Entonces Y_i obtiene sus propiedades distribucionales de la variable ε_i, y si se comparan la distribución de los ε_i con la de las Y_i (Fig. 7.14), se puede ver que la curva es la misma, solo que en Y se encuentra desplazada una cantidad μ en el eje horizontal, debido a la forma del modelo

$Y_i = \mu + \xi_i$.

Adicionalmente, los ξ_i son influenciados por los

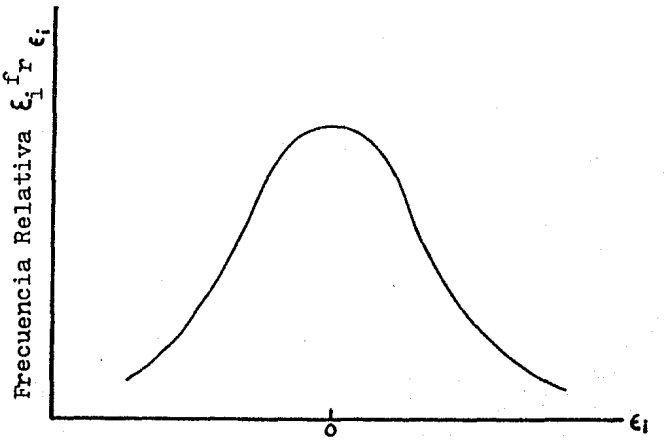


Fig. 7.13

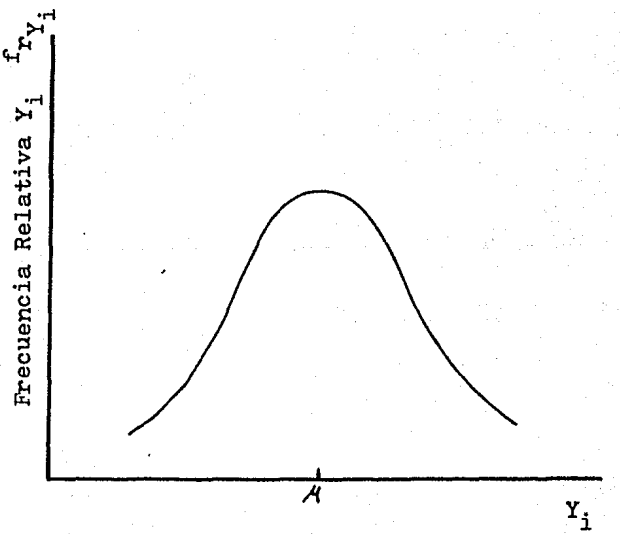


Fig. 7.14

factores no controlados, por lo tanto, poblaciones con igual grado de generalidad deberan tener varianzas iguales o casi iguales; este, precisamente, es el concepto básico de los modelos lineales, - ya que considera el estudio de varias poblaciones con un mismo - grado de generalidad, en las que el modelo de las distribuciones - de frecuencia es normal, con varianza constante e independencia de errores; la Ec. 7.17 se generaliza a:

$$Y_{ijk....l} = \mu(x_j, x_k, ..., x_l) + \epsilon_i \qquad (7.18)$$

en donde $\mu(x_j, x_k, ..., x_l)$ representa la media de una población definida por los factores especificos $x_j, x_k, ..., x_l$. Si se reconoce que la relación es lineal, entonces:

$$(x_j, x_k, ..., x_l) = \sum_{w=1}^{p} \beta_w g_w(x_j, x_k, ..., x_l)$$

donde $g_w(x_j, x_k, ..., x_l)$ son funciones conocidas de las condiciones especificadas, $x_j, x_k, ..., x_l$. Las β_w's son parámetros desconocidos sobre los cuales se enfocara el análisis para su estimación. Si se define $z_w = g_w(x_j, x_k, ..., x_l)$, y efectuando la sumatoria desde - $w = 0$, el modelo queda:

$$Y_i = \beta_o z_o + \beta_1 z_1 + ... + \beta_p z_p + \epsilon_i \qquad (7.19)$$

$$z_o = 1$$

$$Y_i = f(z_1, ..., z_p)$$

Con la Ec. (7.19) se define a todos los modelos - lineales. Si z_w $(w = 1, ..., p)$ actua solo como variable indicadora de la presencia o ausencia de efectos de los factores, se obtienen los llamados modelos de diseños experimentales. En este caso z_w es uno o cero. En cambio si z_w son valores irrestrictos dentro de -

ciertos intervalos, se tienen los modelos de regresión. Si se presentan ambos tipos de comportamiento, los modelos son llamados de covarianza.

β_0 es un parámetro adicional que surge en modelos de regresión al intentar representar a μ, y en modelos de diseño - experimental es μ.

Ahora bien, situando en este contexto al ejemplo número de la sección anterior, se tiene que:

1.- El modelo propuesto $b_o + b_1x$ es un modelo que tratará de representar a μ.

2.- Los coeficientes b_o y b_1 se estiman en - base a la minimización de los cuadrados de - las discrepancias $\varepsilon_i = Y_i - Y_i$.

3.- El método de mínimos cuadrados cumple con la condición de que $\mu_{\varepsilon_i} = 0$.

4.- El modelo $b_o + b_1x$ se traduce en que esa población solo considera una variable cons- - tante (x) por lo cual su grado de generalidad es muy alto, y se debe esperar que su variabilidad también lo sea.

5.- Claramente $b_o + b_1x$ es un caso específico del modelo $Y_i = \beta_o z_o + \beta_1 z_1 + \ldots + \beta_p z_p$ con:

$$z_1 = x \quad y \quad z_2, \ldots, z_p = 0$$

donde $z_2, \ldots, z_p$ implican relaciones funcionales de los factores involucrados $x_j, x_k, \ldots, x_1$

y que se considera no afectan a la media, lo - cual puede ser cierto o no.

Asimismo es necesario que ε_i sea una variable - aleatoria, distribuida normalmente, con varianza σ^2 y ε_i independientes uno de otro.

Supóngase que se tienen diferentes poblaciones en estudio que se definen por el valor de una variable cuantitativa - llamada x. Tres casos particulares de la Ec. 7.19 se daran al considerar que la media de las poblaciones depende de x en las si- - guientes formas:

$$\mu(x) = \beta_0$$

$$\mu(x) = \beta_0 + \beta_1 x$$

$$\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

de manera que los modelos son:

$$Y_i = \beta_0 + \varepsilon_i \qquad (7.20)$$

$$Y_i = \beta_0 + \beta_1 x + \varepsilon_i \qquad (7.21)$$

$$Y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon_i \qquad (7.22)$$

en donde $\varepsilon_i \sim N(0, \sigma^2)$.

Consideremos el modelo 7.22. En este modelo se - considera que, para cada valor de x, se define una población de - valores de Y, con distribución normal, media $\beta_0 + \beta_1 x + \beta_2 x^2$ y varianza constante σ^2. Esto se representa esquematicamente en la -

Fig. 7.15.

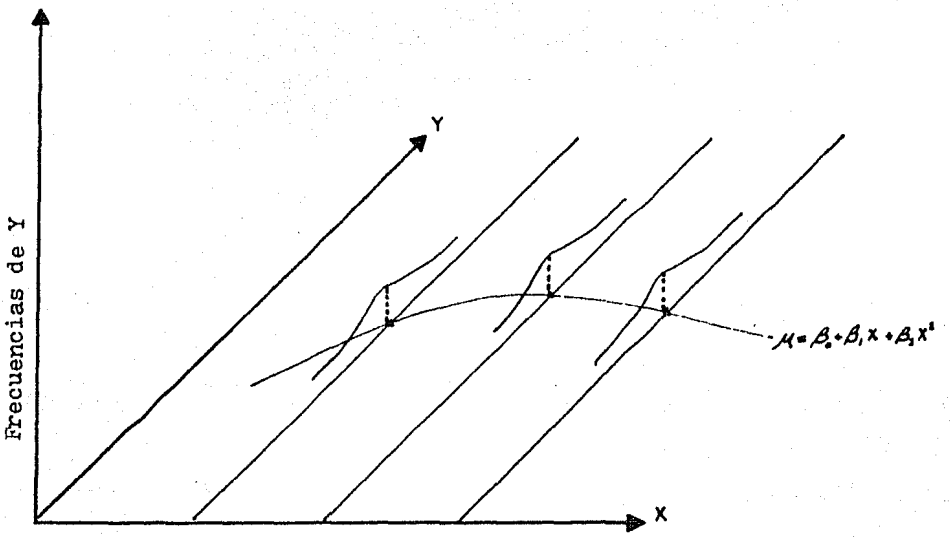


Fig 7.15

La curva $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ esta sobre el plano X-Y, y este plano es el que comunmente se grafica:

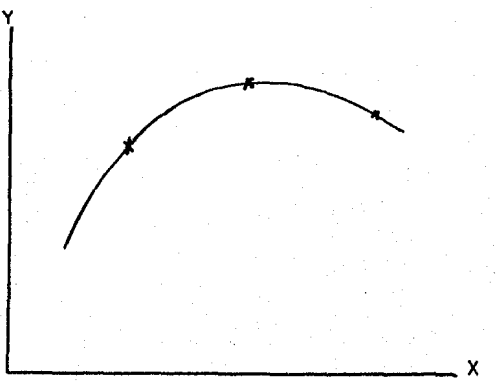


Fig. 7.16

Si suponemos que el modelo anterior es el verdadero, incurriríamos en una "falta de ajuste" si se intentara usar el modelo 7.21:

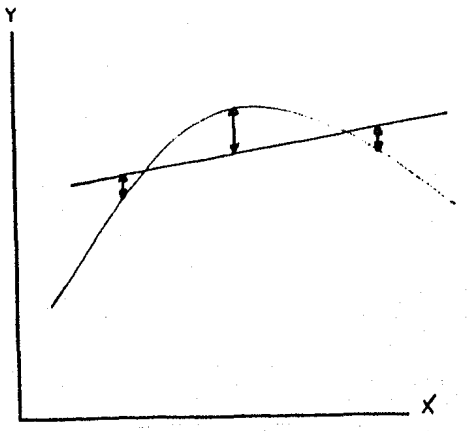


Fig. 7.17

el modelo intenta representar las medias poblacionales en forma lineal con respecto a x:

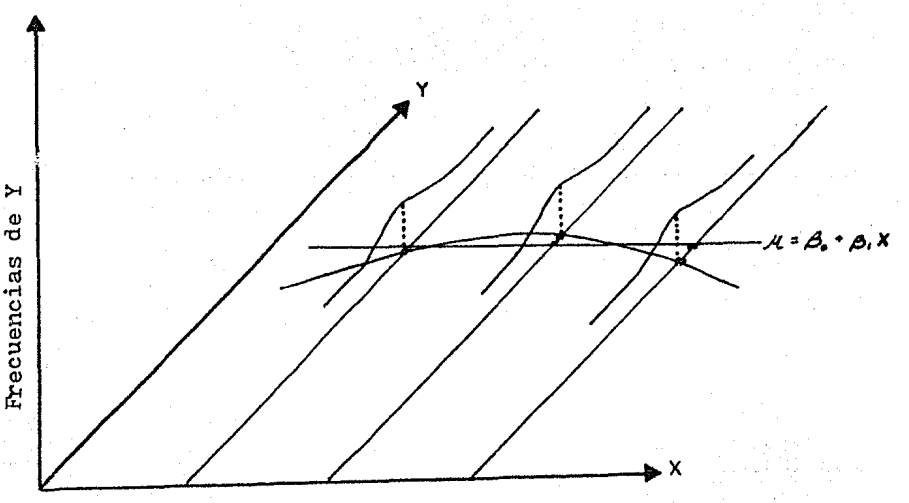


Fig. 7.18

Lo mismo pasaría, pero en mayor proporción, si se intentara usar el modelo 7.20:

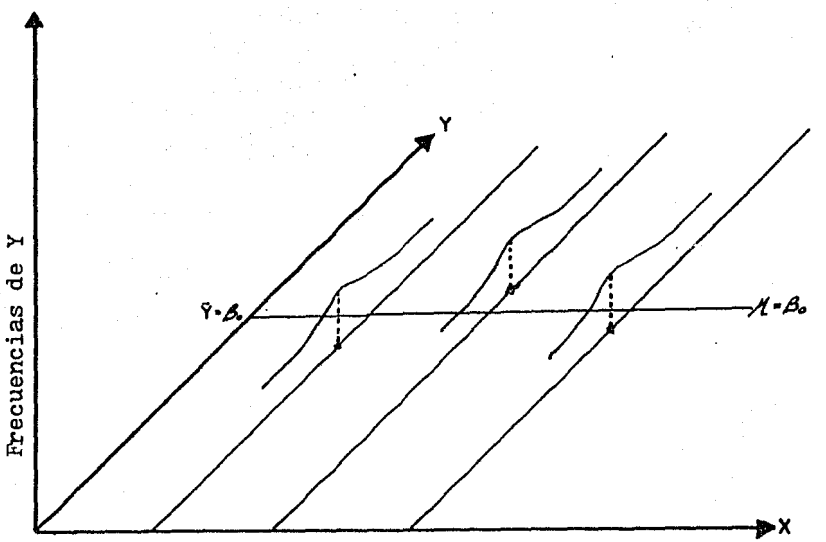


Fig. 7.19

De ahí, surge la necesidad de efectuar un análisis de Varianza, como el realizado en la sección anterior. Con el Análisis de Varianza, nosotros podemos verificar la bondad del modelo que se usa para ajustar las medias poblacionales. Debe notarse, que para cada modelo propuesto, la varianza de las poblaciones de Y para cada x es constante, es decir, poseen el mismo grado de generalidad.

Si z_w, en cambio, se considera como variable que denota ausencia o presencia de categorías cualitativas de efectos, se trabaja con modelos de diseños experimentales, los cuales po-

seen 2 o más grados de generalidad, ya sea que se ignoren o no - ciertas categorías cualitativas.

El ignorar o no ciertas categorías se hace deliberadamente, con el objeto de comparar las medias poblacionales; - imaginemos que se tienen 3 marcas de catalizadores para una reacción específica; obviamente, los catalizadores no pueden representarse en primera instancia como factores cuantitativos. Los catalizadores son la misma sustancia química y se diferencian en el - sentido de que cada uno lo produce una compañía diferente.

Si suponemos que los catalizadores son usados para efectuar la reacción, manteniendo constantes las demás condiciones, tales como temperatura, presión, cantidad de reactivos, tiempo de reacción, etc., se podría considerar que cada catalizador j tiene una población normal de rendimiento (Y_i) con varianza constante entre poblaciones, pero diferentes medias (μ_j); el modelo - sería entonces:

$$Y_{ij} = \mu_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

el término Y_{ij} denota la observación i-ésima de la población j, - donde j = 1,2,3 representa las marcas de catalizador.

Si se ignora el cambio de catalizador, entonces - la población resultante tendrá un mayor grado de generalidad, debido a que el catalizador pasa a ser un factor no controlado; la - varianza de la nueva población será mayor, y la media, llamada media general μ, tendrá un valor específico (Fig. 7.20). Esta pro- - piedad guía a la estimación de efectos especiales en cada una de - las poblaciones estudiadas. Este efecto es la discrepancia entre -

μ y μ_j; a este se le llama efecto del factor marca y se denota - por τ_j, así:

$$\tau_j = \mu_j - \mu \qquad j = 1,2,3$$

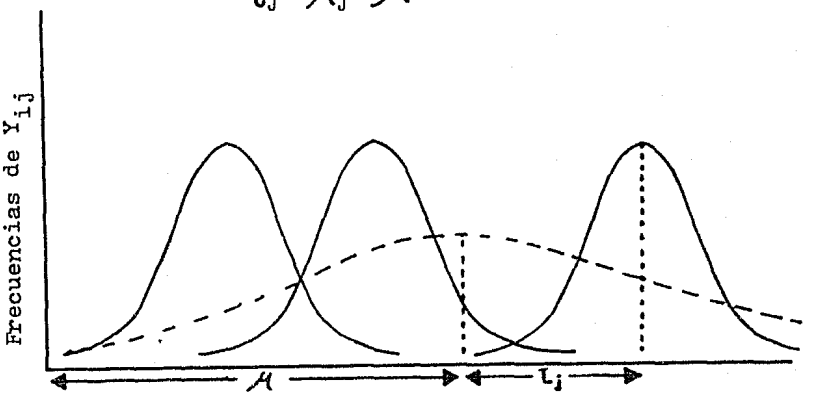


Fig. 7.20

Los valores de τ_j serán positivos si la marca en cuestión aumenta el rendimiento promedio en relación a la media de rendimiento general μ, y negativos si el rendimiento μ_j es menor a μ.

Las 3 poblaciones, entonces, pueden representarse como:

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0,\sigma^2)$$

donde:

Y_{ij} = rendimiento de la medición i-ésima con el catalizador j-ésimo.

τ_j = efecto de la población (rendimiento) - j-ésima.

ε'_{ij} = error aleatorio, producido fundamentalmente por las particularidades específicas de la i-ésima medición en el rendimiento producido por el catalizador j-ésimo, que se genera

por los factores no considerados constantes - al definir la población.

El modelo anterior se puede generalizar con la - Ec. 7.19 como:

$$Y_{ij} = z_o\mu + z_{1j}\tau_1 + z_{2j}\tau_2 + z_{3j}\tau_3 + \varepsilon_{ij} \qquad (7.23)$$

con:

$$z_o = 1 \quad y \quad z_{kj} = \begin{cases} 1 \text{ si } k = j \\ 0 \text{ si } k \neq j \end{cases} \quad (k,j = 1,2,3)$$

y donde se enfatiza que z_w es una cantidad que denota ausencia o presencia de efectos.

Los factores cuantitativos pueden eliminarse a - priori dentro de estos modelos, pero también pueden enmarcarse si los valores entran en el modelo solamente indicando presencia o - ausencia del valor o intervalo de valores determinado.

Así en el caso de un factor como la edad (x_j), - se pueden generar poblaciones donde μ_j sea función de x_j, y x_j sea niños = 1, jovenes = 2 y adultos = 3, utilizando variables indicadoras con valores cero y uno para indicar a que población (grupo - de edad) pertenece cada individuo.

Cuando las poblaciones en estudio se definen por las categorías de un solo factor, como ha ocurrido hasta ahora, se dice que poseen un criterio de clasificación.

Existen diseños que involucran 2 o más criterios de clasificación. En este caso, las poblaciones difieren según la categoría o nivel de los factores involucrados. Si se considera un

diseño con 2 factores (A,B), el modelo puede representarse como:

$$Y_{ijk} = \mu_{jk} + \varepsilon_{ijk} \qquad (7.24)$$

donde

Y_{ijk} = medición i-ésima en la población con nivel j de A y nivel k de B.

μ_{jk} = media de la población con nivel j de A y nivel k de B.

ε_{ijk} = error aleatorio causado por las características específicas particulares de la i-ésima medición en la población con niveles j y k de A y B, respectivamente. O sea, es la desviación que hay entre el valor observado Y_{ijk} y la media μ_{jk}: $\varepsilon_{ijk} = Y_{ijk} - \mu_{jk}$.

Este modelo es general y puede desglosarse según haya interacción o no entre los factores. Considerese la tabla 7.6 y la figura 7.21; se dice que no hay interacción entre los factores si el cambio de nivel en un factor produce una variación en las medias μ_{jk}, constante al considerar los niveles del otro factor, o sea:

$$\mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'}$$

y

$$\mu_{jk} - \mu_{jk'} = \mu_{j'k} - \mu_{j'k'}$$

para cualquier valor de j, j', k y k'.

Valores de μ_{jk}

	FACTOR A	
	j = 1	j = 2
k = 1	100	120
k = 2	104	124

FACTOR B

Tabla 7.6

La Ec. 7.24 se puede transformar con ayuda de la media general μ; μ sería la media de la población si se ignoran los niveles de A y B; μ_j sería la media de la población ignorando los niveles de B y μ_k sería la resultante de ignorar los niveles de A solamente. Consecuentemente, existen 3 grados de generalidad: μ el más alto, μ_j y μ_k donde un factor se ignora (sus niveles), y μ_{jk} el más bajo. Esto se representa en la tabla 7.7.

A partir de esto, se define el efecto principal como la diferencia de la media general, con la media resultante de tomar en cuenta un solo factor, esto es:

$\tau_j = \mu_j - \mu$ efecto principal del factor A en su nivel j.

$\beta_k = \mu_k - \mu$ efecto principal del factor B en su nivel k.

y

$$\mu_{jk} = (\tau_j + \beta_k + \mu)$$

Media de la población	Grado de generalidad
μ	1 (más alto)
μ_j μ_k	2
μ_{jk}	3 (más bajo)

Tabla 7.7

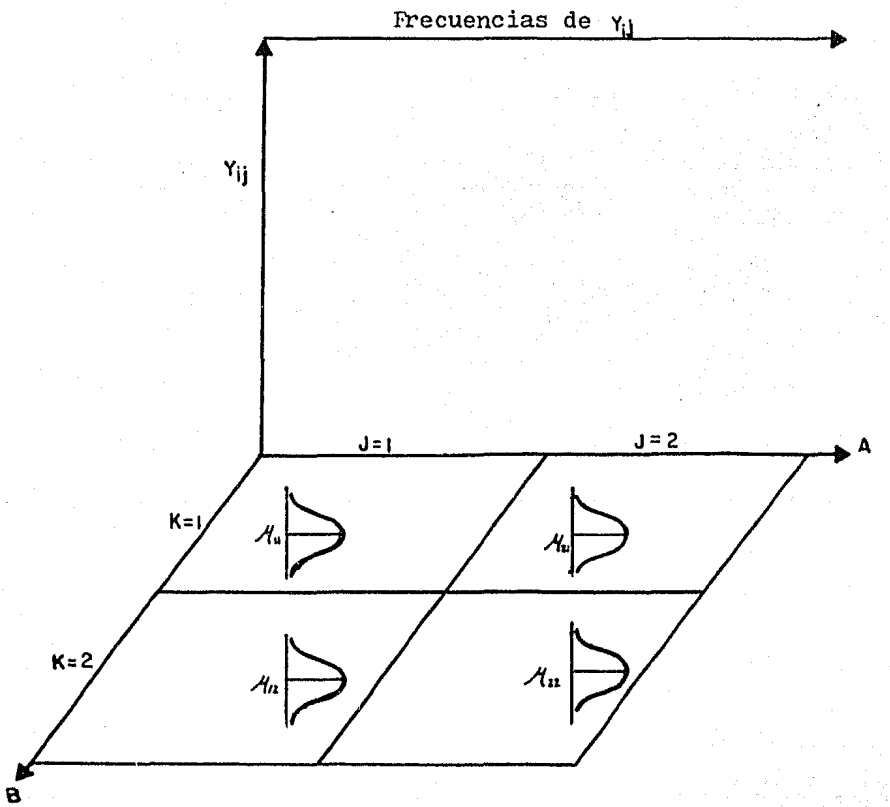


Fig. 7.21

Entonces, la no interacción significa aditividad de los efectos principales.

En caso de que exista interacción, el modelo se complica un poco; para representar μ_{jk} en términos de μ, τ_j y β_k, es necesario introducir un efecto más, el de <u>interacción</u>, y el cual representa la diferencia entre la media real obtenida, μ_{jk}, y la media que se obtendría por la sola adición de los efectos principales:

$$\delta_{jk} = \mu_{jk} - (\tau_j + \beta_k + \mu)$$

por lo tanto

$$\mu_{jk} = \mu + \tau_j + \beta_k + \delta_{jk}$$

y el modelo del diseño con 2 criterios de clasificación con interacción es:

$$Y_{ijk} = \mu + \tau_j + \beta_k + \delta_{jk} + \varepsilon_{ijk} \qquad (7.25)$$

donde

δ_{jk} = efecto de interacción del nivel j de A y el nivel k de B.

Tanto τ_j, β_k y δ_{jk} son constantes desconocidas, pero fijas; existen casos, en los que las constantes no son tales, sino que se comportan como <u>variables</u> aleatorias, dando lugar a los diseños anidados o jerárquicos.

Hay varias condiciones que deben satisfacerse para que los efectos puedan ser considerados aleatorios; en primer -

lugar, los niveles de los factores deberán ser seleccionados al azar de un intervalo permisible de valores; ádemas, deben elegirse al azar las observaciones y los lugares de experimentación, de un número grande de posibilidades. La estimación de efectos se hace prácticamente igual, y la diferencia radica en la utilidad de los resultados; si las conclusiones del experimento se aplican únicamente a los niveles de los factores incluidos en el experimento, estos se consideran de efectos fijos; si en cambio, las conclusiones se quieren extender a muchos niveles del factor, entre los cuales se seleccionaron al azar los niveles estudiados en el experimento, estos se consideran de efectos aleatorios; en este trabajo, solo se consideran factores de efectos fijos.

Cuando existen más de dos criterios de clasificación, se siguen los mismos lineamientos anteriores, aumentando obviamente el número de términos e índices, pero manteniendo en el fondo la forma del modelo 7.25. No obstante, existen procedimientos que cambian la estructura del modelo; estos se usan con el fin de disminuir la magnitud de los errores experimentales, y son muy útiles si los ε_i son muy grandes, pero innecesarios si no es ese el caso. Algunos procedimientos de este tipo se discuten en la sección 7.1.2.4.

7.1.2.3 Regresión Múltiple.- Cuando se trabaja con modelos lineales de regresión, el objetivo primario consiste en obtener los mejores estimadores de las β's desconocidas del modelo propuesto -. Dado que una población consta de un número muy grande de individuos (el cual puede ser infinito), es obvio que no se puede usar con fines prácticos a la población completa para conocer los pará-

metros verdaderos β's. Por lo tanto, se hace necesario tomar una - muestra de la población (aleatoria), con la cual se obtienen números (los estimadores) que den idea de los valores de los parámetros que se desean conocer.

En el caso de los modelos lineales, para poder - estimar los parámetros, se debe especificar el modelo (una recta - en el caso del ejemplo de la sección 7.1.2.1) y se deben cumplir - las suposiciones de homogeneidad de varianzas e independencia de - errores; empero, debe observarse que no es necesario que los errores se distribuyan normalmente.

