Universidad Nacional Autónoma de México

ESCUELA NACIONAL DE ESTUDIOS
PROFESIONALES - ACATLAN

SIGNIFICADO Y USO DE MODELOS ESTADISTICOS LINEALES

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A

ROSA SALAZAR VALDES

ACATLAN, MEXICO 1980

M-0037496

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

E.N.E.P. ACATLAN U.N.A.M.
COORDINACION DEL PROGRAMA
DE INGENIERIA Y ACTUARIA

CAI-C-008-80.

SRITA. ROSA SALAZAR VALDES,
Alumna de la carrera de Actuario,
P r e s e n t e .

De acuerdo a su solicitud presentada con fecha 8 de enero de 1980, me complace notificarle que esta Coordinación tuvo a bien asignarle el siguiente tema de tesis: "SIGNIFICADO Y USO DE MODELOS ESTADISTICOS LINEALES", el cual se desarrollará como sigue:

Introducción
1. Aspectos generales de modelos estadísticos lineales.
2. Principales modelos estadísticos lineales
3. Aplicaciones
4. Conclusiones

Asimismo fue designado como Director de Tesis el -- Señor Act. Jorge Olguin Uribe, Profesor de esta Escuela.

Ruego a usted tomar nota que en cumplimiento de lo especificado en la Ley de Profesiones, deberá prestar servicio social durante un tiempo mínimo de seis meses como requisito básico para sustentar examen profesional, así como de la disposición de la Dirección General de Servicios Escolares en el sentido de que se imprima en lugar visible de los ejemplares de la Tesis, el título del trabajo realizado. Esta comunicación deberá imprimirse en el interior de la Tesis.

A T E N T A M E N T E
"POR MI RAZA HABLARA EL ESPIRITU"
Acatlán Edo. de México a de Febrero de 1980.

ENEP - ACATLAN
COORDINACION DE
INGENIERIA Y ACTUARIA

ING. ALEJANDRO RAMIREZ SECEÑA
Coordinador del Programa
de Ingeniería y Actuaría

A Emanuel que le ha dado
un sentido a mi vida.

A Juan Humberto y Rosa Astrea
por su inagotable apoyo.

A mis hermanos Juan, Sergio,
Mario, José, Octavio y Sara.

Agradezco a Jorge Olguin U.
la orientación y apoyo que
me brindó durante el desa-
rrollo de este trabajo.

I N D I C E

Pág.

INTRODUCCION

La Mátemática como ciencia formal ocupa un lugar preponderante en la estructura general de las ciencias, esto es debido a un sistema ordenado y lógico de justificar y demostrar sus teorías.

Es por ello que la aplicación de la Matemática en las diferentes ciencias sociales se ha incrementado considerablemente. Vemos entre otras la Estadística, que es una de las más poderosas herramientas con las que contamos para describir una situación existente, crear sistemas y modelos, así como para inferir resultados que nos llevarán a hacer predicciones respecto al comportamiento de un fenómeno.

En general la Estadística es aplicable a cualquier campo de estudio en el cual se hagan observaciones. Los métodos estadísticos forman, hoy, una parte importante de muchos campos de la ciencia y se están desarrollando rápidamente dada su gran utilidad. Algunos de estos métodos fueron introducidos por Karl Gauss, pero fueron Karl Pearson y sobre todo R. A. Fisher los que dieron un impulso más fuerte al uso general de los mismos a principios de este siglo. Nos referimos en especial a los modelos matemáticos que toman en cuenta aspectos aleatorios de los fenómenos conocidos

como "modelos estadísticos" o "modelos estocásticos". Los - modelos estadísticos lineales además de ser los más senci -- llos de este tipo, pueden utilizarse en diversas situaciones.

En este trabajo se pretende exponer de la manera - más clara y sencilla (sin enfatizar en los desarrollos matemáticos) el concepto de modelo estadístico lineal presentando sus formas más usadas, así como poner en claro las condiciones bajo las cuales deben emplearse dichos modelos de tal suerte que al ser utilizados se llegue al uso adecuado de -- los mismos.

Asimismo se mencionarán algunas de las principales aplicaciones que tienen dentro de la regresión y el diseño - de experimentos.

Se espera que pueda ser de utilidad para los estudiantes de la carrera de Actuaría como complemento para los cursos de Estadística y Análisis de Regresión; y para carreras como Ingeniería, Biología, Medicina, Psicología, Economía, etc., en donde manejen problemas con variabilidad de tipo aleatorio.

I. ASPECTOS GENERALES DE MODELOS ESTADISTICOS LINEALES

En este capítulo presentamos un glosario de términos (a los cuales haremos referencia durante el desarrollo de este trabajo) con la idea de puntualizar conceptos que son de importancia para el tratamiento de los modelos estadísticos lineales. También daremos un breve recordatorio al tema de Distribuciones de Frecuencias ya que nos ayudará a comprender el uso de la distribución normal.

Llamamos "Fenómenos aleatorios" a aquella clase de fenómenos en los que no es posible determinar una relación exacta entre distintos aspectos de ellos.

En este sentido tenemos que no es posible describir con un modelo matemático exacto, por ejemplo, las relaciones entre la cantidad de fluor en los dientes de los individuos y el número de caries que tengan, o bien entre concentración de oxígeno, plancton y sal en el mar y el número de peces que lo habitan.

Podemos entonces apreciar que para la ocurrencia de ciertos fenómenos intervienen diversos factores que al ser considerados hacen que no podamos determinar de manera exacta los resultados que vamos a obtener.

Al surgir el concepto de probabilidad y establecer

modelos matemáticos utilizando dicho concepto, fue posible - describir las relaciones entre diversos aspectos o modalidades de los fenómenos de tipo aleatorio. Estos modelos fue - ron construidos a raíz de la existencia de cierta regularidad que se manifiesta al estudiar un fenómeno un número gran de de veces en condiciones iguales o muy semejantes.

Un problema central en el descubrimiento de nuevos conocimientos acerca del mundo real consiste en observar algunos (ya que en muchas ocasiones no es posible observar a - todos) de los elementos bajo discusión y, sobre la base de - éstos, hacer una afirmación referente a la totalidad de los mismos. Esto nos lleva a definir el concepto de "población" que es el conjunto de valores posibles (mediciones o conte - os) de una característica particular común en un grupo especificado de seres u objetos. Es importante recalcar que para definir una población se requieren especificar ciertos - factores comunes a todos los individuos u objetos sobre los que se efectúan las mediciones. Así, nos referimos:

a) Al conjunto de indígenas que usan calzado en - la zona Tarasca del Edo. de Michoacán,

b) Al conjunto de balines de 1" de diámetro que - produce diariamente la Cía. X.

En la población a) tenemos los siguientes factores comunes: i)que sean indígenas, ii)que sean de la zona Tarasca, iii)del Edo. de Michoacán y iv)que usen calzado.En la población b) encontramos: i)que sean balines, ii)de 1" de diá-

metro y iii)producidos diariamente por la Cía. X.

En este sentido tenemos que la población puede sufrir variaciones al considerar un mayor (o menor) número de factores comunes en los individuos u objetos pertenecientes a la misma, puesto que al especificar algunos factores, dejamos de enunciar muchos otros que pueden variar entre los individuos u objetos de la población y que darán como resultado una fluctuación en la medición de sus elementos.

Volviendo a nuestros ejemplos, podemos aumentar algunas especificaciones a nuestras poblaciones de la siguiente manera:

a.1) Conjunto de indígenas mayores de 5 años que pertenezcan a la zona Tarasca de Michoacán, que usen calzado, que hablen español y que vivan en el Mun. de Aquila,

b.1) Conjunto de balines de 1" de diámetro producidos diariamente por la Cía. X, que son empacados y distribuidos.

De este modo tenemos que: en a.1) el hecho de que los indígenas sean mayores de 5 años, que hablen español y que vivan en el Municipio de Aquila harán que la población a) disminuya con respecto a la población a.1) ya que eliminaremos a los indígenas menores de 5 años, a los indígenas que vivan fuera del Municipio de Aquila y a los indígenas que aún siendo mayores de 5 años y vivan en el Municipio de Aquila no hablen español. Ahora, en b.1) hemos especificado que además de ser balines de 1" de diámetro producidos diariamen

te por la Cía. X, sean empacados y distribuidos, lo cual hará variar nuestra población.

Obsérvese que, particularmente en la población b.1 puede darse el caso de que el haber aumentado el número de especificaciones de la población original b), no necesariamente hará variar ésta última. Es decir, puede suceder que todos los balines de 1" producidos diariamente en la Cía. X sean empacados y distribuidos. En este caso las poblaciones b) y b.1) no variarán.

Del mismo modo podemos disminuir las especificaciones que fueron enunciadas en a) y b) diciendo:

a.2) Conjunto de indígenas de la zona Tarasca que usan calzado,

b.2) Conjunto de balines producidos diariamente en México.

Tenemos entonces que con a.2) y b.2) aumentan algunos factores que ahora quedan indefinidos y que estuvieron enunciados originalmente en a) y b). Es lógico que en a.2) y b.2) nuestras poblaciones sean mayores, ya que aumenta el número de factores no constantes.

De lo anterior diremos que las poblaciones a.2) y b.2) tienen un "mayor grado de generalidad" que las poblaciones a) y b); y diremos que a.1) y b.1) tienen un "menor grado de generalidad" con respecto a las poblaciones a) y b). Por lo anterior diremos que el concepto de población es "flexible".

Al hablar de población resulta conveniente considerar lo siguiente:

i) Es importante que la población sea susceptible de ser definida con absoluta precisión (sin ambigüedad); por ejemplo en una población agrícola, fijando reglas para definir lo que es un rancho o una hacienda, de manera que pueda decidirse acertadamente si un caso dudoso pertenece o no a la población. Por lo demás puede ser real o hipotética,

ii) En general, al ampliarse el concepto de población, o sea al tener un mayor grado de generalidad, habrá más variabilidad, es decir, se tendrán mayores discrepancias en las mediciones que de ella se obtengan.

En las aplicaciones de los modelos lineales, el concepto de población se emplea repetidas veces. Por lo general, se suponen poblaciones infinitas que en la práctica suelen ser finitas con un gran número de elementos. Luego entonces, para el estudio de las poblaciones se suele tomar una "muestra" que es una parte de la población seleccionada de acuerdo a un plan o regla previamente establecido de "n"* elementos de la población.

Ya obtenida la muestra, sus "n" elementos pueden ser tantos que virtualmente son inútiles, a menos que se condensen o se reduzcan en una forma conveniente. Esto puede hacerse mediante un procedimiento gráfico que permita destacar las peculiaridades del conjunto de valores muestrales.

---

*Ver Referencia bibliográfica No. 3 capítulo 4.

Entonces, se procede a ordenar los "n" valores muestrales -- con la construcción de una "tabla de frecuencias" que divide los datos en un número relativamente reducido [10 a 20] de - intervalos de valores llamados "clases" [categorías] donde - se registrará el número de veces, o sea, la "frecuencia" con la que aparece un dato de la muestra en cada una de las clases. Así, la tabla de frecuencias indica las clases y la -- frecuencia de cada clase.

Una tabla de frecuencias sacrifica parte de la información contenida en un conjunto de datos, es decir, en lugar de conocer el valor exacto de cada dato sólo conocemos - que pertenece a cierta clase. Por otra parte, este tipo de agrupamientos elimina características importantes de los datos y la única ganancia es la legibilidad de los datos, lo - que en general compensa sobradamente la pérdida de informa-- ción.

Para ilustrar la construcción de una tabla de frecuencias consideremos los datos de la Tabla 1.1 que representa las velocidades en kilómetros por hora de 120 automóviles que pasan por un punto de vigilancia con radar en una carretera.

