AGENCY FOR INTERNATIONAL DEVELOPMENT
WASHINGTON, D. C. 20523

BIBLIOGRAPHIC INPUT SHEET

FOR AID USE ONLY

1. SUBJECT CLASSIFICATION

A. PRIMARY: Economics

B. SECONDARY: Agricultural Economics

2. TITLE AND SUBTITLE

Expected value of a sample estimate

3. AUTHOR(S)

Houseman, E.E.

4. DOCUMENT DATE	5. NUMBER OF PAGES	6. ARC NUMBER
1974	145 p.	ARC

7. REFERENCE ORGANIZATION NAME AND ADDRESS

Statistical Reporting Service, U.S. Department of Agriculture
Washington, D.C.

8. SUPPLEMENTARY NOTES *(Sponsoring Organization, Publishers, Availability)*

(In U.S. Dept. of Agriculture, Statistical Reporting Service, SRS no.19)
(Available in English and Spanish; Spanish, 143p.: PN-AAC-142)

9. ABSTRACT

Relevant, introductory mathematics and probability in the context of sample surveys. Although some proofs are presented, the emphasis is more on exposition of mathematical language and concepts than on the mathematics per se and rigorous proofs. Many problems are given as exercises so a student may test his interpretation or understanding of the concepts. Each chapter begins with very simple explanations and ends at a much more advanced level. Chapters I and II provide a background for Chapter III, which is a discussion of expected values of random variables and the original purpose of this monograph. Finally, Chapter IV focuses attention on the distribution of an estimate which is the basis for comparing the accuracy of alternative sampling plans as well as a basis for statements about the accuracy of an estimate from a sample.

10. CONTROL NUMBER

PN-AAC-141 in English, PN-AAC-142 in Spanish

11. PRICE OF DOCUMENT

12. DESCRIPTORS

13. PROJECT NUMBER

14. CONTRACT NUMBER

RSSA/USDA 7-74 GTS

15. TYPE OF DOCUMENT

AID 590-1 (4-74)

VALOR ESPERADO DE UN ESTIMADOR MUESTRAL

por

Earl E. Houseman
Estadístico Matemático

Statistical Reporting Service
United States Department of Agriculture

Washington, D. C. 20250 January 1975

Esta publicación es traducción de EXPECTED VALUE OF A SAMPLE ESTIMATE, editado originalmente en inglés por el Departamento de Agricultura de los EE. UU. de A.

Traducción de Raymond Stadelman, Agencia para el Desarrollo Internacional (AID), Washington, D.C.

Traducción revisada por el Profesor Evelio O. Fabbroni, Centro Interamericano de Enseñanza de Estadística, en Santiago de Chile.

(N. del T.: Puesto que el Profesor Fabbroni tuvo oportunidad de corregir únicamente la penúltima versión de esta traducción, la responsabilidad de cualesquiera errores existentes en ésta recae sobre el traductor)

PROLOGO

El Servicio de Información Estadística (Statistical Reporting Service) durante muchos años ha dedicado esfuerzos a la capacitación de estadísticos agropecuarios oriundos de varias partes del mundo. La mayoría de estos estudiantes han recibido ayuda de la Agencia para el Desarrollo Internacional (AID) bajo sus programas de adiestramiento; sin embargo, muchos de ellos, bajo el patrocinio de la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO), entran al Centro Internacional para Programas Estadísticos de la Oficina del Censo del Departamento de Comercio -- Centro con el cual coopera este Servicio.

Este tratado ha sido preparado por el Servicio de Información Estadística con la cooperación de la AID y del Centro con el propósito de suministrar material mejorado didáctico y de referencia en el ramo de la estadística agropecuaria para el uso no solamente de los estudiantes extranjeros sino también del personal empleado por estas entidades.

HARRY C. TRELOGAN
Administrador, SRS

PREFACIO

El autor es de la opinión de que la teoría elemental de valores esperados de una variable aleatoria debe recibir más atención en los cursos prácticos sobre el muestreo. La teoría que trata de variables aleatorias y funciones de las mismas es la base del muestreo probabilístico. La interpretación de la exactitud de estimaciones hechas de encuestas por muestreo probabilístico se funda en, entre otras cosas, la teoría de valores esperados.

Hay muchos estudiantes cuya carrera exige la aplicación del muestreo probabilístico que han recibido muy poca instrucción en matemáticas y estadística. El adiestramiento en el muestreo debe ir más allá del mero trato de diseños de muestras de una manera descriptiva. Deben incluirse los fundamentos de las matemáticas y la estadística. Esto aumentaría mucho el entendimiento de sesgo, error del muestreo aleatorio, los componentes del error y otros conceptos técnicos; puede aumentar la capacidad de hacer adaptaciones prácticas de los principios del muestreo y del uso correcto de fórmulas; y puede facilitar la comunicación con estadísticos matemáticos y rendirla más provechosa.

El propósito de esta monografía es servir como referencia para la conveniencia de los estudiantes del muestreo. Procura expresar las matemáticas y la probabilidad pertinentes en forma introductoria y en el contexto de encuestas por muestreo. Aunque se presentan algunas pruebas, se hace hincapié en la exposición de lenguaje y conceptos matemáticos más bien que en las matemáticas <u>per se</u> y la prueba rigurosa. Se dan muchos problemas como ejercicios para que el estudiante pruebe su interpretación o entendimiento de los conceptos. La mayor parte de las matemáticas es elemental. Si alguna fórmula parece complicada, es probablemente porque representa una larga secuencia de operaciones aritméticas.

Cada capítulo empieza con explicaciones muy sencillas y termina en un nivel mucho más alto. La mayoría de los estudiantes que hayan estudiado el álgebra elemental no deben tener dificultad respecto a los principios de los capítulos. Los que han cursado matemáticas más avanzadas y estadística pueden repasar rápidamente los principios de los capítulos y dedicar más tiempo al estudio de las partes más difíciles. No es necesario terminar un capítulo antes de seguir al siguiente.

Una explicación de los valores esperados de variables aleatorias, como la dada en el capítulo III, fue la razón inicial que motivó esta monografía. Los capítulos I y II se agregaron como base para el capítulo III. El capítulo IV enfoca atención sobre la distribución de un estimador, que es la base para comparar la exactitud de planes alternativos de muestreo y para opinar sobre la exactitud de una estimación hecha de una muestra. El contenido del capítulo IV se encuentra en textos sobre el muestreo, pero es importante que los estudiantes reciban de más de una fuente la explicación de la distribución de un estimador, especialmente con respecto a estimaciones hechas de encuestas reales por muestreo.

Mi interés y mi experiencia en la instrucción de estudiantes han sido relacionados principalmente con personas que empezaron sus carreras en encuestas agrícolas. Agradezco al Statistical Reporting Service la oportunidad de preparar esta monografía.

