

INT-2540

Gonzalo Garcia.

PRELIMINAR

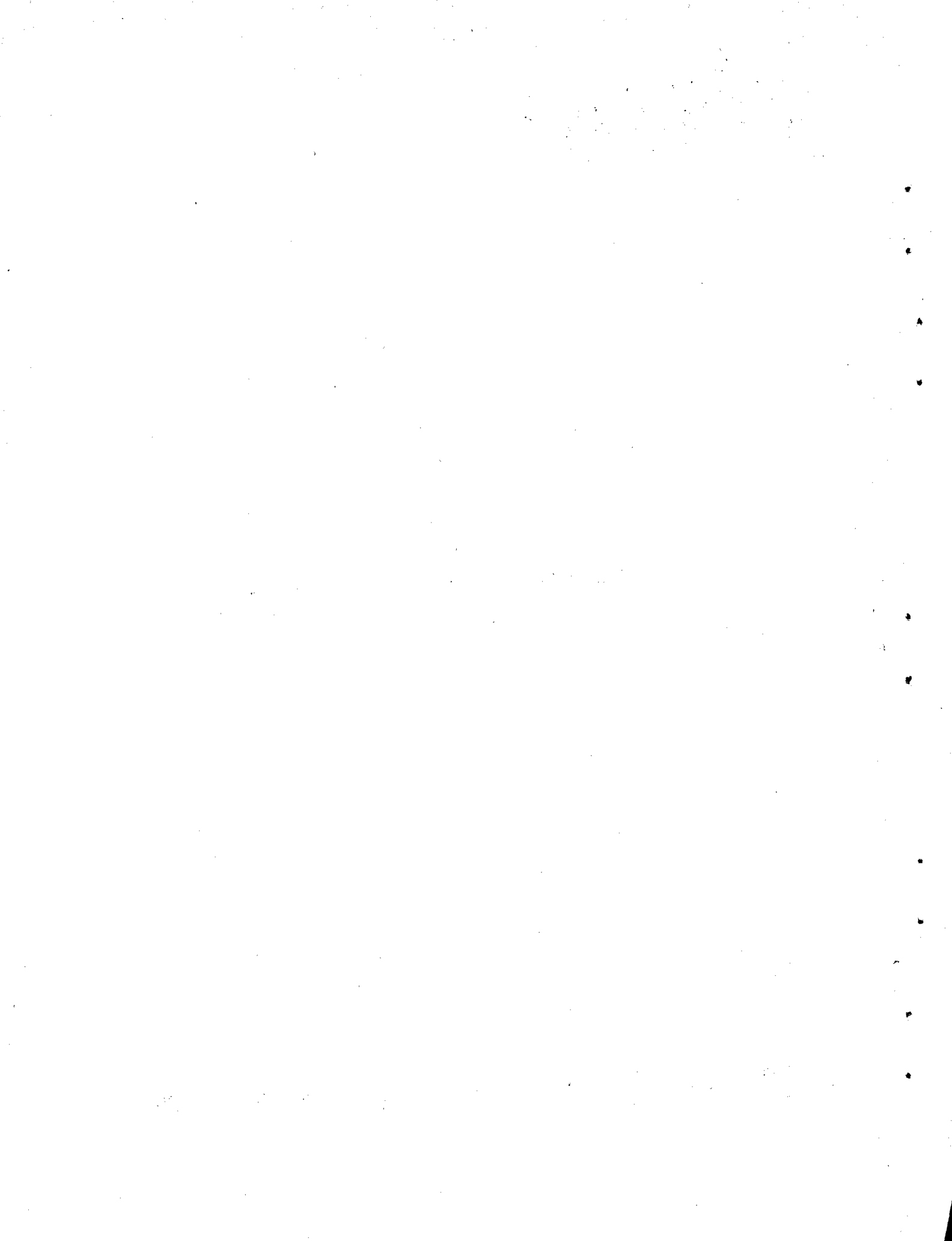
Instituto Latinoamericano de
Planificación Económica y Social
Santiago, diciembre de 1967

CURSO DE ESTADISTICA BASICA PARA PROGRAMACION*

Parte I



* Programa de Capacitación. Profesor, señor Arturo Núñez del Prado B.



INDICE

I. ESTADIGRAFOS DESCRIPTIVOS

	<u>Páginas</u>
A. DISTRIBUCION DE FRECUENCIA	1
1. Generalidades	1
2. Representación gráfica	4
B. ESTADIGRAFOS DE TENDENCIA CENTRAL	8
1. Media aritmética	8
a) Propiedades	9
b) Métodos abreviados de cálculo	11
2. Mediana	13
a) Variable discreta	14
3. Valor modal	18
a) Variable discreta	19
b) Variable continua	19
4. Media geométrica	22
5. Media armónica	23
C. EVALUACION DE LOS ESTADIGRAFOS DE TENDENCIA CENTRAL	24
D. ESTADIGRAFOS DE DISPERSION	25
1. Recorrido de la variable	26
2. Recorrido intercuartílico	26
3. Varianza	26
a) Para datos no agrupados	26
b) Para datos agrupados	26
c) Propiedades de la varianza	27
d) Componentes de la varianza	28
e) Métodos abreviados de cálculo	31
4. Coeficiente de variabilidad	33
E. UTILIZACION DE INDICADORES EN LA PROGRAMACION	35

II. NUMEROS INDICE

A.	<u>El problema general</u>	36
B.	<u>Clases de números indice</u>	36
C.	<u>Fórmulas de cálculo</u>	36
D.	<u>Pruebas sobre los números indice</u>	43
	1. Prueba de reversión de factores	43
	2. Prueba de reversión temporal	44
	3. Prueba circular	44
E.	<u>Base de un número indice</u>	46
F.	<u>Utilización de los números indice</u>	48
	1. La deflactación	48
	2. El deflactor implícito del Producto Bruto	53
	3. La proyección sobre la base de índices de cantidad	58
G.	<u>Indices de comercio exterior</u>	58
	1. Indices de precios	58
	2. Indices de cantidad	59
H.	<u>Algunos indicadores económicos</u>	60
	1. Indice de la relación de términos del intercambio	60
	2. Producto e ingreso bruto	61
	3. Capacidad para importar	63
	4. Tipo de cambio de paridad	63
	5. Transferencias implícitas	65
I.	<u>Etapas de la construcción de numerosos indices</u>	67
	1. Objetivo del indice	67
	2. Determinación de la estructura de consumo	67
	3. Selección de artículos	68
	4. Formas de valuación	68
	5. Variaciones de calidad de los bienes y servicios	68
	6. Base del indice	68
	7. Elección de los métodos de cálculo	69

FINALIDAD DEL CURSO

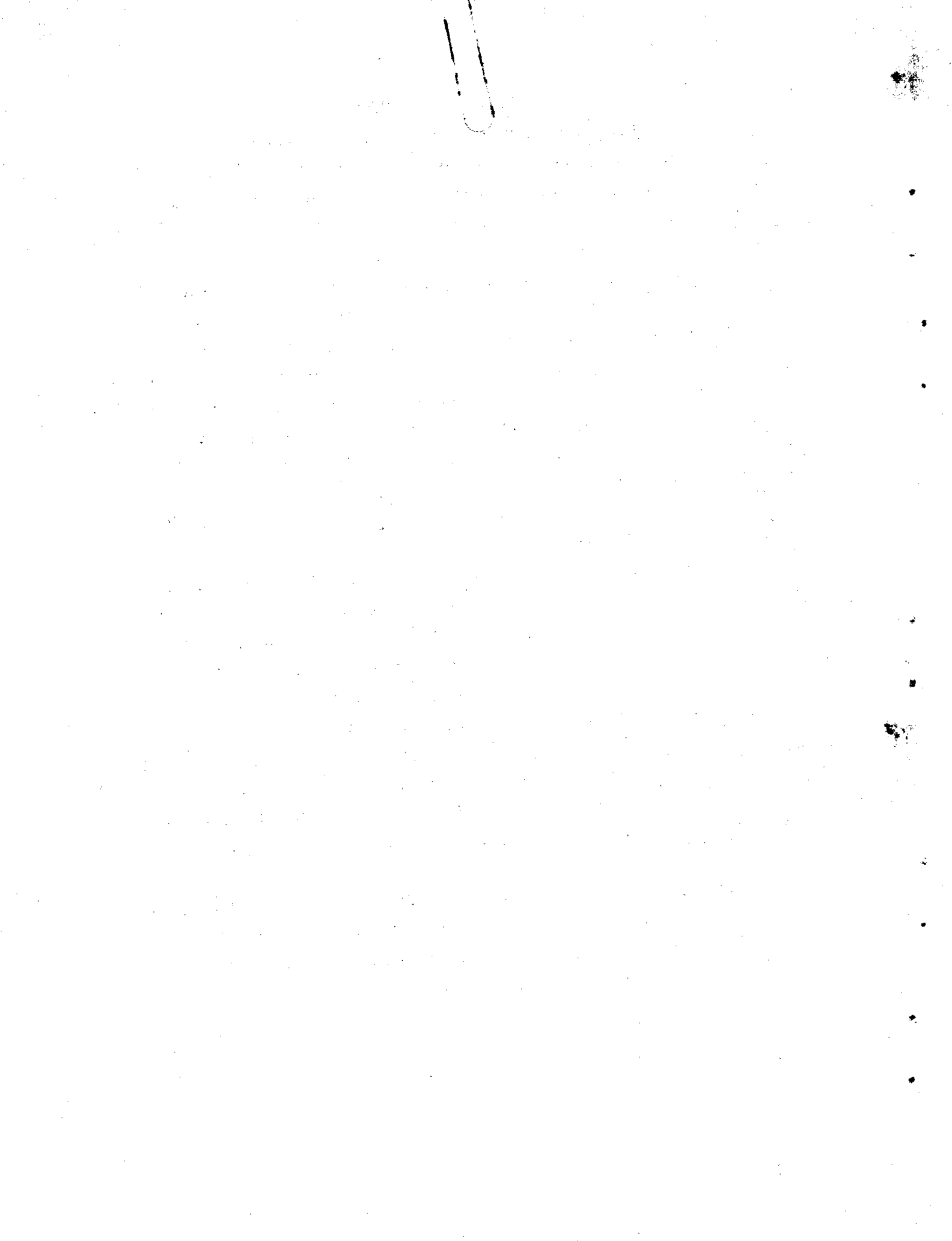
El programa ha sido diseñado para capacitar a los participantes del Curso Básico en materias estadísticas de uso cotidiano en la planificación y la investigación económica general. Por otra parte también se ha contemplado las necesidades de instrumental estadístico que tienen otras asignaturas dentro del Programa de Capacitación.

Principalmente lo que interesa es que el planificador pueda utilizar racionalmente el aparato estadístico y pueda entenderse con estadísticos y econométricos en estudios más profundos. Sería pretencioso pensar que después de un curso de 24 horas de clase e igual número de horas de seminarios y ejercicios prácticos, el alumno haya alcanzado una cierta especialidad en este campo. No, el objetivo es proporcionarles un conjunto de conceptos que garantizarán al planificador detectar las posibilidades y limitaciones de la utilización de indicadores estadísticos, interpretando cabalmente y en su justa dimensión los estadígrafos de uso más frecuente.

Se trata de un curso intensivo, donde el trabajo principal recae en el propio alumno. Muchos conceptos y materias no serán tratados en forma teórica, si no a través de ejercicios prácticos en los seminarios.

Para facilitar el trabajo de los alumnos de manera que dispongan de un resumen de las distintas materias que se presenten a lo largo del curso, se han preparado estos apuntes. Apuntes que constituyen en buena parte resúmenes de otros textos como los Apuntes de Estadística del profesor Pedro Vuskovic, el Curso General de Estadística del profesor Enrique Cansado, Introducción a la Estadística de G.U. Yule y M.G. Kendall, y Estadística General Aplicada de F. Croxton y D. Cowden.

Estos apuntes, no son en absoluto suficientes. El alumno, para conseguir una eficiente capacitación, deberá consultar los textos citados y realizar con toda conciencia y dedicación los ejercicios y seminarios que constituyen la médula del curso.



I. ESTADIGRAFOS DESCRIPTIVOS

A. DISTRIBUCIONES DE FRECUENCIA

1. Generalidades

Un conjunto de datos o masa estadística, es susceptible de ser resumida y clasificada de acuerdo a criterios convenientes. Sea que las informaciones provengan de censos o de muestras relativamente grandes, siempre será ventajoso para el análisis, ya que difícilmente podrán obtenerse conclusiones válidas de una masa estadística no clasificada.

Los tipos de variables que preocuparán la atención del curso serán los siguientes:

- a) Variables cardinales: susceptibles de medición cuantitativa, las que comprenden a:
 - i) continuas: variables que pueden tomar cualquier valor dentro de un intervalo (ingresos, estaturas, distancias, etc.);
 - ii) discretas: variables que sólo toman algunos valores dentro de un intervalo (número de hijos por familia, número de accidentes de tránsito por día, etc.);
- b) Variables ordinales: sólo susceptibles de ordenación pero no de medición cuantitativa (grado de cultura de una persona: muy culta, regularmente culta, poco culta, inculta).

Para cada uno de estos tipos de variables, un conjunto de observaciones, puede dar origen a una distribución de frecuencias. Debe entenderse ésta, como un cuadro o tabla resumen de los datos originales.

En el caso de variables continuas será necesario fijar intervalos de frecuencia para llegar a un efectivo resumen de la información original. El punto medio de cada intervalo se denominará marca de clase y constituirá el valor representativo de cada intervalo. El número de observaciones que correspondan a cada intervalo se denominarán frecuencias absolutas.

/Una tabla

Una tabla de distribución de frecuencia para variable continua y sus símbolos correspondientes, se representa de la siguiente forma:

<u>Ingresos de Profesionales</u>		<u>Número de Profesionales</u>	
<u>Intervalos</u>	<u>Marcas de clase</u>	<u>Frecuencias Absolutas</u>	
Y'_{i-1}	Y'_i	Y_i	n_i
Y'_0	Y'_1	Y_1	n_1
Y'_1	Y'_2	Y_2	n_2
Y'_2	Y'_3	Y_3	n_3
\vdots	\vdots	\vdots	\vdots
Y'_{m-1}	Y'_m	Y_m	n_m

donde: $Y_i = \frac{Y'_{i-1} + Y'_i}{2}$ marca de clase

$n = \sum_{i=1}^m n_i$: número de observaciones

$c_i = Y'_i - Y'_{i-1}$: amplitud del intervalo

Estas tablas pueden ser de amplitud constante o de amplitud variable, dependiendo de los valores que tome c_i .

Cuando se trata de variable discreta o discontinua, la tabla de distribución de frecuencias es de la forma siguiente:

Y_i	n_i
Y_1	n_1
Y_2	n_2
Y_3	n_3
\vdots	\vdots
Y_m	n_m

/Cabe destacar

Cabe destacar que cuando se tiene un número grande de valores distintos de la variable, para abreviar el trabajo, con cierta arbitrariedad y con alguna pérdida de precisión, puede tratarse como una variable continua, formando intervalos de clase.

Por último, en el caso de variables no mensurables, la tabla en cuestión, tendrá una forma como la siguiente:

<u>Variabes</u>	<u>Frecuencia</u>
Característica A	n_A
Característica B	n_B
⋮	⋮
Característica Z	n_Z

El estudiante advertirá que las tablas de distribución de frecuencias, facilitan enormemente el análisis. Es muy ventajoso disponer de informaciones clasificadas en intervalos o en valores específicos de la variable, ya que de esta manera es posible obtener conclusiones gruesas acerca de la variable que se investiga.

Hay que advertir que en gran medida depende del tipo de análisis que se está realizando

Respecto de las frecuencias, es posible y generalmente útil, presentarlas en términos relativos, calculando la proporción que del total de observaciones, corresponde a cada intervalo o marca de clase. Se denominan frecuencias relativas, y se simbolizarán por h_i :

$$h_i = \frac{n_i}{n}$$

Tanto las frecuencias absolutas como las relativas son susceptibles de acumulación respecto de los intervalos o marcas de clase. Las frecuencias absolutas acumuladas se simbolizarán por N_i y se definen:

$$N_i = \sum_{j=1}^i n_j$$

/Las frecuencias

Las frecuencias relativas acumuladas se simbolizarán por H_i y se definen:

$$H_i = \sum_{i=1}^i h_i$$

En general este tipo de frecuencias se acumulan en sentido creciente de la variable y una frecuencia acumulada N_i indica el número de casos u observaciones en que la variable toma valores a lo sumo iguales a Y_i en el caso de variable discreta y a Y'_i en el caso de variable continua. Sin embargo, para ciertos análisis, también es necesario acumular en sentido inverso; de ahí que se hable de frecuencias acumuladas hacia arriba o hacia abajo.

2. Representación gráfica

En general, la representación gráfica de una tabla de distribución de frecuencias, permite visualizar con mayor claridad algunas características de la masa de datos que se investiga. Resulta bastante más fácil transmitir conclusiones a personas no habituadas a la interpretación de distribuciones de frecuencia, cuando se utilizan gráficos estadísticos.

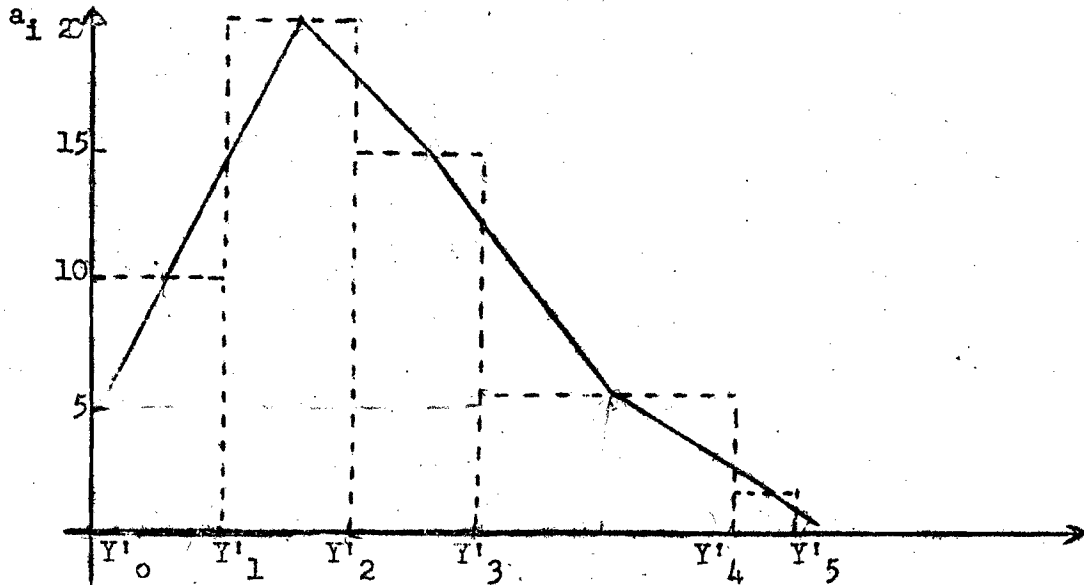
- a) Representación gráfica de variable continua. Utilizando un par de ejes coordenados, en el eje de las abscisas se representará la variable que se estudia, en tanto que en el eje de las ordenadas, se representarán las frecuencias correspondientes. Recuérdese que en este tipo de variables, la frecuencia corresponde a un intervalo y por este hecho se representa mediante una superficie. Por medio de un ejemplo se aclararán estas ideas: Admitase que se tiene la siguiente tabla correspondiente a las edades de los participantes del curso básico:

<u>Edades</u>		<u>Alumnos</u>	<u>Amplitud de Intervalo</u>
Y'_{i-1}	Y'_i	n_i	C_i
18	22	10	4
22 23	26	20	4
26 27	30	16	4
30 31	38	12	8
38 39	40	1	2

/En vista

En vista de que la amplitud más frecuente es 4, puede elegirse ésta como amplitud unitaria; así el cuarto intervalo tendrá dos veces la amplitud unitaria elegida y el quinto intervalo tendrá la mitad de dicha amplitud. La representación gráfica se hará de la siguiente manera:

Gráfico N° 1



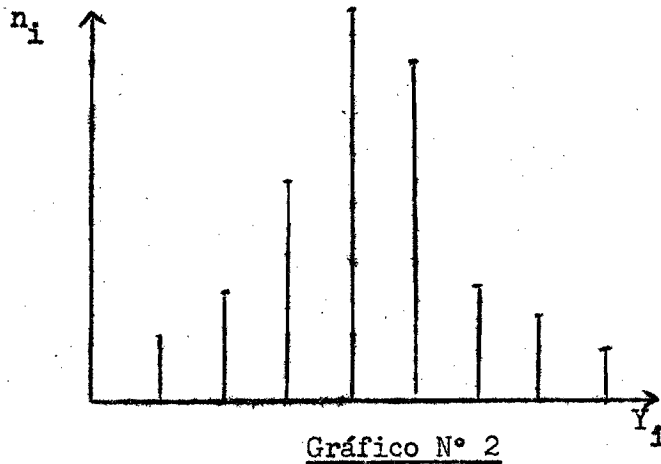
En el gráfico, para calcular la altura de cada rectángulo, se planteó la relación de superficie siguiente:

$$\begin{aligned} \text{superficie} &= \text{base} \times \text{altura} \\ n_i &= c'_i \cdot a_i \end{aligned}$$

donde c'_i es la amplitud unitaria elegida con objeto de diseñar un gráfico adecuado. Perfectamente pudo haberse trabajado con las amplitudes originales, pero habría sido algo más laborioso.

Este tipo de gráficos recibe el nombre de histogramas y la línea quebrada que une los puntos medios de los lados superiores de los rectángulos recibe el nombre de polígono de frecuencias.

b) Representación gráfica de variable discreta. En este caso la frecuencia correspondiente a cada valor de la variable estará representada por una barra vertical.



Naturalmente que se pueden construir, en forma similar, gráficos que relacionen la variable con cualquiera de los tipos de frecuencias que se han visto, relativas, acumuladas, etc.

A continuación se presentará un ejemplo donde se seguirán todos los pasos para llegar a una tabla de distribución de frecuencias completa. Supóngase que se dispone de las siguientes informaciones acerca de los sueldos de los obreros de una fábrica (en dólares por mes):

68	48	53	73	100	80	40	55	65	95	85	35	110	120	60
90	70	40	80	100	70	50	55	70	65	45	80	60	90	50
55	60	30	110	110	90	70	60	45	65	80	85	90	68	72
50	40	45	90	105	108	35	45	50	70	82	84	66	38	48

Una de las primeras decisiones es determinar el número de intervalos que tendrá la tabla. Para ello es necesario considerar el objetivo que se persigue con el estudio de la variable, qué tipo de diferencias o agrupamientos interesaría conocer. Por otra parte, es necesario considerar el recorrido de la variable, es decir, el menor y mayor valor entre los datos que se analizan. Por último, el número de observaciones, de manera que los diferentes intervalos tengan frecuencias en alguna medida significativas. Cuando existan valores escasos muy alejados de lo que podría llamarse una concentración central, puede optarse por dejar los intervalos extremos abiertos. Supóngase que tomando en cuenta las consideraciones

/anteriores se

anteriores se decide clasificar los datos originales en 9 intervalos de amplitud constante. Dado que la diferencia entre los valores extremos (30 y 120) es de 90, la amplitud de los intervalos será igual a 10.

Y'_{i-1}	Y'_i	Observaciones	n_i	h_i	N_i	H_i	N_i^*	H_i^*	Y_i
30,0	40	### //	7	7/60	7	7/60	60	60/60	35
40,1	50	### ####	10	10/60	17	17/60	53	53/60	45
50,1	60	### ///	8	8/60	25	25/60	43	43/60	55
60,1	70	### #### /	11	11/60	36	36/60	35	35/60	65
70,1	80	### /	6	6/60	42	42/60	24	24/60	75
80,1	90	### ////	9	9/60	51	51/60	18	18/60	85
90,1	100	///	3	3/60	54	54/60	9	9/60	95
100,1	110	###	5	5/60	59	59/60	6	6/60	105
110,1	120	/	1	1/60	60	60/60	1	1/60	115
			60	1					

(*) acumuladas "hacia arriba"

Para evitar situaciones ambiguas, cuando se presentan observaciones con valores de la variable que corresponden a los límites de los intervalos, puede seguirse el criterio planteado en el ejemplo de agregar un decimal a la columna de límites inferiores, aunque esto sólo tenga utilidad para fines de clasificación de las observaciones, ya que en los cálculos que posteriormente se tratará, tales decimales son despreciados.

Con lo visto hasta el momento, es posible realizar los primeros análisis de un conjunto dado de datos. Tanto la representación gráfica como la tabulación de las distintas clases de frecuencias ayudan a ensayar los primeros juicios. Es necesario insistir sobre la necesidad de tomar en cuenta, en todas las decisiones respecto de la tabla de frecuencias, la naturaleza del fenómeno que se investiga; sobre todo en lo que se refiere al número de intervalos y a sus amplitudes: constantes o variables. Naturalmente que, una vez clasificados los datos originales, será preciso realizar análisis con mayor profundidad, utilizando instrumentos estadísticos que se detallarán en las próximas páginas.

B. ESTADIGRAFOS DE TENDENCIA CENTRAL

Una vez que se ha conseguido la clasificación de los datos originales donde se destacan sus características más esenciales, será preciso calcular un conjunto de indicadores que caractericen en forma un tanto más precisa, la distribución que se está estudiando. Interesa en primer término disponer de estadígrafos que representen valores centrales en torno de los cuales se agrupan las observaciones. En general se los designa como promedios, y son de extraordinaria utilidad tanto en el análisis de una distribución, como en la comparación entre distribuciones.

1. Media aritmética. Es sin duda, el estadígrafo más utilizado, sobre todo en la cuantificación de variables económicas. Se simbolizará por \bar{Y} ó $M [Y_i]$, y se definirá como:

$$\bar{Y} = M [Y_i] = \frac{\sum_{i=1}^m Y_i n_i}{n}$$

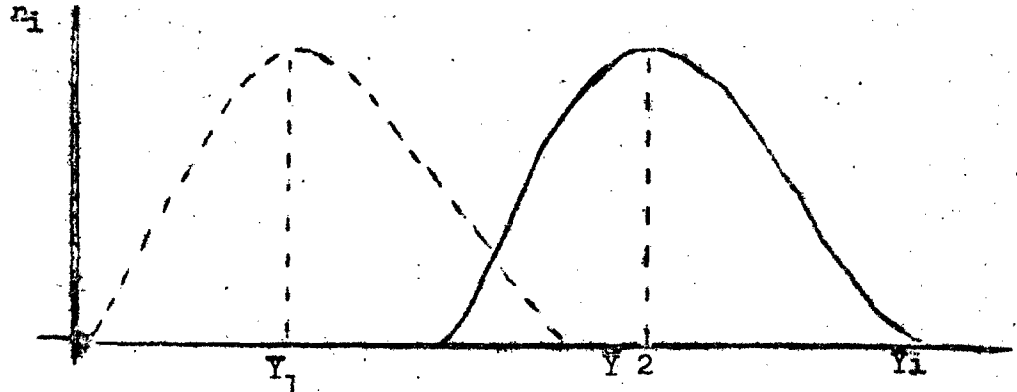
Puede observarse que a cada valor de la variable o marca de clase, se le dá una importancia o peso equivalente a la frecuencia absoluta correspondiente. Esta fórmula de cálculo es para datos agrupados en forma de una distribución de frecuencias. Cuando se desea calcular una media aritmética de datos no agrupados, todas las frecuencias absoluta serán iguales a la unidad, se simbolizará por \bar{X} ó $M [X_i]$ y se definirá como:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

En general siempre es conveniente la agrupación de frecuencias en intervalos, cuando el número de observaciones es relativamente grande. Es evidente que al resumir un conjunto de datos en un número dado de intervalos, se pierde precisión; esta pérdida estará relacionada con la amplitud del intervalo, mientras mayor sea ésta, menos preciso el cálculo. Por ello, en general, para un mismo conjunto de datos, la media aritmética obtenida de los datos originales, que es un calculo exacto, diferirá de la obtenida de una tabla de distribución de frecuencias. La razón estriba en el supuesto de uniformidad de la distribución de frecuencias dentro de cada intervalo, supuesto que generalmente no se cumple. En todo caso esa pérdida de precisión está más que compensada por las ventajas que significa tener una tabla de frecuencias. Por lo demás en ciencias sociales, la precisión necesaria

autoriza, dentro de ciertos límites, a desentenderse de una rigurosidad extrema.

En el gráfico que a continuación se presenta, se tienen dos distribuciones de frecuencias (supuesto un gran número de intervalos pequeños) muy similares, y sin embargo con medias aritméticas muy distintas.



La media aritmética como estadígrafo de tendencia central, indica la posición de la distribución. Cabe advertir sobre este estadígrafo su fuerte sensibilidad en cuanto a valores extremos de la variable. Un valor muy alejado de los valores centrales, aunque poco representativo por ser único, puede hacer variar fuertemente el promedio. Por ello cuando se está utilizando este indicador en un análisis, vale la pena percatarse de la representatividad de los valores extremos, y la influencia que éstos tienen en el resultado. Muchas veces se concluye que es preferible estratificar previamente los datos originales en dos o tres categorías, realizando cálculos de medias aritméticas en forma separada para cada grupo.

a) Propiedades. Se presentarán las propiedades más importantes de la media aritmética

i) Primera propiedad: La suma de las desviaciones ponderadas de los valores de la variable respecto de la media aritmética es cero:

$$\sum_{i=1}^m (Y_i - \bar{Y}) n_i = 0$$

$$\sum_{i=1}^m Y_i n_i - n\bar{Y} = 0$$

$$n\bar{Y} - n\bar{Y} = 0$$

/ii) Segunda Propiedad:

ii) Segunda propiedad: La suma de los cuadrados de las desviaciones ponderadas de los valores de la variable es un mínimo, cuando se toman respecto de la media aritmética. Se iniciará la demostración tomando desviaciones respecto a un valor cualquiera P, para luego concluir que P forzosamente tendrá que ser la media aritmética.

$$\sum_{i=1}^m (Y_i - P)^2 n_i \quad \text{es mínimo}$$

La derivada de esa expresión respecto de P, se iguala a cero.

$$- 2 \sum_{i=1}^m (Y_i - P) n_i = 0$$

$$\sum_{i=1}^m Y_i n_i - nP = 0$$

$$P = \frac{\sum_{i=1}^m Y_i n_i}{n}$$

Existe seguridad que igualando la primera derivada a cero se obtiene un mínimo, por que se llega a un valor concreto. Para que fuera un máximo P tendría que ser infinito.

iii) Tercera propiedad. La media aritmética de una variable más (menos) una constante es igual a la media de la variable más (menos) la constante.

$$M [Y_i \pm K] = M [Y_i] \pm K$$

$$\sum_{i=1}^m \frac{(Y_i + K)n_i}{n} = M [Y_i] \pm K$$

$$\sum_{i=1}^m \frac{Y_i n_i}{n} \pm \frac{nK}{n} = M [Y_i] \pm K$$

$$M [Y_i] \pm K = M [Y_i] \pm K$$

iv) Cuarta propiedad. La media aritmética de una variable multiplicada (dividida) por una constante, es igual a la constante que multiplica (divide) a la media de la variable.

$$M [Y_i K] = KM [Y_i]$$

$$\sum_{i=1}^m \frac{Y_i K n_i}{n} = KM [Y_i]$$

$$K \sum_{i=1}^m \frac{Y_i n_i}{n} = KM [Y_i]$$

$$KM [Y_i] = KM [Y_i]$$

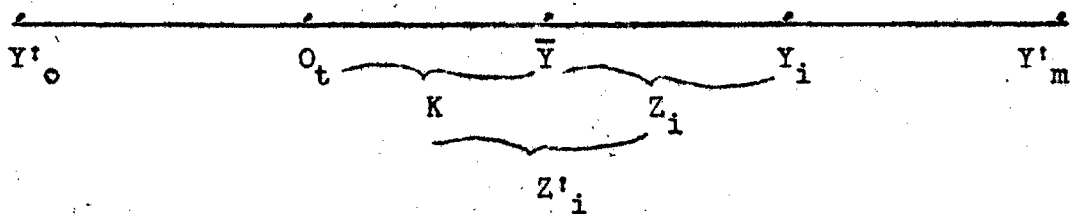
b) Métodos abreviados de cálculo. Se presentarán estos métodos más que por el ahorro de tiempo que puede significar su aplicación, porque recalcan algunos detalles sobre la media aritmética.

i) Primer método abreviado. Se trata de reducir la magnitud de la variable, en términos de desviaciones respecto de un origen de trabajo O_t elegido arbitrariamente. En cuanto a la elección de O_t vale la pena, para que el método sea realmente abreviado, tomar como origen de trabajo, un valor o marca de clase central de la distribución de frecuencias.

Se definirá esta variable reducida en la forma siguiente:

$$Z'_i = Y_i - O_t$$

Gráficamente se fijan los siguientes puntos dentro del recorrido de una variable:



Además, a las desviaciones respecto de la media aritmética, se simbolizará por Z_i , es decir:

$$Z_i = Y_i - \bar{Y}$$

La diferencia entre la media aritmética y el origen de trabajo arbitrariamente elegido se designará por K , es decir:

$$K = \bar{Y} - O_t$$

Por este hecho, observando el gráfico, puede concluirse que:

$$\bar{Y} = O_t + K$$

Se trata de encontrar una expresión para K , en función de desviaciones Z'_i , para disponer de una fórmula de cálculo. En efecto:

$$/Z'_i =$$

$$Z'i = Z_i + K \text{ (multiplicando por } n_i)$$

$$Z'i n_i = Z_i n_i + K n_i \text{ (aplicando sumatoria)}$$

$$\sum_{i=1}^m Z'i n_i = \sum_{i=1}^m Z_i n_i + nK$$

Recuérdese que en la primera propiedad de la media aritmética se demostró que:

$$\sum Z_i n_i = \sum (Y_i - \bar{Y}) n_i = 0$$

Luego:

$$K = \frac{\sum Z'i n_i}{n}$$

La fórmula de cálculo abreviado será

$$\bar{Y} = O_t + \frac{\sum Z'i n_i}{n}$$

En la siguiente tabla de distribución de frecuencias se aplica este método:

	Ahorros			Familias	
Y'i-1	Y'i	Yi	n _i	Z'i	Z'i n _i
0	8	4	8	- 19	- 152
8	20	14	10	- 9	- 90
20	26	23 <u>1/</u>	30	0	0
26	30	28	9	5	45
30	40	35	<u>3</u>	12	<u>36</u>
			60		- 161

1/ Se elige $O_t = 23$

$$\bar{Y} = 23 - \frac{161}{60} = 20,32$$

ii) Segundo método abreviado. Este método en general sólo se aplica con ventaja, en el caso en que la amplitud de los intervalos sea constante. Al igual que en el anterior se trata de trabajar en términos de desviaciones, pero además en este método dichas desviaciones se expresan en unidades de intervalo (dividiéndolas por C)

Esta nueva variable es en consecuencia:

$$Z''_i = \frac{Y_i - O_t}{C} = \frac{Z'_i}{C}$$

de donde: $Z'_i = CZ''_i$

En la fórmula del primer método se reemplaza Z'_i y se tiene :

$$\bar{Y} = O_t + C \frac{\sum Z''_i n_i}{n}$$

/Cabe destacar

Cabe destacar que este método permite obtener Z''_i en forma totalmente mecánica; basta fijar el origen de trabajo para completar todos los valores de Z''_i .

En el siguiente ejemplo podrá apreciarse las ventajas de esta forma de cálculo.

Rentabilidad del Capital (por cientos)		Nº de Empresas				
Y'_{i-1}	Y'_i	Y_i	n_i	Z''_i	$Z''_i n_i$	
2	-	6	4	20	-3	-60
6	-	10	8	40	-2	-80
10	-	14	12	50	-1	-50
14	-	18	16 <u>1/</u>	90	0	0
18	-	22	20	60	1	60
22	-	26	24	<u>40</u>	2	<u>80</u>
			300			- 50

1/ Eligiendo $O_t = 16$

Aplicando la fórmula del segundo método abreviado se tiene:

$$\bar{Y} = O_t + C \sum_{i=1}^m \frac{Z''_i n_i}{n}$$

$$\bar{Y} = 16 - 4 \frac{50}{300} = 15,33$$

Obsérvese que una vez fijado el origen de trabajo las Z''_i se colocan en sucesión decreciente para las marcas de clase menores que O_t y en sucesión creciente para las marcas de clase mayores.

2.- Mediana (Me) Se trata de otro estadígrafo de tendencia central de aplicación muy frecuente. Se define como aquel valor de la variable que supera a no más de la mitad de las observaciones y es superado por no más de la mitad de dichas observaciones. Es un estadígrafo menos sensible que la media aritmética ante valores extremos de la variable; pudiendo calcular este estadígrafo aún en variables de tipo ordinal, como se verá en un ejemplo.

Quando las observaciones no están agrupadas en forma de una tabla de distribución de frecuencias, su cómputo es en extremo sencillo. Basta disponer los valores en orden creciente y ubicar el valor central. Por ejemplo, supóngase que se tienen ordenados los siguientes valores de gastos en consumo de 7

/ familias (en

(en dólares por mes)

40 - 47 - 60 - 70 - 78 - 80 - 90

La mediana será 70 dólares ya que este valor supera a 3 observaciones (40, 47 y 60) que no son más que la mitad (la mitad es 3,5) y a su vez es superada por 3 observaciones (78, 80 y 90) que tampoco son más de la mitad.

Cuando el número de observaciones es par, existen dos valores centrales que satisfacen la definición de mediana. Si en el ejemplo anterior se agrega una familia adicional se tiene:

40 - 47 - 60 - 70 - 78 - 80 - 90 - 180

En este caso tanto 70 como 78 son valores medianos. La mitad de las observaciones es 4 y 70 supera a 3 que no son más de la mitad y es superado por 4 que tampoco es más de la mitad, es exactamente la mitad. Igual cosa ocurre con 78; para evitar ambigüedades se toma como mediana en estos casos el punto medio entre los dos valores medianos. En el ejemplo, la mediana definitiva sería 74 dólares. Obsérvese la poca sensibilidad de este estadígrafo a los valores extremos. Al agregar la 8a observación con un valor bastante más alto que el resto, la mediana ha experimentado apenas un ligero crecimiento, es más, aunque en vez de 180 aquel valor hubiese sido de 5.000, la mediana siempre habría sido 74. En cambio no ocurre lo mismo para la media aritmética que es sumamente sensible a ese tipo de valores extremos; intente el lector su cálculo en uno y otro ejemplo, y constatará una fuerte variabilidad. Si los datos se agrupan en una tabla de frecuencias, el cálculo de la mediana implica computar previamente las frecuencias acumuladas.

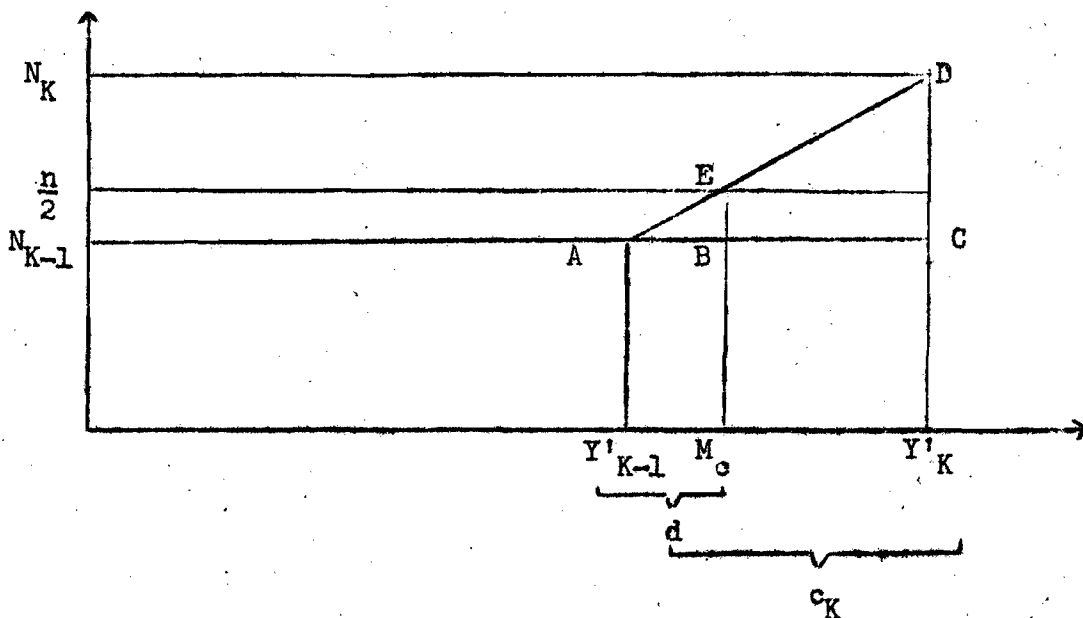
a) Variable discreta. En este caso bastará con identificar la frecuencia acumulada que es inmediatamente mayor a la mitad de las observaciones. La mediana será aquel valor de la variable que corresponda a dicha frecuencia acumulada. Ejemplo:

Número de predios per persona	Número de propietarios	Frecuencia acumulada
Y_i	n_i	N_i
1	200	200
2	160	360
3	150	510
4	100	610
5	80	690
6	40	730
7 y más	20	750
	750	

Siendo $\frac{n}{2} = 375$, la menor frecuencia acumulada que supera este valor es 510, que corresponde al valor 3 de la variable, siendo éste el valor mediano. Dicho valor supera a 360 observaciones que no es más de la mitad y es superado por 260 que tampoco es más de la mitad, satisfaciendo la definición de mediana.

b) Variable continua: En este caso el problema consiste en determinar un punto dentro del intervalo en que se halla comprendida la mediana. La identificación del intervalo en el cual se halla la mediana, es exactamente igual al caso de variable discreta: el intervalo será aquél que corresponda a la frecuencia acumulada inmediatamente superior a la mitad de las observaciones. Como se dijo, el asunto que preocupa es fijar un punto dentro de ese intervalo, que corresponde a la mediana. Para ello se adoptará el supuesto de que las observaciones se distribuyen linealmente dentro del mencionado intervalo.

Gráficamente se tiene lo siguiente:



Sea $Y'_{K-1} - Y'_K$ el intervalo donde se halla la mediana. Luego:

$$Me = Y'_{K-1} + d.$$

Será necesario encontrar una expresión para d en función de frecuencias que son los valores conocidos de que se dispone y por los cuales está determinado este estadígrafo.

Por semejanza de triángulos se tiene que:

$$\frac{AB}{BE} = \frac{AC}{CD}$$

Pero:

$$AB = d$$

$$BE = \frac{n}{2} - N_{K-1}$$

$$AC = C_K$$

$$CD = N_K - N_{K-1} = n_K$$

Reemplazando se tiene:

$$\frac{d}{\frac{n}{2} - N_{K-1}} = \frac{C_K}{n_K}$$

de donde:

$$d = C_K \frac{\frac{n}{2} - N_{K-1}}{n_K}$$

Luego:

$$Me = Y'_{K-1} + C_K \left(\frac{\frac{n}{2} - N_{K-1}}{n_K} \right)$$

Ejemplo: La siguiente tabla muestra la distribución de los coeficientes producto-capital de 310 empresas industriales

Coeficientes		Empresas	F. Acumuladas
Y'_{i-1}	Y'_i	n_i	N_i
0,15	0,20	40	40
0,20	0,30	60	120
0,30	0,42	100	220
0,42	0,50	60	280
0,50	0,70	30	310
		<u>310</u>	

$$\frac{n}{2} = \frac{310}{2} = 155$$

/ La menor

La menor frecuencia acumulada que supera a 155 es $N_K = 220$. Luego: $n_K = 100$,
 $Y'_{K-1} = 0,30$, $C_K = 0,12$ y $N_{K-1} = 120$

Reemplazando estos valores en la fórmula:

$$Me = 0,30 + 0,12 \frac{155 - 120}{100} = 0,30 + 0,036 = 0,336$$

Como una extensión de este estadígrafo, es fácil ampliar el concepto a otros indicadores que dividen la masa de informaciones en otras proporciones y no sólo en mitades como lo hace la mediana.

Se tiene el caso de los cuartiles que dividen las observaciones en cuartas partes. Así, el primer cuartil Q_1 es un valor de la variable que supera a no más de un cuarto de las observaciones, y es superado por no más de tres cuartos de ellas. Para identificar el intervalo en el que se halla Q_1 , habrá que determinar la frecuencia acumulada inmediatamente superior a $\frac{n}{4}$. El cálculo es similar al de la mediana:

$$Q_1 = Y'_{K-1} + C_K \frac{\left(\frac{n}{4} - N_{K-1} \right)}{n_K}$$

Para el tercer cuartil, sucede igual cosa

$$Q_3 = Y'_{K-1} + C_K \frac{\left(3 \frac{n}{4} - N_{K-1} \right)}{n_K}$$

Para identificar el intervalo que comprende a Q_3 , habrá que constatar cuál es la frecuencia acumulada N_K , inmediatamente superior a $3\frac{n}{4}$. No es necesario detallar el segundo cuartil, porque corresponde con la mediana. En forma similar se pueden encontrar estadígrafos que dividan al total de observaciones en décimas partes (deciles) en centésimas partes (percentiles) etc. Las fórmulas correspondientes pueden deducirse por analogía con las de los cuartiles.

Así, el 7° Decil estará dado por:

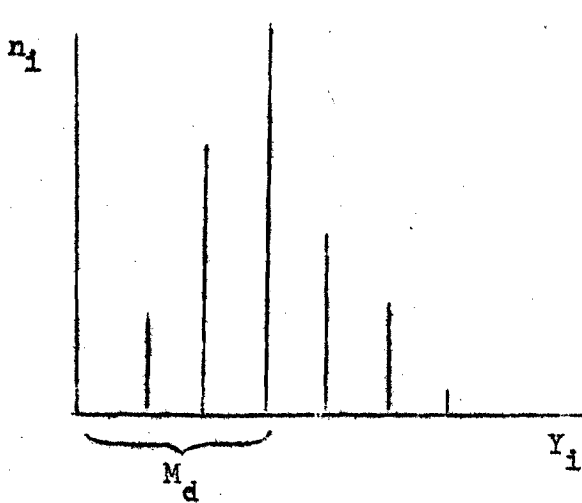
$$D_7 = Y'_{K-1} + C_K \frac{\left(\frac{7n}{10} - N_{K-1} \right)}{n_K}$$

El 35 Percentil estará dado por

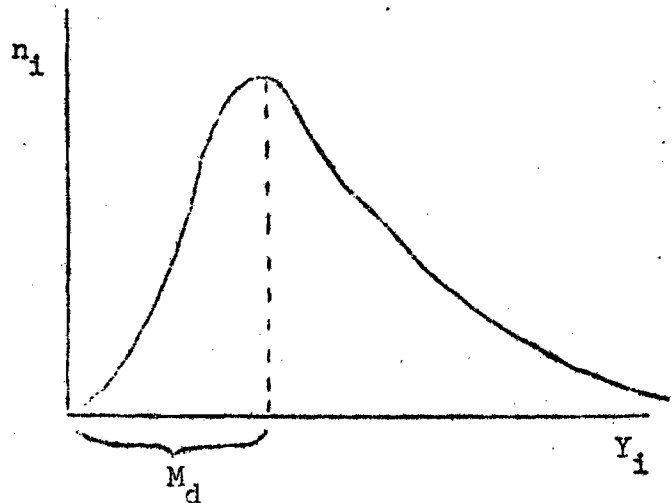
$$P_{35} = Y'_{K-1} + C_K \frac{\left(\frac{35n}{100} - N_{K-1} \right)}{n_K}$$

3. Valor Modal.

Se trata de otro estadígrafo de tendencia central. Tiene un significado bastante preciso y es de extraordinaria utilidad, aunque inexplicablemente poco utilizado en estudios socioeconómicos. Se simbolizara por M_d y se definirá como: aquel valor de la variable al que corresponde la máxima frecuencia. Es muy frecuente que se confunda el valor modal con la frecuencia máxima; recuérdese que es un valor de la variable y por lo mismo se le representa en el eje de las abscisas. Está dado por la frecuencia máxima, pero no se trata de una frecuencia.



Variable Discreta



Variable Continua

/a) Variable discreta

a) **Variable discreta.** Una vez que los datos están agrupados, es posible determinar inmediatamente el valor modal; bastará con fijar el valor de la variable que más se repite.

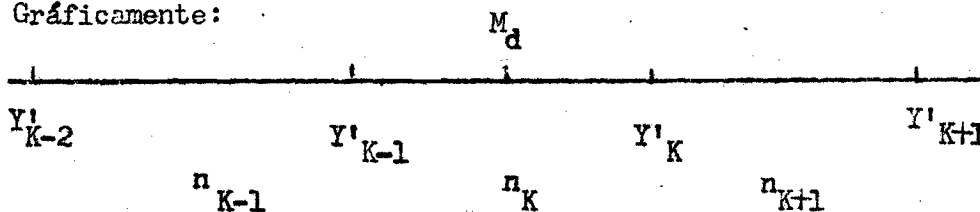
Ejemplo:

Número de cargas familiares	Número de familias
Y_i	n_i
0	80
1	120
2	210
3	380
4	180
5	60
6 o más	40
	<u>1 070</u>

La frecuencia máxima es 380 que corresponde al cuarto valor de la variable. El valor modal en consecuencia, es 3. Este valor modal será tanto más representativo mientras mayor es la frecuencia máxima. Se presentarán algunos casos donde el valor modal pierda significación; es el caso donde hay varios valores de las variables que tienen frecuencias similares. Es así como se califica a las distribuciones de unimodales, bimodales, multimodales, etc.

b) **Variable continua.** En la misma forma que cuando se calculaba la mediana, primero es necesario determinar el intervalo en el que se halla comprendido el valor modal. En este caso bastará ver cuál es el intervalo que tiene la frecuencia máxima. El siguiente paso es determinar un punto dentro de ese intervalo. Existen algunos criterios, un tanto arbitrarios, para deducir fórmulas del valor modal. Uno de tales criterios toma en cuenta la magnitud de las frecuencias de los intervalos contiguos (mayor y menor) al que comprende el valor modal. En otras palabras, dividirá al intervalo en partes inversamente proporcionales a las frecuencias de los intervalos contiguos.

Gráficamente:



$$\frac{Md - Y'_{K-1}}{Y'_K - Md} = \frac{n_{K+1}}{n_{K-1}}$$

La moda estará más cerca del intervalo contiguo que tenga mayor frecuencia.

Despejando Md de la relación anterior, se tiene:

$$Md \cdot n_{K-1} - Y'_{K-1} \cdot n_{K-1} = Y'_K \cdot n_{K+1} - Md \cdot n_{K+1}$$

pero,

$$Y'_K = Y'_{K-1} + C_K$$

$$Md [n_{K-1} + n_{K+1}] = Y'_{K-1} [n_{K+1} + n_{K-1}] + C_K n_{K+1}$$

$$Md = Y'_{K-1} + \frac{C_K n_{K+1}}{n_{K-1} + n_{K+1}}$$

Es necesario destacar que la deducción anterior, no toma en cuenta la amplitud de los intervalos contiguos. En caso de amplitudes muy diferentes, puede distorsionarse el valor de este estadígrafo.

Ejemplo:

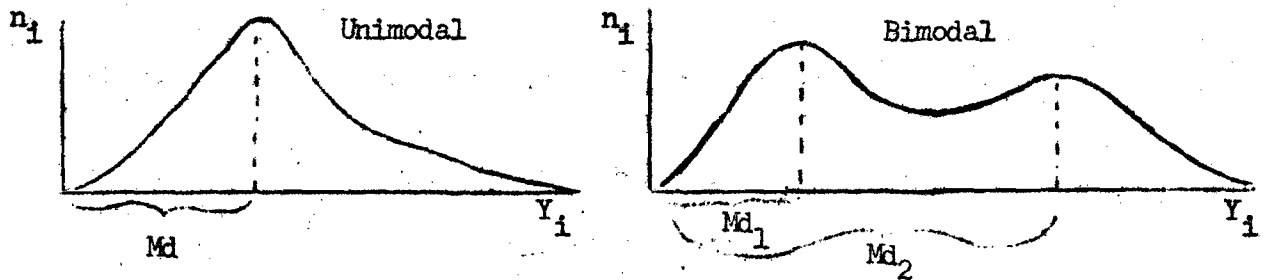
Ingresos de profesionales		Profesionales
Y'_{i-1}	Y'_i	n_i
0	20	25
20	40	45
40	60	80
50	80	60
80	100	40
100	120	15
120 y más		5

Inmediatamente se puede adelantar que la moda se encontrará en el tercer intervalo: 40 - 60. Aplicando la fórmula:

$$Md = 40 + \frac{20(60)}{45 + 60} = 40 + \frac{1200}{105} = 51,43$$

/Si se

Si se supone una gran cantidad de intervalos pequeños para una cierta distribución, gráficamente el valor modal estaría así representado:



Este estadígrafo al igual que la mediana, es susceptible de determinarse para variables cualitativas, ya que basta con constatar la frecuencia máxima. El siguiente ejemplo ilustra esta posibilidad:

Color de automóviles
preferido por los clientes

Clientes encuestados

Blanco	18
Azul	22
Verde	40
Amarillo	25
Rojo	75

El "valor" modal en este caso es el rojo, ya que teniendo la máxima frecuencia, es el preferido por los clientes.