Después de obtener los parámetros, logicamente se desea saber la validez de ellos. Para facilitar la validación, es deseable que los errores se distribuyan normalmente, para de esa - forma utilizar conjuntamente el Análisis de Varianza y las pruebas de hipótesis.

La hipótesis estadística es una suposición que se hace sobre la forma (tipo de función) de una población, o sobre - los parámetros que caracterizan a una forma (de población) específica. Comunmente, se supone conocida la forma de distribución (en este caso, normal) y entonces se plantean hipótesis sobre los parámetros de esa función de distribución.

Dado que se trabaja con muestras, existe la posibilidad de cometer errores al probar una hipótesis. Esos errores - son de dos tipos:

I.- Rechazar una hipótesis cierta (α).

II.- No rechazar una hipótesis falsa.

DECISION	SITUACION REAL	
	Hipótesis Cierta	Hipótesis Falsa
No rechazar la hipótesis	No error	Error tipo II
Rechazar la hipótesis	Error tipo I	No error

Tabla 7.8

El enfoque clásico considera a α (probabilidad de error tipo I), como fijo (.05 o .01) y busca un procedimiento de prueba que minimice la probabilidad de cometer un error tipo II.

A α, se le conoce como nivel de significancia, y es utilizado también para la formación de límites de confianza.

La prueba de hipótesis se desarrolla de la siguiente manera:

1.- Tomar una muestra de la población sobre la cual se quiere probar la hipótesis.
2.- Calcular un estadístico con la muestra.
3.- Con el empleo de la función de distribución de probabilidades de ese estadístico (funciones derivadas del muestreo), evaluar la probabilidad de tener un estadístico como el obtenido, suponiendo cierta la hipótesis.
4.- Si la probabilidad de tener un estadístico como el obtenido, siendo cierta la hipótesis,

es una probabilidad baja (comunmente .05 o .01), se tendrá la siguiente alternativa:

a) La hipótesis es falsa
b) La hipótesis es cierta y ha producido un evento improbable.

Siempre se optara por a), quedando b) como una posibilidad de error tipo I; entonces, se dice que el hecho observado es significativo. La distribución t de Student se utiliza comunmente como el estadístico a calcular.

El nivel de α es arbitrario, y puede usarse una probabilidad más baja si fueran muy serias las consecuencias de un rechazo erróneo de la hipótesis (recordar que α es la probabilidad de rechazar una hipótesis cierta). Sin embargo, al reducir este valor de probabilidad, automaticamente disminuye la posibilidad de rechazar una hipótesis que es falsa.

Estas pruebas llevan consigo la construcción de límites de confianza; estos demarcan un intervalo de valores, dentro del cual se espera que el valor verdadero del parámetro estimado (en este caso las β's) se encuentre, con una probabilidad $(1 - \alpha)$ de que así ciertamente suceda.

Los límites de confianza juegan un papel aún más importante que las pruebas de significancia. Considere el caso en que se desea saber si un producto A es mejor que uno B; si al construir los límites de confianza al 5% (α), la diferencia B - A esta en el intervalo 240-310, logicamente, aunque haya una probabilidad de 1 en 20 de que el valor verdadero no este dentro de

ese intervalo, la diferencia es tan grande con respecto a cero, - que sin duda el producto B es mejor que el A; este tipo de información suele ser más útil que el utilizar pruebas de significancia (F o t).

Continuando en este punto el ejemplo de la sección 7.1.2.1, los intervalos de confianza para b_o y b_1 se construyen de la siguiente forma:

$$b_1 \pm \frac{t(n-2,1-\frac{1}{2}\alpha) * s}{\left\{\sum (X_i - \bar{X})^2\right\}^{\frac{1}{2}}} \qquad (7.26)$$

$$b_o \pm t(n-2,1-\tfrac{1}{2}\alpha) * \left(\frac{\sum X_i^2}{n\sum (X_i - \bar{X})^2}\right)^{\frac{1}{2}} * s \qquad (7.27)$$

donde

$t(n-2,1-\frac{1}{2}\alpha)$ es el valor del estadistico t - al $(1-\frac{1}{2}\alpha)\%$ y con n-2 grados de libertad.

s es el estimador muestral de $\sigma = (s^2)^{\frac{1}{2}}$.

y dado que

$$\frac{s}{\left\{\sum (X_i - \bar{X})^2\right\}^{\frac{1}{2}}} = .0105$$

y la ecuación (7.26) queda como:

$$-.0798 \pm (2.069 * .0105)$$

y por lo tanto

$$-.1015 \leq \beta_1 \leq -.0581$$

lo que significa que β_1 (el valor verdadero) se encuentra en un punto ubicado entre -.1015 y -.0581, esto con una confianza de $(1 - \alpha) = 95\%$.

Para efectuar la prueba de hipótesis, se obtiene un valor de t que se contrastará con el de $t(1 - \frac{1}{2}\alpha)$:

$$\left|t_{b_1}\right| = \frac{(b_1 - \beta_{10})\left\{\sum (X_i - \bar{X})^2\right\}^{\frac{1}{2}}}{s} \qquad (7.28)$$

$$\left|t_{b_o}\right| = \frac{(b_o - \beta_{oo}) * s}{\left(\frac{\sum X_i^2}{n\sum (X_i - \bar{X})^2}\right)^{\frac{1}{2}}} \qquad (7.29)$$

donde β_{10} y β_{oo} son los valores especificados que se desean probar. Cuando esos valores son cero, se dice que se prueba una hipótesis nula*. Para β_1 y β_o la prueba de hipotesis nula queda:

$$\left|t_{b_1}\right| = \frac{b_1 - 0}{.0105} = \frac{-.0798}{.1015} = \left|-7.60\right| = 7.60$$

y dado que $7.60 > 2.069$, la hipótesis es rechazada (con una confianza del 95 %), por lo cual se opta por la hipótesis alternativa

---

* Se usa el valor absoluto de t, cuando la prueba de hipótesis es de dos colas. Esta se hace cuando no se sabe si el valor o tratamiento probado es mejor o inferior a β_{10} o a un tratamiento base, respectivamente. Si se sabe que el valor o tratamiento debe ser mejor que su estandar respectivo, se usa una prueba de una cola; cuando esto ocurre, se debe tomar el doble del valor de t para efectuar la prueba.

$\beta_{10} \neq 0$*.

Si el valor de $\left|t_{b_1}\right|$ hubiera sido más pequeño, - que el valor de tablas, la hipótesis no se podría rechazar, lo - cual no implica que se acepte esta; solo no se puede rechazar. Como opción, se debe plantear otra hipótesis, o bien profundizar en la investigación; otra alternativa es verificar si el intervalo de confianza incluye el valor de β_{10} (en este caso 0).

Para b_o, los resultados son:

$$\frac{\sum X_i^2}{n\sum (X_i - \bar{X})^2} = .4267941$$

$$13.6230 \pm 2.069 * \left(.4267941\right)^{\frac{1}{2}} * .7923$$

$$12.5520 \leq \beta_0 \leq 14.6939$$

$$\left|t_{b_o}\right| = \frac{13.6230 - 0}{(.4267941)^{\frac{1}{2}}} * .7923 = 16.5216$$

$$16.5216 > 2.069$$

por lo tanto, la hipótesis es rechazada.

Asimismo, es posible construir intervalos de con-

---

* El hecho de que $\beta_{10} \neq 0$ significa que la variable asociada al coeficiente b_1 tiene influencia sobre el valor de la respuesta. Observando los límites de confianza (o el signo del estimador) se puede saber si esa influencia va en deterioro o mejoramiento de la res- - puesta.

fianza para la respuesta $\hat{Y}_i$. Para eso, se usa la fórmula:

$$\hat{Y}_k \pm t(n-2, 1-\tfrac{\alpha}{2}) * e.s.e.(\hat{Y}_k)$$

donde

e.s.e.$(\hat{Y}_k)$ = error estandard estimado de $\hat{Y}_k$ (estimated standard error) y se calcula como:

$$e.s.e.(\hat{Y}_k) = s * \left(\frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)^{\frac{1}{2}}$$

entonces, e.s.e.$(\hat{Y}_k)$ es función del valor actual de X_k, donde:

X_k = valor específico de X usado para predecir $\hat{Y}_k$.

$\hat{Y}_k$ = valor <u>medio</u> predicho de Y a X_k

de esta forma, los intervalos de confianza varían según el valor de X_k; la Fig. 7.22 ilustra este efecto, para el ejemplo con nivel de $\alpha = 0.05$.

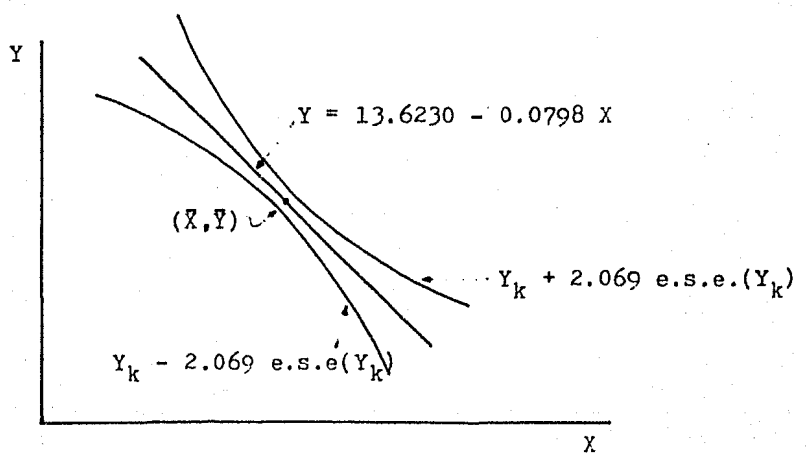


Fig. 7.22

Entonces, el valor verdadero de la media poblacional μ, para una x_k dada, se encuentra en algún punto del intervalo correspondiente; esto con una probabilidad de 1 en 20 de ser errónea la afirmación.

Como se ha visto, las pruebas de hipótesis sirven para validar los parámetros del modelo (uno por uno), mientras que el Análisis de Varianza ayuda a validar la ecuación de regresión (o sea, todos los parámetros a la vez). En el caso particular de una recta, la prueba F para regresión es la misma que la prueba t para $\beta_1 = 0$. El estadístico F sigue una distribución χ^2 y posee los mismos grados de libertad que MS_R y s^2, y el objeto es probar la hipótesis $\beta_1 = 0, \beta_2 = 0, \ldots\ldots, \beta_n = 0$. El valor de F obtenido, deberá ser mayor que el valor de F de tablas. Para el ejemplo, $F_{real} = 57.52$ y $F_{tablas} = (1, n-2, 0.95) = 4.28$, y así $F_{real} > F_{tab.}$ por lo que se rechaza la hipótesis $\beta_1 = 0, \beta_2 = 0, \ldots\ldots, \beta_n = 0$ y se conoce que la ecuación de regresión afecta a la respuesta.

Hasta este punto, se ha desarrollado un método completo para analizar un modelo de la forma $Y = \beta_0 + \beta_1 X$; es deseable, sin embargo, desarrollar un método general, el cual proporcione tanto los parámetros como el subsecuente análisis, de cualquier tipo de modelo lineal. Para hacer esto, es necesario utilizar el álgebra de matrices.

Se define el vector $\underline{Y}$, como el vector de observaciones Y_i*, $\underline{X}$ como la matriz de variables independientes, $\underline{\beta}$ como

---

* Cualquier matriz con una columna es llamada un vector columna; cualquier matriz con una hilera es llamada un vector hilera. Una matriz de 1 * 1 es justamente un número ordinario o escalar.

el vector de parámetros a estimar, y $\underline{\varepsilon}$ como el vector de errores. Para el ejemplo de la sección 7.1.2.1 esto queda:

$$\underline{Y} = \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ 10.36 \\ 11.08 \end{bmatrix} \quad \underline{X} = \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ 1 & 30.8 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 33.4 \\ 1 & 28.6 \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix}$$

Como es conocido, para que 2 matrices puedan multiplicarse, ambas deben ser conformables*; así el producto $\underline{X}\,\underline{\beta}$ del ejemplo sería:

$$\underline{X\beta} = \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 28.6 \end{bmatrix} * \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 & 35.3\,\beta_1 \\ \beta_0 & 29.7\,\beta_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \beta_0 & 28.6\,\beta_1 \end{bmatrix}$$

$$25 * 2 \qquad 2 * 1 \quad = \quad 25 * 1$$

---

* Si $\underline{A}$ es una matriz de n * p, donde n = numero de renglones, y p = número de columnas, esta puede: a) postmultiplicarse por una matriz de p * q resultando una matriz de n * q; b) premultiplicarse por una matriz m * n para dar una matriz de m * p.

La suma de dos matrices se obtiene sumando los elementos correspondientes de ambas matrices:

$$\underline{X}\underline{\beta} + \underline{\varepsilon} = \begin{bmatrix} \beta_0 + 35.3\beta_1 \\ \beta_0 + 29.7\beta_1 \\ \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \\ \beta_0 + 28.6\beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{25} \end{bmatrix} = \begin{bmatrix} \beta_0 + 35.3\beta_1 + \varepsilon_1 \\ \beta_0 + 29.7\beta_1 + \varepsilon_2 \\ \cdot \quad\quad \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \quad\quad \cdot \\ \cdot \quad\quad \cdot \quad\quad \cdot \\ \beta_0 + 28.6\beta_1 + \varepsilon_{25} \end{bmatrix}$$

como se ve, las matrices necesariamente deberan tener las mismas dimensiones. Si 2 matrices o vectores son iguales, sus elementos correspondientes también son iguales. Entonces, una ecuación matricial válida es:

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} \qquad (7.30)$$

que implica que:

$$\begin{aligned} 10.98 &= \beta_0 + 35.3\beta_1 + \varepsilon_1 \\ &\cdot \quad\quad \cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \quad\quad \cdot \\ &\cdot \quad\quad \cdot \quad\quad \cdot \\ 11.08 &= \beta_0 + 28.6\beta_1 + \varepsilon_{25} \end{aligned}$$

o bien

$$Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i \qquad (i = 1,2,3,\ldots\ldots,25) \qquad (7.31)$$

para cada una de las 25 observaciones. Entonces la ecuación matricial (7.30) y la ecuación (7.31) expresan el mismo modelo.

La transpuesta de una matriz se define como aquella matriz en la cual las hileras son las columnas de la matriz - original, y las columnas son las hileras; esto se escribe como $\underline{A}^t$, donde $\underline{A}$ es precisamente la matriz original.

Asi, por ejemplo:

$$\underline{\varepsilon}^t = \left[\varepsilon_1 \quad \varepsilon_2 \quad \varepsilon_3 \ldots\ldots \varepsilon_{24} \quad \varepsilon_{25}\right]$$

$$\underline{Y}^t = \left[10.98 \quad 11.13 \quad 12.51 \ldots\ldots 10.36 \quad 11.08\right]$$

nótese que:

$$\varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \ldots\ldots + \varepsilon_{n-1}^2 + \varepsilon_n^2 = \underline{\varepsilon}^t\underline{\varepsilon}$$

y

$$Y_1^2 + Y_2^2 + Y_3^2 + \ldots\ldots + Y_{n-1}^2 + Y_n^2 = \underline{Y}^t\underline{Y}$$

adicionalmente

$$\underline{X}^t\underline{X} = \begin{bmatrix} 1 & 1 & \ldots. & 1 \\ 35.3 & 29.7 & \ldots. & 28.6 \end{bmatrix} \begin{bmatrix} 1 & 35.3 \\ 1 & 29.7 \\ . & . \\ . & . \\ 1 & 28.6 \end{bmatrix}$$

y esto corresponde a:

$$\underline{X}^t\underline{X} = \begin{bmatrix} 25 & 1315.00 \\ 1315 & 76323.42 \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}$$

el producto X^tY daría:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ 35.3 & 29.7 & \dots & 28.6 \end{bmatrix} \begin{bmatrix} 10.98 \\ 11.13 \\ . \\ 11.08 \end{bmatrix} = \begin{bmatrix} 235.60 \\ 11821.432 \end{bmatrix}$$

$$\underline{X}^t\underline{Y} = \begin{bmatrix} \sum X_i \\ \sum X_i Y_i \end{bmatrix}$$

entonces, las ecuaciones normales pueden ser escritas como:

$$\underline{X}^t\underline{X}\underline{b} = \underline{X}^t\underline{Y}$$

donde $\underline{b}^t = [b_o, b_1]$, los estimadores por mínimos cuadrados de $[\beta_o, \beta_1]$. Para solucionar dichas ecuaciones, se hace uso de la inversa de una matriz, escrita $\underline{A}^-$, donde $\underline{A}$ es la matriz original, que debe ser cuadrada, y con su determinante diferente de cero.

La multiplicación $\underline{A}\underline{A}^- = \underline{A}^-\underline{A}$ siempre dara por resultado la matriz identidad $\underline{I}$, la cual contiene 1's en su diagonal principal, y 0's en sus demás elementos:

$$= \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

De esta forma, la solución de las ecuaciones normales puede escribirse como:

$$(\underline{X}^t\underline{X})^-(\underline{X}^t\underline{X})\underline{b} = (\underline{X}^t\underline{X})^-\underline{X}^t\underline{Y}$$

$$\underline{I}\underline{b} = (\underline{X}^t\underline{X})^-\underline{X}^t\underline{Y}$$

y dado que $\underline{I}\underline{A} = \underline{A}$ para cualquier $\underline{A}$:

$$\underline{b} = (\underline{X}^t\underline{X})^-\underline{X}^t\underline{Y}$$

donde:

$$(\underline{X}^t\underline{X})^- = \begin{bmatrix} \dfrac{\sum X_i^2}{n\sum(X_i - \bar{X})^2} & \dfrac{-\bar{X}}{\sum(X_i - \bar{X})^2} \\ \dfrac{-\bar{X}}{\sum(X_i - \bar{X})^2} & \dfrac{1}{\sum(X_i - \bar{X})^2} \end{bmatrix}$$

$$= \frac{1}{n\sum(X_i - \bar{X})^2} \begin{bmatrix} \sum X_i & -\sum X_i \\ -\sum X_i & n \end{bmatrix}$$

Usando los datos del ejemplo anterior, se halla - que:

$$(\underline{X}^t\underline{X})^- = \begin{bmatrix} .4267941 & -0.0073535 \\ -.0073535 & 0.0001398 \end{bmatrix}$$

y

$$\underline{b} = \begin{bmatrix} .4267941 & -0.0073535 \\ -.0073535 & 0.0001398 \end{bmatrix} \begin{bmatrix} 235.60 \\ 11821.432 \end{bmatrix}$$

$$\underline{b} = \begin{bmatrix} 13.623005 \\ -0.079829 \end{bmatrix}$$

Entonces, con este método, es posible ajustar cualquier modelo lineal en los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ por mínimos cuadrados.

Igualmente, todos los cálculos del Análisis Estadístico como son el Análisis de Varianza y las pruebas de hipótesis, se desarrollan en forma matricial. La tabla de Análisis de Varianza queda como:

FUENTE	SUMA DE CUADRADOS	GRADOS DE LIBERTAD	MEDIA
Regresión	$\underline{b}^t\underline{X}^t\underline{Y}$	p	MS_R
Residual	$\underline{Y}^t\underline{Y} - \underline{b}^t\underline{X}^t\underline{Y}$	n - p	$= s^2$
Total (no corregida)	$\underline{Y}^t\underline{Y}$	n	

En un trabajo de regresión, las preguntas más importantes se hacen con respecto a si incluir ciertos términos en el modelo es benéfico. Esta pregunta puede ser resuelta considerando la porción extra de la suma de cuadrados de la regresión que surge debido al hecho de que el término bajo consideración estuvo

dentro del modelo. La media cuadrada derivada de esta suma extra - de cuadrados puede ser comparada con el estimado, s^2, de σ^2, para ver si esta es significativamente grande. Si es así, el término - deberá ser incluido, si no, el término podrá ser juzgado innecesario y puede ser removido. Un ejemplo de este caso, se vio al formar la tabla 7.3 de Análisis de Varianza, donde $SS(b_1|b_o)$ repre- - sentó la suma extra de cuadrados debida a la inclusión del término $\beta_1 X$ en el modelo. El procedimiento general es el siguiente: supóngase que las funciones Z_1, Z_2,....., Z_p son funciones conocidas de las variables básicas X_1, X_2,..., X_n y que los valores de las X's y sus correspondientes Y's son disponibles. Considerando los dos - modelos siguientes:

$$1.- \quad Y = \beta_o + \beta_1 Z_1 + \beta_2 Z_2 + \ldots\ldots + \beta_p Z_p + \varepsilon$$

de este primer modelo los estimadores por mínimos cuadrados serían $b_o(1)$, $b_1(1)$, $b_2(1)$,....., $b_p(1)$ y $SS(b_o(1), b_1(1), b_2(1),\ldots\ldots, b_p(1)) = S_1$ sería la correspondiente suma de cuadrados. Se supone que el modelo no sufre falta de ajuste*. Entonces, el estimado de σ^2 será s^2, obtenido de los residuos del modelo (1).

$$2.- \quad Y = \beta_o + \beta_1 Z_1 + \beta_2 Z_2 + \ldots\ldots + \beta_q Z_q + \varepsilon \qquad (q < p)$$

Las Z's en este modelo son las mismas que en el - modelo (1) cuando los subíndices son los mismos; hay sin embargo, menos términos en este segundo modelo.

Obteniendo los estimadores por mínimos cuadrados

---

* Ver sección 9

de este modelo: $b_0(2)$, $b_1(2)$, $b_2(2)$,....., $b_q(2)$ la suma de cuadrados sería igual a $SS(b_0(2), b_1(2), b_2(2),....., b_q(2)) = S_2$. - Entonces $S_1 - S_2$ es la suma extra de cuadrados debida a la inclusión de los términos $\beta_{q+1}x_{q+1} + ...+\beta_p x_p$ en el modelo (1). - Dado que S_1 tiene $(p + 1)$ grados de libertad y S_2 tiene $(q + 1)$ grados de libertad, $S_1 - S_2$ tiene $(p - q)$ grados de libertad. Si $\beta_{q+1} = \beta_{q+2} = ... = \beta_p = 0$, entonces $E\left[(S_1 - S_2)/(p - q)\right]^{*} = \sigma^2$. - Si adicionalmente, los errores son normalmente distribuidos, - $(S_1 - S_2)/(p - q)$ tendrá un tipo de distribución χ^2, <u>independiente</u> de s^2. Esto significa que se puede comparar $(S_1 - S_2)/(p - q)$ con s^2 por una prueba $F(p-q, v, \alpha)$ (v = número de grados de libertad de s^2), <u>para probar la hipótesis nula</u>: $\beta_{q+1} = \beta_{q+2} = ... = \beta_p = 0$.

$S_1 - S_2$ puede escribirse como $SS(b_{q+1}, ..., b_p, b_0, b_1, ..., b_q)$, que se lee como la suma de cuadrados de $b_{q+1}, ..., b_p$, dados $b_0, b_1, ..., b_q$. Aplicando este principio, es posible obtener, sucesivamente, para cualquier modelo de regresión, $SS(b_0)$, $SS(b_1|b_0)$, $SS(b_2|b_0, b_1)$,....., $SS(b_p|b_0, b_1, ..., b_{p-1})$; - todas estas sumas de cuadrados estan distribuidas independientemente - mentes de s^2, y por lo tanto, también sus medias cuadradas, dado - que cada una tiene un grado de libertad. Estas medias pueden compararse con s^2 por una serie de pruebas F.

Así, y retomando el ejemplo ya desarrollado, se - probaría la bondad de añadir un término más al modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

---

* Ver apéndice B

donde

X_2 = nuevo factor, cuyos 25 correspondientes valores se encuentran en el apéndice C.

Las siguientes matrices pueden ser construidas:

$$\underline{Y} = \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ 8.4 \\ . \\ . \\ . \\ 10.36 \\ 10.08 \end{bmatrix} \quad \underline{X} = \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ 1 & 30.8 & 23 \\ 1 & 58.8 & 20 \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & 33.4 & 20 \\ 1 & 28.6 & 22 \end{bmatrix} \quad \underline{\beta} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

Usando los desarrollos anteriores:

$$\underline{b} = (\underline{X}^t\underline{X})^{-}\underline{X}^t\underline{Y}$$

donde $\underline{b}$ es el vector de estimadores de $\underline{\beta}$, y dado que el determinante de $\underline{X}^t\underline{X} \neq 0$, entonces:

$$\underline{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \left[\begin{bmatrix} 1 & 1 & 1 & ..... & 1 \\ 35.3 & 29.7 & 30.8 & ..... & 28.6 \\ 20 & 20 & 23 & ..... & 22 \end{bmatrix} \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ 1 & 30.8 & 23 \\ . & . & . \\ 1 & 28.6 & 22 \end{bmatrix} \right]^{-} *$$

$$* \begin{bmatrix} 1 & 1 & 1 & \ldots\ldots & 1 \\ 35.3 & 29.7 & 30.8 & \ldots\ldots & 28.6 \\ 20 & 20 & 23 & \ldots\ldots & 22 \end{bmatrix} \begin{bmatrix} 10.98 \\ 11.13 \\ 12.51 \\ . \\ 11.08 \end{bmatrix}$$

esto da:

$$\begin{bmatrix} b_o \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 25.00 & 1315.00 & 506.00 \\ 1315.00 & 76323.42 & 2653.30 \\ 506.00 & 26353.30 & 10460.00 \end{bmatrix} - * \begin{bmatrix} 235.600 \\ 11821.432 \\ 4831.860 \end{bmatrix}$$

$$= \begin{bmatrix} 2.778747 & -0.011242 & -0.106098 \\ -0.011242 & 0.146207*10^{-3} & 0.175467*10^{-3} \\ -0.106098 & -0.175467*10^{-3} & 0.478599*10^{-2} \end{bmatrix} * \begin{bmatrix} 235.600 \\ 11821.432 \\ 4831.860 \end{bmatrix}$$

que finalmente proporciona:

$$\begin{bmatrix} b_o \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 9.1266 \\ -0.0724 \\ 0.2029 \end{bmatrix}$$

Por lo tanto, la ecuación ajustada por mínimos - cuadrados es:

$$Y = 9.1266 - 0.0724X_1 + 0.02029X_2 \qquad (7.32)$$

Las tablas de Análisis de Varianza serían:

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	MEDIA CUADRADA	F
Total (no corregida)	25	2284.1102		
Media (b_0)	1	2220.2944		
Total (corregida)	24	63.8158		
Regresión b_0	2	54.1871	27.0936	61.8999
Residual	22	9.6287	0.4377	

Tabla 7.9

Si $\alpha = 0.05$ $F(2,22,0.95) = 3.44$, por lo tanto, la ecuación es buena predictora, ya que $F_{real} = 61.8999$ es mayor a $F_{tablas} = 3.44$.