Earl E. Houseman
Estadístico

INDICE

Página

CAPITULO I. NOTACION Y SUMATORIA

1.1 Introducción 1

1.2 Notación y el Símbolo de Sumatoria 1

1.3 Distribuciones de Frecuencias 9

1.4 Algebra 10

1.5 Subíndices Dobles y Sumatorias Dobles 13

1.5.1 Clasificación Cruzada 14

1.5.2 Clasificación Anidada o Jerárquica 21

1.6 El Cuadrado de Una Suma 25

1.7 Sumas de Cuadrados 28

1.7.1 Clasificación Jerárquica 28

1.7.2 Clasificación Crusada 30

CAPITULO II. VARIABLES ALEATORIAS Y PROBABILIDAD

2.1 Variables Aleatorias 33

2.2 Adición de Probabilidades 34

2.3 Multiplicación de Probabilidades 40

2.4 Muestreo con Reposición 41

2.5 Muestreo sin Reposición 43

2.6 Muestras Aleatorias Simples 46

2.7 Algunos Ejemplos del Muestreo Aleatorio Restringido 48

2.8 Muestreo en Dos Etapas 56

CAPITULO III. VALORES ESPERADOS DE VARIABLES ALEATORIAS

3.1 Introducción 62

3.2 Valor Esperado de la Suma de Dos Variables Aleatorias 66

3.3 Valor Esperado de Un Estimador 71

3.4 Variancia de Una Variable Aleatoria 76

3.4.1 Variancia de la Suma de Dos Variables Aleatorias Independientes 76

3.4.2 Variancia de la Suma de Dos Variables Aleatorias Dependientes 77

3.5 Variancia de Un Estimador 79

3.5.1 Probabilidades de Selección Iguales 79

3.5.2 Probabilidades de Selección Desiguales 82

3.6 Variancia de Una Combinación Lineal 83

3.7 Estimación de Variancias 88

3.7.1 Muestreo Aleatorio Simple 88

3.7.2 Probabilidades de Seleccion Desiguales 91

3.8 Razón Entre Dos Variables Aleatorias 92

3.9 Esperanza Condicional 95

3.10 Variancia Condicional 101

CAPITULO IV. LA DISTRIBUCION DE UN ESTIMADOR

4.1 Propiedades de Muestras Aleatorias Simples 110

4.2 Forma de la Distribución en el Muestreo 114

4.3 Diseño de Muestras 116

4.4 Error de Respuesta 122

4.5 Sesgo y Error Estándar 133

CAPITULO I. NOTACION Y SUMATORIA

1.1 INTRODUCCION

Al trabajar con grandes cantidades de datos, es necesario contar con un sistema adecuado de notación. La notación debe identificar los datos por elementos individuales y proveer expresiones matemáticas significativas para una variedad de resúmenes de datos individuales. Este capítulo describe la notación e introduce el álgebra de sumatorias, principalmente con referencia a datos derivados de encuestas por muestreo y de censos. El propósito es familiarizar a los estudiantes con la notación y la sumatoria, más que presentar conceptos estadísticos. Al principio algunas de las expresiones podrán parecer complejas o abstractas, pero no involucran más que series de operaciones de adición, substracción, multiplicación y división. Se incluyen ejercicios para que el estudiante verifique su interpretación de distintas expresiones matemáticas. También se discuten las operaciones algebraicas y se incluyen algunos ejercicios algebraicos. En gran parte, este capítulo puede considerarse como manual de ejercicios para los estudiantes que tengan interés en el muestreo pero que no estén completamente acostumbrados al uso del símbolo de sumatoria, Σ. El conocimiento del lenguaje matemático facilitará mucho el estudio del muestreo.

1.2 NOTACION Y EL SIMBOLO DE SUMATORIA

En esta monografía "elemento" se usará como expresión general para una unidad a la que corresponde una medición. Un elemento puede ser una finca, una persona, una escuela, un tallo de maíz o un animal. Tales unidades se llaman a veces "unidades de observación" o "unidades de información". Generalmente son de interés varias características de un elemento.

"Medida" y "valor" se usarán como términos generales para el valor numérico de una característica especificada de un elemento. Este uso incluye los valores asignados. Por ejemplo, el elemento podría ser una finca y la característica podría ser que en ella se cultive o no trigo. Un valor de 1 podría asignarse a una finca en que se cultiva trigo y un valor de 0 a una finca sin cultivo de trigo. Así, la "medida" o "valor" de una finca con trigo sería 1 y el valor de una finca sin trigo sería 0.

Típicamente, un conjunto de medidas de N elementos se expresará como sigue: $X_1, X_2, \ldots, X_N$, en donde X se refiere a la característica que se mide y el subíndice se refiere a los distintos elementos de la población (o conjunto). Por ejemplo, si hay N personas y la característica X es la estatura de una persona, entonces X_1 es la estatura de la primera persona, etc. Para referirse a cualquier elemento, no a un elemento específico, se usa un subíndice, "i". Así, X_i (léase "equis sub i") quiere decir el valor de X correspondiente a cualquiera de los N elementos. Una expresión común sería "X_i es el valor de X correspondiente al i-ésimo elemento".

La letra griega Σ (sigma mayúscula) generalmente se usa para indicar una suma, dándosele el nombre de "símbolo de sumatoria". Cuando se encuentra en una ecuación, quiere decir "la suma de". Por ejemplo, $\sum_{i=1}^{N} X_i$ representa la suma de todos los valores de X, desde X_1 hasta X_N inclusive; es decir, $\sum_{i=1}^{N} X_i = X_1 + X_2 + \ldots + X_N$. Los límites inferior y superior aparecen debajo y arriba del símbolo de sumatoria. Por ejemplo, para expresar la suma de X para los elementos 11 hasta 20 inclusive, se escribe $\sum_{i=11}^{20} X_i$. También pueden encontrarse notaciones tales como "ΣX_i en donde i = 1, 2, 3, ..., N", indicando que hay N elementos (o valores) en el conjunto identificados por subíndices que constan de los numerales consecutivos desde el 1 hasta el N inclusive; en caso de tratarse de una parte de un conjunto,

podría tenerse una expresión como "ΣX_i en donde i = 11, 12, ..., 20". Generalmente el subíndice del símbolo de sumatoria empieza con 1, de manera que a menudo se verá la sumatoria escrita como $\sum_i^N X_i$. Es decir, aparece únicamente el límite superior de la suma, y se entiende que la numeración empieza con i = 1. Hasta pueden omitirse las indicaciones de ambos límites en el caso de ser perfectamente entendido el conjunto de valores que se suma. Así, se entiende que $\sum_i X_i$ o ΣX_i es la suma de X, o sea el total de valores de X en el conjunto bajo consideración. A veces se omite hasta todo subíndice y se usa ΣX para indicar la suma de todos los valores de X. Se acostumbra adoptar la notación más sencilla que sirva para los propósitos. En esta monografía habrá variación intencional en la notación con el fin de familiarizar a los estudiantes con distintas maneras de representar los datos.

Un promedio generalmente se indica por la presencia de una raya sobre el símbolo. Por ejemplo, $\bar{X}$ (léase "equis raya") quiere decir el promedio de los valores de X. Así, $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$. En este caso, el señalamiento del límite superior, N, de la suma aclara el hecho de que la suma se divide por el número de elementos y que $\bar{X}$ es el promedio de todos los elementos. Sin embargo, $\frac{\Sigma X_i}{N}$ también se interpretaría como el promedio de todos los valores de X, salvo que hubiera indicación en contrario.

<u>No trate de estudiar las matemáticas sin lápiz y papel</u>. Cuando usted no entienda la escritura abreviada, ensaye escribirla en forma detallada. Esto a menudo reduce cualquier ambigüedad y ahorra tiempo.

Aquí hay unos ejemplos de "taquigrafía" matemática:

(1) Suma de los inversos de X: $\sum_{i=1}^{N} \frac{1}{X_i} = \frac{1}{X_1} + \frac{1}{X_2} + \ldots + \frac{1}{X_N}$

(2) Suma de las diferencias entre X_i y una constante, C:

$$\sum_{i=1}^{N}(X_i-C) = (X_1-C)+(X_2-C)+\ldots+(X_N-C)$$

(3) Suma de las desviaciones de X_i respecto al promedio de X:

$$\sum_{i}^{N}(X_i-\bar{X}) = (X_1-\bar{X})+(X_2-\bar{X})+\ldots+(X_N-\bar{X})$$

(4) Suma de los valores absolutos de las diferencias entre X_i y $\bar{X}$ (valor absoluto, indicado por las líneas verticales, quiere decir el valor positivo de la diferencia):

$$\Sigma|X_i-\bar{X}| = |X_1-\bar{X}|+|X_2-\bar{X}|+\ldots+|X_N-\bar{X}|$$

(5) Suma de los cuadrados de X_i: $\Sigma X_i^2 = X_1^2 + X_2^2 + X_3^2 +\ldots+ X_N^2$