El primer paso para construir una tabla de frecuencias consiste en decidir cuántas clases se van a usar y cuáles son los límites de cada clase. Frecuentemente el número de clases empleado depende del número de observaciones, pero es raro que se pueda usar con utilidad menos de 10 o más de

96	123	90	130	95	84	104	65	90	63	80	161
75	53	116	71	93	122	108	119	113	96	86	130
89	83	98	114	77	113	101	78	131	74	98	69
113	96	93	128	140	81	122	114	86	96	137	110
147	128	117	144	126	90	135	126	90	110	101	75
108	78	143	105	72	89	93	56	98	114	68	105
95	90	62	132	99	101	117	110	81	113	99	75
74	149	114	102	120	72	123	104	105	108	119	86
111	111	105	110	98	108	81	113	111	87	104	110
83	120	80	107	155	98	111	86	95	104	98	87

Tabla 1.1

20 clases. Entre otras cosas basamos esta decisión en el - "recorrido" de los datos que es la diferencia entre la observación mayor y la menor. En el ejemplo la observación mayor es 161 y la menor 53 lo cual representa el "rango" de nuestros datos. El recorrido será entonces 108.

Escogiendo intervalos entre los que los datos puedan ser contados podemos elegir 11 clases que tengan por intervalos 47 - 57, 58 - 68, 69 - 79, . . . , 157 - 167. Nótese que en cada caso los intervalos de las clases no se sobreponen y que todas las clases son del mismo tamaño.

Ordenando y agrupando las 120 observaciones obtenemos la siguiente tabla de frecuencias (Tabla 1.2)

Intervalos de clase: x	Conteo	Frecuencia: f	Porcentaje del total
47 - 57	11	2	1.67
58 - 68	1111	4	3.33
69 - 79	1111 1111 11	12	10.00
80 - 90	1111 1111 1111 1111 1	21	17.50
91 - 101	1111 1111 1111 1111 1	21	17.50
102 - 112	1111 1111 1111 1111 111	23	19.17
113 - 123	1111 1111 1111 1111	20	16.67
124 - 134	1111 111	8	6.67
135 - 145	1111	5	4.17
146 - 156	111	3	2.50
157 - 167	1	1	0.83
Totales		120	100.00

Tabla 1.2

Obsérvese que la suma de las frecuencias es igual al tamaño de la muestra.

Es usual calcular la "frecuencia relativa" en la tabla de frecuencias, ya que nos dice que proporción de las observaciones totales cae en cada clase. Su valor se determina dividiendo cada frecuencia de clase entre el total de las frecuencias. Tenemos entonces que la suma de las frecuencias relativas es igual a la unidad. Para ilustrar esto veamos la Tabla 1.3.

Intervalos de clase	Frecuencia	Frecuencia relativa
47 - 57	2	0.017
58 - 68	4	0.033
69 - 79	12	0.100
80 - 90	21	0.175
91 - 101	21	0.175
102 - 112	23	0.192
113 - 123	20	0.167
124 - 134	8	0.067
135 - 145	5	0.041
146 - 156	3	0.025
157 - 167	1	0.008
Totales	120	1.000

Tabla 1.3

Muchas propiedades importantes de las distribuciones de frecuencias tales como la simetría y asimetría, el número de sus máximos, etc., son más fácilmente observados por medio de gráficas. Utilizando la tabla de frecuencias del ejemplo se construyen las siguientes representaciones gráficas:

a) Histograma.- que está formado por rectángulos cuyas bases son los intervalos de clase y las alturas concuerdan con la frecuencia de clase. Ver Fig. 1.1.

b) Perfil.- se representa uniendo las bases superiores de los rectángulos obtenidos en el histograma. Ver Fig. 1.2.

c) Polígono de frecuencias.- se obtiene uniendo -

los puntos medios de las bases superiores de los rectángulos Ver Fig. 1.3.

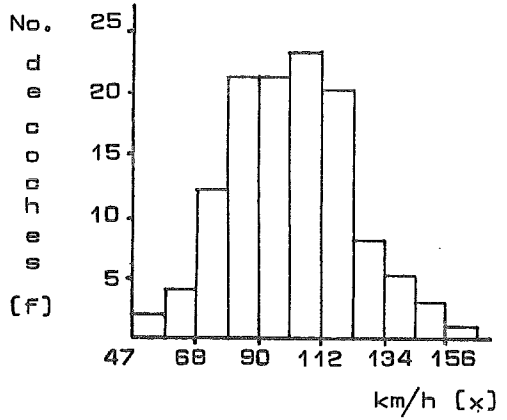

Fig. 1.1

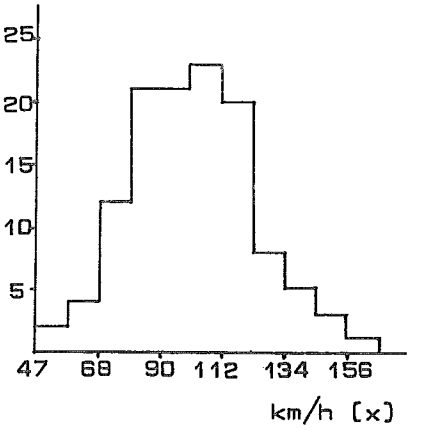

Fig. 1.2

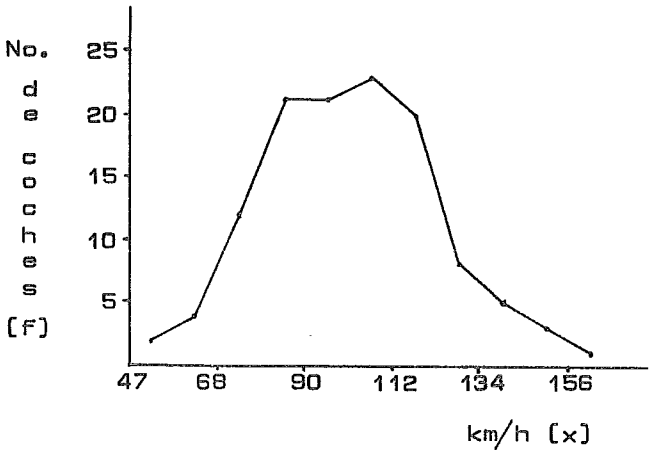

Fig. 1.3

Cabe señalar que las representaciones gráficas a - partir de la tabla de frecuencias nos dan una idea del promedio y sobre todo de la variabilidad de los datos. Para pre-

cisar esto podemos recurrir al cálculo de las medidas de tendencia central y las de dispersión, de las cuales principalmente se emplean la media aritmética y la varianza.

Es conveniente recoger suficiente información para poder obtener el grado de precisión que se desee. Mientras más medidas se efectuen, al dibujar las gráficas de la tabla de frecuencias obtendremos un mejor ajuste a la distribución de la población.

Se ha visto en muchos casos y estudiado en diversas variables que si la muestra es de tamaño grande (si se miden muchos elementos) y los histogramas se construyen agrupando las medidas en intervalos de menor anchura, la forma general de las distribuciones de frecuencias es muy frecuentemente la misma. Ver Fig. 1.4.

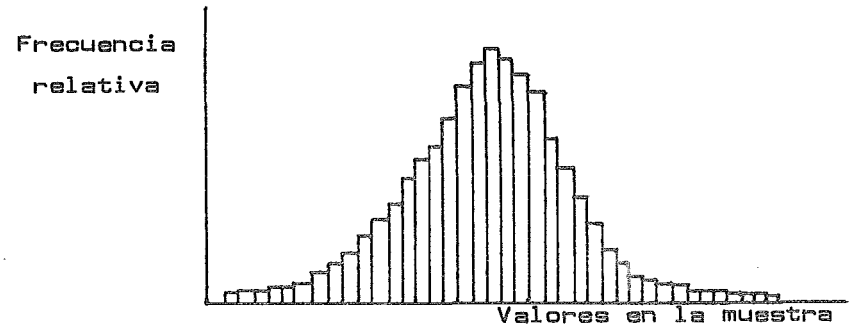

Fig. 1.4

Además se ha observado que en los fenómenos aleatorios existe cierta "regularidad estadística", es decir, al tomar valores grandes de "n" las frecuencias relativas tienden a estabilizarse. Por ejemplo, existe regularidad en las

edades en que las mujeres contraen matrimonio. Al aumentar el tamaño de la muestra se advierte que los valores tienden a ser constantes (se estabilizan) en ciertas edades. En el ejemplo podrían ser entre 18 y 25 años.

En vista de lo anterior se creó un modelo matamáti co para representar lo que se espera que sea la distribución de las frecuencias relativas al considerar un conjunto infinito de mediciones en todos los individuos de la población. Nos referimos al modelo conocido como "distribución normal". El modelo matemático surge al estimar las frecúencias relati vas en el caso límite en que se muestrea a toda la pobla --- ción. De este modo el histograma tiende a una curva conti-- nua. Su gráfica, representada por la Fig. 1.5, es la de una curva simétrica acampanada que se extiende en ambas direccio nes positiva y negativa. Su utilización es la misma que la de cualquier otra curva de distribución.

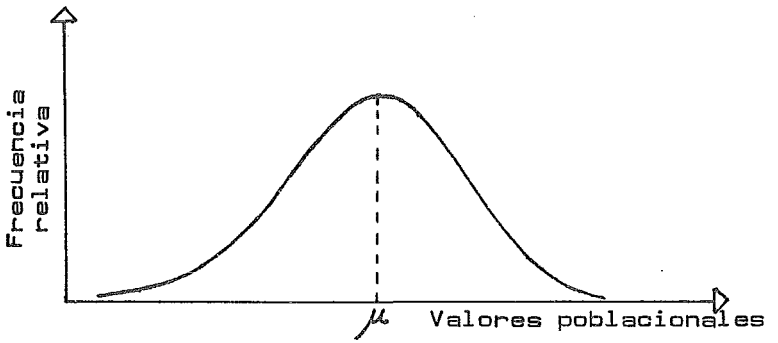

Fig. 1.5

Es conocido que una distribución normal queda ca-- racterizada por el conocimiento de sus dos parámetros: media

y varianza. La media aritmética de toda población, o sea , el número que resultaría al promediar todos los valores de - las mediciones, se representa por μ y se ilustra en la Fig. 1.5. La media muestral se representa por $\bar{x}$. La media indica dónde está situado el centro de la distribución sobre - el eje de las abcisas.

Un promedio en sí mismo no da una clara descrip-- ción de una distribución. Otra medida que ayuda a aclarar - la forma de la distribución es la que indica cómo las observaciones se separan de la media. Esta medida se conoce como dispersión, extensión o variabilidad. La medida de variabilidad usual es la varianza poblacional σ^2 o la desviación estándar poblacional σ . La varianza y desviación estándar - muestrales se representan por s^2 y s respectivamente.

Para indicar que la distribución de frecuencias de una variable y , en toda la población, puede representarse con el modelo normal, se escribe $y \sim N(\mu, \sigma^2)$ que significa: y se distribuye normalmente con media poblacional μ y varianza poblacional σ^2 .

En adelante veremos que la distribución normal desempeña un papel preponderante.

En primer lugar, muchas de las poblaciones que se encuentran en investigaciones realizadas en muy diversos campos parecen tener una distribución bastante aproximada a la normal. Se ha dicho que este fenómeno es razonable si se - tiene en cuenta el Teorema central del límite. Con esto no

se pretende indicar que la mayor parte de las distribuciones encontradas en la práctica sean normales, porque no es éste el caso, pero sí que es muy frecuente encontrar distribuciones aproximadas a la normal.

Otra consideración a favor de la distribución normal es el hecho de que las distribuciones en el muestreo a menudo son lo suficientemente grandes para que el estimador obtenido de ellas se distribuya aproximadamente de una manera normal, lo que resulta cómodo desde el punto de vista analítico.

Cuando poseemos información acerca de dos o más variables relacionadas (o concomitantes*) es natural buscar un modo de expresar las "relaciones funcionales" (de las cuales posteriormente nos ocuparemos). No buscamos únicamente una función matemática que nos diga de qué manera están relacionadas las variables, sino que también queremos saber con que precisión se pueden hacer afirmaciones acerca de el valor de una variable si conocemos los valores de las variables asociadas.

Para escoger una relación funcional particular como representativa de la población bajo investigación se emplean dos métodos: 1) Una consideración analítica del fenómeno que nos ocupa, y 2) un exámen de la distribución de fre--

*La concomitancia entre dos fenómenos revela, pero no necesariamente postula, una relación causal (causa-efecto) entre ellos.

cuencias en forma gráfica de los datos observados.

Con esto pretendemos hacer notar que es de suma importancia tener conocimiento correcto del problema para hacer una selección adecuada de nuestras variables.