Al iniciar el estudio de este estadígrafo, se decía que inexplicablemente era poco utilizado en los análisis. Evidentemente que requiere, para su cómputo, más información que la media aritmética; ello podría explicar parcialmente su poco uso, pero aún disponiendo de información muchas veces se cree suficiente calcular por ejemplo un ingreso per capita y quedarse con un análisis parcial. Sin duda, para caracterizar adecuadamente una distribución, se requiere una serie de estadígrafos cuyas indicaciones se complementan. Por ejemplo, saber que dos países tienen ingresos por habitante de 150 y 200 dólares al año, puede permitir cierto tipo de conclusiones, pero saber además que los valores modales de los ingresos anuales por habitante, son de 140 y 130 dólares respectivamente, permite obtener conclusiones bastante más objetivas. Naturalmente que son necesarios muchos otros antecedentes e indicadores que se presentarán a continuación, para realizar análisis completos. Interesaba destacar la

/necesidad de

necesidad de buscar un conjunto de indicadores que permitieran analizar las distintas facetas de un fenómeno sometido a estudio.

4. Media Geométrica.

Este estadígrafo se define como la raíz de orden n del producto de los n valores de la variable.

Cuando los datos no están agrupados, su fórmula de cálculo es:

$$Mg = \sqrt[n]{x_1 x_2 x_3 \dots x_n} = \sqrt[n]{\prod_{i=1}^n x_i}$$

Para fines prácticos es preferible calcular el logaritmo de la media geométrica y luego el antilogaritmo de ésta.

$$\log Mg = \frac{1}{n} \sum_{i=1}^n \log x_i$$

Si los datos aparecen agrupados, es decir, si las marcas de clase tienen frecuencias superiores a la unidad, se tendrá la siguiente fórmula.

$$Mg = \sqrt[n]{Y_1^{n_1} Y_2^{n_2} Y_3 \dots Y_m} = \sqrt[n]{Y_1^{n_1} Y_2^{n_2} \dots Y_m^{n_m}}$$

$$Mg = \sqrt[n]{\prod_{i=1}^m Y_i^{n_i}}$$

$$\log Mg = \frac{1}{n} \sum_{i=1}^m (\log Y_i) n_i$$

El estadígrafo que se estudia, aparte del inconveniente que representa la laboriosidad de su cálculo, tiene además la limitación de que los valores de la variable deben ser positivos para que sea susceptible de interpretación. Si algún valor de la variable es cero, la media geométrica será cero. Igualmente si existe algún valor negativo el estadígrafo toma un valor imaginario. Pese a estos inconvenientes, para cierto tipo de variables, en especial cronológicas, que sigan una tendencia exponencial, se hace indispensable su uso si se desea calcular valores intermedios, es decir, si se desea interpolar no linealmente. Por ejemplo, si cierta población en 1940 era de 2 millones y medio y en 1960 alcanza a los 4 millones, para calcular la población en 1955, sería indispensable el uso de la media geométrica, si se admite el crecimiento exponencial de la población a una tasa constante.

$$P_{1950} = \sqrt{(2,5) (4,0)} = 3,15 \text{ millones}$$

$$P_{1955} = \sqrt{(3,15) (4,0)} = 3,55 \text{ millones}$$

5. Media Armónica. El último de los estadígrafos de tendencia central que se estudiará en el curso, se define como el recíproco de la media aritmética de los valores recíprocos de la variable.

Para datos no agrupados:

$$M_h = \frac{1}{\sum \frac{1}{x_i} / n} = \frac{n}{\sum \frac{1}{x_i}}$$

Para datos agrupados

$$M_h = \frac{n}{\sum \frac{n_i}{Y_i}}$$

Ejemplo: Un grupo de trabajadores construyen los primeros 120 metros de una avenida con una productividad de 12 metros diarios, en cambio los siguientes 120 metros lo hacen a razón de 18 metros por día. Se pide determinar la productividad diaria en todo el trabajo.

Si se decidiera calcular la media aritmética se tendría:

$$\bar{Y} = \frac{12 + 18}{2} = 15 \text{ metros diarios}$$

Por otra parte en los primeros 120 metros se demoran 10 días y los siguientes 120 metros lo hacen en 6,67 días; es decir, todo el trabajo lo harían en 16,67 días. Si la productividad diaria es de 15 metros, en los 16,67 días construirían un total de 250,05 metros, lo que es inconsistente ya que el trabajo total es solamente de 240 metros. Si se utiliza la media armónica en cambio:

$$M_h = \frac{2}{\frac{1}{12} + \frac{1}{18}} = \frac{72}{3 + 2} = \frac{72}{5} = 14,4 \text{ metros}$$

Trabajando con una productividad media de 14,4 metros por día, en los 16,67 días, se construirán 240 metros.

Como pudo advertirse, la media armónica se aplica en el caso en que se presenta una relación inversa entre las variables implícitas; en el caso del ejemplo se presenta una relación inversa entre la productividad y el tiempo.

$$e = p \times t$$

e: espacio

p: productividad

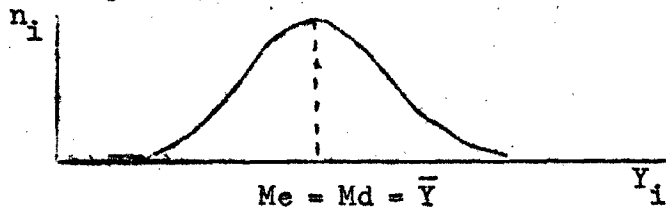
t: tiempo

$$p = e \cdot \frac{1}{t}$$

C - EVALUACION DE LOS ESTADIGRAFOS DE TENDENCIA CENTRAL

En más de una oportunidad se insistió sobre lo indispensable de contar con un conjunto de indicadores sobre la variable que se está estudiando. Los indicadores presentados, denominados en general como de tendencia central, tienen definiciones precisas; por ello muestran aspectos particulares del fenómeno que se estudia. Se trata de un conjunto de estadígrafos complementarios; las conclusiones a que en último término den lugar, deberán ser producto de la consideración simultánea de los valores que alcanzan dichos indicadores.

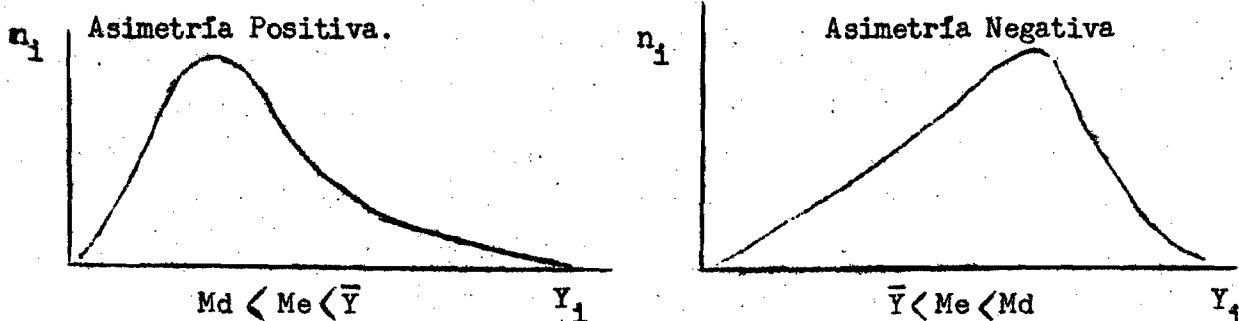
Al analizar la bondad de cada uno de estos indicadores, es preciso tener presente el volumen de observaciones tomadas en cuenta para su cálculo y las limitaciones que tiene cada uno de ellos. Siempre es conveniente complementar el análisis de las cifras, con una representación gráfica de la distribución de frecuencias de la variable. Interesa destacar la posición relativa de la media aritmética, la mediana y el valor modal. La posición relativa de estos estadígrafos, depende de la forma de la distribución. Así si la distribución es simétrica, es decir, si se observa perfecta simetría respecto de un eje central, los tres estadígrafos coinciden



En el caso de distribuciones no simétricas; la posición relativa de los estadígrafos, depende del tipo de asimetría. Así, si la asimetría es positiva, es decir, si la distribución tiene su rama o manto más extendido hacia valores positivos de la variable, la moda será menor que la media aritmética. La mediana por el hecho de dividir la masa de observaciones en dos partes, quedará comprendida entre ambas. Si la asimetría es negativa, es decir cuando la distribución se extiende suavemente hacia valores negativos de la variable, la

/moda superará

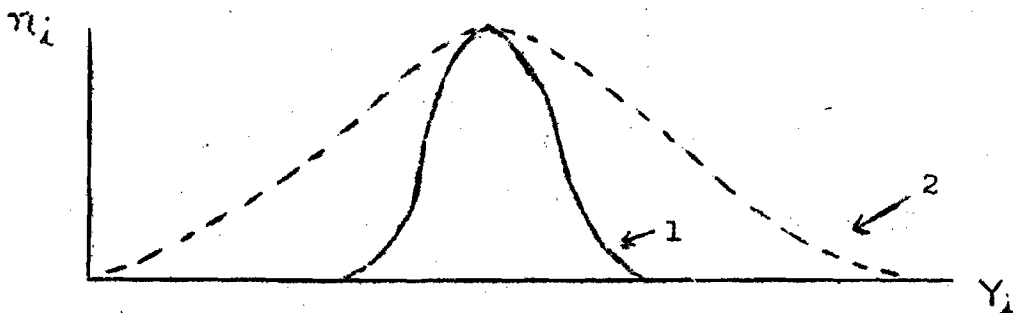
moda superará a la media aritmética, permaneciendo la mediana por la misma razón dada en el otro caso, comprendida entre ambos indicadores. Gráficamente



Recuérdese que la media aritmética es un estadígrafo muy sensible a valores extremos de la variable, de ahí que en un caso sea el mayor de los tres estadígrafos y en otro caso el menor de ellos. La moda, como el valor de la variable que más se repite, tiene en general una clara ubicación. Habría que agregar que su valor depende sobremanera de la amplitud de intervalo elegida y su representatividad sólo se garantiza cuando existe una clara concentración de frecuencias en un intervalo dado.

D. ESTADIGRAFOS DE DISPERSION

Una vez caracterizada la distribución a través de indicadores de tendencia central y en conocimiento del tipo de asimetría, interesa tener indicaciones acerca del grado de heterogeneidad con que la variable se distribuye en un conjunto de observaciones. Dos distribuciones pueden tener iguales estadígrafos de tendencia central, sin embargo pueden mostrar grados de dispersión diferentes, como puede observarse en el gráfico que a continuación se muestra.



Evidentemente en la primera distribución (línea continua) los valores aparecen más concentrados en torno al eje central, en tanto que en la otra aparecen mucho más dispersos. Si ambas distribuciones representaran ingresos de dos poblaciones, se concluiría que en la primera distribución los ingresos son más homogéneos, mientras que en la segunda se observaría gran disparidad /entre ingresos.

entre ingresos altos, medios y bajos.

Demás está destacar la importancia de contar con indicadores que pudieran mostrar este tipo de características en una distribución. Sobre todo en lo que se refiere a distribución de ingresos, de tanta actualidad hoy día, es indispensable contar con indicaciones adecuadas en este sentido.

1. Recorrido de la variable. Cuando se piensa en dispersión, lo primero que viene a la mente es el campo de recorrido de la variable: la diferencia entre el mayor y menor valor de ella. Si bien da una primera idea acerca de la heterogeneidad, tiene el inconveniente que sólo toma los dos valores extremos, sin considerar el conjunto de valores intermedios. Puede suceder que uno de los valores extremos, esté accidentalmente desplazado y no constituya por tanto un valor representativo; en este caso el recorrido sería exagerado y la dispersión aparecería distorsionada. Para iniciar el análisis es conveniente contemplar el recorrido, pero en ningún caso es suficiente.

2. Recorrido intercuartílico. Como una manera de obviar el inconveniente de los valores extremos que presentaba el estadígrafo anterior, se define un nuevo indicador, que toma en cuenta el recorrido entre el 1er. y 3er cuartil.

$$D_q = Q_3 - Q_1$$

Si bien es cierto que este indicador representa un adelanto respecto del anterior, no lo es menos que siempre toma dos valores de la variable, dejando de lado el resto, y en consecuencia la influencia de valores extremos puede, aunque en menor medida, originar algún tipo de deformación en cuanto al grado de dispersión.

3. Varianza. Se define este estadígrafo en virtud de la propiedad de la media aritmética que minimiza la suma de las desviaciones al cuadrado. Se simbolizará por σ^2 ó $V [Y_i]$

a) Para datos no agrupados

$$V [X_i] = \sigma^2 = \frac{\sum (X_i - \bar{X})^2}{n} = \frac{\sum X_i^2}{n} - \bar{X}^2$$

b) Para datos agrupados.

$$V [Y_i] = \sigma^2 = \frac{\sum (Y_i - \bar{Y})^2 n_i}{n} = \frac{\sum Y_i^2 n_i}{n} - \bar{Y}^2$$

/Si bien

Si bien la varianza no tiene un fin "per se" si no que se utiliza en materias que se presentarán posteriormente, da origen a un estadígrafo que sí tiene utilidad e interpretación práctica. Se trata de la desviación típica o estándar que se define como la raíz cuadrada positiva de la varianza.

$$\sigma = + \sqrt{\sigma^2}$$

Mientras más dispersa la variable, mayor será la magnitud de la desviación típica ya que mayores serán los desvíos respecto de la media aritmética, no habiendo posibilidad de compensación de desvíos por tratarse de suma de cuadrados. Este estadígrafo se expresa en las mismas unidades de la variable que se estudia, en tanto que la varianza se expresa en el cuadrado de la unidad de medida.

Como puede observarse en la fórmula, este indicador de dispersión toma en cuenta todos los valores de la variable con sus correspondientes frecuencias o pesos relativos, sin embargo, siempre es sensible a valores extremos. Por ello es conveniente, antes de calcular los estadígrafos, hacer un análisis previo de la tabla de distribución de frecuencias, para visualizar la representatividad de valores extremos y sus posibles efectos en los valores de los estadígrafos.

c) Propiedades de la Varianza

i - Primera propiedad: La varianza de una variable a la cual se le ha sumado (restado) una constante, es igual a la varianza de la variable original.

$V [Y_i \pm K] = V [Y_i]$ Aplicando la definición de varianza a la variable $Y_i + K$, se tiene:

$$\sum_{i=1}^m \frac{(Y_i \pm K - M [Y_i \pm K])^2 n_i}{n} = ./.$$

pero se vio que: $M [Y_i \pm K] = K \pm M [Y_i]$

$$\sum_{i=1}^m \frac{(Y_i \pm K - M [Y_i] \pm K)^2 n_i}{n} = V [Y_i]$$

$$\sum_{i=1}^m \frac{(Y_i - \bar{Y})^2 n_i}{n} = V [Y_i]$$

/ Gráficamente, las

Gráficamente, las dos distribuciones tienen la misma varianza, pese a estar desplazadas en el eje de las abscisas.



ii) Segunda propiedad: La varianza del producto de una constante por una variable, es igual al cuadrado de la constante por la varianza de la variable.

$$V [K Y_i] = K^2 V [Y_i]$$

$$\sum_{i=1}^m \frac{(K Y_i - K\bar{Y})^2 n_i}{n} = \dots$$

$$\sum_{i=1}^m \frac{[K^2 (Y_i - \bar{Y})]^2 n_i}{n} = \dots$$

$$K^2 \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2 n_i}{n} = \dots$$

$$K^2 V [Y_i] = K^2 V [Y_i]$$

d) Componentes de la Varianza. En el caso en que un conjunto de datos haya sido dividido previamente en grandes categorías o estratos, es posible desglosar la varianza en dos componentes muy útiles para el análisis. Admitase que una masa de datos ha sido dividida en L estratos, cada estrato tendrá una media aritmética, una varianza y un número de observaciones que representa la importancia de cada uno de estos estratos. En este caso la variabilidad total puede deberse tanto a variabilidad dentro de cada estrato y a variabilidad entre los diferentes estratos.

i) Intervarianza: Estadígrafo que representa la variabilidad entre los estratos, se define como la varianza entre las medias de los estratos.

$$\sigma_b^2 = V [\bar{Y}_h] = \frac{\sum_{h=1}^L (\bar{Y}_h - \bar{Y})^2 n_h}{n}$$

/donde: \bar{Y}_h

donde: \bar{Y}_h es la media aritmética del estrato h
 \bar{Y} es la media aritmética general

nh es el número de observaciones o tamaño de cada estrato

ii) Intravarianza. Estadígrafo que representa la variabilidad dentro de los estratos. Se define como el promedio de las varianzas de los estratos.

$$\sigma_w^2 = M[\sigma_h^2] = \frac{\sum_{h=1}^L \sigma_h^2 \cdot nh}{n}$$

donde σ_h^2 es la varianza del estrato h, y n_h y n obedecen a las mismas definiciones del caso anterior.

Dado que los dos estadígrafos estudiados son partes componentes de la varianza, a continuación se presenta la correspondiente de mostración.

$$\sigma^2 = \sigma_b^2 + \sigma_w^2$$

$$\frac{\sum_{h=1}^L \sum_{t=1}^{nh} (Y_{ht} - \bar{Y})^2}{n} = \frac{\sum_{h=1}^L (\bar{Y}_h - \bar{Y})^2 nh}{n} + \frac{\sum_{h=1}^L \sigma_h^2 \cdot nh}{n}$$

pero $\sigma_h^2 = \frac{\sum (Y_{hi} - \bar{Y}_h)^2}{nh}$; reemplazando.

$$\sum_{h=1}^L \sum_{i=1}^{nh} (Y_{hi} - \bar{Y})^2 = \sum_{h=1}^L (\bar{Y}_h - \bar{Y})^2 nh + \sum_{h=1}^L \sum_{t=1}^{nh} \frac{(Y_{hi} - \bar{Y}_h)^2 \cdot nh}{nh}$$

Elevando al cuadrado los correspondientes binomios

$$\sum_{h=1}^L \sum_{i=1}^{nh} (Y_{hi}^2 - 2\bar{Y} Y_{hi} + \bar{Y}^2) = \sum_{h=1}^L (\bar{Y}_h^2 - 2\bar{Y} \bar{Y}_h + \bar{Y}^2) nh + \sum_{h=1}^L \sum_{i=1}^{nh} (Y_{hi}^2 - 2\bar{Y}_h Y_{hi} + \bar{Y}_h^2)$$

Aplicando las propiedades de la sumatoria

$$\sum_{h=1}^L \left[\sum_{i=1}^{nh} (Y_{hi}^2) - 2\bar{Y} \sum_{i=1}^{nh} Y_{hi} + nh\bar{Y}^2 \right] = \sum_{h=1}^L \bar{Y}_h^2 nh - n\bar{Y}^2 + \sum_{h=1}^L \left[\sum_{i=1}^{nh} (Y_{hi}^2) - nh\bar{Y}_h^2 \right]$$

$$\sum_{h=1}^L \sum_{i=1}^{nh}$$

$$\sum_{h=1}^L \sum_{i=1}^{nh} Y_{hi}^2 - n\bar{Y}^2 = \sum_{h=1}^L \bar{Y}_h^2 nh - n\bar{Y}^2 + \sum_{h=1}^L \sum_{i=1}^{nh} Y_{hi} - \sum_{h=1}^L nh\bar{Y}_h^2$$

Simplificando términos queda:

$$\sum_{h=1}^L \sum_{i=1}^{nh} Y_{hi}^2 = \sum_{h=1}^L \sum_{t=1}^{nh} Y_{ht}^2$$

Luego $\sigma^2 = \sigma_b^2 + \sigma_w^2$.

Ejemplo: Piénsese en los sueldos y salarios pagados por una fábrica, que tienen una varianza de 137.600. Si las observaciones se clasifican en los estratos: Obreros, empleados administrativos, y técnicos, es posible analizar más a fondo la distribución de ingresos.

Las informaciones de que se dispone serían:

Estratos (h)	Tamaño Estrato nh	Media por Estrato \bar{Y}_h	Varianza por Estrato σ_h^2
Obreros	300	400	160.000
Empleados Adm.	100	400	160.000
Técnicos	100	500	40.000

La media aritmética general será:

$$\bar{Y} = \frac{\sum_{h=1}^L \bar{Y}_h nh}{n} = \frac{2100}{5} = 420$$

$$\sigma_w^2 = \frac{\sum_{h=1}^L \sigma_h^2 nh}{n} = \frac{680.000}{5} = 136.000$$

$$\sigma_b^2 = \frac{\sum_{h=1}^L (\bar{Y}_h - \bar{Y})^2 nh}{n} = \frac{8000}{5} = 1600$$

El cálculo de los anteriores estadígrafos permite concluir que la variabilidad se debe principalmente a heterogeneidad en las remuneraciones dentro de los estratos y no así a diferencias entre estratos; en otros términos las remuneraciones promedio de cada estrato, son bastante homogéneas ya que la intervarianza es pequeña, mientras que las remuneraciones dentro de cada estrato son muy heterogéneas ya que la intravarianza es bastante grande.

/e) Métodos abreviados

(Ver anexo 1)

e) Métodos abreviados de cálculo.

1) Primer método abreviado. Se trata de encontrar una fórmula que reduzca el volumen de operaciones; igual que en el caso de la media aritmética, la variable se expresará en términos de desvíos respecto de un origen de trabajo. Recuérdese que:

$$\sigma^2 = \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2 n_i}{n} = \frac{\sum_{i=1}^m Z_i^2 n_i}{n}$$

ya que $Z_i = Y_i - \bar{Y}$

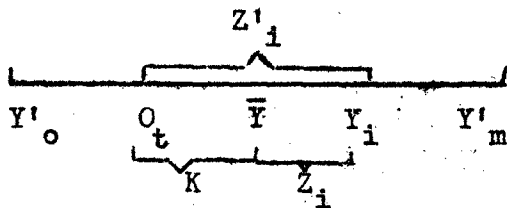
Si se desarrolla el cuadrado del binomio dentro de la sumatoria, se tiene:

$$\sigma^2 = \frac{\sum_{i=1}^m Y_i^2 n_i - 2\bar{Y} \sum_{i=1}^m Y_i n_i + n \bar{Y}^2}{n}$$

pero $\sum_{i=1}^m Y_i n_i = n\bar{Y}$, luego:

$$\sigma^2 = \frac{\sum_{i=1}^m Y_i^2 n_i}{n} - \bar{Y}^2$$

Observando el siguiente gráfico, resulta inmediata la deducción de la fórmula abreviada:



$$Z'_i = Z_i + K$$

Elevando al cuadrado

$$Z'^2_i = Z_i^2 + 2K Z_i + K^2$$

Ponderando por n_i

$$Z'^2_i n_i = Z_i^2 n_i + 2K Z_i n_i + K^2 n_i \quad \text{Aplicando sumatoria}$$

$$\sum_{i=1}^m Z'^2_i n_i = \sum_{i=1}^m Z_i^2 n_i + 2K \sum_{i=1}^m Z_i n_i + nK^2$$

Pero la primera propiedad de la media aritmética decía:

$$\sum_{i=1}^m Z_i n_i = 0, \quad \text{luego}$$

$$\sum_{i=1}^m$$

$$\sum_{i=1}^m Z'_i{}^2 n_i = \sum Z_i{}^2 n_i + nK^2$$

pero $\sum_{i=1}^m Z_i{}^2 n_i = n\sigma^2$ y

$nK = \sum_{i=1}^m Z'_i n_i$ (en el cálculo abreviado de la media aritmética)

Luego:

$$\sigma^2 = \frac{\sum_{i=1}^m Z'_i{}^2 n_i}{n} - \left(\frac{\sum_{i=1}^m Z'_i n_i}{n} \right)^2$$

ii) Segundo método abreviado. Consiste, como se recordará, en expresar las desviaciones, en términos de unidades de intervalo (dividiendo por la amplitud del intervalo).

$$\frac{Y_i - O_t}{c} = \frac{Z'_i}{e} = Z''_i$$

es decir que $Z'_i = c Z''_i$ y reemplazando en la fórmula anterior se obtiene:

$$\sigma^2 = c^2 \left\{ \frac{\sum_{i=1}^m Z''_i{}^2 n_i}{n} - \left(\frac{\sum Z''_i n_i}{n} \right)^2 \right\}$$

En el siguiente ejemplo, se podrá comprobar el ahorro de tiempo que significa la aplicación de estas fórmulas. En el caso del primer método abreviado.

Impuestos por contribuyente		Número de contribuyentes				
Y'_{i-1}	Y'_i	Y_i	n_i	Z'_i	$Z'_i n_i$	$Z'_i{}^2 n_i$
0	20	10	20	-40	-800	32.000
20	40	30	15	-20	-300	6.000
40	60	50 ^{1/}	10	0	0	0
60	80	70	8	20	160	3.200
80	100	90	5	40	200	8.000
			<u>58</u>		<u>-740</u>	<u>49.200</u>

^{1/} $O_t = 50$

Reemplazando estos valores en la fórmula, se obtiene:

$$\sigma^2 = \frac{49.200}{58} - \left(\frac{-740}{58}\right)^2 = 848,3 - 162,8 = 685,5$$

En el caso del segundo método abreviado se tiene:

Y'_{i-1}	Y'_i	Y_i	n_i	Z''_i	$Z''_i n_i$	$Z''_i^2 n_i$
0	20	10	20	-2	-40	80
20	40	30	15	-1	-15	15
40	60	50 \checkmark	10	0	0	0
60	80	70	8	1	8	8
80	100	90	5	2	10	20
			<u>58</u>		<u>-37</u>	<u>123</u>

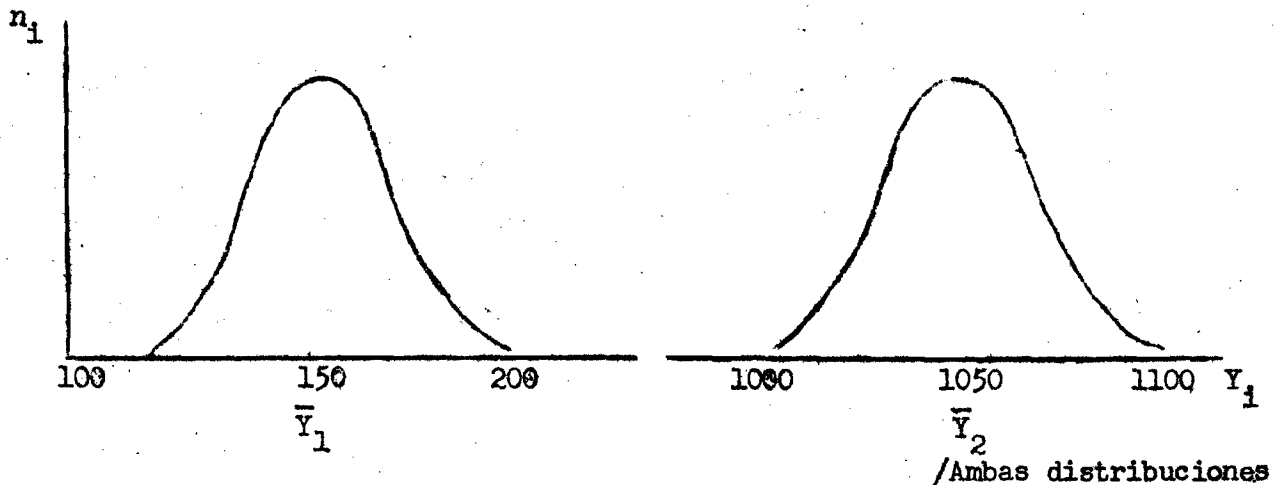
$$\checkmark \quad \sigma_t = 50$$

Aplicando la fórmula respectiva:

$$\sigma^2 = 20^2 \left\{ \frac{123}{58} - \left(\frac{-37}{58}\right)^2 \right\} = 400 \{ 2,12 - 0,407 \} = 685,5$$

4. Coeficiente de Variabilidad: Tanto la varianza como la desviación típica tienen el inconveniente de los estadígrafos absolutos, que en el caso de indicaciones sobre dispersión tiene mucha importancia, no tomar en cuenta la posición de la distribución. Estos estadígrafos, sobre todo en la comparación de distribuciones, pueden deformar las conclusiones.

Obsérvese las dos distribuciones que se tienen a continuación.



Ambas distribuciones muestran la misma dispersión en torno a la media, es decir, tienen igual varianza y desviación típica; sin embargo en términos relativos, una distribución de ingresos donde el menor ingreso es 1.000 y el mayor es 1.100, es mucho más homogénea que otra distribución donde el menor ingreso es 100 y el mayor 200. En un caso la diferencia es 10 %, mientras que en el otro es de 100 %.

Surge la necesidad de disponer de un estadígrafo que tome en cuenta la posición de la distribución. Se define así el coeficiente de variabilidad, como la razón entre la desviación típica y la media aritmética.

$$C.V. = \frac{\sigma}{\bar{Y}}$$

En el ejemplo anterior, si ambas distribuciones tuvieran una desviación típica de 60 por ejemplo, los coeficientes de variabilidad serían:

$$CV_1 = \frac{\sigma_1}{\bar{Y}_1} = \frac{60}{150} = 0,4 = 40 \%$$

$$CV_2 = \frac{\sigma_2}{\bar{Y}_2} = \frac{60}{1050} = 0,057 = 5,7 \%$$

Estos estadígrafos permiten llegar a conclusiones más realistas y ciertas.

El coeficiente de variabilidad o desviación típica relativa como también se le llama, puede tomar valores tan grandes como se quiera, ya que no hay una relación de dependencia entre σ y \bar{Y} . Por otra parte en el caso de una distribución donde la media aritmética fuera negativa no tiene sentido, para fines de calificar la dispersión, considerar el signo. Por ello este estadígrafo podría definirse como el valor absoluto del cociente entre la desviación típica y la media aritmética.

En vista de que las propiedades de la media aritmética y la desviación típica ya fueron analizadas, las propiedades del coeficiente de variabilidad serán el resultado de las propiedades de los indicadores componentes.

Si bien es cierto que para los efectos de calificar la dispersión de una distribución es más apropiado el coeficiente de variabilidad, no debe deducirse de esto que la varianza y la desviación típica carecen de utilidad. Por el contrario, son muy útiles en el tratamiento de materias que se estudiarán posteriormente.

E. UTILIZACION DE INDICADORES DE LA PROGRAMACION

Es muy frecuente escuchar quejas respecto de la escasez de informaciones estadísticas básicas para los países en vías de desarrollo. Admitido este punto, es conveniente también reconocer que no siempre se hace un uso óptimo de tan escasas informaciones. Reconociendo como etapa primaria en un proceso de planificación la realización de diagnósticos, es necesario destacar la enorme utilidad de la estadística descriptiva en lo que se refiere a caracterización de fenómenos en un instante de tiempo. Por una parte, el diagnóstico implica disponer de una perspectiva histórica referida a las variables estratégicas; por otra esa perspectiva histórica debe complementarse con análisis cuantitativos en profundidad en períodos que presenten cambios de orientación y/o ritmo en las tendencias observadas, aparte de una cuantificación detallada en el momento "cero" de un plan. Es en esos puntos donde el instrumental de estadística descriptiva debe ser puesto a disposición del analista. Se ha insistido en la urgencia de contar con un juego de indicadores para las principales variables, cada indicador muestra una faceta de la variable estudiada, un conjunto de ellos permitirá una adecuada e integral calificación de las variables que interesan en el diagnóstico.

En general un conjunto de indicadores, permite realizar análisis de consistencia, llegando de ese modo a una primera evaluación acerca de la calidad y fidelidad de la información que se pretende utilizar.

II. NUMEROS INDICE

A. El problema general

Al definir un número índice como un indicador de la tendencia central de un conjunto de elementos que se expresa generalmente como porcentaje, saltan a la mente las limitaciones de todo estadígrafo.

El cotidiano uso que se hace de este instrumento obliga a plantear previamente algunas de sus limitaciones. Es muy útil no olvidar que se trata de un indicador que pretende reflejar el comportamiento de ciertas variables en forma aproximada, en consecuencia no se trata de una medición exacta. Por otra parte es necesario establecer que un número índice plantea una comparación ya sea en el tiempo o en el espacio, respecto de un punto de referencia denominado base del índice.

A medida que se vayan introduciendo los distintos conceptos que se refieren al conjunto de los números índice, se profundizarán estos planteamientos primarios, ya que la experiencia aconseja que el estudiante tenga ciertas reservas a medida que avanza en este campo, para evitar posteriormente una utilización indiscriminada y sin las aludidas reservas.

B. Clases de números índice

Fundamentalmente, dentro de la Estadística Económica, interesa disponer de indicadores sobre precios, cantidades y valor.

Un índice de precios será un indicador que refleje la variación de los precios de un conjunto de artículos entre dos puntos en el tiempo o en el espacio. Es el caso de un índice de costo de vida.

Un índice de cantidades, será un indicador que refleje la variación en las cantidades de un conjunto de productos entre dos puntos en el tiempo o en el espacio. Por ejemplo un índice de producción industrial.

Por último un índice de valor indica la variación en el valor total de un conjunto de productos entre dos puntos del tiempo o el espacio. Ejemplo, índice de ventas totales comerciales.

C. Fórmulas de cálculo

Ocurre que cuando se trata de analizar la variación, por ejemplo, en el precio de un sólo artículo, no es necesario un indicador especial, basta con expresar la variación en términos porcentuales. Ejemplo:

/Período

<u>Período</u>	<u>Precio del bien</u>	<u>Relación porcentual</u>
1960	20	100
1961	25	125
1962	26	130
1963	30	150

Tomando como punto de referencia el precio del año 1960, y asignándole el valor 100, se calculan por una regla de tres simple, los índices correspondientes a los otros períodos. Así puede decirse que el precio del bien entre 1960 y 1965, ha experimentado un alza de 50 %. El cálculo de un índice, cualquiera que sea, referido a un sólo bien, no necesita pues de un estadígrafo especial.

El problema de los números índice nace cuando se desea averiguar las variaciones de precios o cantidades de un conjunto de artículos. Considérese el siguiente ejemplo:

Bienes	Precios	Precios
	1960	1964
A (metro)	10	15
B (quintal)	100	140
C (litro)	50	60
D (tonelada)	8	6
	<hr/>	<hr/>
	168	221

Para resumir el incremento en los precios de este conjunto de artículos, una primera solución consistiría en sumar los precios en ambos períodos y establecer la variación porcentual entre ambos agregados. Se llegaría de esa manera a calcular un índice agregativo simple. Si se considera el año 1962 como base, se tiene:

$$1960 \quad 1964$$

$$100 \quad 131,5 = \frac{221}{168} \cdot 100$$

Podría concluirse que el conjunto de estos precios ha variado en 31,5 % en el período. Sin embargo el método tiene dos serias limitaciones. Por una parte estará afectado por las unidades a que estén referidos los precios y por otra considera a los cuatro bienes, igualmente importantes, en circunstancias que el bien B puede ser trigo y el bien D comino; no se

/discrimina en

discrimina en este método en cuanto a la importancia relativa de cada artículo. En cuanto a las unidades de medida, considérese el ejemplo anterior, y supóngase que el precio del bien B se refiere al quintal de trigo; si se tuviera el precio del kilo el resultado del índice sería distinto:

Bienes	1960	Precios	1964
A (metro)	10		15
B (kilo)	1		1,4
C (litro)	50		60
D (tonelada)	8		6
	<hr/>		<hr/>
	69		82,4

Asignándole siempre a 1960 el valor 100 como base del índice, se tiene que para 1964 el índice agregativo simple es ahora de 119,42 ($100 \cdot 82,4/69$). Se llega a un resultado distinto sin que haya habido variación de precios alguna respecto del caso anterior, excepto que en ambos periodos se tomó un precio referido ahora a una unidad distinta.

En cuanto a este problema, es posible obviarlo calculando precios relativos. Se asigna a los precios de cada uno de los artículos en el período base, el valor 100 y por regla de tres simple, se calculan los correspondientes al período para el cual interesa conocer el índice. Si se observan los ejemplos anteriores, se tendrá:

Bienes	1960	Precios	1964
A	100		150
B	100		140
C	100		120
D	100		75

Obsérvese que en el bien B, sea cual fuere la unidad de medida, el incremento de su precio es de 40 %. Pero aunque en este método denominado de cifras relativas se obvia el problema de las unidades, todavía persisten problemas que exigen solución: en primer lugar respecto a la elección de un indicador de tendencia central. Se tienen los precios relativos para 1964, pero es necesario resumirlos por medio de un estadígrafo de posición: media aritmética, mediana etc. Todos ellos conducirán en general a

general a resultados distintos. ¿Cuál de ellos tomar? Si bien en cada caso particular habrá un estadígrafo adecuado que satisfaga las necesidades de representabilidad que se tenga, en general se utiliza la media aritmética, principalmente, por la facilidad que implica su manejo algebraico. Es necesario cuidar de que no hayan valores extremos que distorcionen el estadígrafo. En el último ejemplo, si se toma la media aritmética, el índice para 1960 será de 100 y el de 1964 de 121,25, es decir, el conjunto de artículos habrá experimentado un alza de 21,25 % en el período. Obsérvese que la conclusión está referida al conjunto, ya que se trata de un promedio. Cada bien en particular acusa variaciones disparés en sus precios, incluso el bien D muestra decrecimiento.

El tomar cifras relativas en consecuencia, salva el inconveniente de las unidades de medida, pero aun subsiste el problema de la ponderación, ya que a cada artículo debe asignársele la importancia debida.

Tomando como punto de partida los precios relativos se ensayarán algunos criterios de ponderación, para obtener los índices más usuales en la investigación económica.

Si los precios relativos:

$$\frac{p_n}{p_0}$$

donde p_n es el precio en el período dado y p_0 el precio en el período base, se ponderan por los valores del año base: $p_0 q_0$, se obtiene la conocida fórmula de Laspeyres para precios (IPL), es decir:

$$IPL = \frac{\sum \frac{p_n}{p_0} \cdot p_0 q_0}{\sum p_0 q_0} = \frac{\sum p_n q_0}{\sum p_0 q_0}$$

La sumatoria se extiende a todos los artículos considerados en el índice. Como en todo promedio aritmético, se divide por la suma de las ponderaciones.

Un índice de precios de Laspeyres debe interpretarse como el nivel que alcanzan los precios en un año dado, respecto de un año base al que se asigna el valor 100, considerando las cantidades del año base en ambos períodos. En otras palabras se trata de detectar la variación en los precios de una canasta de productos elegidos en el año base y que permanece inalterada en los períodos sucesivos.

/Este índice

Este índice, por lo tanto, tiene un significado bien concreto. El supuesto de que la canasta de productos permanezca en la realidad sin variaciones significativas, es otro problema. El analista tendrá que determinar si el supuesto se cumple o no y por consiguiente si es o no conveniente utilizar un índice de Laspeyres.

Por otra parte, si los precios relativos $\frac{p_n}{p_o}$, se ponderan por valores híbridos: $p_o q_n$, se tiene el índice de precios de Paasche (IPP), que también es de frecuente utilización:

$$IPP = \frac{\sum \frac{p_n}{p_o} \cdot p_o q_n}{\sum p_o q_n} = \frac{\sum p_n q_n}{\sum p_o q_n}$$

Obsérvese que ahora los precios están multiplicados por las cantidades del año que se calcula (q_n). Por este hecho un índice de precios de Paasche debe interpretarse como la variación de los precios de un conjunto de productos, suponiendo constantes las cantidades del año dado. En otros términos, la canasta de productos que se considera, es la del período que se calcula y se toma esta misma canasta para el año base.

Respecto de los índices de valor, por el significado simple que tienen, no requieren de deducciones especiales, ya que sencillamente es el cociente entre los valores del año que se calcula y del año base.

$$IV = \frac{\sum p_n q_n}{\sum p_o q_o}$$

A continuación se considerará un ejemplo en que se calcularán los índices presentados:

Artículos	Año 0		Año 1		Año 2	
	p	q	p	q	p	q
A	10	4	12	5	20	3
B	4	3	4	3	5	3
C	8	10	8	12	7	15
D	20	2	30	2	40	3

p: precio q: cantidad

No hace falta el cálculo en el año 0, base del índice, ya que coinciden precios y cantidades: $p_n = p_o$ y $q_n = q_o$.

/Para los

Para los índices de Laspeyres se tiene:

$$\text{IPL (año 1)} = \frac{\sum p_1 q_0}{\sum p_0 q_0} = \frac{48 + 12 + 80 + 60}{40 + 12 + 80 + 40} = \frac{200}{172} = 116,3$$

$$\text{IPL (año 2)} = \frac{\sum p_2 q_0}{\sum p_0 q_0} = \frac{80 + 15 + 70 + 80}{40 + 12 + 80 + 40} = \frac{245}{172} = 142,4$$

Para los índices de Paasche, suponiendo siempre el año 0 como base del índice:

$$\text{IPP (Año 1)} = \frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{60 + 12 + 96 + 60}{50 + 12 + 96 + 40} = \frac{228}{198} = 115,2$$

$$\text{IPP (Año 2)} = \frac{\sum p_2 q_2}{\sum p_0 q_2} = \frac{60 + 15 + 105 + 120}{30 + 12 + 120 + 60} = \frac{300}{222} = 135,1$$

El índice de valor:

$$\text{IV (año 1)} = \frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{60 + 12 + 96 + 60}{40 + 12 + 80 + 40} = \frac{228}{172} = 115,1$$

$$\text{IV (año 2)} = \frac{\sum p_2 q_2}{\sum p_0 q_0} = \frac{60 + 15 + 105 + 120}{40 + 12 + 80 + 40} = \frac{300}{172} = 174,4$$

En el siguiente cuadro puede apreciarse la diferencia en las indicaciones de uno y otro índice

Años	Indices		
	IPL	IPP	IV
0	100,0	100,0	100,0
1	116,3	115,2	115,1
2	142,4	135,1	174,4

Puede apreciarse que el IPL, muestra mayor crecimiento que el IPP. El primero se considera constante la canasta de productos del año base, en cambio en el segundo, se considera constante la canasta de productos del año que se calcula. Por lo tanto ambos índices indican la variación promedio de los precios bajo supuestos diferentes. Es muy frecuente confundir el significado de estos indicadores, porque no se tienen en cuenta los supuestos de uno y otro.

Con referencia a los índices de cantidad, es necesario hacer el mismo tipo de consideraciones ya que se presentan problemas similares: unidades de medida, ponderaciones, etc.

/Siguiendo el

Siguiendo el mismo criterio de los índices de precios, puede obtenerse el índice de cantidades de Laspeyres (IQL)

$$IQL = \frac{\sum \frac{q_n}{q_o} \cdot p_o q_o}{\sum q_o p_o} = \frac{\sum q_n p_o}{\sum q_o p_o}$$

El índice de cantidades de Paasche será (IQP)

$$IQP = \frac{\sum \frac{q_n}{q_o} q_o p_n}{\sum q_o p_n} = \frac{\sum q_n p_n}{\sum q_o p_n}$$

Mientras el IQL representa la variación en las cantidades suponiendo constantes los precios del año base, el IQP representa la variación de las cantidades suponiendo constantes los precios del año que se calcula. Nuevamente se insiste en la necesidad de no perder de vista estos supuestos, al interpretar un índice.

Las fórmulas presentadas son, como se dijo, las de uso más frecuente. Existen una cantidad extraordinaria de fórmulas de índices que se diferencian unas de otras según los factores de ponderación que se utilicen. Por razones de tiempo sólo se mostrarán las más conocidas.

Marshall - Edgeworth para precios

$$IPM = \frac{\sum p_n (q_o + q_n)}{\sum p_o (q_o + q_n)}$$

Índice de precios de Keyúes

$$IPK = \frac{\sum p_n (q_o \wedge q_n)}{\sum p_o (q_o \wedge q_n)}$$

donde el signo \wedge es infimo y quiere decir que se tome la menor de las cantidades que están a sus costados.

La llamada fórmula "ideal" de Fisher que es la media geométrica de los índices de Laspeyres y de Paasche.

$$JPF = \sqrt{IPL \cdot IPP} = \sqrt{\frac{\sum p_n q_o}{\sum p_o q_o} \cdot \frac{\sum p_n q_n}{\sum p_o q_n}}$$

En estas últimas fórmulas, bastará reemplazar p por q, para obtener las fórmulas correspondientes a índices de cantidad.

D. Pruebas sobre los números índice

Irving Fisher, quien plantea la fórmula per él llamada ideal, propone pruebas para calificar a los números índice.

1. Prueba de reversión de factores

La prueba se basa en un criterio de analogía: lo que es cierto para un producto, deberá ser cierto para un conjunto de ellos. Así para cualquier artículo.

$$\text{precio} \times \text{cantidad} = \text{valor}$$

Esta prueba, como puede verse a continuación, no cumplen las fórmulas de Laspeyres y Paasche

$$IPL \times IQL \neq IV, \text{ en efecto}$$

$$\frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum q_n p_0}{\sum q_0 p_0} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$$

$$IPP \times IQP \neq IV$$

$$\frac{\sum p_n q_n}{\sum p_0 q_n} \times \frac{\sum q_n p_n}{\sum q_0 p_n} \neq \frac{\sum p_n q_n}{\sum p_0 q_0}$$

La fórmula de Fisher si cumple este test.

$$IPF \times IQF = IV$$

$$\left(\frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum p_n q_n}{\sum p_0 q_n} \right)^{1/2} \left(\frac{\sum q_n p_0}{\sum q_0 p_0} \cdot \frac{\sum q_n p_n}{\sum q_0 p_n} \right)^{1/2} = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

Simplificando términos semejantes se tiene:

$$\frac{\sum p_n q_n}{\sum p_0 q_0} = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

Sin embargo, esta prueba, también se satisface con una combinación de índices de precios de Laspeyres y cantidades de Paasche o viceversa, como se comprueba a continuación:

$$IPL \times IQP = IV$$

$$\frac{\sum p_n q_0}{\sum p_0 q_0} \cdot \frac{\sum q_n p_n}{\sum q_0 p_n} = \frac{\sum p_n q_n}{\sum p_0 q_0}$$

$$IPP \times IQL = IV$$

$$/ \frac{\sum p_n q_n}{\sum p_0 q_0}$$

$$\frac{\sum p_n q_n}{\sum p_o q_n} = \frac{\sum q_n p_o}{\sum q_o p_o} = \frac{\sum p_n q_n}{\sum p_o q_o}$$

Estas dos últimas relaciones vale la pena tenerlas presente, por que se utilizan frecuentemente.

2. Prueba de reversión temporal

Nuevamente el criterio de analogía, si el precio de un producto es en el periodo a de 40 y en el periodo b de 50, en el primer periodo se constata que el precio es el 80 % del que se da en el periodo b, y en éste el 125 % del precio del periodo a. Lógicamente el producto de estos porcentajes debe dar 1, es decir:

$$\frac{P_n}{P_o} \times \frac{P_o}{P_n} = 1$$

Esta prueba, no la cumplen las fórmulas de Laspeyres y Paasche.

En efecto:

$$IPL_{b,a} \times IPL_{a,b} \neq 1$$

donde el primer subíndice indica el periodo que se calcula y el segundo indica el periodo base.

$$\frac{\sum p_b q_a}{\sum p_a q_a} \cdot \frac{\sum p_a q_b}{\sum p_b q_b} \neq 1$$

$$IPP_{b,a} \times IPP_{a,b} \neq 1$$

$$\frac{\sum p_b q_b}{\sum p_a q_b} \times \frac{\sum p_a q_a}{\sum p_b q_a} \neq 1$$

La fórmula de Fisher, si cumple la prueba

$$IPF_{b,a} \times IPF_{a,b} = 1$$

$$\left(\frac{\sum p_b q_a}{\sum p_a q_a} \cdot \frac{\sum p_b q_b}{\sum p_a q_b} \right)^{\frac{1}{2}} \left(\frac{\sum p_a q_b}{\sum p_b q_b} \cdot \frac{\sum p_a q_a}{\sum p_b q_a} \right)^{\frac{1}{2}} = 1$$

3. Prueba circular

Si el precio de un producto es de 10 en el primer periodo de 12 en el segundo y de 18 en el tercero, se constata que en el segundo /periodo el

período el precio es 120 % del precio en el primero y en el tercer período 150 % del precio que se da en el segundo. En consecuencia el precio del tercer período es 180 % del que se da en el primero:
 $(120 \%) (150 \%) = 180 \%$

La prueba circular no la cumple ninguna de las fórmulas analizadas. Suponiendo tres períodos para IPL se tiene:

$$IPL_{3,2} \cdot IPL_{2,1} \neq IPL_{3,1}$$

donde el primer subíndice indica el período que se calcula y el segundo el período base.

$$\frac{\sum p_3 q_2}{\sum p_2 q_2} \cdot \frac{\sum p_2 q_1}{\sum p_1 q_1} \neq \frac{\sum p_3 q_1}{\sum p_1 q_1}$$

Esta relación sólo se cumpliría en el caso en que $q_1 = q_2 = q_3$. Sin embargo, la relación planteada se cumple aproximadamente, cuando no existen diferencias significativas en las cantidades en los distintos períodos.

El estudiante podrá comprobar que la prueba circular no es cumplida, por la fórmula de Paasche ni por la de Fisher, siguiendo el mismo proceso de la comprobación para la fórmula de Laspeyres. Respecto de estas pruebas, es conveniente aclarar que el hecho de que hayan índices que no las cumplan no es razón para dejarlos de lado. Interesa más el significado concreto del índice, teniendo en cuenta sus alcances y limitaciones. Es así como la fórmula ideal de Fisher, cumpliendo con las pruebas por él planteadas, no es susceptible de clara interpretación, ya que se trata de la combinación de dos índices que por separado adquieren cabal significado pero al combinarlos presentan dificultades en su calificación.

E. Base de un número índice

Al definir un número índice se ha recalcado que se trata de una comparación de dos puntos en el tiempo o en el espacio. El punto respecto del cual se hace la comparación, recibe el nombre de base de un índice y se le asigna el valor 100, para analizar las variaciones porcentuales. Respecto de la elección del período base hay que tener siempre presente el objetivo que se persigue con el índice. En general se dice que el período base debe ser un período normal. Cabe preguntarse qué se entiende por normalidad en estos casos, cuando en los países en desarrollo los cambios se están sucediendo con mucha frecuencia y la anormalidad es un denominador común. Tal vez sería más sensato al definir el período base pensar en un período en el que no existan accidentes o cambios violentos. Será necesario cambiar la base del índice cuando los supuestos planteados pierdan validez a medida que pasa el tiempo. Es el caso de los índices de costo de vida, donde es necesario cambiar la base toda vez que la estructura de consumo, presente cambios significativos respecto de la que se daba en el período base.

Sobre este mismo asunto, es necesario distinguir dos tipos de base: base fija y base variable. Los índices de base fija son aquellos que mantienen como base un período fijo de referencia, en tanto que los índices de base variable son aquellos que tienen como base el período inmediatamente anterior. Teniendo un índice de base fija es posible calcular el correspondiente de base variable y viceversa; los resultados en general diferirán de los que se obtendrían a partir de los datos originales ya que las fórmulas usuales no cumplen con la prueba circular.

Ejemplo: Supóngase que el índice de Laspeyres para los precios de los materiales de construcción sea el siguiente:

Índice	
base 1960 = 100	
1960	100
1961	110
1962	120
1963	144
1964	180
1965	190

/El correspondiente

El correspondiente índice de base variable sería:

Índice de Base variable

1960	—
1961	110,0
1962	109,1
1963	120,0
1964	125,0
1965	105,5

Otra operación que es muy usual respecto de los índices de base fija es la del empalme. Se trata, como dice su nombre, de empalmar índices con base distinta. Obsérvese el siguiente ejemplo, en que se tiene un índice para el período 1956-1959 con base en 1956 y otro índice de 1959 a 1962 con base en 1959 y se pretende tener una serie para todo el período 1956-1962

Años	Índice base 1956 = 100	Índice base 1959 = 100
1956	100	
1957	120	
1958	150	
1959	180	100
1960		110
1961		132
1962		150

Mediante una sencilla regla de tres, puede completarse cualquiera de las dos series, para tener el movimiento del índice en todo el período.