La pregunta siguiente sería: ¿Que tanto beneficio ha traído consigo la adición de la 2ª variable?. En otras palabras, ¿La variable es útil para la predicción?. Utilizando el principio de suma extra de cuadrados, se puede construir la siguiente tabla análoga a la 7.9:

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	MEDIA CUADRADA	F
Total (no corregida)	25	2284.1102		
Media (b_0)	1	2220.2944		
Total (corregida)	24	63.8158		
Regresión b_0	2	54.1871	27.0936	61.8999
debido a $b_1 \mid b_0$	1	45.5924	45.5924	104.1636
debido a $b_2 \mid b_1, b_0$	1	8.5947	8.5947	19.6361
Residual	22	9.6287	.4377	

Tabla 7.10

Como podrá observarse, la contribución de X_2 es muy importante, tanto que 19.6361 excede a $F(1,22,0.95) = 4.30$, y por lo tanto su introducción a sido significativa. Esta es la llamada prueba F secuencial, en donde las variables son añadidas una a una en etapas, a la ecuación de regresión.

Si se toma en cuenta el orden de entrada, para la ecuación anterior pueden existir dos casos:

a) $SS(b_2 | b_1, b_o)$

b) $SS(b_1 | b_2, b_o)$

Cada uno de ellos, dan una medida del "valor" de añadir el término b_j correspondiente, estando los otros ya presentes. Su media cuadrada se puede comparar con la prueba F, con s^2, como ya se hizo para $SS(b_2 | b_1, b_o)$. A estas, se les conoce como pruebas F parciales. Son de suma importancia, ya que el efecto de una variable, (digamos X_q) en la determinación de la respuesta, puede ser muy grande cuando la ecuacion incluye solamente a X_q. Sin embargo, cuando la misma variable es introducida en la ecuación después de otras variables, podría afectar muy poco a la respuesta, debido al hecho de que X_q estuviera altamente correlacionada con las variables que ya estan en la ecuación de regresión. Para la variable X_1, esta prueba parcial F esta ya incluida en la prueba secuencial F anterior; para la variable X_2, esta prueba parcial seria:

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	MEDIA CUADRADA	F
Total (no corregida)	25	2284.1102		
Media (b_0)	1	2220.2944		
Total (corregida)	24	63.8158		
Regresion b_0	2	54.1871	27.0936	61.8999
debido a $b_2 \mid b_0$	1	18.3424	41.9063	41.9063
debido a $b_1 \mid b_2, b_0$	1	35.8447	35.8447	81.8933
Residual	22	9.6287	.4377	

Tabla 7.11

Se puede observar, que la contribución de X_2 es - más importante que su misma contribución cuando X_1 había sido ya - introducida. Nótese también que esta es reflejada en el valor de F para X_1, en las 2 tablas ($81.8933 < 104.1636$). Sin embargo, X_1 es la variable más importante en ambos casos, dado que su contribución a la reducción de la suma de cuadrados residual es más grande, no importando el orden de introducción de las variables.

Los términos pueden ser introducidos en cualquier agrupamiento lógico también, por ejemplo, en modelos polinomiales (que no consideran interacciones), se pueden construir sumas ex- - tras de cuadrados alternativas, por ejemplo, $SS(b_0)$, SS(términos - de primer orden b_0), SS(términos de segundo orden términos de primer orden, b_0), etc, y comparar esas SS con s^2. Esto da un gran número de opciones para usar este principio, y la gran mayoría de - los métodos especializados lo utilizan.

Otro estadístico que puede utilizarse para verificar la bondad de la introducción de una variable, es el R^2, y -

adicionalmente, se pueden construir intervalos de confianza. R^2 - para el ejemplo, se obtiene como:

$$R^2 = \frac{54.1871}{63.8158} = .844$$

R^2 para el ejemplo de la sección 7.1.2.1 es:

$$R^2 = \frac{45.5924}{63.8158} = .714$$

entonces, se ve que la ecuación de regresión (7.32) explica mejor la variación de los datos que la ecuación (7.13). Sin embargo, este estadístico debe usarse con cuidado ya que es posible siempre - hacer que $R^2 = 1$*.

Los límites de confianza tienen la misma estructura que los calculados en la sección 7.1.2, y asimismo, se interpretan análogamente. Estos se calculan como:

$$b_i \pm t(v, 1-\tfrac{1}{2}) * e.s.e.\ (b_i)$$

donde

e.s.e. = error estandard estimado de b_i, y se calcula como la raíz cuadrada del término - i-diagonal de la matriz $(\underline{X}^t\underline{X})^{-}s^2$, conocida - como matriz de varianza-covarianza, y cuya - estructura es:

$$\begin{bmatrix} V(b_o) & cov(b_o,b_1) & cov(b_o,b_2) \\ cov(b_1,b_o) & V(b_1) & cov(b_1,b_2) \\ cov(b_2,b_o) & cov(b_1,b_2) & V(b_2) \end{bmatrix}$$

---

* Cuando los grados de libertad son pocos, R^2 tiende a 1. El caso límite ocurre cuando df = 0, entonces, necesariamente R^2 es 1.

los intervalos de confianza separados son frecuentemente utilizados, pero ellos son muy susceptibles de malinterpretarse. Considere la Fig. 7.23, la cual ilustra una posible situación que surge al considerar 2 parámetros. La región de confianza conjunta al 95%, para los verdaderos parámetros, β_1 y β_2, es mostrada como una elipse alargada, la cual encierra los valores (β_1, β_2) que los datos obtenidos registran como conjuntamente razonables para los parámetros. Esta, toma en cuenta la correlación entre los estimados b_1 y b_2.

Los intervalos de confianza para β_1 y β_2 son apropiados para los rangos específicos de los parámetros individuales; sin embargo, si un intento es hecho para interpretar esos

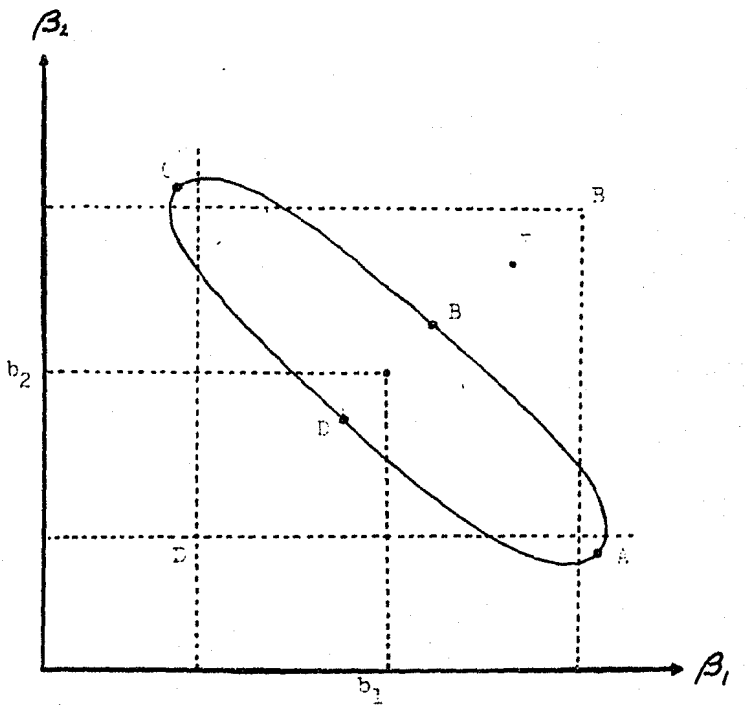


Fig. 7.23

intervalos simultaneamente, este sera erróneo, ya que asi implicitamente se esta definiendo como región conjunta de confianza al rectangulo ABCD, en lugar de la elipse correspondiente. Entonces, un punto E, a primera vista razonable, será completamente erróneo.

La elipse se obtiene de la ecuación matricial:

$$(\underline{\beta} - \underline{b})^t \underline{X}^t \underline{X} (\underline{\beta} - \underline{b}) \leq p s^2 F(p, v, 1-\alpha)$$

recordando que p es el número de variables involucradas. En general, esta ecuación es muy útil solamente cuando p es pequeño (2 o 3), ya que cuando más parámetros estan envueltos, la interpretación es muy dificil, de ahi que los intervalos separados sean muy usados; entonces, se debe tener cuidado en su manejo. Para p = 2, varianza de b_i es diferente a varianza de b_j, y la correlación entre b_i y b_j, calculada como:

$$\rho_{ij} = \frac{cov(b_j, b_i)}{[V(b_i)V(b_j)]^{\frac{1}{2}}}$$

no es pequeña, la situación de la Fig. 7.23 ocurre. Si ρ_{ij} es aproximadamente cero, entonces la región rectangular definida por los intervalos de confianza individuales se aproxima a la región conjunta correcta de confianza*; el alargamiento de la figura dependera de los tamaños relativos de $V(b_i)$ y $V(b_j)$ (Figs. 7.24 y 7.25)

---

* Sera correcta, en términos de si verdaderamente incluye a los parámetros β_1 y β_2 del modelo. Recordar que para $\alpha = 0.05$, existe una posibilidad de uno en 20 de que esa región no sea correcta.

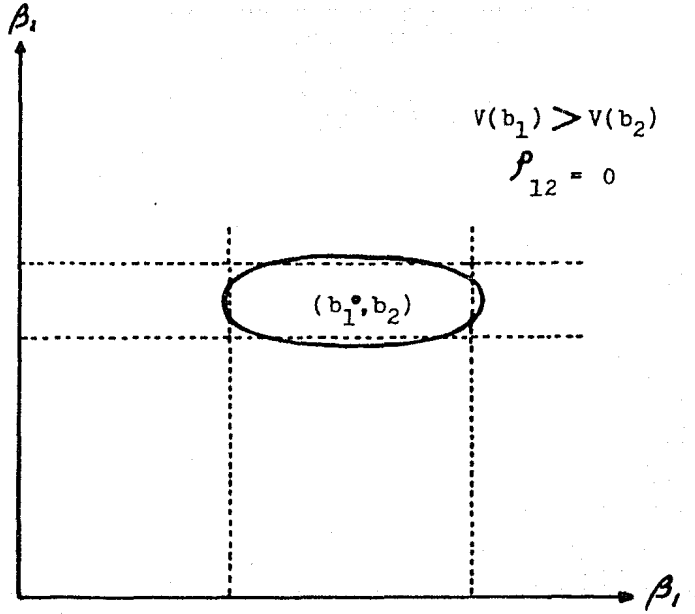


Fig. 7.24

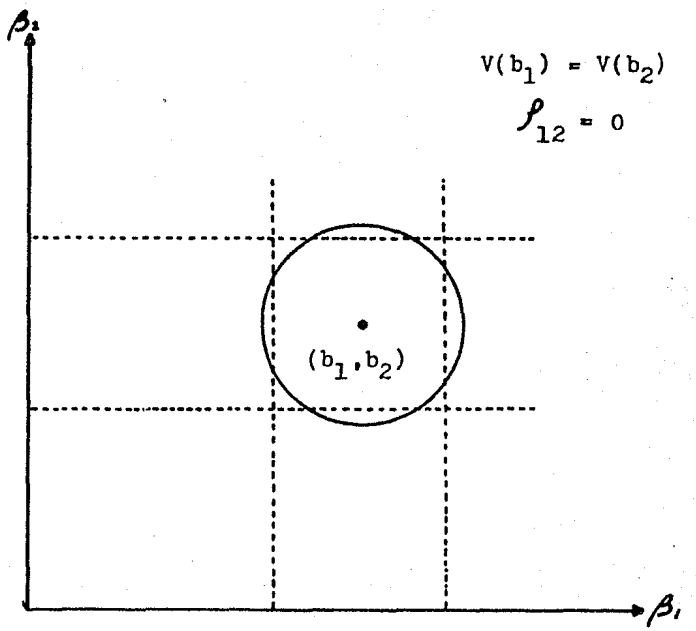


Fig. 7.25

Por último, los límites de confianza para el valor medio de Y a X_k, son obtenidos de:

$$\hat{Y}_k = t(v,1-\tfrac{1}{2}) * s * (\underline{X}_k{}^t(\underline{X}^t\underline{X})^{-}\underline{X}_k)^{\frac{1}{2}}$$

para el caso de la respuesta; esto genera un tipo de curva similar al de la Fig. 7.22. En este caso, no existe riesgo de malainterpretación.

7.1.2.4 Usos de los modelos lineales.- Fundamentalmente, los modelos de regresión tienen 4 tipos de uso: a) Descripción y Explicación, b) Predicción, c) Control y d) Calibración. Los modelos de diseño se emplean para a) Comparación de medias, b) Estudio de efectos y c) Estimación de parámetros poblacionales.

a) Descripción y Explicación.- Los modelos de regresión son valiosos para describir el tipo de asociación entre la variable[13] dependiente Y, y las variables independientes $X_1, X_2, \ldots, X_j, \ldots, X_p$. En este caso lo que se persigue es resumir las tendencias de los datos y encontrar la forma de asociación entre las variables.

Es muy importante señalar que en dicho uso no se pretende establecer relaciones causales, en el sentido de que los valores de las X_j produzcan cambios en los valores de Y. Los modelos de regresión indican únicamente que existe asociación entre las variables y cuál es la forma de dicha asociación; esto se hace de manera empírica, solo con la información de los datos observados.

Puede suceder que existan factores no estudiados que estén causando conjuntamente cambios en los valores de Y y en

los de las X_j. Para citar un ejemplo clásico, señalaremos que en un estudio de poblacion en Estados Unidos se encontró que la frecuencia de cáncer (Y) estaba relacionada con la intensidad del hábito de fumar tabaco (X), a traves de un modelo del tipo $Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$, con $\beta_1 > 0$; sin embargo, los defensores de las compañias de cigarros señalaron, con justa razón, que podria existir un factor génetico, fisiológico o psicológico que produjera ambos efectos, es decir, que aumentara la susceptibilidad al cáncer y al mismo tiempo incrementara el deseo de fumar. Posteriormente, con experimentos más precisos, empleando perros, se encontró una relación causa-efecto que indicaba que el cáncer era causado entre otros factores por el hábito de fumar. Otro caso es el de la relación entre número de cigüeñas en Inglaterra (Y) y la producción de acero (X); esto no indica, necesariamente, que al aumentar la producción de acero se incremente la cantidad de cigüeñas.

No se pretende que los modelos lineales de regresión señalen relaciones causa-efecto a través de una relacional funcional. Sin embargo, si basados en otros conocimientos cientificos ajenos a la estadistica se establece la relación causa-efecto, los modelos de regresión son valiosos auxiliares que permiten simplificar y estudiar dicha relación. Por ejemplo, aquí incluiríamos los modelos llamados "curvas de crecimiento", que son los que ligan la edad o tiempo con la masa o número de individuos producidos al estudiar seres vivos o poblaciones de seres vivos.

Un empleo interesante de los modelos de regresión se da en las relaciones entre factores que puedan pasar a ser hi-

pótesis científicas al suponerse la relación causa-efecto provisionalmente. Así, por ejemplo, si al estudiar a los alumnos de nivel universitario se encuentra que las calificaciones promedio en la universidad (Y) tienen una asociación lineal positiva con el ingreso promedio de los padres (X), se puede plantear como hipótesis que un ingreso bajo produce estudiantes con malas calificaciones y que ingresos altos producen buenas calificaciones. Provisionalmente, y como explicación, se considera que los estudiantes de familias de ingresos bajos tienen que trabajar y dedican menos tiempo al estudio que los del otro tipo de familias; otra alternativa sería que los estudiantes de familias con ingresos bajos tienen una alimentación deficiente, cosa que no sucede con los de ingresos altos. Se requerirá una investigación para probar las hipótesis sugeridas y que la regresión solo ayudó a plantear.

Ejemplos como el anterior son muy útiles en economía, donde el empleo de modelos lineales es muy abundante en el área conocida como econometría; por ejemplo, el estudio del precio de los automóviles Y, como función de ingreso per cápita X_1, $X_2 = X_1^2$, costo de tipo colectivo X_3 e índice de costo de alimentos X_4.

b) Predicción.- Un empleo sumamente importante de los modelos de regresión es la posibilidad de predecir el valor que tendrá Y_i, o la media de la población del conjunto de Y_i que se genera cuando se especifican las condiciones del proceso mediante los valores de X_{ji}.

Un ejemplo lo constituyen las llamadas series de tiempo, donde se trata de predecir cuál será el valor de una ca-

racterística Y que depende del tiempo. Por ejemplo, el precio de los tomates para un año futuro, donde las X_j ($j = 1,\ldots,p$) son funciones del tiempo y/o de observaciones anteriores; o bien, estimar la demanda de petróleo en función del tiempo. Un enfoque diverso en las series de tiempo se logra mediante la teoría de procesos estocásticos.

Otro ejemplo se presenta cuando una fábrica necesita predecir o pronosticar la producción (Y) que tendrá para ciertos costos, concentrados y calidades (las X_j).

La predicción se emplea también para diferentes situaciones; por ejemplo, si se liga el rendimiento de un cultivo en una región (maíz, trigo, caña de azúcar, etc.), que sería la variable Y, con lluvia (X_1), cantidad de nitrógeno (X_2), de fósforo (X_3) aplicado al suelo y características agrónomas tales como incidencia de plagas (X_5), días necesarios para la floración (X_6), etc., se podrá predecir el rendimiento conociendo los valores de las X_j ($j = 1,\ldots,p$) que se tiene en un momento dado.

Para una buena predicción es necesario que la variabilidad de los errores ε_i, sea lo más pequeña posible. Esto no significa que sea conveniente introducir muchas variables independientes en el modelo, ya que pueden aparecer algunas que no son <u>explicativas</u>, o sea que, aunque reducen <u>poco</u> la variabilidad de los errores, pueden incrementar la variabilidad de las predicciones al estimar los parámetros ligados a las variables no explicativas.

También debe tenerse en cuenta que es factible lograr una reducción de la variabilidad del error mediante la in-

troducción de muchas variables independientes, hasta el grado de tener error cero cuando las variables independientes son n - 1, siendo n el numero de observaciones; en este caso, hay un ajuste perfecto; sin embargo, en virtud de que es para las observaciones y_i particulares que se tienen, no se ajustará del mismo modo a otras observaciones. Por dicha razón, este modelo con ajuste perfecto no funciona para la predicción, de manera que no es recomendable medir e incluir como variables independientes todo lo que pueda tener influencia sobre la variable dependiente; se debe concretar en cambio a estudiar aquellas variables independientes que, con base en conocimientos previos, se supone que poseen una influencia preponderante para determinar los valores de las observaciones en la variable dependiente.

Existe en literatura muy especializada varios procedimientos para seleccionar de entre un conjunto de posibles variables independientes, uno de los mejores subconjuntos (stepwise, backward, etc). Los criterios que se utilizan son: el valor de R^2, las pruebas F, y otros.

c) Control.- Un modelo de regresión puede servir para encontrar cuáles son los valores de las x_j ($j = 1,.....,p$) que pueden optimizar, de acuerdo con algún criterio, los valores de la variable dependiente Y. En esta área se hallan las llamadas funciones de producción, las cuales constituyen una extensión de las llamadas superficies de respuesta.

Dos ejemplos típicos de lo que significa el control se presentan a continuación:

1) Se intenta encontrar la ración óptima (la más económica) para la producción de carne en los pavos; para lograrlo se encuentra, por regresión, una función que ligue la carne producida Y y las cantidades de varios alimentos como maíz, soya, etc. (las X_j); posteriormente, en función de los costos y, con ayuda de programación lineal, se obtiene la ración óptima.

2) Se investiga el rendimiento del maiz (Y) bajo diferentes dosis de fertilizantes, por ejemplo, cantidad de nitrógeno (X_1), de fósforo (X_2), de potasio (X_3), de agua (X_4) y funciones de esas variables como X_1^2, X_2^3, X_1X_2, X_3^4, $X_5^{.62}$, etc.; evaluando el costo de los fertilizantes y del agua se puede determinar cuál es la combinación de valores de las X_j, que producen un rendimiento óptimo desde el punto de vista económico.

En estos ejemplos se establece la ecuación de regresión y se buscan los valores de las X_j que controlen los cambios en el sentido deseado.

En la industria se pueden usar los modelos de regresión para controlar, es decir, para optimizar procesos de producción. En la del acero, por ejemplo, es factible relacionar la producción o resistencia del acero Y, con caracteristicasdel proceso, tales como temperatura de fundición X_1, cantidad de hierro X_2, cantidad de carbon X_3, etc. De este modo se pueden determinar los valores de X_j que optimizan Y, en un contexto económico.

En términos generales, si bien se puede relacionar, mediante un modelo de regresión, una caracteristica de interes económico Y, con un conjunto de factores X_j o variables independientes susceptibles de modificarse según el arbitrio humano,

se optimizará la característica Y variando los valores de los factores X_j. A este proceso estadístico se le conoce como control. - Para esto, el modelo de regresión deberá ser satisfactorio desde el punto de vista estadístico, con varianza de errores baja y cumpliendo las suposiciones hechas en el planteamiento del modelo. - Por otra parte, el modelo podrá contemplar algunas variables independientes que no sean controlables: lluvia, temperatura ambiental de la región, concentración de materias primas, etc., pero habrá otras variables independientes factibles de ser controladas.

d) Calibración.- La regresión se emplea también - en el problema de "calibración", en el que la Y es una característica aleatoria* fácil de medir que depende de una variable no - aleatoria difícil de medir X; la relación de dependencia se puede determinar mediante una regresión del tipo

$$Y_i = \sum_{j=1}^{p} \beta_j g_j(X_i) + \varepsilon_i$$

El modelo se usa para predecir el valor de X que ha producido una Y determinada. Por ejemplo, si se desea saber el contenido de calcio X en una sustancia, se pueden medir ciertas - bandas características al analizarla con un espectrofotómetro (Y). Si se preparan algunas muestras con cantidades de calcio conocidas X_i y se determina la Y_i para cada muestra, se ajusta una re- - gresión como la ya mencionada. El modelo obtenido se empleará para encontrar el valor de X, que es el contenido de calcio de una mu-

---

* Ver secciones 7.1.2.2 y 7.1.1

estra nueva desconocida, sólo midiendo las bandas en el espectrofótometro Y.

En general, dicho procedimiento se aplica para la calibración de aparatos científicos y métodos de determinación de características no aleatorias, pero en los que el aparato o el método de medición estan sujetos a fluctuaciones aleatorias que producen variabilidad en los resultados.

a) Comparación de medias.- Los modelos estadísticos lineales de diseño de experimentos tienen como objetivo principal comparar las medias de las poblaciones estudiadas; por ejemplo, al comparar en el trópico la producción de leche de cinco poblaciones formadas considerando igual número de razas de ganado, o bien, tipos de manejo o combinaciones de ambos criterios; lo que interesa es comparar las medias de producción de leche de esas poblaciones.

Otro ejemplo es la comparación de medias de calificaciones de los alumnos universitarios sometidos a métodos diversos de enseñanza, los cuales constituyen las diferentes poblaciones.

En la industria se pueden comparar las medias de producción de varias poblaciones generadas al considerar diferentes tipos de operación de las fabricas.

En medicina es frecuente la comparación de las medias de poblaciones de ciertos caracteres fisiológicos o morfológicos tales como presión arterial, contenido de azúcar en la sangre, tamaño de órganos de hipertrofias, etc., y las poblaciones que se generan al considerar cierto tipo de droga o tipos de trata-

miento médico en general.

En todos los casos las hipotesis planteadas son del tipo:

$$\mu_1 = \mu_2 = \ldots = \mu_t$$

esto es,

$$\mu_j = \mu^* \quad \text{para } j = 1,\ldots,t$$

Esta hipotesis, al incorporarse a los modelos vistos en la sección 3.6, genera una hipotesis del tipo

$$\alpha_1 = \alpha_2 = \ldots\ldots = \alpha_t$$

o

$$\tau_1 = \tau_2 = \ldots\ldots = \tau_t$$

o sea que la hipótesis de igualdad de medias se confirma mediante la hipótesis de igualdad de efectos de poblaciones. No siempre se prueba la llamada "nulidad de efectos", sobre todo con ciertos casos restringidos, aunque autores como Searle consideran que no es practico probar hipótesis de nulidad.

Una hipótesis de nulidad seria:

$$\alpha_1 = \alpha_2 = \ldots\ldots = \alpha_t = 0$$

Si una hipótesis del tipo $\mu_1 = \mu_2 = \ldots\ldots = \mu_t$ es rechazada (mediante el análisis de varianza), se puede deber a que μ_1 es diferente del resto y $\mu_2 = \mu_3 = \ldots\ldots = \mu_t$; a que todas son distintas entre sí; o bien que $\mu_1 = \mu_2 \neq \mu_3 = \mu_4 = \mu_5 \neq \mu_6$, etc., Como se puede ver, existe un número muy grande de posibili-

dades en las cuales no se cumple $\mu_1 = \mu_2 = \ldots = \mu_t$. En estos casos se recurre a cierto tipo de pruebas estadisticas, que permiten probar un conjunto de hipótesis que consisten en la igualdad de parejas de medias de poblaciones, y establecer un orden entre los valores de las μ_j, lo que obviamente permitirá inferir sobre las poblaciones que optimicen los valores de Y. Un ejemplo sería:

$$\mu_1 = \mu_2 \cdot \mu_2 = \mu_3 \cdot \mu_1 \neq \mu_3$$

Obsérvese que en este caso no actúa la ley de transitividad según la cual "dos cosas iguales a una tercera son iguales entre si", debido a que no se trata de una igualdad algebraica sino de una igualdad "estadistica", es decir, con algún grado de probabilidad de ser cierta y otro de ser falsa.

b) Estudio de efectos.- Un diseño experimental puede planearse para estudiar las relaciones entre varios factores cualitativos o cuantitativos que sirven como criterios para definir las poblaciones bajo estudio y sus respectivas medias[10,25].