Con respecto a la información es necesario que ésta sea de tipo cuantitativo. De no ser así, debemos evaluarla de alguna manera para hacerla manejable.

La información cuantitativa la podemos encontrar en áreas tales como Ingeniería, Economía, Medicina, etc., por ejemplo: El precio de un producto es expresado en No. de pesos, el desempleo es expresado en número de personas, la longitud de una carretera se indica en No. de kilómetros, la "vida" de un tractor se mide en número de horas trabajadas, la respuesta a un estímulo se expresa en número de ocurrencias, etc.

En cuanto a la organización de los datos, si la información es poca (15,20 datos) puede ser manual; pero si no lo es puede recurrirse a hojas de registro, tarjetas perforadas, etc., y a un sistema electrónico de procesamiento de datos de tal forma que puedan manejarse rápida y eficazmente.

Hemos visto que en muchas ocasiones resulta impráctico, imposible o muy costoso y lento trabajar con la población total que representan nuestros datos, por lo cual se procede a extraer una muestra de la población bajo estudio. Consideremos lo siguiente con respecto a la muestra poblacional:

a) Cuando la muestra no contiene las características principales de la población en estudio, la media de los datos no reflejará la realidad. Luego entonces, nuestras conclusiones serán inadecuadas.

b) Si la muestra no fue bien seleccionada y se tomaron demasiados datos de cierta característica y pocos de otra de mayor importancia, la interpretación que demos a nuestros datos no corresponderá a los propósitos del problema en estudio.

c) Otro riesgo es el que se refiere a la cantidad de la información, pues puede llevarnos a supuestos falsos. Por ejemplo, en cuanto a la tendencia. Supóngase que una persona sugiere que se pronostiquen ventas sabiendo que éstas fueron en ascenso durante el trimestre de enero a marzo de la siguiente menera: 2, 3.5 y 5 millones de pesos respectivamente. Mientras que otra persona sugiere que se pronostiquen ventas considerando que de abril a junio descendieron en la siguiente forma: 5, 3 y 1.5 millones de pesos respectivamente. Es claro que los datos no son suficientes para hacer una buena predicción.

Si podemos controlar lo antes mencionado, podremos interpretar la información correctamente de tal forma que sea posible extraer conclusiones válidas y decisiones rezonables.

II. PRINCIPALES MODELOS ESTADISTICOS LINEALES

II.1 Modelo de regresión lineal simple.

La regresión lineal tiene principalemente dos finalidades:

1) Estimar una función lineal que relacione dos o más variables, y

2) Describir el tipo de asociación entre la variable y (dependiente o respuesta) y las variables x_1, x_2, . . . , x_p (independientes).

Cuando en la relación hay una variable independiente se dice que se trata de regresión lineal simple; si en la relación intervienen dos o más variables independientes se llama regresión lineal múltiple.

Debe quedar claro que y_i es el valor de una variable aleatoria cuya distribución depende de x_i.

Ahora nos ocuparemos de los llamados modelos lineales simples.

Sean y_1, y_2, . . . , y_n variables aleatorias observables y tales que $y_i = \alpha + \beta x_i + \varepsilon_i$ en donde α y β son parámetros desconocidos; x_i, variables matemáticas (no aleatorias) y ε_i variables aleatorias no observables, cuya media es cero y tienen varianza dada por σ^2

[σ^2 no es función de α , β o x_i]. Estos supuestos definen un modelo lineal simple.

Analicemos cada uno de los supuestos anteriores.

i) y_i puede representar el estudio de cualquier variable observable que presente variación aleatoria en sus valores. Se considera que los valores de la variable provienen de un número teóricamente infinito de valores posibles -[población]. Así, y_i puede ser por ejemplo el rendimiento de un tractor, o la distancia que ha recorrido una partícula durante cierto lapso de tiempo, o bien, y_i puede representar los valores obtenidos al hacer mediciones de una característica física determinada, digamos la longitud de una varilla.*

ii) Las observaciones y_i no tendrán relación alguna entre ellas de modo que si modificamos alguna esto no afectará a las demás. Diremos entonces que son independientes.

iii) Tenemos que

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad [1]$$

donde

α y β son parámetros desconocidos los cuales se estimarán dependiendo de si:

a) Las variables aleatorias y_i se distribuyen normalmente, entonces para un

*Decimos "los valores obtenidos" y no "el valor obtenido" -- porque se condidera que debido a que la medición resulta de un proceso aleatorio -ya que la medida es afectada por temperatura, humedad, luz con la que se hace la lectura, escala - de medida empleada, etc.- tendremos diferentes valores de y_i.

valor dado de x ; la variable aleatoria (v.a.) y se distribuye normalmente con media $\alpha + \beta x$ y varianza σ^2 ; y los parámetros α y β estimados serán:

$$\hat{\beta} = \frac{\Sigma x y}{\Sigma x^2} \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

con $\bar{x} = \frac{1}{n}\Sigma x_i$ y $\bar{y} = \frac{1}{n}\Sigma y_i$

b) Las v.a. no se distribuyen normalmente, en cuyo caso se recurre al método de estimación de mínimos cuadrados y se llega a que $\hat{\alpha}$ y $\hat{\beta}$ son los mismos que en - a).

x_i Son variables matemáticas que pueden tomar cualquier valor dentro de los núme-- ros reales.

ε_i Es considerado como aleatorio, o sea que no pueda predecirse su valor en un momento dado. Además es no observable, con - media 0 y varianza σ^2. (posteriormente se verá con más detalle).

En la mayoría de las situaciones en que la relación de y_i respecto a x_i es aproximadamente lineal, lo que nos interesa principalmente es la relación entre las $x's$ y la media de la distribución correspondiente de las $y's$ que como vimos está dada por $\alpha + \beta x$ para x determinada. - Llamémosle μ a esa media.

μ es una constante característica de los miem - bros de la variable en estudio y_i ; así, al decir que las y_i fluctuarán en torno a un valor μ queremos decir que pa

ra ciertas condiciones que llamaremos "controladas", los valores observados y_i tenderán a un valor μ. Aclaremos esto con un ejemplo. Al estudiar el rendimiento en kilómetros por litro de un automóvil, debemos tener en cuenta que intervendrán varios factores que nos permitirán llegar a determinar un valor como rendimiento. Entonces, para determinadas condiciones controladas tales como marca, modelo, velocidad y carretera constantes, debe existir en el rendimiento cierta tendencia a un valor en especial μ que depende de dichas condiciones.

Además de las condiciones controladas cuyo valor hace que las observaciones y_i fluctúen alrededor de μ, existen otras condiciones que intervienen para determinar el valor de las observaciones y_i. Podemos llamarles condiciones "no controladas". Precisamente a ellas se debe que los valores de las observaciones y_i fluctúen alrededor de [y no sean iguales a] un valor μ. Estas fluctuaciones son los ε_i. En nuestro ejemplo podrían ser la manera de conducir, la forma de efectuar las mediciones, la temperatura ambiental, la altitud, etc. Todos estos factores pueden ser agrupados y se los considerará aleatorios debido a su gran número. Así, podemos reescribir el medelo [1] de la siguiente manera:

$$y_i = \mu + \varepsilon_i \qquad [2]$$

donde

y_i variable en estudio,

μ constante característica de los miembros de la población,

ε_i error aleatorio por las características específicas de la i-ésima observación.

De este modo ε_i es la desviación entre y_i y la media de la población μ, así $\varepsilon_i = y_i - \mu$. Entonces, el modelo dado por (2) es una identidad, ya que

$$y_i = \mu + \varepsilon_i = \mu + y_i - \mu$$

Analizando un momento el ejemplo puede observarse que los ε_i más frecuentes son los valores cercanos a cero y son menos frecuentes al alejarse de cero hacia un lado o hacia otro. Los errores aleatorios ε_i no son previsibles con exactitud por lo que se dificulta la predicción de su efecto como conjunto, sin embargo no por ello debe pensarse que no tienen cierta regularidad estadística. Además es importante observar que los errores pueden ocurrir tanto en un sentido como en otro, es decir, ε_i tendrá valores positivos y negativos; pero al estudiar un gran número de valores ε_i su promedio estará cercano a cero de manera que al dar oportunidad de que aparezcan todos los tipos de fluctuaciones aleatorias de un fenómeno, éstas serán tales que se anularán unas a otras dando un promedio de cero. Este punto constituye una de las suposiciones básicas en los modelos estadísticos lineales (m.e.l.).

Al hacer un gran número de mediciones observamos que el promedio de los valores ε_i estará cercano a cero y que serán menos frecuentes los valores de ε_i más lejanos a cero. Se tendrá así que el valor de ε_i fluctua alrededor de 0. Si se representaran graficamente las frecuencias relativas de los ε_i obtendríamos una figura parecida a la Fig. 2.1.

Puesto que el modelo postulado es $y_i = \mu + \varepsilon_i$, si a cada valor de ε_i se le suma la constante μ, se tendrá una variación aleatoria en las sumas resultantes, o sea en los valores de y_i. De modo que los valores de y_i son aleatorios a causa de los errores aleatorios ε_i.

Ahora si representamos los valores de y_i contra sus frecuencias de ocurrencia, se tendrá la Fig. 2.2.

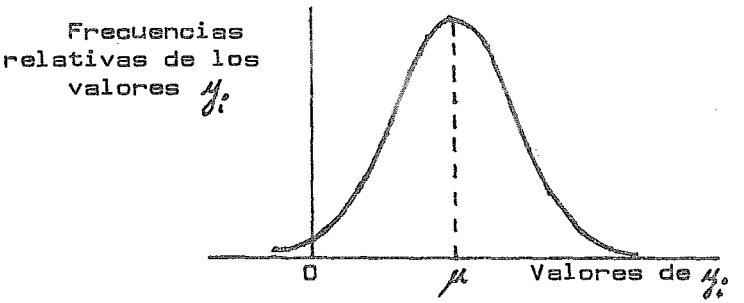

Fig. 2.2

Al comparar las figuras 2.1 y 2.2 se observa que la diferencia entre ellas radica únicamente en la posición de la curva, ya que al sumarse la constante μ a los valores ε_i para constituir los valores de y_i, lo que se hace es recorrer (desplazar) la curva una cantidad μ. La dispersión o variabididad, ya sea de los valores ε_i o y_i, es i--

déntica ya que no se alteran las frecuencias de ocurrencia - de ningún valor.

A partir de los modelos (1), (2) y los que posteriormente veremos podemos estar interesados en obtener la distribución de los estimadores α y β y deducir conclusiones importantes acerca de los mismos, o bien hacer pruebas de hipótesis que nos ayuden a validar nuestras suposiciones. Dichos aspectos se salen de los objetivos de este trabajo - por lo cual no nos ocuparemos de ellos.

Hemos visto que la distribución normal es un modelo adecuado para representar las frecuencias relativas de los valores de un gran número de poblaciones. El concepto de población como ya se explicó, es flexible; esto es, dependiendo de las especificaciones que se hagan sobre la población, tendremos poblaciones con un mayor o menor grado de generalidad. Por lo tanto, si cambiamos las especificaciones que definen a una población es de esperarse que los parámetros μ y σ de la distribución cambien.

Ahora bien, si de un concepto de población pasamos a otro de mayor grado de generalidad, se pueden esperar cambios imprevistos en la media provocados por la intromisión - de nuevos miembros a la población original. Sin embargo en la varianza puede esperarse un incremento, puesto que en la población de mayor grado de generalidad aumenta la variabilidad de las condiciones particulares del individuo u objeto - en el que se efectúa cada medición y_i, incrementando con -

esto la variabilidad de las ocurrencias de los y_i. Al recordar que los ε_i son las discrepancias de una medición con relación a la media de la población producida a causa de las condiciones específicas relativas al individuo objeto de la medición, tenemos que al aumentar el grado de generalidad de la población se incrementa la variabilidad de los ε_i que como vimos, se refleja en las y_i.