(6) Suma de cuadrados de las desviaciones de X respecto a $\bar{X}$:

$$\Sigma(X_i-\bar{X})^2 = (X_1-\bar{X})^2 +\ldots+ (X_N-\bar{X})^2$$

(7) Promedio de los cuadrados de las desviaciones de X respecto a $\bar{X}$:

$$\frac{\sum_{i=1}^{N}(X_i-\bar{X})^2}{N} = \frac{(X_1-\bar{X})^2+\ldots+(X_N-\bar{X})^2}{N}$$

(8) Suma de productos de X por Y:

$$\sum_{i=1}^{N} X_iY_i = X_1Y_1 + X_2Y_2 +\ldots+ X_NY_N$$

(9) Suma de cocientes de X dividido por Y:

$$\Sigma\frac{X_i}{Y_i} = \frac{X_1}{Y_1} + \frac{X_2}{Y_2} +\ldots+ \frac{X_N}{Y_N}$$

(10) Suma de X dividida por la suma de Y:

$$\frac{\Sigma X_i}{\Sigma Y_i} = \frac{X_1 + X_2 + \ldots + X_N}{Y_1 + Y_2 + \ldots + Y_N}$$

(11) Suma de los primeros N dígitos:

$$\sum_{i=1}^{N} = 1 + 2 + 3 +\ldots+ N$$

(12)

$$\sum_{i=1}^{N} iX_i = X_1 + 2X_2 + 3X_3 + \ldots + NX_N$$

(13)

$$\sum_{i=1}^{6} (-1)^iX_i = -X_1+X_2-X_3+X_4-X_5+X_6$$

Ejercicio 1.1. Dado un conjunto de cuatro elementos con los siguientes valores de X: $X_1 = 2$, $X_2 = 0$, $X_3 = 5$, $X_4 = 7$, y para probar su entendimiento de la notación de sumatoria, calcule los valores de las siguientes expresiones algebraicas:

Expresión	Respuesta
(1) $\sum_{i=1}^{4} (X_i + 4)$	30
(2) $\Sigma 2(X_i - 1)$	20
(3) $2\Sigma(X_i - 1)$	20
(4) $\Sigma 2X_i - 1$	27
(5) $\bar{X} = \frac{\Sigma X_i}{N}$	3,5
(6) ΣX_i^2	78
(7) $\Sigma(-X_i)^2$	78
(8) $(\Sigma X_i)^2$	196
(9) $\Sigma(X_i^2 - X_i)$	64
(10) $\Sigma(X_i^2) - \Sigma X_i$	64
(11) $\Sigma i(X_i)$	45
(12) $\Sigma(-1)^i(X_i)$	0
(13) $\sum_{i=1}^{4} (X_i^2 - 3)$	66
(14) $\sum_{i=1}^{4} X_i^2 - \sum_{i=1}^{4} (3)$	66

Nota: $\sum_{i=1}^{4} (3)$ significa "la suma de cuatro valores 3".

(15) $\Sigma(X_i - \bar{X})$	0

Expresión (continuación)	Respuesta
(16) $\dfrac{\Sigma(X_i - \bar{X})^2}{N - 1}$	$\dfrac{29}{3}$
(17) $\dfrac{\Sigma(X_i^2 - 2X_i\bar{X} + \bar{X}^2)}{N - 1}$	$\dfrac{29}{3}$
(18) $\dfrac{\Sigma X_i^2 - N\bar{X}^2}{N - 1}$	$\dfrac{29}{3}$

Definición 1.1. La variancia de X cuando $X = X_1, X_2, \ldots, X_N$ se define en cualquiera de las dos maneras siguientes:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N} \qquad \text{o} \qquad S^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1}$$

La razón de la existencia de dos definiciones se explicará en el capítulo III. Las fórmulas de la variancia proveen medidas del grado en que los valores de X varían (desvían) de su promedio. La raíz cuadrada de la variancia de X se llama la "desviación estándar de X". El papel importante que juegan estas definiciones de variancia y desviación estándar en la teoría del muestreo se volverá obvio a medida que estudie muestreo. La variancia de una estimación derivada de una muestra es una de las medidas que se necesitan para juzgar la exactitud de la estimación y para evaluar diseños alternativos de muestreo. Gran parte del álgebra y de la notación en este capítulo se relaciona con el cálculo de variancias. En el caso de los planes complejos de muestreo, las fórmulas para la variancia son complejas. Este capítulo deberá ayudar a aclarar para el estudiante el lenguaje matemático usado en el muestreo.

Definición 1.2. "Población" es un término estadístico que se refiere a un conjunto de elementos del que se escoge una muestra (a menudo se usa "universo" en vez de "población").

Algunos ejemplos de poblaciones son conjuntos de fincas, tiendas, estudiantes, hogares, fabricantes y hospitales. Una definición completa de una población es una especificación detallada de los elementos que la componen. Los datos que han de recolectarse también tienen que ser definidos. Los problemas relacionados con la definición de poblaciones que han de ser investigadas deben recibir mucha atención en cursos sobre muestreo. De una población definida se escoge una muestra de elementos, se obtiene información acerca de cada elemento en la muestra y de la muestra se sacan conclusiones con respecto a la población. Casi todas las poblaciones estudiadas con encuestas por muestreo son finitas, de manera que las matemáticas y el tratamiento en esta monografía se limitan a poblaciones finitas.

Al tratar de la teoría del muestreo, es importante distinguir entre los datos que corresponden a los elementos en una muestra y los datos que corresponden a los elementos de la población entera. Muchos autores usan letras mayúsculas al referirse a la población y minúsculas al referirse a la muestra. Así, $X_1, \ldots, X_N$ representarían los valores de alguna característica, X, para los N elementos de la población, y $x_1, \ldots x_n$ representarían los valores de X que se encuentren en una muestra de n elementos. Los subíndices en $x_1, \ldots, x_n$ simplemente señalan los distintos elementos en una muestra y no corresponden a los subíndices en $X_1, \ldots, X_N$ que señalan los elementos de la población. En otras palabras, x_1 podría ser cualquiera de los X_i. De manera que

$\frac{\sum_i^N X_i}{N} = \bar{X}$ representa el promedio de X en la población y $\frac{\sum_i^n x_i}{n} = \bar{x}$ representa una media muestral. En este capítulo usaremos solamente mayúsculas, salvo en el caso de constantes y subíndices, porque el énfasis

rincipal será puesto sobre la representación simbólica de datos para un conjunto de elementos y sobre el álgebra. Para este propósito es suficiente empezar con los datos correspondientes a un conjunto de elementos, sin interesarnos en que los datos correspondan a una muestra de elementos o a todos los elementos de una población.

Las letras X, Y, Z a menudo se usan para representar distintas características (variables), a la vez que las primeras letras del alfabeto se usan comúnmente para las constantes. No hay reglas fijas con respecto a la notación. Por ejemplo, cuatro variables o características podrían llamarse X_1, X_2, X_3, X_4. En este caso, X_{1i} podría usarse para representar el i-ésimo valor de la variable X_1. Típicamente, los autores usan la notación que les convenga para sus problemas. No es práctico estandarizar completamente la notación.

<u>Ejercicio 1.2</u>. En la lista de expresiones en el ejercicio 1.1, encuentre la variancia de X; es decir, encuentre S^2. Suponga que X_4 sea 15 en vez de 7. ¿Cuánto cambia la variancia de X? (Respuesta: de $9\frac{2}{3}$ a $44\frac{1}{3}$.)