Años	Índice base 1956 = 100	Índice base 1959 = 100
1956	100,0	55,6
1957	120,0	66,7
1958	150,0	83,3
1959	180,0	100,0
1960	198,0	110,0
1961	237,6	132,0
1962	270,0	150,0

/Debe advertirse

Debe advertirse que este tipo de empalmes, significa tan sólo una aproximación que puede ser muy defectuosa dependiendo de la similitud de las bases y de sus supuestos. En todo caso siempre es posible recurrir a estos empalmes en tanto se tenga conciencia de sus limitaciones.

F. Utilización de los números índice

Un número índice indica la evolución de precios, cantidades y valores, para un conjunto de productos. Prestan en consecuencia la utilidad inmediata de reflejar la tendencia de los cambios y ritmos de los conceptos señalados. Ese sólo hecho ya justifica su cómputo y su periódica utilización en la investigación socio económica. Sin embargo prestan además otros servicios sobre los que es conveniente hacer algunos comentarios.

1. La deflactación: Las alteraciones en los sistemas y niveles de precios que se presentan dentro de la actividad económica, originan dificultades en la comparación de valores monetarios que corresponden a períodos de tiempo distanciados. No es mucho lo que se puede deducir de la comparación de valores nominales, es decir, de valores expresados en unidades monetarias de distinto poder adquisitivo. Para poder llegar a conclusiones válidas acerca del comportamiento de una variable que represente "valor", será necesario expresar los montos monetarios nominales, en unidades homogéneas. La transformación aludida recibe el nombre de deflactación, y con esta operación se pretende eliminar, exclusivamente, el efecto de alteraciones en los precios.

El proceso de la deflactación exige disponer de un índice deflactor, es decir, de un indicador que proporcione una pauta de las alteraciones en los precios que dicen relación con la variable que se pretende deflactar. Es de fundamental importancia recordar que no existe un índice deflactor único; cada variable, estrictamente, debería tener un deflactor adecuado. La disponibilidad de sólo un reducido número de índices de precios, determina que éstos sean utilizados indiscriminadamente para una serie de propósitos. Aunque este hecho puede adolecer de errores conceptuales, muchas veces se justifica el procedimiento, porque interesa conocer un orden de magnitud antes que un valor exacto, siempre que se tenga conciencia de las limitaciones del método. En todo caso, aun dentro de las escasas disponibilidades de índices de

/precios, es

precios, es posible llegar a soluciones aceptables ya sea eligiendo racionalmente un índice de precios como deflactor o combinando y ponderando en forma adecuada dos o más índices. Este último proceso, si bien puede no conducir a soluciones ideales, al menos puede representar una disminución de las posibles distorsiones.

Antes de tener idea de este asunto de la deflactación, cuando se desea transformar unidades monetarias heterogéneas (unidades de cada período) en unidades monetarias homogéneas (unidades del período base) y permitir de este modo la comparación en el tiempo, lo primero que se piensa, es expresar los montos monetarios nominales en unidades de moneda extranjera de valor más o menos estable, dólares, libras, etc. Sobre este procedimiento caben algunas objeciones. Los Gobiernos tienen herramientas que les permiten fijar los tipos de cambio con las monedas extranjeras en forma arbitraria, arbitraria para los fines que aquí se comenta, y no representan generalmente las alteraciones en los niveles de precios. En Chile hay una experiencia reciente; en el transcurso de menos de dos años, la unidad monetaria chilena ha llegado a sufrir una devaluación enorme, (de E\$1,053 en noviembre de 1962 a E\$3,400 en abril de 1963, por dólar norteamericano tipo de cambio corredor). Si un valor dado en escudos se expresa en dólares en ambas fechas, se concluiría que entre los dos períodos mencionados dicho valor se reduciría a la tercera parte. Si bien esto puede ser cierto para una persona que va a gastar sus ingresos en EE.UU., no lo es para quien efectúa sus desembolsos en Chile, porque en ese período, ningún índice de precios propiamente tal ha sufrido un aumento de 200 por ciento.

Por otra parte, una segunda objeción a este procedimiento, es el hecho que aun las economías más estables pueden sufrir cierto grado de inflación. Los dos argumentos presentados descalifican, en general, la conversión a unidades monetarias extranjeras como una alternativa de la deflactación. La mecánica de la deflactación implica dividir los montos monetarios nominales, por el índice de precios elegido como deflactor adecuado. La razón por la que debe dividirse, estriba en la siguiente regla de tres. Si en el año n existen un valor nominal VN_n y un índice de precios IP_n , cuál sería este valor expresado en unidades monetarias

/de igual

de igual poder adquisitivo que las del año base? En otros términos, cuál sería este valor si el índice de precios no hubiera variado?

El planteamiento queda así reducido:

$$\begin{array}{ccc} IP_n & & VN_n \\ 100 & & x \end{array}$$
$$x = \frac{VN_n}{IP_n} \cdot 100 \Rightarrow \text{Valor real}$$

Desde otro punto de vista, se justifica la deflactación pensando en los componentes de un valor: precio por cantidad.

$$\text{Índice de precios} \times \text{cantidad} = \text{Valor (100)}$$

$$\text{Cantidad} = \frac{\text{Valor (100)}}{\text{Índice de precios}} \Rightarrow \text{Valor real}$$

Ahora bien, la evolución de las cantidades en el tiempo esta libre, por así decirlo, de influencias monetarias directas, mostrando la evolución física, real, de una serie, que es justamente lo que se pretende con la deflactación.

Los valores reales así obtenidos, vienen expresados en unidades monetarias que tienen un poder adquisitivo del año que es la base del índice deflactor.

Con referencia a este último planteamiento, es necesario distinguir dos tipos de base. Se llamará base propiamente tal al período que corresponde al diseño del índice, donde se diseña la muestra, se establecen las ponderaciones, etc. Por otra parte, se utilizará la expresión "base aritmética" para referirse a cualquier período, que por transformación lineal, se le haya asignado el valor 100. Los cambios aritméticos de base implican hasta cierto punto una arbitrariedad que puede originar algún tipo de perturbaciones en las comparaciones. Sus efectos serán tanto más fuertes, cuanto más distante sea la base propiamente tal del índice deflactor. Cuando exista conciencia que los supuestos de la construcción y diseño del índice siguen siendo válidos, los cambios aritméticos de base no introducirán deformaciones significativas. Por ello, antes de deflactar será necesario decidir de qué año serán las unidades

/monetarias en

monetarias en que interesa expresar los valores reales. Es recomendable, si se cumple la condición de constancia de los supuestos de la construcción del índice, expresar los valores nominales en unidades monetarias del año más reciente, lo que se consigue mediante un índice deflactor que tenga por base el último año en el sentido cronológico. La utilidad de la sugerencia anotada radica en el hecho de que el investigador tiene una visión reciente y tal vez objetiva del sistema de precios imperante, por lo que es probable que se facilite la obtención de conclusiones. Es necesario destacar que un proceso de deflactación conduce a valores reales que pueden tener dos interpretaciones: una expresión física o un poder de compra. Una variable monetaria está compuesta por una suma de valores del tipo $\sum p_n q_n$, si esta serie se deflacta por un índice de precios de los productos considerados en la serie nominal, el resultado será una expresión física de la serie. En efecto, utilizando un índice deflactor de Paasche, se tiene:

$$\frac{\text{Valor Nominal}}{\text{IPP}} = \frac{\sum p_n q_n}{\frac{\sum p_n q_n}{\sum p_0 q_n}} = \sum p_0 q_n$$

El resultado es evidentemente un quantum, es decir, cantidades del período n , valorizadas a precios del período base. Si se hubiera deflactado por un índice de precios de Laspeyres, se tendría:

$$\frac{\text{Valor Nominal}}{\text{IPL}} = \frac{\sum p_n q_n}{\frac{\sum p_n q_0}{\sum p_0 q_0}} = \text{IQP} \cdot \sum p_0 q_0$$

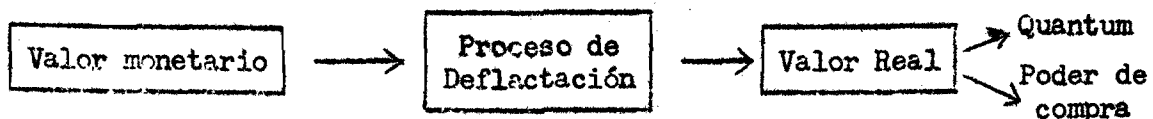
Resultado que equivale a proyectar un valor en el período base, a través de un índice de cantidad de Paasche. También representa una evolución física de la serie, aunque con connotaciones diferentes al caso anterior. Cabe hacer notar que cuando se desea llegar a una expresión física rigurosa, sólo existe un tipo de deflactor adecuado, aquél que contiene los productos incluidos en la serie nominal.

Por otra parte, cuando se quiere obtener un poder de compra, es necesario especificar el uso que se dará a un monto monetario; así, si el sueldo de un empleado en diferentes períodos se deflacta por un índice de precios

/al consumidor

al consumidor, el resultado sería un poder de compra en términos de la canasta de productos elegida en el índice deflactor. Si el sueldo de dicho empleado se deflacta por índice de valores bursátiles, el resultado será un poder de compra en términos de acciones y bonos. Es el uso que se dará a un monto monetario el que determina el tipo de poder de compra resultante.

Resumiendo esquemáticamente se tiene:



A continuación y a vía de ejemplo, se presentan los resultados de un proceso de deflactación:

Años	Sueldo de un empleado (U.M. de c/año) A	Índice de precios al consumidor (base 1963 = 100) B	Sueldo Real (U.M. de 1963) $\frac{A}{B} \cdot 100$
1958	400	36,5	1 096
1959	480	50,6	946
1960	600	56,5	1 062
1961	680	60,9	1 117
1962	720	69,3	1 039
1963	900	100,0	900

La deflactación presentada implica haber elegido como deflactor, el índice de precios al consumidor. Tal vez, de los valores reales no pueda deducirse categóricamente que el empleado haya sufrido un detrimento de esa magnitud en su poder de compra. Esto hubiera ocurrido si dicho empleado gastara todo su ingreso en la forma que lo hace el empleado típico elegido como padrón en el índice de precios al consumidor. Si el empleado en cuestión sólo gasta en manutención el 60 por ciento de sus ingresos y el 40 por ciento restante lo dedica, por ejemplo, a la construcción de una vivienda, la deflactación se realizaría en dos partes: la primera parte (60%) se deflactaría por el índice de precios al consumidor y el saldo /debería deflactarse

debería deflactarse por un índice de precios de insumos de la construcción; de esta manera se obtendría la expresión real de su poder de compra, en términos del uso que dará a sus ingresos.

Cuando se ha realizado una deflactación, es necesario tener presente que los valores reales así obtenidos son simplemente aproximaciones, que serán tanto mejores cuanto más representativo sea el índice de los precios que se cancelarán con el monto monetario. Esto no quiere decir que sea indispensable "construir" índices si no se dispone de los más adecuados, lo que se pretende es recalcar la necesidad de selección de los índices disponibles, y en todo caso, el tener presente las limitaciones que acuse una deflactación obligada por un índice que no sea el mejor. Es responsabilidad de los organismos autorizados y de los principales usuarios la elaboración de índices de precios de la actividad económica.

La deflactación, mecánicamente es un asunto trivial, pero la elección de deflatores adecuados requiere mucha atención. Véase por ejemplo lo que sucede con los índices de producción y ventas industriales de Chile.^{1/} Observando las cifras siguientes debe admitirse que no existe la concordancia que debiera existir, no encontrándose razones que expliquen la diferencia. Es muy probable que el desajuste radique en la calidad del deflactor utilizado, antes que en las variaciones de stocks.

Período	Índice de producción industrial Base 1959 = 100	Índice de ventas industriales reales Base 1959 = 100
1959	100,0	100,0
1960	103,5	98,3
1961	106,1	110,6
1962	113,8	127,7
1963		
Enero	108,4	120,9
Abril	122,2	129,8
Agosto	112,8	132,1

^{1/} "Los índices de producción industrial manufacturera y ventas reales industriales. Un comentario acerca de su comportamiento". Cesar Molestina, mayo de 1963.

De la simple observación de estas series surgen las contradicciones anotadas. Si se admite que el índice de producción industrial es un indicador confiable, el índice de ventas sería el que adolece de defectos. Estos defectos pueden deberse a dos causas no excluyentes: que el índice de ventas nominales no sea lo suficientemente acertado, por errores inherentes o ajenos al muestreo o que el deflactor utilizado no tiene la relación apropiada con los precios de los bienes industriales, o una combinación de ambas causas.

2. El deflactor implícito del Producto Bruto. No es necesario citar la imperiosa necesidad de disponer de un sistema de contabilidad social expresada en precios constantes. Respecto de esto, se tratará principalmente la deflactación del Producto Bruto.^{1/}

Es importante clasificar la idea de deflactor implícito. Este aparece como resultado de un proceso de deflactación cuando se trata de cumplir con alguna restricción. Principalmente se presenta cuando una variable global ha sido desglosada en componentes, con objeto de identificar un índice deflactor adecuado, con cada componente. En estos casos la restricción aparece como suma de componentes que reproducen la variable global. Es necesario agregar que el deflactor implícito se origina en el hecho de que la restricción (muchas veces definición) debe ser satisfecha tanto en valores nominales o corrientes como en valores reales o constantes; entonces se dice que el proceso de deflactación es coherente. El deflactor resultante de una deflactación coherente tiene el calificativo de implícito, porque justamente está implícito en el cumplimiento de la restricción.

Supóngase que X, Y y Z representan valores sujetos a la restricción:

$$X + Y + Z = W$$

restricción que se satisface en valores corrientes. Si además se desea que esta restricción sea satisfecha en valores constantes, se tiene:

$$\bar{X} + \bar{Y} + \bar{Z} = \bar{W}$$

^{1/} Ver "Discusiones conceptuales y prácticas sobre la deflactación", Arturo Núñez del Prado, ILPES, 1967.

/donde la

donde la barra sobre el símbolo se utiliza para indicar que se trata de valores reales, es decir, que:

$$\bar{X} = \frac{X}{IPX}$$

$$\bar{Y} = \frac{Y}{IPY}$$

$$\bar{Z} = \frac{Z}{IPZ}$$

La suma de $\bar{X} + \bar{Y} + \bar{Z}$ reproduce el valor real de W, es decir, \bar{W} . ¿Cuál deberá ser el deflactor de W para que el valor real resultante \bar{W} coincida con la suma de $\bar{X} + \bar{Y} + \bar{Z}$? En efecto, se tiene que:

$$\frac{X}{IPX} + \frac{Y}{IPY} + \frac{Z}{IPZ} = \frac{W}{IPW}$$

Basta con despejar IPW de la anterior relación:

$$IPW = \frac{W}{\frac{1}{IPX} \cdot X + \frac{1}{IPY} \cdot Y + \frac{1}{IPZ} \cdot Z}$$

Se comprueba que el deflactor implícito IPW es el promedio armónico de los deflactores componentes, donde los factores de ponderación son justamente los valores nominales. Otra forma de calcular el deflactor implícito es comparar \bar{W} que se obtiene como suma de $\bar{X} + \bar{Y} + \bar{Z}$ con W.

$$IPW = \frac{W}{\bar{W}}$$

Para determinar el deflactor implícito del Producto Bruto, habrá que pensar previamente en alguna definición. Si la restricción es del siguiente tipo:

$$P = C + I + E - M$$

Se determinará los deflactores para Consumo, Inversión, Exportaciones e Importaciones y se obtendrá por suma el Producto Real (\bar{P}).

$$\bar{P} = \frac{C}{IPC} + \frac{I}{IPI} + \frac{E}{IPE} - \frac{M}{IPM}$$

El deflactor implícito del Producto (DI) según esta restricción estará dado por la relación:

$$DI = \frac{P}{\bar{P}}$$

Puede verificarse al mismo tiempo que el deflactor implícito también corresponde con la media armónica de los deflactores componentes.

/Fórmula

$$DI = \frac{P}{\frac{1}{IPC} \cdot C + \frac{1}{IPI} \cdot I + \frac{1}{IPE} \cdot E - \frac{1}{IPM} \cdot M}$$

Si la restricción fuera de otro tipo, por ejemplo que la suma de los valores agregados sectoriales reproducen el Producto Bruto, se tendrá otro deflactor implícito sujeto a la restricción propuesta. Habrá que discutir previamente la posibilidad de encontrar deflatores adecuados para el valor agregado. Una alternativa, por cierto muy discutible, es deflactar el valor agregado de cada sector, por los precios de los bienes que el sector produce. El resultado sería un poder de compra de los valores agregados sectoriales, en términos de los bienes que produce cada sector. Puede argumentarse en favor de este método, sosteniendo que la contrapartida física del valor agregado es justamente la cantidad de bienes producidos. Sin embargo, la justificación es en extremo débil. En todo caso, en los cursos de Contabilidad Social se muestran las ventajas y desventajas de este y otros métodos de deflactación. Para ilustrar, a continuación se detalla este procedimiento.

Si se desea llegar a expresiones reales, en cada rama de actividad, debería deflactarse por un índice de precios relacionado directamente con dichas ramas de actividad. Así el valor agregado por el sector industrial, debería deflactarse por un índice de precios de bienes industriales, el valor agregado por el sector agropecuario, se deflactaría por un índice de precios de bienes agropecuarios. De esta manera se obtendrían valores reales en cada rama que representaría una aproximación a la evolución física de lo producido por cada sector. Sumando los valores reales de todas las ramas de actividad, se tendría el Producto Bruto en términos reales. Comparando los Productos Brutos en valores nominales y reales se obtiene el llamado deflactor implícito del Producto, que no es otra cosa que un índice general promedio de los precios que rigen en la actividad económica.

Sean: PN_{ij} el producto nominal del sector "i" en el año "j".

IP_{ij} el índice de precios del sector "i" en el año "j".

PR_{ij} el producto real del sector "i" en el año "j".

/De la

De la deflactación resulta:

$$PR_{1j} = \frac{PN_{1j}}{IP_{1j}} \cdot 100$$

Si se suman todos los productos reales por sector en un año cualquiera "j", se tiene el Producto Bruto Real en el año "j".

$$PR_j = \sum_{i=1}^L PR_{1j} = \sum_{i=1}^L \frac{PN_{1j}}{IP_{1j}} \cdot 100$$

donde $i = 1, 2 \dots L$ representa el número de sectores considerados. Por otra parte, se tiene que el Producto Bruto Real PR_j , se obtiene deflactando el Producto Bruto Nominal PN_j , por el deflactor implícito DI_j , todos los conceptos referidos a un período j, es decir:

$$PR_j = \frac{PN_j}{DI_j} \cdot 100$$

Reemplazando en la igualdad anterior, se tiene:

$$\frac{PN_j}{DI_j} \cdot 100 = \sum_{i=1}^L \frac{PN_{1j}}{IP_{1j}} \cdot 100$$

Despejando DI_j ,

$$DI_j = \frac{PN_j}{\sum_{i=1}^L \frac{PN_{1j}}{IP_{1j}}}$$

que justamente corresponde con la definición de media armónica de los deflatores sectoriales considerando como ponderaciones los productos nominales de cada sector, ya que:

$$PN_j = \sum_{i=1}^L PN_{1j}$$

/A continuación

A continuación se presentará un ejemplo para ilustrar el proceso de deflatación de los productos sectoriales.

Supóngase que la actividad económica ha sido dividida en cuatro sectores para los que se dispone de las siguientes informaciones:

Producto Nominal

(unidades monetarias corrientes)

<u>Sectores</u>	<u>1960</u>	<u>1961</u>	<u>1962</u>
Minería	200	300	400
Agricultura	300	350	400
Industria	200	250	300
Servicios	500	600	700
PRODUCTO BRUTO NOMINAL	<u>1.200</u>	<u>1.500</u>	<u>1.800</u>

Indices de precios (base 1960 = 100)

Productos mineros	100	130	150
Productos agrícolas	100	110	120
Productos industriales	100	120	130
Servicios	100	110	120

Deflatación el producto de cada sector, por el índice de precios correspondiente, se tendrá los productos reales de cada sector:

Producto Real

(unidades monetarias constantes de 1960)

<u>Sectores</u>	<u>1960</u>	<u>1961</u>	<u>1962</u>
Agricultura	200	230,8	266,7
Minería	300	318,2	333,3
Industria	200	208,3	230,8
Servicios	500	545,5	583,3
PRODUCTO BRUTO REAL	<u>1.200</u>	<u>1.302,8</u>	<u>1.414,1</u>

Para calcular el deflactor implícito, recuérdese que:

$$DI_j = \frac{PN_j}{PR_j} \cdot 100$$

/Deflactor implícito

	1960	1961	1962
Deflactor implícito	100	115,1	127,3

Como se demostró, estos valores equivalen a los promedios armónicos de los índices deflatores.

Sobre la deflactación del Producto Bruto es conveniente advertir que puede prestarse a manejos caprichosos, según sea la base de los índices deflatores. En otras palabras, el Producto Bruto Real, mostrará distintas tasas de crecimiento según la base de los deflatores. Aparentemente esto no tendría por qué ocurrir, ya que un cambio aritmético de la base del índice deflactor no afectaría las variaciones reales en el Producto Bruto. Mediante un ejemplo que pretende exagerar la situación antes que representar un hecho real, se ilustra este planteamiento.

Para abreviar, supóngase que la actividad económica ha sido dividida en dos sectores: primario y secundario. Los datos son los siguientes:

	Producto Bruto Nominal (u.m. corrientes)	
	1956	1962
Sector primario	1.000	1.600
Sector secundario	1.000	1.800

	Índices deflatores (base 1956 = 100)	
	1956	1962
Sector primario	100	110
sector secundario	100	300

Si se calcula el Producto Bruto real, en precios constantes de 1956, se tiene:

	Producto Bruto Real (u.m. constantes)	
	1956	1962
Sector primario	1.000	1.450
Sector secundario	1.000	600
	<u>2.000</u>	<u>2.050</u>

Entre ambos años, el crecimiento es 2,5 por ciento, si se valora el Producto en precios de 1956. Si aritméticamente se cambia la base de los deflatores y en consecuencia el Producto se valora a precios de 1962, se llega a una variación porcentual diametralmente opuesta.

/Los nuevos

Los nuevos índices deflatores (con base en 1962) serán los siguientes:

Índices deflatores (base 1962 = 100)

	1956	1962
Sector primario	91,0	100
Sector secundario	33,3	100

Deflactando los Productos nominales, por los índices anteriores se tiene:

Producto Bruto Real (u.m. constantes)

Sector primario	1.100	1.600
Sector secundario	3.000	1.800
	<u>4.100</u>	<u>3.400</u>

Puede observarse un decrecimiento de 17 por ciento en el mismo período. Nótese que para los sectores, considerados en forma aislada, no ocurre esta incompatibilidad, ya que sólo se presenta al comparar sumas de sectores que tienen crecimientos nominales distintos y cuyos respectivos deflatores también muestran variaciones desiguales. Los resultados serán tanto más contradictorios, en uno y otro caso, cuanto más distintos sean los índices deflatores y mayores sus variaciones. Lo que ocurre es que se está sumando valores que no son rigurosamente homogéneos, dados los cambios aritméticos de la base de los índices. Los resultados a que se ha llegado no son sorprendentes si se piensa en la limitación de los cambios de base aludidos. Por otra parte la estructura del Producto Bruto nominal es distinta en ambos períodos. En el primer caso los dos índices deflatores son iguales a 100 en 1956 en circunstancias que el Sector Primario aporta con el 50 por ciento al Producto Bruto y en el 50 por ciento el Sector Secundario. En el segundo caso los dos índices deflatores son iguales a 100 en 1962 y el Sector Primario aporta con un 47 por ciento al Producto y con 53 por ciento el sector secundario. Esta es la razón de que para los sectores considerados en forma aislada, no se presenten las inconsistencias anotadas, y sí se presenten al comparar sumas deflactadas por índices distintos. La inconsistencia será mayor mientras más distintas sean las variaciones de los deflatores.

3. La proyección sobre la base de índices de cantidad. Un procedimiento corrientemente utilizado para proyectar el Producto Bruto real por sectores, es sobre la base de índices de cantidad. Consiste en hacer variar el producto de un período dado en conformidad al índice de producción correspondiente. Así, si se sabe que para 1962, el producto generado por la Industria fue de 5 000 unidades monetarias de ese año y se dispone del índice de producción industrial para los años 1962 a 1965, se puede suponer que habrá correspondencia entre las variaciones de dicho Producto sectorial y del índice de cantidad.

Años	Producto Industrial (precios de 1962)	Índice de Producto Ind. base 1962 = 100	Estimación del Producto Ind. (precios de 1962)
1962	5 000	100	5 000
1963	-	198	5 400
1964	-	120	6 000
1965	-	135	6 750

Este tipo de estimaciones, para períodos cortos de tiempo parece ser justificada. Lo que no hay que perder de vista es que el índice de producción muestra más bien las variaciones de la Producción Global antes que del Producto (V. Agregado); la metodología de estimación presentada tiene como supuesto la constancia en los coeficientes de insumo producto, constancia que puede no ser real en una época de tanto cambio tecnológico.

G. Índices de Comercio Exterior

En lo referente al intercambio de bienes y servicios con el resto del mundo, es indispensable cuantificar indicadores acerca de precios, cantidades, valores, etc., relacionados con exportaciones e importaciones. El tipo de problemas que debe enfrentarse en estos casos, son los inherentes a los números índices, más algunas precauciones que es necesario tomar por circunstancias que atañen al comercio exterior.

1. Índices de precios. En general es necesario homogeneizar los datos básicos, ya que no siempre están sujetos a criterios uniformes de valuación: precios CIF, FOB, precios expresados en monedas diferentes, valores nominales de retorno u otro tipo de valorización,

/en que

en qué momento se considera que la importación o exportación está consumada: en tránsito, en aduana etc.

Para el cálculo de índices de precios, puede utilizarse fórmulas de Laspeyres y Paasche. Sin embargo, al utilizar la fórmula de Laspeyres si algún producto deja de transarse, implica suponer que su precio bajó a cero, lo que constituye un evidente falseamiento. Habrá que cuidar que sólo se tomen en cuenta en el cálculo de los productos que se transan tanto en el período base como en el período dado. Por otra parte, no es necesario tomar esta precaución, si se aplica la fórmula de Paasche, ya que ésta se elimina automáticamente; el producto que deja de transarse y no quedará comprendido en ninguno de los factores $p_n q_n$ ni $p_n q_o$. Por lo anterior, para índices de precios se acostumbra utilizar fórmulas de Paasche. En estadísticas de comercio exterior, se acostumbra designar a estos índices como índices de valor unitario, por el tipo de informaciones que se tienen. No es el precio y la cantidad de un artículo, si no el precio promedio o valor unitario que correspondería a un artículo proveniente de una partida de artículos no totalmente homogéneos. Así si la importación de 100 automóviles de distintas marcas representa un valor CIF de 400.000 unidades monetarias, el precio promedio por automóvil es de 4.000 u.m. Es en este sentido que se prefiere designar a estos índices como índices de valor unitario en vez de precios, aunque metodológicamente no hay diferencias.

2. Índices de cantidad. Nuevamente aquí hay que hacer referencia a la denominación especial de índices de quantum que se les da en comercio exterior por razones similares a las anotadas en el punto anterior.

En este caso, si un producto deja de importarse o exportarse, quiere decir que la cantidad transada bajó a cero, lo que queda ahora bien reflejado en la fórmula de Laspeyres. Por esa razón se acostumbra utilizarla en el cálculo de índices de quantum de comercio exterior. Además, el utilizar fórmulas de Laspeyres o Paasche para quantum y valor unitario respectivamente garantiza el cumplimiento de la prueba de reversión de factores planteada por Fisher. En virtud de ello, basta con conocer el valor total de exportaciones o importaciones y uno cualquiera de los dos índices mencionados, para deducir inmediatamente el otro. Recuérdese que:

IPP x IQL = IV

La laboriosidad que representa el cálculo de estos índices induce a la utilización de muestras, en vez de indagaciones totales.

En el caso de índices de precios se supone que los precios de los artículos incluidos en la muestra, representan los precios de los no incluidos. Naturalmente que será necesario calcular los errores de muestreo correspondientes para hacer un racional uso del indicador. Por lo que toca a los índices de quantum, el hecho de que en todos los períodos se disponga del valor total de las importaciones, permite cuantificar la representatividad que tenga la muestra utilizada en cada período mediante el cociente.

$$\frac{\sum p_n q_n \text{ (para la muestra)}}{\sum p_n q_n \text{ (para el total)}} = \text{Representatividad}$$

Posteriormente, si se desea calcular un índice de quantum de Laspeyres, se cuantifica $\sum q_n p_n$ en la muestra, y para proyectar al total, se divide por la representatividad (implica la aplicación de una regla de tres simple). De esa manera se tiene la estimación del numerador de la fórmula para el total poblacional. En cuanto al denominador no hay problema alguno, ya que en cada período se dispone del total de importaciones y exportaciones. Si se desea calcular un índice de quantum de Paasche, será necesario ajustar al 100 por ciento la expresión del denominador de la fórmula, calculada con datos muestrales.

H. Algunos indicadores económicos

Se presentarán algunos indicadores de muy frecuente uso en la literatura económica. Si bien es cierto que la mayoría de ellos serán analizados con mucha más profundidad en el curso de Contabilidad Social, es conveniente examinarlos desde el punto de vista estadístico.

1. Índice de la relación de términos del intercambio. Se define como el cociente entre el índice de precios de exportaciones y el índice de precios de importaciones, ambos referidos a la misma base. En consecuencia, este estadígrafo indica la evolución en el tiempo de la relación de precios que se dio en el período base; en otros

/términos representa

términos representa variaciones en la capacidad de compra de un volumen de exportaciones. Da la respuesta a la siguiente interrogante en cuánto ha aumentado o disminuido el número de unidades que es necesario exportar, para financiar la importación del mismo volumen de productos que se importaba en el año base. Se trata de un concepto estrictamente relativo. No se puede concluir que la relación de intercambio sea buena o mala, sino mejor o peor que la que existía en el año base. El índice de la relación de precios de intercambio queda entonces definido como:

$$\text{IRI} = \frac{\text{Indice de precios de exportaciones}}{\text{Indice de precios de importaciones}} = \frac{\text{IPX}}{\text{IPM}}$$

2. Producto e Ingreso bruto. Cuando estos conceptos se consideran en valores constantes, ambas magnitudes no coinciden. El Producto es una medida del valor de los bienes y servicios debidos al esfuerzo productivo interno. El Ingreso, tiene que ver con el intercambio con el exterior, desde el momento que parte de la producción de un país se exporta y parte de los insumos y bienes de consumo deben adquirirse en el exterior. Por consiguiente estas transacciones con el exterior afectan, por la relación de precios de intercambio, la disponibilidad de bienes y servicios que satisfacen la demanda de la población. A medida que se deteriora la relación de intercambio, hay una transferencia al exterior de parte del esfuerzo productivo interno. En valores constantes, la diferencia entre ambos conceptos, recibe el nombre de efecto de la relación de precios del intercambio, luego:

$$\text{Ingreso Bruto} = \text{Producto Bruto} + \text{Efecto de la relación de precios de intercambio.}$$

En símbolos:

$$\text{IB} = \text{PB} + \text{EFI}$$

donde:

Efecto de la relación de precios del intercambio = poder de compra de exportaciones - quantum de exportaciones.

Es decir:

$$\text{EFI} = \text{PEX} - \text{QX}$$

Además:

/Poder de

Poder de compra de exportaciones = Quantum de exportaciones x
x índice de la relación de intercambio

$$PEX = QX (IRI)$$

Dado que el producto del Quantum de exportaciones y el índice de precios de exportaciones es equivalente al valor corriente de las exportaciones, el Poder de compra también puede definirse como:

$$PEX = \frac{\text{Valor corriente de las exportaciones}}{\text{Índice de precios de importaciones}}$$

en otros términos, el valor nominal o corriente de las exportaciones queda deflactado por el índice de precios de las importaciones. Por otra parte el Quantum de las exportaciones es el valor de las exportaciones a precios constantes.

$$QX = \sum q_n p_o$$

donde la sumatoria se extiende a todos los rubros de exportación, es decir:

$$QX = \frac{\text{Valor corriente de exportaciones}}{\text{Índice de precios de exportaciones}}$$

Reemplazando las expresiones correspondientes en la fórmula de EFI se tiene:

$$EFI = QX (IRI) - QX = QX [IRI - 100]$$

A continuación se cuantificarán estos conceptos mediante un ejemplo, para hacer resaltar su importancia. Supóngase que se tienen las siguientes informaciones:

	1960	1962	1964
a) Producto Bruto (u.m. corrientes)	1.000	1.200	1.500
b) Deflactor público (base 1960 = 100)	100	115	140
c) Índice de precios de exportaciones	100	110	110
d) Índice de precios de importaciones	100	120	130
e) Valor corriente de las exportaciones	200	240	280

Para calcular el EFI, se tiene

	1960	1962	1964
f) PB real (u.m. de 1960) (a/b)	1.000	1.043	1.071
g) IRI (base 1960 = 100) (c/d)	100	92	85
h) QX (u.m. de 1960) (e/c)	200	218	255
i) QX IRI = PEX (h.g)	200	201	217
j) EFI (i - h)	-	- 17	- 38
k) YB real (f - j)	100	1.026	1.033

3. Capacidad para importar. Cuando se piensa en términos de valores corrientes, la capacidad para importar no es otra cosa que el valor total de las exportaciones, más el ingreso neto de capitales extranjeros.

La capacidad para importar a precios constantes estará dada por:

Quantum de exportaciones	QX
+ Radicación de capitales extranjeros	RC
+ Efecto de la relación de precios del intercambio	EFI
Capacidad total de pagos sobre el exterior	
- Remesas de utilidades e intereses	
- Salida de capitales extranjeros	
Capacidad para importar	CPI

La capacidad para importar a precios constantes, también puede determinarse deflactando la capacidad para importar a precios corrientes por el índice de precios de importaciones. La capacidad para importar a precios constantes es:

$$CPI = QX + QX \left[\frac{IRI - 100}{100} \right] + RC \text{ (neta, a precios corrientes)}$$

$$CPI = QX \left[\frac{IRI}{100} \right] + \frac{RC \text{ (neta a precios corrientes)}}{IPM}$$

$$CPI = QX \cdot \frac{IPX}{IPM} + \frac{RC}{IPM}$$

$$CPI = \frac{\text{Valor cte. de las export.} + RC \text{ (neta a precios ctes)}}{\text{Índice de precios de importaciones}}$$

$$CPI = \frac{\text{Capacidad para importar a precios corrientes}}{\text{Índice de precios de importaciones}}$$

4. Tipo de cambio de paridad. Como una aplicación de números es conveniente tratar la forma de convertir monedas de distintos países a unidades homogéneas con propósitos de comparación. Utilizar los tipos de cambio oficiales para ello, puede originar distorsiones serias, en la medida que exista más de un tipo de cambio (discriminación de áreas, cambiarias), o que el tipo de cambio unico esté sobre o subvaluado.

Una alternativa teórica consistiría en la determinación de una canasta de productos, resultando el tipo de cambio entre dos monedas, como la relación de valores que sería necesario gastar para

/adquirir dicha

adquirir dicha canasta. Este método tiene inconvenientes de tipo práctico; la determinación de los artículos que conformarían la canasta significa un serio problema, sobre todo si se piensa en la enorme variación de las preferencias de los consumidores en los distintos países.

Un método que puede ser utilizado con alguna ventaja, es la proyección de un tipo de cambio base, en un período donde no se haya advertido sobre o subvaluaciones monetarias, sin cambios múltiples y en general sin accidentes que descalifiquen a ese período base. La aludida proyección se efectúa sobre la base de las modificaciones en la relación de precios de los dos países, cuyo tipo de cambio se pretende determinar.

Sean los países A y B, y se desea estimar el número de unidades monetarias de A, por unidad monetaria de B. El tipo de cambio de paridad en un año cualquiera n, estará dado por

$$\text{Tipo de cambio de paridad año } n = \frac{\text{Tipo de cambio año base} \cdot \frac{\text{Deflactor implícito de A}}{\text{Deflactor implícito de B}}}{1}$$

Ejemplo: Supóngase que el tipo de cambio base en el año 0, fué de 5 u.m. de A por 1 u.m. de B, siendo los deflatores implícitos los siguientes:

Año	D.I. país A	D.I. país B
0	100	100
1	120	110
2	150	120
3	200	140
4	300	150

El tipo de cambio de paridad para los años siguientes será:

Año	Relación de deflatores $\frac{DI_A}{DI_B}$	Tipo de cambio de paridad
0	100,0	5,00
1	109,1	5,45
2	125,0	6,25
3	142,9	7,14
4	200,0	10,00

/Evidentemente,

Evidentemente, el método da estimaciones, que serán tanto más confiables en la medida que el tipo de cambio base haya sido elegido adecuadamente, y los deflatores implícitos sean representativos de las variaciones de precios en ambos países.

5. Transferencias implícitas. El hecho que en la actividad económica se presenten transacciones intersectoriales en que cada sector tiene precios distintos, origina cierto tipo de transferencias de producto de un sector a otro, implícitas en las transacciones que efectúan cuanto se contabilizan a precios constantes.

Aquellos sectores cuyos precios crecen menos que el promedio general de precios representado por el deflactor implícito, están transfiriendo parte de su producto hacia aquellos sectores que tienen precios que crecen más que el promedio de precios.

Evidentemente que la suma algebraica de las transferencias será nula, por cuanto la ganancia de unos sectores tiene como contrapartida la pérdida de otros.

Se definen las transferencias implícitas para cada sector, como la diferencia entre la producción valorizada a precios del sector y esa misma producción valorizada a precios promedios representados como se dijo, por el deflactor implícito:

Transferencias implícitas = Producción a precios corrientes - Producción a precios constantes (Deflactor implícito)

en símbolos:

$$T \text{ Imp} = p_n q_n - p_o q_n \text{ (DI)}$$

donde n y o son el período que se calcula y el período base respectivamente.

$$\sum_{h=1}^L T \text{ Imp} (h) = \sum_{h=1}^L p_n q_n - \sum_{h=1}^L p_o q_n \text{ (DI)} = 0$$

Donde L es el número de sectores considerados, h = 1, 2, 3, L

$$\sum_{h=1}^L T \text{ Imp} (h) = \text{Producción nominal} - \text{Producción real} \cdot \frac{\text{Producción no.}}{\text{Producción real}}$$

- 0

/Ejemplo:

Ejemplo:

Sectores	Producción nominal		Indice de cantidad base 0 = 100	Producción real	Indice de precios base 0 = 100
	año 0	año 1			
1	500	1.080	120	600	180
2	600	800	110	660	121
3	300	350	100	300	117
4	400	420	90	360	117
		<u>2.650</u>		<u>1.920</u>	

El deflactor implícito para el año 1 será:

$$DI = \frac{\text{Producción nominal año 1}}{\text{Producción real año 1}} = \frac{2.650}{1.920} = 138$$

Las transferencias implícitas por sector:

Sectores	Producción nominal	Producción real (DI)	Transferencias implícitas
	$p_1 q_1$	$p_0 q_1$ (DI)	
1	1.080	828	+ 252
2	800	911	- 111
3	350	414	- 64
4	420	497	- 77
	<u>2.650</u>	<u>2.650</u>	-

I. Etapas de la construcción de números índices

Es interesante e ilustrativo seguir cada uno de los pasos para la confección de indicadores acerca de precios o cantidades. Los pasos que se detallarán estarán referidos al diseño de un índice de costo de vida por constituir un caso bastante general y complejo.

1. Objetivo del índice: Es fundamental calificar con toda precisión el objetivo del indicador. En el caso de un índice de costo de vida, es necesario determinar a quiénes estará referido ese índice: si a la población en su conjunto, a una ciudad, a profesionales, a obreros, a campesinos, etc. De esto dependerá qué tipo de artículos conformarán la muestra de productos componentes del índice. Es imprescindible no perder de vista el objetivo básico: los usos principales que tendrá el indicador.
2. Determinación de la estructura de consumo. Es conveniente, al calcular un índice de costo de vida, clasificar los gastos: alimentación, vestuario, habitación, varios, etc. De esta manera es posible calcular índices separadamente para cada componente. Este desglose, aparte de que permite realizar análisis de las variaciones de precios en forma más detallada es muy útil en la selección de deflatores adecuados. Estos índices desglosados son susceptibles de combinación, para la obtención del índice general. Las ponderaciones de cada componente se establecen mediante una muestra. Se selecciona una muestra, de unidades familiares en el caso del costo de vida, provenientes de la población para la cual se confecciona el índice por ejemplo, la población de obreros y empleados de una ciudad. Esta muestra, en lo posible debe ser aleatoria y de un tamaño que represente fidelidad y alta confianza. En cada una de las unidades elegidas en la muestra, se lleva un registro de los gastos en bienes y servicios por tipo de bien o servicio, donde se recopilen los precios pagados y las cantidades consumidas. Vale la pena llevar este registro durante un tiempo largo, por lo general de un año de manera de captar las variaciones en los consumos en las diferentes estaciones. De esta manera es posible llegar a obtener ponderaciones por componentes y por tipos de bienes y servicios.

3. Selección de artículos. En las anotaciones que se hacen en los registros o "Libretas de consumo" acerca de los distintos bienes que se consumen, en general aparecerá una gran cantidad de productos, muchos de ellos obedecerán a consumos esporádicos o accidentales. Se hace necesario seleccionar los productos más importantes, de consumo habitual y representativos de las preferencias de los consumidores. Para seleccionar estos productos y servicios, no es conveniente un método aleatorio, es preferible decidir qué productos serán considerados en el índice, tomando en cuenta el volumen del gasto en cada bien o servicio. En esa forma se llegará a una lista de 100 a 300 productos para los cuales se tiene tabulada su importancia relativa o ponderación.
4. Formas de valuación. Otra etapa importante en la confección de índices de precios, es la decisión acerca de los tipos de precios que se considerarán. Sabido es que los precios varían enormemente según la fuente donde se adquieren los productos; por otra parte es corriente que para artículos considerados de primera necesidad, hayan precios oficiales y precios reales muy distantes entre sí. En el caso de índices de costo de vida los precios debieran tomarse en las fuentes donde el consumidor adquiere los productos: almacenes, ferias, etc. Sobre el particular es necesario tomar decisiones previas en forma categórica.
5. Variaciones de calidad de los bienes y servicios. Otro aspecto fundamental es la especificación precisa, hasta donde sea posible, acerca de la calidad de los productos, de manera que se pueda controlar a través del tiempo la invariabilidad de sus características. Es muy frecuente, sobre todo en los artículos controlados, que se "incrementen" los precios reales, permaneciendo fijos los precios nominales, vía una disminución de la calidad de los productos.
6. Base del índice. En general cuando se planteó el tema se advirtió sobre la necesidad de elegir como base un período donde estuvieran ausentes las modificaciones y circunstancias violentas. Ahora hay que agregar que en el caso de índices de costo de vida, el período base debe ser modificado, toda vez que la estructura de
/consumo haya

consumo haya cambiado significativamente. En los tiempos actuales, donde las innovaciones tecnológicas cambian con extraordinaria rapidez, determinando cambios en las preferencias de los consumidores la base de un índice de costo de vida no debiera tener una antigüedad superior a los 6 a 8 años.

Sin embargo, en las fórmulas de cálculo, para evitar cambios de base muy seguidos que significan altos costos, se contempla la posibilidad de introducir nuevos artículos, toda vez que se detecten cambios en las preferencias de los consumidores.

7. Elección de los métodos de cálculo. En general es necesario tomar decisiones sobre la fórmula de cálculo. Corrientemente, cuando se trata de índices de costo de vida se acostumbra utilizar fórmulas de Laspeyres, porque implica considerar constantes las ponderaciones del año base. La utilización de una fórmula de Paasche significaría un esfuerzo muy grande, ya que en cada período (generalmente cada mes) habría que reponderar "hacia atrás", aparte de que sería necesario calcular estas ponderaciones a medida que pasa el tiempo.

La decisión acerca de qué fórmula se utilizará estará condicionada en primer lugar al objetivo que se persigue con el índice y a la factibilidad de su uso desde el punto de vista del costo y la oportunidad con que se entreguen los resultados.

En Chile, la Dirección de Estadística y Censos calcula el índice de costo de vida sobre la base de una modificación a la fórmula de Laspeyres.

$$I_i = I_{i-1} \left(\frac{\sum q_0 p_{i-1} \left(\frac{p_i}{p_{i-1}} \right)}{\sum q_0 p_{i-1}} \right)$$

Esta fórmula implica el cálculo del índice en un período i, sobre la base del índice en el período anterior, afectándolo por la variación que acusen los precios. La fórmula tiene la ventaja de que se pueden introducir nuevos artículos en términos de sus especificaciones, o cambiar la fuente de información.

III ANALISIS DE REGRESION

A. METODO DE LOS MINIMOS CUADRADOS.

Probablemente uno de los temas estadísticos de mayor utilización en la planificación, es el referente al análisis de regresión y correlación.

Es de extraordinaria utilidad conocer la forma en que están relacionadas las variables que son objeto de análisis, es decir, la función matemática capaz de representar tal relación.

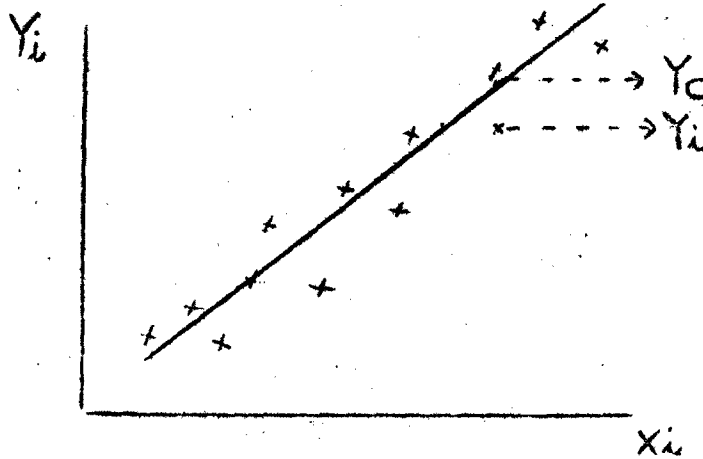
Conociendo tal función, es posible estimar el comportamiento de la variable objeto de estudio denominada variable dependiente o predictando, en términos de las variaciones de otra variable denominada independiente o predictor. De lo anterior se deduce que la regresión debe aplicarse a variables que tengan una relación lógica, es decir, que exista razonablemente dependencia entre las variables. Teóricamente, a cualquier par de variables puede encontrárseles una función matemática o ecuación de regresión que las relacione, pero sólo será de utilidad en el caso que haya una relación de causalidad entre dichas variables.

- 1.- Regresión simple: Se denomina de esta manera, a la metodología de obtención de ecuaciones, donde sólo intervienen dos variables: una dependiente o predictando y otra independiente o predictor. Una vez que por medio del análisis lógico se ha comprobado la existencia de una relación de causalidad entre las variables es necesario determinar cuál es la función matemática que representa adecuadamente la relación. Para ello es indispensable disponer de informaciones acerca de los valores que ha alcanzado cada una de las variables en distintos períodos, si se trata de un análisis histórico cronológico o en distintos lugares si se trata de un corte transversal en el tiempo. Con las informaciones obtenidas, que deben ser en un número suficiente para garantizar un buen ajuste, se construirá un gráfico y se podrá decidir si la función adecuada es una recta, una hipérbola, una parábola, etc.

/Una vez

Una vez que se ha decidido cuál es la función adecuada para el ajuste de regresión, es posible determinar los parámetros de la función elejida.

a) Línea recta: Si al representar los puntos en un gráfico, éstos representan satisfactoriamente una recta como en el cuadro siguiente:



es necesario calcular los parámetros o coeficientes de regresión de dicha recta.

$$Y_C = a x + b$$

para poder determinar los valores de a y b, se recurre al método de los mínimos cuadrados, que cumple la condición de minimizar la siguiente expresión:

$$\sum_{i=1}^n (Y_i - Y_C)^2$$

donde Y_i : es un valor observado

Y_C : es un valor calculado por la ecuación de regresión

n: es el número de observaciones

Si se reemplaza Y_C por $ax_i + b$ dentro de la sumatoria, es posible, derivando, encontrar los valores de los coeficientes de regresión a y b que satisfacen la condición. En efecto, llamemos Z a la

/expresión: $Z =$

expresión:

$$Z = \sum (Y_i - a X_i - b)^2$$

Se trata de derivar parcialmente respecto de cada uno de los parámetros

$$\frac{\partial Z}{\partial b} = 2 \sum (Y_i - a X_i - b) (-1) = 0$$

Aplicando las propiedades de la sumatoria se tiene:

$$\sum Y_i = a \sum X_i + n b$$

que es la primera ecuación normal.

$$\frac{\partial Z}{\partial a} = 2 \sum (Y_i - a X_i - b) (-X_i) = 0$$

Aplicando propiedades de la sumatoria

$$\sum Y_i X_i = a \sum X_i^2 + b \sum X_i$$

que es la segunda ecuación normal.

Obsérvese que se tienen dos ecuaciones normales y dos incógnitas. Se trata de un sistema de ecuaciones que permiten calcular los parámetros o coeficientes de regresión.

1^a Ecuación Normal: $\sum Y_i = a \sum X_i + nb$

2^a Ecuación Normal: $\sum Y_i X_i = a \sum X_i^2 + b \sum X_i$

} Sistema

Donde $\sum Y_i$, es la suma de los valores observados de la variable

dependiente; $\sum X_i$, es la suma de los valores observados de la variable independiente y n es el número de observaciones. En este caso el

sistema está formado por dos ecuaciones, porque sólo hay dos parámetros por determinar. El signo del coeficiente de regresión que corresponde con la pendiente de la recta (a), determina si la regresión es directa

/o inversa

o inversa. Si "a" es positivo, quiere decir que ante incrementos de la variable predictor, corresponde incrementos de la variable predictando. Si el signo de "a" es negativo, ante incrementos de la variable predictor habrá decrementos de la variable predictando y se dice que la regresión es inversa.

Hasta el momento se ha estado planteando una regresión de " Y en X ", es decir, considerando a Y como variable dependiente y a X como variable independiente. Donde se trataba de minimizar:

$$\sum_{i=1}^n (Y_i - Y_C)^2$$

Puede perfectamente plantearse una regresión de " X en Y ", donde lo que interesa minimizar es:

$$\sum_{i=1}^n (X_i - X_C)^2$$

siendo $X_C = a Y_i + b$

Las ecuaciones normales en este caso, por analogía serán:

$$\sum X_i = a \sum Y_i + nb$$

$$\sum X_i Y_i = a \sum Y_i^2 + b \sum Y_i$$

Téngase presente que los parámetros de la regresión de " Y en X ", serán distintos a los parámetros de la regresión de " X en Y ". Por ello suele distinguirse a estos parámetros de la siguiente manera:

a_{YX} : coeficiente de regresión de Y en X

a_{XY} : coeficiente de regresión de X en Y

En general, al analizar la relación de las variables cuya regresión se pretende determinar, se puede especificar cuál es la variable dependiente y cuál es la variable independiente. Una vez tomada la decisión, se denominará con Y_i a la variable dependiente o predictando y con X_i a la

/variable independiente

variable independiente Y o predictor, para evitar confusiones. A continuación se plantea un ejemplo que permitirá aclarar algunos aspectos que son difíciles de expresar literalmente. El lenguaje de los símbolos es claro y no permite malos entendidos ni interpretaciones equivocadas.

Ejemplo: En los últimos años las ventas de una empresa han crecido por razones de una intensa campaña de promoción de ventas. Las variables en cuestión han tenido el siguiente comportamiento en el tiempo.

Años	Ventas Y_i	Gasto en Propaganda X_i
1958	100	10
1959	150	14
1960	200	21
1961	210	22
1962	300	28
1963	500	45
1964	600	55

Interesa determinar la función matemática o ecuación de regresión que relaciona a estas variables. Representando estos valores en un gráfico, se concluirá que la recta representa adecuadamente la relación de las variables. Para determinar los parámetros de la recta, se plantean las ecuaciones normales:

$$\sum Y_i = a \sum X_i + nb$$
$$\sum Y_i X_i = a \sum X_i^2 + b \sum X_i$$

Luego es necesario tabular los valores que interesan para reemplazar en estas ecuaciones normales. A continuación se procede con este punto.