Por ejemplo, al estudiar la población dada por las estaturas de los indígenas de 10 años que habitan en la zona Tarasca del Edo. de Michoacán, se podría tener una distribución normal con media μ de 130 cm. y una desviación estándar de 20 cm. Al tener la población un mayor grado de generalidad, como ocurre al considerar las estaturas de los indígenas Tarascos de 10 años que habitan fuera del Edo. de Michoacán (en los Edos. de Gto., Méx., Col. y Gro.), se espera que ocurra un cambio en los parámetros; μ podría aumentar o disminuir dependiendo de varios factores que estarán dados en general por las condiciones de vida de los nuevos indígenas considerados, sin embargo, es razonable que σ aumente ya que habrá mayor variabilidad en las observaciones y_i.

Considerando varias poblaciones con el mismo grado de generalidad por ejemplo la longitud de la mano en hombres adultos que habitan en los Estados de a_1) Veracruz, a_2) Chiapas y a_3) Sonora, podemos esperar cambios en las μ, pero muy probablemente tendrán las mismas varianzas puesto que el tipo de factores no especificados al definir la población y

que intervienen para determinar los $\mathcal{E}_i$, o sea, los factores particulares asociados con la medición y_i , siguen siendo los mismos. En el ejemplo estos factores no especificados serían el tipo de alimentación, constitución genética, tipo de actividad, situación geográfica, etc. Luego, la variabilidad de esos factores es esencialmente la misma. Esto nos lleva a mencionar otro supuesto básico que se hace al tratar con modelos estadísticos lineales que es la "homogeneidad de las varianzas poblacionales" (homocedasticidad).

Podemos presentar en forma más compacta el ejemplo si consideramos a μ como función de los factores especificados al definir la población y a σ como función de los factores no especificados, de la siguiente manera:

$$y_{1i} \sim N\left(\mu(a_1,b,c),\quad \sigma_1^2(d,e,f,\ldots)\right)$$
$$y_{2i} \sim N\left(\mu(a_2,b,c),\quad \sigma_2^2(d,e,f,\ldots)\right)$$
$$y_{3i} \sim N\left(\mu(a_3,b,c),\quad \sigma_3^2(d,e,f,\ldots)\right)$$

donde

y_{1i}, y_{2i} y y_{3i}	son las longitudes de la mano,
b	señala hombres,
c	señala adultos,
a_1	Veracruzanos,
a_2	Chiapanecos,
a_3	Sonorenses,
d	tipo de alimentación,
e	constitución genética,

f tipo de actividad,
etc.; otros factores no constantes.

II.2 Modelo de regresión lineal múltiple.

Dentro de los modelos estadísticos lineales el concepto básico es el 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 observaciones.

Dado que el rasgo esencial en los modelos lineales es el que considera que las medias de la población dependen de los factores que definen a las poblaciones, el modelo - - $y_{i\cdot} = \mu + \varepsilon_i$ para una población se generaliza al modelo:

$$y_{ijk\ldots m} = \mu\,(\chi_j, \chi_k, \ldots, \chi_m) + \varepsilon_i \qquad (3)$$

donde

$y_{ijk\ldots m}$ es la variable en estudio con ciertas especificaciones $j, k, \ldots, m$

$\mu(\chi_j, \chi_k, \ldots, \chi_m)$ representa la media de una población definida por los factores específicos dados por $\chi_j, \chi_k, \ldots, \chi_m$ y señala su dependencia en dichos - factores,

ε_i tienen una distribución normal, son independientes con media 0 y varianza σ^2, i.e. $\varepsilon_i \sim N(0, \sigma^2)$.

Se considera que esa dependencia de las medias al

ser definidos ciertos factores específicos, es de tipo li-- neal [en los parámetros], o sea que cumple

$$\mu\,[x_j, x_k, \ldots, x_m] = \sum_{\omega=1}^{p} \beta\omega\, g\omega[x_j, x_k, \ldots, x_m] \qquad [4]$$

donde

$\beta\omega$ son parámetros desconocidos que deben ser estimados,

$g\omega[x_j, \ldots, x_m]$ con $\omega = \overline{1,p}$; son funciones conocidas especificadas por $x_j, x_k, \ldots, x_m$ que son las modalidades de los factores que intervienen al definir una población específica.

Seleccionando adecuadamente las $g\omega[x_j, x_k, \ldots, x_m]$, lo cual estará basado en el conocimiento previo del fenómeno en estudio, podemos obtener una relación satisfactoria entre las modalidades dadas por los factores $x_j, x_k, \ldots, x_m$ [que en el caso de regresión lineal son de tipo cuantitativo] y el va-- lor de la media en la población definida por dichas modalidades. Así, combinando [3] y [4] y considerando las observa-- ciones hechas, el modelo lineal general queda de la siguiente forma:

$$y_{ijk\ldots m} = \sum_{\omega=1}^{p} \beta\omega\, g\omega[x_j, x_k, \ldots, x_m] + \varepsilon_i \qquad [5]$$

Este modelo es el que tiene un mayor grado de -- flexibilidad.

Debe aclararse que se canservan las mismas suposi-

ciones que en el modelo de regresión lineal simple en cuanto a que los valores $\mu_{ijk\ldots m}$ resultan al considerar la media de una población definida por los factores específicos $\chi_j, \chi_k, \ldots, \chi_m$, pero además, a causa de la ocurrencia de los errores aleatorios ε_i .

El modelo [5] suele representarse en forma vectorial así:

$$Y = \chi\beta + \varepsilon \qquad [6]$$

donde

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \chi = \begin{bmatrix} \chi_{11} & \chi_{12} & \cdots & \chi_{1p} \\ \chi_{21} & \chi_{22} & \cdots & \chi_{2p} \\ \vdots & \vdots & & \vdots \\ \chi_{n1} & \chi_{n2} & & \chi_{np} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

tales que

Y	es un vector observado aleatorio,
χ	una matriz $n\chi p$ de cantidades fijas - conocidas,
β	es un vector $p\chi 1$ de parámetros desconocidos,
ε	es un vector aleatorio.

Al igual que en el modelo de regresión lineal simple pueden ocurrir dos casos relacionados con la distribución del vector ε :

Caso a) Cuando cada ε_i se distribuye normalmente

con media 0 y varianza igual a σ^2 y que los ε_i son conjuntamente independientes,

Caso b) Cuando el valor esperado de cada $\varepsilon_i = 0$, la matriz de covarianza de ε_i es $\sigma^2 I$ en donde σ^2 es desconocida.

Bajo la teoría normal en el Caso a) y utilizando el Teorema de Gauss-Markoff en el Caso b), se llegan a estimar β y σ^2.*

II.3 Modelos lineales en diseño de experimentos.

Es sabido que resulta prácticamente imposible que se cumplan de manera exacta las suposiciones de los modelos planteados. Una característica común de los experimentos en muy diversos campos de la investigación es que los efectos de los tratamientos experimentales varían de un ensayo a otro cuando se repiten. Esta variación introduce cierto grado de incertidumbre en cualquiera de las conclusiones que se obtienen de los resultados, los cuales están afectados no solamente por la acción de los tratamientos sino también por las variaciones extrañas que tienden a encubrir sus efectos. El término de errores experimentales se aplica frecuentemente a estas variaciones, donde la palabra errores no es sinó-

*Ver Referencia bibliográfica No. 11 pags. 394 - 402.

nimo de equivocaciones, sino que incluye todos los tipos de variación extraña.

Se pueden distinguir dos fuentes principales de errores experimentales. La primera es la variabilidad inherente al material experimental al cual se aplican los tratamientos. La población puede ser una parcela, un paciente en un hospital o una porción de masa, un grupo de aves en un corral o un lote de semillas. Es característico de tales poblaciones que produzcan diversos resultados aun cuando se sujeten al mismo tratamiento. Estas diferencias bien sean grandes o pequeñas, contribuyen a formar los errores experimentales. La segunda fuente de variabilidad es la falta de uniformidad en la conducción física del experimento, es decir la diferencia en poder uniformizar la técnica experimental.

En consecuencia, la investigación de los métodos para incrementar la exactitud de los experimentos en base al mejor diseño de los mismos ha desempeñado un papel importante en la investigación experimental.

El diseño de un experimento es la secuencia completa de pasos tomados de antemano para asegurar que los datos apropiados se obtendrán de modo que permitan un análisis objetivo que conduzca a elecciones válidas con respecto al problema establecido.

El propósito de cualquier diseño experimental es proporcionar una cantidad máxima de información pertinente

al problema bajo investigación. Sin embargo también es conveniente que el diseño, plan o programa de prueba sea tan -- simple como sea posible.

En el caso de los diseños experimentales la forma dada por [4] es válida, sólo que ahora las funciones $g\omega[x_j, x_k, \ldots, x_m]$ serán funciones indicadoras con valores cero o uno según la presencia o ausencia de los parámetros $\beta\omega$, que se interpretan como los efectos de las distintas modalidades que definen a la población a través de $x_j, x_k, \ldots, x_m$. A estas modalidades se les llama "niveles de los factores estudiados".

En la mayoría de las investigaciones, el investigador trata con más de una variable independiente y con los -- cambios que ocurren con la variable dependiente cuando varía una o más de las variables independientes. En el lenguaje - de diseño experimental una variable independiente se conoce como un "factor". En el siguiente ejemplo se enlistan cinco factores para el estudio del lavado de ropa en casa: 1) tipo de agua, 2) temperatura del agua, 3) duración del tiempo de lavado, 4) tipo de máquina lavadora y 5) clase de agente limpiador. Algunos de los niveles de los factores para el - ejemplo anterior pueden ser respectivamente: dos tipos de agua, dos temperaturas del agua, tres tiempos de lavado, dos tipos de máquinas lavadoras y tres clases de agente limpiador. Por lo tanto para cada factor 1, 2 y 4 habrán dos nive

les y cada factor 3 y 5 aparecerá en tres niveles.

A la suposición de que la forma definida por [4] - como $\mu[x_j, x_k, \ldots, x_m]$ se exprese como función lineal en los parámetros $\sum_{\omega} \beta\omega\, g\omega[x_j, x_k, \ldots, x_m]$ se le denomina "aditividad de efectos". La aditividad se expresa en los diseños experimentales en los que las $g\omega$ son variables indicadoras con valores cero y uno; en ese caso la expresión es una suma de -- parámetros.

Entonces, si redefinimos a las $g\omega[x_j, x_k, \ldots, x_m]$ como variables independientes $z\omega$, el modelo [5] quedará:

$$y_i = \sum_{\omega=1}^{p} \beta\omega\, z_{\omega i} + \varepsilon_i \qquad [7]$$

Con este modelo definimos en general a todos los - modelos lineales. De este modo, surgen tres divisiones al - considerar el conjunto de las $z\omega$:

1) Si son introducidas como variables indicadoras de la presencia o ausencia de modalidades cualitativas de -- los efectos que definen a las poblaciones, se obtienen los - modelos de diseños experimentales donde las $z\omega$ toman valo-- res cero para indicar ausencia y uno para indicar presencia de dichas modalidades,

2) Si las $z\omega$ son variables reales que pueden tomar cualquier valor dentro de ciertos intervalos dados por - $x_j, x_k, \ldots, x_m$ [llamados región de exploración], se obtienen los modelos de regresión, y

3) Es posible obtener una combinación de los dos

tipos de comportamiento dados por 1) y 2) para el conjunto - $\mathcal{Z}\omega$ en los llamados modelos de covarianza o modelos mixtos. Para el modelo [7] se tiene que:

-La base que argumenta la normalidad descansa en - el tipo de variables estudiado y la experiencia previa con - variables similares. Se supone que las poblaciones genera-- das en cada punto del espacio de exploración, o sea para ca- da conjunto de condiciones especificadas por las variables - independientes [sean de clasificación en modelos de diseño - de experimentos o cuantitativas en los modelos de regresión lineal] son normales.

-Para la homogeneidad de varianzas la base es la - consideración de que las poblaciones y_i y en consecuencia - de $\mathcal{E}_i$, generadas en los puntos del espacio de n dimensio- nes de las $\mathcal{Z}\omega$, tienen los mismos factores no definidos y - por consiguiente igual tipo de variabilidad.

-El concepto $\mathcal{E}_i$ sigue siendo el mismo que se discu tió en II.1. En cuanto a la independencia, se considera que la manera de tomar mediciones se haga de preferencia aleato- riamente. El proceso de aleatorización tiene como uno de -- sus objetivos introducir independencia entre los errores.