Una hipótesis de gran importancia práctica es la de "no interacción" entre dos o más factores cuantitativos que sirven como criterios para definir las poblaciones bajo estudio. Por ejemplo, al investigar la depresión en sujetos drogadictos, se podrá conocer si el efecto de la droga se modifica al tratarse de hombres o de mujeres, o bien en un proceso industrial, si el efecto de concentración de reactivos sobre la produccion se modifica o no con el tiempo de proceso. Cuando no es rechazada la hipótesis de "no interacción", los resultados relativos a un factor de clasificación en estudio respecto a un nivel de los otros factores

se pueden generalizar a los otros niveles de los demás factores. - Cuando la hipótesis de no "interacción" es rechazada, los resultados de un factor se ven modificados por los niveles de otro(s) - factor(es). Por ejemplo, si se estudian ciertas variedades de plantas en diferentes lugares, la "no interacción" indicará que la mejor variedad en general es tambien la mejor en cada uno de los lugares estudiados; o sea que las relaciones entre las medias de las variedades son las mismas en los diferentes lugares. Si la hipótesis de no interacción es rechazada, ello indicará que para cada - lugar se tiene cierto tipo de relaciones entre las medias de variedades. Posiblemente, la mejor variedad en general no sea la mejor en todos los lugares, lo que indicará la necesidad de estudiar las relaciones entre medias de variedades en cada lugar por separado.

Otro aspecto de interés práctico es cuando se investiga el patrón de cambio que siguen las medias de poblaciones - generadas al modificar los niveles de un factor con respecto a los niveles dados; por ejemplo, si el cambio en las medias es de tipo lineal, cubico, cuadrático, etc.; en general, con las medias estimadas de poblaciones de un experimento se puede investigar la relación funcional entre dichas medias y los niveles de los factores - empleados en el experimento; esto es, lo que se conoce como "Superficie de Respuesta". De hecho, se usan primero los modelos de diseños experimentales y posteriormente, considerando las medias de tratamientos estimadas como variables dependientes y los niveles - de los factores cuantitativos como independientes, se emplea un modelo de regresión.

Es posible también comparar ciertos grupos de poblaciones, considerados como una sola población con mayor grado de

generalidad, con otro grupo de poblaciones, a través de los llamados "contrastes". Esto es un valioso auxiliar para investigar la naturaleza de los cambios en las medias de las poblaciones al variar los diversos factores en estudio.

En ocasiones se estudia la interrelación de variables reales (como las de regresión) y las cualitativas de clasificación del diseño. Este es otro de los posibles usos de los modelos de covarianza; por ejemplo, al analizar los rendimientos de diferentes poblaciones de plantas, definidas por distintas prácticas de cultivo o variedades (tratamientos), puede interesar la comparación de medias de rendimiento de las poblaciones (tratamientos) en un mismo valor de número de plantas presentes y grado de enfermedad; esto se consigue en forma teórica empleando un modelo de covarianza, donde las covariables son las medidas del número de plantas y del grado de enfermedad. La información anterior sirve para investigar el efecto de los tratamientos sobre el rendimiento a través del número de plantas y grado de enfermedad comparando las diferencias de medias de las poblaciones (tratamientos) "ajustadas" por las covariables, o sea comparando un valor común de éstas con las diferencias de medias de tratamientos, pero ignorando las covariables. Por ejemplo, si al comparar tratamientos hay alguno que produce una poblacion con media de rendimiento mayor que las otras cuando la comparación se hace con un valor teórico constante de las covariables; mientras que al ignorar las covariables ese mismo tratamiento produce una media igual o menor que las demás. Esto indicará que el tratamiento en cuestión tiene potencialidad de alto rendimiento siempre y cuando por -

otros medios se pueda lograr un grado de enfermedad y un número de plantas como el logrado teóricamente con el análisis de covarianza.

c) Estimación de parámetros.- Los diseños experimentales pueden estar planeados para estimar el valor absoluto de medias de poblaciones, aunque esto es poco común. Por ser de utilidad práctica es muy frecuente la estimación de las varianzas de efectos aleatorios, a las que se llama componentes de varianza. - Una posibilidad es estimar la variabilidad con la cual los toros transmiten la potencialidad para la producción de leche. Con inseminación artificial, un macho puede tener muchos hijos a la vez; - si las vacas hijas llegan a producir leche, la calidad del toro se mide por la cantidad de leche de sus hijas, lo que es, en cierto - grado, independiente del aspecto físico del macho. En este caso - es importante estimar la variabilidad con la cual se presenta la - producción de leche. Así, el modelo del ejemplo puede ser:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

donde

Y_{ijk} producción de leche anual promedio de la vaca que es la k-ésima hija del toro j-ésimo - en el hato i-ésimo.

ε_{ijk} error aleatorio causado por las particularidades de la k-ésima hija del toro j-ésimo en el hato iésimo, ε_{ijk} $(0, \sigma^2)$.

α_i efecto del hato i-ésimo, o sea la diferencia $\mu_j - \mu$, donde μ_j es la media de la po-

blación teórica de descendientes del toro j-ésimo. Como a su vez los toros forman una población, se considera que β_j es un efecto aleatorio con varianza σ^2_β . Y es precisamente la estimacion de - σ^2_β la que es de importancia práctica.

7.1.3 METODOLOGIA DE SUPERFICIE DE RESPUESTA.

7.1.3.1 Conceptos básicos de optimización.- En cualquier campo de la ciencia, existen muchas soluciones para un problema - especifico. Seleccionar la mejor solución de entre ese gran número de potenciales soluciones no es de ninguna manera un problema nuevo, y menos en el área de Ingeniería Química.

Comúnmente, las respuestas en el pasado se basaban en cierta intuición o experiencia; ahora, gracias al desarrollo de la matemática, existen diversas técnicas o métodos para determinar las condiciones óptimas de un proceso. Una de esas técnicas la constituye la metodología de superficie de respuesta, tema central de este trabajo.

El primer punto importante por cubrir en cualquier técnica de optimización, es la definición de que es lo que hay que optimizar. Para esto, comúnmente se escoge una función objetivo, - que relacione costos, rendimiento, pureza y algunos otros crite- - rios para la optimización. Esta función puede ser expresada en la - forma de una ecuación tal como:

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3 + \ldots..$$

donde

X_1, X_2, X_3, ... factores o criterios usados para la optimización.

Y = función objetivo.

En seguida, debe de definirse que es el óptimo. En la función objetivo, este puede ser un máximo o un mínimo. Si la función es diferenciable, estos pueden obtenerse en base a derivadas sucesivas parciales. Se debe tomar en cuenta, también, la naturaleza del máximo, o mínimo, en cuestión. Estos pueden ser locales, globales o absolutos*; adicionalmente, se podría localizar un óptimo en relación con el valor de las X's, pero el valor de Y máximo (o mínimo) no es verdadero o practicable (por ejemplo, una concentración negativa).

Existen ciertas reglas acerca de como optimizar una función; si esta consta de una sola variable, las reglas son:

1.- Existen valores extremos (máximo o mínimo) solamente donde $dY/dX = 0$ o donde dY/dX no existe. Esta regla es necesaria pero no suficiente, dado que otras condiciones deben ser alcanzadas.

2.- Si en el punto determinado de la regla 1 ciertas derivadas desaparecen, entonces la próxima derivada, que no desaparezca, es examinada por signo. Digamos que todas las derivadas hasta la n-derivada desaparecen:

$$dY/dX = d^2Y/dX^2 = d^3Y/dX^3 = \ldots = d^nY/dX^n = 0$$

---

* Absoluto es en todos los valores posibles de Y.

entonces la próxima derivada $d^{n\,1}Y/dX^{n\,1}$ es o positiva o negativa. Si n es un valor par, - hay un punto de inflexión. Si n es impar, la próxima derivada (la n + 1) es examinada; si esta es negativa, un máximo existe; si esta - es positiva, un mínimo existe.

3.- Cuando dY/dX no existe (esto es, hay discontinuidad*), los alrededores del punto extremo deben ser explorados. La primera derivada, dY/dX, es investigada dentro de esa área, cuando X incrementa su valor a traves de los alrededores del punto crítico. El signo de - dY/dX debe ser registrado: Si el signo de - la derivada pasa de más a menos, un máximo - existe; si el signo va de menos a más, un mínimo existe; si el signo no cambia, no hay - punto extremo.

Para más de una variable, el desarrollo de tales reglas es muy complejo. Para el caso de funciones de 2 variables - estas serían[4,5]:

1.- Evaluar las derivadas parciales $\partial Z/\partial X$ y - $\partial Z/\partial Y$.

2.- Si $\partial Z/\partial X = 0$ y $\partial Z/\partial Y = 0$, entonces un punto extremo existe y las etapas 4 y 5 deben seguirse. Si tanto $\partial Z/\partial X$ o $\partial Z/\partial Y$ no son cero, enton-

---

* Esto no en todos los casos.

ces no hay máximo ni mínimo.

3.- Evaluar $\partial^2 Z/\partial X^2$, $\partial^2 Z/\partial Y^2$ y $\partial^2 Z/(\partial X \partial Y)$, y evaluar el término M, donde M es

$$(\partial^2 Z/\partial X^2)(\partial^2 Z/\partial Y^2) - (\partial^2 Z/\partial X \partial Y)^2$$

4.- Observar la siguiente tabla, para determinar si una condición extrema existe. Esto constituye la condicion suficiente para un extremo:

M		Optimo
Positivo	$\partial^2 Z/\partial Y^2$ y $\partial^2 Z/\partial X^2$ son pos.	Min.
	$\partial^2 Z/\partial Y^2$ y $\partial^2 Z/\partial X^2$ son neg.	Max.
Negativo	.....	Punto silla
Cero	.....	Sin decisión

Aunque existen otro tipo de reglas, el procedimiento anterior será suficiente para analizar un modelo con 2 variables independientes.

7.1.3.2 Diseños Experimentales.- Los diseños experimentales estadísticos constituyen una herramienta más, enmarcada dentro de la metodología de superficie de respuesta. Por medio de diseños, se planean las condiciones en las que se va a efectuar un experimento y el número de estos a efectuar. Dichas condiciones, son especificadas en tal forma, que los experimentos proporcionan mayor información por experimento, que los experimentos no planeados, reduciendo su tiempo de ejecución y aumentando su eficiencia.

Asimismo, se provee de esta forma de un método or-

ganizado de colección y análisis de información. Frecuentemente, las conclusiones de un experimento diseñado estadísticamente son evidentes sin necesidad de un extensivo análisis estadístico. De esta forma, la informacion es más "digerible" y la autocritica - de los resultados guia a conclusiones más creibles, que si los - experimentos se hubieran realizado sin planeación, originando - que el análisis estadistico de estos últimos, fuera más complejo.

Como última caracteristica principal, la técnica provee medios para estimar las interacciones entre variables experimentales, guiando a predicciones más razonables de la respuesta en áreas no directamente cubiertas por la experimentación.

La mayoria de los diseños experimentales han sido desarrollados en conexión con la investigación agricola; esas técnicas también son aplicables a experimentación industrial y - tecnológica, pero existen ciertas condiciones bajo las cuales la técnica debe aplicarse en forma ligeramente diferente de como se hace en agricultura. La principal diferencia la constituye la velocidad. En el campo de la agricultura, los experimentos estan - restringidos comúnmente a uno por año. Existe entonces una gran - oportunidad de planear ese experimento, llevarlo a cabo y analizarlo, antes de que la planeación del próximo experimento sea - - iniciada. Dado que solo se tiene un experimento por temporada, - es evidente que se debe planear uno que sea lo suficientemente - completo (y por lo tanto, complejo) para asi obtener el máximo de información en un tiempo razonable. Igualmente, el procedimiento entero puede ser supervisado por una oficina científica, tal vez con respaldo universitario.

Sin embargo, esto no es posible hacerlo en la -

industria. Una simple máquina produce 1200 focos en una hora, y un riel de ferrocarril puede ser elaborado en 10 minutos. Igualmente, cuando existe una vasta producción en serie, la supervisión no es una actividad en manos de científicos, sino de personal más dedicado a la práctica.

Llevar a cabo un experimento diseñado bajo esas condiciones requiere un considerable esfuerzo en organización, y las personas, ya sea de producción, planeación o estimación, solamente se inclinan a tomar en cuenta tales métodos de experimentación, si ellas son convencidas de su utilidad y además de si son capaces de comprender los resultados útiles logrados.

Dado que los experimentos tienen una mayor velocidad de respuesta, los diseños pueden ser sencillos, con el objeto de que suministren la información en forma clara y concisa, ya sea para seguir la experimentación en otras condiciones de operación, o bien para determinar los niveles de los factores ya muy cercanos al óptimo. Igualmente, se hace indispensable que el número de experimentos sea pequeño, ya que el costo de cada uno de ellos pudiera ser muy elevado, situación que se torna difícil si no se cuenta más que con los recursos propios de la empresa.

Debido a este tipo de situaciones, los diseños más utilizados en el área de Ingeniería Química, son los siguientes:

a) Diseños Factoriales
b) Diseños Factoriales Fraccionales
c) Diseños de Composición Central

a) Diseños Factoriales.- En un experimento factorial se investigan simultaneamente los efectos de cierto número de factores. Los tratamientos constan de todas las combinaciones que puedan formarse de los distintos factores, esto es, número de tratamientos = M^n, donde M es el número de niveles, y n es el número de factores.

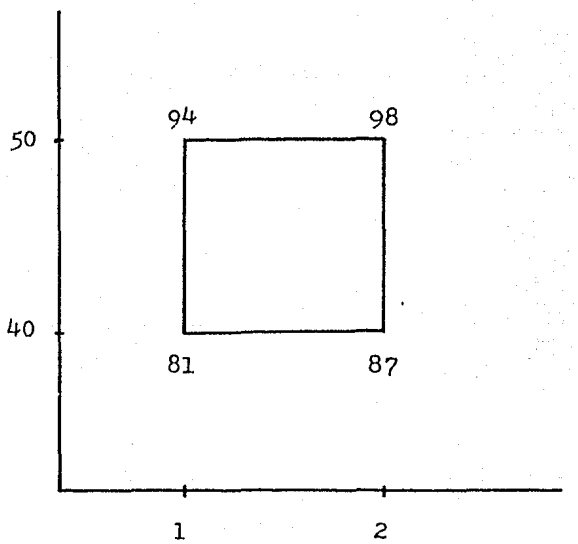
En un 2^2 factorial, los experimentos son corridos a 2 niveles, para cada una de las 2 variables, dando un total de - $2^2 = 4$ corridas. La Fig. 7.26 representa esquematicamente este diseño. Al nivel con menor valor se le conoce como nivel bajo, y - convencionalmente se representa por un signo menos (-). El otro - nivel, alto, se representa por un signo más (+). Los valores - (X_1, X_2) que definen los puntos del diseño, pueden ser absolutos o, para simplificar los cálculos, codificados.

Para ilustrar lo anterior, considere el siguiente ejemplo: en un reactor, se produce una sustancia con una determinada pureza, la cual constituye la respuesta Y. Las variables involucradas son temperatura y tiempo de reacción, más especifica- - mente, tiempo de residencia promedio. El esquema del diseño y sus resultados se presentan en la Fig. 7.27.

Se pueden resumir los resultados de la siguiente manera: Considerando el efecto de la temperatura, se puede decir - que el aumento de 10^oC, produjo una pureza mayor de 13% cuando el tiempo fue 1 hora, que el aumento obtenido en 2 horas de residencia (11%). Estas cifras se llaman efectos simples de la temperatura y representan el tipo de información que podría desearse, por - ejemplo, para aconsejar un aumento de temperatura, en reactores -

que siempre han tenido una hora de tiempo de residencia, e igualmente, considerando el efecto simple del tiempo, se puede decir - que este aumentó la pureza un 6% en reactores con temperatura baja, y un 4% con temperatura alta.

Hay, sin embargo, otra manera de ver los resultados; sucede a veces que los efectos de los factores son independientes; con esto se quiere decir que la respuesta a la temperatu-



Temperatura

		40	50	Media	Diferencia
Tiempo	1	81	94	87.5	13
	2	87	98	92.5	11
Media		84	96	90.0	
Diferencia		6	4		

Fig. 7.27

ra es la misma, ya sea que el tiempo de residencia sea una o 2 horas, y por lo tanto, la diferencia en pureza será la misma, ya sea que la reacción se lleve a cabo en uno u otro tiempo. En este caso, los efectos simples de la temperatura son estimaciones de la misma cantidad, y difieren únicamente por errores experimentales. Con esta suposición, se podrían promediar las estimaciones, y el resultado se llama el efecto principal de la temperatura, el cual sería 11 + 13 = 24/2 = 12%. Alternativamente, el efecto principal del tiempo sería 5%.* Dado que el efecto principal es estimado en base a un mayor número de tratamientos, su varianza será menor que la de los efectos simples**.

En consecuencia, si se esta seguro de que los factores operan independientemente, el resumen anterior que se dió en términos de efectos simples, puede ser reemplazado con otro que sea a la vez más conciso y preciso. Este podría ser como sigue: El aumento de temperatura en 10°C aumentó la pureza en 12%, o bien, se aumento 5% la pureza con un aumento de 1 hora en el tiempo de residencia. Vale la pena repetir, que estos resultados, cuando los

---

* Observese que el efecto principal de t se estima como: t = ((Tt) + (t) - (T) - (1))/2, donde (Tt) = experimento con los niveles altos de temperatura y tiempo, (T) = experimento con nivel alto de temperatura, (t) = experimento con nivel alto de tiempo, y (1) = experimento con nivel bajo de todos los factores.

** El error estándar de un efecto principal es $1/(2)^{\frac{1}{2}}$ veces el error estándar del de un efecto simple, para el caso de un experimento factorial 2^2.

factores son independientes, son los mejores, ya que ambos utilizan los 4 puntos del experimento para su estimación.

En otras palabras, la información integra del experimento esta contenida en los efectos principales.

Como puede observarse, estos resultados tienen su base teórica en los puntos discutidos en la sección 7.1.2.2. Ahí mismo, se planteó la posibilidad de que los factores interactúen; esto solo puede averiguarse mediante el conocimiento de los procesos por los cuales los factores producen sus efectos. En el caso anterior, si ΔH^{o} de la reacción es positivo, un incremento en la temperatura llevara consigo un aumento en el valor de la constante de equilibrio y es de esperarse, que si se da un tiempo suficiente para llegar a este equlibrio, la concentración del producto aumentaría. Esto es, es muy probable que exista interacción entre los factores.

Aun cuando exista interacción, la información que provee un diseño factorial es de suma importancia, ya que así podemos revelar esta relación, y saber que a diferentes niveles de otros factores, la respuesta no será la misma para el nivel de un factor determinado, cosa que no pueden detectar métodos como el del factor único.

Para codificar los factores se puede utilizar la fórmula:

$$X = \frac{t - \mu}{d}$$

donde

μ = media de los niveles considerados

t = valor del nivel a codificar

d = $\frac{1}{2}$ de la amplitud del rango de valores cubierto por los niveles bajo y alto, o bien - cualquier valor arbitrario que simplifique - los cálculos.

X = valor de la variable t codificado.

Igualmente, la respuesta puede ser transformada - por una fórmula como:

$$Z = Y - W$$

donde

W = cualquier número arbitrario

Esto puede hacerse, ya que este tipo de transformaciones son lineales. Para el ejemplo anterior los cálculos po- - drían ser:

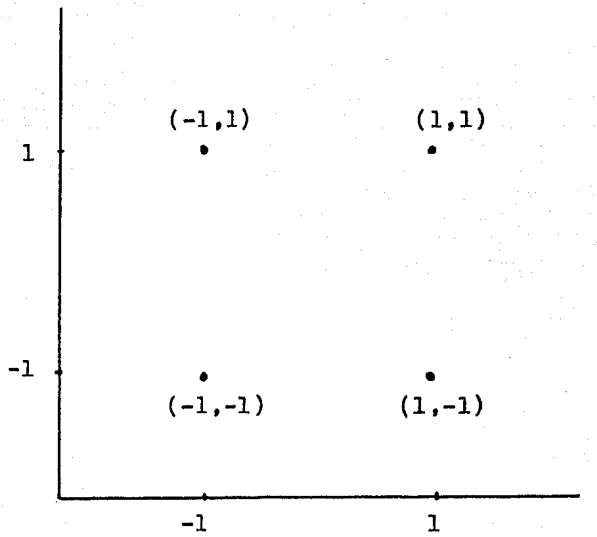
$$t_{1+} = \frac{50 - 45}{5} = 1 \qquad t_{2+} = \frac{2 - 1.5}{.5} = 1$$

$$t_{1-} = \frac{40 - 45}{5} = -1 \qquad t_{2-} = \frac{1 - 1.5}{.5} = -1$$

$$Z = 100 - Y$$

Con esta última transformación, la búsqueda del -- óptimo pasa a ser la búsqueda de un mínimo en lugar de un máximo.

Cuando el número de factores aumenta se recomienda



utilizar una tabla de diseño, para identificar las corridas; para - 2 y 3 factores, estas son:

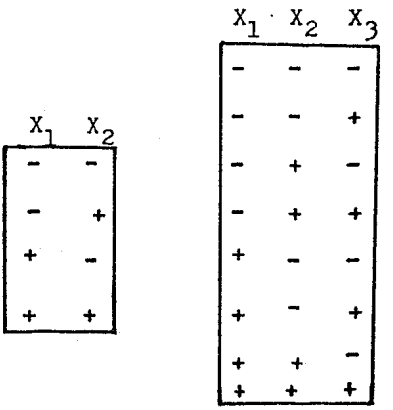


Fig. 7.29

La desventaja de los diseños 2^n factoriales, estriba en el hecho de que un gran número de corridas son necesarias

cuando n aumenta. Por ejemplo, cuando n = 10, 2^{10} = 1024, lo cual en la gran mayoría de los casos, representa un elevadísimo costo, y por lo tanto, un diseño impracticable. Para este tipo de situaciones, se utilizan los diseños factoriales fraccionales.

Como se dijo antes, los diseños factoriales se utilizan para estimarefectos principales, y efectos de interacción, pero de ninguna manera efectos de curvatura. Así, considere el diseño de la figura 7.31. Con 4 puntos, es posible ajustar los si-

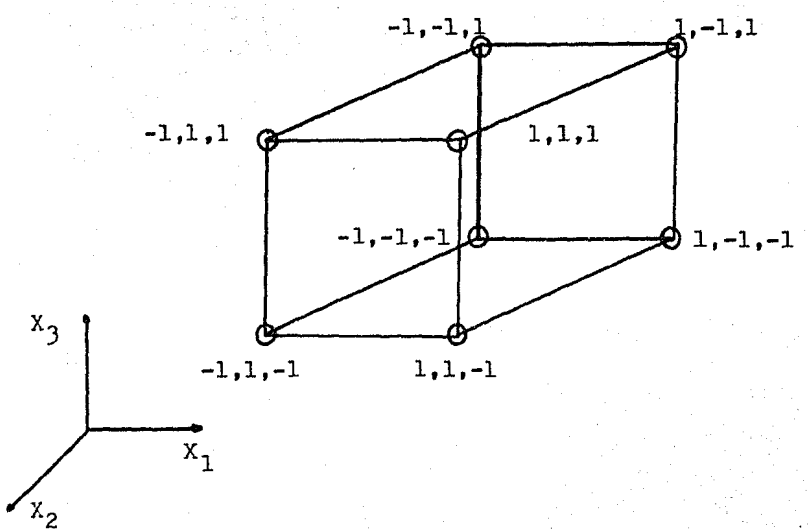


Fig. 7.30

guientes modelos, agotando los grados de libertad.

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3X_2$$

$$Y = b_o + b_1X_1 + b_2X_2 + b_{11}X_1^2$$

$$Y = b_o + b_1X_1 + b_2X_2 + b_{22}X_2^2$$

Los dos últimos, matemáticamente daran un resultado, pero este puede ser completamente erróneo. Ningún punto del diseño contempla la posibilidad de estimar efectos de curvatura; para

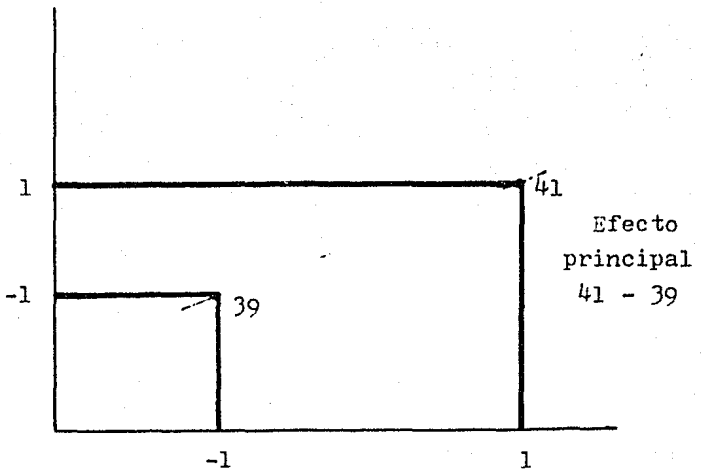


Fig. 7.31

hacer esto, sería necesario agregar un punto, el cual comúnmente se sitúa en el centro de los niveles del factor del cual se desea estimar la curvatura (fig. 7.32).

En un diseño 2^2 se opta por colocar un punto en el centro de este (0,0).