<u>Ejercicio 1.3</u>. Dados cuatro elementos con los siguientes valores de X y de Y

$X_1 = 2$	$X_2 = 0$	$X_3 = 5$	$X_4 = 7$
$Y_1 = 2$	$Y_2 = 3$	$Y_3 = 1$	$Y_4 = 14$

encuentre los valores de las siguientes expresiones:

Expresión	Respuesta	Expresión	Respuesta
(1) $\Sigma X_i Y_i$	107	(6) $\Sigma(X_i - Y_i)$	-6
(2) $(\Sigma X_i)(\Sigma Y_i)$	280	(7) $\Sigma X_i - \Sigma Y_i$	-6
(3) $\Sigma(X_i - \bar{X})(Y_i - \bar{Y})$	37	(8) $\Sigma(X_i - Y_i)^2$	74
(4) $\Sigma X_i Y_i - N\bar{X}\bar{Y}$	37	(9) $\Sigma(X_i^2 - Y_i^2)$	-132
(5) $\frac{1}{N}\Sigma\frac{X_i}{Y_i}$	1,625	(10) $\Sigma X_i^2 - \Sigma Y_i^2$	-132

Expresión	Respuesta	Expresión	Respuesta
(11) $\{\Sigma(X_i - Y_i)\}^2$	36	(12) $(\Sigma X_i)^2 - (\Sigma Y_i)^2$	-204

1.3 DISTRIBUCIONES DE FRECUENCIAS

Varios elementos en un conjunto de N podrían tener el mismo valor con respecto a alguna característica, X. Por ejemplo, muchas personas tienen la misma edad. Sean X_j una edad específica y N_j el número de personas que tienen la edad X_j en una población (conjunto) de N personas. Entonces $\sum_{j=1}^{K} N_j = N$, en donde K es el número de edades distintas en la población. Además, $\Sigma N_j X_j$ es la suma de las edades de las N personas en la población y $\frac{\Sigma N_j X_j}{\Sigma N_j}$ representa el promedio de las edades de las N personas. Una lista de X_j y N_j se llama "la distribución de frecuencias de X", porque N_j es el número de veces (la frecuencia) que la edad X_j se encuentra en la población.

O, si queremos, podemos usar X_i para representar la edad de la i-ésima persona en una población de N personas. Fíjese que la j era subíndice de edad. Ahora usamos la i como índice de individuos, y la edad media se escribiría $\frac{\Sigma X_i}{N}$. Fíjese que $\Sigma N_j X_j = \Sigma X_i$ y que $\frac{\Sigma N_j X_j}{\Sigma N_j} = \frac{\Sigma X_i}{N}$. La elección de cualquiera de estas dos representaciones simbólicas de la edad de personas en la población es cuestión de conveniencia y propósito.

<u>Ejercicio 1.4</u>. Hay 20 elementos en un conjunto (es decir, N = 20) y los valores de X para los 20 elementos son 4, 8, 3, 7, 8, 8, 3, 3, 7, 2, 8, 4, 8, 8, 3, 7, 8, 10, 3, 8.

(1) Haga una lista de los valores de X_j y N_j donde j es un índice para los valores 2, 3, 4, 7, 8, 10. Esta lista es la distribución de frecuencias de X.

(2) ¿Cuál es el valor de K?

Interprete y verifique lo siguiente mediante la ejecución de los cálculos indicados:

$$(3)\quad \sum_{i=1}^{N} X_i = \sum_{j=1}^{K} N_j X_j$$

$$(4)\quad \frac{\Sigma X_i}{N} = \frac{\Sigma N_j X_j}{\Sigma N_j} = \bar{X}$$

$$(5)\quad \frac{\Sigma (X_i - \bar{X})^2}{N} = \frac{\Sigma N_j (X_j - \bar{X})^2}{\Sigma N_j}$$

1.4 ALGEBRA

En la aritmética y el álgebra elemental, el orden seguido al sumar o multiplicar números no afecta los resultados. Las leyes bien conocidas de la aritmética, al extenderse al álgebra que involucra el símbolo de sumatoria, conducen a las siguientes reglas o teoremas importantes:

Regla 1.1 $\Sigma(X_i - Y_i + Z_i) = \Sigma X_i - \Sigma Y_i + \Sigma Z_i$

o $\Sigma(X_{1i} + X_{2i} + \ldots + X_{Ki}) = \Sigma X_{1i} + \Sigma X_{2i} + \ldots + \Sigma X_{Ki}$

Regla 1.2 $\Sigma a X_i = a \Sigma X_i$ en donde <u>a</u> es una constante

Regla 1.3 $\Sigma(X_i + b) = \Sigma X_i + Nb$ en donde <u>b</u> es una constante

Si no le es obvio que las ecuaciones dadas arriba son correctas, escriba en forma detallada ambos miembros de las igualdades y fíjese que la diferencia entre el lado izquierdo y el de la derecha es el orden en que la suma ha sido ejecutada. Fíjese que el uso de los paréntesis en la regla 1.3 significa que b aparece N veces en la suma. Es decir,

$$\sum_{i=1}^{N} (X_i + b) = (X_1 + b) + (X_2 + b) + \ldots + (X_N + b)$$
$$= (X_1 + X_2 + \ldots + X_N) + Nb$$

Con base en la regla 1.1 podemos escribir

$$\sum_{i=1}^{N} (X_i + b) = \sum_{i=1}^{N} X_i + \sum_{i=1}^{N} b$$

La expresión $\sum_{i=1} b$ quiere decir "sume el valor de b, que ocurre N veces".

Por lo tanto, $\sum_{i=1}^{N} b = Nb$.

Fíjese que, si la expresión hubiera sido $\sum_i^N X_i + b$, entonces b sería una cantidad que habría de agregarse a la suma $\sum_i^N X_i$.

En muchas ecuaciones aparecerá $\bar{X}$; por ejemplo, en $\sum_i^N \bar{X}X_i$ o en $\sum_i^N (X_i - \bar{X})$. Desde que $\bar{X}$ es constante con respecto a la suma, $\Sigma\bar{X}X_i = \bar{X}\Sigma X_i$. Así es que $\sum_i (X_i - \bar{X}) = \sum_i X_i - \sum_i \bar{X} = \sum_i X_i - N\bar{X}$. Por definición,

$$\bar{X} = \frac{\sum_i X_i}{N} \text{ ; por lo tanto, } N\bar{X} = \sum_i X_i, \text{ y } \sum_i (X_i - \bar{X}) = 0.$$

Para poder trabajar con una expresión como $\sum_i^N (X_i + b)^2$, tenemos que elevar al cuadrado la cantidad entre paréntesis antes de sumar, así:

$$\begin{aligned} \sum_i (X_i + b)^2 &= \Sigma(X_i^2 + 2bX_i + b^2) \\ &= \Sigma X_i^2 + \Sigma 2bX_i + \Sigma b^2 \qquad \text{(regla 1.1)} \\ &= \Sigma X_i^2 + 2b\Sigma X_i + Nb^2 \qquad \text{(reglas 1.2 y 1.3)} \end{aligned}$$

Verifique este resultado mediante el uso de notación detallada. Empiece con $(X_1 + b)^2 + \ldots + (X_N + b)^2$.

Es muy importante observar las reglas ordinarias del álgebra referentes al uso de los paréntesis. Los estudiantes frecuentemente cometen errores a causa de no prestarle suficiente atención a la colocación de los paréntesis o a la interpretación de ellos. Hasta que conozca bien las reglas dadas, practique la conversión de fórmulas simbólicas en expresiones detalladas y viceversa. Estudie cuidadosamente los ejemplos siguientes:

(1) $\Sigma(X_i)^2 \neq (\Sigma X_i)^2$

El miembro izquierdo de esta desigualdad es la suma de los cuadrados de X_i; el miembro derecho es el cuadrado de la suma de X_i, siendo aquí indispensable el uso de los paréntesis. El miembro izquierdo, en cambio, podría escribirse simplemente como ΣX_i^2.

(2) $\Sigma(\frac{X_i}{N})^2 = \frac{\Sigma X_i^2}{N^2}$

Se aplica la regla 1.2.