Si la forma dada en [4] denominada "relación fun-- cional" es correcta, ninguna de las variables no incluidas _ en el modelo ni en las especificaciones generales tiene un e fecto preponderante, sino efectos pequeños, lo cual tiene -- que ver con la distribución de los errores $\mathcal{E}_i$. Por lo tan

to la elección de las funciones $g_\omega = z_\omega$ se basa en conocimientos previos sobre el fenómeno.

Si la relación funcional es incorrecta por existir variables o funciones de las variables consideradas que sean de importancia y que no estén incluidas en el modelo, entonces los estimadores de las pruebas de hipótesis resultarán afectados de manera imprevisible. También cabe mencionar que aunque se tenga una relación funcional inadecuada, la estricta aleatorización del proceso de muestreo con el cual se obtienen los valores de $y_{ijk\ldots m}$ garantiza que el modelo resulte válido, aunque la variabilidad de los errores resulte mayor que cuando la relación funcional es correcta. En este caso, o sea cuando la relación funcional es incorrecta, los errores resultantes pueden no estar normalmente distribuidos.

Cuando expusimos los modelos de regresión lineal simple y regresión lineal múltiple se trataron modelos que un experimentador utilizaría si estuviese interesado en encontrar una fórmula a partir de la cual pudiera predecir el valor de un factor en estudio, por medio de otro relacionado con él. Así, por ejemplo el experimentador puede desear predecir la dureza y de un metal, conociendo la temperatura T y el tiempo t de cierta operación química. El modelo sería $y = \mu[T, t] + \varepsilon$. En lo futuro consideraremos una situación algo diferente, en la que el interés no radica en predecir un valor de un factor utilizando valores de fac-

tores relacionados con él, sino en comparar los efectos de - dos o más factores.

II.3.1 Modelo de diseños con un criterio de clasificación.

Presentamos un ejemplo a partir del cual obtendremos el modelo. Supongamos que el director de una fábrica desea comprar máquinas para realizar cierta operación en un - proceso de producción. Hay tres compañías que fabrican tales máquinas y toma a prueba una de cada compañía a fin de - determinar cual de las tres es la más apropiada a sus fines. El director hace que menejen las máquinas varios de sus hombres durante unos días para descubrir con cual de las tres - se consigue mayor producción por día.

En este sencillo experimento la medida que interesa es el número de productos por día, y el factor único utilizado (lo cual nos lleva a utilizar un modelo de diseño con un criterio de clasificación) es el tipo de máquina.

Imaginemos que se utilizan seis hombres en el experimento, asignando al azar dos para cada tipo de máquina. - Habrá entonces dos observaciones para cada una de las tres - máquinas, siendo cada observación la cantidad producida por la máquina en un día. Los datos podrán ser los que aparecen en la tabla siguiente:

Número de máquina

1	2	3
64	41	65
39	48	57

Lo interesante es saber si las máquinas son o no - diferentes en cuanto al número de elementos que son capaces de producir. Se piensa que con cada máquina se tiene una población normal de producción por día, con varianza constante pero diferentes medias μ_j. El modelo es:

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

en donde

y_{ij} denota la observación i-ésima de la población j, en que $j = 1,2,3$ representa las máquinas,

μ_j media de la máquina j con $j = 1,2,3$,

ε_{ij} se considera que están normal e independientemente distribuidos con media 0 y varianza σ^2, i.e. $\varepsilon_{ij} \sim NID\ [0, \sigma^2]$.

Ahora, si construimos una población ignorando el - cambio de máquina, se tendrá una nueva población con cierta media μ y varianza en general mayor que σ^2. Esto último se debe a que nuestra nueva población tendrá un grado de generalidad superior a las tres anteriores ya que ahora hay un factor más que no es constante: la máquina; por lo que se espera mayor variabilidad en esta población general. A la me-

dia μ de esta población más general se le llama "media general".

En cada una de las poblaciones estudiadas se puede definir un "efecto especial" al considerar que la media de la población particular cambia con relación a la media de la población general. En el ejemplo son las peculiaridades específicas de la población estudiada -dadas por modalidades o niveles de factores que la definen- cuyo grado de generalidad es menor que en la población general. En el ejemplo son las peculiaridades específicas inherentes a una máquina.

Así, en el ejemplo considerado se define un efecto específico de máquina como la discrepancia entre μ_{j° y μ ; se le llama efecto del factor máquina y se denota por τ_{j° ; luego entonces

$$\tau_{j^\circ} = \mu_{j^\circ} - \mu$$

donde

μ_{j° es la media de la población de valores de y manteniendo constante la máquina j° .

Los valores τ_{j° serán positivos y negativos: positivos para las máquinas que aumenten la producción diaria en promedio con relación a la media de la población de todas las producciones diarias con cualquiera de las máquinas consideradas; negativos cuando la producción media diaria de una máquina vaya en detrimento con respecto a la media general.

Entonces el modelo para nuestro ejemplo puede es-- cribirse

$$\begin{aligned} y_{11} &= \mu + \tau_1 + \varepsilon_{11} \\ y_{12} &= \mu + \tau_1 + \varepsilon_{12} \\ y_{21} &= \mu + \tau_2 + \varepsilon_{21} \\ y_{22} &= \mu + \tau_2 + \varepsilon_{22} \\ y_{31} &= \mu + \tau_3 + \varepsilon_{31} \\ y_{32} &= \mu + \tau_3 + \varepsilon_{32} \end{aligned} \qquad [8]$$

y en forma más compacta

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij} \qquad [9]$$

con $i = 1,2$; $j = 1,2,3$; $\varepsilon_{ij} \sim$ NID $[0, \sigma^2]$ y en donde

y_{ij} producción diaria de la medición i-ésima -- con la máquina j-ésima,

μ media general,

τ_j efecto de la población (máquina) j-ésimo,

ε_{ij} error aleatorio producido fundamentalmente por las particularidades específicas de la i-ésima medición en una producción obtenida con la máquina j-ésima, que se genera por - los factores no considerados como constan-- tes al definir la población.

Con notación matricial el sistama de ecuaciones -- (8) se escribirá

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix} \qquad [10]$$

en donde los unos y ceros denotan presencia o ausencia res-- pectivamente de características cualitativas de efectos. El modelo [10] para n observaciones y p parámetros se escribe

$$Y = X\beta + \varepsilon \qquad [11]$$

donde

Y es un vector aleatorio observable $n \times 1$,

X es una matriz $n \times p$ de rango $q < p < n$ y que - contiene sólo ceros y unos,

β vector de parámetros desconocidos $p \times 1$,

ε vector de variables aleatorias no observa-- bles $n \times 1$.

El modelo [11] representa el modelo general de diseño experimental* y además está implícito en el modelo general de todos los modelos lineales dado por [7].

Daremos otro ejemplo para ilustrar el modelo de diseño con un criterio de clasificación. Supongamos que cierta compañía desea determinar si existe alguna diferencia entre los distintos métodos de fabricación de acero. Imaginemos que se examina una muestra de n_i trozos de cable fabricado por el proceso i, para cada uno de t procesos y que la resistencia a la rotura del j-ésimo trozo de cable fabricado por el i-ésimo proceso es y_{ij}, entonces podemos escribir

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad \text{con } i = 1,2,\ldots, \quad j = 1,2,\ldots, \quad \varepsilon_{ij} \sim \text{NID}\,[0, \sigma^2)$$

*Consultar Referencia bibliográfica No. 6 pags. 226 - 253.

donde

y_{ij} medición de la resistencia a la rotura del j-ésimo trozo de cable fabricado por el proceso i-ésimo,

μ es la media general de la población de la resistencia a la ruptura,

τ_i efecto del i-ésimo proceso,

ε_{ij} error aleatorio no observable debido a las variaciones incontrolables en los procesos de fabricación y medición.

Los ejemplos anteriores ilustran modelos de diseños con un criterio de clasificación porque las poblaciones en estudio difieren por los niveles o categorías de un factor: en el primer ejemplo las diferencias obtenidas son debidas a las peculiaridades específicas inherentes al factor máquina, y en el segundo ejemplo las diferencias obtenidas provienen del factor cable de acero.

Para ilustrar gráficamente un modelo de diseño con un criterio de clasificación veamos la Fig. 2.3.

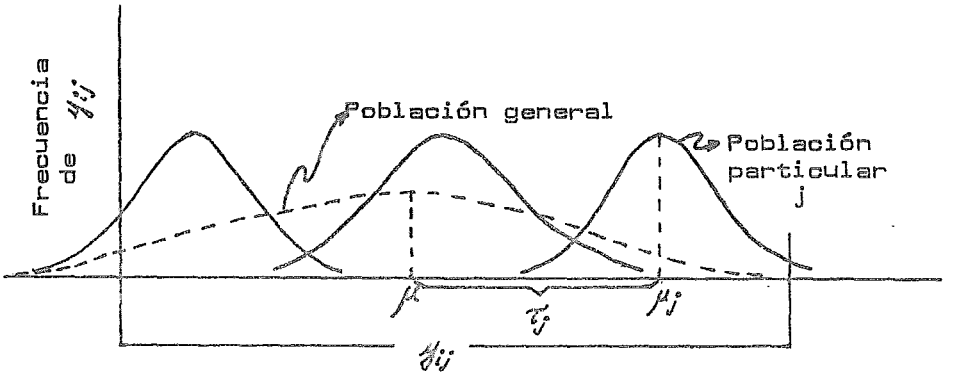

Fig. 2.3

II.3.2 Modelo de diseños con dos criterios de clasificación.

Tal vez se haya observado que el primer ejemplo dado en la sección II.3.1 tenía un diseño poco satisfactorio. El inconveniente se debía a la presencia de un factor externo: la pericia de los diversos obreros que han de intervenir necesariamente en el experimento. Si la producción de una máquina resulta relativamente grande ¿se deberá ello a la máquina o a la superioridad del grupo particular de trabajadores que se le asignó?

En el lenguaje de diseño de experimentos, los efectos debidos a máquinas y los debidos a grupos de trabajadores están confundidos* y no hay modo de distinguir entre estos dos factores. La dificultad desaparece al diseñar un nuevo experimento considerándolo como bifactorial.

Supongamos ahora que intervienen sólo cinco hombres en el experimento y que cada uno de ellos trabaja un día en cada una de las tres máquinas. El orden en que ha de trabajar un hombre dado en las tres máquinas se fijará aleatoriamente. Los datos quedarán clasificados en la Tabla 2.1 y tendrán doble entrada, es decir, con los niveles de los dos factores que llamaremos α y β.

En cada casilla o celda de la tabla se genera una población normal con una media que depende de los niveles de

*Se dice que dos o más efectos se confunden en un experimento si es imposible separar los efectos cuando se lleva a cabo el subsecuente análisis estadístico.

Valores de μ_{jk}

Hombre (Factor $\mathcal{B}$)	Máquina (Factor α) 1	2	3
1	53	47	57
2	56	50	63
3	45	47	54
4	54	47	57
5	49	53	58

Tabla 2.1

los factores -es decir, de la celda que se trate- y de que la varianza de las distintas poblaciones (una para cada celda) sea constante. Esto último se basa en la suposición de que las poblaciones de cada una de las celdas tienen el mismo grado de generalidad.

El modelo se representa identificando los niveles de los factores mediante índices:

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk}$$

donde

j, k niveles de los factores de las máquinas y los hombres,

μ_{jk} media de la población generada de la celda j, k, i.e., al definir el nivel j del factor máquina y el nivel k del factor hombre como constantes en los miembros de la población,

ε_{ijk} desviación de la medición i-ésima de la población en la celda j, k con respecto a su media μ_{jk}. Esto se origina por las peculia

ridades del i-ésimo elemento estudiado en - esa población, ocasionadas principalmente - por los factores no señalados al definir la población. Por lo tanto $\varepsilon_{ijk} \sim$ NID $(0, \sigma^2)$.

En nuestro ejemplo hemos considerado que se tienen tres niveles del factor a a saber $j = 1,2,3$ y cinco niveles del factor b que son $k = 1,2,3,4,5$. Entonces, según se ignoren uno u otro de los factores o ambos a la vez, tendremos poblaciones con varios grados de generalidad. Así, si se ignoran los niveles del factor a se tendrán cinco poblaciones una para cada nivel del factor b con medias μ_k ; y ciertas varianzas que en general serán mayores que σ^2. Si se ignoran los niveles del factor b se contará con tres poblaciones, una para cada nivel del factor a , con media μ_j y varianza posiblemente mayor que σ^2. Si se ignoran los niveles de ambos factores se obtiene una población con mayor grado de generalidad con media μ (media general).