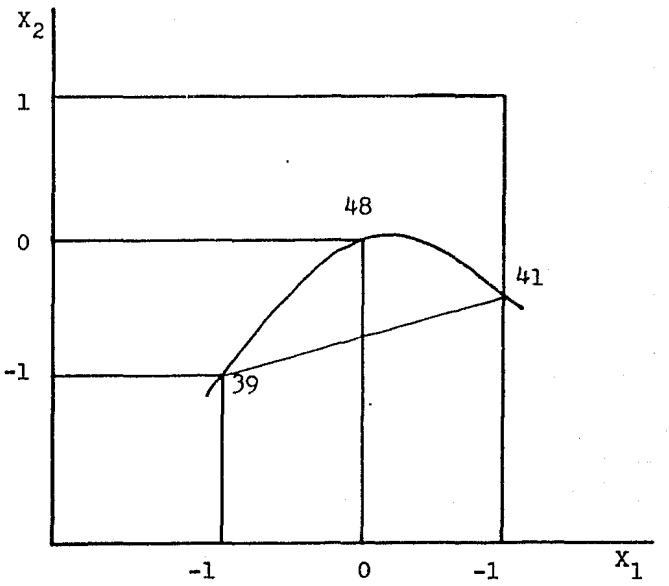


Fig. 7.32

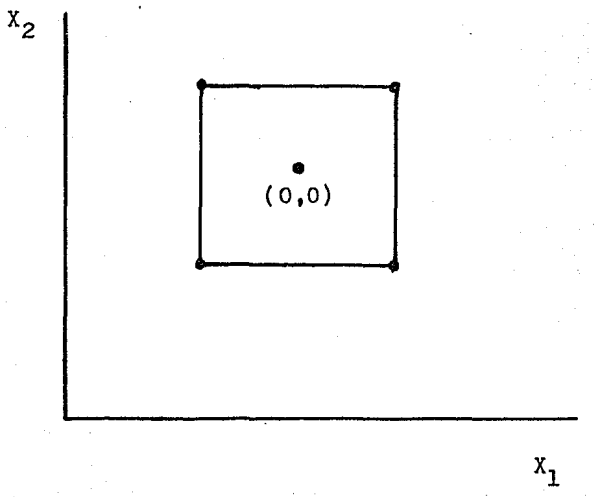


Fig. 7.33

Sin embargo, para estimar efectos de curvatura, - son más adecuados los diseños de composición central.

b) Diseños Factoriales Fraccionales.- El principal atractivo de los diseños factoriales fraccionales consiste en que - permiten incluir simultáneamente 5 o más factores en un experimento de tamaño práctico, de tal manera que el investigador puede determinar rápidamente cuales factores tienen un efecto importante sobre el resultado. Sin embargo, la disminución del tamaño del experimento no se obtiene sin sacrificar otros aspectos. Los resultados de - un experimento factorial fraccional estan propensos a mala inter- - pretación, que no se presentaría en un factorial completo. Para entender esto, considere el siguiente caso:

Se tiene un factorial 2^3 en el cual solamente se - prueban 4 combinaciones de tratamientos a,b,c y abc. Esta es la mitad de una repetición completa y equivale a un 2^{3-1}. Conforme a la notación ya empleada, los efectos principales e interacciones podrían estimarse a traves del cuadro siguiente:

	EFECTOS FACTORIALES						
Combinación de Tratamientos (*).	A	B	C	AB	AC	BC	ABC
a	+	-	-	-	-	+	+
b	-	+	-	-	+	-	+
c	-	-	+	+	-	-	+
abc	+	+	+	+	+	+	+

Para el efecto principal de A, se suman los ren- - dimientos de las unidades que contienen una a, y se restan los ren-

(*) Niveles altos.

dimientos de las unidades que no la contienen*; esto es:

$$A = (abc) + (a) - (b) - (c)$$

y de la misma manera para los efectos principales de B y C. Puede observarse en el cuadro que las comparaciones A, B y C son totalmente diferentes, u ortogonales. Por lo tanto, las estimaciones de los 3 efectos principales son independientes.

Para calcular la interacción AB, se sigue la regla de "pares contra nones". Esto es, se suman los rendimientos de las unidades que contienen un número par de letras, a, b y se restan los rendimientos de las unidades que contienen un número non. Esto da:

$$AB = (abc) + (c) - (a) - (b)$$

Pero esta es la misma cantidad que se usa para estimar el efecto principal de C. El nombre de alias se da a dos efectos factoriales representados por la misma comparación (o tratamiento). Así, C y AB son alias. Se escribe C = AB. El cuadro anterior muestra también que:

$$AC = B \qquad BC = A$$

La interacción triple no se puede estimar. Si estuvieran disponibles las 8 combinaciones de tratamientos, ABC podría calcularse de la siguiente manera:

$$ABC = (abc) + (a) + (b) + (c) - (ab) - (ac) - (bc) - (1)$$

---

* Además de un factor específico, por el cual se dividen las comparaciones. En el caso de A, ese factor es 2^{13}.

Obsérvese, que se han escogido las 4 combinaciones de tratamientos que llevan el signo más en esta expresión, - para que ellos sean los tratamientos a efectuar en este diseño - factorial fraccional, llamado mitad de repetición*; ABC se llama contraste de definición, ya que fue el efecto (o contraste) usado para dividir el factorial en 2 mitades de repetición.

Como resultado del uso de una mitad de repetición, se pierde completamente un efecto factorial, ABC, y se deja cada - efecto principal completamente mezclado con una de las interacciones dobles.

Si el experimento muestra un efecto aparente de A, no hay forma de conocer los resultados de si el efecto se debe - realmente a A, a la interacción BC, o a la mezcla de los dos. Esta clase de confusión se presenta siempre en experimentos factoriales fraccionales, ya que cada efecto factorial tiene siempre uno o más alias; al interpretar los resultados, el experimentador se enfrenta al problema de decidir a que alias se debe atribuir un efecto.

En algunos tipos de investigación, la experiencia previa con los mismos factores, o un conocimiento de la naturaleza de sus acciones, pueden conducir al investigador a predecir con - toda confianza que sus interacciones no tienen importancia. Así, - la ambigüedad desaparece: Se concluye que un efecto aparente de A, de hecho, se debe a A, y lo mismo para B y C. Con estas consideraciones, la mitad de repeticion provee estimaciones independientes de los tres efectos principales.

---

* Se llama así, porque el número de experimentos es la mitad del de un factorial completo. Asimismo, existen cuartos de repetición, etc.

Si se desea una reducción mayor del experimento, se puede utilizar un cuarto de repetición, con un número de corridas igual a 2^{n-2}. En un cuarto de repetición, sin embargo, cada efecto esta confundido con 2 efectos más.

Para entender mejor los conceptos anteriores, se desarrollara un cuarto de repetición, y sus posibles implicaciones, de un experimento factorial 2^6.

Un cuarto de repetición, contendrá $2^{6-2} = 2^4 = 16$ corridas o unidades experimentales. Para obtener esas 16 corridas se necesitan definir 2 contrastes de definición, los cuales dividiran el factorial completo, primero en una mitad de repetición, de 32 corridas, y luego en el cuarto de repetición deseado, con 16 corridas. Dado que los contrastes de definición no podrán ser estimados, como pasó con la interacción ABC anteriormente, es aconsejable tomar una interacción de alto orden, como ABCDEF y ABCDE; sin embargo, esto debe hacerse con cuidado, ya que podriamos, sin querer, dejar de estimar un efecto principal. Para los contrastes elegidos, los bloques quedarian:

abcdef	abcd	abce	abcf	abde	abdf	abef	acde
acdf	acef	adef	bcde	bcdf	bcef	bdef	cdef
ab	ac	ad	ae	af	bc	bd	be
bf	cd	ce	cf	de	df	ef	(1)

Tabla 7.12

para obtener este bloque de corridas, se hace lo siguiente: se multiplica el contraste de definición, ABCDEF, con las 64 combinaciones del factorial completo; los términos cuadráticos, se eliminan, que-

dando solo los términos a la primera potencia. Así, por ejemplo, - ABCDEF * ABC = $A^2B^2C^2DEF$ = DEF. Estos son los alias de cada una - de las 32 corridas, donde las 32 son aquellas que contienen 6,4,2 o ninguna de las letras ABCDEF. Así, ABCD, que es una de las corridas, tiene como alias ABCDEF * ABCD = EF, que también forma - parte de las 32 unidades.

Para formar el cuarto de repetición, y si se toma como contraste el efecto ABCDE, se escogen, de las 32 unidades anteriores, aquellas que contengan 5, 3 o 1 de las letras ABCDE. Se obtendran los siguientes 16 tratamientos:

abcdef	abcf	abdf	abef	acdf	acef	adef	bcdf
bcef	bdef	cdef	af	bf	cf	df	ef

Estas son las 16 combinaciones que tienen signo - positivo en la expresión de ABCDEF y tambien en la expresión de - ABCDE. Nótese, sin embargo, que las 16 combinaciones contienen la letra F. Se ha producido, inadvertidamente un diseño en el cual el efecto principal de F no puede estimarse, ya que no hay una sola - corrida en la cual F este a nivel bajo; de aquí, podemos concluir que, al no poderse estimar, F se ha convertido en un contraste de definición; si se hubiera usado F en lugar de ABCDE para construir el cuarto de repetición, se hubiese obtenido el mismo bloque de 16 tratamientos.

Este ejemplo es una ilustración de la regla general:

"En el sistema 2^n, cualquier efecto de 2 factores puede usarse como contraste de definición para dividir un experimento -

factorial en cuartos de repetición. Su interacción generalizada - actúa también como un contraste de definición y no puede estimarse del cuarto de repetición".

La interacción generalizada de ABCDEF y ABCDE es la multiplicacion ABCDEF * ABCDE = $A^2B^2C^2D^2E^2F$, o sea F. Por lo - tanto, la elección de los contrastes no ha sido satisfactoria, y se deben escoger otros.

En un cuarto de repeticion, los alias se obtienen al multiplicar por los 3 contrastes de definición, los 16 tratamientos obtenidos. Por ejemplo, los alias de A son A * ABCDEF = BCDEF, A * ABCDE = BCDE y A * F = AF.

Al buscar un diseño mejor para un cuarto de repetición de un 2^6 factorial, puede desearse que los contrastes de - definición, todos los alias de los efectos principales y de las - interacciones de 2 factores sean interacciones de orden superior - que puedan considerarse despreciables. Sin embargo, en este caso, esto no es posible realizarlo. Si se escogen los contrastes nuevos como los tratamientos ABCE, ABDF y ABCE * ABDF = CDEF, las 16 corridas y sus respectivos alias serían las resultantes de la tabla 7.13.

Aislando las interacciones de dos factores (tabla 7.14), este diseño puede usarse en varias situaciones. Si todas - las interacciones de 2 factores se consideran relativamente pequeñas, se estiman solamente los efectos principales y las comparaciones restantes proveen grados de libertad para el análisis estadístico. El análisis de varianza se simplifica como sigue:

Efectos principales 6 grados de libertad

Tratamiento	Alias		
abcdef	df	ce	ab
acf	bef	bcd	ade
abdf	cdef	(1)	abce
aef	bcf	bde	acd
cdef	abdf	abce	(1)
bcf	aef	acd	bde
abce	(1)	cdef	abdf
bef	acf	ade	bcd
acd	bde	bcf	aef
ab	ce	adf	abcdef
ade	bcd	bef	af
ce	ab	abcdef	df
bcd	ade	acf	bef
df	abcdef	ab	ce
bde	acd	aef	bc
(1)	abce	abdf	cdef

Tabla 7.13

Interacciones de 2 factores	Alias		
AB	CE	DF	ABCDEF
AC	BE	BCDF	ADEF
AD	BF	BDCE	ACEF
AE	BC	BDEF	ACDF
AF	BD	BCEF	ACDE
CD	EF	ABDE	ABCF
CF	DE	ABEF	ABCD

Tabla 7.14

Error residual (de las interacciones)	9 D. F.
Total	15 D. F.

Alternativamente, pueden estimarse algunas de las interacciones de 2 factores, ya que sus alias son despreciables. - Suponga que E y F no se interaccionan con ninguno de los otros - factores. En los grupos de alias de la Tabla 7.14, los primeros 6 grupos pueden considerarse como estimaciones de AB, AC, AD, BC, BD y CD, respectivamente. El séptimo grupo, que contiene CF y DE, contribuye a la estimación del error. Esto es, hasta donde es posible llegar en el desenmarañamiento de las interacciones de 2 - factores. Los D. F. se subdividen como:

Efectos principales	6 D. F.
Interacciones Dobles de A,B,C,D	6 D. F.
Error (de las interacciones restantes)	3 D. F.
Total	15 D.F.

Como se puede apreciar, los D. F. son escasos.

El siguiente ejemplo, ilustra el riesgo de mala-interpretación en diseños factoriales fraccionales:

Hay 3 factores. A no tiene efecto, mientras que B y C tienen efectos positivos y una interacción positiva. Se supondrá que B provee un incremento de 40 cuando se aplica al nivel más bajo de C y de 60 al nivel más alto de C, mientras que C provee un incremento de 20 y 40 a los niveles más bajo y alto de B. Con un nivel básico de 100 para la combinación (1) y sin error - experimental, los resultados para los 8 tratamientos son como -- sigue:

(a)	100	(1)	100
(b)	140	(ab)	140
(c)	120	(ac)	120
(abc)	180	(bc)	180

Estos resultados son desconocidos. Si se ensaya la mitad de repetición, con ABC como contraste (columna izquierda), los efectos principales se estiman como:

$$A = \frac{((abc) + (a) - (b) - (c))}{2} = 10$$

$$B = \frac{((abc) + (b) - (a) - (c))}{2} = 50$$

$$C = \frac{((abc) + (c) - (a) - (b))}{2} = 30$$

si se escoge como contraste a (1), los resultados serían:

$$A = -10$$
$$B = 50$$
$$C = 30$$

Las dos mitades de repetición dan los mismos resultados para B y C, siendo además el resultado correcto para las 8 combinaciones de tratamientos. Sin embargo, la primera mitad de repetición muestra un aumento erróneo de 10 para A, mientras que la segunda mitad muestra una disminucion errónea de 10. Ningúno de los bloques provee una señal acerca de la presencia de la interacción BC (a menos que el experimentador suponga que ésta podría ser una interpretación de los resultados).

El daño que se haga depende de las decisiones que se tomen a partir de estos resultados. Se considererán brevemente

tres situaciones:

1.- Si el experimento es primordialmente del tipo tamiz, diseñado para obtener los factores más importantes, puede llegarse a la conclusión que B y C son importantes y que A lo es relativamente menos. Una experimentación más avanzada se restringe a B y C y se supone que revele la interacción BC. Se ha hecho poco o ningun daño.

2.- Si el experimento es parte de un programa de investigación básica, el investigador puede concluir equivocadamente, ya sea que A tenga un efecto benéfico, o que tenga un efecto perjudicial. Si no se efectuara una experimentación posterior, pueden perderse energías y tiempo elaborando teorías incorrectas que expliquen el "efecto" de A o en discusión del tema.

3.- En un programa de investigación aplicada, el propósito puede ser seleccionar la mejor combinación de niveles entre los factores. En la primera mitad de repetición, (abc) obtiene el rendimiento más elevado. Si los resultados de esta mitad de repetición se usan para especular sobre la respuesta de las combinaciones de tratamientos (ab), (ac), y (bc), que no se probarón, todas estas combinaciones de tratamientos podrán predecirse como inferiores a -

(abc). Por ejemplo, (bc) se predice como 10 unidades inferior a (abc), ya que A es causa de un incremento de 10 en esta mitad de repetición. De aquí que se recomiende (abc) como óptimo.

Si se elige la segunda mitad de repetición, (bc) da el mejor rendimiento observado. Podrá predecirse que sea superior a una combinación como (abc), que no se probó, ya que - A produce una disminución de 10 en la segunda mitad de repetición.

Realmente, ambas selecciones (abc) y (bc) dan el rendimiento correcto más alto. o sea, 180.

La duda estriba en cual debe ser el nivel de A. Si el nivel que se señala para A supone - una diferencia sustancial en costo, y si el - investigador llega a conclusiones definitivas sin una mayor experimentación, tiene la posibilidad de recomendar un "óptimo" innecesa- - riamente caro.

Por lo dicho antes, claramente se ve que no puede hacerse una aseveración concisa con relación al peligro de una mala interpretación de los resultados de un factorial fraccionado. Las consecuencias pueden o no tener importancia. En general, los diseños fraccionados serán convenientes si el riesgo de confusión por la presencia de las interacciones de los factores, es pequeño.

c) Diseños de composición central.- Este tipo de diseños, se utiliza para estimar efectos de curvatura, y por tanto, es un elemento de los diseños de 2^0 orden.

El rasgo caracteristico de un diseño de este tipo, es que los puntos utilizados para estimar los efectos de 2^0 orden, estan situados en la periferia* del espacio definido por un diseño factorial, y su varianza con respecto a Y, es la misma que la de - los puntos factoriales:

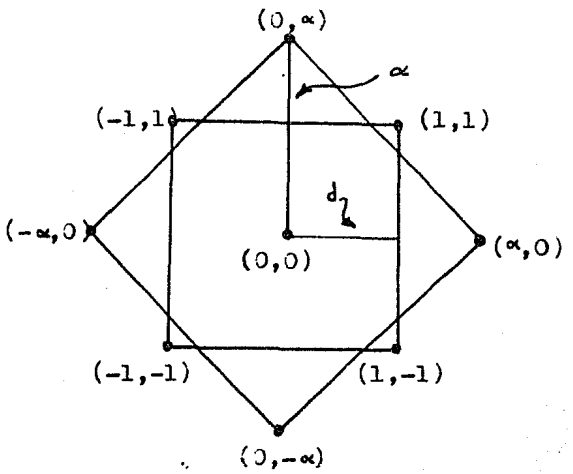


Fig. 7.34

Diseño de composición central para n = 2

para el caso de 3 factores, la representacion esquemática la constituye la Fig. 7.35, en la cual, los puntos periféricos, o estre- - lla, son 2 * 3 = 6.

Asimismo, los diseños de composicion central proveen de puntos "replicados", con el objeto de probar una posible -

---

* A una distancia α situada entre 1 y $2^{n/4}$ veces la distancia d.

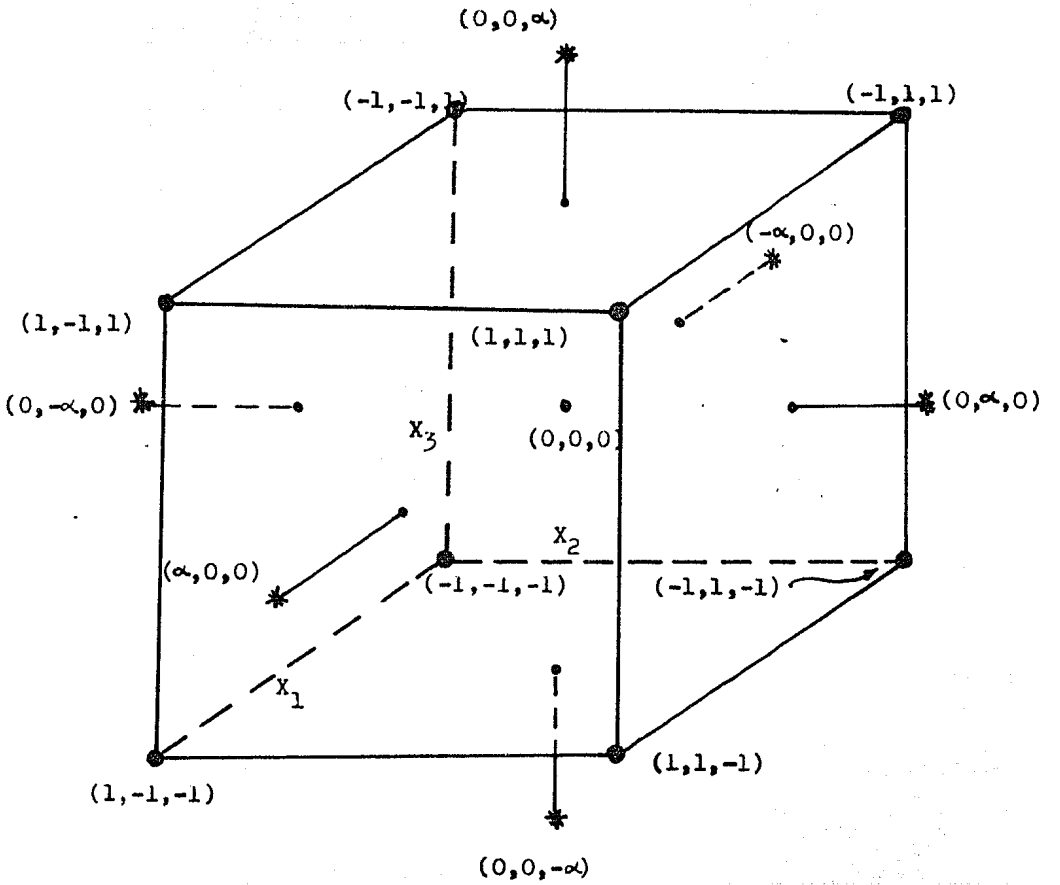


Fig. 7.35

Diseño de composición central para n = 3

'falta de ajuste'; esos puntos, son corridos al nivel (0,0). El número de ellos es arbitrario, siendo 5 o 6, los más adecuados.

El número de corridas, así como los niveles de los factores involucrados son:

	Número de Corrida	X_1	X_2
Para n = 2	1	-1	-1
	2	-1	1
	3	1	-1
	4	1	1
	5	$-\alpha$	0
	6	$+\alpha$	0
	7	0	$-\alpha$
	8	0	$+\alpha$
	9	0	0
	.	.	.
	n	0	0

Para n mayor a 3, el número de corridas se determina por $2^n + 2 * n$; así un diseño de composición central de 5 factores tendrá $2^5 = 32$ corridas factoriales + $2 * 5 = 10$ corridas estrella. A este número, se le debe aumentar el número elegido de corridas "replicadas", a nivel (0,0). El nivel de α deberá estar entre 1 y $2^{5/4} = 2.378$ veces d.

En general, este tipo de diseño debe practicarse en las últimas etapas del programa de investigación, ya que el objetivo final de este, es determinar los niveles óptimos de los factores en estudio. Estos últimos, deberán ser los verdaderamente

	Número de corrida	X_1	X_2	X_3
	1	-1	-1	-1
	2	-1	-1	1
	3	-1	1	-1
	4	-1	1	1
	5	1	-1	-1
	6	1	-1	1
	7	1	1	-1
	8	1	1	1
Para n = 3	9	$-\alpha$	0	0
	10	0	$-\alpha$	0
	11	0	0	$-\alpha$
	12	α	0	0
	13	0	α	0
	14	0	0	α
	15	0	0	0
	16	.	.	.
	n	0	0	0

preponderantes, ya que el análisis se enfoca a modelos de 2^o y hasta 3^o orden. El modelo completo de 2^o orden para 2 variables es:

$$Y = b_o + b_1X_1 + b_2X_2 + b_{12}X_1X_2 + b_{11}X_1^2 + b_{22}X_2^2 \qquad (X_o = 1)$$

y para 3 variables, es:

$$Y = b_o + b_1X_1 + b_2X_2 + b_3X_3 + b_{12}X_1X_2$$

$$b_{13}X_1X_3 + b_{23}X_2X_3 + b_{11}X_1^2 + b_{22}X_2^2 + b_{33}X_3^2 \qquad (X_o = 1)$$

Dado el número de términos, el análisis de dichos

es complejo, y puede involucrar costos innecesarios, si alguna de las variables incluidas no es preponderante.

7.1.3.3 Experimentación Secuencial.- Un experimento secuencial es aquel en que el experimentador puede detenerse después de cada observación, y examinar los resultados acumulados hasta ese momento, antes de decidir si se continua con el experimento.

En otras palabras, el análisis puede hacerse secuencialmente del mismo modo que el experimento. Esto se debe a 2 características propias de los experimentos:

1) Los tratamientos son aplicados a las unidades experimentales en alguna secuencia de tiempo definida y
2) El proceso de medida es rápido, de tal forma que el rendimiento o respuesta de cualquier unidad es conocido antes de que el experimentador trate la siguiente unidad, en la secuencia de tiempo.

Debido a las características anteriores, la experimentación secuencial es altamente adecuada al área de Ingeniería*; enmarcada dentro de la metodologia de superficie de respuesta, la experimentación secuencial reune todos los elementos en una secuencia de pasos lógicos para poder llegar a determinar los niveles óptimos de variables a estudiar en un programa experimental completo.

Una de esas herramientas, y que no se ha mencion-

---

* Ver sección 7.1.3.2

do hasta ahora, es el método del gradiente, o de la máxima pendiente*. Este metodo, consiste en buscar la dirección en la cual, se trasladará la zona de experimentacion de un conjunto a otro de niveles de las X_i.

Este conjunto se escoge de tal manera, que ocurra el máximo (o minimo) incremento esperado en Y. Para el cálculo matemático**, la dirección que se toma es precisamente la señalada por el gradiente de la función objetivo, el cual se define como:

$$\nabla\phi = \frac{\partial \phi}{\partial X_1} + \frac{\partial \phi}{\partial X_2} + \cdots\cdots + \frac{\partial \phi}{\partial X_n}$$

donde

n = número de factores involucrados.

Si la ecuación es lineal de 1^o orden, y sin interacciones:

$$\nabla Y = b_o + b_1X_1 + b_2X_2 + \cdots\cdots + b_nX_n$$

la dirección buscada es calculada en forma muy simple, como:

$$\frac{\partial Y}{\partial X_1} + \frac{\partial Y}{\partial X_2} + \cdots\cdots + \frac{\partial Y}{\partial X_n}$$

y esto da:

$$\nabla Y = b_1 + b_2 + \cdots\cdots + b_n$$

De esta forma, la zona de experimentación se mueve b_1 unidades en el eje X_1, b_2 unidades en el eje X_2, ..., b_n unida-

---

* También llamado de Box y Wilson[6,7,8]

** Referencia 19 Capitulo 8.

des en el eje X_n (situado en el hiperplano).