Y_i X_i

Y_i	X_i	$Y_i X_i$	X_i^2
100	10	1000	100
150	14	2100	196
200	21	4200	441
210	22	4620	484
300	28	8400	784
500	45	22500	2025
<u>600</u>	<u>55</u>	<u>33000</u>	<u>3025</u>
2060	195	75820	7055

Las ecuaciones normales en valores serán:

$$2060 = 195 a + 7 b$$

$$75820 = 7055 a + 195 b$$

Resolviendo el sistema

$$a \hat{=} 11,4$$

$$b \hat{=} - 26,1$$

La ecuación de ajuste queda en consecuencia así expresada:

$$Y_C = 11,4 X_i - 26,1$$

Por medio de esta ecuación se puede determinar valores calculados de la variable dependiente ante cualquier valor de la variable independiente. Naturalmente que al realizar estimaciones, por ejemplo para calcular el probable volumen de ventas ante un desembolso en propaganda de 100 ($X_t = 100$), hay que tener en cuenta el campo de validez de la regresión. No escapará a la atención del lector el hecho de que aumentos sucesivos de propaganda no siempre acarrearán mayores volúmenes de venta, porque puede llegar un momento de saturación del mercado u otra objeción por el estilo. Es necesario en consecuencia al realizar estimaciones, la verificación del cumplimiento de los supuestos implícitos en los datos disponibles. Por ello al realizar una estimación, es indispensable advertir que sólo tendrá validez, si es que se sigue manteniendo la tendencia de los puntos observados en el período histórico.

/b) Potencial

b) Potencial. Una función muy utilizada en proyecciones, por su flexibilidad, es la denominada función potencial o de elasticidad. Su expresión matemática es la siguiente:

$$Y = b x^a$$

Para determinar las ecuaciones normales se procede en forma similar al caso de la recta, linealizando previamente mediante la aplicación de logaritmos:

$$\log Y_C = \log b + a \log X_i$$

$$\log Y_C = b' + a \log X_i \quad \text{donde } b' = \log b$$

En este caso se trata de minimizar la expresión:

$$Z = \sum_{i=1}^n (\log Y_i - \log Y_C)^2$$

es decir:

$$Z = \sum (\log Y_i - a \log X_i - b')^2$$

Derivando respecto de cada uno de los parámetros e igualando los resultados a cero, se obtendrán las dos ecuaciones normales.

$$\frac{\partial Z}{\partial b'} = 2 \sum (\log Y_i - a \log X_i - b') (-1) = 0$$

$$\frac{\partial Z}{\partial a} = 2 \sum (\log Y_i - a \log X_i - b') (-\log X_i) = 0$$

Aplicando las propiedades de la sumatoria a ambas derivadas, se tiene:

$$\sum \log Y_i = a \sum \log X_i + nb'$$

$$\sum \log Y_i \log X_i = a \sum (\log X_i)^2 + b' \sum \log X_i$$

que forman el sistema de dos ecuaciones normales que permitirán el cálculo de los dos parámetros. Evidentemente el método es un tanto

/laborioso cuando

laborioso cuando se tienen muchas observaciones ya que es necesario trabajar en los logaritmos con al menos 5 decimales para evitar aproximaciones que pueden implicar serios desajustes.

c) Exponencial. Principalmente, cuando se desea calcular tasas de crecimiento tomando en cuenta todos los puntos observados en el período histórico se recurre a la función:

$$Y = a b^t \text{ donde } b = 1 + i$$

t: tiempo en períodos

Aplicando logaritmos a la anterior expresión:

$$\log Y_C = \log a + t_i \log b$$

Como en los casos anteriores interesa minimizar la expresión

$$Z = \sum_{i=1}^n (\log Y_i - \log Y_C)^2$$

$$Z = \sum (\log Y_i - \log a - t_i \log b)^2$$

$$\frac{\partial Z}{\partial \log a} = 2 \sum (\log Y_i - \log a - t_i \log b) (-1) = 0$$

$$\frac{\partial Z}{\partial \log b} = 2 \sum (\log Y_i - \log a - t_i \log b) (-t_i) = 0$$

Aplicando las propiedades de la sumatoria, se obtienen las dos ecuaciones normales.

$$\sum \log Y_i = n \log a + \log b \sum t_i$$

$$\sum t_i \log Y_i = \log a \sum t_i + \log b \sum t_i^2$$

El caso general de la función exponencial, es el cálculo de tasas de crecimiento cuando se considera el tiempo como variable independiente. Sin embargo, puede considerarse cualquier otra variable y ajustar la función sin hacer referencia a tasas de crecimiento.

d) Parábola: Esta conocida función, se ajusta en forma similar a los casos anteriores.

$$/ Y = a x^2$$

$$Y = a x^2 + b x + C$$

Dado que la forma general contiene tres parámetros, será necesario determinar tres ecuaciones normales para determinar los valores de a, b, c. Estas tres ecuaciones normales provienen de la derivación parcial respecto de cada uno de dichos parámetros. Interesa minimizar la expresión:

$$Z = \sum_{i=1}^n (Y_i - Y_C)^2$$

$$Z = \sum (Y_i - a x_i^2 - b x_i - c)^2$$

Derivando respecto de a, b, y c, se tiene:

$$\frac{\partial Z}{\partial c} = 2 \sum (Y_i - a x_i^2 - b x_i - c) (-1) = 0$$

$$\frac{\partial Z}{\partial b} = 2 \sum (Y_i - a x_i^2 - b x_i - c) (-x_i) = 0$$

$$\frac{\partial Z}{\partial a} = 2 \sum (Y_i - a x_i^2 - b x_i - c) (-x_i^2) = 0$$

Aplicando las propiedades de la sumatoria, se tienen las siguientes ecuaciones normales:

$$1^{\text{a}} \text{ Ecuación normal: } \sum Y_i = a \sum X_i^2 + b \sum X_i + n C$$

$$2^{\text{a}} \text{ Ecuación normal: } \sum Y_i X_i = a \sum X_i^3 + b \sum X_i^2 + C \sum X_i$$

$$3^{\text{a}} \text{ Ecuación normal: } \sum Y_i X_i^2 = a \sum X_i^4 + b \sum X_i^3 + C \sum X_i^2$$

En vista que en el período histórico, se tienen los valores de Y_i y X_i , es necesario tabular todas las sumatorias que aparecen en las ecuaciones normales. Resolviendo el sistema, se tiene determinado el valor de cada uno de los tres parámetros.

/e) Hipérbola

c) Hipérbola equilátera: Para el ajuste de algunas funciones de demanda, y por la propiedad que tiene de que cualquier punto de la función subtiende superficies iguales con los ejes de coordenadas, su aplicación es bastante frecuente. Su expresión matemática es:

$$Y_C = \frac{a}{x}$$

En vista que sólo tiene un parámetro, será necesario calcular una ecuación normal, minimizando la expresión:

$$Z = \sum (Y_i - Y_C)^2$$

$$Z = \sum \left(Y_i - \frac{a}{X_i} \right)^2$$

Derivando respecto de a

$$\frac{\partial Z}{\partial a} = 2 \sum \left(Y_i - \frac{a}{X_i} \right) \left(-\frac{1}{X_i} \right) = 0$$

Aplicando las propiedades de la sumatoria se tiene:

$$\text{Ecuación Normal } \sum Y_i / X_i = a \sum 1/X_i^2$$

f) Otras funciones. Dentro de la investigación económica, a veces es preciso ajustar particulares funciones. La metodología de la obtención de ecuaciones normales es similar a los casos vistos. Por ejemplo la función:

$$\log Y_C = a \bar{x} + b$$

Siempre se tratará de minimizar la expresión

$$Z = \sum_{i=1}^n (\log Y_i - \log Y_C)^2$$

$$Z = \sum (\log Y_i - a x_i - b)^2$$

/ donde

Donde:

$$\frac{\partial Z}{\partial b} = 2 \sum (\log Y_i - a x_i - b) (-1) = 0$$

$$\frac{\partial Z}{\partial a} = 2 \sum (\log Y_i - a x_i - b) (-X_i) = 0$$

Las ecuaciones normales serán:

$$\sum \log Y_i = a \sum X_i + n b$$

$$\sum X_i \log Y_i = a \sum X_i^2 + b \sum X_i$$

Resolviendo el sistema, es posible determinar el valor de los parámetros.

Siempre es conveniente seguir esta metodología, para funciones en que sus derivadas no compliquen demasiado las expresiones que aparecen en las ecuaciones normales.

2.- Regresión múltiple. Ocurre que a veces es necesario encontrar funciones en que se relacionen una variable dependiente y dos o más variables independientes, de ahí el calificativo de múltiple. En este caso se adoptará una simbología especial, para designar cada una de las variables y parámetros:

X_1 : variable dependiente

$X_2, X_3 \dots X_p$: variables independientes

Así, si se trata de un caso de correlación múltiple donde se consideren dos variables independientes, la función se expresará de la siguiente manera:

$$X_{1.23} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

donde:

$X_{1.23}$

$X_{1.23}$: indica la variable dependiente que se relaciona con las variables X_2 y X_3 , esa es la razón de los subíndices.

$a_{1.23}$: coeficiente de posición (término libre) del plano de regresión donde se consideran la variable dependiente X_1 y las variables independientes X_2 y X_3

$b_{12.3}$: coeficiente de regresión que multiplica a la variable X_2 , cuando además se considera la variable X_3 .

$b_{13.2}$: coeficiente de regresión que multiplica a la variable X_3 , cuando además se considera la variable X_2

Es sencillo extender esta notación al caso en que se consideren 3 o más variables e independientes. En el caso de tres variables independientes (X_2, X_3, X_4) la función quedará así simbolizada:

$$X_{1.234} = a_{1.234} + b_{12.34} X_2 + b_{13.24} X_3 + b_{14.23} X_4$$

El estudiante, por analogía con el caso anterior puede interpretar cada uno de estos símbolos.

Cuando se desea ajustar una función de este tipo a una serie de datos, el método de los mínimos cuadrados implica hacer mínima la expresión.

$$\sum_{i=1}^n (X_{1i} - X_{1.23})^2$$

donde X_{1i} son los valores observados y $X_{1.23}$ son los valores calculados de la variable dependiente. Las ecuaciones normales en el caso de dos variables independientes, se obtiene minimizando la siguiente expresión:

$$Z = \sum (X_{1i} - a_{1.23} - b_{12.3} X_2 - b_{13.2} X_3)^2$$

Para ello, se deriva parcialmente respecto de cada uno de los parámetros, igualando los resultados a cero.

$$\frac{\partial Z}{\partial a_{1.23}} =$$

$$\frac{\partial Z}{\partial a_{1.23}} = 2 \sum (X_1 - a_{1.23} - b_{12.3} X_2 - b_{13.2} X_3)^2 (-1) = 0$$

$$\frac{\partial Z}{\partial b_{12.3}} = 2 \sum (X_1 - a_{1.23} X_2 - b_{13.2} X_3)^2 (-X_2) = 0$$

$$\frac{\partial Z}{\partial b_{13.2}} = 2 \sum (X_1 - a_{1.23} - b_{12.3} X_2 - b_{13.2} X_3)^2 (-X_3) = 0$$

Aplicando las propiedades de la sumatoria se tienen las siguientes tres ecuaciones normales que formarán el sistema para calcular el valor de cada uno de los tres parámetros.

$$\sum X_1 = b_{12.3} \sum X_2 + b_{13.2} \sum X_3 + n a_{1.23}$$

$$\sum X_1 X_2 = b_{12.3} \sum X_2^2 + b_{13.2} \sum X_2 X_3 + a_{1.23} \sum X_2$$

$$\sum X_1 X_3 = b_{12.3} \sum X_2 X_3 + b_{13.2} \sum X_3^2 + a_{1.23} \sum X_3$$

Tabulando los valores de las sumatorias que aparecen en el sistema, se podrá resolver para cada parámetro.

B. CONSIDERACIONES PRACTICAS

En páginas anteriores se ha detallado la metodología de obtención de ecuaciones normales por el método de los mínimos cuadrados, para el tipo de funciones de más frecuente uso en la planificación. Ahora se pretende establecer algunas consideraciones que es necesario tomar, desde el punto de vista práctico, al realizar los mencionados ajustes.

1. Respecto del tipo de función. Si se piensa que una de las principales aplicaciones de la regresión, es la proyección en el tiempo o en el espacio donde no se tengan valores de la variable en estudio, y donde no queda otra alternativa que conformarse con estimaciones provenientes de extrapolación de funciones ajustadas por regresión, deberá admitirse

/la necesidad

La necesidad de disponer de funciones sencillas que contengan un reducido número de variables y parámetros. Recuérdese que una función complicada, de muchas variables y parámetros, se parecerá más bien a una interpolación, a una función que se aproximará al mayor número de puntos observados. Para determinar tendencia no tiene sentido la interpolación. Recuérdese que para proyectar una variable dependiente, es necesario disponer de estimaciones para todas las variables independientes. Disponer de estimaciones para muchas variables independientes suele ser en extremo difícil y en todo caso existe alta probabilidad de cometer errores. En cambio una función sencilla, como las analizadas en páginas anteriores, es susceptible de representar cabalmente una tendencia de la relación de la variable dependiente con la o las variables independientes.

2. Respecto del número de observaciones. Un buen ajuste implica disponer de una cantidad significativa de puntos observados. El conjunto de puntos observados representa una muestra de la relación de las variables en el tiempo o el espacio. Mientras más grande esta muestra, es decir mientras mayor número de puntos se tenga, tendrá más representatividad y menor será la probabilidad de cometer errores. Cuando se está analizando una ecuación de regresión, una de las primeras cuestiones a aclarar debe ser el número de observaciones, para que con este antecedente se califique en parte, lo significativo de la regresión.

3. Respecto de la laboriosidad de cálculo. El estudiante comprobará a través de la realización de seminarios, la laboriosidad que representan los cálculos de regresión. En la práctica, cuando ya se tienen aclarada la parte conceptual, donde constituye una fuerte ayuda la realización de ejercicios en forma manual, es útil recurrir a los computadores electrónicos, ya que una vez entregadas las informaciones originales, en brevísimo tiempo es posible disponer de cálculos exactos, ya que los programas de regresión están previamente diseñados. Por otra parte respecto de la deducción de las ecuaciones normales, puede significar cierta demora su obtención sobre la base de las derivadas parciales. Existe una regla nemotécnica para hallar ecuaciones normales en funciones lineales

/respecto de

respecto de los parámetros. La regla es la siguiente:

Para la primera ecuación normal, multiplique la función a ajustar por el coeficiente del primer parámetro, y luego aplíquese el operador sumatoria a la función. Para la segunda ecuación, multiplíquese toda la función por el coeficiente del segundo parámetro y luego aplíquese el operador sumatoria. Así sucesivamente para todas las ecuaciones normales que se deba obtener.

Ejemplos: Se obtendrán las ecuaciones normales de la función:

$$\log Y = a \log X + \log b$$

Multiplicando ambos miembros de la ecuación por el coeficiente de $\log b$ que es uno y aplicando sumatoria, se tiene la primera ecuación normal.

$$\sum \log Y_i = a \sum \log X_i + n \log b$$

Multiplicando ambos miembros de la ecuación por $\log X$ que es el coeficiente del otro parámetro y aplicando sumatoria, se tiene:

$$\sum \log Y_i \log X_i = a \sum (\log X_i)^2 + \log b \sum \log X_i$$

que es la segunda ecuación normal. Comparando estas dos ecuaciones normales con las obtenidas por derivación parcial en la parte 1, b se concluye que son idénticas.

Si se quisiera obtener la ecuación normal de una recta que pasa por el origen, se tiene:

$Y = a x$ (recta que pasa por el origen ya que no tiene término libre; coeficiente de posición 0.)

Multiplicando por X_i que es el coeficiente del único parámetro y aplicando sumatoria, se tiene:

$$\sum Y_i X_i = a \sum X_i^2$$

Por el proceso de derivación, se llega al mismo resultado. En efecto:

$$Z = \sum_{i=1}^n (Y_i - Y_c)^2 = \sum (Y_i - aX_i)^2$$

$$\frac{\partial Z}{\partial a} = 2$$

$$\frac{\partial Z}{\partial a} = 2 \sum (Y_i - a X_i) (-X_i) = 0$$

$$\sum Y_i X_i = a \sum X_i^2$$

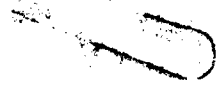
En las páginas siguientes, se tratarán conceptos referentes al análisis de correlación, conceptos que permitirán cuantificar el grado de asociación entre las variables que se estudian y la validez de las proyecciones a través de las ecuaciones de regresión.

PRELIMINAR
Instituto Latinoamericano de
Planificación Económica y Social
Santiago, diciembre de 1967

CURSO DE ESTADISTICA BASICA PARA PROGRAMACION*

Parte II

* Programa de Capacitación. Profesor, señor Arturo Núñez del Prado B.



CONTENIDO

A.	Objetivos del Análisis de Correlación	1
B.	Tipos de Correlación.	2
	1. Atendiendo al número de variables.	2
	2. Atendiendo a la forma de la función.	3
	3. Atendiendo a la relación de variables.	3
C.	El Coeficiente de Correlación	3
	1. Definición	3
D.	Limitaciones de la Correlación.	4
E.	Correlación Rectilínea.	5
	1. Representación de las magnitudes que determinan las varianzas.	5
	2. Método abreviado de cálculo.	9
	3. Otras fórmulas de cálculo.	11
	4. Correlación por rangos	14
E.	Correlación no Rectilínea	20
F.	Correlación Múltiple.	25
	1. Correlación en un Plano de Regresión	25
	2. Correlación en un hiperplano de regresión.	28
	3. Correlación múltiple logarítmica	30
G.	Etapas de la Construcción de un Modelo de Regresión	31
H.	Métodos de Estimación por medio del Coeficiente de Elasticidad.	35
	1. Presentación conceptual.	35
	2. Tipos de elasticidad	39
	3. Métodos de cálculo	39
	4. La ecuación de regresión y el coeficiente de elasticidad como instrumentos de proyección	42
	5. Relaciones y propiedades del coeficiente de elasticidad.	46
	6. Elasticidad de regresión múltiple.	48
	7. Limitaciones en la utilización de coeficientes de elasticidad.	51
I.	Modelos de Regresión.	53
	1. Introducción	53
	2. El modelo lineal de dos variables.	55
	3. Pruebas de hipótesis	70

60

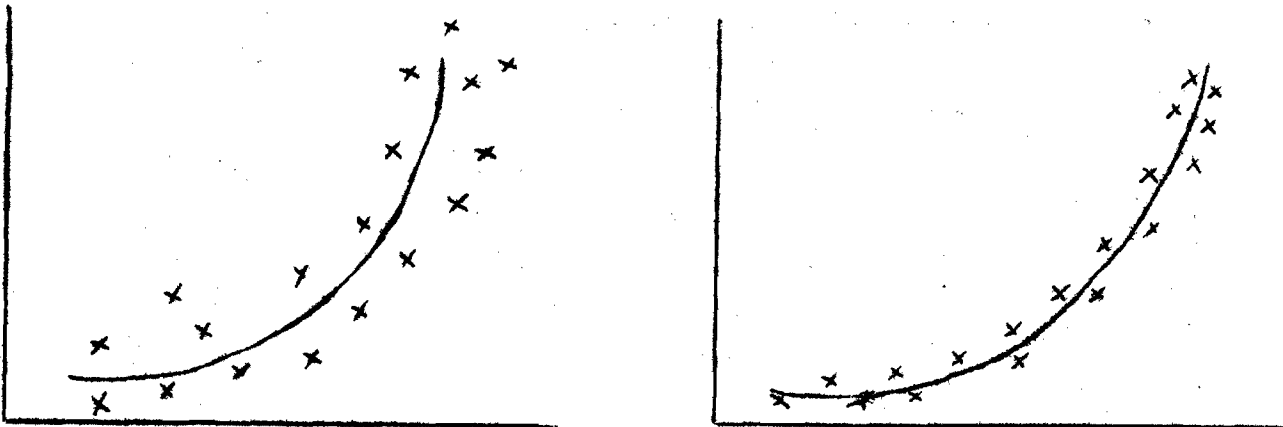
IV. CORRELACION

A. Objetivos del Análisis de Correlación

En el anterior capítulo se presentaron las técnicas del ajuste de funciones por el método de mínimos cuadrados. Una vez determinada la función, es necesario especificar si hay asociación entre las variables consideradas y en qué medida están asociadas. En el caso en que las variables estén fuertemente asociadas, la ecuación de regresión puede ser utilizada para explicar el comportamiento de la variable dependiente en términos de las variaciones que experimente la variable independiente. Por ejemplo, el incremento del volumen de venta de artefactos eléctricos puede ser explicado por aumentos en los niveles de ingreso, por variaciones en los precios, por modificaciones en los tipos de cambio, etc. Por otra parte, el instrumento de la regresión y correlación puede ser empleado en la estimación de valores de la variable dependiente (predictando), supuesto conocidas las variaciones de la variable independiente (predictor). En general, los planes de desarrollo especifican los niveles de ingreso por habitante que se pretende alcanzar en los próximos períodos; con tales datos y la ecuación de regresión del caso, pueden estimarse magnitudes de las variables que muestren alta asociación con el ingreso, como ser el consumo, la importación de alimentos, la reinversión de utilidades, etc. En todo caso, la validez de una proyección por regresión depende del grado de asociación entre las variables; si la asociación es alta, la estimación tiene base de fundamento, y si la asociación es débil, la proyección no tiene justificación.

Recuérdese que para determinar la ecuación de regresión es necesario contar con antecedentes sobre los valores que han tomado las variables. La representación gráfica de estos valores ayuda a especificar el tipo de función. En esta etapa ya puede adelantarse algo acerca del grado de asociación. Obsérvese los dos diagramas siguientes:

/Diagramas.



En el primer diagrama de dispersión los puntos están más alejados de la función que en el segundo. La proximidad de los puntos observados a la función es lo que determina el grado de asociación.

El objetivo básico del análisis de correlación es, pues, evidente: se trata de disponer de un indicador cuantitativo del grado de asociación que respalde la ecuación de regresión que se pretende utilizar. De hecho un conjunto de puntos que muestren la relación de un par de variables puede ser representada por cualquier función, pero una representación adecuada sólo se consigue cuando va garantizada por una asociación estrecha entre las variables.

B. Tipos de Correlación

En forma similar a la clasificación de los tipos de regresión que se presentó en el anterior capítulo, se puede distinguir los siguientes tipos de correlación:

1. Atendiendo al número de variables:

- a) Correlación simple.- Cuando se estudia el grado de asociación entre un par de variables: predictor y predictando.
- b) Correlación múltiple.- Cuando se estudia el grado de asociación que simultáneamente existe entre la variable dependiente y dos o más variables independientes.
- c) Correlación parcial.- En el caso de correlación múltiple, la cuantificación de la asociación neta entre dos variables, una vez que se elimina estadísticamente la influencia de otras variables independientes.

/2. Atendiendo

2. Atendiendo a la forma de la función. Según el tipo de ecuación de regresión se tiene correlación rectilínea, parabólica, potencial, exponencial, logarítmica, etc.
3. Atendiendo a la relación de variables.
 - a) Correlación directa o positiva; cuando ante aumentos en la variable independiente corresponden aumentos de la variable dependiente.
 - b) Correlación inversa o negativa; cuando ante aumentos en la variable independiente corresponden disminuciones de la variable dependiente.

C. El Coeficiente de Correlación

1. Definición. Un coeficiente de correlación indica el grado de asociación entre las variables. Se simbolizará por r y se definirá de la siguiente manera:

$$r = \left(\frac{S_{Yc}^2}{S_Y^2} \right)^{1/2} = \frac{S_{Yc}}{S_Y}$$

Donde S_{Yc}^2 representa a la varianza explicada, es decir, a aquella parte de la varianza total explicada por la ecuación de regresión y S_Y^2 representa a la varianza total tal cual se definió en la primera parte del curso, es decir:

$$S_{Yc}^2 = \frac{\sum (Yc - \bar{Y})^2}{n} \quad (Yc: \text{valor calculado})$$

$$S_Y^2 = \frac{\sum (Yi - \bar{Y})^2}{n} \quad (Yi: \text{valor observado})$$

Como puede observarse ambas varianzas expresan un promedio de cuadrados de desviaciones respecto de la media aritmética y su cómputo no difiere del que se realiza para una varianza cualquiera. Lo que ocurre es que la variabilidad total se descompone en dos fuentes: la varianza explicada y la varianza no explicada que se define:

$$S_{Ys}^2 = \frac{\sum (Y_i - Y_c)^2}{n}$$

Lógicamente la suma de la varianza explicada y la varianza no explicada reproducen la varianza total, como se demuestra más adelante. La raíz de la varianza no explicada, por el hecho de ser un indicador del grado de dispersión de los puntos observados respecto de los puntos calculados por la ecuación de regresión, recibe el nombre de error de proyección y se utiliza para fijar intervalos de confianza.

Observando la fórmula del coeficiente de correlación, éste puede interpretarse como la proporción que representa la desviación típica explicada dentro de la desviación típica total.

D. Limitaciones de la Correlación

La rapidez y sencillez con que se presentó el tema puede hacer que la correlación se interprete sin salvedades y con ilimitados alcances. Es conveniente por ello plantear los siguientes puntos.

1. Un alto coeficiente de correlación no necesariamente determina causalidad entre las variables. Dos variables pueden aparecer correlacionadas por casualidad y no porque exista una relación de dependencia entre ellas.
2. En cuanto a las variables, es necesario que aparezcan depuradas de las influencias de otras variables. Dos series nominales pueden mostrar estrecha asociación porque hay una tercera variable: alzas de precios, que exageran el grado de asociación. Por ello es conveniente trabajar con series reales, per capita, de manera de hacer más significativa la correlación.
3. Dos series pueden también arrojar coeficientes de correlación cercanos a uno, porque el tamaño de muestra es insuficiente. En un caso extremo en que sólo se tomen dos puntos, el coeficiente de correlación rectilíneo mostrará en general un valor igual a la unidad, pero esto no garantiza la adecuada significación. La calificación del grado de asociación no puede dejar de considerar el número de puntos utilizados en el estudio.

/4. Desde

4. Desde el punto de vista del tipo de función, sobre todo cuando se tiene por objetivo la predicción de una variable, es conveniente trabajar con funciones sencillas capaces de representar la tendencia de la nube de puntos. Si se tiene una función complicada con muchos parámetros y muchas variables independientes, posiblemente se obtenga un alto coeficiente de correlación, porque la función, dada su complejidad, pasará muy cerca de los puntos observados. Sin embargo, la correlación pierde validez como garantía de una adecuada predicción; la estimación de los valores de las variables independientes, es decir, la fijación de las variables exógenas se hace más difícil cuando éstas son numerosas.
5. No debe olvidarse que la predicción por regresión y correlación es válida en tanto sigan en vigencia los supuestos y circunstancias implícitos en los datos y antecedentes disponibles. Predecir por regresión la producción agrícola de los próximos períodos, por ejemplo, haciendo caso omiso de una eventual reforma agraria, probablemente conducirá a estimaciones alejadas de la realidad. Es importante, al realizar estimaciones, dejar en claro los supuestos básicos y admitir que cualquier desviación de estos supuestos exige una revisión del modelo de proyección o del modelo de análisis según sea el caso.
6. Por último, vale la pena destacar que los modelos de regresión y correlación significan una permanente revisión de supuestos y acumulación de nuevos antecedentes que permitan ajustar el modelo a las nuevas circunstancias.

E. Correlación Rectilínea

Es conveniente presentar en detalle los conceptos expuestos en forma general aplicados al caso específico de la correlación rectilínea.

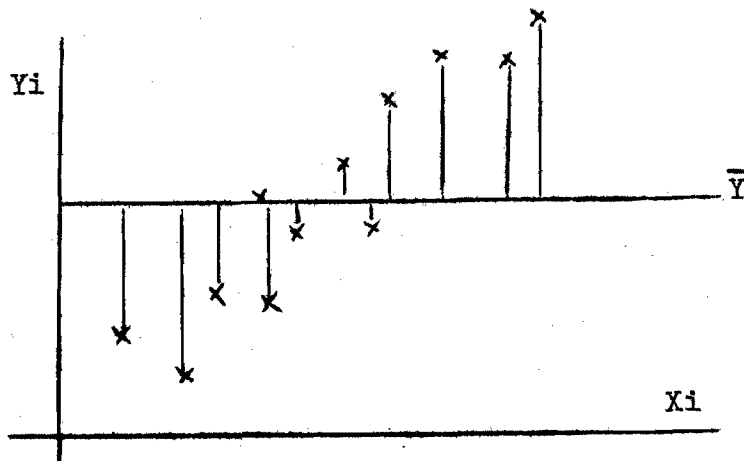
1. Representación de las magnitudes que determinan las varianzas.

Se dijo que la varianza denominada total correspondía exactamente al concepto en la primera parte del curso; en efecto:

$$s_Y^2 = \frac{\sum (y_i - \bar{Y})^2}{n}$$

/En el

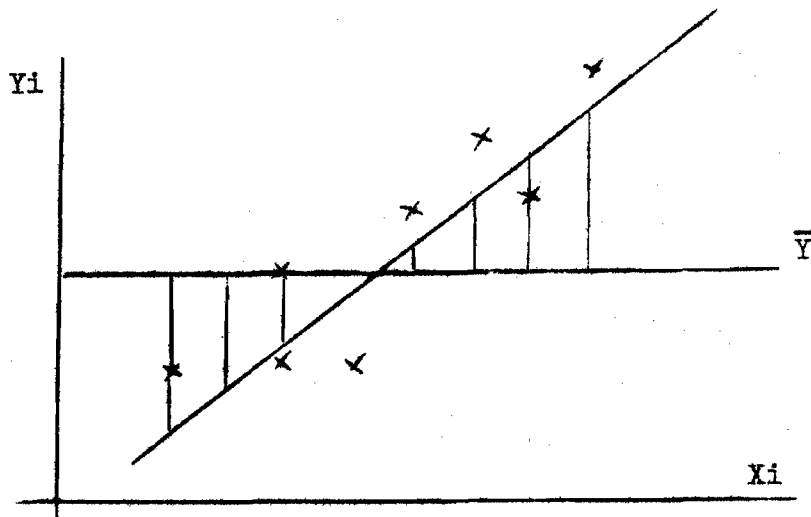
En el siguiente diagrama puede observarse las desviaciones que toma en cuenta este estadígrafo:



La varianza explicada está dada por las desviaciones de los valores calculados respecto de la media aritmética.

$$S_{Y_c}^2 = \frac{\sum (Y_c - \bar{Y})^2}{n}$$

Gráficamente:

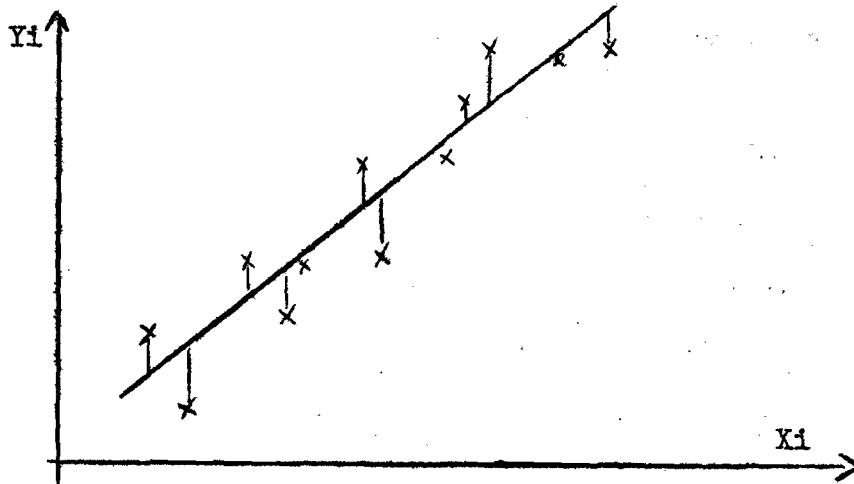


La varianza no explicada está dada por las desviaciones de los valores observados respecto de los valores calculados.

/Fórmula

$$S_{Ys}^2 = \frac{\sum (Y_i - Y_c)^2}{n}$$

Gráficamente:



Evidentemente:

varianza

$$S_Y^2 = S_{Yc}^2 + S_{Ys}^2$$

ya que:

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

$$\sum Y_i^2 - 2\bar{Y} \sum Y_i + n\bar{Y}^2 = \sum Y_c^2 - 2\bar{Y} \sum Y_c + n\bar{Y}^2 + \sum Y_i^2 - 2\sum Y_i Y_c + \sum Y_c^2$$

Por otra parte $\sum Y_i = \sum Y_c$ ya que:

$$\sum Y_i = a \sum X_i + nb$$

1ª ecuación normal.

$$Y_c = aX_i + b$$

Ecuación de regresión rectilínea.

Aplicando el operador sumatoria:

$$\sum Y_c = a \sum X_i + nb$$

$$\text{Luego } \sum Y_c = \sum Y_i$$

/La relación

La relación original puede simplificarse en consecuencia:

$$2 \sum Y_i Y_c - 2 \bar{Y} \sum Y_i = 2 \sum Y_c^2 - 2 \bar{Y} \sum Y_c$$

pero,

$$\sum Y_i = \sum Y_c$$

y queda:

$$\sum Y_i Y_c = \sum Y_c^2$$

Pero:

$$Y_c = a X_i + b$$

$$Y_c^2 = a^2 X_i^2 + 2ab X_i + b^2$$

$$Y_c^2 = a^2 X_i^2 + ab X_i + ab X_i + b^2$$

$$Y_c^2 = a(a X_i^2 + b X_i) + b(a X_i + b)$$

$$\sum Y_c^2 = a(a \sum X_i^2 + b \sum X_i) + b(a \sum X_i + nb)$$

Las expresiones entre paréntesis son las ecuaciones normales de una recta, luego

$$\sum Y_c^2 = a \sum X_i Y_i + b \sum Y_i$$

Por otra parte:

$$\sum Y_i Y_c = \sum Y_i (a X_i + b)$$

ya que:

$$Y_c = a X_i + b$$

$$\sum Y_i Y_c = a \sum X_i Y_i + b \sum Y_i$$

Luego:

$$\sum Y_i^2 = \sum Y_i Y_c$$

y por consiguiente:

$$S_Y^2 = S_{Y_c}^2 + S_{Y_b}^2$$

/De esto

De esto se deduce que el valor numérico del coeficiente de correlación, o de su cuadrado que se denomina coeficiente de determinación, fluctúa entre 0 y 1.

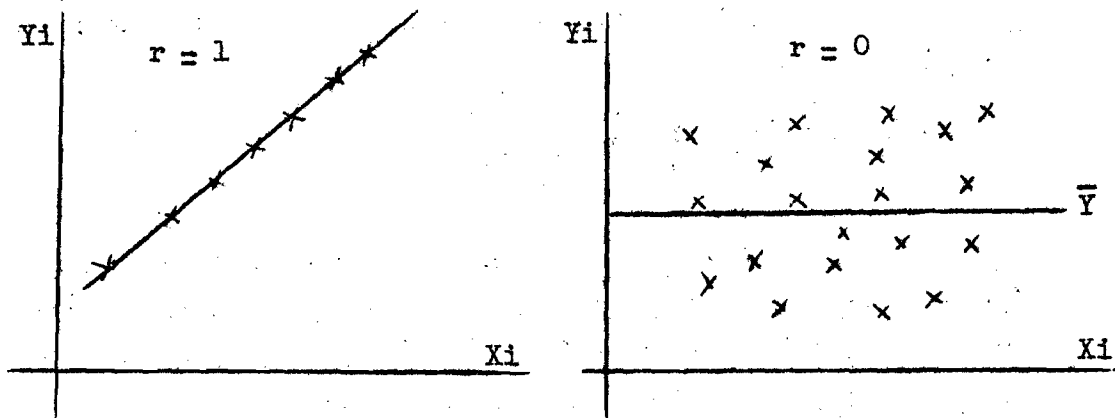
$$0 \leq r^2 \leq 1$$

$$0 \leq r \leq 1$$

En correlación rectilínea se asigna el signo positivo de la raíz cuando se trata de correlación directa y el signo negativo si la correlación es inversa. En este caso, entonces, los límites serán:

$$-1 \leq r \leq 1$$

El coeficiente de correlación tomará el valor uno cuando todos los puntos observados estén situados sobre la ecuación de regresión, y tomará el valor cero cuando la ecuación de regresión coincida con una paralela al eje de las abscisas a la altura de la media aritmética.



2. Método abreviado de cálculo

El cálculo del coeficiente de correlación basado en las varianzas, es decir, en la definición, implica cuantificar los valores calculados por la ecuación de regresión, lo que en sí representa un trabajo bastante laborioso. El método que a continuación se presenta, aprovecha los cálculos que se han debido realizar

/para la

para la determinación de los parámetros de la ecuación de regresión. Recordando la definición:

$$r^2 = \frac{s_{Yc}^2}{s_Y^2} = \frac{\sum (Yc - \bar{Y})^2}{\sum (Yi - \bar{Y})^2}$$
$$= \frac{\sum Yc^2 - 2 \bar{Y} \sum Yc + n \bar{Y}^2}{\sum Yi^2 - 2 \bar{Y} \sum Yi + n \bar{Y}^2}$$

Obsérvese que $\sum Yc = \sum Yi = n \bar{Y}$

$$= \frac{\sum Yc^2 - 2 n \bar{Y}^2 + n \bar{Y}^2}{\sum Yi^2 - 2 n \bar{Y}^2 + n \bar{Y}^2} = \frac{\sum Yc^2 - n \bar{Y}^2}{\sum Yi^2 - n \bar{Y}^2}$$

Será necesario encontrar una expresión para $\sum Yc^2$. En el anterior punto se demostró que:

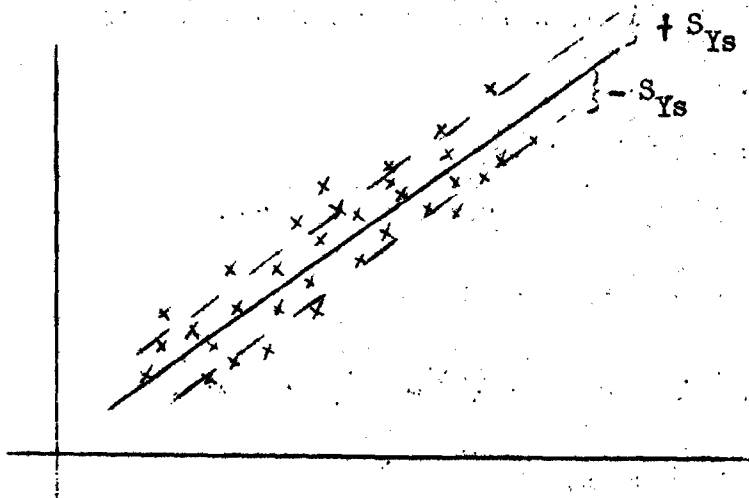
$$\sum Yc^2 = a \sum Xi Yi + b \sum Yi$$

Reemplazando esta nueva expresión en la última fórmula de r^2 , se tiene:

$$r^2 = \frac{a \sum Xi Yi + b \sum Yi - n \bar{Y}^2}{\sum Yi^2 - n \bar{Y}^2}$$

La fórmula anterior, como se dijo, tiene la ventaja de utilizar cálculos que debieron hacerse para el ajuste por el método de los mínimos cuadrados, a excepción de $\sum Yi^2$. El numerador de esta fórmula equivale a n veces la varianza explicada, y el denominador equivale a n veces la varianza total. Por consiguiente, para obtener el cuadrado del error de proyección (varianza no explicada), será necesario restar el numerador del denominador y dividir la diferencia por n . La utilización del error de proyección en la predicción o estimación de intervalos tiene la siguiente interpretación gráfica:

/Gráfico



Detrás de esta interpretación está el supuesto de que las diferencias de valores observados a valores calculados tienen una distribución de probabilidad normal. Por ese hecho pueden establecerse niveles de confianza a probabilidades de acierto en las estimaciones. Si se suma y resta una vez el error de proyección, el intervalo resultante implica un nivel de confianza de 68 por ciento; si se suma y resta dos veces el error de proyección, el nivel de confianza será de 95 por ciento; si se suma y resta tres veces el error de proyección, el nivel de confianza será de 99 por ciento, etc.

3. Otras fórmulas de cálculo

Existen otras fórmulas para cuantificar el grado de asociación, entre ellas la fórmula llamada momento-producto que conduce a su vez a expresar el coeficiente de correlación como la media geométrica de los coeficientes de regresión angulares.

Dada la ecuación de regresión:

$$Y_c = a X_i + b$$

aplicando el operador media aritmética, se tiene:

$$\bar{Y} = a \bar{X} + b$$

de donde: $b = \bar{Y} - a \bar{X}$

/Por otra

Por otra parte, las ecuaciones normales para la recta son:

$$1) \sum Y_i = a \sum X_i + nb$$

$$ii) \sum X_i Y_i = a \sum X_i^2 + b \sum X_i$$

Dividiendo la segunda ecuación por n, queda:

$$\frac{\sum X_i Y_i}{n} = a \frac{\sum X_i^2}{n} + b \bar{X}$$

Reemplazando el valor de $b = \bar{Y} - a \bar{X}$, se tiene:

$$\begin{aligned} \frac{\sum X_i Y_i}{n} &= a \frac{\sum X_i^2}{n} + (\bar{Y} - a \bar{X}) \bar{X} \\ &= a \frac{\sum X_i^2}{n} + \bar{X} \bar{Y} - a \bar{X}^2 \end{aligned}$$

$$\frac{\sum X_i Y_i}{n} - \bar{X} \bar{Y} = a \left(\frac{\sum X_i^2}{n} - \bar{X}^2 \right)$$

Observando estas expresiones, se concluye que el primer miembro no es otra cosa que la covarianza de las variables Y_i y X_i , y la expresión dentro del paréntesis es la varianza de la variable independiente, es decir:

$$C[X_i Y_i] = a V[X_i]$$

Nótese que esta expresión corresponde a una ecuación de regresión de Y en X, es decir, donde X_i es la variable predictor y Y_i es la variable predictando. Para especificar la fórmula en este sentido, el coeficiente angular "a" tendrá la siguiente expresión:

$$a_{YX} = \frac{C[X_i Y_i]}{V[X_i]}$$

Por analogía, si la ecuación de regresión fuera de X en Y se tendría:

$$X_c = a Y_i + b$$

/Dado que:

Dado que:

$$C(X_i Y_i) = C(Y_i X_i) = \frac{\sum X_i Y_i}{n} - \bar{Y} \bar{X}$$

donde el orden de los factores no altera el producto numérico, se tiene:

$$a_{XY} = \frac{C[\sum X_i Y_i]}{V[\sum Y_i]}$$

En resumen, hasta ahora se dispone de fórmulas para los coeficientes de regresión en términos de varianzas, covarianzas y medias aritméticas, que son útiles para la obtención de valores numéricos y para las demostraciones que a continuación se presentan,

La fórmula abreviada del coeficiente de correlación es:

$$r^2 = \frac{a \sum X_i Y_i + b \sum Y_i - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2}$$

dado que

$$b_{YX} = \bar{Y} - a_{YX} \bar{X}$$

reemplazando se tiene:

$$\begin{aligned} r^2 &= \frac{a_{YX} \sum X_i Y_i + (\bar{Y} - a_{YX} \bar{X}) \sum Y_i - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2} \\ &= \frac{a_{YX} \sum X_i Y_i + n \bar{Y}^2 - n a_{YX} \bar{X} \bar{Y} - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2} \\ &= \frac{a_{YX} \{ \sum X_i Y_i - n \bar{X} \bar{Y} \}}{\sum Y_i^2 - n \bar{Y}^2} \end{aligned}$$

dividiendo numerador y denominador por n se tiene:

$$r^2 = \frac{a_{YX} \left\{ \frac{\sum X_i Y_i}{n} - \bar{X} \bar{Y} \right\}}{\frac{\sum Y_i^2}{n} - \bar{Y}^2} = a_{YX} \frac{C[\sum X_i Y_i]}{V[\sum Y_i]}$$

/Luego

Luego

$$r^2 = a_{YX} a_{XY}$$

$$r = \pm \sqrt{a_{YX} a_{XY}}$$

De otra manera

$$r^2 = \frac{(C[\bar{X}_i \bar{Y}_i])^2}{V[\bar{X}_i] V[\bar{Y}_i]} \therefore r = \frac{C[\bar{X}_i \bar{Y}_i]}{S_{X_i} S_{Y_i}}$$

Han sido deducidas dos fórmulas adicionales para el coeficiente de correlación. La primera está dada por la media geométrica de los coeficientes angulares de regresión, y la segunda tiene como numerador a la covarianza, que es un momento de orden uno-uno respecto del origen, y como denominador al producto de las desviaciones típicas de las variables; de ahí el nombre de momento-producto que se le da a esta fórmula.

Se ha puesto énfasis en la necesidad de distinguir el sentido de la regresión y correlación, es decir, si se trata de "Y sobre X" o de "X sobre Y" por el hecho de que hay análisis donde puede presentarse cierta reversibilidad en la causalidad.

4. Correlación por rangos

Un caso particular de la correlación rectilínea es la llamada correlación por rangos u ordenamientos. Resulta que existe una cantidad de variables que no son susceptibles de medición exacta y, sin embargo, pueden ser susceptibles de ordenarse cualitativamente. Por ejemplo, una selección de candidatos a un cargo basada en entrevistas personales, puede conducir a ordenamientos de los

/candidatos, por

candidatos (por ejemplo, de mejor a peor) por cada uno de los entrevistadores. El análisis de correlación por rangos determinará si estos ordenamientos son coincidentes o dispares, y cuál es la magnitud de la coincidencia o disparidad. El asunto es asignar a cada candidato un número de orden y determinar el grado de asociación entre dos ordenamientos. Este asunto puede ser enfrentado con las fórmulas generales vistas para la correlación rectilínea. Sin embargo, en este caso se consigue alguna ventaja de cálculo por el hecho de que las variables tomarán valores enteros equidistanciados. Siguiendo el ejemplo, si dos supervisores hubieran ordenado a 8 postulantes en la siguiente forma:

Postulantes	Ordenamiento del Entrevistador 1 u_{1i}	Ordenamiento del Entrevistador 2 u_{2i}
A	4°	3°
B	2°	4°
C	7°	8°
D	6°	5°
E	3°	2°
F	8°	7°
G	5°	6°
H	1°	1°

El análisis de la correlación por rangos proporcionará un indicador cuantitativo acerca de la disparidad o coincidencia de los ordenamientos. Obsérvese que las variables son los números naturales en ambos tipos de ordenamientos. Este hecho permite concluir que:

- i) las medias aritméticas de los dos ordenamientos serán iguales, es decir: $M[u_{1i}] = M[u_{2i}]$.
- ii) las varianzas de ambos ordenamientos serán iguales, es decir: $V[u_{1i}] = V[u_{2i}]$.

Una de las fórmulas generales para el coeficiente de correlación rectilínea era la siguiente:

$$r = \frac{C[X_i Y_i]}{S_{X_i} S_{Y_i}} \therefore r^2 = \frac{(C[X_i Y_i])^2}{V[X_i] V[Y_i]}$$

Para las deducciones posteriores es importante establecer previamente cuál es la varianza de una suma de variables:

$$V[X_i + Y_i] = \frac{\sum (X_i + Y_i - \bar{X} - \bar{Y})^2}{n} = \frac{\sum [(X_i - \bar{X}) + (Y_i - \bar{Y})]^2}{n}$$

$$V[X_i + Y_i] = \frac{\sum (X_i - \bar{X})^2}{n} + \frac{\sum (Y_i - \bar{Y})^2}{n} + \frac{2\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$V[X_i + Y_i] = V[X_i] + V[Y_i] + 2 C[X_i Y_i]$$

ya que:

$$\begin{aligned} 2 \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n} &= \frac{2 \sum (X_i Y_i - \bar{X} Y_i - \bar{Y} X_i + \bar{X} \bar{Y})}{n} \\ &= 2 \left[\frac{\sum X_i Y_i}{n} - \bar{X} \frac{\sum Y_i}{n} - \bar{Y} \frac{\sum X_i}{n} + \bar{X} \bar{Y} \right] \\ &= 2 \left[\frac{\sum X_i Y_i}{n} - \bar{X} \bar{Y} - \bar{Y} \bar{X} + \bar{X} \bar{Y} \right] \\ &= 2 \left[\frac{\sum X_i Y_i}{n} - \bar{X} \bar{Y} \right] = 2 C[X_i Y_i] \end{aligned}$$

Por analogía la varianza de la diferencia de dos variables será:

$$V[X_i - Y_i] = V[X_i] + V[Y_i] - 2 C[X_i Y_i]$$

Dadas las variables ordenamiento u_{1i} , u_{2i} , se puede establecer una relación de diferencia entre ellas:

$$/d_1 =$$

$$d_i = u_{1i} - u_{2i}$$

$$V[d_i] = V[u_{1i}] + V[u_{2i}] - 2C[u_{1i} u_{2i}]$$

El coeficiente de correlación en términos de estas variables será:

$$r = \frac{C[u_{1i} u_{2i}]}{\sqrt{V[u_{1i}] V[u_{2i}]}}$$

$$r = \frac{C[u_{1i} u_{2i}]}{V[u_{1i}]} \quad \text{ya que } V[u_{1i}] = V[u_{2i}]$$

Despejando de la relación $V[d_i]$, se tiene que:

$$\begin{aligned} C[u_{1i} u_{2i}] &= (V[u_{1i}] + V[u_{2i}] - V[d_i]) \frac{1}{2} \\ &= (2V[u_{1i}] - V[d_i]) \frac{1}{2} \end{aligned}$$

Por otra parte:

$$d_i = u_{1i} - u_{2i}$$

$$M[d_i] = M[u_{1i}] - M[u_{2i}] = 0$$

Luego,

$$\begin{aligned} V[d_i] &= M[d_i^2] - (M[d_i])^2 \\ &= \frac{\sum d_i^2}{n} \end{aligned}$$

Entonces:

$$r = \frac{C[u_{1i} u_{2i}]}{V[u_{1i}]} = \frac{V[u_{1i}] - V[d_i]}{V[u_{1i}]} \frac{1}{2}$$

$$= 1 - \frac{V\sqrt{d_i}}{2V\sqrt{u_{1i}}}$$

Pero la varianza de $u_{1i} \sqrt{V(u_{1i})}$, es la varianza de los n primeros números naturales.

$$\begin{aligned} V\sqrt{u_{1i}} &= \frac{\sum u_{1i}^2}{n} - \bar{u}_1^2 \\ &= \frac{n(n+1)(2n+1)}{6n} - \left(\frac{n(n+1)}{2n}\right)^2 \\ &= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)(n+1)}{4} \\ &= (n+1) \left[\frac{2n+1}{6} - \frac{n+1}{4} \right] \\ &= (n+1) \left[\frac{4n+2-3n-3}{12} \right] \\ &= (n+1) \left(\frac{n-1}{12} \right) = \frac{n^2-1}{12} \end{aligned}$$

Luego:

$$r = 1 - \frac{V\sqrt{d_i}}{2\left(\frac{n^2-1}{12}\right)}$$

$$r = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

En el ejemplo propuesto en páginas anteriores el cálculo se haría de la siguiente manera:

/Postulantes

Postulantes	u_{1i}	u_{2i}	d_i	d_i^2
A	4	3	1	1
B	2	4	-2	4
C	7	8	-1	1
D	6	5	1	1
E	3	2	1	1
F	8	7	1	1
G	5	6	-1	1
H	<u>1</u>	<u>1</u>	<u>0</u>	<u>0</u>
	36	36	0	10

Aplicando la fórmula anterior resulta

$$r = 1 - \frac{6(10)}{8(63)} = 0,88$$

El resultado del indicador muestra que los dos ordenamientos están bastante asociados, no existiendo en general discrepancias significativas.

A veces puede utilizarse con ventaja este tipo de indicador, aún en casos de variable cuantificable. Puede ocurrir que si el número de datos es muy grande y las variables toman valores que dificultan el cálculo numérico, sea conveniente ordenar las observaciones de acuerdo a sus valores numéricos y establecer la correlación entre los ordenamientos. Evidentemente esta simplificación implica rigidez por la introducción de supuestos adicionales: correlación rectilínea y que los ordenamientos reflejan adecuadamente la distribución de las variables originales. Sin embargo, muchas veces basta con saber si existe o no asociación sin interesar refinar mucho el análisis; para este tipo de estudios puede prestarse esta conversión arbitraria de variables.