Partiendo de las medias de las diferentes poblaciones consideradas se definen los "efectos" del modo siguiente:

$\tau_j = \mu_j - \mu$ es el efecto del factor a (o efecto de a) en su nivel j .

$\rho_k = \mu_k - \mu$ es el efecto del factor b (o efecto de b) en su nivel k .

En los diseños de dos o más criterios de clasificación, en los que se forman tablas de doble o múltiple entra-

da, se tiene la posibilidad de considerar que los efectos : i) no interactúen, o sea que no se modifican mutuamente, o - bien que ii) interactúen, esto es, que debido a la influen-- cia combinada de dos o más factores resulta un efecto adicio nal. Interacción es la respuesta diferencial a un factor en combinación con niveles variables de un segundo factor apli- cado simultáneamente.

i) La no interacción de efectos en el diseño con dos criterios de clasificación se basa en el hecho de que un cambio de nivel en un factor produce una variación constante en las medias μ_{jk} al considerar los niveles del otro factor, o sea

$$\mu_{jk} - \mu_{j',k} = \mu_{jk'} - \mu_{j',k'} \qquad \text{para todo } j', j'', k, k' \quad (12)$$

lo que equivale a

$$\mu_{jk} - \mu_{jk'} = \mu_{j'k} - \mu_{j'k'} \qquad \text{para todo } j', j'', k, k' \quad (13)$$

Por ejemplo, supongamos el factor a con los nive- les $j' = 1,2$ y el factor b con los niveles $k = 1,2,3$. Los - valores de μ_{jk} de estos factores se dan en la siguiente ta-- bla:

		Factor a	
		1	2
Factor b	1	8	6
	2	5	3
	3	2	0

Entonces, utilizando [12] tenemos que para todo - $k = 1,2,3$ y $k' = 1,2,3$ se cumple

$$\mu_{1k} - \mu_{2k} = \mu_{1k'} - \mu_{2k'} = 2$$

y

$$\mu_{2k} - \mu_{1k} = \mu_{2k'} - \mu_{1k'} = -2$$

Utilizando [13] tenemos que para todo $j' = 1,2$ y - $j'' = 1,2$ se cumple

$$\mu_{j'1} - \mu_{j'2} = \mu_{j''1} - \mu_{j''2} = 3$$
$$\mu_{j'1} - \mu_{j'3} = \mu_{j''1} - \mu_{j''3} = 6$$
$$\mu_{j'2} - \mu_{j'3} = \mu_{j''2} - \mu_{j''3} = 3$$

El cambio de niveles de los factores a y b produce alteraciones iguales en las medias de las poblaciones para todos - los niveles b y a respectivamente.

Luego, puesto que se cumplen las condiciones [12] y [13] se dice que no hay interacción en los factores a y b.

ii) Interacción.- se dice que hay interacción entre dos factores si los cambios de nivel de un factor afectan de manera diferente las μ_{jk} al cambiar los niveles del otro factor. O sea que si la proporcionalidad entre los valores de μ_{jk} se pierde, entonces habrá interacción entre los - factores considerados. Basta con que un valor de μ_{jk} no cumpla con las condiciones [12] y [13] para que surja la inter-

acción. Así, en el ejemplo considerado al inicio de la sección II.3.2 podemos verificar que existe interacción con sólo confirmar que

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

para $j = 1$, $j' = 3$, $k = 1$ y $k' = 2$. Entonces,

$$\mu_{11} - \mu_{31} = -4 \qquad \mu_{12} - \mu_{32} = -7$$

por lo tanto existe interacción entre los factores a y b del ejemplo mencionado.

Luego entonces, para el caso de no interacción tenemos que la media de la población jk es la suma de la media general y los efectos τ_j y ρ_k, es decir

$$\mu_{jk} = \mu + \tau_j + \rho_k$$

y el modelo del diseño con dos criterios de clasificación -- sin interacción es

$$y_{ijk} = \mu + \tau_j + \rho_k + \varepsilon_{ijk} \qquad [14]$$

donde

y_{ijk}	medición i-ésima con nivel j de a y nivel k de b,
μ	media general,
τ_j	efecto del nivel j de a,
ρ_k	efecto del nivel k de b,

ε_{ijk} error aleatorio por las características específicas particulares de la i-ésima medición en la población con niveles j de a y k de b. Entonces $\varepsilon_{ijk} \sim N[0, \sigma^2]$ y $\varepsilon_{ijk} = y_{ijk} - \mu_{jk}$.

El modelo [14] es una identidad ya que

$$y_{ijk} = \mu + \tau_j + \rho_k + \varepsilon_{ijk} = \mu_{jk} + \varepsilon_{ijk}$$

$$y_{ijk} = \mu_{jk} + y_{ijk} - \mu_{jk}$$

En el caso de que exista interacción, para presentar μ_{jk} en términos de μ, τ_j y ρ_k se introducirá un efecto adicional: δ_{jk} [denominado efecto de interacción], así:

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

y el efecto δ_{jk} será

$$\delta_{jk} = (\tau\rho)_{jk} = \mu_{jk} - \mu_j - \mu_k + \mu$$

Por lo tanto el modelo de diseño con dos criterios de clasificación con interacción es

$$y_{ijk} = \mu + \tau_j + \rho_k + \delta_{jk} + \varepsilon_{ijk} \qquad [15]$$

donde y_{ijk}, μ, τ_j, ρ_k y ε_{ijk} son los mismos que en [14] y

δ_{jk} es el efecto de interacción del nivel j de a y el nivel k de b.

M-0037496

En general, las suposiciones básicas de los mode-- los vistos en esta sección son las mismas que las consideradas anteriormente: la normalidad, que descansa en el hecho - de juzgar que cada casilla o celda de las tablas de clasificación múltiple tienen población normal con varianza constante, y que las medias dependen de la casilla específica a través de efectos principales e interacciones; la homocedasticidad, que resulta al considerar que las poblaciones en cada - celda tienen el mismo grado de generalidad.

Es posible extender los modelos [14] y [15] a tres o más criterios de clasificación con o sin interacción, siempre y cuando se cumplan las suposiciones antes mencionadas y se expresen los nuevos modelos en términos de variables indicadoras de presencia o ausencia de un efecto principal o interacción en cada una de las mediciones realizadas. Así, un diseño con tres criterios de clasificación con interacción - se escribiría

$$y_{ijkm} = \mu_{jkm} + \varepsilon_{ijkm}$$

o más explícitamente

$$y_{ijkm} = \mu + \tau_j + \rho_k + \lambda_m + [\tau\rho]_{jk} + [\tau\lambda]_{jm} + [\rho\lambda]_{km} + [\tau\rho\lambda]_{jkm} + \varepsilon_{ijkm}$$

y sin interacción

$$y_{ijkm} = \mu + \tau_j + \rho_k + \lambda_m + \varepsilon_{ijkm}$$

en donde $i = 1,2,\ldots,I$ $\quad k = 1,2,\ldots,K$

$j = 1,2,\ldots,J$ $\quad m = 1,2,\ldots,M$

para ambos casos.

En los modelos [11], [14] y [15] se han presentado los llamados efectos de poblaciones τ_j y ρ_k y su interacción δ_{jk}. Estos valores son considerados constantes fijas particulares, aunque desconocidas, de una situación experimental determinada por lo cual dichos modelos se denominan de "efectos fijos".

Mencionaremos algunos ejemplos en los cuales pueden ser utilizados los modelos vistos en esta sección.

-En experimentación agronómica, un tratamiento puede referirse a los factores a) a una marca de fertilizante y b) a una cantidad de fertilizante.

-En experimentación de nutrición animal, por ejemplo el aumento de peso en ganado lanar, los factores considerados pueden ser a) el sexo de los animales, b) el padre del animal experimental y c) el tipo de proteina.

-En estudios psicológicos y sociológicos podemos considerar los factores a)edad, b) sexo, c) grado de escolaridad, etc.

-Para estudiar el rendimiento de cierto proceso químico podemos considerar a)la temperatura a la cual se ejecuta el proceso y b) la cantidad de catalizador usada.

-En la investigación y desarrollo concerniente a baterías, los tratamientos pueden ser varias combinaciones -

de α) la cantidad de electrolito y β) la temperatura a la cual fue activada la batería.

II.3.3 Diseño experimental en bloque azarizado.

Todos los modelos lineales presentados son generalizaciones de la sección II.1 y son casos particulares de - (4). En todos ellos, el término ε_i representa el error aleatorio generado por los factores no especificados al definir las poblaciones. Si la variabilidad de los ε_i es grande el error experimental puede reducirse adoptando una o más de las técnicas siguientes: 1) usando material experimental - más homogéneo o por la estratificación cuidadosa del mate--- rial disponible, 2) utilizando información proporcionada por variables aleatorias cuya relación funcional sea correcta , 3) teniendo más cuidado al dirigir el experimento, 4) usando un diseño experimental más eficiente.

El propósito de esta sección es estudiar técnicas que reduzcan la magnitud de los errores experimentales en - los modelos lineales ligados a los diseños experimentales, - para lo cual utilizaremos el concepto de "bloque". Por bloqueo damos a entender la distribución de los elementos de la población en estudio (unidades experimentales) en bloques , - de tal manera que las unidades dentro de un bloque sean relativamente homogéneas, de esta manera la mayor parte de la va

riación predecible entre las unidades queda confundida con - el efecto de los bloques.

La introducción de bloques en los diseños permite disminuir la variabilidad que presentan los $\mathcal{E}_i$ en los modelos, mediante la identificación de ciertos factores que aunque no sean objeto de estudio producen variación considera-- ble en las mediciones y_i. Originalmente el concepto de - bloque se introdujo en los modelos de diseño experimental pero puede extenderse a modelos de regresión, esto es, si redefinimos el modelo lineal general dado en II.2 por (3) y re-- presentamos a las funciones $g\omega$ por $v_j, v_k, \ldots, v_m$, tendremos que

$$y_{ijk\ldots m} = \beta_0 + \beta_j v_j + \beta_k v_k + \ldots + \beta_m v_m + \mathcal{E}_i \qquad (16)$$

donde

$\{\beta\}$ son los efectos que interesa estudiar,

$\{v\}$ pueden ser variables indicadoras o fun-- ciones reales de variables reales.

Tenemos entonces que si existen otros efectos (diferentes al $\{\beta\}$) que no son objeto de estudio, pero que producen variaciones sistemáticas preponderantes, se llamarán $z_g, \ldots, z_r$. Por lo tanto el concepto de bloqueo implica expresar dichos efectos explícitamente en el modelo.

Cuando las $z_g, \ldots, z_r$ representan condiciones de - factores cualitativos se denominan bloques. En el caso de -

que algunas z sean valores de factores cuantitativos se tendrá un modelo de covarianza (como se vió en la subdivisión 3 de la sección II.3) donde el objeto de introducir las covariables -las z cuantitativas- es la reducción de la magnitud de los errores en el modelo. En los dos casos las z se introducen para reducir la variación de los ε_i. Luego, al incorporar las variables z de bloqueo o covarianza en (16) se tiene

$$y_{ij k \ldots m_{\beta} \ldots r} = \beta_0 + \beta_j x_j + \ldots + \beta_m x_m + \beta_\beta z_\beta + \ldots + \beta_r z_r + \varepsilon_i^*$$

en donde los ε_i^* tendrán fluctuaciones aleatorias más pequeñas que en el modelo dado por (16), principalmente por tratarse ahora de poblaciones con menor grado de generalidad, ya que al definirlas se han especificado un mayor número de factores, es decir, a los factores $\{\beta\}$ que nos interesan les agregamos los de bloqueo (o covarianza).

El concepto de bloques al azar surgió al considerar que todos los tratamientos* aparecieran una vez en cada bloque donde los elementos de la población en estudio fueran relativamente homogéneos, para asegurar que todos los tratamientos tuvieran la oportunidad de mostrar sus resultados dentro de cada bloque. Los tratamientos son asignados al azar a los elementos poblacionales de cada bloque.