Para n = 2, esto se representa esquematicamente en la siguiente figura:

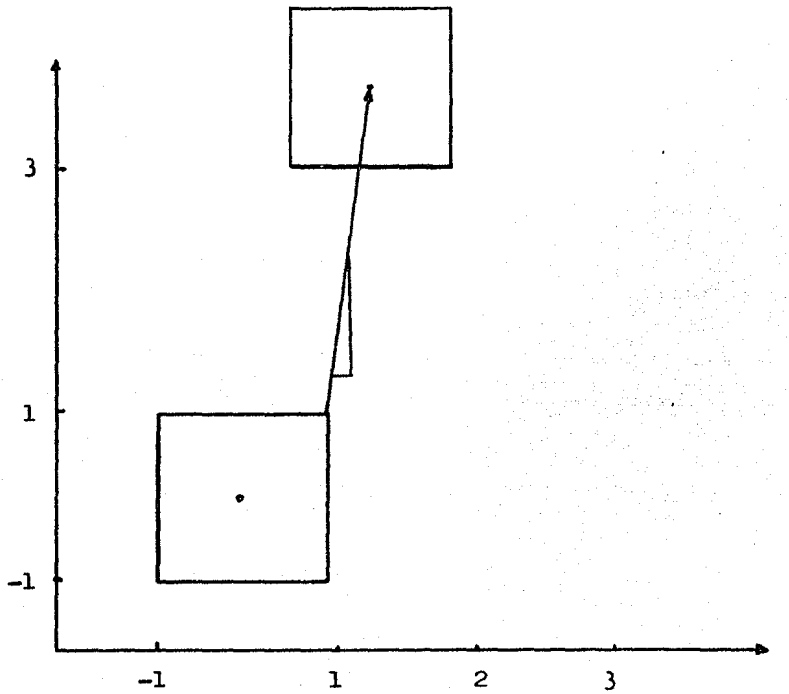


Fig. 7.36

Si lo que se desea es minimizar la función, lo - unico que se debe hacer es cambiar el sentido del gradiente, o sea tomar $\nabla\phi$ como $-\nabla\phi$;

Utilizando experimentación secuencial, la secuencia lógica de pasos a seguir en la optimización se puede resumir en la figura siguiente:

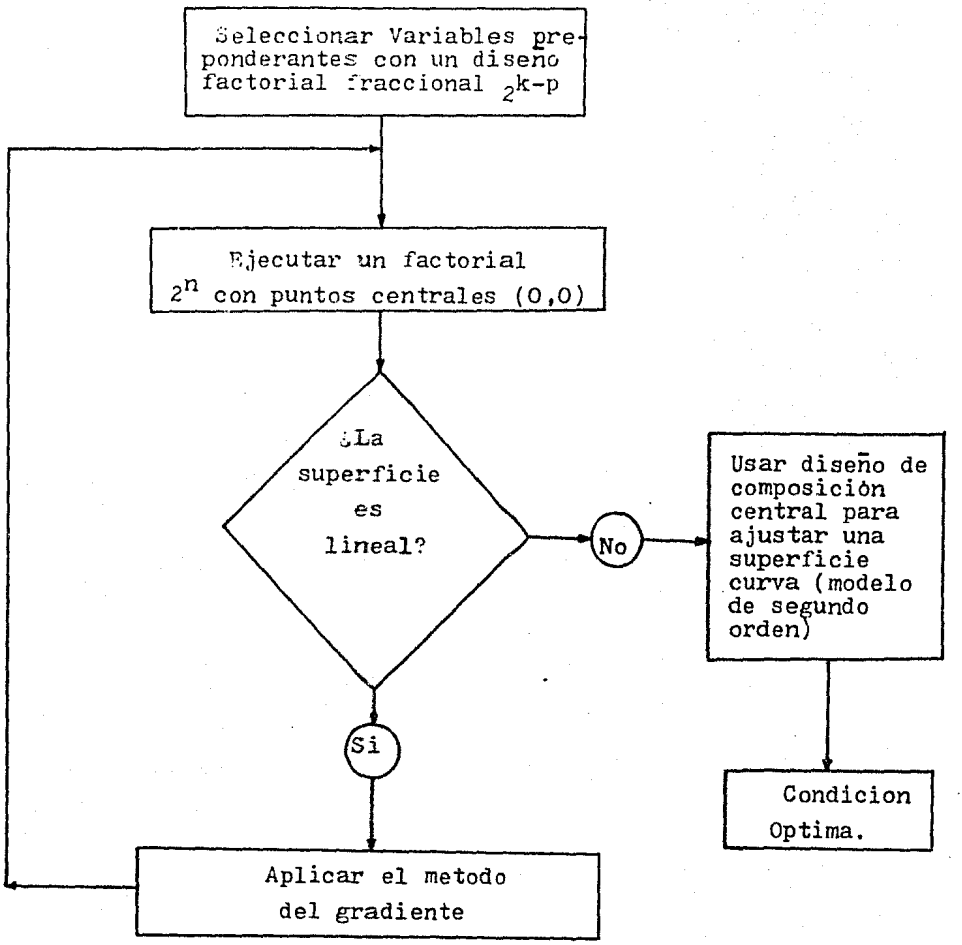
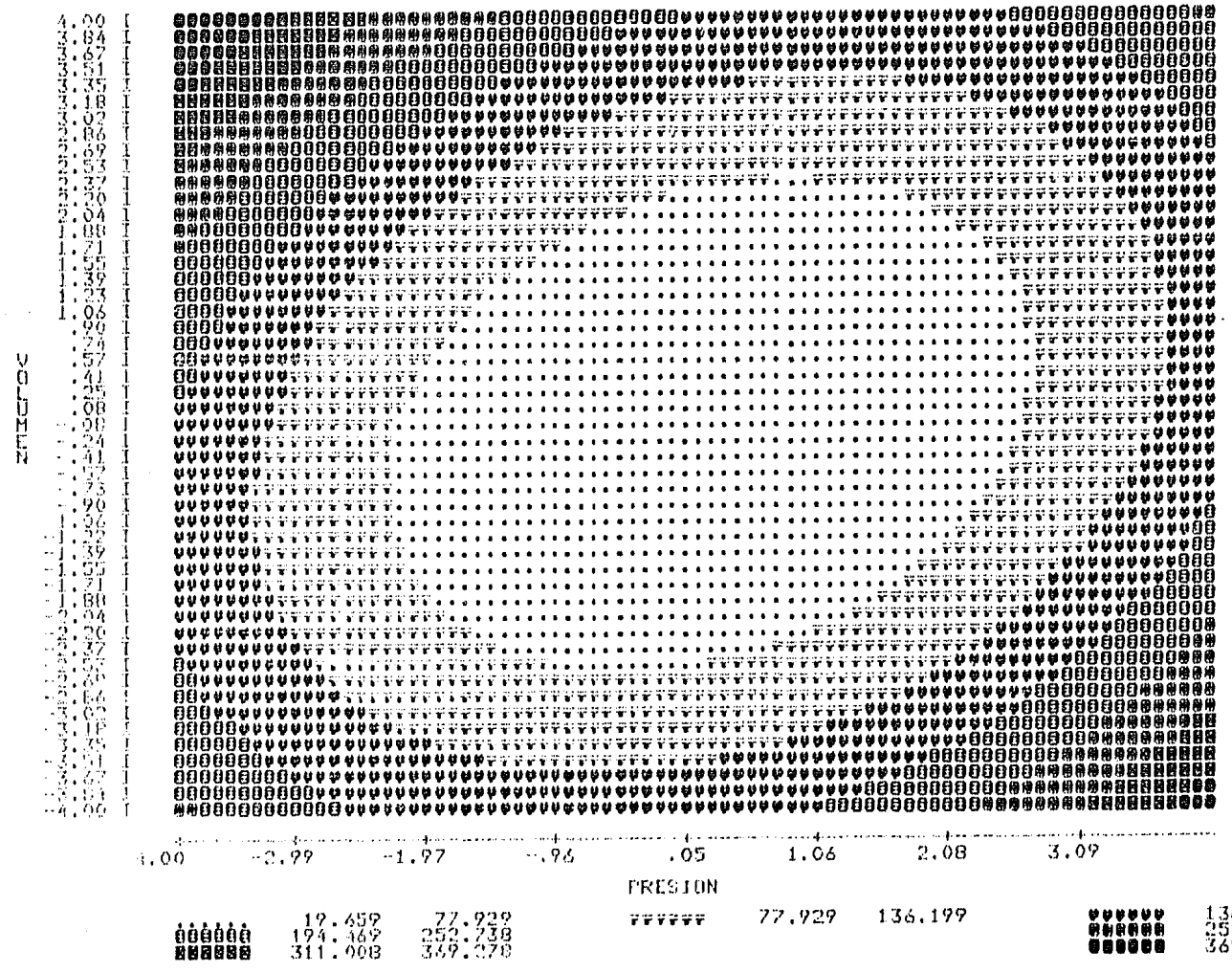

Seleccionar Variables preponderantes con un diseño factorial fraccional 2^{k-p}
Ejecutar un factorial 2^n con puntos centrales (0,0)
¿La superficie es lineal?
No
Usar diseño de composición central para ajustar una superficie curva (modelo de segundo orden)
Si
Condicion Optima.
Aplicar el metodo del gradiente

Fig. 7.37

Como puede observarse, se incluyen todas las herramientas vistas y analizadas durante este capitulo. Es en la etapa final, al usar los diseños de composicion central, cuando el programa graficador de superficies de respuesta es ampliamente usado, con el objeto de visualizar la direccion del gradiente. Hay que recordar, que el diagrama solo es bidimensional, y es una representación de un fenomeno que se graficaria en 3 dimensiones, siendo el tercer eje, el de la respuesta o rendimiento en estudio.

La siguiente sección consta de algunos ejemplos de gráficas obtenidas con el programa, y la descripción global de este, se encuentra en el Apéndice D.

7.2 EJEMPLOS



PRL

VALOR MAXIMO DE Y= 427.54800 VALOR MINIMO DE Y= 19.6591 STOP

El programa desarrollado posee un alto grado de flexibilidad, ya que maneja los renglones y columnas que se especifiquen, asi como el número de contornos: la primera de las siguientes hojas muestra una grafica de 12 renglones por 25 columnas, y 9 contornos; para ella los datos se codifican de la siguiente manera:

Registro 1:

Col. 1	8	15	22	29
133.61	0.0	0.0	12.25	-3.09

Col. 36

0.0

0 sea, la ecuación es

$$Y = 133.61 + 12.25*X1**2 - 3.09*X2**2$$

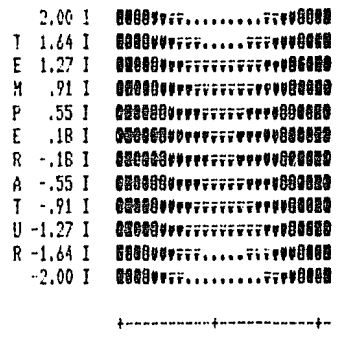
En el mismo registro 1 se leen los nombres de las variables indepen dientes, 11 para el eje vertical y 10 caracteres para el eje horizontal; para la variable dependiente se tienen 10, pero este dato no se usa.

Para el registro 2, la cuerda de entrada es:

9012025 -2.0 2.0 -2.0 2.0 1

que especifica 9 contornos, 12 renglones, 25 columnas, nivel bajo de X1 de -2, y alto de 2, y nivel bajo de X2 de -2, con valor de nivel alto de 2.

La siguiente gráfica muestra como se incrementa el número de renglones, los datos son los mismos, solo N1 cambia, de 12

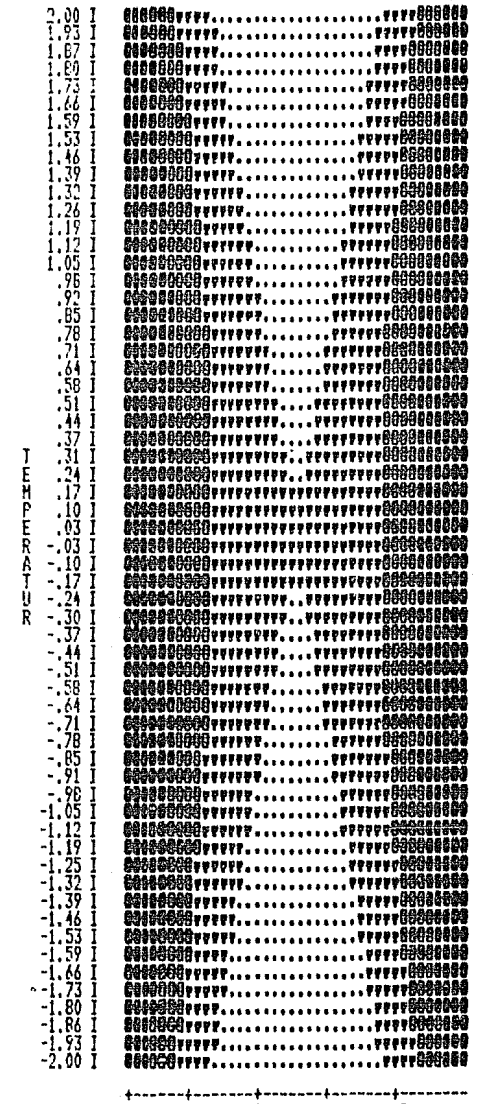


.....	121.238	128.045	▒▒▒▒▒▒	128.045	134.853	▒▒▒▒▒▒	134.853	141.661
▒▒▒▒▒▒	141.661	148.469	▒▒▒▒▒▒	148.469	155.277	▒▒▒▒▒▒	155.277	162.085
▒▒▒▒▒▒	162.085	168.893	▒▒▒▒▒▒	168.893	175.701	▒▒▒▒▒▒	175.701	182.508

RRL

VALOR MAXIMO DE Y= 182.50830 VALOR MINIMO DE Y= 121.2376 STOP

E.1



.....	121.270	133.537	vvvvvv	133.537	145.805	@@@@@@	145.805	158.072
######	158.072	170.339				@@@@@@	170.339	182.607

RRL

VALOR MAXIMO DE Y= 182.60650 VALOR MINIMO DE Y= 121.2699 STOP

a 60, y el número de contornos es 5.

Se debe entonces, al analizar una gráfica, tomar en cuenta la distorsión ocasionada por el aumento o disminución de las columnas o los renglones; esto se ejemplifica con el siguiente ejemplo:

Ecuación: - 12X1**2 - 12X2**2

Cuerda de especificaciones:

50017038 -2.0 2.0 -2.0 2.0 1

que especifica 5 contornos, 17 renglones y 38 columnas; la gráfica resultante es la E.3.

Este caso es el más correcto, ya que la gráfica es cuadrada, y da la apariencia verdadera de la superficie. Si la gráfica se alargara, se obtendría una elipse, cosa que no es verdad. - (Gráfica E.4).

También se tiene que tomar en cuenta el tipo de - impresión utilizado; las gráficas anteriores usan letra chica, mientras que la E.5 utilizó grande; además, se utilizo un separamiento mayor de los renglones (6 por pulgada) en lugar de los 8 por pulgada que se venian usando. La gráfica E.6 muestra el efecto de usar - uno u otro espaciamiento.

Por lo que respecta a los límites, el mayor numero de columnas es 110, y el mayor número de renglones es 60 (aunque puede ser mayor si se hacen variaciones dentro del sistema operativo). Una gráfica de ese tamaño, con letra pequeña es la E.7.

En lo que concierne al número de renglones, el mínimo es 3, y en las columnas es 20. La gráfica E.8 es un ejemplo de una de 3 renglones.

Notesé, que para hacer la comparación, se ha usado siempre la misma función.

Para obtener un cuadrado exacto, se deberan especificar el número de columnas doble al numero de renglones, en letra - chica, y 1.5 veces en letra grande. La gráfica E.9 tiene 40 renglones y 80 columnas.

Las siguientes gráficas, corresponden a las ecuaciones siguientes:

		Contornos
E.10	-60 X1 **2	7
E.11	20 - 9*X1 **2 + 5*X2 **2	7
E.12	20 + 4*X1 **2 + 5*X2 **2 -5 * * X1 * X2 (interacción)	4

la cuerda de entrada para esta seria:

4035082 10.0014.88 3.00 7.00 1

E.13	20 - 12*X1 - 5*X1**2 + X2	7

para esta curva, el primer registro seria:

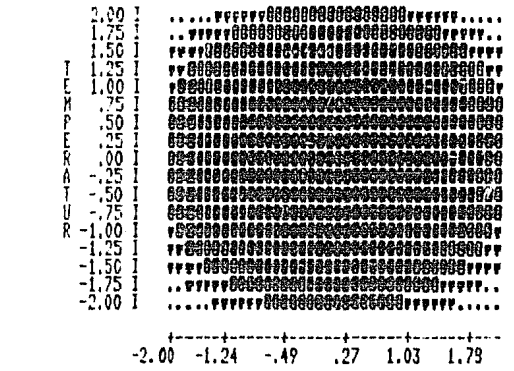
20.0 -12.0 1.0 -5.0 0.0 0.0

E.14	20 + 10*X1 **2 + 10*X2 **2 - 5(X1*X2) - 3*X1 + 3*X2	7

los registros serian:

20.0 -3.0 3.0 10 10 -5.0

7035082 -4.0 4.0 -4.0 4.0 1 VOLUMEN PRESION RENDIMIENT
I———I———I———
10 espac.

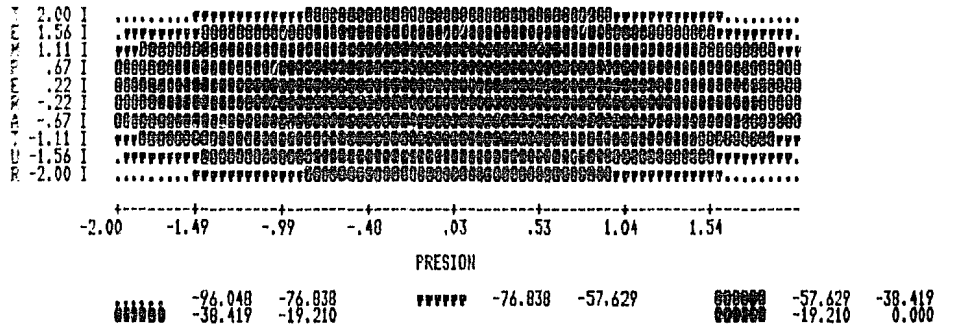


-96.048 -76.838 -76.838 -57.629 -57.629 -38.419
-38.419 -19.210 -19.210 0.000

RRL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -96.0480 STOP

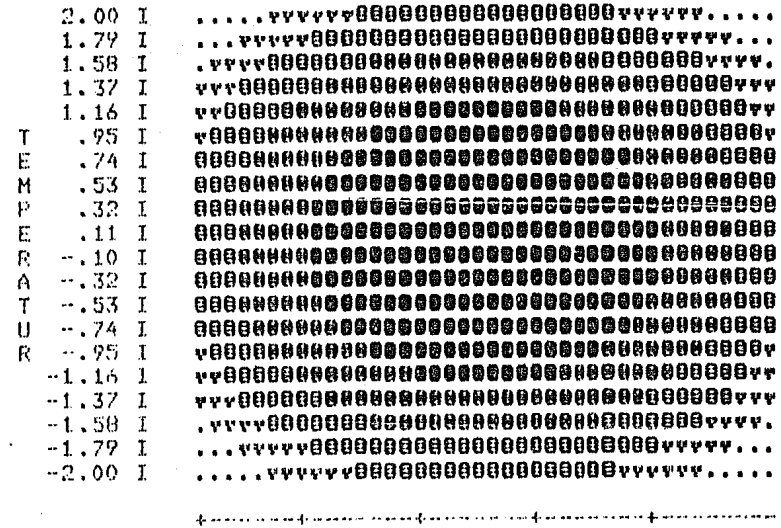
E.3



RRL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -96.0480 STOP

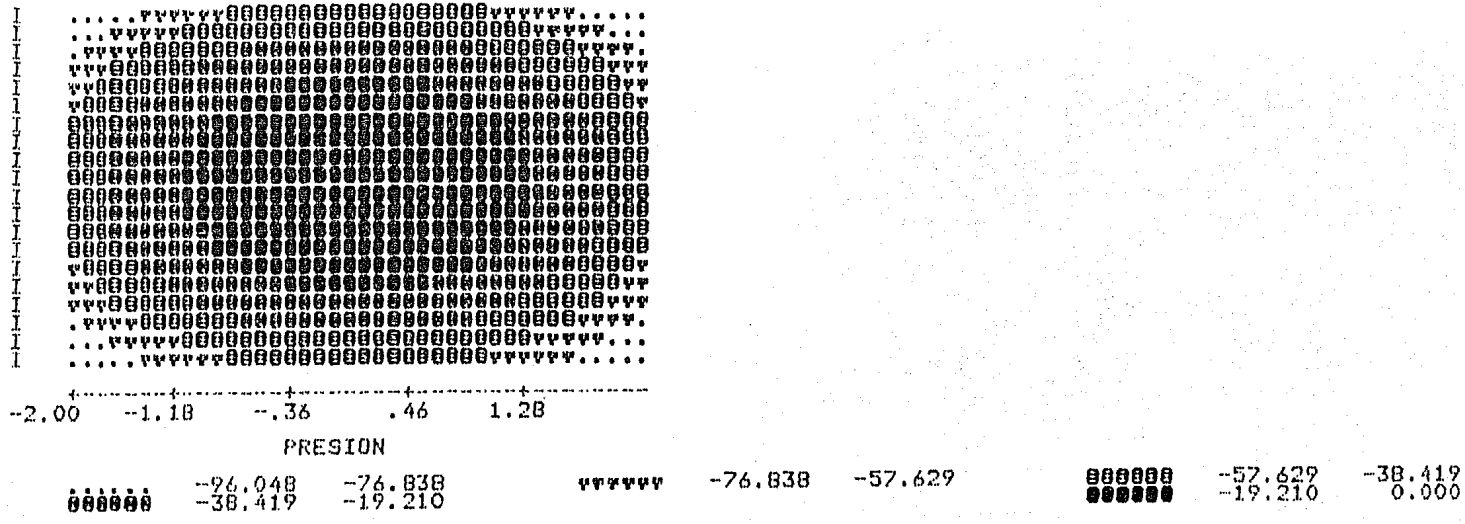
E.4



.....	-96.048	-76.838	vvvvvv	-76.838	-57.629	@@@@@@	-57.629	-38.419
@@@@@@	-38.419	-19.210				######	-19.210	0.000

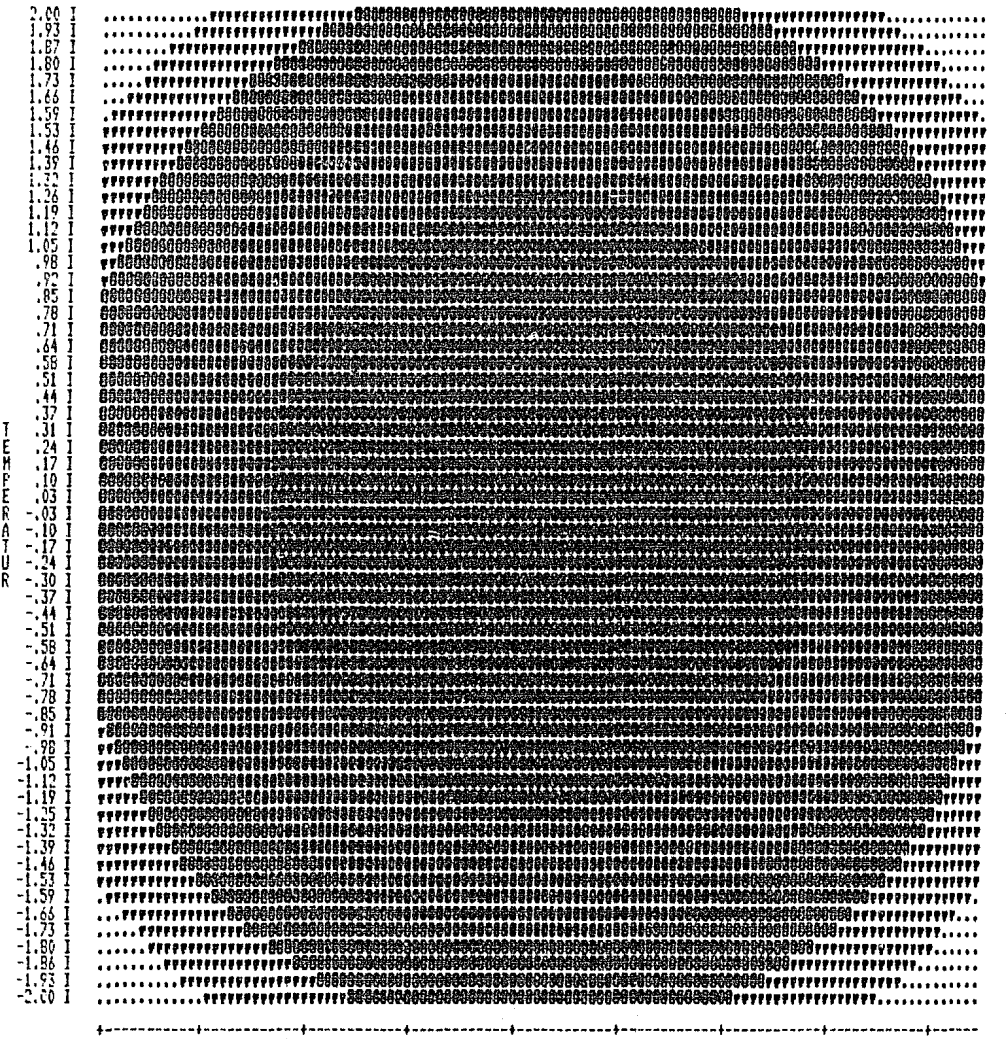
RRL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -96.0480 STOP



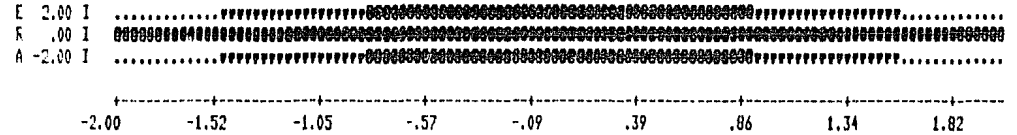
RRL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -96.0480 STOP