(3) $\Sigma(X_i + Y_i)^2 \neq \Sigma X_i^2 + \Sigma Y_i^2$

Hay que elevar al cuadrado la cantidad entre paréntesis antes de sumar.

(4) $\Sigma(X_i^2 + Y_i^2) = \Sigma X_i^2 + \Sigma Y_i^2$

Se aplica la regla 1.1.

(5) $\Sigma X_i Y_i \neq (\Sigma X_i)(\Sigma Y_i)$

El miembro de la izquierda es la suma de productos; el miembro de la derecha es el producto de sumas.

(6) $\Sigma(X_i - Y_i)^2 = \Sigma X_i^2 - 2\Sigma X_i Y_i = \Sigma Y^2$

(7) $\sum_i^N a(X_i - b) \neq a\sum_i^N X_i - ab$

(8) $\sum_i^N a(X_i - b) = a\sum_i^N X_i - Nab$

(9) $a(\sum_i^N X_i - b) = a\sum_i^N X_i - ab$

(10) $\Sigma X_i(X_i - Y_i) = \Sigma X_i^2 - \Sigma X_i Y_i$

<u>Ejercicio 1.5</u>. Demuestre las siguientes ecuaciones (suponiendo en todos los casos que $i = 1, 2, \ldots, N$):

(1) $\Sigma(X_i - \bar{X}) = 0$

(2) $\Sigma\frac{X_i Y_i}{X_i^2} = \Sigma\frac{Y_i}{X_i}$

(3) $N\bar{X}^2 = \frac{(\Sigma X_i)^2}{N}$

(4) $\sum_{i=1}^{N} (aX_i + bY_i + C) = a\Sigma X_i + b\Sigma Y_i + NC$

Nota: Las ecuaciones (5) y (6) serán muy familiares en el futuro.

(5) $\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - N\bar{X}^2$

(6) $\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \Sigma X_i Y_i - N\bar{X}\bar{Y}$

(7) $\Sigma(\frac{X_i}{a} + Y_i)^2 = \frac{1}{a^2} \Sigma(X_i + aY_i)^2$

(8) Sea $Y_i = a + bX_i$; demuestre que $\bar{Y} = a + b\bar{X}$ y que $\Sigma Y_i^2 = Na(a + 2b\bar{X}) + b^2\Sigma X_i^2$.

(9) Suponga que $X_i = 1$ para N_1 elementos de un conjunto, y que $X_i = 0$ para N_0 de los elementos. El número total de elementos en el conjunto es $N = N_1 + N_0$. Sean $\frac{N_1}{N} = P$, y $\frac{N_0}{N} = Q$. Demuestre que $\frac{\Sigma(X_i - \bar{X})^2}{N} = PQ$.

(10) $\Sigma(X_i - d)^2 = \Sigma(X_i - \bar{X})^2 + N(\bar{X} - d)^2$. Sugerencia: convierta $(X_i - d)^2$ en la forma $\{(X_i - \bar{X}) + (\bar{X} - d)\}^2$. Recuerde de sus estudios del álgebra elemental que $(a + b)^2 = a^2 + 2ab + b^2$, y considere a $(X_i - \bar{X})$ como a , y a $(\bar{X} - d)$ como b. ¿Para qué valor de d será mínima la cantidad $\Sigma(X_i - d)^2$?

1.5 SUBINDICES DOBLES Y SUMATORIAS DOBLES

Cuando hay más de una característica para un conjunto de elementos, las distintas características podrían distinguirse mediante el uso de una letra distinta para cada una, o mediante un subíndice. Por ejemplo, X_i e Y_i podrían representar el número de hectáreas de trigo sembrado y el número de hectáreas de trigo cosechado en la i-ésima finca. O podría usarse X_{ij}, donde i es el subíndice para las características y j el subíndice para los elementos; es decir, X_{ij} sería el valor de la característica X_i para el j-ésimo elemento. Sin embargo, cuando se van a

tratar de la misma manera los datos sobre cada una de varias características para un conjunto de elementos, podría ser innecesario usar notación que distinga las características. Por lo tanto, uno podría, digamos, calcular $\frac{\Sigma(X_i - \bar{X})^2}{N - 1}$ para todas las características.

Se necesita más de un subíndice cuando los elementos son clasificados de acuerdo con más de un criterio. Por ejemplo, X_{ij} podría representar el valor de la característica X para la j-ésima finca en el i-ésimo municipio; o X_{ijk} podría ser el valor de X para el k-ésimo hogar en la j-ésima manzana de la i-ésima ciudad. Como otro ejemplo, suponga que el procesamiento de datos para fincas involucra la clasificación de fincas por tamaño y tipo. Podríamos usar X_{ijk} para representar el valor de la característica X para la k-ésima finca en el subconjunto de fincas clasificadas como tipo j y tamaño i. Si N_{ij} es el número de fincas clasificadas como tipo j y tamaño i, entonces $\frac{\sum_{k}^{N_{ij}} X_{ijk}}{N_{ij}} = \bar{X}_{ij\cdot}$ es el promedio del valor de X para el subconjunto de fincas clasificadas como tipo j y tamaño i.

Hay dos tipos de clasificación: la cruzada y la "anidada" o jerárquica. Ambos tipos se encuentran a menudo en el mismo problema. Sin embargo, los trataremos por separado. Un ejemplo de la clasificación jerárquica es fincas dentro de municipios, municipios dentro de Estados, y Estados dentro de regiones. La clasificación cruzada consiste en exhibir los datos en dos o más dimensiones, como se demuestra en la sección que sigue.

1.5.1 CLASIFICACION CRUZADA

Como ejemplo de clasificación cruzada y de sumatoria con dos subíndices, suponga que estamos tratando de las áreas de K cultivos sembrados

en un conjunto de N fincas. Representemos con X_{ij} el área del i-ésimo cultivo en la j-ésima finca, donde i = 1, 2, ..., K, y j = 1, 2, ..., N. En este caso, los datos pueden arreglarse en una matriz de K por N, como sigue:

Fila (i)	Columna (j)					Total de Fila
	1		j		N	
1	X_{11}	...	X_{1j}	...	X_{1N}	$\sum_j X_{1j}$
	⋮		⋮		⋮	
i	X_{i1}	...	X_{ij}	...	X_{iN}	$\sum_j X_{ij}$
	⋮		⋮		⋮	
K	X_{K1}	...	X_{Kj}	...	X_{KN}	$\sum_j X_{Kj}$
Total de Columna	$\sum_i X_{i1}$		$\sum_i X_{ij}$		$\sum_i X_{iN}$	$\sum_{ij}\sum X_{ij}$

La expresión $\sum_j^N X_{ij}$ (o $\sum_j X_{ij}$) significa la suma de los valores de X_{ij} correspondientes a un valor fijo de i. De manera que, con referencia a la matriz, $\sum_j X_{ij}$ es el total de los valores de X en la i-ésima fila; o, con referencia al ejemplo de fincas y áreas bajo cultivo, $\sum_j X_{ij}$ sería el área total en todas las fincas ocupada por el i-ésimo cultivo, sea cual fuere este. De manera similar, $\sum_i^K X_{ij}$ (o $\sum_i X_{ij}$) es el total de columna para la j-ésima columna, que en el ejemplo es el total del área que dedica la j-ésima finca a los K cultivos bajo consideración. La suma de todos los valores de X podría escribirse como $\sum_{ij}^{KN}\sum X_{ij}$ o $\sum_{ij}\sum X_{ij}$.