---

*Conjunto particular de condiciones experimentales que deben imponerse a la población en estudio dentro de los confines del diseño seleccionado.

Para ilustrar esto, supóngase que se van a compa-- rar seis variedades de avena con respecto a sus rendimientos disponiéndose para tal efecto de 30 parcelas experimentales. Sin embargo, hay evidencias de que existe una tendencia en - la fertilidad de norte a sur, siendo más fértiles las parce- las del norte. De acuerdo con esto, parece razonable agru-- par a las parcelas en cinco bloques de seis parcelas cada u- no, de manera que el primer bloque contenga las seis parce-- las más fértiles, el siguiente grupo contendrá las siguien-- tes parcelas más fértiles y así sucesivamente hasta el quin- to bloque (más al sur) que contendrá a las seis parcelas me- nos fértiles. Una vez hecho esto, se asignarán al azar las seis variedades a las parcelas de cada bloque, haciendo una nueva aleatorización en cada bloque.

Fisher propuso el modelo matemático para el diseño en bloques al azar de la siguiente manera:

$$y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

con $i = 1,2,\ldots,t$ y $j = 1,2,\ldots,r$; siendo

y_{ij} medición (que puede ser rendimiento, incre- mento de peso, número de adultos, etc.) del tratamiento i en el bloque j ,

μ efecto de la media general,

τ_i efecto del tratamiento i , que se expresa con valores positivos para tratamientos con valores promedios superiores a la media ge- neral, y con valores negativos para trata--

mientos con promedios menores a la media general del experimento; o sea $\tau_i = \mu_i - \mu$ en donde μ_i es la media de la población dada por el tratamiento i-ésimo, que puede -- ser un nivel fijo de un factor o una combinación de niveles de varios factores en estudio,

β_j efecto del bloque j^o que se expresa similarmente a τ_i con $\beta_j = \mu_j - \mu$, donde μ_j es la media de la población de mediciones - que se pueden efectuar en el bloque que representa constancia de algunos factores,

ε_{ij} error aleatorio que surge por el efecto conjunto de todos los factores no controlados en el diseño y que causan heterogeneidad en las poblaciones, ε_{ij} tienen una distribu--- ción normal con independencia y homogenei-- dad de varianzas lo cual se verifica por - tratarse del estudio de poblaciones con el mismo grado de generalidad.

Para explicar el concepto de ortogonalidad diremos que: como se pretende que todos los tratamientos aparezcan en todos los bloques, la media aritmética de las mediciones en un tratamiento acumulará los efectos de bloque de igual - forma para cualquier tratamiento, de manera que la diferen-- cia entre medias de tratamiento no arrastra efectos de blo-- que. Esto puede expresarse como sigue: sin pérdida de generalidad supongamos los tratamientos 2 y 5 como un ejemplo. - Sea $\overline{y}_{2\cdot}$ = media de tratamientos No. 2 = $\dfrac{y_{21} + y_{22} + \ldots + y_{2j} + \ldots + y_{2n}}{n}$

llamémosle ecuación [17]. Según el modelo

$$\begin{aligned} y_{21} &= \mu + \tau_2 + \beta_1 + \varepsilon_{21} \\ y_{22} &= \mu + \tau_2 + \beta_2 + \varepsilon_{22} \\ &\vdots \\ y_{2j} &= \mu + \tau_2 + \beta_j + \varepsilon_{2j} \\ &\vdots \\ y_{2n} &= \mu + \tau_2 + \beta_n + \varepsilon_{2n} \end{aligned} \qquad [18]$$

Sustituyendo las ecuaciones [18] en la [17] se obtiene

$$\bar{y}_{2\cdot} = \frac{n\mu + n\tau_2 + \sum_{j=1}^{n} \beta_j + \sum_{j=1}^{n} \varepsilon_{2j}}{n}$$

$$\bar{y}_{2\cdot} = \mu + \tau_2 + \frac{\sum_{j=1}^{n} \beta_j}{n} + \frac{\sum_{j=1}^{n} \varepsilon_{2j}}{n} \qquad [19]$$

Trabajando análogamente para $\bar{y}_{5\cdot}$, se tiene

$$\bar{y}_{5\cdot} = \mu + \tau_5 + \frac{\sum_{j=1}^{n} \beta_j}{n} + \frac{\sum_{j=1}^{n} \varepsilon_{5j}}{n} \qquad [20]$$

Por lo tanto, la diferencia entre las medias del tratamiento 2 y del 5 será, utilizando las ecuaciones [19] y [20], como sigue

$$\begin{aligned} \bar{y}_{2\cdot} - \bar{y}_{5\cdot} &= \mu + \tau_2 + \frac{\sum_{j=1}^{n} \beta_j}{n} + \frac{\sum_{j=1}^{n} \varepsilon_{2j}}{n} - \mu - \tau_5 - \frac{\sum_{j=1}^{n} \beta_j}{n} - \\ &\quad - \frac{\sum_{j=1}^{n} \varepsilon_{5j}}{n} \\ &= \tau_2 - \tau_5 + \frac{\sum_{j=1}^{n} \varepsilon_{2j}}{n} - \frac{\sum_{j=1}^{n} \varepsilon_{5j}}{n} \end{aligned}$$

Lo cual nos lleva a afirmar que la diferencia de medias entre los tratamientos 2 y 5 no contiene efectos de bloque, -- sino únicamente los efectos de los tratamientos 2 y 5, más - la diferencia de los promedios de los errores presentes en

las parcelas (o de las subdivisiones que se trate) con tratamientos 2 y 5. Esto se conoce como ortogonalidad entre bloques y tratamientos.

Cuando el número de tratamientos por probarse es grande, de modo que las unidades experimentales varían en alto grado dentro del bloque, entonces se puede prever que, al poner todos los tratamientos en un bloque completo grande que los contenga, se tendrá una heterogeneidad muy fuerte entre las unidades experimentales, dentro del bloque completo (con todos los tratamientos). En este caso, lo que se propone es considerar bloques más pequeños a pesar de que no contengan a todos los tratamientos (bloques incompletos), pero buscando siempre homogeneidad de las unidades experimentales dentro de los bloques incompletos.

III. APLICACIONES

Las primeras aplicaciones de la Estadística se limitaban únicamente a determinar el punto donde la tendencia general era evidente (si es que existía), de una gran cantidad de datos observados. Al mismo tiempo, en muchas ciencias se hizo énfasis de que en lugar de hacer estudios individuales, deberían hacerse estudios del comportamiento de grupos de individuos. Los modelos estadísticos satisficieron tal necesidad pues, los grupos concuerdan consistentemente con el concepto de población. Así pues, podemos utilizar modelos estadísticos lineales* en los siguientes campos.

Por ejemplo, el agricultor siempre está enfrentado al problema de mantener un alto nivel de productividad en sus cosechas. Para ayudarlo, el agrónomo efectuará un sinnúmero de experimentos a fin de determinar la diferencia de utilidades entre las diferentes variedades de productos agrícolas, efectos de los distintos fertilizantes, y del mejor método de cultivo. Con base a los resultados de estos experimentos, el agrónomo espera hacer recomendaciones exactas y útiles al agricultor.

Relacionada con la agronomía está el cultivo de plantas donde se pretende el perfeccionamiento de variedades

*Siempre y cuando se satisfagan las condiciones que se han mencionado.

o híbridos. La elección del diseño depende de los conceptos que implica; la uniformidad del suelo, la exactitud y precisión de las estimaciones particulares juzgadas necesarias para obtener los resultados requeridos; el tiempo, recursos y dinero disponibles y posiblemente otros factores. Los datos recopilados son entonces analizados de acuerdo con el plan del experimento, el cual fue diseñado para hacer posible una comparación adecuada entre las técnicas por probar.

El modelo estadístico empleado debe por supuesto, tener una iterpretación lógica del proceso biológico en consideración y de la manera en que el experimento se ha efectuado a fin de que se tengan resultados útiles.

Otras áreas de investigación en las que se hace buen uso de los modelos estadísticos lineales son la cría de aves de corral, cría de animales y nutrición de animales donde uno de los usos más importantes es la separación de los efectos del medio ambiente y hereditarios. Por ejemplo, en el campo de la nutrición animal, se han previsto muchos experimentos para descubrir la importancia de las vitaminas en las diferentes fases de la producción animal. En tales investigaciones se eligen varios grupos de animales, tan homogéneos como sea posible, para la investigación. Tales grupos homogéneos se forman, usualmente, por consideración de criterios tales como edad, peso, sexo, herencias, vigor y nutrición previa. Un grupo ya revisado se escoge y alimenta a una ración normal. Los grupos restantes son alimentados a

diferentes niveles de la vitamina en cuestión, uno de ellos se alimenta con una ración que contenga más vitaminas que la ración normal y otro grupo se alimenta con una ración que contenga muy poca o nada, de la vitamina en cuestión. El resto de los grupos son alimentados con raciones comprendidas entre los extremos. Los animales se conservan durante un cierto periodo bajo la alimentación arbitrariamente elegida y el investigador registra datos tales como incremento diario de peso, economía obtenida, viabilidad, etc. Si el experimento ha sido diseñado de acuerdo con los principios estadísticos establecidos, pueden obtenerse conclusiones de gran valor para el ganadero.

Actualmente en la Ingeniería han sido utilizados modelos estadísticos lineales en temas como el calor transferido por unidad de tiempo a través de un material aislante, control de producción, tolerancias en partes de máquinas, estudios de corrosión, control de calidad, y muchos otros problemas especializados de investigación y desarrollo.

En ciencias sociales, tales como Sociología, Demografía, Psicología, los modelos estadísticos también tienen una amplia aplicación. Probablemente, las comparaciones más importantes en economía de la producción se hacen cuando dos o más características se estudian o miden simultáneamente. Esto implica técnicas como regresión y correlación que son herramientas inestimables para el economista.

Ahora, pasamos a señalar los principales usos que

tienen los modelos estadísticos lineales dentro de la regresión y el diseño de experimentos.

III.1 Regresión.

III.1.1 Predicción.

Como se ha visto los modelos de regresión son valiosos para describir el tipo de asociación entre la variable dependiente y y la(s) variable(s) independiente(s) x_1, x_2, ..., x_n. En este caso lo que se persigue es resumir la tendencia de los datos y encontrar la forma de asociación de las variables. Es importante recalcar que en dicho uso no se pretende establecer relaciones causales en el sentido de que los valores de las x's produzcan cambios en los valores de y; se trata únicamente de indicar la asociación entre las variables y cual es la forma de dicha asociación.

La predicción de la variable y_i sólo será válida si en el período de predicción se mantienen constantes las condiciones con las que se probó el modelo, si estas se alteran será necesario hacer las consideraciones pertinenetes para que funcione el modelo original, o bien crear un nuevo modelo.

El valor de la predicción dependerá de la magnitud de su posible error, por ello, para una buena predicción es necesario que la variabilidad de los errores ε_i sea lo más pequeña posible.

Así pues, podemos desear predecir: el peso en Kg.

de un individuo que tiene cierta altura en cm., el número de hijos que tendrán las familias con determinado ingreso, el monto total de préstamos sobre pólizas en cierta compañía de seguros durante el próximo año, la cantidad de óxido que se formará en la superficie de un metal calentado en un horno durante un intervalo especificado de tiempo a 200 grados C., o la medida de deformación de un anillo sometido a una fuerza de compresión de 1000 lbs., o bien, podemos predecir la demanda que tendrá dentro de un semestre cierta mercancía con un precio determinado.

III.1.2 Control.

Aunque existe la tendencia a considerar la garantía de calidad como un tema reciente, no hay nada nuevo sobre la idea básica de hacer un producto de calidad caracterizado por un alto grado de uniformidad. Durante siglos, los artesanos hábiles se han esforzado por hacer productos que se distinguieran por su calidad superior y, una vez que se lograba cierto grado de calidad, eliminar, hasta donde fuera posible, cualquier variabilidad entre sus productos que los hiciera anormales. Nótese que la palabra calidad cuando es usada técnicamente, se refiere a alguna propiedad medible o contable de un producto, tal como el diámetro externo de un rodamiento de bolas, la resistencia a la rotura de un hilo, el número de imperfecciones en una pieza de tela, la potencia de una droga, etc.