La facilidad de cálculo del coeficiente de correlación por rangos, tiene como contrapartida una seria limitación. Se supone que las distancias o diferencia de atributos es constante entre los casos /considerados. En

considerados. En el ejemplo visto, esto quiere decir que la diferencia entre el postulante H y el postulante B es la misma que la que existe entre el postulante B y el postulante E, etc., para cada uno de los ordenamientos. En la práctica difícilmente se cumplirá este supuesto, pero como se ha dicho, este tratamiento es adecuado para variable de atributos donde la jerarquización u ordenamiento es la única forma de discriminación y donde se observa con menos rigurosidad las limitaciones aludidas.

E. Correlación no Rectilínea

Si bien el caso particular de la correlación rectilínea es útil en la presentación de los conceptos del análisis de regresión y correlación, su aplicación práctica es algo restringida por el hecho de que en los estudios socioeconómicos las relaciones entre las variables toman formas que en general difícilmente pueden ser representadas en forma adecuada por una línea recta.

De la misma manera que distinguíamos diversas relaciones no rectilíneas en el capítulo de regresión, aquí desde ese mismo punto de vista se expondrán los coeficientes de correlación respectivos. En correlación no rectilínea no tiene utilidad distinguir entre directa e inversa, por el hecho de que pueden haber tramos donde la relación sea directa y otros donde sea inversa.

1. Si la función es del tipo:

$$Y_c = a \log X + b$$

es decir, si cambios relativos de X determinan cambios absolutos de Y, el coeficiente de determinación tendrá la expresión general:

$$r^2 = \frac{S_{Y_c}^2}{S_Y^2} = \frac{\sum (Y_c - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum Y_c^2 - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2}$$

pero,

$$Y_c^2 = a^2 (\log X_i)^2 + 2ab \log X_i + b^2$$

$$/Y_c^2 =$$

$$Y_c^2 = a \left\{ a(\log X_i)^2 + b \log X_i \right\} + b \left\{ a \log X_i + b \right\}$$
$$\sum Y_c^2 = a \left\{ \underbrace{a \sum (\log X_i)^2 + b \sum \log X_i}_{\text{Ecuación normal}} \right\} + b \left\{ \underbrace{a \sum \log X_i + nb}_{\text{Ecuación normal}} \right\}$$

$$\sum Y_c^2 = a \sum Y_i \log X_i + b \sum Y_i$$

$$r^2 = \frac{a \sum Y_i \log X_i + b \sum Y_i - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2}$$

Nótese que este coeficiente de determinación resulta de la relación entre Y_i y $\log X_i$ que será distinto al que resulta de la relación entre Y_i y X_i (siendo X_i el antilogaritmo de $\log X_i$).

Para el cálculo del error de proyección se procede de manera similar al caso de correlación rectilínea, es decir, basta dividir por n la diferencia entre el numerador de la fórmula del coeficiente de determinación (n veces la varianza explicada) y el denominador (n veces la varianza total). La raíz cuadrada de esta diferencia dividida por n será el error de proyección. Nuevamente, el error de proyección está dado teniendo en cuenta la proyección de Y_i , en términos de la variable independiente $\log X_i$, que será distinto al error que se dé al proyectar Y_i en términos de X_i .

2. Si la función es del tipo:

$$Y = fd^X \quad \text{si } \log f = b; \log d = a$$
$$\log Y = aX + b$$

es decir, una función de las llamadas exponenciales, el procedimiento para encontrar las fórmulas del coeficiente de correlación y del error de proyección es similar al caso anterior. Obsérvese que en esta función, variaciones absolutas de la variable X determinan variaciones relativas de Y . La fórmula general del coeficiente de determinación es la siguiente:

$$r^2 = \frac{\sum (Y_c - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum Y_c^2 - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2}$$

/Obsérvese que

Obsérvese que en la función aparece el logaritmo de Y_i . Por este hecho la fórmula particular será:

$$r^2 = \frac{\sum (\log Y_c)^2 - n \overline{\log Y}^2}{\sum (\log Y_i)^2 - n \overline{\log Y}^2}$$

donde,

$$\overline{\log Y} = M[\log Y_i] = \frac{\sum \log Y_i}{n}$$

Dado que: $\log Y_c = aX_i + b$

$$\begin{aligned} \sum (\log Y_c)^2 &= a^2 \sum X_i^2 + ab \sum X_i + ab \sum X_i + nb^2 \\ &= a \left\{ a \sum X_i^2 + b \sum X_i \right\} + b \left\{ a \sum X_i + nb \right\} \\ &= a \sum (\log Y_i) X_i + b \sum \log Y_i \end{aligned}$$

$$r^2 = \frac{a \sum X_i \log Y_i + b \sum \log Y_i - n \overline{\log Y}^2}{\sum (\log Y_i)^2 - n \overline{\log Y}^2}$$

Cabe destacar nuevamente que el coeficiente de correlación calculado según la última expresión diferirá del calculado según la expresión general. En la fórmula general se establece la asociación entre Y_i y X_i , en cambio, en el caso particular se establece la correlación entre el logaritmo de Y_i y la variable X_i . El error de proyección en uno y otro caso se obtiene calculando la raíz de la varianza no explicada que es la ene-ava parte de la diferencia entre el numerador y el denominador de la fórmula del coeficiente de determinación.

†3. En una función potencial del tipo

$$Y_c = b X^a$$

cuya expresión logarítmica es:

$$\log Y_c = \log b + a \log X_i$$

donde variaciones relativas de la variable independiente determinan

/variaciones también

variaciones también relativas de la variable dependiente, pueden establecerse dos fórmulas para el coeficiente de correlación que conducirán a resultados distintos:

$$r^2 = \frac{\sum (\log Y_c - \overline{\log Y})^2}{\sum (\log Y_i - \overline{\log Y})^2}$$

y la otra con los antilogaritmos respectivos.

$$r^2 = \frac{\sum (Y_c - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

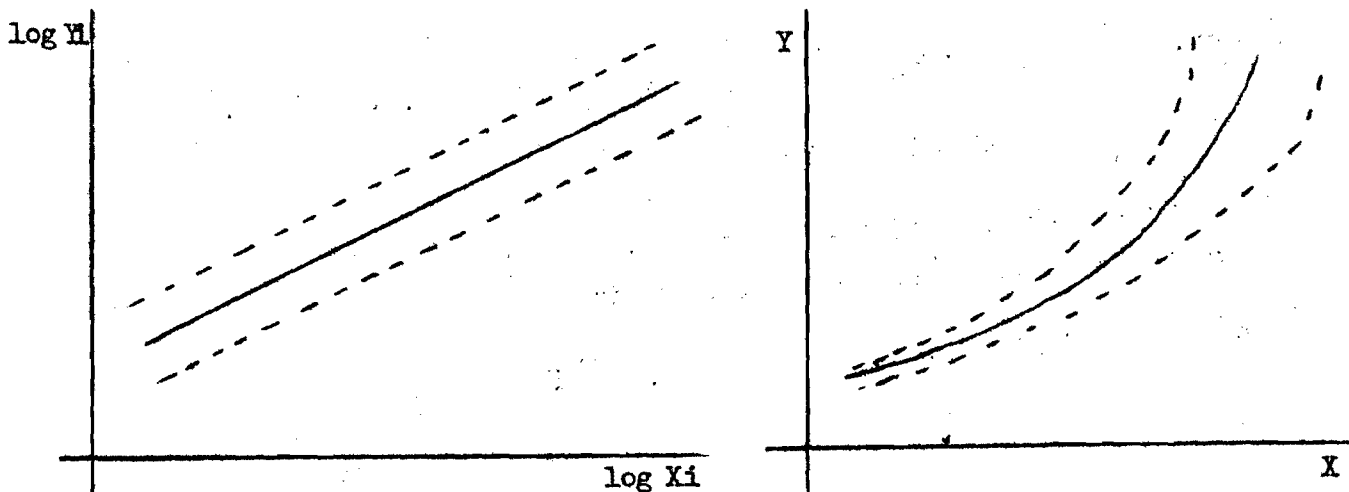
Para el caso de la correlación logarítmica la fórmula abreviada de cálculo se obtiene de la misma forma que las anteriores.

$$r^2 = \frac{a \sum \log X_i \log Y_i + \log b \sum \log Y_i - n \overline{\log Y}^2}{\sum (\log Y_i)^2 - n \overline{\log Y}^2}$$

De esta fórmula también puede obtenerse el error de proyección logarítmico con el procedimiento ya conocido.

En el caso de correlación logarítmica la estimación de intervalos se hace en la misma forma que en el caso rectilíneo, pero teniendo en cuenta que para los valores reales el desvío contemplado representa una proporción constante de los valores dados por la ecuación de regresión.

La interpretación gráfica es la siguiente:



/En escalas

En escalas logarítmicas aparece una diferencia constante. Los antilogaritmos de estos valores conducen a la representación en escala natural donde puede observarse que el intervalo es cada vez más grande, lo que corresponde a una proporción constante del semiancho del intervalo respecto de los valores dados por la ecuación de regresión.

4. Correlación parabólica

Dada la función:

$$Y_c = a X^2 + b X + c$$

El coeficiente de determinación está dado por:

$$r^2 = \frac{\sum (Y_c - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{\sum Y_c^2 - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2}$$

$$Y_c^2 = a^2 X^4 + b^2 X^2 + c^2 + 2 a b X^3 + 2 a c X^2 + 2 b c X$$

$$Y_c^2 = \underline{a^2 X^4} + \underline{a b X^3} + \underline{a b X^3} + \underline{b^2 X^2} + a c X^2 + \underline{a c X^2} + c^2 + \underline{b c X} + b c X$$

$$Y_c^2 = a \{ a X_i^4 + b X_i^3 + c X_i^2 \} + b \{ a X^3 + b X_i^2 + c X_i \} + c \{ a X^2 + b X + c \}$$

Aplicando el operador sumatoria:

$$\sum Y_c^2 = a \{ a \sum X_i^4 + b \sum X_i^3 + c \sum X_i \} + b \{ a \sum X_i^3 + b \sum X_i^2 + c \sum X_i \} + c \{ a \sum X^2 + b \sum X + n c \}$$

Las expresiones dentro de los paréntesis corresponden con las ecuaciones normales de la parábola, es decir:

$$\sum Y_c^2 = a \{ \sum Y_i X_i^2 \} + b \{ \sum Y_i X_i \} + c \{ \sum Y_i \}$$

/La fórmula

La fórmula abreviada de cálculo para el coeficiente de correlación será:

$$r^2 = \frac{a \sum Y_i X_i^2 + b \sum Y_i X_i + c \sum Y_i - n \bar{Y}^2}{\sum Y_i^2 - n \bar{Y}^2}$$

El error de proyección se calcula por el método expuesto para los casos anteriores.

F. Correlación Múltiple

Cuando se tiene dos o más variables independientes, la necesidad de disponer de indicaciones acerca de la asociación que simultáneamente tiene la variable dependiente con las variables independientes, conduce a la obtención de coeficientes de correlación múltiples. Si bien son necesarios para el análisis los coeficientes de correlación simples, es preciso complementar este conjunto de indicadores con un estadígrafo que resuma simultáneamente los grados de asociación simples.

1. Correlación en un plano de regresión

Si se tienen dos variables independientes, la ecuación de regresión es de la forma:

$$X_{1c} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

El coeficiente de correlación se obtiene siempre a partir de la fórmula general que ahora tendrá la siguiente simbología:

$$R_{1.23}^2 = \frac{S_{1c \ 1.23}^2}{S_{X1}^2} = \frac{\sum (X_{1c} - \bar{X})^2}{\sum (X_{1i} - \bar{X})^2} = \frac{\sum X_{1c}^2 - n \bar{X}_1^2}{\sum X_{1i}^2 - n \bar{X}_1^2}$$

Donde $S_{1c \ 1.23}^2$ es la varianza explicada por las variables X_2 y X_3 y S_{X1}^2 es la varianza total de la variable dependiente.

$$X_{1c}^2 = \frac{a_{1.23}^2}{1.23} + \frac{b_{12.3}^2}{12.3} X_2^2 + \frac{b_{13.2}^2}{13.2} X_3^2 + \frac{a_{1.23} b_{12.3}}{1.23} X_2 + \frac{a_{1.23} b_{13.2}}{1.23} X_3 + \frac{a_{1.23} b_{13.2}}{1.23} X_3 + \frac{a_{1.23} b_{12.3}}{1.23} X_2 X_3 + \frac{b_{12.3} b_{13.2}}{12.3} X_2 X_3$$

$$+ b_{12.3} b_{13.2} X_2 X_3$$

$$\begin{aligned} X_{1c}^2 = & a_{1.23} \left\{ a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3 \right\} + \\ & + b_{12.3} \left\{ b_{12.3} X_2^2 + a_{1.23} X_2 + b_{13.2} X_2 X_3 \right\} \\ & + b_{13.2} \left\{ b_{13.2} X_3^2 + a_{1.23} X_3 + b_{12.3} X_2 X_3 \right\} \end{aligned}$$

Aplicando sumatoria se tiene:

$$\begin{aligned} \sum X_{1c}^2 = & a_{1.23} \left\{ n a_{1.23} + b_{12.3} \sum X_2 + b_{13.2} \sum X_3 \right\} \\ & + b_{12.3} \left\{ b_{12.3} \sum X_2^2 + a_{1.23} \sum X_2 + b_{13.2} \sum X_2 X_3 \right\} \\ & + b_{13.2} \left\{ b_{13.2} \sum X_3^2 + a_{1.23} \sum X_3 + b_{12.3} \sum X_2 X_3 \right\} \end{aligned}$$

Las expresiones dentro de los paréntesis corresponden con las ecuaciones normales de un plano de regresión. En efecto:

$$\sum X_{1c}^2 = a_{1.23} \sum X_1 + b_{12.3} \sum X_1 X_2 + b_{13.2} \sum X_1 X_3$$

La fórmula abreviada del coeficiente de determinación queda en consecuencia,

$$R_{1.23}^2 = \frac{a_{1.23} \sum X_1 + b_{12.3} \sum X_1 X_2 + b_{13.2} \sum X_1 X_3 - n \bar{X}_1^2}{\sum X_1^2 - n \bar{X}_1^2}$$

En correlación múltiple no tiene sentido el signo de R, ya que puede haber variables que influyan positiva o negativamente en la variable dependiente.

En cuanto a la forma de cálculo del error de proyección, no difiere de los vistos anteriormente. La diferencia entre numerador y denominador es n veces la varianza no explicada.

/En el

En el caso de correlación múltiple se presenta un problema particular originado en la necesidad de disponer de indicaciones sobre la asociación neta existente entre la variable dependiente y cada una de las variables independientes. El coeficiente de correlación múltiple indica el grado de asociación que simultáneamente se presenta entre la variable dependiente y las variables independientes. Un coeficiente de correlación simple indica el grado de asociación entre dos variables: dependiente e independiente, pero sin eliminar o depurar estadísticamente la asociación entre ambas variables de la influencia de otras variables que actúan a través de la variable independiente. Por ejemplo, puede haber una alta correlación entre la cantidad vendida de un artículo y su precio; pero esta asociación puede disminuir sustancialmente al eliminar explícitamente la influencia de la variable precio de un sustituto.

Este concepto de asociación neta o depurada se cuantifica a través del coeficiente de correlación parcial que en el caso de tres variables se define de la siguiente manera:

$$r_{12.3} = \left(\frac{S_{Xc\ 1.23}^2 - S_{Xc\ 1.3}^2}{S_{Xs\ 1.3}^2} \right)^{\frac{1}{2}}$$

$r_{12.3}$ representa la asociación entre las variables X_1 y X_2 , eliminando estadísticamente la influencia de la variable X_3 . En efecto, si se observa el numerador, se concluye que representa el incremento en la varianza explicada al incluir la variable X_2 . Este incremento se compara con la varianza que dejaba sin explicar la variable X_3 . Sustituyendo el numerador por varianzas totales y no explicadas se tiene:

$$\begin{aligned} r_{12.3}^2 &= \frac{S_{X1}^2 - S_{Xs\ 1.23}^2 - S_{X1}^2 + S_{Xs\ 1.3}^2}{S_{Xs\ 1.3}^2} \\ &= 1 - \frac{S_{Xs\ 1.23}^2}{S_{Xs\ 1.3}^2} \end{aligned}$$

El otro coeficiente de correlación parcial se define:

$$r_{13.2}^2 = \frac{S_{Xc\ 1.23}^2 - S_{Xc\ 1.2}^2}{S_{Xs\ 1.2}^2} = 1 - \frac{S_{Xs\ 1.23}^2}{S_{Xs\ 1.2}^2}$$

Con todos estos estadígrafos en el caso de tres variables se tiene un conjunto de indicadores complementarios que permiten obtener conclusiones objetivas.

Por una parte se dispone de tres coeficientes de correlación simple:

r_{12} , r_{13} , r_{23} . Además dos coeficientes de correlación parcial:

$r_{12.3}$ y $r_{13.2}$. Por último un coeficiente de correlación múltiple:

$r_{1.23}$. Por otra parte, se dispone de todos los errores de proyección correspondientes que permitirán obtener intervalos para las proyecciones.

Las relaciones que se plantean entre estos coeficientes de correlación permiten realizar análisis de consistencia. Cualquier coeficiente de correlación parcial es menor y a lo sumo igual que un coeficiente de correlación simple, por la eliminación explícita de la influencia de otras variables:

$$r_{ij.K} \leq r_{ij}$$

Un coeficiente de correlación múltiple será siempre mayor o al menos igual que un coeficiente de correlación simple, por el hecho de que aquél toma en cuenta un mayor número de variables independientes que explican la variabilidad de la variable dependiente.

$$R_{i\ jk} \geq r_{ij}$$

$$R_{i\ jk} \geq r_{ik}$$

2. Correlación en un hiperplano de regresión. Cuando se tienen más de dos variables independientes, se presenta el caso general de la correlación múltiple. Los conceptos analizados para el caso de

/tres variables

tres variables son también aplicables al caso general. Lo que ocurre es que si se consideran muchas variables independientes se dificulta un tanto el análisis y el cálculo de estadígrafos si no se dispone de computadores, resulta en extremo laborioso. En todo caso, a continuación se presentan las fórmulas de los estadígrafos más importantes para una ecuación lineal que considera 3 variables independientes, es decir:

$$X_{1c} = a_{1.234} + b_{12.34} X_2 + b_{13.24} X_3 + b_{14.23} X_4$$

$$r_{1.234}^2 = \frac{a_{1.234} \sum X_1 + b_{12.34} \sum X_1 X_2 + b_{13.24} \sum X_1 X_3 + b_{14.23} \sum X_1 X_4 - n \bar{X}_1^2}{\sum X_1^2 - n \bar{X}_1^2}$$

El error de proyección se calcula recordando que la diferencia entre numerador y denominador de la fórmula anterior es n veces la varianza no explicada.

Los coeficientes de correlación parcial, aislando el efecto de dos variables, son los siguientes:

$$r_{12.34}^2 = \frac{S_{Xc\ 1.234}^2 - S_{Xc\ 1.34}^2}{S_{Xs\ 1.34}^2}$$

$$r_{13.24}^2 = \frac{S_{Xc\ 1.234}^2 - S_{Xc\ 1.24}^2}{S_{Xs\ 1.24}^2}$$

$$r_{14.23}^2 = \frac{S_{Xc\ 1.234}^2 - S_{Xc\ 1.23}^2}{S_{Xs\ 1.23}^2}$$

/Pueden definirse

Puede definirse otro tipo de coeficientes de correlación parcial donde se aísla el efecto de una variable:

$$r_{123.4}^2 = \frac{S_{Xc\ 1.234}^2 - S_{Xc\ 1.4}^2}{S_{Xs\ 1.4}^2}$$

$$r_{134.2}^2 = \frac{S_{Xc\ 1.234}^2 - S_{Xc\ 1.2}^2}{S_{Xs\ 1.2}^2}$$

$$r_{124.3}^2 = \frac{S_{Xc\ 1.234}^2 - S_{Xc\ 1.3}^2}{S_{Xs\ 1.3}^2}$$

3. Correlación múltiple logarítmica

La linealidad de los planos de regresión vistos anteriormente, puede no ser adecuado en muchos problemas de estimación. En esos casos es conveniente probar con otro tipo de funciones, como por ejemplo:

$$X_{1c} = a_{1.23} + b_{12.3} X_2^2 + b_{13.2} X_3^{\frac{1}{2}}$$

La metodología de obtención de ecuaciones normales para determinar los parámetros de regresión y la deducción de las fórmulas de coeficientes de correlación múltiples y parciales, es la misma que se presentó en páginas anteriores. Con ecuaciones del tipo que se presenta puede representarse adecuadamente la relación entre las variables.

Sin embargo, una función que es muy utilizada en los problemas de predicción es la llamada función logarítmica:

$$X_{1c} = \alpha X_2^\beta X_3^\gamma$$

Para poder aplicar la metodología de los mínimos cuadrados, es
/necesario previamente

necesario previamente "linealizar" esta función, mediante la aplicación de logaritmos.

$$\log X_{1c} = \alpha + \beta \log X_2 + \gamma \log X_3$$

La función así linealizada es similar al caso de correlación múltiple lineal; la única diferencia radica en que en los cálculos deben tomarse los logaritmos de las variables. El coeficiente de correlación múltiple logarítmico, por analogía con el caso de correlación múltiple lineal es: (siendo $\log X = X^*$)

$$R_{\log(1.23)}^2 = \frac{\alpha^* \sum \log X_1 + \beta \sum \log X_1 \log X_2 + \gamma \sum \log X_1 \log X_3 - n \overline{\log X_1}^2}{\sum (\log X_1)^2 - n \overline{\log X_1}^2}$$

La diferencia entre numerador y denominador resulta ser n veces la varianza no explicada. Esta función tiene mucha aplicación por el hecho de que los parámetros β y γ son coeficientes de elasticidad entre las variables X_1 y X_2 y las variables X_1 y X_3 respectivamente. Esta ecuación será tratada con más detalle, en el capítulo correspondiente a las proyecciones por coeficientes de elasticidad.

En cuanto a los coeficientes de correlación parcial, tampoco se presentan diferencias, ya que la metodología de deducción y cálculo no varía; no hay que descuidar eso sí, el trabajo con los logaritmos de las variables que deben tener una precisión equivalente a los 6 decimales.

G. Etapas de la Construcción de un Modelo de Regresión

Es necesario distinguir dos tipos de modelos de regresión en cuanto al objetivo que persiguen: los modelos de análisis, utilizados para /cuantificar relaciones

cuantificar relaciones y explicar adecuadamente lo que sucedió con una variable en términos de otras variables que tienen influencia sobre aquélla y los modelos predictivos que además de ser útiles en el análisis, están diseñados para "predecir" o estimar valores de la variable dependiente en términos de las variables independientes supuesto conocido su comportamiento. Además, es conveniente distinguir entre modelos temporales y atemporales. Los primeros son aquéllos que analizan y estiman valores en el tiempo, por ejemplo, estimación de los precios agrícolas del próximo año en función de las siembras y la política de importaciones. Los modelos atemporales, en cambio, no toman en cuenta ni explícita ni implícitamente la variable tiempo, son cortes transversales. Sería el caso de la estimación de los consumos familiares en función de la variable ingreso, pero teniendo como datos los consumos e ingresos de una muestra en un momento o período dado.

La metodología que a continuación se presenta tiene aplicación general. Sin embargo, se aclarará aquellos puntos que son más delicados en uno y otro tipo de modelo.

1. El primer punto, obviamente es la determinación clara y precisa del objetivo del estudio. Es necesario especificar los objetivos de la investigación general y los objetivos del análisis de regresión y correlación en particular. En esencia es necesario dar respuesta a las interrogantes. ¿En qué se utilizará el modelo? ¿Qué se pretende demostrar por medio de la regresión y correlación?
2. Una vez aclarado el primer punto básico, es necesario hacer un análisis lógico respecto de qué variables deben entrar en el análisis. Debe tomarse en cuenta en principio todas aquéllas que pueden razonablemente estar asociadas a nuestra variable en estudio.
3. A continuación se procede a la recolección de las estadísticas, ya sea históricas si se trata de modelos temporales o las estadísticas del caso si trata de un modelo atemporal.
4. Un análisis de la calidad de los datos recolectados es indispensable.

/En este

En este punto ya quedan eliminadas algunas de las variables que en principio fueron seleccionadas, por el hecho de que sus valores pueden no ser confiables. Por otra parte, pueden haber algunas variables que pese a ser confiables, no pueden ser tomadas en cuenta porque constituyen muy pocas observaciones. En este punto hay que decidir cuál es el número de datos u observaciones que puede considerarse como mínimo. Recuérdese que tamaños de muestra insuficientes conducen a resultados erróneos. En los modelos temporales no puede pensarse en un número inferior a las 10 o 12 observaciones (puntos en el tiempo). Además, en los modelos predictivos el número de variables independientes está condicionado por la posibilidad de disponer con cierta confianza de valores futuros de tales variables.

5. Las variables restantes deben ser depuradas de otras variables que actúan a través de éstas. Como anteriormente se apuntó, es indispensable trabajar con series que representen valor real o "quantum". La consideración de valores nominales exagera la correlación por el hecho de que la variable inflación o alzas de precios puede actuar sobre la variable dependiente y simultáneamente sobre las variables independientes. Es conveniente también, en lo posible, representar las series en términos por habitante, si es que no hubiera un propósito específico para no hacerlo.
6. Una vez que se dispone de las estadísticas de las principales variables depuradas, es necesario determinar la forma y cuantificar el grado de la asociación simple que cada una de estas variables tenga con la variable dependiente que se estudia. Puede ser conveniente también calcular los coeficientes de correlación simple entre las variables independientes para detectar las posibles dependencias que existan entre ellas. A esta altura del análisis ya se tiene bastante definido el campo de la posible metodología que se utilizará finalmente. Por lo menos, estará decidido si se tratará de correlación simple o múltiple.

7. Otro punto de gran importancia es la determinación de la forma general de la función. Si se trata de correlación simple, es de utilidad la representación gráfica, es decir, con la ayuda del diagrama de dispersión puede solucionarse adecuadamente este problema. Si se trata, en cambio, de correlación múltiple, hay que considerar principalmente los coeficientes cuantificados en el punto 5 y las formas particulares de relación entre las variables. A veces se dispone de modelos teóricos ya probados, donde tan sólo se precisa comprobar si tal teoría corresponde al caso que se estudia; por ejemplo, la función consumo de Friedman, donde ya se tienen especificadas las variables independientes y la forma de la función, restando únicamente calcular el valor de los parámetros. El caso más corriente es el de determinar la función (formulación de la teoría) primero en términos conceptuales y segundo, cuantificando resultados. En los modelos temporales un punto delicado es la especificación de los "desfases" entre las variables. Por ejemplo, la producción del período t podría depender de la inversión del período $t - a$, donde a indicaría el tiempo de maduración de la inversión. La representación gráfica por parejas de variables (dependiente e independiente) puede ayudar a la especificación mencionada.
8. El paso siguiente es el de la cuantificación de estadígrafos: medias, varianzas, coeficientes de correlación simples, múltiples, parciales, errores de proyección y, por último, la estimación en los modelos predictivos y el análisis en los modelos descriptivos. Es conveniente también calcular por medio de la ecuación de regresión los valores de la variable dependiente en términos de los valores conocidos de la variable independiente, para compararlos con valores observados y analizar la bondad del ajuste. Es, tal vez, el punto más descuidado en los análisis de regresión y correlación, las formulaciones de pruebas de consistencia entre los estadígrafos calculados. Por otra parte, es aquí donde cabe calificar el análisis a la luz de las cuantificaciones apropiadas.

/Es conveniente

Es conveniente comparar la magnitud de los errores con los valores calculados, estableciendo porcentualmente la cuantía de los probables desvíos.

9. Finalmente, en la presentación de los resultados es imprescindible destacar:

- a) Clara definición de las variables
- b) Tamaño de muestra y tipo de modelo
- c) Forma de la función
- d) Estadígrafos pertinentes.

No se debe dejar de señalar las limitaciones particulares del método, los supuestos utilizados y las fuentes de obtención de informaciones.

H. Métodos de Estimación por Medio del Coeficiente de Elasticidad

1. Presentación conceptual.

Un método muy utilizado en proyecciones de variables socioeconómicas es el que utiliza el coeficiente de elasticidad entre las variables. El coeficiente de elasticidad se define como:

$$E = \frac{\frac{dY}{Y}}{\frac{dX}{X}} = \frac{dY}{dX} \cdot \frac{X}{Y} = Y' \frac{X}{Y}$$

Como puede observarse, el coeficiente de elasticidad es una medida de cambios porcentuales experimentados por una variable Y (dependiente) ante cambios porcentuales de una variable X (independiente).

Implícita en la definición está la función que relaciona ambas variables. Desde el punto de vista estricto, se trata de un cociente entre cambios porcentuales infinitesimales. Cuando se trata de estimar valores de una variable, no interesan los cambios demasiado pequeños, sino que los cambios de significación.

El objetivo inmediato será entonces, encontrar funciones donde el coeficiente de elasticidad sea constante en cualquier punto de la función. Solamente tal tipo de funciones podrán ser utilizadas

/en la

en la proyección, ya que de otra manera el coeficiente de elasticidad variará para cada punto de la función haciendo impracticable la proyección.

Si la función es una recta, el coeficiente de elasticidad no es constante como se ve a continuación.

$$Y = aX + b$$

$$\frac{dY}{dX} = a$$

$$E = \frac{dY}{dX} \cdot \frac{X}{Y} = a \cdot \frac{X}{Y}$$

pero $Y = aX + b$ ∴

$$E = a \frac{X}{aX + b}$$

Como puede observarse, el coeficiente de elasticidad E está en función de X, y entregará un valor distinto para cada valor de X.

Si la función es una hipérbola equilátera, se tiene:

$$Y = \frac{a}{X} = aX^{-1}$$

$$\frac{dY}{dX} = -aX^{-2}$$

$$E = \frac{dY}{dX} \cdot \frac{X}{Y} = -aX^{-2} \cdot \frac{X}{\frac{a}{X}}$$

$$E = -aX^{-2} \cdot \frac{X^2}{a} = -1$$

El resultado se interpreta de modo que aumentos porcentuales en la variable independiente, determinan disminuciones de igual magnitud porcentual en la variable dependiente. En este resultado,

/en consecuencia,

en consecuencia, puede estimarse cualquier valor de la variable dependiente supuesto conocido un valor de la variable independiente. En la función potencial también puede verificarse la constancia del coeficiente de elasticidad. En efecto:

$$Y = b X^a$$

$$\frac{dY}{dX} = ab X^{a-1}$$

$$E = \frac{dY}{dX} \cdot \frac{X}{Y} = ab X^{a-1} \cdot \frac{X}{Y}$$

pero $Y = b X^a$. . .

$$E = ab X^{a-1} \frac{X}{b X^a} = a$$

El hecho de que el coeficiente de elasticidad sea constante en esta función hace que se la utilice periódicamente en las proyecciones. La proyección tiene su base en lo siguiente:

Dada la función:

$$Y = b X^a$$

Aplicando logaritmos:

$$\log Y = \log b + a \log X$$

Las relaciones correspondientes al año 0 (base de proyección) y al año n (periodo en el que se quiere estimar la variable dependiente), son las siguientes:

$$\log Y_0 = \log b + a \log X_0$$

$$\log Y_n = \log b + a \log X_n$$

Restando la primera de la segunda se tiene:

$$\log Y_n - \log Y_0 = a(\log X_n - \log X_0)$$

/El antilogaritmo

El antilogaritmo de la relación anterior:

$$\frac{Y_n}{Y_0} = \left(\frac{X_n}{X_0} \right)^a$$

Observando la fórmula puede concluirse que los cambios porcentuales en la variable dependiente son equivalentes a los cambios porcentuales en la variable independiente, elevados a la potencia a . A veces, erróneamente suele interpretarse la relación potencial. Por ejemplo, si $a = 2$, se dice que un cambio de 50 por ciento en X determinará un cambio de 100 por ciento en Y . Evidentemente, la conclusión es falsa, porque ella supone una relación de linealidad entre las variables que está lejos de presentarse en este caso.

Los datos necesarios para proyectar mediante este método son: disponer del coeficiente de elasticidad (a), conocer el valor base y dado de la variable independiente (X_0 y X_n) o al menos su variación porcentual. Con estos datos puede aplicarse la fórmula citada anteriormente:

$$\frac{Y_n}{Y_0} = \left(\frac{X_n}{X_0} \right)^a$$

Por ejemplo, si $a = 2$

$$Y_0 = 100$$

$$X_0 = 200$$

$$X_n = 300$$

Reemplazando:

$$\frac{Y_n}{100} = \left(\frac{300}{200} \right)^2 = 2,25$$

$$Y_n = 225$$

2. Tipos de Elasticidad

Es necesario distinguir el tipo de elasticidad según las variables que se toman en cuenta. Así, si la variable Y representa consumos y la variable X representa ingresos, se habla de elasticidad ingreso del consumo o de la demanda. Por otra parte, si la variable Y representa consumo y la variable X representa consumo específico de un bien o conjunto de bienes similares, se habla de elasticidad gasto del consumo específico. Estos dos son los más conocidos y utilizados conceptos. Sin embargo, según las denominaciones de las variables puede hablarse de otros distintos tipos de elasticidad, como por ejemplo, elasticidad de la tributación al ingreso, elasticidad del ahorro al Producto, de las importaciones al tipo de cambio, etc.

3. Métodos de Cálculo

A continuación se presentarán las formas de cálculo del coeficiente de elasticidad. Cuando se está utilizando la forma general de proyección;

$$\frac{Y_n}{Y_0} = \left(\frac{X_n}{X_0} \right)^a$$

implícitamente se están aceptando dos supuestos: a) que las variables están relacionadas mediante la función potencial; b) que el coeficiente de correlación entre los logaritmos de X y de Y sea significativo. El cumplimiento de estos supuestos garantiza una buena proyección.

De lo anterior se deduce que una forma de obtención del coeficiente de elasticidad consiste en ajustar la función potencial por el método de mínimos cuadrados a los datos retrospectivos de que se disponga. Es decir, dada

$$Y = b X^a \quad \underline{1/}$$

1/ En la parte final de este capítulo hay adjunta una exposición sobre las diferencias de las proyecciones a través de ecuaciones de regresión y de coeficientes de elasticidad.

/el ajuste

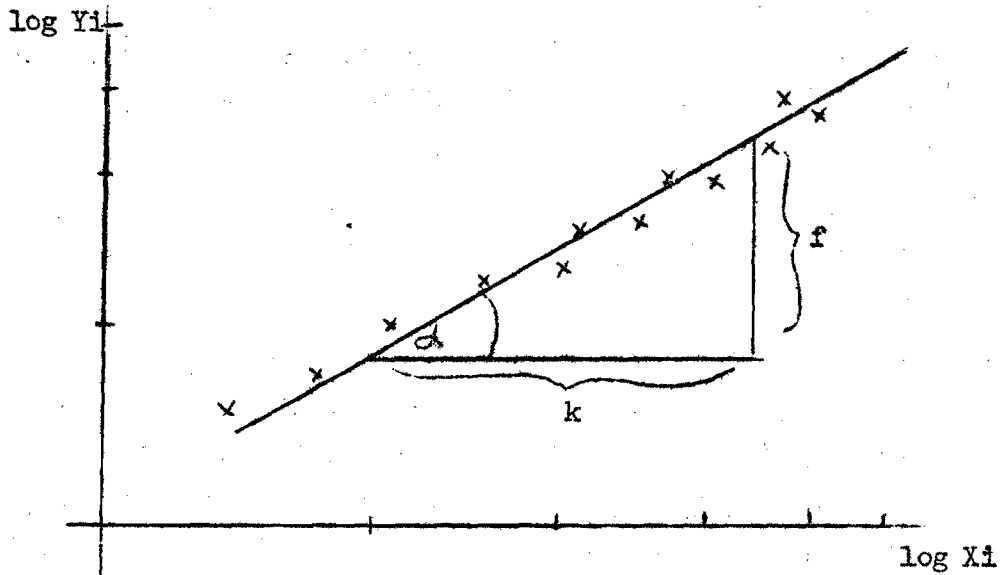
el ajuste a una nube de puntos conocida permitirá calcular los parámetros. El valor de a corresponde, como se demostró, al coeficiente de elasticidad.

Existe una forma aproximada de estimar este coeficiente de elasticidad por el método gráfico. En la expresión:

$$\log Y = \log b + a \log X$$

el coeficiente de elasticidad es el coeficiente angular de la recta logarítmica.

Si los puntos retrospectivos de que se dispone se representan en escalas logarítmicas, es posible a simple vista ajustar la recta tratando de aproximarse a la recta minimocuadrática.



La recta "ajustada a ojo" puede entregar estimaciones muy cercanas al valor efectivo con un ahorro de tiempo considerable. La manera de obtener el valor de " a " es la siguiente:

$$a = T \operatorname{tg} \alpha = \frac{f}{k}$$

donde f es el cateto opuesto al ángulo α en el triángulo del gráfico medido en centímetros u otras unidades de longitud, y k

es el

es el cateto adyacente al ángulo α , también medido en las mismas unidades de f . El cociente dará la inclinación de la recta logarítmica y, por consiguiente, el coeficiente de elasticidad. La bondad de esta estimación depende de la habilidad y cuidado que se tenga para hacer pasar la recta por entre los puntos, de manera de minimizar el cuadrado de las diferencias.

Cualquiera de los dos métodos anteriores supone disponer de estadísticas retrospectivas. Ocurre frecuentemente que es necesario proyectar variables para las cuales no es posible recopilar suficientes antecedentes de manera de garantizar una cierta representatividad del coeficiente de elasticidad.

En estos casos es corriente utilizar comparaciones internacionales, eligiendo países que tengan similitudes marcadas con el país para el cual se precisa hacer la proyección. Por ejemplo, es posible utilizar el coeficiente de elasticidad gasto del consumo de artefactos eléctricos en Colombia, al realizar una primera estimación para Chile. Entre ambos países existen características comunes en cuanto a población y su concentración, nivel de ingreso, grado de industrialización, etc. Otra manera sería seleccionar un conjunto de países dentro de un rango de nivel de ingreso comparable al del país en cuestión y calcular el coeficiente de elasticidad por los métodos anteriores, contando con informaciones de estos países en vez de las estadísticas retrospectivas que se mencionaron. A este método se le conoce con el nombre de estimación del coeficiente de elasticidad por medio de datos internacionales. El último método mencionado tiene por objetivo generalmente la estimación de coeficientes de elasticidad ingreso de la demanda. Finalmente, otra manera de cuantificar coeficientes de elasticidad, principalmente elasticidad gasto, es la realización de muestras en un período dado, donde se averigua los valores que toman las variables que interesa analizar y proyectar. Si bien el hecho de calcular un coeficiente de corte transversal en el tiempo y utilizarlo en proyecciones hacia el futuro, tiene limitaciones,

/hay que

hay que reconocer que si se tiene buen cuidado de definir las unidades muestrales de manera que reflejen internamente las condiciones de una cierta dinámica, las proyecciones no serán distorsionadas seriamente. Previamente deben realizarse pruebas sobre la racionalidad y consistencia de los resultados alcanzables.

4. La ecuación de regresión y el coeficiente de elasticidad como instrumentos de proyección.^{1/}

Como se ha visto, la ecuación de regresión puede utilizarse directamente como instrumento para estimar valores futuros de la variable dependiente, una vez planteadas determinadas hipótesis sobre el comportamiento de la variable independiente. Cabría discutir entonces, qué diferencia existiría entre utilizar la ecuación de regresión o el coeficiente de elasticidad como instrumento de proyección, por ejemplo, de la demanda de un bien en función del ingreso.

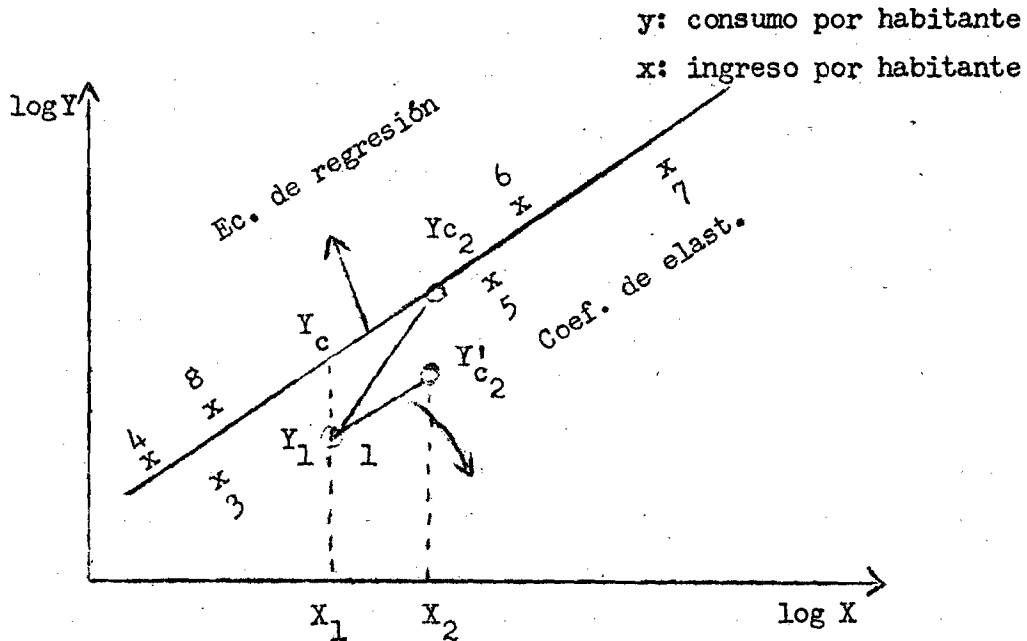
En primer término, la ecuación de regresión es de aplicabilidad mucho más general, ya que podría utilizarse cualquiera que fuese la forma de la relación que se admita entre las dos variables. Desde este punto de vista, el concepto de elasticidad no sería sino un caso particular, en el que como se ha visto, se admite una relación logarítmica.

Podría suceder, sin embargo, que en determinados casos prácticos no pudiera disponerse de la ecuación de regresión correspondiente; en cambio, sí podría ser posible utilizar una estimación del coeficiente de elasticidad. Si sólo se tienen las cifras globales de consumo e ingreso para un solo período reciente, podrían por ejemplo, utilizarse coeficientes de elasticidad deducidos de las experiencias de otros países de condiciones similares.

Aún si se dispusiera de los dos instrumentos - ecuación de regresión y coeficiente de elasticidad-ingreso - y se admitiera una relación

^{1/} El punto 4 es una copia textual de los apuntes del Profesor Pedro Vuscković.

logarítmica, los resultados de las proyecciones a que conducirían uno y otro serían en la generalidad de los casos diferentes. Supóngase, por ejemplo, que las comparaciones correspondientes se hayan referido al consumo de determinado bien en países con distinto nivel de ingreso (países 1, 2, 3, en el gráfico siguiente, siendo el país 1 el que interesa para la proyección).



En la medida en que la relación entre las dos variables esté más alejada de la línea de regresión en el país que interesa para las proyecciones, mayor sería la diferencia a que se llegue utilizando la ecuación de regresión y el coeficiente de elasticidad como instrumento de proyección. Si, como en el gráfico anterior, la relación está en ese país por debajo de la línea de regresión, ello significaría que existe allí un consumo relativamente bajo (en comparación con el nivel de ingreso) del bien de que se trate; una estimación del consumo futuro basado en la ecuación de regresión supondría que tal situación se eliminaría, y el consumo tendería a aumentar no sólo por efecto del incremento del ingreso, sino también para superar ese retraso relativo; la

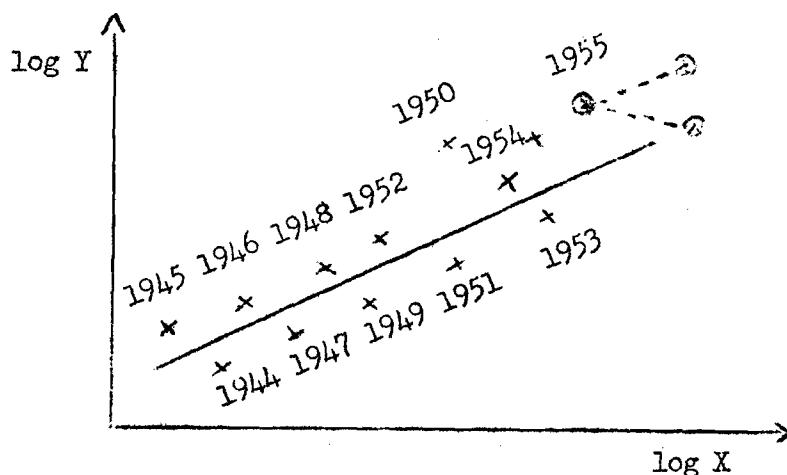
/utilización del

utilización del coeficiente de elasticidad, en cambio, equivaldría a admitir que el consumo aumentará sólo por el efecto ingreso, pero que continuará registrándose un consumo relativamente bajo (en comparación con el nuevo nivel de ingreso).

En otras palabras, al aumentar el ingreso de X_1 a X_2 en la ecuación de regresión, se admite un aumento del consumo de Y_1 a Y_{c2} . En el primer caso, se supone que el nuevo nivel del consumo corresponderá exactamente al valor dado por la ecuación de regresión; en el segundo, se admite que continuará existiendo una discrepancia entre el valor teórico (dado por la ecuación de regresión y el valor efectivo proporcionalmente igual al que existiría en el período base).

Es difícil juzgar en términos generales cuál de los dos métodos podría ser más adecuado ante una situación de esta índole. Si el nivel relativamente bajo del consumo en el país considerado es atribuible a limitaciones de la oferta, u otros factores de carácter temporal, podría ser más adecuada la proyección basada en la ecuación de regresión; si se debe, en cambio, a diferencias en hábitos de los consumidores, a factores climáticos u otros de carácter relativamente permanente, sería más adecuada la proyección basada en el coeficiente de elasticidad. Aún en el primer caso sería necesario tener en cuenta si el período a que se refiere la proyección es lo suficientemente largo como para que lleguen a eliminarse los efectos adversos de los factores temporales. Las diferencias anotadas entre los dos métodos de proyección podrían presentarse también en el caso en que todo el análisis se hubiera basado en series cronológicas correspondientes a un mismo país, puesto que las cifras correspondientes al período que se tome como base seguramente serán diferentes de los valores teóricos dados por la ecuación de regresión.

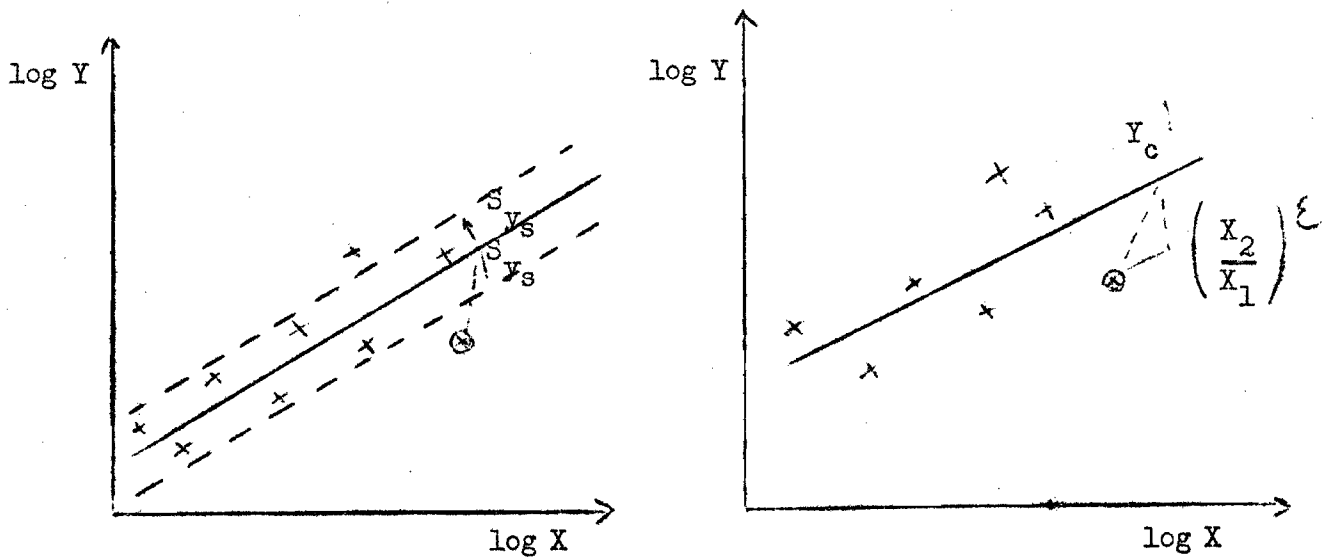
/Gráfico



La interpretación sería, sin embargo, algo diferente, tratándose de un mismo país, el consumo relativamente bajo (o relativamente elevado) registrado en el período base sería una mayor probabilidad atribuible a factores de carácter temporal. En consecuencia, podría considerarse como más adecuada la proyección basada en la ecuación de regresión.

De cualquier modo, en términos generales la utilización de la ecuación de regresión y del coeficiente de elasticidad conducirá en la mayor parte de los casos a proyecciones diferentes, sin que resulte posible precisar cuál de las dos tendría que considerarse más adecuada. Esto puede conducir a la proyección de un intervalo probable para la variable dependiente, basado no en la magnitud del error standard de estimación, sino en la diferencia entre la proyección obtenida con la ecuación de regresión y la proyección a que conduce la aplicación del coeficiente de elasticidad. El gráfico anterior y el que viene a continuación ilustran esta alternativa. En el primer caso se utilizan la ecuación de regresión y el error standard de estimación para proyectar el intervalo correspondiente.

/Al utilizar



Al utilizar $Y_c - S_{ys}$ se está admitiendo no sólo que se elimina el bajo consumo relativo registrado en el período base o en el país correspondiente (si se trata de una comparación internacional), sino además se estima como probable que llegue en el período cubierto por la proyección a registrarse un consumo relativamente elevado. Es evidente que las posibilidades prácticas de que esto aconteciera son muy limitadas. En el segundo caso, se utilizan la ecuación de regresión y el coeficiente de elasticidad, y se proyecta un intervalo delimitado por estos dos valores. De este modo se estima un intervalo más amplio por debajo de la línea de regresión y ninguno por encima de ésta, lo que parecería más lógico en una situación como la supuesta en esos gráficos.

5. Relaciones y propiedades del coeficiente de elasticidad.

a) El coeficiente de elasticidad ingreso del consumo está relacionado con las propensiones media y marginal al consumo de la siguiente manera:

$$E = \frac{dY}{dX} \cdot \frac{X}{Y}$$

Si Y = consumo

X = ingreso

/Fórmula

$$\frac{dY}{dX} = \text{Propensión marginal a consumir}$$

$$\frac{X}{Y} = \text{Inverso de la propensión media al consumo}$$

Luego,

$$E = \text{Propensión marginal} \cdot \frac{1}{\text{Propensión media}}$$

b) Si el consumo total X se divide en consumos parciales $u_1, u_2, u_3 \dots u_k$, de manera que,

$$\sum_{i=1}^k u_i = X,$$

la media aritmética ponderada de las elasticidades gasto respectivas será igual a la unidad.

La definición de la elasticidad gasto en estos términos es la siguiente:

$$E_{gi} = \frac{du_i}{dX} \cdot \frac{X}{u_i}$$

donde $i = 1, 2, \dots, k$ son los componentes del consumo total.

$$\sum E_{gi} W_i = 1$$

donde $W_i = \frac{u_i}{X}$ (participación porcentual del consumo específico dentro del consumo total).

Reemplazando E_{gi} y W_i por las definiciones, se tiene:

$$\sum \frac{du_i}{dX} \cdot \frac{X}{u_i} \left(\frac{u_i}{X}\right) = 1$$

$\sum d(u_i) = dX$, dado que la suma de las diferenciales es /equivalente a

equivalente a la diferencial de la suma:

$$d[\sum u_i] = d X$$

recordando que $\sum u_i = X$

se tiene:

$$d X = d X$$

con lo que se comprueba la proposición enunciada. /

6. Elasticidad en regresión múltiple

Es conveniente presentar el caso del cálculo de elasticidades simultáneas para más de una variable independiente. Es muy frecuente tratar con funciones potenciales múltiples al enfrentar problemas de análisis económico. Determinar, por ejemplo, en forma simultánea cómo juegan las elasticidades con respecto al precio y al ingreso en sus relaciones con la cantidad vendida. Cuál es la elasticidad de la tributación respecto a variaciones en las tasas y variaciones en el ingreso. El tratamiento simultáneo implica evitar la superposición que podría presentarse cuando se hacen cálculos parciales por separado.