"Sumatoria doble" quiere decir "la suma de sumas". La separación de una sumatoria doble en sus componentes puede facilitar mucho el entendimiento de ella. Aquí hay dos ejemplos:

$$(1)\quad \sum_{ij}^{KN}\sum X_{ij} = \sum_{j}^{N} X_{1j} + \sum_{j}^{N} X_{2j} + \ldots + \sum_{j}^{N} X_{Kj} \tag{1.1}$$

Con referencia a la matriz anterior, la ecuación (1.1) expresa el gran total como la suma de los totales de las filas.

$$(2)\quad \sum_{ij}^{KN}\sum X_{ij}(Y_{ij}+a) = \underbrace{\sum_{j}^{N} X_{1j}(Y_{1j}+a)} + \ldots + \sum_{j}^{N} X_{Kj}(Y_{Kj}+a) \tag{1.2}$$

$$\sum_{j}^{N} X_{1j}(Y_{1j}+a) = X_{11}(Y_{11}+a) + \ldots + X_{1N}(Y_{1N}+a)$$

En las ecuaciones (1.1) y (1.2) la suma doble está escrita como la suma de K sumas parciales; es decir, una suma parcial para cada valor de i.

<u>Ejercicio 1.6</u>. (a) Escriba una ecuación similar a la ecuación (1.1) que exprese el gran total como la suma de los totales de las columnas.

(b) La ecuación (1.2) abarca KN términos, $X_{ij}(Y_{ij}+a)$. Escriba estos términos en la forma de una matriz.

Las reglas presentadas en la sección 1.4 también son aplicables a la suma doble, como se ve:

$$\sum_{ij}^{KN}\sum X_{ij}(Y_{ij}+a) = \sum_{ij}^{KN}\sum X_{ij}Y_{ij} + a\sum_{ij}^{KN}\sum X_{ij} \tag{1.3}$$

Estudie la ecuación (1.3) con referencia a la matriz solicitada en el ejercicio 1.6(b). Para poder entender completamente la ecuación (1.3), puede ser necesario escribir los pasos intermedios para convertir el miembro del lado izquierdo de la igualdad en el del lado derecho.

Con el objeto de simplificar la notación, se usa comúnmente un sistema de notación con puntos. Por ejemplo:

$$\sum_{j} X_{ij} = X_{i\cdot}$$

$$\sum_{i} X_{ij} = X_{\cdot j}$$

$$\sum_{ij}\sum X_{ij} = X_{\cdot\cdot}$$

El punto en $X_{i.}$ indica que, además de la i, hay otro subíndice involucrado y $X_{i.}$ se interpreta como la suma de los valores de X correspondiente a un valor fijo de i. Similarmente, $X_{.j}$ es la suma de X correspondiente a un valor fijo de j, y $X_{..}$ representa una suma que abarca ambos subíndices. Como ya se ha dicho, un promedio se indica mediante una raya. Así es que $\bar{X}_{i.}$ es el promedio de X_{ij} correspondiente a un valor fijo de i; es decir, $\frac{\sum_{j=1}^{N} X_{ij}}{N} = \bar{X}_{i.}$, y $\bar{X}_{..}$ representaría el promedio de todos los valores de X_{ij}; es decir, $\frac{\sum_{ij}\sum X_{ij}}{NK}$.

Aquí hay una ilustración de como la notación de puntos puede simplificar una expresión algebraica. Suponga que se desea referirse a la suma de los cuadrados de los totales de filas en la matriz antes mencionada. Esta suma se escribiría como $\sum_i (X_{i.})^2$. La suma de cuadrados de los promedios de filas sería $\sum_i (\bar{X}_{i.})^2$. Sin la notación de puntos, las expresiones correspondientes serían $\sum_i^K (\sum_j^N X_{ij})^2$ y $\sum_i^K (\frac{\sum_j^N X_{ij}}{N})^2$. Es muy importante usar correctamente los paréntesis. Por ejemplo, la expresión $\sum_i^K (\sum_j^N X_{ij})^2$ no es la misma que $\sum_{ij}^{KN}\sum X_{ij}^2$. Incidentalmente, ¿cuál es la diferencia entre estas dos expresiones?

Usando la notación de puntos, la variancia de los promedios de filas podría escribirse como sigue:

$$V(\bar{X}_{i.}) = \frac{\sum_i^K (\bar{X}_{i.} - \bar{X}_{..})^2}{K - 1} \qquad (1.4)$$

donde V representa "variancia" y $V(\bar{X}_{i.})$ es una expresión de la variancia de $\bar{X}_{i.}$. Sin la notación de puntos o algo equivalente, una fórmula para la variancia de los promedios de filas se vería mucho más complicada.

Ejercicio 1.7. Escriba una ecuación, como la (1.4), que represente la variancia de los promedios de columnas.

Ejercicio 1.8. Dados los valores siguientes de X_{ij}

i	j			
	1	2	3	4
1	8	11	9	14
2	10	13	11	14
3	12	15	10	17

encuentre los valores de las siguientes expresiones algebraicas:

Expresión	Respuesta	Expresión	Respuesta
(1) $\sum_{j}^{N} X_{1j}$	42	(9) $K\sum_{j}^{N}(\bar{X}_{j.} - \bar{X}_{..})^2$	54
(2) $\frac{\sum_{j}^{N} X_{2j}}{N}$	12	(10) $\sum_{ij}^{KN}\sum(X_{ij} - \bar{X}_{.j} - \bar{X}_{i.} + \bar{X}_{..})^2$	6
(3) $\bar{X}_{3.}$	13,5	(11) $\sum_{ij}^{KN}\sum X_{ij}^2 - \frac{(\sum_{ij}^{KN}\sum X_{ij})^2}{KN}$	78
(4) $\sum X_{i4}$	45		
(5) $\sum_{ij}^{KN}\sum X_{ij}$	144	(12) $\frac{\sum_{i}^{K} X_{i.}^2}{N} - \frac{(\sum_{ij}^{KN}\sum X_{ij})^2}{KN}$	18
(6) $\bar{X}_{..}$	12		
(7) $\sum_{ij}^{KN}\sum(X_{ij} - \bar{X}_{..})^2$	78	(13) $\sum_{j}^{N}(X_{1j} - \bar{X}_{1.})^2$	21
(8) $N\sum_{i}^{K}(\bar{X}_{i.} - \bar{X}_{..})^2$	18	(14) $\sum_{ij}^{KN}\sum(X_{ij} - \bar{X}_{i.})^2$	60

Ilustración 1.1. Con el fin de introducir otro aspecto de notación, refiérase a la matriz de la página 15 y suponga que los valores de X en la primera fila son multiplicados por a_1, los valores de X en la segunda fila por a_2 y así sucesivamente. La matriz entonces sería:

$$
\begin{array}{ccccc}
a_1X_{11} & \cdots & a_1X_{1j} & \cdots & a_1X_{1N} \\
\vdots & & \vdots & & \vdots \\
a_iX_{i1} & \cdots & a_iX_{ij} & \cdots & a_iX_{iN} \\
\vdots & & \vdots & & \vdots \\
a_KX_{K1} & \cdots & a_KX_{Kj} & \cdots & a_KX_{KN}
\end{array}
$$

El término general puede escribirse como a_iX_{ij} porque el subíndice de $\underline{a}$ y el subíndice $\underline{i}$ en X_{ij} son el mismo. El total de los KN valores de a_iX_{ij} es $\sum_{ij}^{KN}\sum a_iX_{ij}$. Puesto que a_i es constante con respecto a la suma que involucra j, podemos colocar a_i antes del símbolo de sumatoria $\sum_j^N$. Es decir, $\sum_{ij}\sum a_iX_{ij} = \sum_i a_i \sum_j X_{ij}$.

<u>Ejercicio 1.9</u>. Vuelva a la matriz de valores de X_{ij} en el ejercicio 1.8. Suponga que $a_1 = -1$, $a_2 = 0$, $a_3 = 1$.