RRL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -96.0480 STOP

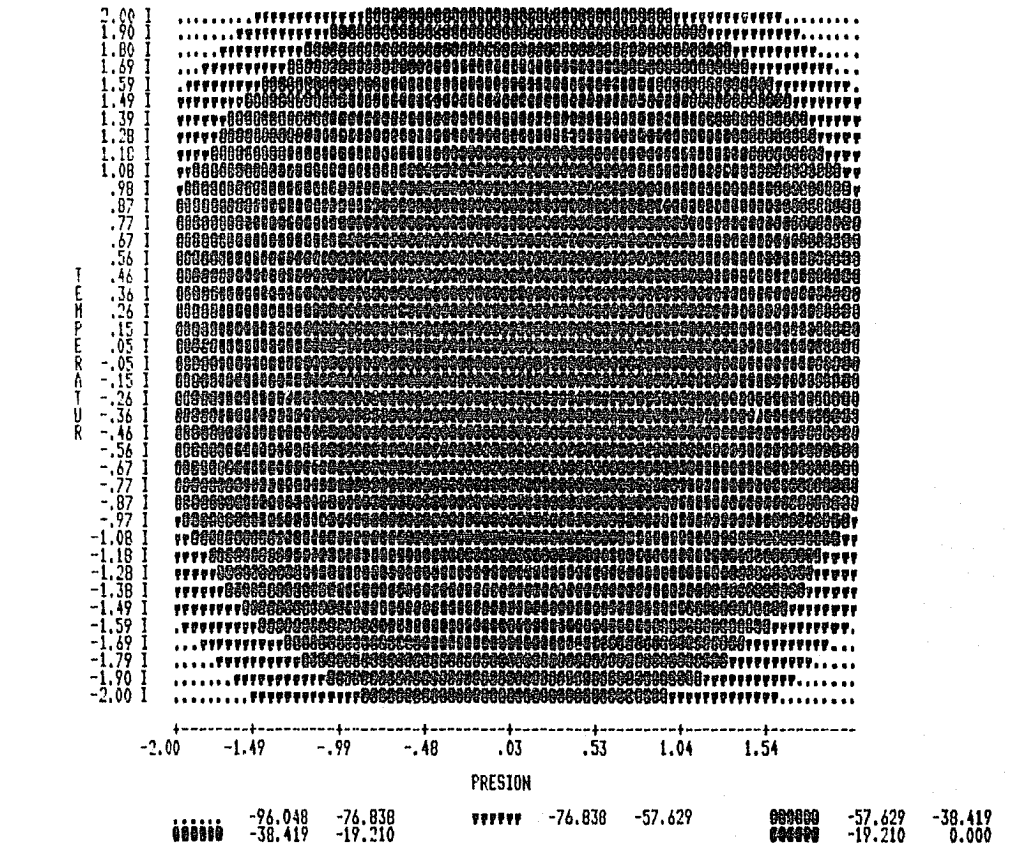


.....	-96.048	-76.838	▮▮▮▮▮▮	-76.838	-57.629	▮▮▮▮▮▮	-57.629	-38.419
▮▮▮▮▮▮	-38.419	-19.210				▮▮▮▮▮▮	-19.210	0.000

RRL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -96.0480 STOP

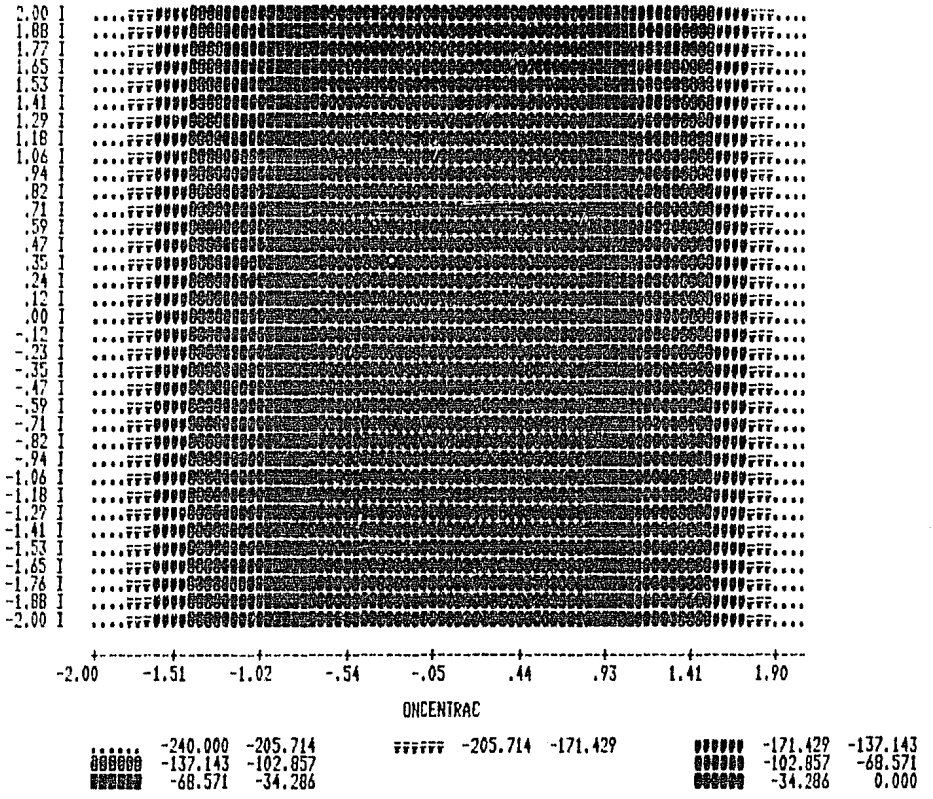
E.8



RRL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -96.0480 STOP

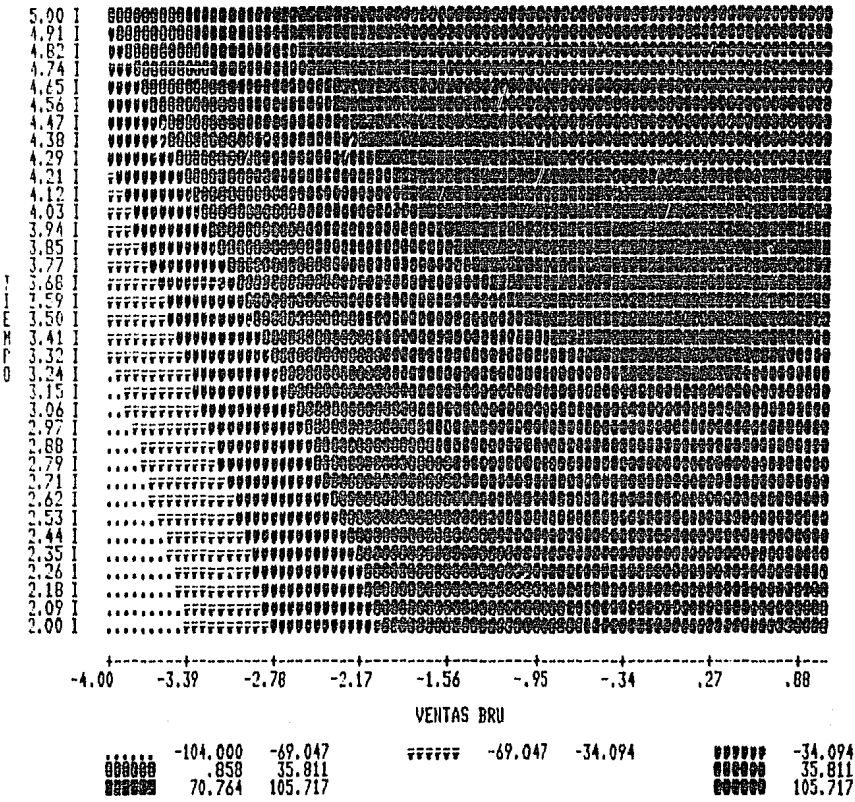
E.9



RPL

VALOR MAXIMO DE Y= .00000 VALOR MINIMO DE Y= -240.0000 STOP

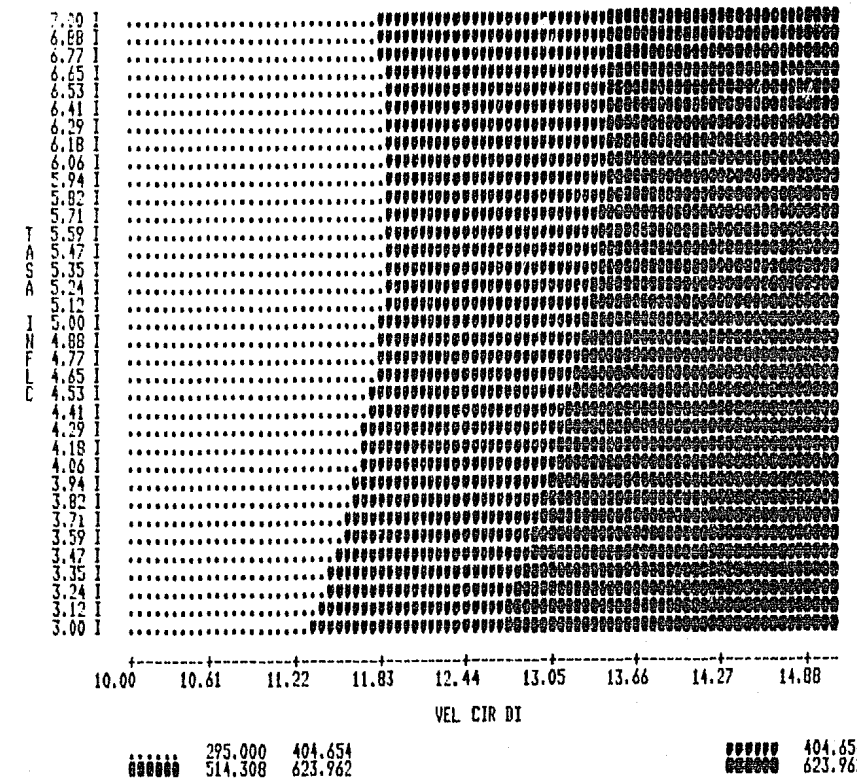
E.10



ERL

VALOR MAXIMO DE Y= 140.66960 VALOR MINIMO DE Y= -104.0000 STOP

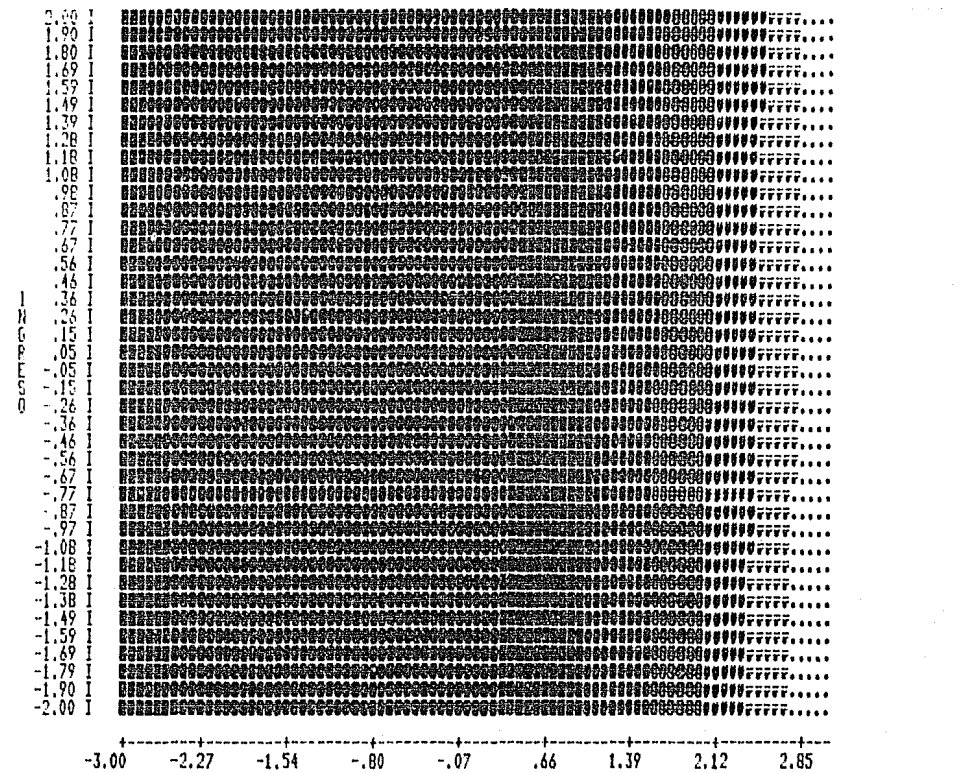
E.11



RRL

VALOR MAXIMO DE Y= 733.61600 VALOR MINIMO DE Y= 295.0000 STOP

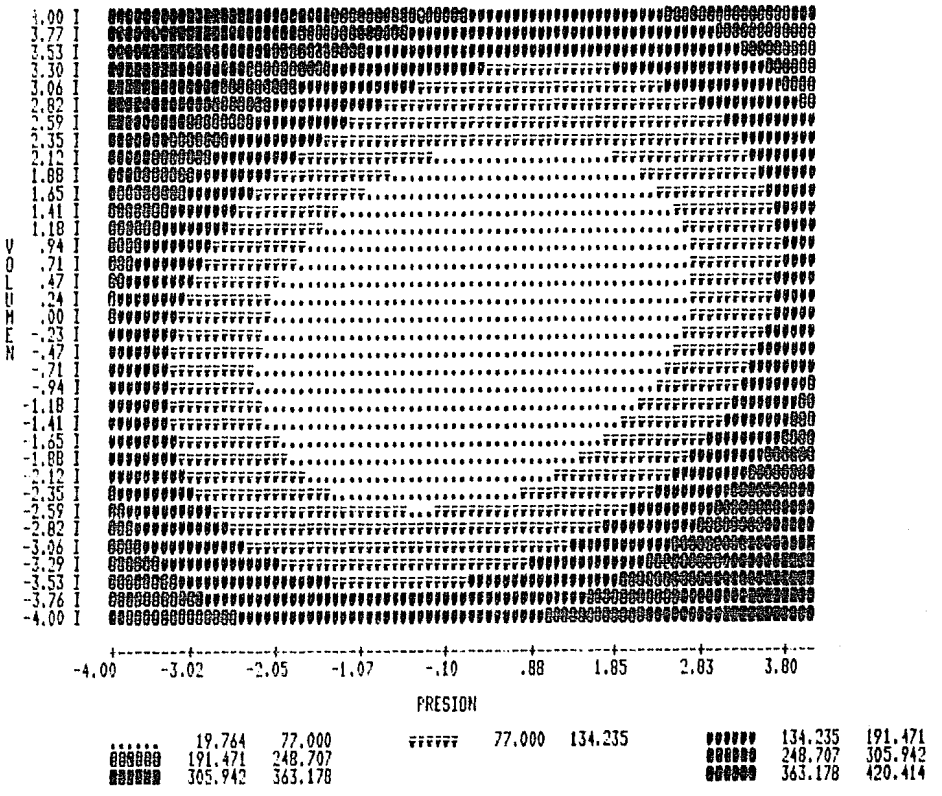
F.12



RRL

VALOR MAXIMO DE Y= 29.19670 VALOR MINIMO DE Y= -63.0000 STOP

E.13



RRL

VALOR MAXIMO DE Y= 420.41380 VALOR MINIMO DE Y= 19.7640 STOP

E.14

8

CONCLUSIONES

A lo largo del trabajo, se han expuesto primordialmente los conceptos y la interpretación de la metodología de superficie de respuesta, con el objeto de que los conocimientos puedan ser llevados a la práctica, en un tiempo corto y sin menoscabo de la eficiencia de la propia metodología. Su aplicación correcta se traducirá en un fuerte ahorro de los recursos asignados para la optimización del proceso de interés.

Por lo que respecta al programa, y como se ha constatado, este es de fácil manejo, y además flexible, ya que su tamaño y el número de contornos puede ser variado, cumpliendo adicionalmente con las especificaciones que poseen actualmente los programas de este tipo. Es de resaltar, que el programa ocupa menos de 20 KB de memoria para su ejecución, en una microcomputadora con longitud de palabra fija de 8 bits, por lo que es más eficiente que cualquiera de los rentados en el país.

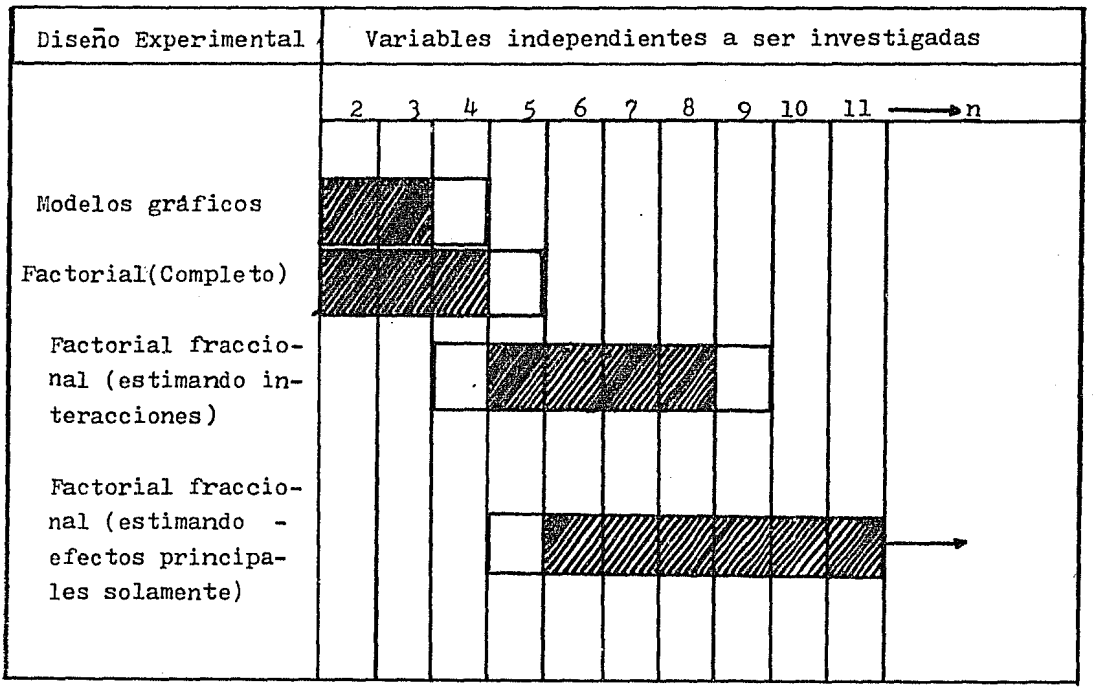
Asimismo, la comprensión y el uso de los modelos lineales, es indispensable para poder llevar a cabo investigación en campos como la termodinámica, en donde ellos son utilizados como introducción a los modelos no lineales, los cuales son a su vez el antecedente para poder formular modelos teóricos. De esta forma, este trabajo puede ser utilizado en esa etapa introductoria, para aquellas personas que se interesen en el campo de la Investigación de Operaciones.

9

PROPUESTAS Y RECOMENDACIONES

Comúnmente, al iniciar un programa de investiga- - ción, se tienen dudas sobre el tipo de diseño experimental a usar, el cual depende fuertemente de las metas a alcanzar en dicho pro- - grama experimental, y del número de factores involucrados.

En estos casos, se propone aplicar el esquema siguiente, el cual ayuda a tener una idea más clara de como manejar - el programa experimental. Dicho esquema, contempla implicitamente - el costo y tiempo de ejecución de experimentos, así como el mínimo deseable de información a obtener para el número especifico de factores involucrados.



Por otra parte, si se desea aumentar la precisión del plan experimental, debido a que la magnitud de los errores ε_i -

es muy alta, existen tanto diseños más sofisticados, como métodos alternativos de análisis; entre ellos se encuentran:

1.- Homogeneización de técnica experimental[13].

2.- Agrupación planeada (bloques)[30].

3.- Mediciones adicionales[13].

4.- Uso de experimentos repetidos y pruebas de falta de ajuste[19].

5.- Transformaciones no lineales sobre la variable dependiente[2,14].

Por último, se hace necesario resaltar, que el experimentador no debe perder de vista el sentido lógico de los procesos que estudia. Es frecuente observar casos, en los que para obtener una conclusión, se hicierón gran cantidad de análisis complejos de los resultados, siendo que con solo entender y razonar las implicaciones lógicas de ellos, se podría haber llegado a la misma conclusión. Es indispensable, por tanto, evitar envolverse en una maraña de números y pasos a ejecutar, y no perder de vista, nunca, el sentido común.

APENDICE A.- Probabilidad

La observación se considera una parte importante de un experimento. Los resultados verdaderos de una observación - son llamados las consecuencias o desenlaces del respectivo experimento, y a la totalidad de las posibles consecuencias de un experimento se le llama espacio muestral y se denota por S.

En problemas que incluyen fenómenos aleatorios, - es conveniente representar los desenlaces de un experimento como - puntos en un espacio de una o más dimensiones; esos puntos, constituyen los elementos del espacio muestral. Así por ejemplo, si - un experimento consiste en examinar una junta soldada, los resultados pueden ser: intacta (0) o rota (1), pudiendose representar - en un espacio muestral unidimensional (Fig A.1a). Si consiste en - examinar 2 juntas, habrá 4 posibles consecuencias como se muestra en el espacio bidimensional de la Fig. A.2. En general, en un circuito con n-juntas soldadas, hay 2^n posibles desenlaces que representan 2^n puntos en un espacio n-dimensional.

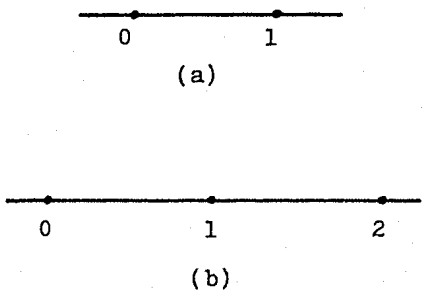


Fig. A.1

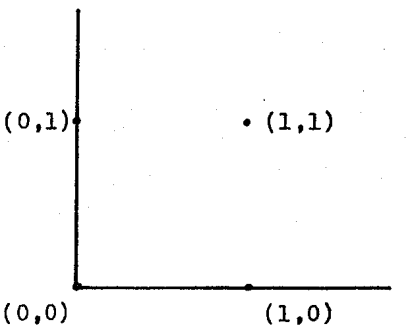


Fig. A.2

Esta configuración geométrica no es única. Así por ejemplo, los desenlaces del experimento anterior se pueden representar en un espacio unidimensional como 3 puntos (0,1,2). Aquí, el punto 1 de la Fig. A.1b representaría a los puntos (1,0) y (0,1) de la Fig. A.2; sin embargo, es altamente deseable, que los espacios muestrales tengan elementos que no puedan ser adicionalmente divididos, es decir, que no representen 2 o más desenlaces que no sean distinguibles en alguna forma.

Los espacios muestrales pueden clasificarse de acuerdo al número que ellos contienen. Entonces, puede haber espacios finitos, e infinitos. Un espacio muestral infinito es el conjunto de números reales.

Adicionalmente, un espacio infinito puede ser contable o no; si se mide la resistencia de una junta soldada en Ohms, el espacio muestral podrá consistir de puntos sobre una escala continua (un intervalo sobre la recta de los números reales). Los elementos de este espacio no pueden ser contados, esto es, no puede haber correspondencia uno a uno con el espacio de los números naturales (que es contable). Este tipo de espacio es llamado continuo, mientras que los otros son llamados discretos.

Los espacios muestrales pueden subdividirse, y a los subconjuntos resultantes se les denomina eventos. Si 2 eventos son independientes entre sí, es decir, no tienen elementos comunes, ellos son mutuamente exclusivos.

Entonces si E es un evento, $E \subseteq S$ (E es subconjunto de S). En un diagrama de Venn, la relación se representa:

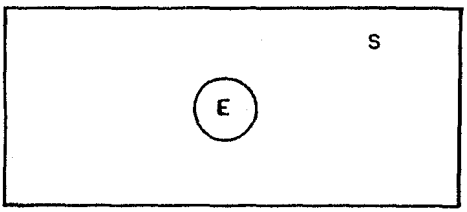


Fig. A.3

esto significa que $E \cup S = S$.

En la anterior fórmula, implícitamente hacemos uso de una función conjunto: la función conjunto aditiva. En esta, los elementos del dominio son conjuntos, y los elementos del rango son números reales. Esta función, asigna un número real a cada subconjunto A de un conjunto dado. La forma o valor del número asignado depende de la definición particular. El caso más sencillo es aquel en que el número asignado es igual al número de elementos contenidos dentro del subconjunto A. Por ejemplo, se puede suponer que una escuela tiene 50 profesores, que son clasificados de acuerdo a si ellos son casados (M) o no (M'), y a si sin graduados o no (G) y (G'), para de esa forma, estimar su nivel académico.