Sea la función:

$$Y = \alpha X^\beta W^\gamma$$

La elasticidad X de Y se encontrará derivando parcialmente la función respecto de X y multiplicando por la relación $\frac{X}{Y}$. En otras palabras, aplicando la definición de elasticidad:

$$E = \frac{d Y}{d X} \cdot \frac{X}{Y}$$

En la función que se presenta habrá dos elasticidades; una que relaciona X con Y y otra que relaciona W con Y.

Para el primer caso se tiene:

/Fórmula

$$\frac{dY}{dX} = \alpha W^{\gamma} \beta X^{\beta-1}$$

$$E_X = \alpha W^{\gamma} \beta X^{\beta-1} \cdot \frac{X}{Y}$$

pero $Y = \alpha X^{\beta} W^{\gamma}$

Luego,

$$E_X = \alpha W^{\gamma} \beta X^{\beta-1} \cdot \frac{X}{\alpha X^{\beta} W^{\gamma}}$$

$$E_X = \beta$$

Para el segundo caso se tiene:

$$E_W = \alpha X^{\beta} \gamma W^{\gamma-1} \cdot \frac{W}{\alpha X^{\beta} W^{\gamma}}$$

$$E_W = \gamma$$

Los exponentes de la función potencial múltiple corresponden con los conceptos de elasticidad respectivos.

Para determinar la magnitud de los parámetros α, β, γ , donde los dos últimos corresponden con las elasticidades mencionadas, se sigue el método tradicional del ajuste por mínimos cuadrados. Previamente será necesario "linealizar" la función mediante la aplicación de los logaritmos, es decir:

$$\log Y = \log \alpha + \beta \log X + \gamma \log W$$

Interesa minimizar la expresión:

$$Z = \sum (\log Y_i - \log Y_c)^2$$

$$Z = \sum (\log Y_i - \log \alpha - \beta \log X - \gamma \log W)^2$$

y haciendo:

/Fórmula

$$\frac{\partial z}{\partial \log \alpha} = 0$$

$$\frac{\partial z}{\partial \beta} = 0$$

$$\frac{\partial z}{\partial r} = 0$$

se tienen las tres ecuaciones normales que permitan determinar los parámetros. Estas ecuaciones normales son:

$$\sum \log Y_i = n \log \alpha + \beta \sum \log X_i + r \sum \log W_i$$

$$\sum (\log Y_i) \log X_i = \log \alpha \sum \log X_i + \beta \sum (\log X_i)^2 + r \sum \log W_i \log X_i$$

$$\sum \log Y_i \log W_i = \log \alpha \sum \log W_i + \beta \sum \log X_i \log W_i + r \sum (\log W_i)^2$$

donde Y_i , W_i , X_i son los valores observados de las tres variables, ya sea que correspondan a valores en el tiempo o en el espacio (temporal o atemporal). Los límites de las sumatorias corresponden al total de observaciones que simultáneamente sobre las tres variables se disponga.

Con una función ajustada de esa manera, pueden realizarse predicciones y análisis entre las variables. Para las proyecciones, al igual que en el caso de regresión simple, habrá la alternativa de hacerlo o a través de la ecuación de regresión, o a través de los coeficientes de elasticidad.

Para la proyección por medio de la ecuación de regresión, bastará con fijar exógenamente el comportamiento de las variables independientes y reemplazar tales valores en la función. Si se desea proyectar a través de los coeficientes de elasticidad se tiene para el período 0:

$$\log Y_0 = \log \alpha + \beta \log X_0 + r \log W_0$$

para el período n:

$$\log Y_n = \log \alpha + \beta \log X_n + r \log W_n$$

/restando ambas

restando ambas ecuaciones:

$$\log Y_n - \log Y_0 = \beta (\log X_n - \log X_0) + \gamma [\log W_n - \log W_0]$$

El antilogaritmo de la anterior relación conduce a:

$$\frac{Y_n}{Y_0} = \left(\frac{X_n}{X_0}\right)^\beta \left(\frac{W_n}{W_0}\right)^\gamma$$

que es la fórmula básica de proyección a través de coeficientes de elasticidad en el caso de más de una variable independiente. Como ejemplo de una proyección de este tipo, admítanse los siguientes casos: Y variable que representa la recaudación efectiva tributaria. X variable que representa la tasa tributaria promedio. W representa el Producto Geográfico real.

Si se tienen estimaciones de que el producto crecerá en los próximos cinco años en 20 por ciento siendo la elasticidad producto de la tributación unitaria, y se desea subir la tasa promedio en 44 por ciento siendo la elasticidad tasa de la tributación equivalente a 0,5, el incremento porcentual de la recaudación tributaria será:

$$\frac{Y_n}{Y_0} - 1 = (1,44)^{0,5} \cdot (1,20)^1 - 1$$

$$\frac{Y_n}{Y_0} - 1 = (1,2)(1,2) - 1 = 1,44 - 1 = 0,44$$

Evidentemente que el tratamiento puede extenderse a más de dos variables independientes. La metodología presentada es de fácil extensión a tales casos.

7. Limitaciones en la utilización de coeficientes de elasticidad. Pese a la profusa, tal vez exagerada, utilización de coeficientes de elasticidad en las proyecciones económicas, hay que reconocer que su aplicación en países donde los cambios político-económicos se suceden con inusitada frecuencia, tiene serias limitaciones. Se tratará el caso principal de los coeficientes de elasticidad ingreso.

/Un coeficiente

Un coeficiente de este tipo calculado, por ejemplo, con estadísticas retrospectivas, lleva implícita la distribución de ingresos y estructura de preferencias de los consumidores en el período que comprenden las estadísticas retrospectivas. La objeción inmediata se presenta cuando se piensa en proyecciones al futuro donde precisamente el objetivo puede ser cambiar esa distribución de ingresos. Igual objeción puede hacerse a las proyecciones por medio de coeficientes de elasticidad calculados a partir de muestras de corte transversal en el tiempo. De la misma manera los coeficientes de elasticidad resultantes de comparaciones internacionales llevan implícita la distribución de ingresos de los países considerados.

El problema de las proyecciones, supone un método de aproximaciones sucesivas. Los tres métodos planteados para calcular coeficientes de elasticidad no son métodos alternativos, sino que complementarios. Un coeficiente calculado a través de estadísticas retrospectivas puede ser corregido a través de muestra sucesivas que detecten los cambios en la distribución de ingresos. Por otra parte, la utilización de coeficientes "internacionales" supone seguir las distribuciones de ingresos de los países considerados, y en determinados casos esa tendencia puede no estar demasiado apartada de los planes que sobre esta materia se tenga. En general, es posible llegar a proyecciones razonables considerando el problema como un método iterativo sujeto a revisiones periódicas. En los modelos temporales, el transcurso del tiempo proporciona nuevas informaciones, que, al ser tomadas en cuenta, modifican las proyecciones de medio y largo plazo. Ese debiera ser el verdadero sentido de las proyecciones y no la estimación esporádica. En las instituciones más avanzadas existen equipos de técnicos que están dedicados permanentemente a la realización, corrección, revisión e integración de modelos predictivos. Como regla general para calificar una proyección, debiera establecerse lo siguiente. Por una parte, debe haber un mínimo razonable de confiabilidad /en la

en la proyección, porque malas proyecciones en general pueden ocasionar serios perjuicios. Por otra parte, aunque no se disponga de estimaciones certeras, disponer de aproximaciones, aunque burdas, siempre será mejor que el desconocimiento o la ignorancia de lo que puede ser un fenómeno en el futuro.

I. Modelos de Regresión^{1/}

1. Introducción

Anteriormente se presentaron los temas de regresión y correlación desde el punto de vista de su interpretación conceptual y utilización por parte del economista y/o del planificador. La presentación anterior, en consecuencia, no contemplaba, sino muy indirectamente, un tratamiento probabilístico que permitiera "probar" los valores resultantes de los parámetros respecto a hipótesis que sean pertinentes. En esta nueva presentación del tema se introduce un término estocástico que representa la magnitud y sentido de los desvíos de los valores observados respecto de los valores calculados por las ecuaciones de regresión, cuando se considera la totalidad de las observaciones. La introducción y tratamiento de variables estocásticas y la asociación con distribuciones de probabilidad, conducen a planteamientos más generales y rigurosos sobre estos temas. Nuevamente surgirá el problema de muestreo que tiene directa relación con la representatividad y significación de los estadígrafos asociados al análisis de regresión y correlación. Un modelo de regresión implica de cualquier manera una simplificación de la realidad. Al establecer la dependencia de una cierta variable respecto de otras, pueden establecerse una cantidad innumerable de variables independientes que pueden influir sobre la variable dependiente. Recolectar todas las informaciones necesarias de todas las variables ligadas, es prácticamente imposible.

1/ Notas de clase extractadas de los textos: Econometric Methods de J. Johnston, Estadística Matemática de P. Hoel, e Introducción a la Teoría de la Estadística de M. Mood.

Más aún, si se piensa que habrá una serie de variables cualitativas cuya influencia no es susceptible de cuantificación. Por lo demás, esa lista exhaustiva seguramente no tendrá sentido, ya que la mayoría de las variables ayudarán muy poco a explicar el comportamiento de la variable dependiente, y puede optarse por elegir las más influyentes dejando el resto de las influencias representadas por el término estocástico u . Cabe esperar que este término tenga en general valores pequeños porque habrá cierta compensación de influencias del resto de las variables; unas influirán positivamente y otras en forma negativa, siendo el efecto neto " u " un valor no demasiado grande. Por ello es posible pensar en esta variable estocástica como una variable con una distribución de probabilidad centrada en cero y con varianza finita σ_u^2 , como supuesto y apelando al teorema central del límite, puede asociársela a una distribución normal que satisface las condiciones antes mencionadas. Por otra parte, puede justificarse además la introducción del término estocástico admitiendo que pese a considerar las más importantes variables, existe un elemento impredecible de aleatoriedad en las actitudes humanas que solamente puede caracterizarse en forma adecuada, mediante un término aleatorio. Finalmente, existe una justificación ulterior para la introducción de términos aleatorios, que dice relación con los errores de observación o de medida, siempre que no constituyan un sesgo propiamente tal y por el contrario, tengan un carácter aleatorio.

Dentro del análisis económico no se necesita forzar mucho la mente para encontrar la importancia de los planteamientos anteriores; el estudio de relaciones entre las variables, cantidad demandada, ingreso, precio del bien, precio de los sustitutos, precio de los bienes complementarios, funciones de producción en general, etc., constituyen las bases del análisis económico. La necesidad de explicitar estas relaciones de dependencia y cuantificar estadísticamente los parámetros, estadígrafos de análisis y sus correspondientes errores, llevan a considerar los modelos que aquí se presentan.

2. El modelo lineal de dos variables.

a) Presentación.

Con propósitos de simplificación didáctica se considerará el modelo lineal de dos variables, y dentro de esta familia de modelos, se destacará con detalle el modelo rectilíneo, ya que permite una presentación más clara de los diferentes conceptos que interesa caracterizar.

En los capítulos anteriores se habría tratado la ecuación de regresión de la recta bajo la forma:

$$Y = \beta X + \alpha$$

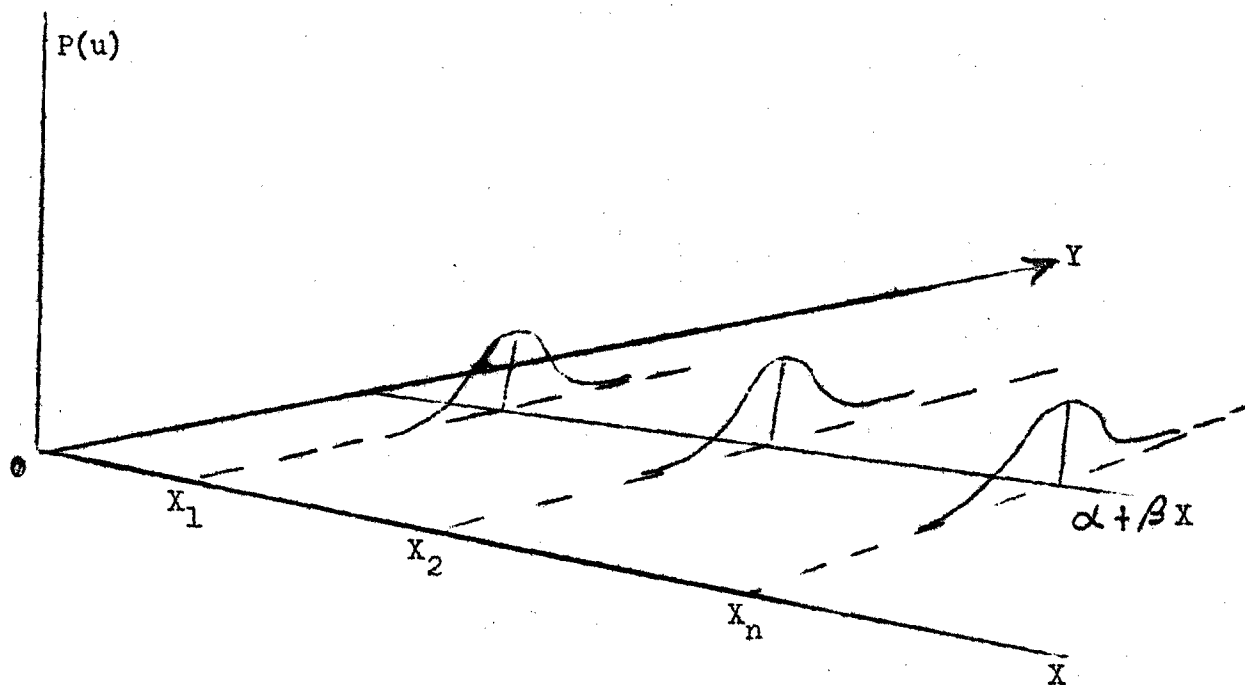
La introducción del término de desvío estocástico modifica la ecuación de la siguiente manera:

$$Y = \beta X + \alpha + u$$

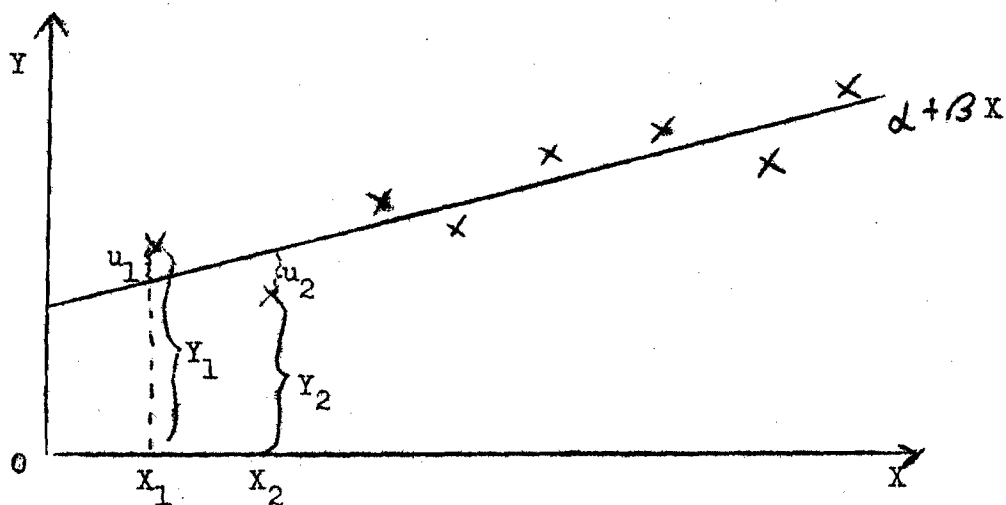
donde u es la variable estocástica que puede tomar valores positivos o negativos. Los supuestos que además se hacen sobre la variable u son: que tiene media aritmética nula, varianza constante e independiente de X y que los diferentes valores de u son independientes entre sí. El supuesto de constancia de la varianza de u (σ_u^2) y su independencia respecto de los valores de X , tienen el objetivo de simplificar el tratamiento y no constituyen limitación, ya que tales supuestos son levantados en un tratamiento más general.

Para aclarar estas ideas, piénsese, por ejemplo, en que se desea determinar los parámetros de la regresión rectilínea entre ingresos (X_i) y consumos familiares (Y_i). Admítase que se tienen varios niveles de ingreso familiar $X_1, X_2 \dots X_n$ (de menor a mayor). Para cada uno de estos valores de ingreso familiar corresponderá una cantidad de valores de consumos familiares, unos serán más bajos que el valor dado por la ecuación de regresión, y otros serán más altos. La distribución de probabilidad de estos desvíos es la que se supone con media nula y varianza constante. En el siguiente diagrama puede observarse la representación del ejemplo propuesto y los conceptos señalados.

/Gráfico



Si en vez de tener informaciones para el total de la población de familias se extrajera una muestra de n familias, una de cada nivel de ingreso, estos puntos muestrales pueden ser representados en el siguiente diagrama.



La aleatoriedad de los desvíos u_i , es una garantía de que el resto de las variables dejadas de lado, no tienen una influencia /significativa en

significativa en el análisis. Si por otra parte los desvíos hubiesen mostrado tendencia determinada en cuanto a su signo, habría querido decir que una variable importante no ha sido tomada en cuenta.

El problema que frecuentemente se enfrenta es aquél en que se dispone de n pares de valores muestrales (ingreso y consumo familiar por ejemplo), pero se desconoce la función.

$$Y = \alpha + \beta X_i + u_i$$

calculada para toda la población. El caso real, en consecuencia, implica desconocer los parámetros α y β , y por otra, la necesidad de estimarlos. Para ello se plantea el siguiente conjunto de hipótesis (supuestos del modelo).

$$Y = \alpha + \beta X_i + u_i \quad [I]$$

$i = 1, 2, 3 \dots n$

$$E(u_i) = 0 \quad \text{para todo } i = 1, 2, 3 \dots n$$

$$E(u_i u_j) = 0 \quad \text{si } i \neq j = 1, 2 \dots n$$

$$E(u_i u_j) = \sigma_u^2 \quad \text{si } i = j = 1, 2, 3 \dots n$$

Como se dijo, α , β y σ_u^2 , son los parámetros desconocidos que deberán ser estimados a partir de los datos muestrales. Posteriormente, y a la luz de esos resultados, podrán formularse pruebas de hipótesis acerca de los valores de dichos parámetros. Para la estimación de α y β , se dispone del conocido método de ajuste de los mínimos cuadrados que como se recordará cumple con el requisito de minimizar la expresión siguiente:

$$Z = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

donde Y_i son los valores observados y \hat{Y}_i son los valores calculados
/por la

por la ecuación de regresión. Resolviendo el sistema de ecuaciones normales provenientes de la derivación parcial de Z respecto de cada uno de los parámetros, igualada a cero, se determinará el valor de cada uno de estos parámetros, es decir:

$$\frac{\partial Z}{\partial \hat{\alpha}} = 0 \quad \text{y} \quad \frac{\partial Z}{\partial \hat{\beta}} = 0$$

Con ello, se habrá determinado la ecuación:

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X \quad \text{[II]}$$

donde $\hat{\alpha}$ y $\hat{\beta}$ son estimadores de α y β respectivamente.

Las ecuaciones normales que resultan de aplicar el método de los mínimos cuadrados son:

$$\begin{aligned} \sum_{i=1}^n Y_i &= n \hat{\alpha} + \hat{\beta} \sum_{i=1}^n X_i \\ \sum X_i Y_i &= \hat{\alpha} \sum X_i + \hat{\beta} \sum_{i=1}^n X_i^2 \end{aligned} \quad \text{[III]}$$

Recuérdese que la ecuación de regresión pasará por el punto en que se cortan las medias aritméticas de las variables, es decir:

$$\bar{Y} = \hat{\alpha} + \hat{\beta} \bar{X} \quad \text{[IV]}$$

Restando II de IV se tiene:

$$\hat{Y} - \bar{Y} = \hat{\beta} (X - \bar{X})$$

Denominando con letras minúsculas a las desviaciones respecto de las medias aritméticas, se tendrá:

$$x_i = X_i - \bar{X} \quad y_i = Y_i - \bar{Y} \quad \hat{y}_i = \hat{Y}_i - \bar{Y}$$

Recuérdese que:

$$M[Y_i] = M[\hat{Y}_i] = \bar{Y}$$

/La ecuación

La ecuación de los mínimos cuadrados puede escribirse en consecuencia:

$$\hat{y} = \hat{\beta} x \quad [V]$$

Por otra parte, denominando por e_i los desvíos de los valores muestrales respecto de los valores calculados por la ecuación de regresión muestral, se tiene:

$$e_i = Y_i - \hat{Y}_i$$

Sumando y restando la media aritmética:

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

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta} x_i$$

Elevando al cuadrado y aplicando sumatoria:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\beta} x_i)^2$$

Para minimizar la expresión anterior se deriva respecto de $\hat{\beta}$.

$$\begin{aligned} \frac{\partial \sum e_i^2}{\partial \hat{\beta}} &= 2 \sum (y_i - \hat{\beta} x_i)(-x_i) = 0 \\ &= -\sum y_i x_i + \hat{\beta} \sum x_i^2 = 0 \end{aligned}$$

Luego:

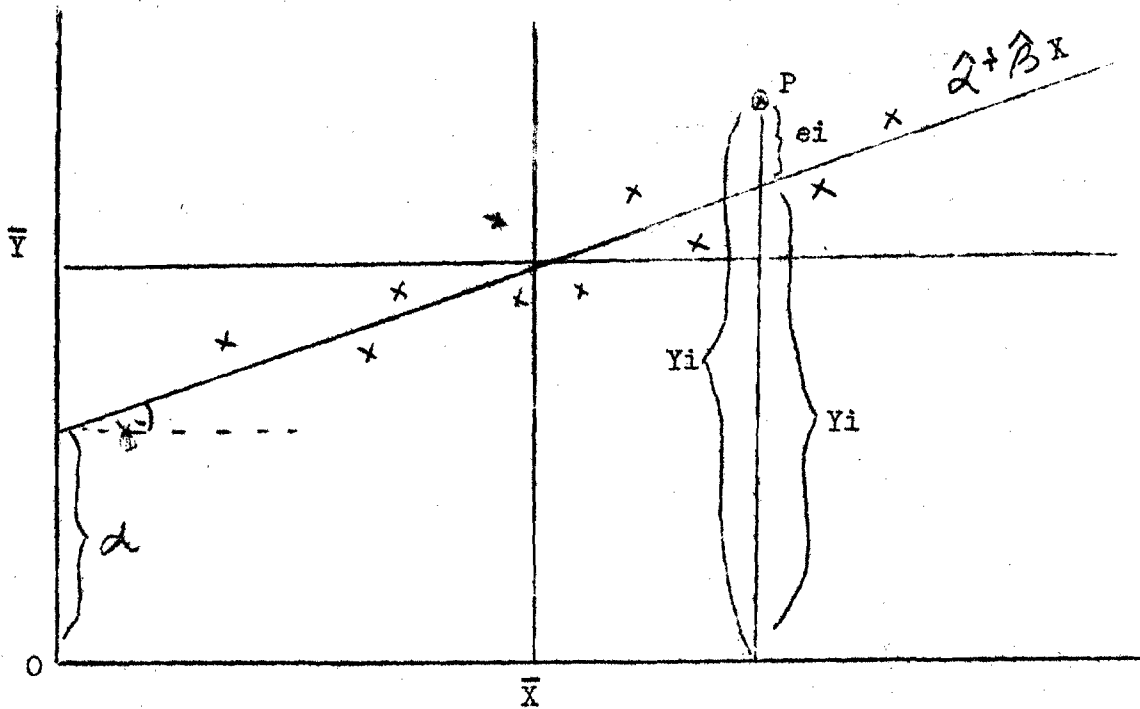
$$\hat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2} \quad [VI]$$

La expresión de equivalencia de $\hat{\alpha}$ se obtiene de IV:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

En el diagrama siguiente se representan los conceptos analizados:

/Gráfico



donde $\text{tg } \gamma = \hat{\beta}$

b) Propiedades de los estimadores.

A continuación se presentarán algunas propiedades de los estimadores $\hat{\alpha}$ y $\hat{\beta}$, para ello se supondrá que la variable independiente es un conjunto de valores fijos y que la variable dependiente es la que toma valores distintos en cada muestra, dados los valores X_i en forma exógena. El primer punto es convenir acerca de la linealidad de los estimadores por mínimos cuadrados respecto de las observaciones sobre la variable Y_i .

Se tenía que:

$$\begin{aligned} \hat{\beta} &= \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{pero } y_i = Y_i - \bar{Y} \quad \therefore \\ &= \frac{\sum x_i Y_i}{\sum x_i^2} - \frac{\bar{Y} \sum x_i}{\sum x_i^2} \quad \text{dado que } \sum_{i=1}^n x_i = 0 \end{aligned}$$

y haciendo:

/Fórmula

$$W_i = \frac{x_i}{\sum x_i^2} \quad \text{se tiene:}$$

$$\hat{\beta} = \sum_{i=1}^n W_i Y_i \quad \text{[VII]}$$

Dado que W_i está en función de X , también se considerará como valores fijos en todas las muestras repetidas que se extraigan.

Además:

$$\sum W_i = \sum \frac{x_i}{\sum x_i^2} = \frac{\sum x_i}{\sum x_i^2} = 0 \quad \text{[VIII a]}$$

Por otra parte:

$$W_i^2 = \frac{x_i^2}{(\sum x_i^2)^2}$$

$$\sum W_i^2 = \frac{\sum x_i^2}{(\sum x_i^2)^2} = \frac{1}{\sum x_i^2} \quad \text{[VIII b]}$$

Finalmente

$$\sum W_i x_i = \sum W_i X_i = 1 \quad \text{[VIII c]}$$

ya que:

$$\sum W_i (X_i - \bar{X}) = \sum W_i X_i - \bar{X} \sum W_i$$
$$\text{y } \sum W_i = 0$$

Para el otro parámetro $\hat{\alpha}$ se tiene:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

$$\hat{\alpha} = \bar{Y} - \bar{X} \cdot \sum W_i Y_i$$

$$\hat{\alpha} = \frac{\sum Y_i}{n} - \bar{X} \sum W_i Y_i$$

$$\hat{\alpha} = \sum \left(\frac{Y_i}{n} - \bar{X} W_i Y_i \right)$$

$$\hat{\alpha} = \sum \left(\frac{1}{n} - \bar{X} W_i \right) Y_i$$

[IX]

/Como podrá

Como podrá concluirse al final de las demostraciones siguientes, los estimadores $\hat{\alpha}$ y $\hat{\beta}$ son estimadores insesgados de los parámetros α y β (poblacionales).

En efecto, se tenía que:

$$Y_i = \alpha + \beta X_i + u_i$$

$$\hat{\beta} = \sum W_i Y_i$$

$$\hat{\beta} = \sum W_i (\alpha + \beta X_i + u_i)$$

$$\hat{\beta} = \alpha \sum W_i + \beta \sum W_i X_i + \sum W_i u_i$$

$$\hat{\beta} = 0 + \beta + \sum W_i u_i$$

[X]

Aplicando el operador esperanza matemática:

$$E(\hat{\beta}) = E(\beta) + E(\sum W_i u_i)$$

dado que W_i también son valores fijos por ser función de X_i , y recordando las propiedades de la sumatoria:

$$E(\hat{\beta}) = \beta + \sum W_i E(u_i)$$

pero por la hipótesis inicial $E(u_i) = 0$

Luego:

$$E(\hat{\beta}) = \beta$$

y se concluye que $\hat{\beta}$ es un estimador insesgado y consistente de β .

Respecto del otro parámetro la demostración es similar:

En IX se tenía:

$$\hat{\alpha} = \sum \left(\frac{1}{n} - \bar{X} W_i \right) Y_i$$

pero $Y_i = \alpha + \beta X_i + u_i$

Luego:

$$\hat{\alpha} = \sum \left(\frac{1}{n} - \bar{X} W_i \right) (\alpha + \beta X_i + u_i)$$

/Fórmula

$$\hat{\alpha} = \alpha - \alpha \bar{X} \sum W_i + \beta \bar{X} - \beta \bar{X} \sum W_i X_i + \sum \left(\frac{1}{n} - \bar{X} W_i \right) u_i$$

pero $\sum W_i = 0$ $\sum W_i X_i = 1$

$$\hat{\alpha} = \alpha + \sum \left(\frac{1}{n} - \bar{X} W_i \right) u_i \quad [XI]$$

$$E(\hat{\alpha}) = E(\alpha) + E \left[\sum \left(\frac{1}{n} - \bar{X} W_i \right) u_i \right]$$

La esperanza matemática, nuevamente, se aplica solamente al término u_i , ya que el resto son fijos.

$$E(\hat{\alpha}) = \alpha + \sum \left(\frac{1}{n} - \bar{X} W_i \right) E(u_i)$$

pero $E(u_i) = 0$ (hipótesis original.)

Luego:

$$E(\hat{\alpha}) = \alpha$$

Luego $\hat{\alpha}$ es un estimador insesgado y consistente de α .

Para determinar las varianzas de cada uno de estos estimadores, se tiene de [X].

$$\hat{\beta} = \beta + \sum_{i=1}^n W_i u_i$$

Recuérdese que:

$$V(\hat{\beta}) = E[(\hat{\beta} - \beta)^2]$$

$$\hat{\beta} - \beta = \sum W_i u_i$$

$E[(\hat{\beta} - \beta)^2] = E[(\sum W_i u_i)^2]$ desarrollando el cuadrado de la sumatoria:

$$= E[(W_1 u_1)^2 + (W_2 u_2)^2 + (W_n u_n)^2 +$$

$$+ 2 W_1 u_1 W_2 u_2 + \dots + 2 W_{n-1} u_{n-1} W_n u_n]$$

/Nuevamente el

Nuevamente el operador esperanza matemática sólo afecta la variable u_i , ya que se suponen fijos los valores de X_i que determinan W_i .

$$\begin{aligned} E[(\hat{\beta} - \beta)^2] &= E\left[\sum W_i^2 u_i^2\right] + 2E\left[\sum_{i \neq j} W_i u_i W_j u_j\right] \\ &= \sum W_i^2 E(u_i^2) + 2\left[\sum W_i W_j E(u_i u_j)\right] \end{aligned}$$

pero $E(u_i^2) = \sigma^2$

$E(u_i u_j) = 0$ para $i \neq j$

Luego:

$$E[(\hat{\beta} - \beta)^2] = \sigma^2 \sum_{i=1}^n W_i^2$$

$$V(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2} u^2$$

[XII]

Para el otro estimador se tiene de [XI]

$$\hat{\alpha} = \alpha + \sum \left(\frac{1}{n} - \bar{X} W_i\right) u_i$$

Por otra parte,

$$V(\hat{\alpha}) = E[(\hat{\alpha} - \alpha)^2]$$

$$V(\hat{\alpha}) = E[(\hat{\alpha} - \alpha)^2] = E\left\{\sum \left[\left(\frac{1}{n} - \bar{X} W_i\right) u_i\right]^2\right\}$$

$$V(\hat{\alpha}) = E\left[\sum \left(\frac{1}{n} - \bar{X} W_i\right)^2 u_i^2\right]$$

$$V(\hat{\alpha}) = \sum \left(\frac{1}{n} - \bar{X} W_i\right)^2 E(u_i^2)$$

$$V(\hat{\alpha}) = \sigma^2 \sum \left(\frac{1}{n} - \bar{X} W_i\right)^2$$

$$V(\hat{\alpha}) = \sigma^2 \sum \left(\frac{1}{n^2} - \frac{2\bar{X} W_i}{n} + \bar{X}^2 W_i^2\right)$$

$$V(\hat{\alpha}) = \sigma^2 \left[\sum \frac{1}{n^2} - 2\bar{X} \frac{\sum W_i}{n} + \bar{X}^2 \sum W_i^2\right]$$

$V(\hat{\alpha}) =$

$$V(\hat{\alpha}) = \sigma^2 u^2 \left[\frac{1}{n} + \bar{X}^2 \sum W_i^2 \right] \quad \text{ya que } \sum W_i = 0$$

$$V(\hat{\beta}) = \sigma^2 u^2 \left[\frac{1}{n} + \bar{X}^2 \frac{1}{\sum x_i^2} \right] \quad \text{ya que } \sum W_i^2 = \frac{1}{\sum x_i^2}$$

$$V(\hat{\alpha}) = \sigma^2 u^2 \left[\frac{\sum x_i^2 + \bar{X}^2 n}{n \sum x_i^2} \right]$$

$$V(\hat{\beta}) = \sigma^2 u^2 \left[\frac{\sum x_i^2 - n \bar{X}^2 + \bar{X}^2 n}{n \sum x_i^2} \right]$$

$$V(\hat{\alpha}) = \sigma^2 u^2 \frac{\sum x_i^2}{n \sum x_i^2} \quad \text{[XIII]}$$

Ya se dispone de fórmulas para la varianza de ambos estimadores:

$V(\hat{\alpha})$ y $V(\hat{\beta})$.

Para la determinación de la covarianza de los estimadores, se tiene:

$$\begin{aligned} \text{Cov}(\hat{\alpha}, \hat{\beta}) &= E[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)] \\ &= E(\hat{\alpha}\hat{\beta}) - E(\hat{\alpha})E(\hat{\beta}) \\ &= E(\hat{\alpha}\hat{\beta}) - \alpha\beta \end{aligned}$$

ya que:

$$\begin{aligned} E[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)] &= E(\hat{\alpha}\hat{\beta} - \hat{\alpha}\beta - \alpha\hat{\beta} + \alpha\beta) \\ &= E(\hat{\alpha}\hat{\beta}) - \beta E(\hat{\alpha}) - \alpha E(\hat{\beta}) + \alpha\beta \\ &= E(\hat{\alpha}\hat{\beta}) - \beta\alpha - \alpha\beta + \alpha\beta \\ &= E(\hat{\alpha}\hat{\beta}) - \alpha\beta \end{aligned}$$

Luego si:

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = E[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)]$$

de [X] se tiene:

$$\hat{\beta} - \beta =$$

$$\hat{\beta} - \beta = \sum W_i u_i$$

De [XI] se tiene:

$$\hat{\alpha} - \alpha = \sum \left(\frac{1}{n} - \bar{X} W_i \right) u_i$$

$$\hat{\alpha} - \alpha = \frac{\sum u_i}{n} - \bar{X} \sum W_i u_i$$

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = E \left[\sum \left(\frac{1}{n} - \bar{X} W_i \right) u_i \sum W_i u_i \right]$$

$$\text{Cov}(\hat{\alpha}, \hat{\alpha}) = E \left[\sum W_i u_i - \bar{X} \left(\sum W_i u_i \right)^2 \right]$$

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\bar{X} \left(\sum W_i^2 \right) \sigma^2 u^2$$

Recuérdese que se demostró que:

$$E \left[\left(\sum W_i u_i \right)^2 \right] = \sigma^2 \sum W_i^2$$

Finalmente, ya que $\sum W_i^2 = \frac{1}{\sum x_i^2}$:

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X} \sigma^2 u^2}{\sum x_i^2}$$

[XIV]

c) Fórmulas prácticas de cómputo.

Hasta el momento, en lo que se refiere a varianzas y covarianza, se dispone de las siguientes fórmulas:

$$V(\hat{\beta}) = \frac{\sigma^2 u^2}{\sum x_i^2}$$

$$V(\hat{\alpha}) = \frac{\sigma^2 u^2 \sum x_i^2}{n \sum x_i^2}$$

$$\text{Cov}(\hat{\alpha}, \hat{\beta}) = -\frac{\bar{X} \sigma^2 u^2}{\sum x_i^2}$$

Observando tales fórmulas se concluye que no son útiles para

/el cómputo

el cómputo, por el hecho de que están en función de $\sigma^2 u^2$ que es la varianza de los desvíos de los valores observados Y_i , respecto de los valores calculados por la ecuación de regresión poblacional:

$$Y_i = \alpha + \beta X_i + u_i$$

Como sólo se dispone de estimaciones de α y β que son $\hat{\alpha}$ y $\hat{\beta}$, en otras palabras, como la ecuación que es posible ajustar es la ecuación de regresión muestral, no es factible calcular los desvíos u_i y menos su varianza. Será necesario disponer de un estimador de $\sigma^2 u^2$ para que dichas fórmulas sean operativas. Para ello se utilizarán las desviaciones muestrales e_i .

$$e_i = y_i - \hat{\beta} x_i$$

Si se promedian por otra parte los valores:

$$Y_i = \alpha + \beta X_i + u_i, \text{ se tiene:}$$

$$\bar{Y} = \alpha + \beta \bar{X} + \bar{u}$$

restando se llega a:

$$Y_i - \bar{Y} = y_i - \beta(x_i) + (u_i - \bar{u})$$

Luego:

$$e_i = \beta x_i + u_i - \bar{u} - \hat{\beta} x_i$$

$$e_i = x_i(\beta - \hat{\beta}) + u_i - \bar{u}$$

$$e_i = -x_i(\hat{\beta} - \beta) + u_i - \bar{u}$$

Elevarlo al cuadrado y sumando:

$$\sum e_i^2 = \underbrace{(\hat{\beta} - \beta)^2 \sum x_i^2}_A + \underbrace{\sum (u_i - \bar{u})^2}_B - \underbrace{2(\hat{\beta} - \beta) \sum x_i(u_i - \bar{u})}_C$$

$$\sum e_i^2 = A + B - 2C$$

$$E(\sum e_i^2) = E(A) + E(B) - 2E(C)$$

$$\text{Pero: } A = (\hat{\beta} - \beta)^2 \sum x_i^2$$

$$/E(A) =$$

$$E(A) = E[(\hat{\beta} - \beta)^2 \sum x_i^2]$$

$$E(A) = \sum x_i^2 E[(\hat{\beta} - \beta)^2]$$

$$E(A) = \sum x_i^2 v(\hat{\beta})$$

$$E(A) = \sum x_i^2 \frac{\sigma_u^2}{\sum x_i^2} = \sigma_u^2$$

$$E(A) = \sigma_u^2$$

Por otra parte:

$B = \sum (u_i - \bar{u})^2$ que es el numerador de la varianza de Cochran (cuasivarianza).

$$E(B) = (n - 1) \sigma_u^2$$

En efecto:

$$\begin{aligned} E[\sum (u_i - \bar{u})^2] &= E[\sum u_i^2 - n \bar{u}^2] \\ &= E[\sum u_i^2 - n \left(\frac{\sum u_i}{n}\right)^2] \\ &= E[\sum u_i^2 - \frac{(\sum u_i)^2}{n}] \\ &= \sum^n E(u_i^2) - \frac{1}{n} E(u_1 + u_2 + \dots + u_n)^2 \\ &= n \sigma_u^2 - \frac{1}{n} E[\sum u_i^2 + 2 \sum_{i \neq j} u_i u_j] \\ &= n \sigma_u^2 - \frac{1}{n} E[\sum u_i^2] - 2 E[\sum_{i \neq j} (u_i u_j)] \\ &= n \sigma_u^2 - \frac{1}{n} \sum E(u_i^2) - 2 \sum \underbrace{E(u_i u_j)}_0 \\ &= n \sigma_u^2 - \frac{1}{n} n \sigma_u^2 \\ &= (n - 1) \sigma_u^2 \end{aligned}$$

/Luego

Luego $E(B) = (n - 1)\sigma^2 u^2$

Para la tercera expresión:

$$C = (\hat{\beta} - \beta) \sum_{i=1}^n x_i (u_i - \bar{u})$$

$$E(C) = E\left[(\hat{\beta} - \beta) \sum_{i=1}^n x_i (u_i - \bar{u})\right]$$

$$\text{pero } \hat{\beta} - \beta = \frac{\sum u_i x_i}{\sum x_i^2}$$

$$E(C) = E\left[\frac{\sum u_i x_i}{\sum x_i^2} (\sum x_i u_i - \bar{u} \sum x_i)\right]$$

ya que: $\sum x_i = 0$

$$E(C) = E\left[\frac{(\sum u_i x_i)^2}{\sum x_i^2}\right]$$

$$= E\left[\frac{(u_1^2 x_1^2 + u_2^2 x_2^2 + \dots + u_n^2 x_n^2 + 2 u_1 x_1 u_2 x_2 + \dots)}{\sum x_i^2}\right]$$

$$= E\left[\sum u_i^2 x_i^2 + 2 \sum_{i \neq j} u_i x_i u_j x_j\right] \frac{1}{\sum x_i^2}$$

$$= \left[\sum x_i^2 E(u_i^2) + 2 \sum_{i \neq j} x_i x_j \underbrace{E(u_i u_j)}_0\right] \frac{1}{\sum x_i^2}$$

$$= \sigma^2 \sum x_i^2 \cdot \frac{1}{\sum x_i^2} = \sigma^2 u^2$$

se tenía que

$$E(\sum e_i^2) = E(A) + E(B) - 2E(C)$$

reemplazando en esta igualdad las expresiones encontradas, se tiene:

$$E(\sum e_i^2) = \sigma^2 u^2 + \sigma^2 u^2 (n - 1) - 2 \sigma^2 u^2$$

$$E(\sum e_i^2) = (n - 2) \sigma^2 u^2$$

/Luego

Luego:

$$E\left(\frac{\sum e_i^2}{n-2}\right) = \sigma_u^2$$

En consecuencia, se dispone de un estimador insesgado de σ_u^2 que es:

$$\hat{\sigma}_u^2 = \frac{\sum e_i^2}{n-2} \quad [XV]$$

Luego, las estimaciones de las varianzas verdaderas: $V(\hat{\alpha})$, $V(\hat{\beta})$ y la covarianza $Cov(\hat{\alpha}, \hat{\beta})$ serán (simbolizando la estimación por medio de un acento circumflejo sobre el operador):

$$\hat{V}(\hat{\alpha}) = \frac{\sum e_i^2 \sum x_i^2}{(n-2)n \sum x_i^2} = \frac{\hat{\sigma}_u^2 \sum x_i^2}{n \sum x_i^2}$$

$$\hat{V}(\hat{\beta}) = \frac{\sum e_i^2}{(n-2) \sum x_i^2} = \frac{\hat{\sigma}_u^2}{\sum x_i^2}$$

$$\hat{Cov}(\hat{\alpha}, \hat{\beta}) = \frac{-\bar{x} \sum e_i^2}{(n-2) \sum x_i^2} = \frac{-\bar{x} \hat{\sigma}_u^2}{\sum x_i^2}$$

La comparación de los errores standard de $\hat{\alpha}$ y $\hat{\beta}$, con los valores estimados de los parámetros $\hat{\alpha}$ y $\hat{\beta}$, ya da una primera idea acerca de la calidad de los estimadores. Un indicador previo resulta de calcular el coeficiente de variabilidad de muestreo de $\hat{\alpha}$ y $\hat{\beta}$, es decir:

$$CV[\hat{\alpha}] = \frac{[\hat{V}(\hat{\alpha})]^{1/2}}{\hat{\alpha}}$$

$$CV[\hat{\beta}] = \frac{[\hat{V}(\hat{\beta})]^{1/2}}{\hat{\beta}}$$

3. Pruebas de hipótesis.

Evidentemente, una vez que se dispone de los estimadores muestrales $\hat{\alpha}$ y $\hat{\beta}$, será necesario establecer pruebas respecto de hipótesis que /sea conveniente

sea conveniente formular de manera de establecer la significación estadística de aquellos valores. Sin embargo, para tratar este punto con alguna rigurosidad, es indispensable plantear previamente algunos temas:

a) Estimadores máximo verosímiles. Dada una cierta función de verosimilitud:

$$V(X_1, X_2, X_3 \dots X_n, \Theta) = f(X_1, \Theta) f(X_2, \Theta) \dots f(X_n, \Theta)$$

$\hat{\Theta}$ será un estimador máximo verosímil de Θ si se encuentra un Θ tal que maximice la función de verosimilitud. Dado que los valores X_i se suponen fijos, la función de verosimilitud, resulta una función que sólo depende de Θ . La función será $V(\Theta)$ que se tratará de maximizar y que estará en función de los X_i considerados como fijos. Por ejemplo, si se tiene la función:

$$f(X, \Theta) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_i - \Theta)^2}$$

La función de verosimilitud será:

$$V = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_1 - \Theta)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_2 - \Theta)^2} \dots \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_n - \Theta)^2}$$

$$V = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_i - \Theta)^2}$$

$$V = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2} \sum (X_i - \Theta)^2}$$

Para encontrar el máximo de V , se derivará parcialmente la función respecto de Θ , lo que obviamente implicará una maximización y no

/un mínimo

un mínimo, dada la forma de la función planteada.

$$\frac{\partial V}{\partial \theta} = \left(\frac{1}{\sqrt{2}}\right)^n e^{-\frac{1}{2} \sum (x_i - \theta)^2} (\sum x_i - \theta) \cdot L e = 0$$

Puede observarse que la expresión:

$$e^{-\frac{1}{2} \sum (x_i - \theta)^2}$$

no puede ser nula, ya que e es una constante positiva, al igual que los otros términos excepto:

$$\sum (x_i - \theta) = 0$$

$$\theta = \frac{\sum x_i}{n}$$

Si se piensa en una función del tipo:

$$f(x, \theta) = \frac{e^{-\theta} \theta^x}{x!}$$

El estimador por máximo de verosimilitud de θ , se encontrará derivando e igualando a cero la función máximo verosímil.

$$V = \frac{e^{-n\theta} \theta^{\sum x_i}}{\prod_{i=1}^n x_i!} \quad \text{donde } \prod \text{ es operador productoria.}$$

$$\frac{\partial V}{\partial \theta} = \frac{1}{\prod x_i!} \left[e^{-n\theta} (-n) \theta^{\sum x_i} + e^{-n\theta} x_i \theta^{\sum (x_i) - 1} \right] = 0$$

$$\theta^{\sum x_i - 1} e^{-n\theta} \left[-\theta n + \sum x_i \right] = 0$$

no puede ser cero

$$\text{Luego: } \theta = \frac{\sum x_i}{n} = \bar{x}$$

/con lo

con lo que llega a la conocida función de probabilidad discreta de Poisson.

b) Estimadores máximo verosímiles en el modelo de regresión rectilíneo.

Sobre la variable estocástica u_i se habían hecho los supuestos de independencia, media nula y varianza constante σ_u^2 . Si se admite el ulterior supuesto de normalidad en distribución de frecuencias, pueden deducirse estimadores máximo verosímiles de los coeficientes de regresión. La función de verosimilitud será:

$$V = \frac{1}{\sigma_u^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma_u^2} \sum_{i=1}^n u_i^2}$$

Recordando que:

$$Y_i = \alpha + \beta X_i + u_i \quad i = 1, 2 \dots n$$

$$u_i = Y_i - \alpha - \beta X_i$$

La función de verosimilitud para la muestra estará dada por:

$$V = \frac{1}{(\sigma_u^2 2\pi)^{n/2}} e^{-\frac{1}{2\sigma_u^2} \sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2}$$

Aplicando logaritmos (base e), se tiene:

$$\log V = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma_u^2 - \frac{1}{2\sigma_u^2} \sum_{i=1}^n (Y_i - \alpha - \beta X_i)^2$$

Derivando parcialmente respecto de α y β se tiene:

$$\begin{aligned}\frac{\partial \log V}{\partial \alpha} &= -\frac{1}{2\sigma_u^2} \sum (Y_i - \alpha - \beta X_i)(-1) \\ &= \frac{1}{\sigma_u^2} \sum (Y_i - \alpha - \beta X_i) = 0 \\ \sum Y_i &= n\alpha + \beta \sum X_i\end{aligned}$$

$$\begin{aligned}\frac{\partial \log V}{\partial \beta} &= \frac{1}{2\sigma_u^2} \sum (Y_i - \alpha - \beta X_i)(-X_i) \\ &= \frac{1}{\sigma_u^2} \sum X_i(Y_i - \alpha - \beta X_i) = 0 \\ \sum Y_i X_i &= \alpha \sum X_i + \beta \sum X_i^2\end{aligned}$$

Obsérvese que los resultados coinciden con las ecuaciones normales del método de los mínimos cuadrados, pero además son estimadores máximo verosímiles bajo los supuestos planteados; se les designará por $\hat{\alpha}$ y $\hat{\beta}$.

Derivando la función de verosimilitud respecto del otro parámetro σ^2 , se tiene:

$$\begin{aligned}\frac{\partial \log V}{\partial \sigma_u^2} &= -\frac{n}{2} \frac{2\sigma_u}{\sigma_u^2} - \frac{1}{2} (-2\sigma_u^{-3}) \sum (Y_i - \alpha - \beta X_i)^2 \\ &= -\frac{n}{\sigma_u} + \frac{1}{\sigma_u^3} \sum (Y_i - \alpha - \beta X_i)^2 = 0\end{aligned}$$

$$\sigma_u^2 = \frac{1}{n} \sum (Y_i - \alpha - \beta X_i)^2$$

Una estimación de σ_u^2 podrá obtenerse utilizando $\hat{\alpha}$ y $\hat{\beta}$ como estimadores máximo verosímiles de α y β .

$$\bar{\sigma}_u^2 = \frac{1}{n} \sum (Y_i - \bar{\alpha} - \bar{\beta} X_i)^2$$

c) Intervalos de confianza para α y β . Si se piensa que se extraen una cantidad de muestras distinta, se concluirá que se generarán igual número de valores $\hat{\alpha}$ y $\hat{\beta}$. Aceptando el supuesto que la distribución de probabilidad de ambos estadígrafos es normal alrededor de sus respectivas esperanzas matemáticas α y β , ya que:

$$E(\hat{\alpha}) = \alpha$$

$$E(\hat{\beta}) = \beta$$

es posible establecer una relación de distribuciones que será útil al análisis posterior.

Además, es necesario suponer una distribución para el estadígrafo $n \bar{\sigma} u^{-2} / \sigma u^2$. Evidentemente la distribución X^2 (ji dos) es la que mejor representa el comportamiento de esa variable, considerando $n - 2$ grados de libertad. Dado lo anterior y recordando que la distribución "t" de Student relaciona una distribución normal con una distribución X^2 siempre que ambas sean independientes de la siguiente manera:

$$t = \frac{u \sqrt{r}}{v}$$

donde u se distribuye normalmente $(N, 01)$ y v^2 en forma de una X^2 con r grados de libertad. El estadígrafo t tendrá una distribución de Student con r grados de libertad. Esta distribución de Student está dada por:

$$f(t) = c \left(1 + \frac{t^2}{r}\right)^{-\frac{(r+1)}{2}}$$

donde c es una constante apropiada para que $f(t)$ sea una función de densidad (área unitaria). Se trata, como se recordará, de una distribución estandarizada con media nula y tiende a igualarse a una distribución normal a medida que crece el número de grados de libertad.

/Para establecer

Para establecer, entonces, intervalos de confianza para α , se asociará la variable u al estadígrafo:

$$u = \frac{(\hat{\alpha} - \alpha)}{[V(\hat{\alpha})]^{1/2}} = \frac{\hat{\alpha} - \alpha}{\left[\frac{\sigma_u^2 \sum x_i^2}{n \sum x_i^2} \right]^{1/2}} = \frac{(\hat{\alpha} - \alpha) \sqrt{n \sum x_i^2}}{\sigma_u \sqrt{\sum x_i^2}}$$

donde u tiene distribución normal estandarizada.