Calcule:

(1) $\sum_{ij}\sum a_iX_{ij}$

(2) $\sum_{ij}\sum \frac{a_iX_{ij}}{N}$

(3) $\sum_{ij}\sum a_iX_{ij}^2$ (Respuesta: -296)

Demuestre algebraicamente que

(4) $\sum_{ij}\sum a_iX_{ij} = \sum_j X_{3j} - \sum_j X_{1j}$

(5) $\sum_{ij}\sum \frac{a_iX_{ij}}{N} = \bar{X}_{3.} - \bar{X}_{1.}$

(6) $\sum_{ij}\sum a_iX_{ij}^2 = \sum_j X_{3j}^2 - \sum_j X_{1j}^2$

<u>Ejercicio 1.10</u>. Estudie la siguiente ecuación y escriba las sumas en forma detallada, en caso de que esto sea necesario para asegurarse que la ecuación es correcta:

$$\sum_{ij}^{KN}\sum(aX_{ij} + bY_{ij}) = a\sum_{ij}\sum X_{ij} + b\sum_{ij}\sum Y_{ij}$$

<u>Ilustración 1.2</u>. Suponga que

$Y_{ij} = X_{ij} + a_i + b_j + c$, donde $i = 1, 2, \ldots, K$, y $j = 1, 2, \ldots, N$.

Los valores de Y_{ij} pueden presentarse en forma de matriz como sigue:

$$\begin{array}{ccc}
Y_{11} = X_{11} + a_1 + b_1 + c & \cdots\cdots\cdots\cdots & Y_{1N} = X_{1N} + a_1 + b_N + c \\
\vdots & & \vdots \\
 & Y_{ij} = X_{ij} + a_i + b_j + c & \\
\vdots & & \vdots \\
Y_{K1} = X_{K1} + a_K + b_1 + c & \cdots\cdots\cdots\cdots & Y_{KN} = X_{KN} + a_K + b_N + c
\end{array}$$

Fíjese que a_i es una cantidad que varía de fila a fila pero es constante en cada fila; b_j varía de columna a columna pero es constante en cada columna. Al aplicar las reglas para el uso de los símbolos de sumatoria, tenemos

$$\sum_j Y_{ij} = \sum_j (X_{ij} + a_i + b_j + c)$$

$$= \sum_j X_{ij} + Na_i + \sum_j b + Nc$$

$$\sum_i Y_{ij} = \sum_i (X_{ij} + a_i + b_j + c)$$

$$= \sum_i X_{ij} + \sum_i a_i + Kb_j + Kc$$

$$\sum_{ij}\sum Y_{ij} = \sum_{ij}\sum (X_{ij} + a_i + b_j + c)$$

$$= \sum_{ij}\sum X_{ij} + N\sum_i a_i + K\sum_j b_j + KNc$$

<u>Ilustración 1.3</u>. Hemos notado que $\Sigma(X_iY_i)$ <u>no es</u> igual a $(\Sigma X_i)(\Sigma Y_i)$. Véase (1) y (2) en el ejercicio 1.3, y (5) en la página 12. Pero $\sum_{ij}\sum X_iY_j = (\sum_i X_i)(\sum_j Y_j)$, donde $i = 1, 2, \ldots, K$, y $j = 1, 2, \ldots, N$. Esto se ve

claramente si escribimos los términos de $\sum_{ij}\sum X_i Y_j$ en forma de matriz:

	Totales de filas
$X_1Y_1 + X_1Y_2 + \ldots + X_1Y_N$	$X_1\Sigma Y_j$
$+ X_2Y_1 + X_2Y_2 + \ldots + X_2Y_N$	$X_2\Sigma Y_j$
$\vdots$	
$+ X_KY_1 + X_KY_2 + \ldots + X_KY_N$	$X_K\Sigma Y_j$
Suma de totales de filas:	$\Sigma X_i \Sigma Y_j$

La columna a la derecha indica la suma de los términos en cada fila. La suma de estos totales es $X_1\Sigma Y_j+\ldots+X_K\Sigma Y_j = (X_1+\ldots+X_K)\Sigma Y_j = \Sigma X_i \Sigma Y_j$. Se podría lograr el mismo resultado final al sumar primero las columnaa. Frecuentemente las sumas intermedias son de interés primario.

Ejercicio 1.11. Verifique que $\sum_{ij}\sum X_i Y_j = (\Sigma X_i)(\Sigma Y_j)$, usando los valores de X e Y dados en el ejercicio 1.3. En el ejercicio 1.3 el subíndice de X y el de Y eran el mismo. Pero este no es el caso en la expresión $\sum_{ij}\sum X_i Y_j$.

Ejercicio 1.12. Demuestre las siguientes ecuaciones:

$$(1)\quad \sum_{ij}^{KN}\sum (a_i X_{ij} + b_j)^2 = \sum_i^K a_i^2 \sum_j^N X_{ij}^2 + 2\sum_i^K a_i \sum_j^N b_j X_{ij} + K\sum_j^N b_j^2$$

$$(2)\quad \sum_{ij}^{KN}\sum a_i (X_{ij} - \bar{X}_{i.})^2 = \sum_i^K a_i \sum_j^N X_{ij}^2 - N\sum_i^K a_i \bar{X}_{i.}^2.$$

$$(3)\quad \sum_{ij}^{KN}\sum a_i (X_{ij} - \bar{X}_{i.})(Y_{ij} - \bar{Y}_{i.}) = \sum_i^K a_i \sum_j^N X_{ij} Y_{ij} - N\sum_i^K a_i \bar{X}_{i.} \bar{Y}_{i.}$$

1.5.2 CLASIFICACION ANIDADA O JERARQUICA

La presencia de un subíndice doble no necesariamente implica que se pueda hacer una clasificación cruzada significativa de los datos. Por ejemplo, X_{ij} puede representar el valor de X correspondiente a la

j-ésima finca en el i-ésimo municipio. En este caso, la j simplemente identifica una finca dentro de un municipio. No hay correspondencia entre, digamos, la finca número 5 en un municipio y la finca número 5 en otro municipio. En verdad, el número total de fincas varía de un municipio a otro. Suponga que hay K municipios y N_i fincas en el i-ésimo municipio. El total de X correspondiente al i-ésimo municipio podría expresarse como $X_{i.} = \sum_j^{N_i} X_{ij}$. En este caso, la expresión $\sum_i X_{ij}$ no tiene significado. El total de todos los valores de X es $\sum_{ij}^{KN_i}\sum X_{ij}$.

Cuando la clasificación es jerárquica, el orden de los subíndices y el orden de los símbolos de sumatoria deben ser del orden superior de clasificación al inferior. Por esto, en el ejemplo citado arriba, el subíndice indicativo de fincas está a la derecha y el símbolo de sumatoria correspondiente a este subíndice también está a la derecha. En la expresión $\sum_{ij}^{KN_i}\sum X_{ij}$, la suma con respecto a i no puede hacerse antes de sumar con respecto a j. Pero cuando la clasificación es cruzada, las sumas se pueden hacer en cualquier orden.