Un modelo de regresión puede servir para encontrar cuales son los valores de las x's que pueden optimizar, de acuerdo con algún criterio, los valores de la variable dependiente y. Para ello se establece la ecuación de regresión y se buscan los valores de las x's que controlen los cambios en y en el sentido deseado.

Es sorprendente como dos partes aparentemente idénticas, hechas bajo condiciones cuidadosamente controladas, de la misma fuente de materia prima y fabricadas sólo con diferencia de segundos por la misma máquina, pueden ser diferentes en muchos aspectos. En realidad, cualquier proceso de fabricación, aunque sea bueno, se caracteriza por un cierto grado de variabilidad que es de una naturaleza aleatoria y que no se puede eliminar completamente.

Cuando la variabilidad presente en un proceso de producción está limitada a la variación aleatoria, se dice que el proceso está bajo control estadístico. Tal estado se consigue buscando y eliminando todas las causas que originan variaciones de otra clase (que son las variaciones atribuibles) que pueden deberse a operarios poco entrenados, materias primas de baja calidad, ajustes indebidos de las máquinas, partes usadas, etc. Como los procesos de fabricación raramente se encuentran libres de este tipo de defectos, es importante tener algún método sistemático de detectar las desviaciones notables de un estado de control estadístico cuando estas se presentan. Para ello se emplean principalmente las cartas de control (ver referencia 13 pags. 519 a

527].

Daremos un ejemplo típico que nos ilustre la utilización del control dentro de la regresión. Supongamos que se desea encontrar la ración óptima de alimento -la más económica- que debe dársele a los puercos para obtener una mejor producción de carne. Para lograrlo se encuentra, por regresión, una función que ligue la carne producida y y las cantidades de varios alimentos x's. Luego entonces, en función de los costos de los alimentos y con la ayuda de programación lineal, se obtiene la ración óptima.

III.1.3 Calibración.

Estudiando las técnicas de medición, existe un problema muy importante, que consiste en determinar las relaciones entre mediciones de la misma cantidad realizadas por diferentes técnicas. A tal problema se le conoce como calibración. Hay dos tipos de calibración:

i) La calibración comparativa, que surge cuando se necesita estandarizar o graduar una técnica o instrumento de medición contra otra u otro, como por ejemplo la calibración de termómetros.

En la calibración de un número de instrumentos que proporcionan mediciones de un tipo similar, usualmente no hay una medida clásica (estándar) a la cual se deben referir aunque algunas veces algún instrumento puede definirse como el estándar. Las mediciones generalmente serán efectuadas a

especímenes de diferentes clases de material, cuyas propiedades puede esperarse que cubran el rango de valores sobre los cuales la calibración va a ser aplicada. Esta condición es esencial para la aplicación de cualquier método de estimación de las relaciones de calibración comparativa.

ii) La calibración absoluta, en donde se tienen dos técnicas: una tradicional y otra no tradicional. Se relacionan las mediciones tomadas de acuerdo a la técnica tradicional con las tomadas de acuerdo a la no tradicional, lo cual podemos expresarlo diciendo que y es una medición de una característica aleatoria que es fácil de medir y depende de una variable aleatoria difícil de medir x. Deseamos entonces efectuar sólo las mediciones fáciles y de éstas estimar las difíciles. Lo anterior puede hacerse de la siguiente manera: Se obtienen "n" pares de valores $(x_1, y_1), \ldots, (x_n, y_n)$ donde las x's y las y's son los valores observados de x y y sobre el mismo objeto; con estos "n" pares se hace la regresión de y sobre x. De ahí en adelante, al obtener una y^* futura, se utilizará la línea de regresión ajustada anteriormente para calibrar el valor o los posibles valores de x^* del o de los cuales pudo provenir y. Ejemplos:

-Se trata de la medición de la concentración de sodio en una muestra de materiales. La estimación de la concentración por medio de métodos tradicionales es precisa pero tediosa, mientras que la lectura del fotómetro es relativamente rápida. El objeto del experimento de calibración es

estimar la relación por medio de la cual las lecturas del fotómetro se pueden convertir en las mediciones de concentración de sodio que se hubieran obtenido si se hubieran utilizado los métodos tradicionales.

La población "A" en este caso está formada por diversos tipos de materiales, cada uno de los cuales tiene concentraciones de sodio diferentes. Se toma una muestra estadística de tamaño "n", donde cada elemento de dicha muestra es una muestra química de elementos de "A". A cada uno de los elementos de la muestra estadística, se le mide la concentración de sodio por medio del método tradicional, formándose con esto las x's ; después, midiendo para cada elemento de la muestra estadística la concentración de sodio por medio del fotómetro, se obtienen las y's. A continuación se efectúa la regresión de y sobre x . En base a esta regresión se puede estimar la medición de la concentración de sodio que se hubiera obtenido por métodos tradicionales, basandose exclusivamente en la medición obtenida por medio del fotómetro. Esto siempre y cuando el material sobre el que se estuvieron efectuando las mediciones pertenezca a la población "A", pues de otra manera no se tendría una base para hacer tal estimación.

-Los antropólogos al efectuar una excavación y encontrar restos humanos hacen estimaciones acerca de la edad del individuo. Para hacer estas estimaciones el antropólogo lleva a cabo mediciones y's de cráneos de individuos de edad

$x's$ conocida pertenecientes a la misma población estadística (tipo racial, grupo étnico, etc.), y entonces en base a estas parejas de datos $(x_1, y_1), \ldots, (x_n, y_n)$ efectúa una regresión de y sobre x que le permitirá estimar la edad desconocida x del individuo (perteneciente a la población estudiada) con cráneo de medida y_0 .

III.2 Diseño de experimentos.

III.2.1 Comparación de medias.

Como se ha visto, los modelos estadísticos lineales de diseño de experimentos tiene como objetivo principal comparar las medias de las poblaciones estudiadas bajo distintos tratamientos.

En efecto, lo que interesa en este caso es comparar, por ejemplo: las medias de producción de frijol en 28 parcelas utilizando 4 fertilizantes, o las medias en cuanto a la resistencia a la tensión en la producción de acero usando diferentes cantidades de carbón. Podemos interesarnos en la comparación de medias de producción de varias poblaciones generadas al considerar distintos tipos de procedimientos de operación dentro de las fábricas. O bien, la comparación de las medias del número de inversionistas de los bancos al variar las tasas de interés; en un hospital la comparación de medias de pacientes leprosos al ser tratados con distintos medicamentos.

Para hacer estas comparaciones necesitamos plantear hipótesis del tipo $\mu_1 = \mu_2 = \ldots = \mu_n$ y hacer un análisis de varianza.

III.2.2 Estudio de efectos.

Una aplicación muy importante en los diseños experimentales es la que se refiere al estudio de las relaciones entre varios factores, ya sea de tipo cualitativo o cuantitativo, que sirven como criterio para definir las poblaciones bajo estudio y sus respectivas medias. Para ello una herramienta muy útil es la verificación de la hipótesis de no interacción entre los factores que intervienen directamente para definir la población bajo estudio. Por ejemplo, al estudiar la depresión de sujetos drogadictos, se podrá conocer si el efecto de la droga se modifica al tratarse de hombres o de mujeres.

Otro aspecto de interés surge al investigar el patrón de cambio que siguen las medias de las 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, cúbico, etc.

IV. CONCLUSIONES

A lo largo de este trabajo se ha visto que el modelo $y_i = \mu + \varepsilon_i$ es útil para representar las ocurrencias de los valores de una población, y el valor de μ depende de las especificaciones que tenga la misma. También la variabilidad presente en los $\{\varepsilon_i\}$, que se transmite a las $\{y_i\}$, se determina por dichas especificaciones. Esta variabilidad depende del tipo de fenómeno estudiado y del grado de generalidad de la población, esto es, de los factores que no quedan especificados al definirla. El modelo de regresión lineal múltiple es una extensión de los conceptos anteriores y tiene un grado máximo de flexibilidad. Con el modelo lineal general (7) definimos todos los modelos lineales. Los supuestos básicos que lo fundamentan son: relación funcional correcta (aditividad), homocedasticidad, independencia de errores y normalidad. Con este modelo podemos representar un espacio de "n" dimensiones ("n" variables $z\omega$). Si las $\{z\omega\}$ se les introduce como variables indicadoras de la presencia o ausencia de modalidades cualitativas de los efectos que definen a las poblaciones, se obtienen los modelos de diseños experimentales. Cuando las poblaciones en estudio difieren por los niveles o categorías de un factor, el cual está determinado como constante al definir las poblaciones, tendre-

mos modelos de diseños con un criterio de clasificación. Si se consideran poblaciones con categorías o niveles de dos o más factores tendremos los llamados diseños con dos (o más) criterios de clasificación. Para reducir la magnitud de los errores experimentales utilizamos los diseños experimentales en bloques al azar.

Los modelos estadísticos lineales antes mencionados son de gran utilidad, independientemente que se apliquen en el análisis de fenómenos físicos, en el estudio de mediciones educacionales, en el estudio de datos provenientes de experimentos biológicos, o del análisis cuantitativo del material en Economía. El agrónomo, el químico, el biólogo -- (por mencionar algunos) tratan de eliminar las diversas fuentes de los factores que afectan el fenómeno que estudian. - Con todo, muchas perturbaciones (los ε_i) están siempre presentes; entonces, éstas se tratarán de reducir al mínimo de manera que las afirmaciones que se hagan acerca del fenómeno en estudio sean razonables..

Para que los modelos estadísticos lineales sean utilizados con ventaja, se propone:

1) Que la persona que desee utilizarlos esté consciente de las suposiciones básicas en que se fundamentan -- los modelos estadísticos lineales y, además, que sea versado en la materia objeto del campo en el que la investigación va a realizarse para poder decidir si es conveniente la utilización de dichos modelos en un fenómeno dado.

2) Si la decisión ha sido utilizar modelos estadísticos lineales, se deberá verificar si las suposiciones básicas planteadas son representativas en el estudio de cierto fenómeno. Se debe proceder con criterio científico, o sea, mediante modelos provisionales que deberán ser confrontados con las observaciones prácticas del mundo real; el modelo se irá modificando hasta llegar a ser satisfactorio desde el punto de vista práctico. En él las suposiciones deberán cumplirse aproximadamente y producir errores lo suficientemente pequeños como para garantizar su uso práctico.

3) Por último, es importante saber cómo organizar la información mediante medios de representación gráfica y tabular, y tener conocimiento de cómo establecer rutinas económicas y funcionales para el manejo y computación de los datos.

BIBLIOGRAFIA

1. Abad de S., A. y Servín A., L., "Introducción al Muestreo", Limusa, S. A., México, 1978.

2. Aranda O.,Fco. J. y Valencia R., Gvo. J., "Un estudio en la teoría de la calibración estadística", Tesis profesional de Actuario, Fac. de Ciencias, UNAM, 1974.

3. Cochran W., G., "Técnicas de Muestreo", CECSA, México, 1976.

4. Cochran W., G. y Cox G., M., "Diseños Experimentales", Trillas,S.A., México, 1978.

5. Dixon W., J. y Massey F., J., "Introducción al Análisis Estadístico", McGraw-Hill, México, 1979.

6. Graybill F.,A., "An Introduction to Linear Statistical Models" Vol. 1, McGraw-Hill Book Co., New York, 1961.

7. Méndez R., I., "Comparación de Medias en Poblaciones", Comunicaciones Técnicas, IIMAS, UNAM,1978.

8. Méndez R., I., "Lineamientos Generales para la Planeación de Experimentos", Comunicaciones Técnicas, IIMAS, UNAM, 1978.

9. Méndez R., I., "Modelos Estadísticos Lineales", Focavi-Conacyt, 1976.

10. Miller, I. y Freund, J. E., "Probabilidad y Estadística para Ingenieros",Reverté S.A., México, 1973.

11. Mood A., M. y Graybill F., A., "Introducción a la Teo - ría de la Estadística", Aguilar, S.A., México, 1976.

12. O'Reilly F., J., "Predicción en Regresión Lineal", Comunicaciones Técnicas, IIMAS, UNAM, 1979.

13. Ostle, Bernard, "Estadística Aplicada", Limusa-Willey, México, 1977.

M-0037496