Por otra parte,

$$v^2 = \frac{\sum e_i^2}{\sigma_u^2} = \frac{(n-2) \hat{\sigma}_u^2}{\sigma_u^2}$$

$$v = \frac{(\sqrt{n-2}) \hat{\sigma}_u}{\sigma_u} \therefore \frac{v}{\sqrt{n-2}} = \frac{\hat{\sigma}_u}{\sigma_u}$$

Recordando que:

$$t = \frac{u}{v/\sqrt{r}}$$

$$r = n - 2$$

se tiene:

$$t = \frac{\frac{(\hat{\alpha} - \alpha) \sqrt{n \sum x_i^2}}{\sigma_u \sqrt{\sum x_i^2}}}{\frac{\hat{\sigma}_u}{\sigma_u}} = \frac{(\hat{\alpha} - \alpha) \sqrt{n \sum x_i^2}}{\hat{\sigma}_u \sqrt{\sum x_i^2}}$$

Con este último resultado, es posible obtener intervalos de confianza para α y, por lo tanto, establecer regiones críticas y de aceptación para el parámetro en cuestión.

$$\text{Prob} (t_{p/2} \leq t \leq t_{1-p/2}) = 1 - p$$

siendo α el nivel de significación que se desea tolerar (probabilidad de error). Por lo tanto, $1 - p$ será el nivel de confianza o probabilidad de acierto. Reemplazando el valor de t , se tiene:

/Ecuación

$$\text{Prob} \left[t_{p/2} \leq \frac{(\hat{\alpha} - \alpha) \sqrt{n \sum x_i^2}}{\hat{\sigma}_u \sqrt{\sum x_i^2}} \leq t_{1-p/2} \right] = 1 - p$$

$$\text{Prob} \left[\frac{t_{p/2} \hat{\sigma}_u \sqrt{\sum x_i^2}}{\sqrt{n \sum x_i^2}} \leq \hat{\alpha} - \alpha \leq t_{1-p/2} \frac{\hat{\sigma}_u \sqrt{\sum x_i^2}}{\sqrt{n \sum x_i^2}} \right] = 1 - p$$

$$\text{Prob} \left[\hat{\alpha} - t_{p/2} \frac{\hat{\sigma}_u \sqrt{\sum x_i^2}}{\sqrt{n \sum x_i^2}} \leq \alpha \leq \hat{\alpha} + t_{1-p/2} \frac{\hat{\sigma}_u \sqrt{\sum x_i^2}}{\sqrt{n \sum x_i^2}} \right] = 1 - p$$

Una vez fijado el nivel de significación p , quedan determinados los valores de $t_{p/2}$ y $t_{1-p/2}$. Al nivel de confianza de $1 - p$, el intervalo de estimación α queda:

$$\hat{\alpha} \pm t_{p/2} \frac{\hat{\sigma}_u \sqrt{\sum x_i^2}}{\sqrt{n \sum x_i^2}}$$

ya que $t_{p/2} = -t_{1-p/2}$

Además, puede establecerse la región de aceptación de una cierta hipótesis al determinar el nivel de significación p . Con este valor y el tamaño de muestra restado en 2 unidades ($n - 2$) que constituyen los grados de libertad, pueden encontrarse los valores de $t_{p/2}$ y $t_{1-p/2}$, si se trata de una hipótesis de igualdad.

Corresponderá verificar si el estadígrafo

$$t = \frac{(\hat{\alpha} - \alpha) \sqrt{n \sum x_i^2}}{\hat{\sigma}_u \sqrt{\sum x_i^2}}$$

está o no comprendido en la región de aceptación, para probar la validez de la hipótesis planteada y concluir al respecto.

De manera similar, puede obtenerse intervalos de confianza para el otro coeficiente de regresión β .

Bajo el supuesto que $\hat{\beta}$ tiene una distribución normal estandarizada $N(0,1)$ dada por:

$$u = \frac{\hat{\beta} - \beta}{\sqrt{v(\hat{\beta})}} = \frac{\hat{\beta} - \beta}{\frac{\sigma_u}{\sqrt{\sum x_i^2}}} = \frac{(\hat{\beta} - \beta) \sqrt{\sum x_i^2}}{\sigma_u}$$

$$v^2 = \frac{\sum e_i^2}{\sigma_u} = \frac{(n-2) \hat{\sigma}_u^2}{\sigma_u^2}$$

Tendrá una distribución "Ji dos" con $n - 2$ grados de libertad.

$$\frac{v}{\sqrt{n-2}} = \frac{\hat{\sigma}_u}{\sigma_u}$$

El estadígrafo

$$t = \frac{u \sqrt{r}}{v} = \frac{u}{\frac{v}{\sqrt{r}}}$$

$$t = \frac{(\hat{\beta} - \beta) \sqrt{\sum x_i^2}}{\sigma_u} \cdot \frac{\sigma_u}{\hat{\sigma}_u}$$

$$t = \frac{(\hat{\beta} - \beta) \sqrt{\sum x_i^2}}{\hat{\sigma}_u}$$

Siguiendo los mismos pasos del caso anterior, se llega a establecer el intervalo de confianza para β , dado un nivel de significación p .

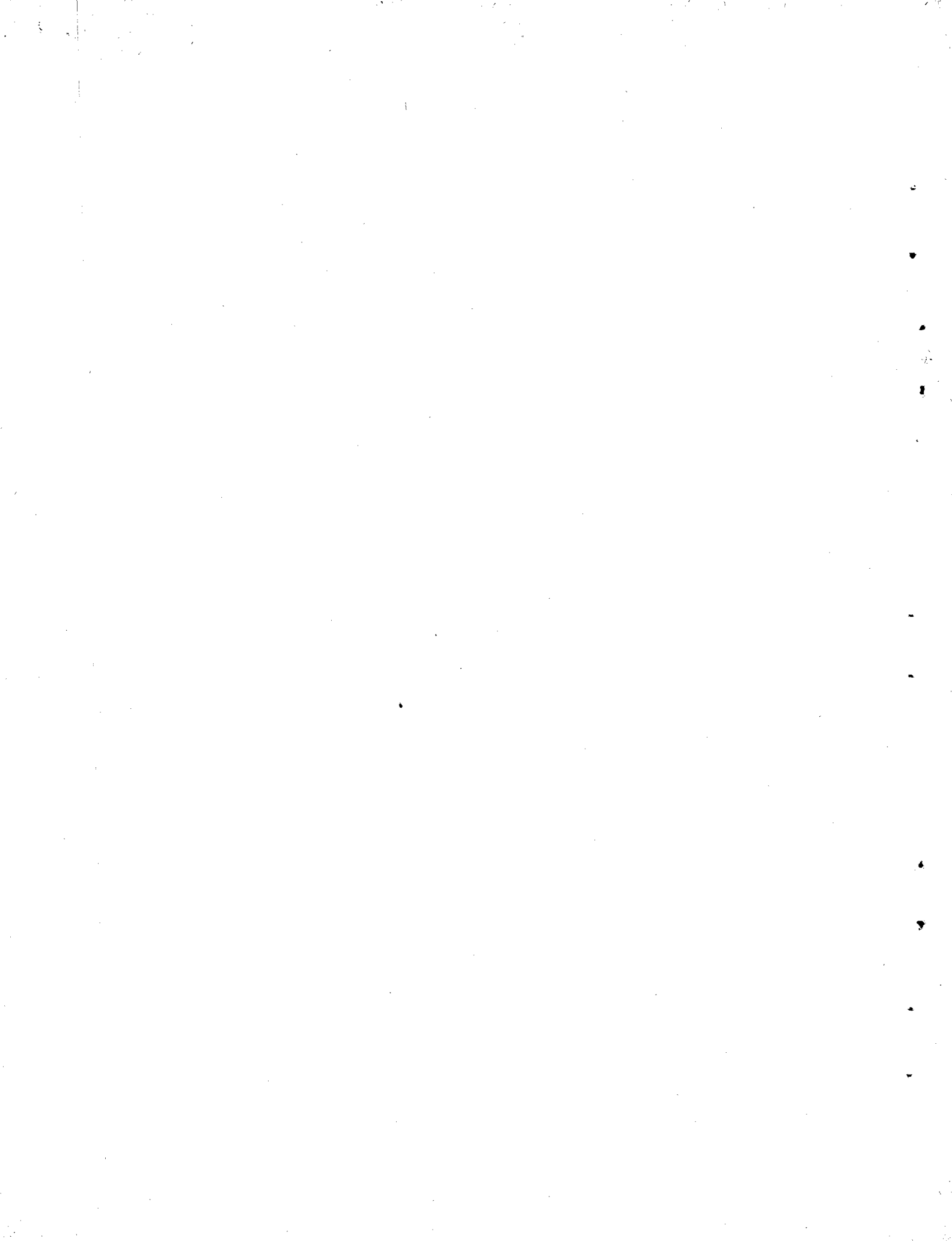
$$\hat{\beta} \pm t_{p/2} \frac{\hat{\sigma}_u}{\sqrt{\sum x_i^2}}$$

Igualmente, con este resultado es posible establecer zonas de aceptación y de rechazo para probar hipótesis.

1

11

11



PRELIMINAR

Instituto Latinoamericano de
Planificación Económica y Social
Santiago, diciembre de 1967

CURSO DE ESTADISTICA BASICA PARA PROGRAMACION*

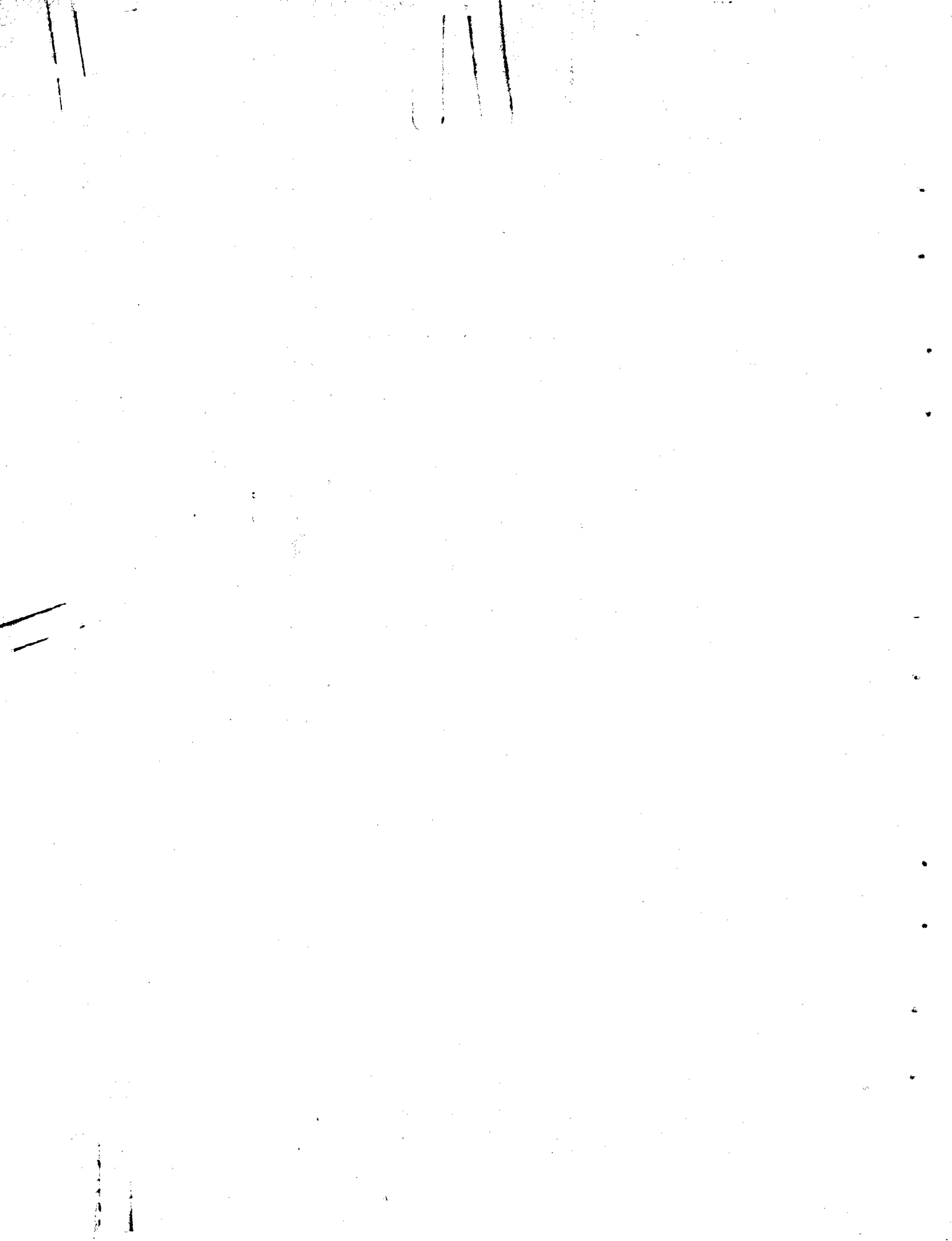
Parte III

* Programa de Capacitación. Profesor, señor Arturo Núñez del Prado B.



CONTENIDO

A.	Introducción	1
B.	Fundamentos Teóricos del Muestreo.	2
C.	Determinación del Tamaño de la Muestra	9
	1. Homogeneidad de la población.	9
	2. Precisión de la estimación.	10
	3. Nivel de confianza.	10
	4. Tamaño de la población.	10
	5. Recursos.	10
D.	La Estimación de Proporciones.	14
E.	Nociones de Muestreo Aleatorio Estratificado	17
	1. Estimadores e intervalos de confianza	18
	2. Afijaciones y tamaños de muestra.	21
	i) Afijación proporcional	21
	ii) Afijación óptima	22
	iii) Afijación óptima económica	24
	iv) Afijación arbitraria	25
F.	Encuestas industriales	26
	1. Determinación clara y precisa de los objetivos.	26
	2. Delimitaciones del marco muestral	27
	3. Elección del diseño muestral.	29
	4. Cálculo del tamaño de muestra	31
	5. Selección de las unidades muestrales.	34
	6. Entrenamiento de enumeradores	35
	7. Colaboración de la población informante	35
	8. Organización del trabajo en el terreno.	36
	9. Sistematización de datos.	37
	10. Determinación del costo total	37
	11. Publicación de resultados	38
	DEMOSTRACIONES MATEMATICAS DEL MUESTREO ALEATORIO SIMPLE.	39
	FORMULARIOS	
	Formulario sobre Muestreo Aleatorio Simple.	47
	A. Variable cuantitativa.	47
	B. Variable cualitativa	49
	Formulario sobre Muestreo Aleatorio Estratificado	51
	A. Variables cuantitativas.	51
	B. Variable cualitativa	54



V. ELEMENTOS DE MUESTREO

A. Introducción

Cuando se acepta la idea de planificar, implícitamente se admite la necesidad de contar con más, mejores, y periódicas informaciones básicas. En el proceso de planificación, desde el diagnóstico hasta la reformulación de los planes, es indispensable contar con magnitudes e indicadores que muestren lo que en verdad está ocurriendo en una actividad económica. Ahora bien, hay varias alternativas de captación de información: los censos, el análisis de casos típicos y las muestras. Al considerar el tipo y fidelidad de las informaciones necesarias, el costo de obtenerlas y el tiempo que demandará su recolección, surgen nítidas las ventajas del muestreo. No debe interpretarse con esto, que el muestreo sea un sustituto del censo. Informaciones básicas para el total de una población o universo siempre serán necesarias; de otra manera no será posible diseñar una muestra con reales ventajas y suficiente garantía. Lo que ocurre es que mediante una muestra pueden obtenerse informaciones que resultarían poco menos que imposibles con una enumeración total de la población: justamente por problemas de costo y tiempo. Por otra parte los censos se realizan cada 11 o más años y las muestras pueden ser utilizadas para estimar parámetros en períodos intermedios. Conocer el tamaño de una población y otras características básicas es indispensable para extraer una buena muestra representativa. Se concluye pues, que censo y muestra son dos métodos complementarios y no excluyentes en el proceso de captación de información.

El objeto de la primera parte de este trabajo es mostrar, a los profesionales planificadores, las posibilidades y limitaciones de esta técnica por una parte, y plantear los principales conceptos envueltos en el muestreo. La segunda parte está destinada a analizar los principales problemas que se presentan en las encuestas industriales y señalar las posibles alternativas de solución.

A esta altura es indispensable advertir que existe una gran profusión bibliográfica sobre el tema, pero es difícil encontrar un trabajo que deje de lado todo el tratamiento matemático y se centre en la parte práctica. El planificador debe conocer los elementos básicos del muestreo, sus /limitaciones y

limitaciones y alcances, de modo que pueda determinar qué tipo de investigaciones son susceptibles de enfrentarse por muestreo y pueda decidir con base fundada sobre las alternativas que le ofrezca el muestrista o especialista en muestreo. El experto en muestreo puede presentar alternativas para una investigación dada, que fácilmente pueden implicar costos que van desde los 5 o 10 mil dólares hasta los 50 mil o más, dependiendo el grado de precisión de los niveles de confianza o probabilidad de acierto, del tipo de diseño muestral, etc. Sobre esas alternativas el planificador debe estar en condiciones de tomar una decisión racional.

Se intentará cumplir con los objetivos señalados en líneas anteriores, prescindiendo hasta donde sea posible del aparato matemático. El trabajo está diseñado para ser interpretado contando con conocimientos de estadística descriptiva, elementos de probabilidades y álgebra superior.

En los anexos finales se presentan las principales demostraciones matemáticas y formularios completos de los principales diseños muestrales.

Se piensa que con estos antecedentes, el planificador industrial estará en condiciones de interpretar en su justa dimensión las reales posibilidades de las técnicas muestrales en general, de calificar las estimaciones resultantes y de tener una posición realista en los problemas en que deberá tomar decisiones.

B. Fundamentos Teóricos del Muestreo^{1/}

Para presentar los elementos básicos del muestreo, se tomará como punto de partida el diseño muestral aleatorio simple por dos razones principales: se trata de un diseño sencillo, de fácil interpretación, y constituye la base de los diseños muestrales aleatorios.

Por población, universo o marco muestral se entenderá el conjunto de elementos de quienes se extraerá información y sobre quienes se podrá generalizar la conclusión obtenida a partir de la muestra. La población podrá estar formada por personas, empresas, familias, áreas, objetos, etc. El número de elementos que componen la población se designará por N.

^{1/} Se utilizará la simbología de W.G. Cochran, para facilitar al lector interesado la consulta del Libro "Técnicas de Muestreo" del autor citado.

Por muestra se entenderá una parte representativa de esa población y se designará por n.

Será necesario definir la variable que se investigará: ingresos, valor agregado, productividad, etc., y las unidades en que se expresará.

Por y_i , se denominará el valor de la variable de la i-ésima unidad de la población o de la muestra ($i = 1, 2, 3 \dots n \dots N$).

El número posible de muestras de composición distinta que se podrá obtener, estará dado por el número combinatorio C_n^N . Se insiste que se trata de muestras de composición distinta, en cuanto a sus elementos, aunque muchas de estas muestras pueden tener la misma media aritmética.

Los estadígrafos que interesa definir son los siguientes:

\bar{Y} : media aritmética de la población

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N}$$

\bar{y} : media aritmética de la muestra

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

σ^2 : varianza de la población

$$\sigma^2 = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N}$$

S^2 : varianza de Cochran de la población

$$S^2 = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}$$

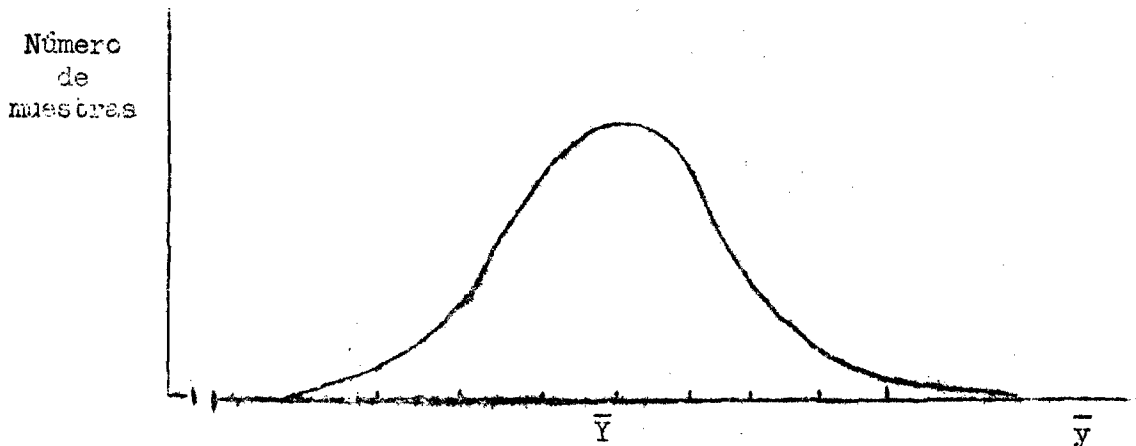
$/s^2 =$

s^2 : varianza de Cochran de la muestra

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

Ahora bien, si el número de muestras distintas que se puede obtener es $\binom{N}{n}$, habrá igual número de medias aritméticas muestrales: \bar{y}_h , donde $h = 1, 2, 3 \dots \binom{N}{n}$.

Si se extraen todas esas muestras computándose sus medias aritméticas y clasificándolas en una tabla de distribución de frecuencias, se comprobará que en general, estas medias aritméticas tienen distribución muy aproximada a la distribución normal. Se recalca en general, porque esto no se cumple cuando la muestra es muy pequeña (menos de 30 - 40 elementos)^{1/}, o la población es muy reducida o con características poco frecuentes. En el siguiente gráfico se ilustra este comentario:



Si se promedian todas estas medias aritméticas, (E: esperanza matemática) se encuentra que el promedio es exactamente igual a la media aritmética poblacional, es decir: (demostración en el apéndice)

$$E(\bar{y}_h) = \frac{\sum_{h=1}^{\binom{N}{n}} \bar{y}_h}{\binom{N}{n}} = \bar{y}$$

^{1/} El desconocimiento de la varianza poblacional (S^2) es otro aspecto a considerar. Para investigaciones en el campo industrial, en donde generalmente las muestras superarán los 30 o 40 elementos, no es muy peligroso aceptar esa simplificación teórica. /Por eso

Por eso a la media aritmética muestral (\bar{y}_h) se le llama estimador insesgado de la media aritmética poblacional.

De la misma manera, a todas las posibles muestras se le puede computar una varianza (s^2_h). Si se promedian estas varianzas, el resultado es la varianza de Cochran para la población:

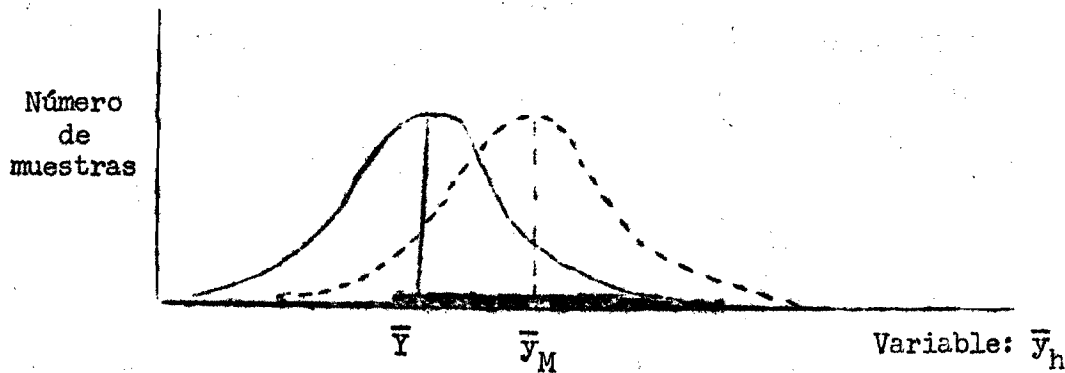
$$E(s^2_h) = \frac{\sum_{h=1}^N s^2_h}{\binom{N}{n}} = S^2$$

Por ello a la varianza de Cochran para la muestra, se le llama estimador insesgado de la varianza de Cochran para la población. Esta igualdad sólo es válida para el caso en que las varianzas hayan sido calculadas con denominadores $N - 1$ (Poblacional) y $n - 1$ (muestral) y esa es la razón para restarle una unidad al denominador de las varianzas.

En la práctica sólo se extrae una muestra y sobre la base de ella se hacen las estimaciones para la población. La selección de esa única muestra se hace, en este diseño muestral, en forma aleatoria y sin reposición, es decir, todas y cada una de las posibles muestras, tiene la misma probabilidad de ser elegida, y cada muestra estará formada por elementos distintos. Por ejemplo, en una encuesta industrial no habrá empresas repetidas en la muestra. Cada empresa aparecerá una y sólo una vez en la muestra. En resumen, este diseño muestral es, pues, equiprobabilístico y sin reposición.

Ahora bien, con el cómputo de la media aritmética de la muestra, se tiene una estimación de la media aritmética poblacional, es decir, una estimación puntual. Es prácticamente imposible que la media aritmética muestral coincida exactamente con la media aritmética poblacional. Es necesario determinar un rango o intervalo a partir de la media aritmética muestral, donde se espera esté comprendida la media aritmética poblacional. A ese intervalo específicamente se le denominará intervalo de confianza y a su mitad se simbolizará por d , que corrientemente se le denomina semi-ancho del intervalo de confianza. Lo anterior puede interpretarse gráficamente de la siguiente manera:

/Gráfico



La curva continua representa la distribución de las medias muestrales en torno a la media poblacional. Si se ha seleccionado una muestra que tiene por media aritmética \bar{y}_M , se traslada la distribución con eje en \bar{y}_M , en el gráfico está representada por la línea punteada. A partir de \bar{y}_M se determina el intervalo de confianza (zona sombreada sobre el eje de las abscisas).

El intervalo de confianza tiene dos componentes:

El coeficiente de confianza que está determinado por la probabilidad de acierto de la estimación, y tiene su origen en la estructura de una distribución normal. En una distribución normal, en el intervalo comprendido entre $\bar{Y} \pm \sigma$ se encuentra el 68 por ciento de los casos, entre $\bar{Y} \pm 2\sigma$ se encuentra el 96 por ciento de los casos.^{1/} El coeficiente de confianza es el número de veces que se debe tomar σ a derecha e izquierda de \bar{Y} para agrupar, entre medio, cierto porcentaje de casos. Asociado al concepto de coeficiente de confianza está el nivel de confianza o probabilidad de acierto (porcentaje de casos comprendidos en cierto intervalo central). A continuación se dan los coeficientes de confianza más frecuentemente utilizados en una distribución normal, los que se simbolizarán por "t".

Nivel de confianza	98%	95%	90%	80%	68%
Coef. de confianza (t)	2,33	1,96	1,65	1,28	1,0

El otro componente del intervalo es el que representa la dispersión

^{1/} Estas proporciones deben interpretarse respecto del número posible de muestras a obtener. $\binom{N}{n}$

de la variable. En una distribución normal corriente, por ejemplo la distribución de una población, ese componente está representado por σ . La distribución normal que ahora se está tratando es bastante particular: en primer lugar la variable está dada por medias aritméticas que provienen de muestras, en segundo lugar se trata de una distribución teórica ya que se suponen computadas todas las posibles medias aritméticas. El estadígrafo que indica la dispersión de esta particular variable estará dado por:

$$v(\bar{y}_h) = \frac{\sum_{h=1}^{(N)} (\bar{y}_h - \bar{Y})^2}{(N)}$$

Nótese que se trata de la fórmula de una varianza cualquiera, pero donde la variable está dada por todas las posibles medias aritméticas muestrales.

Se puede, matemáticamente demostrar que

$$v(\bar{y}) = \frac{\sum_{h=1}^N (\bar{y}_h - \bar{Y})^2}{(N)} = \frac{S^2}{n} \left(1 - \frac{n}{N}\right)$$

que constituye la fórmula de cálculo práctico de la varianza verdadera del estimador de la media aritmética. Este estadígrafo se conoce también con el nombre de cuadrado del "error standard verdadero". En general, cuando se realizan estimaciones sobre media aritmética, no se dispone del valor de la varianza poblacional de Cochran. En esos casos se utiliza la varianza de la muestra s^2 , en virtud de que es un estimador insesgado de S^2 . Cuando ocurre tal cosa, es decir cuando se utiliza s^2 en vez de S^2 , el estadígrafo toma el nombre de varianza estimada del estimador de la media aritmética, y se simboliza así:

$$v(\bar{y}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right)$$

a) Intervalos de confianza:

Con los elementos mencionados, ya se puede cuantificar el intervalo /de confianza

de confianza para la media aritmética. Este intervalo estará dado por:

$$\bar{y} \pm d$$

donde d es la magnitud del semi-ancho del intervalo de confianza y está dado por:

$$d = t \sqrt{v(\bar{y})}$$

Como se recordará, t es el coeficiente de confianza que corresponde a un nivel de confianza. En general el nivel de confianza fluctúa entre 80 por ciento y 95 por ciento, en la práctica generalmente se toma un 90 o 95 por ciento de confianza, lo que en otros términos indicaría que en 95 de cada 100 muestras la media aritmética poblacional estará dentro del intervalo de confianza. $V(\bar{y})$ indicaba la dispersión de las medias muestrales en torno a la media poblacional. Con los elementos vistos ya se está en condiciones de estimar una media aritmética poblacional: Ejemplo: se ha tomado una muestra de 30 de las 108 empresas pesqueras de un país. Se pretende estimar el número promedio de obreros por empresa. Para las 30 empresas de muestra se calculó una media aritmética de 46 obreros por empresa y una varianza de 18. El número promedio de empresas por obrero en toda la población estará comprendido entre:

$$\bar{y} \pm d, \text{ es decir}$$

$$y \pm t \sqrt{v(\bar{y})}$$

Los datos que se conocen son:

$$N = 108; \quad n = 30; \quad \bar{y} = 46; \quad s^2 = 18.$$

$$v(\bar{y}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right) = \frac{18}{30} \left(1 - \frac{30}{108}\right) = 0,433$$

Si se toma como nivel de confianza un 90 por ciento, el correspondiente valor de "t" será de 1,65. El intervalo de confianza estará dado por:

$$46 \pm 1,65 \sqrt{0,43}$$

lo que equivale a decir que la media aritmética poblacional estará comprendida entre 44,92 y 47,08, con 90 por ciento de probabilidad de que tal proposición sea verdadera.

/Para la

Para la estimación de un total, es decir, de la suma de los valores de la variable de cada uno de los elementos de la población, se procede a multiplicar el intervalo para la media aritmética, por N; en virtud de que:

El total (Y) estará dado por:

$$Y = N \bar{Y}$$

y el total estimado (\hat{Y}):

$$\hat{Y} = N \bar{y}$$

Si en el ejemplo anterior se deseara encontrar el intervalo para el total de obreros ocupados en las 108 empresas, con el mismo nivel de confianza se tendrá:

$$N \bar{y} \pm N t \sqrt{v(\bar{y})}$$

que equivale a:

$$(108)(46) \pm (108)(1,65)(0,656)$$

El total de la población estará comprendido entre: 4851 y 5085, con 90 por ciento de probabilidad de que tal cosa sea cierta.

C. Determinación del tamaño de muestra

Hasta ahora se ha supuesto un tamaño de muestra dado, interesa pues analizar brevemente cuáles son los elementos que determinan la magnitud de n.

Fundamentalmente hay cuatro elementos técnicos que condicionan el tamaño de una muestra. Por otra parte hay un quinto elemento de extraordinaria importancia práctica: el monto de recursos financieros y elementos humanos y materiales sin cuyo concurso no es posible garantizar estimaciones confiables.

1. Homogeneidad de la población:

Es fácil interpretar que el grado de homogeneidad indicado por la magnitud de S^2 , condicionará el tamaño de muestra. Donde la variable en la población se distribuya uniformemente, sólo será necesario unos pocos elementos de muestra para tener una idea bastante precisa de lo que ocurre en la población para la variable que se investiga. En cambio en poblaciones

/muy heterogéneas

muy heterogéneas para la variable investigada, será necesario un tamaño de muestra bastante grande para poder realizar estimaciones sin riesgo de grandes errores. Hay, pues, una relación directa entre n y S^2 .

2. Precisión de la estimación:

Nuevamente el título sugiere algo obvio: mientras más precisa una estimación (menor tamaño de "d"), más grande deberá ser el tamaño de muestra. Se había adelantado que la precisión estaba dada por la magnitud de "d". Es el investigador: el planificador, sociólogo, economista, etc., que investiga el comportamiento de una variable, el que tiene que decidir sobre la magnitud del máximo de desviación respecto de la media aritmética verdadera. Se concluye pues que hay una relación inversa entre tamaño de muestra y precisión.

3. Nivel de confianza.

El nivel de confianza que representaba la probabilidad de que la estimación sea verdadera, también tiene una relación directa con el tamaño de muestra, a través del coeficiente de confianza t . Mientras mayor probabilidad de acierto se desea tener, más grande deberá ser el tamaño de muestra.

4. Tamaño de la población.

Otro elemento que es indispensable analizar es la relación que existe entre los tamaños de muestra y población. De poblaciones numerosas, cabrá esperar muestras grandes y vice-versa.

Del equilibrio de todas estas condicionantes, se determina la magnitud del tamaño de una muestra, como se verá en páginas posteriores.

5. Recursos.

Este elemento, si bien no entra dentro de la determinación "técnica" del tamaño de muestra, juega, en nuestros países, un papel de primera línea. Existe toda una problemática entre la compatibilización del tamaño de muestra determinado técnicamente, y el tamaño de muestra que resultaría del monto de recursos, principalmente financieros, asignados para la investigación.

En general el monto de recursos determina una muestra menor que la que resultaría de la aplicación de las fórmulas. La muestra menor

/determinará una

determinará una menor precisión; será necesario analizar si esa precisión resultante garantiza la obtención de estimaciones adecuadas. Es en este punto donde hay que decidir: o se renuncia a la precisión especificada técnicamente y se acepta la resultante de la limitación de recursos, o se destinan mayores recursos financieros. Las dos alternativas restantes son: una combinación de los dos procesos anteriores, o la decisión de no llevar adelante la investigación por estos métodos.^{1/}

Las fórmulas del tamaño de muestra se deducen fácilmente a partir del intervalo de confianza:

Recuérdese la fórmula del intervalo de confianza para la media aritmética:

$$d = t \frac{S}{\sqrt{n}} (1 - \frac{n}{N})^{1/2}$$

Después de elevar al cuadrado, se despeja "n" y se tiene:

$$n = \frac{(\frac{t S}{d})^2}{1 + (\frac{t S}{d})^2 \cdot \frac{1}{N}}$$

Esta es la fórmula del tamaño de una muestra aleatoria simple, que garantiza la estimación de una media aritmética con los requisitos de precisión y confianza o probabilidad de acierto especificados.

Es necesario prestar atención a estos conceptos mencionados ; t, como se había adelantado es el coeficiente de confianza que provenía de la distribución normal estandarizada (media = 0 y varianza = 1) y estaba automáticamente determinado al elegir el nivel de confianza. Una interpretación para este concepto en este caso, sería: la proporción dentro de todas las muestras posibles, que entregan resultados para la media aritmética, que no difieran respecto de la media aritmética verdadera, en más del valor "d" especificado.

^{1/} Ver: "El costo en las investigaciones muestrales", artículo del autor, en la revista Economía N° 83, Facultad de Ciencias Económicas de la U. de Chile.

En cuanto a la precisión representada por la magnitud de "d", se había adelantado que quien está en mejores condiciones para decidir sobre el desvío máximo que se puede tolerar, es el investigador, en este caso el planificador, porque sabe cuáles serán los fines de la estimación. No está de más señalar que una alta precisión (d pequeño) sólo se podrá conseguir a expensas de una muestra grande, y que muestras pequeñas sólo pueden entregar resultados poco precisos. Los anteriores juicios son válidos, cuando el grado de heterogeneidad en la población (magnitud de S^2) es apreciable. En poblaciones homogéneas, una muestra pequeña puede ser suficiente para lograr una aceptable precisión. Todo lo anterior exige una decisión del planificador, que para este fin deberá agregar dos elementos importantes: el costo de la investigación y el tiempo que demandará.

El muestrista se limitará a presentarle alternativas, el planificador deberá elegir la más conveniente.

Observando la fórmula de tamaño de muestra, se puede visualizar que cuando la población es bastante grande, el denominador tiende a la unidad, y el numerador puede aceptarse como una adecuada aproximación del tamaño de muestra.

$$n_0 = \left(\frac{t S}{d}\right)^2$$

El lector debe haberse planteado una interrogante: en las fórmulas para tamaño de muestra se supone conocida la varianza. ¿Qué sentido tiene determinar una muestra para estimar una media aritmética si se supone conocida la varianza y esto implica conocer la media aritmética? Lo que sucede es que el supuesto no es hipotético y la conclusión no carece de sentido. Hay muchos estudios que exigen investigaciones muestrales periódicas. Con estas investigaciones se pueden tener buenas estimaciones de S^2 y porque en muchos casos la varianza puede ser poco cambiante, lo que no siempre ocurre con la media (recuérdese que: $V(y_i + K) = V(y_i)$ y que $M(y_i + K) = K + M(y_i)$).

Bueno, pero el caso más frecuente es que se desconozca el valor de S^2 . El muestrista dispone de métodos que le permiten estimar la magnitud de S^2 ,

/en este

en este caso el planificador deberá tomar nota de que se trata de una estimación que puede constituir una fuente de posibles errores. Es necesario que una vez realizada la muestra, se compulse el valor estimado de S^2 y la varianza de la muestra. Si el valor estimado de S^2 es menor que la varianza de la muestra (s^2), no hay motivo de preocupación, excepto que se ha tomado tal vez una mayor muestra que la necesaria y esto puede estar compensado en parte por la mayor precisión que se consigue. Pero si la varianza de la muestra es apreciablemente mayor que el valor estimado para fines de cálculo del tamaño de muestra, la situación deberá ser motivo de mayor investigación. Habrá que recalcular el tamaño de muestra en función de s^2 . En esta etapa será necesaria otra decisión: se toma una muestra complementaria o se renuncia a la precisión especificada primitivamente.

Por último, en muchas investigaciones resulta que los datos censales disponibles escasamente proporcionan un valor de la magnitud de la población. En estos casos, a veces como primera aproximación, se utilizan varianzas de la variable que se investiga provenientes de otras poblaciones, otros países u otras áreas dentro del mismo país. Nótese que tal procedimiento puede contener serios errores y deberá ser analizado concienzudamente. Además es necesario recalcar que estas suposiciones son con el sólo objeto de tener una primera aproximación del tamaño de la muestra. Posteriormente, si se realiza la encuesta, deberá compararse las magnitudes de las varianzas y proceder en la forma indicada en páginas anteriores.

Quando no es posible tener estimaciones previas de S^2 , a veces se opta por una muestra piloto o de iluminación, cuyo tamaño generalmente está condicionado por el costo que ésta involucra. Dicha muestra de iluminación puede dar una idea del posible tamaño definitivo de muestra. Con todos los conceptos que se han señalado se pretende sistematizar en un cuadro las alternativas que se le presentarán al planificador para que decida sobre el tamaño de muestra definitivo, a la luz de los objetivos que tiene planteados. Sólo se incluyen los principales elementos de decisión, así, se omite presentar alternativas en cuanto a nivel de confianza porque éste generalmente tiene poca variación (entre 90 y 95 por ciento), y su omisión no distorsionaría la decisión.

TAMAÑO DE MUESTRA	DESVIO MAXIMO	COSTO (Miles de dólares)	TIEMPO (días)
150	0,10	4 - 5	30 - 35
600	0,05	8 - 9	60 - 70
938	0,04	9 - 10	70 - 85
1666	0,03	15 - 17	120 - 140
3750	0,02	22 - 25	210 - 240
6000 ^{1/}	0,00	50 - 56	400 - 500

^{1/} Censo

El cuadro anterior puede corresponder por ejemplo a alternativas de tamaño de muestra para una investigación en el sector industrial, donde la variable estratégica sea la estimación del coeficiente producto capital. Se ha supuesto para fines de ilustrar el ejemplo, que la varianza de los coeficientes producto-capital es 0,375, y en todos los casos se ha considerado una probabilidad de acierto de 96 por ciento. He aquí un caso de ocurrencia periódica en la práctica.

¿Puede un investigador decidir racionalmente, seleccionando la mejor alternativa sin contar con ideas fundamentales de los conceptos que entran en juego? Evidentemente que decisiones realistas y racionales sólo provienen de investigadores que conocen los elementos básicos de la teoría de muestras.

D. La estimación de proporciones

Dentro de los diagnósticos y en la preparación de los programas hay cierto tipo de variables cuyo tratamiento puede ser indispensable. Es el caso de las variables cualitativas. Por ejemplo la estimación de la proporción de empresas que son sociedades anónimas, la proporción de empresas que cuentan con más de 100 obreros, el número de obreros que tienen salarios superiores al promedio del país, etc. El lector puede imaginarse una serie de casos donde la estimación de la proporción puede ser útil. No es otra cosa que un caso particular del diseño anterior; aquí la variable puede estar clasificada en dos categorías: posee la característica que se investiga,

/o no

como la posee. Por convención se asigna el valor 1 a aquellas unidades en la población que tienen la característica y 0 si no la tienen. Con esa convención, la proporción en la población será:

$$P = \frac{\sum_{i=1}^N y_i}{N}$$

La proporción de los elementos que no poseen la característica que se investiga será:

$$Q = 1 - P$$

En la muestra este estadígrafo estará dado por:

$$p = \frac{\sum_{i=1}^n y_i}{n} \quad \text{y} \quad q = 1 - p$$

Si se observa las fórmulas en ambos casos, corresponden a las conocidas fórmulas de una media aritmética. En este caso también las proporciones muestrales se distribuyen en general en forma aproximadamente normal, en torno a la proporción poblacional. Donde hay que poner un poco de atención es en la varianza de la proporción. Como se verá en seguida, en este caso la varianza tiene dos límites: sólo puede tomar valores comprendidos entre 0 y $0.25^{1/}$. Cuando la variable es cuantitativa, la varianza sólo tenía límite inferior: no podía ser negativa, pero podía formar un valor tan grande como se quiera.

Es fácil deducir la fórmula de la varianza de la proporción, partiendo de la fórmula conocida para la varianza de Cochran.

$$s^2 = \frac{\sum_{i=1}^N (y_i - Y)^2}{N - 1}$$

Desarrollando el cuadrado y reduciendo términos semejantes, se tiene:

1/ Para poblaciones no demasiado pequeñas. La varianza máxima se tiene cuando $P = Q$.

$$/s^2 =$$

$$s^2 = \frac{\sum_{i=1}^N y_i^2}{N-1} - \frac{N \bar{Y}^2}{N-1}$$

Se trata de reemplazar los valores de esta expresión, por los que resultan de considerar la proporción en el caso en que la variable toma valores cero o uno.

Si

$$P = \frac{\sum_{i=1}^N y_i}{N}, \text{ puede asociarse con } \bar{Y}$$

Además:

$$\sum_{i=1}^N y_i = N P \quad y$$

$$\sum_{i=1}^N y_i^2 = \sum_{i=1}^N y_i \quad \text{porque la variable sólo toma valores}$$

cuyos cuadrados son los mismos valores. Reemplazando en la fórmula de Cochran para la varianza se tiene:

$$s^2 = \frac{N P - N P^2}{N - 1} = \frac{N P Q}{N - 1}$$

Por analogía, la varianza de la muestra estará dada por:

$$s^2 = \frac{n p q}{n-1}$$

Para la determinación de los intervalos de confianza, se utilizan las mismas fórmulas, teniendo cuidado de reemplazar el valor de

$$S^2 \text{ por } \frac{N P Q}{N-1} \quad y \quad s^2 \text{ por } \frac{n p q}{n-1}$$

En todo caso, en el anexo formulario de este trabajo se incluyen

/fórmulas detalladas,

fórmulas detalladas, cuya consulta puede ser beneficiosa.

E. Nociones de muestreo aleatorio estratificado

Este diseño muestral no ofrece complicación alguna. Es simplemente la combinación de varios diseños aleatorios simples.

Se trata de estratificar la población, es decir, dividirla en partes excluyentes, de acuerdo a la característica que se investiga. Así, por ejemplo, si se quiere estimar el valor agregado del sector industrial, sería conveniente dividir la población de industrias en 4 o 5 estratos de manera que en cada estrato las unidades tengan características homogéneas en cuanto a valor agregado, es decir, en cada estrato los valores agregados de cada industria deberían estar dentro de cierto rango. De esta manera se minimiza la magnitud de la varianz^{1/}a en cada estrato y, por consiguiente, se trata de disminuir el error de muestreo. Es importante insistir que el criterio de estratificación debería estar en directa relación al objetivo de la encuesta. Por ejemplo si se quiere estimar un coeficiente producto capital, un posible criterio de estratificación podría ser según el número de turnos diarios; un primer estrato comprendería a las industrias que trabajan un turno diario, el segundo estrato las que lo hacen a dos turnos diarios, el tercer estrato las de tres turnos y un último estrato que comprendería a aquellas industrias que no tienen un número de turnos definido.

La principal ventaja de este diseño muestral, es su eficiencia comparada con otros diseños muestrales: para un mismo tamaño de muestra el error de muestreo es en general menor cuando se estratifica la población. La otra ventaja importante es que permite realizar estimaciones no sólo para la población total sino para cada uno de los estratos, lo que sin duda, especialmente en el campo industrial, permite afinar conclusiones. La desventaja es la necesidad de tener un conocimiento anticipado de las características generales del universo que se investigará; de la calidad de las informaciones básicas dependerá la bondad del criterio de estratificación.

^{1/} Recuérdese que los componentes de la varianz^{1/}a son la intervarianza y la intravarianza. Dado que para los errores de muestreo se promedian las varianzas de cada estrato, en el hecho sólo se está tomando en cuenta la intravarianza.

En cuanto al número de estratos, hay que considerar que mientras mayor sea, más precisas serán las estimaciones y una mayor desagregación en los resultados, permitirá obtener conclusiones específicas y detalladas. Por otra parte, es difícil disponer de antecedentes que permitan clasificar adecuadamente (estratificar) la población. En la práctica con los antecedentes que en general se disponen, lo más que puede hacerse, es dividir la población en gruesas categorías. Incluso si se pudiera estratificar en un número grande de estratos, la complejidad administrativa y la organización del trabajo de campo puede significar trabajo adicional considerable.

En las investigaciones industriales, normalmente se toman entre 4 y 7 estratos.

1. Estimadores e intervalos de confianza.

En cada estrato se podrá computar los siguientes estadígrafos:

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h} \quad \text{Media aritmética del estrato } h$$

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1} \quad \text{Varianza de Cochran del estrato } h$$

$$Y_h = N_h \bar{Y}_h \quad \text{Total del estrato } h$$

La media de la población estará dada por

$$\bar{Y} = \frac{\sum_{h=1}^L \bar{Y}_h N_h}{N} \quad \text{donde} \quad \sum_{h=1}^L N_h = N$$

y la varianza de la población

$$S^2 = \frac{\sum_{h=1}^L (\bar{Y}_h - \bar{Y})^2 N_h}{N - 1} + \frac{\sum_{h=1}^L S_h^2 \cdot (N_h - 1)}{N - 1}$$

/Para la

Para la muestra se tendrían los siguientes estadígrafos:

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \quad \text{media aritmética de la muestra del estrato } h$$

$$s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$$

Ahora bien, dado que en cada estrato se extraen muestras aleatorias simples, tanto la media de la muestra (\bar{y}_h) como la varianza, son respectivamente estimadores insesgados de la media aritmética (\bar{Y}_h) y de la varianza (S_h^2) del estrato:

$$E(\bar{y}_h) = \bar{Y}_h$$

$$E(s_h^2) = S_h^2$$

Si ahora se combinan todos los estratos, la media de la muestra será:

$$\bar{y}_n = \frac{\sum_{h=1}^L \bar{y}_h \cdot n_h}{n} \quad \text{donde} \quad \sum_{h=1}^L n_h = n$$

y el estimador insesgado de la media aritmética poblacional será:

$$\bar{y}_{st} = \frac{\sum_{h=1}^L \bar{y}_h \cdot N_h}{N}$$

Es fácil verificar que:

$$E(\bar{y}_{st}) = \bar{Y}$$

1/ En general \bar{y}_n no es estimador insesgado de \bar{Y} , sólo lo es cuando $\frac{n_h}{n} = \frac{N_h}{N}$.

/lo que

lo que justifica que \bar{y}_{st} , sea estimador insesgado de \bar{Y} .

En la misma forma se pueden combinar las varianzas del estimador de la media. Recuérdese que en muestreo aleatorio simple se tenía:

$$V(\bar{y}) = \frac{S^2}{n} \left(1 - \frac{n}{N}\right)$$

Este estadígrafo para un estrato cualquiera será:

$$V(\bar{y}_h) = \frac{S_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

Promediando las $V(\bar{y}_h)$ para todos los estratos se tendrá la varianza del estimador de la media poblacional:

$$V(\bar{y}_{st}) = \frac{\sum_{h=1}^L N_h^2 V(\bar{y}_h)}{N^2} = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{S_h^2}{n_h} \quad 1/$$

Cuando no se dispone de los valores de las varianzas de los estratos (S_h^2), se utilizan los resultantes de las muestras de cada estrato (s_h^2) y se tiene:

$$v(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{s_h^2}{n_h}$$

que es el estimador de la varianza del estimador de la media aritmética poblacional.

En igual forma que en el muestreo aleatorio simple, el intervalo de confianza para la media estará dado por:

$$\bar{y}_{st} \pm d$$

1/ Fórmula general de la varianza del estimador de la media aritmética poblacional.

/donde d

$$\text{donde } d = t \sqrt{V(\bar{y}_{st})}$$

El intervalo de confianza para el total estará dado por

$$N \bar{y}_{st} \pm d$$

$$\text{donde } d = t N \sqrt{V(\bar{y}_{st})}$$

2. Afijaciones y tamaños de muestra.

Es importante considerar lo que en muestreo se denomina afijación; el significado de esta expresión es el de distribución y asignación de la muestra total en cada uno de los estratos. Por afijación, entonces, se entenderá el proceso que permite distribuir un tamaño de muestra dado (n) entre los estratos, de manera de tener una muestra en cada estrato (n_h). Hay distintas maneras de "afijar" una muestra; a continuación se expondrán las principales.

1) Afijación proporcional. Se trata de distribuir una muestra dada de tamaño n , entre los estratos, en forma proporcional al tamaño de cada estrato. El tamaño de muestra en cada estrato estará dado por:

$$n_h = \frac{N_h}{N} \cdot n$$

La justificación de este método radica en el hecho de que de mayores estratos se extraerán mayores tamaños de muestra.

Cuando se sigue una afijación proporcional, la fórmula general de la varianza del estimador $V(\bar{y}_{st})$, puede simplificarse reemplazando n_h por $\frac{N_h}{N} n$ en la forma:

$$V(\bar{y}_{st})_{A.P.} = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2$$

/donde

donde

$$f = \frac{n}{N} \text{ y } W_h = \frac{N_h}{N}$$

De la fórmula anterior puede despejarse n y se tendrá la fórmula del tamaño de muestra cuando se decide previamente seguir un diseño muestral estratificado por afijación proporcional:

$$n = \frac{n_o}{1 + \frac{n_o}{N}} \quad \text{donde } n_o = \frac{\sum_{h=1}^L W_h S_h^2}{V(\bar{y}_{st})}$$

Es necesario aclarar que cuando se trata de determinar un tamaño de muestra, es necesario tener estimaciones de la varianza de la variable que se desea investigar (S_h^2), ya sea por antecedentes de otras encuestas, por comparaciones con otras variables de distribución similar, por muestras de iluminación o por otros métodos estadísticos de estimación. En la fórmula aparece $V(\bar{y}_{st})$ que es el cuadrado del error de muestreo que se puede tolerar. La manera de especificar este valor es a través del intervalo de confianza.

$$d = t \sqrt{V(\bar{y}_{st})} \quad \text{de donde}$$

$$V(\bar{y}_{st}) = \frac{d^2}{t^2}$$

Recuérdese que d es el desvío máximo que se puede tolerar, y está expresado en las mismas unidades de la variable que se investiga; este valor lo determina el planificador. t es el coeficiente de confianza que está dado por el nivel de confianza o probabilidad de acierto en la estimación. El nivel de confianza también es materia de decisión del investigador. Será necesario plantear alternativas para diferentes magnitudes de d . Con un cuadro similar al presentado en el anterior diseño muestral, se puede decidir en forma objetiva.

ii) Afijación óptima. La distribución de una muestra en forma proporcional al tamaño del estrato puede adolecer de ciertos defectos.

/Es posible

Es posible que existan estratos muy grandes pero bastante homogéneos, y al contrario, puede ocurrir que hayan pequeños estratos sumamente heterogéneos, o ambas cosas a la vez. Si en estos casos se sugiere una afijación proporcional, sucedería que de los grandes estratos homogéneos se extraería una muestra más que suficiente con el correspondiente desperdicio de recursos, y en los pequeños estratos altamente heterogéneos se extraería muestras de tamaño insuficiente. Para evitar tales desajustes, la afijación óptima^{1/} distribuye la muestra total (n) entre los estratos tomando simultáneamente el tamaño y el grado de heterogeneidad del estrato. La fórmula para afijar óptimamente una muestra es:

$$n_h = \frac{N_h S_h}{\sum_{h=1}^L N_h S_h} \cdot n \quad \underline{2/}$$

Observando la fórmula se comprueba que la distribución de n entre los estratos es proporcional, simultáneamente, al tamaño y al grado de variabilidad de la variable en el estrato.