La distribución de esos 50 profesores se muestra esquematicamente en el diagrama de Venn de la Fig. A.4; con él, el valor de N(A) puede ser determinado para cualquiera de los 16 subconjuntos posibles en que pueden clasificarse los profesores, donde N(A) es el número de elementos contenidos en A. Los números dentro de la Fig. A.4 son: el número de profesores casados no graduados, el número de profesores casados graduados, el número de profesores solteros graduados y el número de solteros que no son graduados, esto es:

$$N(M \cap G') = 20$$
$$N(M \cap G) = 10$$
$$N(M' \cap G) = 5$$
$$N(M' \cap G') = 15$$

Para hallar el número de profesores que son casados, se debe añadir el número de profesores casados graduados al número de profesores casados que no son graduados, y se obtendría:

$$N(M) = N(M \cap G) + N(M \cap G') = 10 + 20 = 30$$

igualmente, se puede hallar el número de profesores graduados:

$$N(G) = N(M \cap G) + N(M' \cap G) = 10 + 5 = 15$$

y dado que $N(S) = 50$, por substracción:

$$N(M') = N(S) - N(M) = 50 - 30 = 20$$
$$N(G') = N(S) - N(G) = 50 - 15 = 35$$

La función conjunto aditiva asigna a la unión de dos conjuntos que no tienen elementos en común, un número que es igual a la suma de los números asignados a los conjuntos individuales. Conjuntos que no tienen elementos en común son llamados disjuntos, y como se apuntó antes, eventos que corresponden a conjuntos disjuntos son llamados eventos mutuamente exclusivos. Cuando algún par de conjuntos A y B tienen elementos en común, se aplica la fórmula general:

$$N(A \cup B) = N(A) + N(B) - N(A \cap B) \qquad \text{Ec. (A.1)}$$

que en el ejemplo anterior da:

$$N(M \cup G) = N(M) + N(G) - N(M \cap G) = 30 + 15 - 10 = 35$$

para el número de profesores casados, o graduados, o ambos. Nótese que se sustrajo el número de profesores casados graduados debido a que ellos fuerón contados 2 veces, una cuando se contó a los casados y otra cuando se contó a los graduados.

Usando el concepto de función conjunto aditivo, se puede ahora definir la probabilidad de un evento. Dado un espacio muestral S y un evento A en S, se define P(A), la probabilidad de A, como el valor de una función conjunto aditiva P, llamada la función de probabilidad. Para que una función conjunto sea función de probabilidad, deberá satisfacer las siguientes condiciones:

Axioma 1.- $0 \leq P(A) \leq 1$ para cada evento A en S.

Axioma 2.- $P(S) = 1$

Axioma 3.- Si A y B son eventos mutuamente exclusivos, entonces $P(A \cup B) = P(A) + P(B)$.

Concluyendo:

El primer axioma establece que la función probabilidad asigna a cada evento A en S un número real de 0 a 1 (comunmente a P(A) se le identifica como el número deelementos de A dividido entre el número de elementos de S). El segundo axioma establece que al espacio muestral se le asigna el número 1 y expresa la idea de que la probabilidad de un evento cierto, un evento que siempre sucede, es igual a 1, y el tercero establece que la función probabilidad debe ser aditiva, y con el uso de inducción ma-

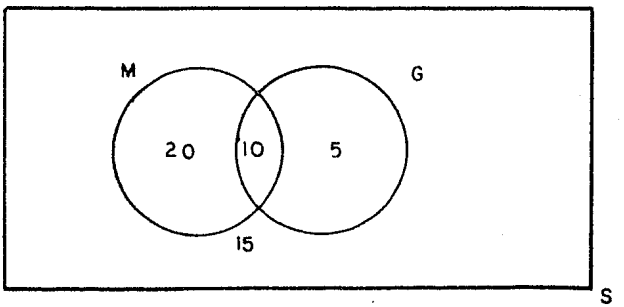


Fig. A.4

temática, puede extenderse a cualquier número finito de eventos - mutuamente exclusivos:

$$P(A_1 \cup A_2 \cup \ldots\ldots \cup A_n) = P(A_1) + P(A_2) + \ldots\ldots + P(A_n)$$

donde $A_1, A_2, \ldots\ldots, A_n$ son mutuamente exclusivos dentro de S.

APENDICE B.- Distribución Normal

Una de las funciones de probabilidad más importantes, es la correspondiente a la de la distribución normal, la cual es:

$$F(x) = \frac{1}{b(2)^{\frac{1}{2}}} \exp\left\{ - \frac{(x - a)^2}{2b^2} \right\} \quad -\infty \leq x \leq \infty \quad (B.1)$$

donde

a es la media de la población que se define como $a = E(x)$.

b^2 es la varianza de la población.

$E(x)$ es la esperanza matemática de la variable aleatoria x, y se calcula como:

$$E(x) \int_{-\infty}^{\infty} xf(x)dx$$

donde

$f(x)$ = función de distribución de probabilidad.

Conceptualmente, $E(x)$ es el valor esperado de x, cuando x es una variable aleatoria. Observese que $E(x)$ en poblaciones normalmente distribuidas es μ, la media verdadera de la población.

Para la función B.1, se cumple:

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

entonces, la probabilidad para una x dada (o rango de valores de x) esta dada por el área bajo la curva definida por la Ec. b.1.

Para ilustrar esto, considere el siguiente ejemplo hipotético:

Existen 6 particulas, numeradas del 1 al 6, en 2 - elementos de volumen A y B. En la tabla B.1 se identifican las 64 - distribuciones discernibles que resultan para tal sistema. Este número puede obtenerse de

$$\Omega = M^N \qquad (B.2)$$

donde

Ω = número de disposiciones discernibles

M = número de celdas discernibles

N = número de particulas discernibles

La Ec. B.2 señala la dependencia del numero de - particulas (N) sobre el número de distribuciones discernibles que - se pueden realizar.

En mecanica estadistica, se le llama microestado a una distribución individual discernible. Para el ejmplo actual, - existen 64 microestados. Se usa el término macroestado para designar un grupo de microestados con caracteristicas comunes. Así, un - macroestado puede estar constituido por todos los microestados con el mismo número de particulas en el compartimiento A y el mismo número en el compartimiento B. De esta forma, se pueden definir los -

Distribución	Compartimiento A	Compartimiento B
0-6		123456
1-5	1	23456
1-5	2	13456
1-5	3	12456
1-5	4	12356
1-5	5	12346
1-5	6	12345
2-4	12	3456
2-4	13	2456
2-4	14	2356
2-4	15	2346
2-4	16	2345
2-4	23	1456
2-4	24	1356
2-4	25	1346
2-4	26	1345
2-4	34	1256
2-4	35	1246
2-4	36	1245
2-4	45	1236
2-4	46	1235
2-4	56	1234
3-3	123	456
3-3	124	356
3-3	125	346
3-3	126	345
3-3	134	256
3-3	135	246
3-3	136	245
3-3	145	236
3-3	146	235
3-3	156	234
3-3	234	156
3-3	235	146
3-3	236	145
3-3	245	136

Continua

Distribución	Compartimiento A	Compartimiento B
3-3	246	135
3-3	256	134
3-3	345	126
3-3	346	125
3-3	356	124
3-3	456	123
4-2		
5-1	Estas son imagenes (reflejadas) de las	
0-6	tres primeras	

Tabla B.1

siguientes 7 macroestados:

Macroestado NA - NB	Número de Microestados
3 - 3	20
4 - 2	15
2 - 4	15
5 - 1	6
1 - 5	6
6 - 0	1
0 - 6	1
	64

NA = Número de partículas en A

NB = Número de partículas en B

Debe hacerse notar, que las disposiciones o arreglos de las mismas partículas dentro de un volumen dado no constituyen microestados adicionales. Así, la dispocisión 234561 en el microestado 0 - 6 es idéntica a la disposición 123456.

El número de combinaciones para colocar N partículas en 2 celdas con M partículas en una y N - M en la segunda es dado por*:

$$C_N^{M,N-M} = \frac{N!}{(N-M)!\,M!}$$

así, para N 6 y M 5

$$C_6^{5,1} = \frac{6!}{1! * 5!} = 6$$

---

* Recordar que $0! = 1$

que efectivamente, es el número de microestados para esa disposición.

Si cada una de las seis partículas tiene igual probabilidad de ir al compartimiento A que al B, cualquiera de los 64 microestados tendrá igual probabilidad de ocurrir. Sin embargo, se puede observar que la probabilidad de encontrar el macroestado 6 - 0 es considerablemente menor que la correspondiente al 3 - 3, ya que aquel ocurre solo en 1 de las 64 posibilidades. Su probabilidad es, pues, de 1/64, comparada con una probabilidad de 20/64 para la configuración 3 - 3. En estas condiciones la probabilidad de encontrar cualquier macroestado dado esta relacionada directamente con el número de microestados que contiene:

Macroestado		Número de	Fracción del tiempo
A	B	microestados	en el macorestado
6	0	1	1/64
5	1	6	3/32
4	2	15	15/64
3	3	20	5/16
2	4	15	15/64
1	5	6	3/32
0	6	1	1/64

Tabla B.2

Puede observarse que la probabilidad obtenida en la tabla B.2, es aproximadamente el área bajo la curva de la Fig. B.1, si se considera como 1 toda el área bajo la curva.

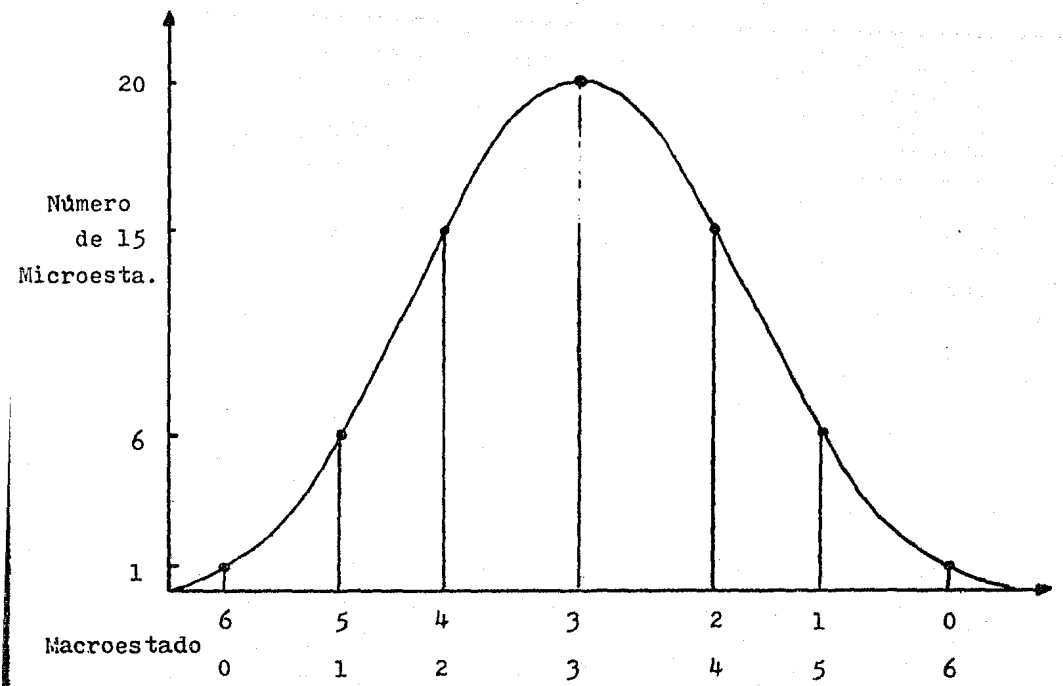


Fig. B.1

Si el número de particulas aumenta a 20, existen:

$$2^{20} = 1048576 \text{ microestados}$$

y un total de 21 macroestados. En este caso, aún cuando el macroestado más probable (10 - 10) ocurre con una probabilidad de solo .176, los 3 macroestados más probables tienen una probabilidad combinada de 0.496. Asi, un sistema tal pasaria aproximadamente una mitad de su tiempo en esos macroestados.

Al aumentar el número de particulas, a un orden análogo al del número de Avogadro, se encuentra que la probabilidad de obtener una disposición distribuida "más o menos" uniforme, se hace abrumadora.

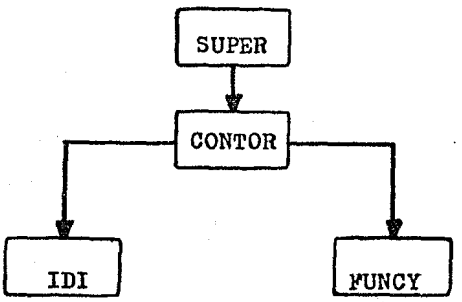
Aunque están ocurriendo continuamente leves desviaciones desde la distribución más probable, esas variaciones son tan pequeñas que no se tienen instrumentos suficientemente sensibles para medirlas; en otras palabras, la probabilidad de cualquier desviación medibles desde una distribución esencialmente uniforme es tan pequeña que puede ser despreciada sin error. Así por ejemplo, la probabilidad de observar una variación de un 0.001% en la densidad de 1 cm^3 de aire es menor que 10^{-10^8} y no es probable que sea observada en billones de años. Se encuentra así, que la concentración espontanea de todas las moléculas gaseosas en una parte del recinto no es totalmente imposible, sino que es tan improbable de que acontezca, que su probabilidad es realmente despreciable.

Apéndice C

Y	X_1	X_2
10.98	35.3	20
11.13	29.7	20
12.51	30.8	23
8.40	58.8	20
9.27	61.4	21
8.73	71.3	22
6.36	74.4	11
8.50	76.7	23
7.82	70.7	21
9.14	57.5	20
8.24	46.4	20
12.19	28.9	21
11.88	28.1	21
9.57	39.1	19
10.94	46.8	23
9.58	48.5	20
10.09	59.3	22
8.11	70.0	22
6.83	70.0	11
8.88	74.5	23
7.68	72.1	20
8.47	58.1	21
8.86	44.6	20
10.36	33.4	20
11.08	28.6	22

APENDICE D.- Estructura Global del Programa

El programa consta de 4 módulos, cada uno de los - cuales tiene una finalidad perfectamente definida; la relación entre estos módulos, puede representarse asi:



El módulo llamado SUPER, constituye el programa - principal. Su función es la de introducir los datos, por medio de 2 lecturas. En la primera, se leen los coeficientes del modelo lineal

$$Y = b_o + b_1X_1 + b_2X_2 + b_{12}X_1X_2 +$$

$$+ b_{11}X_1^2 + b_{22}X_2^2$$

e igualmente, los nombres de las variables independientes X_1 y X_2 - involucradas, y el correspondiente nombre de la respuesta Y.

En la segunda lectura, se leen los siguientes parámetros:

N	Número de contornos que se desea incluya la gráfica:

$$2 \leq N \leq 9$$

N es entero.

N1 — Número de renglones a imprimir, en el eje vertical. (X_2 es el correspondiente al eje vertical) N1 es entero y

$$12 \leq N1 \leq Z$$

$$Z \begin{cases} 60 \\ \\ 120 \end{cases} \text{ Según el S.O.}$$

N2 — Número de columnas a imprimir en el eje horizontal (X_1 es la variable correspondiente). N2 es entero y:

$$15 \leq N2 \leq 115$$

X11 — Valor del nivel bajo de X_1, que corresponderá al menor valor a graficar en el eje horizontal; X_1 deberá cumplir con:

$$-9999 \leq X_1 \leq 9999$$

Si la variable supera estos límites, se aconseja modificarla*.

X12 — Valor del nivel alto de X_1, que corresponderá al mayor valor a graficar en el eje vertical.
Necesariamente X12 > X11

X21 Valor del nivel bajo de X_2, que corresponderá al menor valor a graficar en el eje vertical; X_2 debera cumplir con:

$$-999 \leq X_2 \leq 999$$

Si la variable supera estos limites, se aconseja modificarla*

X22 Valor del nivel alto de X_2, que corresponderá al mayor valor a graficar en el eje vertical.
Necesariamente X22 > X21

IPAR Código de opción:

0 No se desea la gráfica, sino solo - obtener el máximo y el mínimo de Y

1 Se desea la grafica y el valor maximo y minimo de Y

Después de esto, el control se transfiere a CONTOR. En esta subrutina, se definen los caracteres de impresion compuestos, y que consisten de los siguientes caracteres elementales:

.	punto
-	menos
=	igual
+	más
0	letra 0
X	letra X
*	asterisco

@	arrova
#	número

En base a ellos, y al número de contornos, N, especificado, se forma la combinación del carácter compuesto a imprimir; por ejemplo, el caracter θ es la combinación de = y la letra O.

Para cada par ordenado (X_1,X_2) es necesario asignar un caracter compuesto, el cual corresponde al intervalo de valores - dentro del cual, la respuesta Y en ese punto, se encuentra ubicada. - Para lograr esto, la subrutina CONTOR hace referencia a la funcion - entera IDI, la cual especifica el caracter (o caracteres) a imprimirse en ese punto, información que puede obtener gracias a la funcion - real FUNCY, la cual evalúa la función en ese punto.

La subrutina CONTOR, también imprime los ejes respectivos, y asimismo, al pie de la gráfica, el código compuesto de impresión, y el rango de valores que abarca. Por último, el control regresa al programa original SUPER, que imprime el valor mínimo y - máximo de Y encontrado en los rangos especificados de X_1 y X_2.

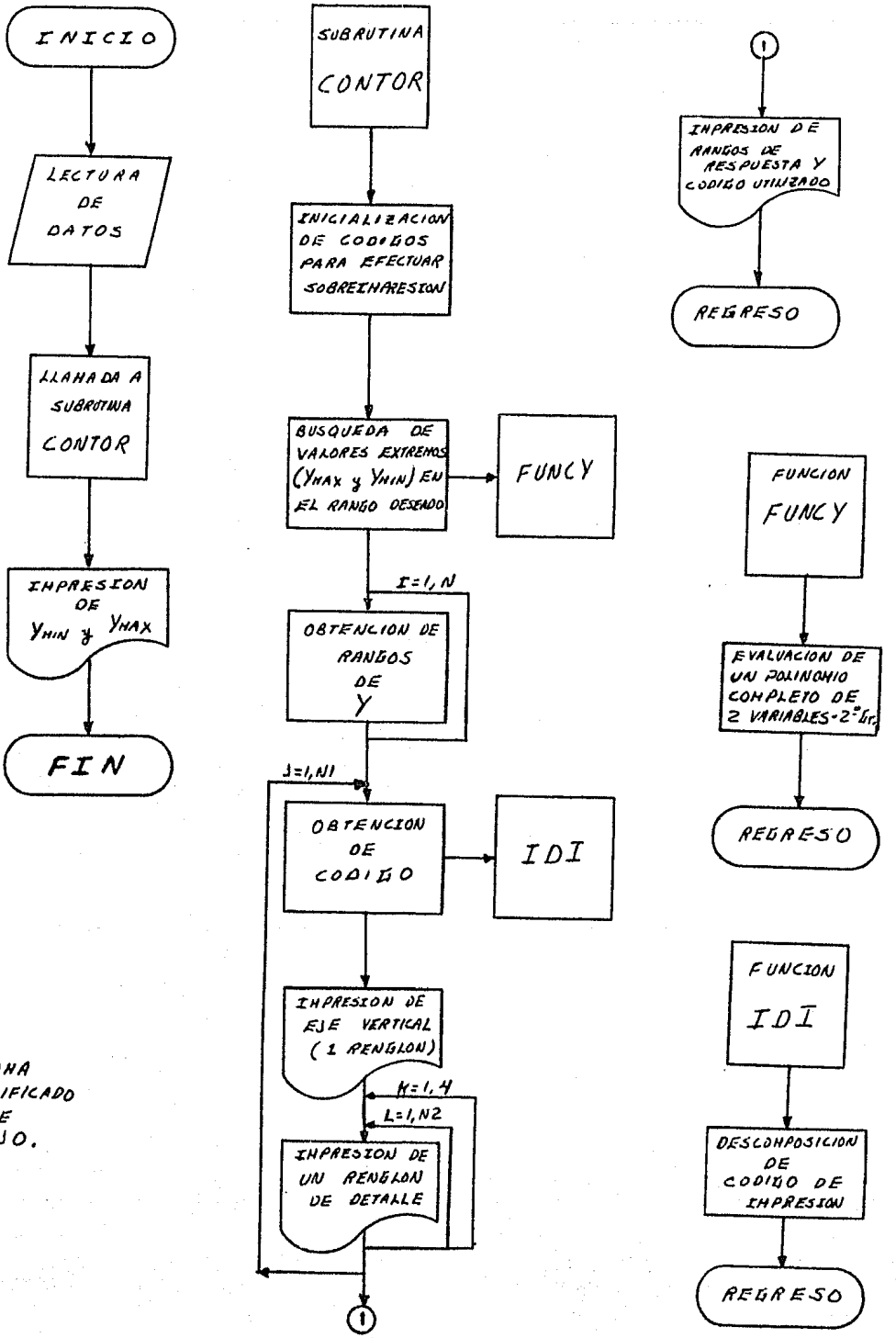


DIAGRAMA SIMPLIFICADO DE FLUJO.

BIBLIOGRAFIA

1.- Atkinson, A. C.

Statistical Designs for Pilot Plant and Laboratory Experiments - Part I

Chemical Engineering

Mayo 9, 1966, pp 149-154

2.- Bacon, David W.

Making the Most of a One-shot Experiment

Industrial and Engineering Chemistry

Vol. 62, No. 7, pp 37-34 (Julio 1970)

3.- Balzhiser, Richard E. et al.

Termodinamica Quimica para Ingenieros

(Trad. Hernández Sanchez Juan L.)

Editorial Prentice-Hall International; Valparaiso, Chile; 1974, pp 35-49

4.- Boas, Arnold H.

What Optimizacion is all About

Chemical Engineering

Diciembre 10, 1962, pp 147-152

5.- Boas, Arnold H.

Optimizing Multivariable Functions

Chemical Engineering

Marzo 4, 1963, pp 97-104

6.- Box, G. E. P.

The Exploration and Explotation of Response Surface. Some General Considerations and Examples

Biometrics

Vol. 10, Marzo 1954, pp 16-60

7.- Box, G. E. P. and P. V. Youle
An Example of the Link Between the Fitted Surface and the Basic Mechanism of the System
Biometrics
Vol. 11, Septiembre 1955, pp 287-323

8.- Box, G. E. P. and Draper, Norman R.
A Basis for the selection of a Response Surface Design
American Statistical Association Journal
Vol. 54, Septiembre 1959, pp 622-654

9.- Burch, J. G. and Strater F. R.
Sistemas de Información
(Trad. Ricardo Calvet Pérez)
Editorial Limusa
México, 1981, pp 491-504

10.- Burtis, C. A. et al
Optimization of Kinetic Method by Response Surface Methodology and Centrifugal Analysis and Aplication to the Enzymatic Measurement of Ethanol
Analytical Chemistry
Vol. 53, 1981, pp 1154-1159

11.- Cardenas, A. F. et al
Ciencias de la Computación Vol. I
Editorial Limusa-Wiley
México, 1972

12.- Cardenas, A. F. et al
Ciencias de la Computación Vol. II
Editorial Limusa-Wiley
México 1972, pp 15-167

13.- Cochran, William G. and Cox, Gertudre M.
Diseños Experimentales
(Trad. Centro de Estadistica y Cálculo del Colegio de Posgraduados de la Escuela - Nacional de Agricultura) 2ª ed.
Editorial Trillas
México, 1981

14.- Draper, Norman R. and Smith, Harry
Applied Regresión Analysis 2ª ed.
Editorial Wiley Interscience
New York, 1981

15.- Fisher, R. A.
The Design of Experiments 4ª ed.
Editorial Oliver and Boyd
Edimburgo, 1947, Caps. 2 y 3

16.- Galton, F.
Family Likeness in stature
Procs. Royal Society
London, 1886, pp 40-47

17.- Hamaker, H. C.

Experimental Design in Industry

Biometrics

Vol. 11, Septiembre 1955, pp 257-258

18.- Hill, W. J. and Hunter, W. G.

A Review of Response Surface Methodology a Literature Survey

Technometrics

Vol. 8, 1966, pp 571-579

19.- Himmenblau, D. M.

Process Analysis by Statistical Methods

Editorial Wiley

New York, 1970, pp 230-257

20.- Hopkin, David and Moss, Barbara

Automata

Editorial Macmillan Press LTD

London, 1976, pp 1-66

21.- Hunter, J. S.

In Plant Experiment. Need not Disturb Your Operation

Chemical Engineering

Marzo 28, 1966, pp 111-118

22.- Hunter, J.S. and Box. G. E. P.

Multifactor Experimental Design for Exploring Response Surfaces

Annal. of Math. Statistical

Vol. 28, 1957, pp 195-241

23.- Hunter, William G.

Statistical Designs for Pilot-Plant and Laboratory Experiments Part II.

Chemical Engineering

Junio 6, 1966, pp 159-164

24.- Isaacson, William B.

Statistical Analyses for Multivariable Systems

Chemical Engineering

Junio 29, 1970, pp 69-75

25.- Jhonson, W. W.

A Least-Squares Method of Interpreting Magnetic Anomalies Caused by Two-dimensional Structures

26.- Karson, M. J. et al

Minimun Bias Estimation and Experimental Design for Response Surfaces

Technometrics

Vol. 11, 1969, pp 461-475

27.- Lewis II, P. M. et al

Compiler Design Theory

The Systems Programming Series

Editorial Addison-Wesley

U.S.A., 1976

28.- Manual de Uso

Statistical Analysis System

Version 76.5, Raleigh, N.C., USA

29.- McCracken, Daniel D.
Programación FORTRAN IV
(Trad. Jairo Osuna Suárez) 2ª ed.
Editorial Limusa
México, 1981

30.- Méndez R., Ignacio
Modelos Estadísticos Lineales. Interpretación y Aplicaciones 2ª ed.
Editorial CONACYT
México, 1981

31.- Miller, Irwin and Freund, Jhon E.
Probability and Statistics for Engineers
Editorial Prentice-Hall
Englewood Cliffs, 1965

32.- Murphy, Thomas D.
Design and Analysis of Industrial Experiments
Chemical Engineering
Junio 6, 1977, pp 168-182

33.- Mustacchi, Carlos y Moresi, Mauro
A Strategy to Obtain Semi-emprical Correlations for Deterministic Systems
Chemical Engineering Science
Vol. 35, 1980, pp 737-741

34.- Read, D. R.

The Design of Chemical Experiments

Biometrics

Vol. 10, Marzo 1954, pp 1-15

35.- Remington

Bioestadistica

Editorial Trillas, México 1978

36.- Sefa, D. and Stanley D.

Cowpea Proteins. 1. Use of Response Methodology in Predicting Cowpea (Vigna Unguiculuta) Protein Extractability

J. Agric. Food Chem.

Vol. 27, No 6, 1979, pp 1238-1234

37.- Weigand F. Eckehard

Conformational Relaxation as Limitation of Chemical Models.

J. of American Chemical Society

Vol. 101, No 24, 1979, pp 7195-7198