Reemplazando este valor de n_h por la expresión recién comentada, en la fórmula general de la varianza del estimador para la media aritmética, se tiene:

$$V(\bar{y}_{st})_{A.O.} = \frac{(\sum_{h=1}^L W_h S_h)^2}{n} - \frac{\sum_{h=1}^L W_h S_h^2}{N}$$

Si no se dispone de los valores de S_h^2 , se podrá introducir el correspondiente s_h^2 de la muestra, con lo que se tendrá el estimador de la varianza del estimador de la media:

^{1/} El calificativo de óptima tiene su origen en el hecho de que mediante esta afijación se consigue el menor error de muestreo.

^{2/} Existe una demostración que justifica esta fórmula, basándose en la minimización de la varianza del estimador.

$$v(\bar{y}_{st})_{A.O.} = \frac{\sum_{h=1}^L W_h s_h^2}{n} - \frac{\sum_{h=1}^L W_h s_h^2}{N}$$

Si de la fórmula de la varianza del estimador se despeja n , se tiene la fórmula del tamaño de muestra cuando se decide aplicar un diseño muestral estratificado por afijación óptima:

$$n = \frac{(\sum_{h=1}^L W_h S_h)^2}{V(\bar{y}_{st}) + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$$

Los elementos que conforman esta fórmula ya han sido tratados. Se insiste que $V(\bar{y}_{st})$ es el cuadrado del error de muestreo que se toleraría. Para especificar su valor, habrá que plantearse un desvío máximo d y una probabilidad de acierto que determinará el valor de t .

$$V(\bar{y}_{st}) = \frac{d^2}{t^2}$$

iii) Afijación óptima económica. Aparte de considerar simultáneamente el tamaño y variabilidad del estrato (N_h y S_h^2), a veces es recomendable introducir un tercer elemento: el costo por unidad encuestada en cada estrato. Sucede que en algunas investigaciones hay diferencias sustanciales en cuanto a las facilidades de acceso a la información por coleccionar y puede ser justificado tomar este elemento que tiene su origen en limitaciones financieras. Cabe aclarar que tal procedimiento sólo se justifica en la medida que su introducción no signifique pérdida de representatividad en las muestras de los estratos. La forma de "afijar" una muestra dada, obedece en este caso a la siguiente relación,

$$n_h = \frac{N_h S_h / \sqrt{C_h}}{\sum N_h S_h / \sqrt{C_h}} \cdot n$$

/donde C_h

donde C_h es el costo de investigar una unidad en el estrato h .^{1/}

iv) Afijación arbitraria. Finalmente, cuando se "decide" distribuir la muestra sobre la base del buen sentido, cuidando de no perder representatividad y tomando supuestos tentativos en cuanto a heterogeneidad de los estratos, es decir, dirigiendo la distribución sobre la base de un conocimiento ilustrado de la población, se dice que la afijación es arbitraria. Para la determinación de los intervalos de confianza se utiliza la fórmula general de la varianza del estimador.

Es necesario advertir que el diseño muestral estratificado va cobrando mayores ventajas, sobre todo en cuanto a eficiencia, a medida que se tiene un mayor conocimiento de la población. De la adecuada elección del criterio de estratificación depende en gran medida la fidelidad y precisión de los resultados. Dado que las encuestas industriales, en un proceso de planificación, deben tener cierta periodicidad, permitirá conseguir cada vez mejores informaciones con lo que se facilitará la tarea iterativa de acercarse, con los estimadores, a valores que estén muy próximos a los que se dan en la realidad.

^{1/} La fórmula de afijación óptima económica tiene su origen en la minimización de la varianza del estimador, sujeta a una restricción lineal del costo.

F. Encuestas industriales

En gran parte de la bibliografía disponible sobre muestreo, aparece en general, un esquema de las etapas que se deben cumplir en una investigación muestral. Pero puede observarse al mismo tiempo que obedecen a circunstancias que precisamente no parecen ser las que se dan en los países subdesarrollados, específicamente los de América Latina, desde el punto de vista de la carencia de informaciones básicas. Si bien es cierto que las etapas son inevitables, cualesquiera sean las circunstancias que se contemplen, no lo es menos que hay necesidad de adaptar y jerarquizar cada punto, en función de las dificultades y posibilidades específicas de la planificación industrial en estos países.

A continuación se detallarán las diferentes etapas y facetas de una investigación del sector industrial, a través de técnicas muestrales. Se recalcará sobre aquellos problemas que la experiencia sobre investigaciones del sector industrial aconseja poner el mayor cuidado, detallando hasta donde sea posible la compatibilización de los conceptos teóricos con las metodologías y sus alternativas prácticas:

1. Determinación clara y precisa de los objetivos

Es de fundamental importancia especificar en la mejor forma posible, cuáles son los fines que persigue la investigación muestral. El diseño de la muestra obedecerá en cada caso a consideraciones distintas. Es diferente diseñar una muestra para averiguar la magnitud que, de la capacidad instalada en el sector industrial se utiliza efectivamente, en comparación a indagar la estructura de insumos importados que tienen las industrias dinámicas, o a investigar la situación de este sector, detectar sus principales escollos y disponer de antecedentes para proyecciones a nivel agregado. Desde otro punto de vista, una misma muestra, no puede ser igualmente eficiente si se trata de una primera investigación sobre relaciones interindustriales, que si se trata de corregir coeficientes técnicos cuando se tienen sospechas de que éstos puedan haber variado. Es necesario tomar conciencia que, según lo que se desea averiguar y la ponderación que la investigación tenga dentro del análisis general, se podrá diseñar en cada caso, una muestra adecuada. No hay pues

/diseños "matrices"

diseños "matrices" en toda su extensión; lo más que puede hacerse en aras de una cierta orientación, es fijar gruesas líneas dentro de las cuales hay una infinidad de alternativas. Más aún, otra condicionante es la situación del sector en cuanto al conocimiento que de él se tiene. En dos encuestas industriales en que se persiguen los mismos objetivos, los diseños muestrales pueden ser muy diferentes si se considera la cantidad, calidad y antigüedad de las informaciones de que se dispongan en uno y otro caso.

Recién, cuando haya completo acuerdo en el equipo investigador sobre lo que se desea averiguar, se podrá iniciar el diseño muestral propiamente tal, siempre que éste sea factible. Recuérdese que el muestreo no es apto para toda averiguación; hay casos donde es saludable renunciar a utilizar este procedimiento, porque es preferible desconocer algo, antes que tener una información apreciablemente errada.

Una clara delimitación de objetivos, permitirá centrar la investigación en los puntos más importantes. Debe dejarse de lado la idea de aprovechar la encuesta para averiguar todo aquello que tiene y "puede llegar" a tener importancia en el futuro; ese frecuente criterio es altamente perjudicial, porque hace de la encuesta un conjunto de preguntas en las que se diluyen los objetivos centrales. Basta analizar los formularios de encuestas a la industria, y en no pocos casos se encontrarán este tipo de defectos. Aquí el planificador debe tomar en consideración los puntos señalados; armonizando, compatibilizando y delimitando los objetivos.

2. Delimitación del marco muestral.

Las estimaciones que proporcione una muestra se generalizan para una población. El campo de donde proviene una muestra, es decir, el marco muestral, exige una muy precisa definición. En el sector industrial, es especialmente peligroso dejar de ser acucioso en esta etapa. Como se sabe es necesario definir lo que se entenderá por unidad industrial. Así, corrientemente se divide este sector en industria manufacturera que puede estar constituida por las empresas industriales que cuentan con cinco o más obreros, e industria artesanal las que no alcanzan a dicha cifra. Pero aquí el problema aún no estará resuelto, porque dentro de las empresas manufactureras

/pueden haber

pueden haber unidades que se dediquen tanto a actividades industriales como a comerciales, agrícolas, etc. Será necesario fijar un criterio que permita clasificarlas nítidamente en una u otra categoría. Es indispensable contar con un rol, directorio o empadronamiento donde estén inscritas todas las unidades sobre las cuales se desea investigar. Normalmente los países disponen de censos industriales que son utilizados como marco de la muestra; sobre este punto es necesario destacar que dichos censos no pueden hacerse anualmente, cuando más se realizan cada cinco y más años y en consecuencia un trabajo adicional que no puede evitarse es la actualización de los directorios. Hay empresas que se forman después del censo, otras que desaparecen y finalmente las que cambian de giro y tamaño dentro de la misma industria; estos cambios deben ser detectados antes de la extracción de la muestra. Es frecuente que en las encuestas industriales se desee tener alguna clasificación, al menos sobre las principales variables, según las ramas o agrupaciones industriales. Sobre lo mencionado surge en algunos casos otro problema: el de la identificación de todas y cada una de las unidades que conforman la población, dentro de dichas categorías.

Un criterio de clasificación podría encontrarse en la escritura jurídica de formación de la empresa, en la que por regla general se destaca el giro o actividad, pero ocurre que en algunas legislaciones sobre el particular, el cambio de giro exige una serie de trámites e impuestos que hacen que los industriales al formar una empresa, aprovechen de especificar una serie de actividades o giros potenciales, para evitarse molestias y erogaciones en un eventual cambio de actividad. Este hecho obliga en muchos casos a prescindir de esa fuente de clasificación y puede no quedar otro camino que una inspección censal. Muchas veces, la misma muestra puede ser utilizada para "ajustar" tentativamente el directorio industrial, en cuanto a clasificación de industrias por giro o tamaño. No está demás señalar que en estos casos puede ser necesaria una muestra complementaria por razones obvias.

Cuando no se dispongan de marcos muestrales, ni siquiera medianamente aceptables como ocurre con la industria artesanal, corrientemente se "estiman" los valores de sus correspondientes variables mediante asociación y ajustes

/correctivos con

correctivos con las pequeñas empresas manufactureras. Es probable que tales estimaciones no estén muy distantes de los valores efectivos. En todo caso, para establecer los coeficientes de ajuste, es necesario averiguar las similitudes y diferencias, cualitativas y cuantitativas, entre la pequeña industria manufacturera y la industria artesanal. No puede prescindirse de un empadronamiento por áreas de este tipo de unidades industriales; lo que ocurrirá, en general, es que sólo será necesaria una pequeña muestra para tener las estimaciones necesarias para realizar los mencionados ajustes, o para cuantificar directamente los estadígrafos que interesen para este sub-sector.

Finalmente, en muchos casos no será posible determinar para la industria artesanal, el marco de la muestra definitivo, y tendrá que recurrirse a marcos muestrales más generales, siempre que tengan una relación plausible con la investigación.

3. Elección del diseño muestral

Dadas las características de los sectores industriales y pensando siempre en función de la recolección de informaciones para la planificación de la industria, no hay mucha posibilidad de elección en cuanto a los diseños muestrales. El muestreo estratificado calza perfectamente bien con los objetivos que se plantean en este tipo de estudios; tanto en lo que concierne a la eficiencia (menor error comparado con otros diseños muestrales para igual tamaño de muestra) como en lo que toca a la posibilidad de tener estimaciones para cada estrato o sub-estrato.

Ahora bien, en cuanto a las informaciones básicas de las que normalmente disponen los países a través de los censos industriales, cabe advertir que prácticamente hay un criterio obligado de estratificación: por tamaño, asimilando éste al número de obreros con que cuenta cada unidad industrial. Vale la pena advertir sobre este punto que no siempre se justifica este tipo de estratificación, a no ser que se piense en términos de factibilidad, es decir, porque puede no ser posible estratificar de otra manera, aunque los conceptos teóricos así lo indiquen. Si se piensa que la finalidad principal que debe cumplir un criterio de estratificación es la de hacer lo más homogéneas posibles

/las unidades

las unidades de cada estrato, de modo de disminuir la magnitud de la varianza, surge inmediatamente la pregunta: ¿Qué tipo de homogeneidad? Debe elegirse una variable estratégica o clave, para agrupar las unidades de la población según los valores que toman en cuanto a dicha variable. Agrupar las unidades según el número de obreros implica admitir que esta variable es representativa, en cuanto a clasificación, de las otras variables que interesan, como el valor agregado, los insumos, la capacidad utilizada, etc. El avance tecnológico produce cierto escepticismo para pronunciarse sobre la validez de ese supuesto. La mecanización puede disminuir la ocupación al mismo tiempo que aumentar considerablemente, por ejemplo el valor agregado. Según uno u otro criterio, una industria con esas características puede quedar clasificada en estratos muy diferentes. Si el objetivo central es estudiar la absorción de recursos principalmente mano de obra, su productividad y composición, el primer criterio sería adecuado, pero si lo que realmente interesa es tener estimaciones sobre el producto industrial, su estructura y dinamismo, parecería preferible utilizar, siempre que fuera factible, el segundo criterio.

En general en una encuesta industrial se plantean una serie de objetivos; de su jerarquización dependerá la elección del criterio adecuado. Si hay más de un objetivo importante, como ocurre con frecuencia, habrá que buscar un criterio que compatibilice ambos, tal vez no se consiga hacerlo en forma ideal, pero al menos podrá lograrse una correspondencia aceptable.

Una vez que la población ha quedado estratificada de acuerdo al criterio elegido, será indispensable establecer nuevas clasificaciones para propósitos de planificar el desenvolvimiento del sector. Dentro de cada estrato podrá requerirse sub-clasificaciones que conformarán sub-estratos, como por ejemplo en industrias tradicionales y dinámicas, según ramas industriales, según ubicación geográfica, muy importante dentro de planes de integración regional, etc. De esta manera se podrá disponer de una serie de clasificaciones cruzadas que permitirán al programador, realizar análisis a distinto nivel de agregación. Los diagnósticos serán más acertados, las metas serán más detalladas y concretas y el control se facilitará muchísimo. La reformulación de los planes se hará sobre la base de un sólido conocimiento de lo ocurrido.

4. Cálculo del tamaño de muestra

Demás está insistir sobre la trascendencia de esta fase. Prácticamente en este punto es donde se conjugan todos los elementos conceptuales vistos en la primera parte. De otro lado, es donde más cuenta respetar los principios teóricos, al confrontarse a una realidad que no se complace con los supuestos simplificadores que debieron hacerse en busca de organicidad y claridad de exposición.

Recuérdese que principalmente, desde un punto de vista puramente técnico, hay tres elementos fundamentales que determinan el tamaño de una muestra: la magnitud de la población, la variabilidad con que se distribuye la característica que se investiga, y la precisión que se desea tener. Así presentado el asunto es extraordinariamente sencillo, pero analizando con un poco más de detenimiento se comprenderá la complejidad que representa. En lo que se refiere a tamaño de la población, el problema radica principalmente en la definición de la unidad a investigarse y en la delimitación del marco muestral, ambos aspectos ya fueron tratados en el punto 2.

El problema crucial dice relación con la variabilidad de la población, es decir con el estadígrafo S^2 . Prescindiendo por un momento de discutir sobre si se cuenta o no con esa información, es interesante discutir sobre cuál será la variable para la que es necesario cuantificar o disponer de una varianza. Revisando cualquier encuesta industrial, se comprobará que se indaga sobre una cantidad grande de variables; pues bien ¿La varianza de cuál de ellos se tomará? Habrán algunas variables cuya distribución es muy homogénea, en tanto que habrán otras que presentan enorme heterogeneidad. En rigor, habrá que elegir sólo una varianza. Puede pensarse en elegir aquélla que presente el más alto valor; es difícil que ello conduzca a resultados prácticos, normalmente en ese caso se calculará un tamaño de muestra excesivamente grande, sin posibilidades de financiamiento. Hay variables, cuya varianza es tan grande, que el cálculo del tamaño de muestra, con una precisión razonable, equivale en el hecho a casi "censar" la población. En todo caso, constituye la alternativa más defendible desde un punto de vista teórico. Las limitaciones que ofrece una investigación en la realidad, obliga en general a desecharla. Deberá

/pensarse en

pensarse en elegir una variable estratégica o representativa del conjunto de variables. En primer lugar es sumamente difícil de disponer de varianzas para todas las variables que interesen investigar, además, como se vió, aún en el hipotético caso que se dispusiera de tales estadígrafos, la muestra resultante podría exceder en demasía el presupuesto disponible. En la práctica es un poco más realista empezar determinando el tamaño de muestra en función de los recursos, y luego calcular "técnicamente" el tamaño de muestra en términos de la variable elegida como representativa, es decir, la que más interese en la investigación. De la comparación de los tamaños de muestra resultantes por esos dos métodos diferentes, nacerá el ajuste que será necesario hacer en aras de la factibilidad de la encuesta. Habrán tres alternativas: se aumenta la cantidad de recursos, se disminuye razonablemente la precisión, o ambas cosas en alguna medida.

De cualquier manera, la muestra que resulte, sólo garantizará la estimación de la media aritmética, con la precisión y confianza especificadas (d y t) de aquella variable cuya varianza se ha tomado en cuenta para el cálculo de n . Para aquellas variables que tengan una varianza menor que la de la variable clave, el tamaño de muestra será más que suficiente, es decir, se conseguirá mayor precisión que la especificada; en cambio para aquellas variables con mayor varianza, el tamaño de muestra será insuficiente, en consecuencia para dichas variables se estará especificando implícitamente una precisión menor (mayor d). Una vez realizada la encuesta y en conocimiento de las varianzas muestrales de todas las variables, deberá procederse a calcular los intervalos de confianza. Es frecuente encontrarse con sorpresas desagradables; hay algunos intervalos, cuya magnitud es tan extraordinariamente grande, que tomar la estimación puntual, implica serios riesgos. El planificador deberá tomar conciencia de estos hechos, para que en todo momento contemple, en la utilización de tales estimadores, sus bondades y limitaciones. Es muy frecuente escuchar la frase de que basta tener un orden de magnitud para formarse una idea, pero ocurre que a veces estos "órdenes de magnitud" tienen un rango de variabilidad tan amplio, que situándose en uno u otro límite del intervalo, las conclusiones pueden ser del todo diferentes. Esta situación ocurre

/particularmente en

particularmente en las investigaciones industriales a niveles muy desagregados, donde las poblaciones son demasiado pequeñas.

La elección de la variable clave es una trascendencia innecesaria de recalcar. En la encuesta sobre la industria en Centroamérica, para fines del cálculo del tamaño de muestra, se ha tomado como variable representativa, el tamaño de las industrias a través del número de obreros que ocupan. No sería aventurado adelantar que habrán otras variables como el valor agregado, los insumos, etc., que tendrán mayor variabilidad que la variable elegida. En esos casos, con seguridad que la precisión será mucho menor que la especificada para estimar el número promedio de obreros por industria. Lo interesante será analizar cuál es la precisión efectivamente alcanzada y la confianza que ello merece.

En cuanto a la precisión, indicada por la magnitud del semi-ancho del intervalo de confianza (d), en otras palabras el desvío máximo tolerable, constituye un grado de libertad en el simplificado modelo del cálculo de una muestra estratificada. Puede teóricamente tomar cualquier valor positivo, lo interesante es determinar un valor, un desvío, de manera que no signifique por una parte un rango demasiado amplio que no permita obtener conclusiones generales, y por otra que no sea tan innecesariamente pequeño que exija un tamaño de muestra prohibitivo. En verdad, en una encuesta industrial sólo será necesario determinar la precisión deseada para la variable clave, ya que para el resto de las variables, habrá dejado de ser un grado de libertad, ya que el tamaño de muestra es único para toda la investigación. La precisión alcanzada para las otras variables habrá que calcularla en función de ese único tamaño de muestra y de la desviación típica (s) de cada una de las variables. Como en general cada variable tiene distinto grado de heterogeneidad, tendrá también una diferente precisión. En la fijación del desvío máximo a tolerar para la variable representativa, el planificador tendrá que tomar una decisión. Si por ejemplo se sabe que el valor agregado de la industria en Chile, fué en 1963 de 811 millones de escudos de 1960, el promedio por industria será del orden de 115.000 escudos de 1960. Si además se supone que el valor agregado por industria crece al 5 por ciento anual,

/el promedio

el promedio para 1964, será de 121.000 escudos de 1960 aproximadamente. La pregunta que debe plantearse el planificador es: si se quisiera estimar este promedio ¿Cuál sería el desvío máximo (d) que se podría tolerar?; 10.000, 20.000 ó 25.000 escudos? Compulsando los tamaños de muestra resultantes para cada alternativa, los objetivos de la investigación, y el presupuesto disponible, se tendrá que tomar esa importante decisión. Cuando se tiene conciencia que el resto de las variables por investigar tiene una heterogeneidad mucho mayor que la de la variable clave, será necesario considerar en el tamaño de muestra un desvío menor al necesario, en la medida que no encarezca en exceso la encuesta industrial. Una vez calculado el tamaño definitivo de la muestra, el siguiente paso será la distribución de ella, entre los estratos y/o sub-estratos por la correspondiente afijación elegida.

5. Selección de las unidades muestrales

Se había advertido que el muestreo estratificado no era más que una combinación de diseños muestrales aleatorios simples; la selección de las unidades deberá ser hecha en forma aleatoria. Sin embargo, la extracción por medio de una tabla de números aleatorios puede significar un procedimiento muy engorroso como primera desventaja, y luego que la concentración industrial alrededor de las zonas urbanas, determina que una gran parte de la muestra provenga justamente de esas áreas. En el hecho lo mencionado en segundo término no constituye una desventaja, a no ser que se tome en cuenta que la muestra no entregará resultados por zonas geográficas, aspecto que debiera interesar al programador. En tales casos puede justificarse seleccionar muestras sistemáticas, es decir, una de cada k industrias de la población (siendo $k = \frac{N}{n}$). En el caso de la encuesta industrial de la Corporación de Fomento de Chile, se siguió una extracción sistemática de norte a sur del país, garantizando de esta manera cierta representatividad de zonas geográficas que de otra manera no la hubieran tenido. La extracción de una muestra sistemática puede asociarse con una muestra aleatoria, siempre que la población no presente un determinado ordenamiento coincidente con el intervalo de sistematización y que la iniciación de la selección sea aleatoria, es decir, que una de las k primeras unidades sea elegida al azar. En tal caso, podrá utilizarse

/los mismos

los mismos estimadores de las muestras aleatorias, o las aproximaciones correspondientes al muestreo sistemático.

6. Entrenamiento de enumeradores

El enumerador, deberá ser una persona que conociendo perfectamente los objetivos de la encuesta, debe tener alguna experiencia en materia de industrias. En general, en dichas encuestas, se inquiriere sobre una serie de datos técnicos, muchos de los cuales son dados directamente por el respondiente; el enumerador debe estar en condiciones de calificar estas respuestas. En las encuestas de opinión pública, en materias de política, religión, simpatías, actitudes, etc., es correcto que el enumerador sea un elemento en lo posible neutro hacia las respuestas del encuestado. En las encuestas industriales, sobre todo en las preguntas que dicen relación con magnitudes, el encuestador tendrá una participación activa y verificadora. Incluso para la realización de encuestas industriales donde se averigüe por datos técnicos, es conveniente que el encuestador sea un técnico entendido en la materia; de otro modo se corre el riesgo de obtener respuestas deficientes, que pueden estropear el trabajo.

En todo caso será necesario un proceso de adiestramiento, probando en el terreno la eficiencia de cada enumerador, sobre la base de indagaciones que exijan una cabal interpretación, teniendo previamente a disposición los datos que averiguará en el terreno cada encuestador.

7. Colaboración de la población informante

En las encuestas industriales, es de una significación muy grande conseguir una actitud positiva de la población que se investiga, imprescindible para garantizar una cierta fidelidad en la masa de informaciones a recolectar. Es frecuente que se advierta a la población la trascendencia de la muestra, y una manera de hacer más efectiva esa colaboración, es realizar conjuntamente la investigación con asociaciones de industriales u organismos de probado prestigio, Universidades, Institutos, etc. De una u otra manera existe siempre un margen considerable de no respuesta. Sobre el particular hay que destacar que si la no respuesta tiene su origen en una desaparición o cambio de estrato de la unidad que se pretende investigar, no debe ser motivo de preocupación, ya que puede tomarse como una muestra representativa de una parte del universo que desaparece o cambia de estrato. El problema es muchísimo más serio cuando

/existiendo la

existiendo la unidad, hay negativa de dar respuesta. Basar las estimaciones solamente en poblaciones de respondientes, puede ocasionar sesgos de alguna magnitud. En esos casos las estimaciones realizadas, podrán ser generalizadas sólo a la población de respondientes y no a la población total. Para tener informaciones sobre la población de no respondientes, habrá que disponer de encuestas de seguimiento u otros métodos indirectos. Sobre este problema de la no respuesta y además de él, sobre la respuesta defectuosa, es necesario una consideración detenida. La necesidad de planificar y para ello de contar con informaciones serias, periódicas y oportunas, son cuestiones de aceptación general. Para poder contar con informaciones de tales atributos, parece que no hay otra salida, que en las legislaciones de los países se contemple la obligación y las sanciones correspondientes, en forma similar a una declaración de impuestos.

No puede esperarse a que las poblaciones tomen conciencia del agudo problema. Esto implicaría la elaboración previa de planes nacionales de estadística, donde se coordinarían todas las necesidades de información de los diferentes organismos de planificación, ejecución y control.

8. Organización del trabajo en el terreno

La programación de la recolección de informaciones debe plantearse en términos de las peculiaridades geográficas, concentración industrial en la urbe, etc. Debiera haber un contacto permanente entre la oficina de diseño de la muestra y el equipo de encuestadores. Siempre se presentarán tipos de respuesta o características de la industria que no se han contemplado en el proceso de adiestramiento. En tales casos es preferible solucionar el problema consultando a los directores de la investigación. En encuestas industriales a nivel nacional, es conveniente que en los lugares donde no hay oficinas a cargo de los que dirigen la encuesta, la enumeración debe ser realizada por personas altamente calificadas para este tipo de trabajos. En la práctica, la recolección de datos en provincias, es realizada por personal de las Juntas de Planificación u organismos similares, dejando la investigación en la capital o sede de la encuesta, a personal temporal contratado para este efecto.

Es indispensable que se implanten ciertos controles en forma simultánea

a la recolección de informaciones. De esta manera se previene que se cometan errores voluntarios e involuntarios. Verificar una pequeña parte del número de encuestas que realiza cada enumerador, puede ser un medio para disminuir un posible sesgo. Es importante que se realice esta forma de control, simultáneamente al proceso de extracción de información, para que puedan tomarse medidas oportunas de corrección. Los controles ex-post, son en general tardíos y no se alcanza a remediar el problema, si no a expensas de mucho mayores gastos y demoras perjudiciales.

9. Sistematización de datos

Toda la masa de informaciones extraídas deberá ser codificada, tabulada, clasificada, interpretada y verificada. En esta altura es donde se deberá diseñar el nivel de desagregación y tipos de clasificaciones necesarias. Por una parte, las informaciones a nivel infimo de desagregación, la unidad de muestreo, no permite obtener conclusiones, por otra parte, estimaciones globales sólo dan una idea muy general de la situación de la industria en sus diversos aspectos. Es imprescindible pues combinar estimadores a distintos niveles de desagregación; unas y otras son necesarias cuando se pretende fijar una política y estrategia de desarrollo industrial. Conocer por ejemplo el coeficiente producto capital para toda la industria manufacturera es un dato de mucha utilidad, pero además es necesario conocer sus componentes, ya sea por ramas industriales, por ubicación geográfica, por tamaño de establecimiento, etc. Con la utilización de máquinas perforadoras y computadores electrónicos, la sistematización de datos se hace mucho más amplia y rápida. Previamente a la introducción de los datos en los computadores, aquéllos debieron ser sometidos a un estricto control, máxime si fueron extraídos por enumeradores que no tenían un cabal conocimiento técnico-económico de la industria. Este control en general se realiza por comparación entre industrias similares, por antecedentes que se tienen acerca de los probables resultados que arrojaría la encuesta en cada caso y por revisión analítica en cuanto a lo razonables que pudieran ser esas informaciones.

10. Determinación del costo total:

Para la realización de una investigación muestral, se ha debido contar con un determinado presupuesto, el que será necesario comparar con el costo real

de la investigación. En la misma forma en que se ha realizado el presupuesto, contemplando cada una de las principales etapas, se deberá realizar la comparación con la realidad. De esta manera se acumularán experiencias útiles para posteriores investigaciones, ya que en todas las diferencias que se produzcan, habrá que analizar sus causas; así se tomará una mayor conciencia de la trascendencia de cada etapa, lo que puede redundar en mayor aprovechamiento de los recursos disponibles en futuras investigaciones.

11. Publicación de resultados:

En la publicación de los resultados obtenidos mediante muestras, es fundamental indicar las principales características del diseño utilizado: criterio de estratificación, tamaño de muestra, tipo de afijación, estratos en que se hizo censo, precisión para cada una de las variables principales, la probabilidad de acierto, formas de selección, etc. De esta manera los usuarios conocerán las bondades y limitaciones de que son objeto las estimaciones resultantes. Incluso es necesario presentar un anexo con una descripción detallada de toda la investigación, detallando los principales problemas, y las formas de solución.

DEMOSTRACIONES MATEMATICAS DE MUESTREO ALEATORIO SIMPLE

I. La media de la muestra, \bar{y} , es un estimador insesgado de \bar{Y} , media de la población.

Demostración: se sabe que

$$A) \quad E(\bar{y}_h) = \frac{\sum_h \bar{y}_h}{\binom{N}{n}} = \frac{\sum_h (y_1 + y_2 + y_3 + \dots + y_n)/h}{n} \cdot \frac{1}{\frac{N!}{n! (N-n)!}}$$

donde $h = 1, 2, 3, \dots, \binom{N}{n}$
 $i = 1, 2, 3, \dots, n \dots N$

Para poder evaluar la suma que aparece en el numerador, es necesario averiguar en cuántas muestras aparece un valor específico cualquiera y_i . Aparte de ese valor y_i , habrá otras $(N-1)$ unidades disponibles para el resto de la muestra y otros $(n-1)$ lugares a ocupar en la muestra. Luego el número de muestras que contienen el valor y_i estará dado por:

$$C_{n-1}^{N-1} = \frac{(N-1)!}{(n-1)!(N-n)!}$$

Por lo tanto:

$$B) \quad \sum_{h=1}^{\binom{N}{n}} (y_1 + y_2 + y_3 \dots + y_n)/h = \frac{(N-1)!}{(n-1)!(N-n)!} (y_1 + y_2 + y_3 + \dots + y_N)$$

Nótese que, por el anterior artificio, la suma de los valores, se extiende hasta y_N , porque el valor específico y_i , puede ser cualquiera de los valores poblacionales.

Reemplazando la expresión obtenida en B en el numerador de la igualdad señalada con A, se tiene:

$$E(\bar{y}_h) =$$

$$E(\bar{y}_h) = \frac{(N-1)!}{(n-1)!(N-n)!} (y_1 + y_2 + y_3 + \dots + y_N) \cdot \frac{1}{n} \cdot \frac{1}{N!} \cdot n!(N-n)!$$

$$E(\bar{y}_h) = \frac{(N-1)!}{(n-1)!(N-n)!} (y_1 + y_2 + y_3 + \dots + y_N) \frac{1}{n} \frac{n!(N-n)!}{N}$$

Simplificando factoriales:

$$E(\bar{y}_h) = \frac{(y_1 + y_2 + y_3 + \dots + y_N)}{N} = \bar{Y}$$

II. Demostración

$$E \left[(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2 \right] = \frac{n(N-1)}{N} S^2$$

Donde $S^2 = \frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1}$ (Varianza de Cochran)

La esperanza propuesta es igual a:

$$c) \frac{\sum_{h=1}^{\binom{N}{n}} (y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2}{\binom{N}{n}}$$

Nótese que los índices de la sumatoria se extienden a todas las posibles muestras.

Utilizando el mismo artificio que en la demostración I, el numerador se puede evaluar de la siguiente manera:

$$\binom{N-1}{n-1} \left[(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2 \right]$$

/Reemplazando esta

Reemplazando esta expresión, en el numerador de C se tiene:

$$\frac{\binom{N-1}{n-1} [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + (y_3 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2]}{\binom{N}{n}}$$

$$= \frac{(N-1)!}{(n-1)! (N-n)!} \cdot \frac{n! (N-n)!}{N!} [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2]$$

Simplificando se obtiene:

$$\frac{n}{N} [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2]$$

Multiplicando y dividiendo por N-1:

$$\frac{(N-1)n}{N} \left[\frac{(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2}{N-1} \right] =$$

$$\frac{(N-1)n}{N} \left[\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1} \right]$$

La expresión dentro del paréntesis es lo que se había llamado S^2 (varianza de Cochran).

Luego, queda demostrado que:

$$E [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2] = \frac{n(N-1)}{N} S^2$$

$$\text{III. } E [(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \dots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})]$$

$$= -\frac{n(n-1)}{2N} S^2$$

La esperanza indicada significa:

$$\frac{\sum_{h=1}^n \binom{N}{n} [(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \dots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})]}{\binom{N}{n}} / h$$

/Para evaluar

Para evaluar la suma del numerador, debe pensarse en el mismo artificio utilizado en las demostraciones I y II, teniendo en cuenta que como ahora se trata de productos de dos desviaciones, habrá dos valores específicos en la población que podrán ocupar dos lugares en la muestra. La evaluación de la suma del numerador estará dada, pues, por:

$$\begin{aligned} & \binom{N-2}{n-2} \left[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_2 - \bar{Y}) + (y_3 - \bar{Y}) + \dots + (y_{N-1} - \bar{Y})(y_N - \bar{Y}) \right] = \\ & = \binom{N-2}{n-2} \frac{\left[(y_1 - \bar{Y})(y_2 - \bar{Y}) + \dots + (y_{N-1} - \bar{Y})(y_N - \bar{Y}) \right]}{\binom{N}{n}} = \\ & = \frac{(N-2)!}{(n-2)! (N-n)!} \cdot \frac{(N-n)! n!}{N!} \left[(y_1 - \bar{Y})(y_2 - \bar{Y}) + \dots + (y_{N-1} - \bar{Y})(y_N - \bar{Y}) \right] \end{aligned}$$

Simplificando factoriales y multiplicando y dividiendo por 2 se tiene:

$$\frac{n(n-1)}{2N(N-1)} \left[2 \left\{ (y_1 - \bar{Y})(y_2 - \bar{Y}) + \dots + (y_{N-1} - \bar{Y})(y_N - \bar{Y}) \right\} \right]$$

A la expresión dentro del paréntesis cuadrado, le sumaremos y restaremos los términos:

$$(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2$$

Luego

$$\begin{aligned} & \frac{n(n-1)}{2N(N-1)} \left[2 \left\{ (y_1 - \bar{Y})(y_2 - \bar{Y}) + \dots + (y_{N-1} - \bar{Y})(y_N - \bar{Y}) \right\} \right. \\ & + \left. \left\{ (y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2 \right\} \right. \\ & \quad \text{suma} \\ & - \left. \left\{ (y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2 \right\} \right] \\ & \quad \text{resta} \end{aligned}$$

Observando los términos dentro del paréntesis cuadrado, se concluye que la expresión que tiene signo negativo es el numerador de la varianza de Cochran, y los dos primeros que tienen signo positivo, son el desarrollo del cuadrado

/de una

de una suma de desviaciones, ya que aparecen las desviaciones al cuadrado y los dobles productos correspondientes, es decir

$$\sum_{i=1}^N (y_i - \bar{Y})^2 + 2 \sum_{i \neq j=1}^N (y_i - \bar{Y})(y_j - \bar{Y}) = \left[\sum_{i=1}^N (y_i - \bar{Y}) \right]^2 = 0$$

por tratarse del cuadrado de la suma de las desviaciones respecto de la media aritmética (es decir, el cuadrado de cero). Con esta reducción, la expresión queda de la siguiente manera:

$$\begin{aligned} & \frac{n(n-1)}{2N(N-1)} \left[- \left\{ (y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + (y_3 - \bar{Y})^2 + \dots + (y_N - \bar{Y})^2 \right\} \right] \\ = & - \frac{n(n-1)}{2N(N-1)} \left[\sum_{i=1}^N (y_i - \bar{Y})^2 \right] \\ = & - \frac{n(n-1)}{2N} \left[\frac{\sum_{i=1}^N (y_i - \bar{Y})^2}{N-1} \right] \\ = & - \frac{n(n-1)}{2N} S^2 \end{aligned}$$

IV.

$$V[\bar{y}] = E(\bar{y}_h - \bar{Y})^2 = \frac{\sum_{h=1}^N (\bar{y}_h - \bar{Y})^2}{\binom{N}{n}} = \frac{S^2}{n} \left(1 - \frac{n}{N}\right)$$

Empezamos la demostración por esta identidad

$$\begin{aligned} n(\bar{y} - \bar{Y}) &= n \cdot \frac{\sum_{i=1}^n y_i}{n} - n\bar{Y} \\ &= y_1 + y_2 + y_3 + \dots + y_n - n\bar{Y} \\ &= y_1 - \bar{Y} + y_2 - \bar{Y} + y_3 - \bar{Y} + \dots + y_n - \bar{Y} \end{aligned}$$

Elevando al cuadrado ambos miembros se tiene:

$$n^2 (\bar{y} - \bar{Y})^2 = \left[(y_1 - \bar{Y}) + (y_2 - \bar{Y}) + (y_3 - \bar{Y}) + \dots + (y_n - \bar{Y}) \right]^2$$

/Fórmula

$$n^2 (\bar{y} - \bar{Y})^2 = (y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2 +$$

$$+ 2 \left[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \dots + (y_{n-1} - \bar{Y})(y_n - \bar{Y}) \right]$$

Aplicando el operador esperanza matemática:

$$n^2 E(\bar{y} - \bar{Y})^2 = E \left[(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \dots + (y_n - \bar{Y})^2 \right] +$$

$$+ 2E \left[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \dots + (y_{n-1} - \bar{Y})(y_n - \bar{Y}) \right]$$

Los dos términos de la derecha, tienen expresiones conocidas, demostradas en los puntos II y III.

Reemplazando tales expresiones se tiene:

$$n^2 E(\bar{y} - \bar{Y})^2 = \frac{n(N-1)}{N} S^2 + 2 \left[-\frac{n(n-1)}{2N} S^2 \right]$$

$$n^2 E(\bar{y} - \bar{Y})^2 = \frac{n(N-1)}{N} S^2 - \frac{n(n-1)}{N} S^2$$

$$n^2 E(\bar{y} - \bar{Y})^2 = \frac{nS^2 \left[(N-1) - (n-1) \right]}{N}$$

$$n^2 E(\bar{y} - \bar{Y})^2 = nS^2 \frac{(N-n)}{N}$$

$$E(\bar{y} - \bar{Y})^2 = \frac{S^2}{n} \left(1 - \frac{n}{N} \right)$$

Esta es la fórmula de cálculo práctico de la varianza de la media de una muestra aleatoria simple. (Cuadrado del error standard de estimación).

V. Demostración de que la varianza de una muestra aleatoria simple es una estimación insesgada de la varianza de la población (ambas con denominadores restados en una unidad).

$$E(s_h^2) = E \left[\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} \right]$$

/Sumando y

Sumando y restando la media poblacional:

$$(n-1) E (s_i^2) = E \left[\sum_{i=1}^n (y_i - \bar{y} \pm \bar{Y})^2 \right]$$

$$(n-1) E (s_i^2) = E \left[\sum_{i=1}^n (y_i - \bar{Y}) - (\bar{y} - \bar{Y}) \right]^2 *$$

pero si hacemos:

$$y_i - \bar{Y} = W_i; \quad \bar{y} - \bar{Y} = \bar{W}$$

tenemos:

$$\begin{aligned} \sum_{i=1}^n [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2 &= \sum_{i=1}^n (W_i - \bar{W})^2 \\ &= \sum_{i=1}^n W_i^2 - 2\bar{W} \sum_{i=1}^n W_i + n\bar{W}^2 \\ &= \sum_{i=1}^n W_i^2 - 2\bar{W} (n\bar{W}) + n\bar{W}^2 \\ &= \sum_{i=1}^n W_i^2 - n\bar{W}^2 \\ &= \sum_{i=1}^n (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \end{aligned}$$

Reemplazando esta expresión, en aquella marcada con el asterisco, se tiene:

$$\begin{aligned} (n-1) E (s^2) &= E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2 \right] \\ &= E \left[\sum_{i=1}^n (y_i - \bar{Y})^2 \right] - n E (\bar{y} - \bar{Y})^2 \end{aligned}$$

Teniendo en cuenta las demostraciones II y IV y reemplazando dichas expresiones en el miembro de la derecha de la igualdad anterior:

/Fórmula

$$\begin{aligned}(n-1) E (s^2) &= \frac{n(N-1)}{N} S^2 - n \frac{S^2}{n} \left(1 - \frac{n}{N}\right) \\ &= \frac{n(N-1)}{N} S^2 - S^2 \left(1 - \frac{n}{N}\right) \\ &= S^2 \frac{n(N-1) - (N-n)}{N} \\ &= S^2 \frac{Nn - n - N + n}{N} \\ &= S^2 \frac{N(n-1)}{N}\end{aligned}$$

Se obtiene finalmente:

$$E (s^2) = S^2$$

FORMULARIOS

Formulario sobre muestreo Aleatorio Simple

A. Variable cuantitativa

N: Número de elementos que componen la población.

n: Número de elementos que componen la muestra.

y_i : Valor de la variable en la i-ésima unidad de la población.
(i = 1, 2, 3, N).

y_i : Valor de la variable en la i-ésima unidad de la muestra.
(i = 1, 2, 3, n).

\bar{Y} : Media aritmética poblacional:

$$\bar{Y} = \frac{\sum_{i=1}^N y_i}{N}$$

\bar{y} : Media aritmética de la muestra:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

{ Estimador insesgado de la media aritmética poblacional }

σ^2 : Varianza de la población:

$$\sigma^2 = \frac{\sum_{i=1}^N (y_i - Y)^2}{N}$$

S^2 : Varianza de Cochran para la población:

$$S^2 = \frac{\sum_{i=1}^N (y_i - Y)^2}{N - 1} = \frac{\sum_{i=1}^N y_i^2}{N - 1} - \frac{N \bar{Y}^2}{N - 1}$$

s^2 : Varianza de Cochran para la muestra:

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

Estimador insesgado de la varianza de Cochran para la población.

/Y: Total

Y: Total poblacional:

$$Y = N \bar{Y} = \sum_{i=1}^N y_i$$

\hat{Y} : Estimador del total poblacional

$$\hat{Y} = N \bar{y}$$

$V(\bar{y})$: Varianza verdadera del estimador de la media o cuadrado de la desviación standard verdadera:

$$V(\bar{y}) = \frac{S^2}{n} \left(1 - \frac{n}{N}\right)$$

$v(\bar{y})$: Estimador de la varianza del estimador de la media:

$$v(\bar{y}) = \frac{s^2}{n} \left(1 - \frac{n}{N}\right)$$

$V(\hat{Y})$: Varianza verdadera del estimador del total:

$$V(\hat{Y}) = N^2 V(\bar{y})$$

$v(\hat{Y})$: Estimador de la varianza del estimador del total:

$$v(\hat{Y}) = N^2 v(\bar{y})$$

n_o : Tamaño de muestra (primera aproximación):

$$n_o = \frac{t^2 S^2}{d^2}$$

en que: "t" es el coeficiente de confianza y "d" es el semi-ancho del intervalo de confianza.

$$d = t \sqrt{V(\bar{y})}$$

n: Tamaño de la muestra (definitivo):

$$\frac{n_o}{1 + \frac{n_o}{N}}$$

B. Variable cualitativa (dos categorías)

La variable solamente puede tomar dos valores: 1 si el elemento (en la población 0 en la muestra) posee la característica que se investiga, y 0 si no la posee.

P: Proporción en la población:

$$P = \frac{\sum_{i=1}^N y_i}{N}$$

p: Proporción en la muestra:

$$p = \frac{\sum_{i=1}^n y_i}{n} \quad \text{Estimador insesgado de la proporción poblacional.}$$

S²: Varianza de Cochran para la población:

$$S^2 = \frac{N}{N-1} PQ \quad Q = 1 - P$$

s²: Varianza de Cochran para la muestra:

$$s^2 = \frac{n}{n-1} pq \quad \text{Estimador insesgado de la varianza de Cochran para la población.}$$
$$q = 1 - p$$

A: Número de elementos que poseen la característica en la población:

$$A = NP = \sum_{i=1}^N y_i$$

\hat{A} : Estimador del número de elementos que poseen la característica en la población:

$$\hat{A} = N p$$

/V(p):

$V(p)$: Varianza verdadera del estimador de la proporción o cuadrado de la desviación standard verdadera:

$$V(p) = \frac{P Q}{n} \left(\frac{N - n}{N - 1} \right)$$

$v(p)$: Estimador de la varianza del estimador de la proporción:

$$v(p) = \frac{p q}{n-1} \left(\frac{N - n}{N} \right)$$

$V(\hat{A})$: Varianza verdadera del estimador del número de elementos que poseen la característica que se investiga:

$$V(\hat{A}) = N^2 V(p)$$

$v(\hat{A})$: Estimador de la varianza del estimador del número de elementos que poseen la característica que se investiga:

$$v(\hat{A}) = N^2 v(p)$$

n_o : Tamaño de la muestra (primera aproximación)

$$n_o = \frac{t^2 P Q}{d^2}$$

en que: "t" es el coeficiente de confianza y "d" es el semi-ancho del intervalo de confianza.

$$d = t \sqrt{V(p)}$$

n : Tamaño de muestra (definitivo):

$$n = \frac{n_o}{1 + \frac{n_o - 1}{N}}$$

Formulario sobre Muestreo Aleatorio Estratificado

A. Variables cuantitativas

N : Número de elementos que componen la población

n : Número de elementos que componen la muestra

N_h : Número de elementos del estrato h-ésimo

n_h : Número de elementos de la muestra del estrato h-ésimo

y_{hi} : Valor de la i-ésima unidad del estrato h-ésimo

$i = 1, 2, 3, \dots, N_h; h = 1, 2, 3, \dots, L$

y_{hi} : Valor de la i-ésima unidad de la muestra del estrato h-ésimo.

$i = 1, 2, 3, \dots, n_h; h = 1, 2, 3, \dots, L$

\bar{Y}_h : Media aritmética del estrato h-ésimo.

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$$

\bar{Y} : Media aritmética de la población

$$\bar{Y} = \frac{\sum_{h=1}^L \bar{Y}_h N_h}{N}$$

\bar{y}_h : Media aritmética de la muestra del estrato h-ésimo

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \quad \text{Estimador insesgado de } \bar{Y}_h$$

\bar{y} : Media aritmética de la muestra total:

$$\bar{y} = \frac{\sum_{h=1}^L \bar{y}_h n_h}{n}$$

$\sqrt{y_{st}}$:

\bar{y}_{st} : Estimador insesgado de la media aritmética poblacional

$$\bar{y}_{st} = \frac{\sum_{h=1}^L \bar{y}_h N_h}{N}$$

S_h^2 : Varianza de Cochran para el estrato h-ésimo

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)^2}{N_h - 1} = \frac{\sum y_{hi}^2}{N_h - 1} - \frac{N_h \bar{y}_h^2}{N_h - 1}$$

s_h^2 : Varianza de Cochran para la muestra del estrato h-ésimo
(estimador insesgado de S_h^2)

$$s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} = \frac{\sum y_{hi}^2}{n_h - 1} - \frac{n_h \bar{y}_h^2}{n_h - 1}$$

Y: Total poblacional

$$Y = N \bar{Y} = \sum_{h=1}^L \sum_{i=1}^{N_h} y_{hi}$$

\hat{Y}_{st} : Estimador del total poblacional

$$\hat{Y}_{st} = N \bar{y}_{st}$$

$V(\bar{y}_{st})$: Varianza verdadera del estimador de la media

a) Fórmula general:

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^L N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

b) Para afijación proporcional

$$V(\bar{y}_{st}) = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 \quad \text{donde:}$$

$$f = \frac{n}{N} \quad \text{y} \quad W_h = \frac{N_h}{N}$$

/c) Para

c) Para afijación óptima

$$V(\bar{y}_{st}) = \frac{\left[\sum_{h=1}^L W_h S_h \right]^2}{n} - \frac{\sum_{h=1}^L W_h S_h^2}{N}$$

$v(y_{st})$: Varianza estimada del estimador de la media (Las mismas fórmulas anteriores, pero utilizando s_h^2 en vez de S_h^2 , por ser estimador insesgado)

n_h : Tamaño de la muestra en el estrato h-ésimo

a) Afijación proporcional

$$n_h = \frac{N_h}{N} \cdot n$$

b) Afijación óptima

$$n_h = \frac{N_h S_h}{\sum N_h S_h} \cdot n$$

c) Afijación óptima económica

$$n_h = \frac{N_h S_h / \sqrt{c_h}}{\sum N_h S_h / \sqrt{c_h}} \cdot n$$

n : Tamaño de la muestra; para estimar la media aritmética.

a) Afijación proporcional

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{donde } n_0 = \frac{\sum_{h=1}^L W_h S_h^2}{V(\bar{y}_{st})}$$

b) Afijación

b) Afijación óptima

$$n = \frac{\left[\sum_{h=1}^L W_h S_h \right]^2}{V(\bar{y}_{st}) + \frac{1}{N} \sum_{h=1}^L W_h S_h^2}$$

B. Variable cualitativa (dos categorías)

La variable solamente puede tomar dos valores: 1 si el elemento (en la población o en la muestra) posee la característica que se investiga, y 0 si no la posee.

(La nomenclatura es similar a la presentada en el punto A).

P_h : Proporción en el estrato h-ésimo

$$P_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h} = \frac{A_h}{N_h}$$

P : Proporción en la población

$$P = \frac{\sum_{h=1}^L P_h N_h}{N}$$

p_h : Proporción en la muestra del estrato h-ésimo

$$p_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} = \frac{a_h}{n_h} \quad \text{Estimador insesgado de } P_h$$

p : Proporción de la muestra total

$$p = \frac{\sum_{h=1}^L P_h n_h}{n}$$

$/P_{st}$:

P_{st} : Estimador insesgado de la proporción poblacional

$$P_{st} = \frac{\sum_{h=1}^L p_h N_h}{N}$$

A : Número de elementos que poseen la característica en población

$$A = NP$$

A : Estimador del total poblacional (número de elementos)

$$A = N P_{st}$$

S_h^2 : Varianza del estrato h-ésimo

$$S_h^2 = \frac{N_h P_h Q_h}{N_h - 1}$$

$V(p_{st})$: Varianza verdadera del estimador de la proporción poblacional

a) Fórmula general

$$V(p_{st}) = \frac{1}{N^2} \sum_{h=1}^L \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}$$

b) Afijación proporcional

$$V(p_{st}) = \frac{N - n}{N} \frac{1}{n N} \sum_{h=1}^L \frac{N_h^2 P_h Q_h}{N_h - 1}$$

$$= \frac{1 - f}{n} \sum_{h=1}^L W_h P_h Q_h$$

c) Afijación óptima

$$V(p_{st}) = \frac{\left(\sum_{h=1}^L W_h \sqrt{\frac{N_h P_h Q_h}{N_h - 1}} \right)^2}{n} - \frac{\sum_{h=1}^L W_h \frac{N_h P_h Q_h}{N_h - 1}}{N}$$

/V(A):

$V(\hat{A})$: Varianza verdadera del estimador del total

$$V(\hat{A}) = N^2 V(p_{st})$$

$v(p_{st})$: Varianza estimada del estimador de la proporción. Se utilizan las mismas fórmulas anteriores, pero sustituyendo:

$$\frac{N_h P_h Q_h}{N_h - 1} \text{ por } \frac{n_h P_h Q_h}{n_h - 1}, \text{ por ser estimador insesgado.}$$

$v(\hat{A})$: Estimador de la varianza del estimador del total

$$v(\hat{A}) = N^2 v(p_{st})$$

n_h : Tamaño de la muestra en el estrato h-ésimo

a) Afijación proporcional

$$n_h = \frac{N_h}{N} \cdot n$$

b) Afijación óptima

$$n_h = \frac{N_h (P_h Q_h)^{1/2}}{\sum N_h (P_h Q_h)^{1/2}} \cdot n$$

c) Afijación óptima económica

$$n_h = \frac{N_h (P_h Q_h / c_h)^{1/2}}{\sum N_h (P_h Q_h / c_h)^{1/2}} \cdot n$$

n : Tamaño de la muestra para estimar una proporción

a) Afijación proporcional

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{donde} \quad n_0 = \frac{\sum W_h P_h Q_h}{V(p_{st})}$$

/b) Afijación

b) Afijación óptima

$$n = \frac{n_0}{1 + \frac{1}{N V(p_{st})} \sum W_h P_h Q_h}$$

$$\text{donde } n_0 = \frac{\left[\sum W_h \sqrt{P_h Q_h} \right]^2}{V(p_{st})}$$