En el ejemplo de K municipios y N_i fincas en el i-ésimo municipio (y en ejemplos similares), pueden considerarse los datos como arreglados en filas (o en columnas):

$$\begin{array}{llll} X_{11}, & X_{12}, & \ldots, & X_{1N_1} \\ X_{21}, & X_{22}, & \ldots, & X_{2N_2} \\ \vdots & \vdots & & \vdots \\ X_{K1}, & X_{K2}, & \ldots, & X_{KN_K} \end{array}$$

Aquí hay dos sumas dobles que han sido descompuestas para su examen:

$$(1) \quad \sum_{ij}^{KN_i}\sum (X_{ij} - \bar{X}_{..})^2 = \underbrace{\sum_j^{N_1} (X_{1j} - \bar{X}_{..})^2}_{} + \ldots + \sum_j^{N_K} (X_{Kj} - \bar{X}_{..})^2 \qquad (1.5)$$

$$\sum_j^{N_1} (X_{1j} - \bar{X}_{..})^2 = (X_{11} - \bar{X}_{..})^2 + \ldots + (X_{1N_1} - \bar{X}_{..})^2$$

La ecuación (1.5) es la suma de cuadrados de las desviaciones, $(X_{ij}-\bar{X}_{..})$, de los valores de X_{ij} respecto al promedio global. Hay $\sum_i^K N_i$ valores de X_{ij}, y $\bar{X}_{..} = \dfrac{\sum_i\sum_j^{KN_i} X_{ij}}{\sum_i^K N_i}$. Si no hubiera interés en identificar los datos por municipios, bastaría un solo subíndice. La ecuación (1.5) entonces sería $\sum_i^N (X_i - \bar{X})^2$.

$$(2)\quad \sum_i\sum_j^{KN_i}(X_{ij}-\bar{X}_{i.})^2 = \underbrace{\sum_j^{N_1}(X_{1j}-\bar{X}_{i.})^2}_{\sum_j^{N_1}(X_{1j}-\bar{X}_{1.})^2 = (X_{11}-\bar{X}_{1.})^2+\ldots+(X_{1N_1}-\bar{X}_{1.})^2} +\ldots+ \sum_j^{N_K}(X_{Kj}-\bar{X}_{K.})^2 \qquad (1.6)$$

En la ecuación (1.6) ¿reconoce $\sum_j^{N_1}(X_{1j}-\bar{X}_{1.})^2$? Involucra únicamente el subconjunto de elementos en donde i = 1; es decir, X_{11}, X_{12}, ..., X_{1N_1}. Fíjese que $\bar{X}_{1.}$ es el promedio de los valores de X en este subconjunto. Por lo tanto, $\sum_j^{N_1}(X_{1j}-\bar{X}_{1.})^2$ es la suma de los cuadrados de las desviaciones de los valores de X en este subconjunto respecto al promedio del conjunto. La suma doble es la suma de K términos, y cada uno de los K términos es una suma de cuadrados correspondiente a un subconjunto de X, siendo i el subíndice de los subconjuntos.

<u>Ejercicio 1.13</u>. Sea X_{ij} el valor de X correspondiente a la j-ésima finca en el i-ésimo municipio. Además, sea K el número de municipios y N_i el número de fincas en el i-ésimo municipio. Suponga que los valores de X son como sigue:

$X_{11} = 3$	$X_{12} = 1$	$X_{13} = 5$	
$X_{21} = 4$	$X_{22} = 6$		
$X_{31} = 0$	$X_{32} = 5$	$X_{33} = 1$	$X_{34} = 2$

Encuentre el valor de cada una de las siguientes expresiones:

Expresión	Respuesta		
(1) $\sum_{i}^{K} N_i$	9		
(2) $\sum_{i}\sum_{j}^{KN_i} X_{ij}$	27		
(3) $X_{..}$ y $\bar{X}_{..}$	27	3	
(4) $\sum_{j}^{N_1} X_{1j} = X_{1.}$	9		
(5) $X_{2.}$ y $X_{3.}$	10	8	
(6) $\bar{X}_{1.}$, $\bar{X}_{2.}$ y $\bar{X}_{3.}$	3	5	2
(7) $\frac{\sum N_i \bar{X}_{i.}}{\sum N_i}$	3		
(8) $\sum_{i}^{K}(\sum_{j}^{N_i} X_{ij})^2$ o $\sum_{i}^{K} X_i^2$	245		
(9) $\sum_{i}\sum_{j}(X_{ij} - \bar{X}_{..})^2$	36		
(10) $\sum_{j}^{N_1}(X_{1j} - \bar{X}_{i.})^2$	8		
(11) $\sum_{j}^{N_i}(X_{ij} - \bar{X}_{i.})^2$	8, 2 y 14 cuando i = 1, 2 y 3, respectivamente		
(12) $\sum_{i}\sum_{j}^{KN_i}(X_{ij} - \bar{X}_{i.})^2$	24		
(13) $\sum_{i}^{K} N_i(\bar{X}_{i.} - \bar{X}_{..})^2$	12		
(14) $\sum_{i}^{K} \frac{(\sum_{j}^{N_i} X_{ij})^2}{N_i} - \frac{(\sum_{i}\sum_{j}^{KN_i} X_{ij})^2}{\sum N_i}$	12		
(15) $\sum_{i}^{K} N_i \bar{X}_{i.}^2 - N\bar{X}_{..}^2$	12		

Las expresiones (14) y (15) del ejercicio 1.13 son representaciones simbólicas de la misma cosa. Por definición,

$$\sum_j^{N_i} X_{ij} = X_{i.}, \quad \sum_i\sum_j^{KN_i} X_{ij} = X_{..} \quad y \quad \sum_i N_i = N$$

La substitución en (14) nos da

$$\sum_i^K \frac{X_{i.}^2}{N_i} - \frac{X_{..}^2}{N} \tag{1.7}$$

También por definición, $\frac{X_{i.}}{N_i} = \bar{X}_{i.}$ y $\frac{X_{..}}{N} = \bar{X}_{..}$. Por lo tanto,

$\frac{X_{i.}^2}{N_i} = N_i\bar{X}_{i.}^2$ y $\frac{X_{..}^2}{N} = N\bar{X}^2$. Por esto, por substitución, la ecuación (1.7) se convierte en $\sum_i^K N_i\bar{X}_{i.}^2 - N\bar{X}_{..}^2$.

<u>Ejercicio 1.14</u>. Demuestre las siguientes igualdades:

(1) $\sum_i\sum_j^{KN_i} X_{i.}X_{ij} = \sum_i^K X_{i.}^2$

(2) $\sum_i\sum_j^{KN_i} \bar{X}_{i.}(X_{ij} - \bar{X}_{i.}) = 0$

(3) $\sum_i^K N_i(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_i^K N_i\bar{X}_{i.}^2 - N\bar{X}_{..}^2$

Fíjese que esto iguala las expresiones (13) y (15) del ejercicio 1.13. La demostración es parecida a la que se pidió en la parte (5) del ejercicio 1.5.

(4) $\sum_i\sum_j^{KN_i} (a_iX_{ij} - b_i)^2 = \sum_i^K a_i^2\sum_j^{N_i} X_{ij}^2 - 2\sum_i^K a_ib_iX_{i.} + \sum_i^K N_ib_i^2$

1.6 EL CUADRADO DE UNA SUMA

En estadística es a menudo necesario trabajar algebraicamente con el cuadrado de una suma. Por ejemplo,

$$(\Sigma X_i)^2 = (X_1 + X_2 + \ldots + X_N)^2 = X_1^2 + X_1X_2 + \ldots + X_2^2 + X_2X_1 + \ldots + X_N^2 + X_NX_1 + \ldots$$

Los términos en el cuadrado de la suma pueden escribirse en forma de matriz:

$$
\begin{array}{ccccccc}
X_1X_1 & X_1X_2 & \cdots & X_1X_j & \cdots & X_1X_N \\
X_2X_1 & X_2X_2 & \cdots & X_2X_j & \cdots & X_2X_N \\
\vdots & \vdots & & \vdots & & \vdots \\
X_iX_1 & X_iX_2 & \cdots & X_iX_j & \cdots & X_iX_N \\
\vdots & \vdots & & \vdots & & \vdots \\
X_NX_1 & X_NX_2 & \cdots & X_NX_j & \cdots & X_NX_N
\end{array}
$$

El término general en esta matriz es X_iX_j, en donde X_i y X_j vienen del mismo conjunto de X, a saber, $X_1, \ldots, X_N$. Por lo tanto, i y j son subíndices del mismo conjunto. Fíjese que los términos que forman la diagonal principal de la matriz son los cuadrados de los valores de X y podrían escribirse como ΣX_i^2 . Es decir, en la diagonal principal i = j, y $X_iX_j = X_iX_i = X_i^2$. Los demás términos son todos productos de algún valor de X multiplicado por algún otro valor de X. En estos términos los subíndices nunca son iguales. Por lo tanto, la suma de todos los términos que no están en la diagonal principal puede escribirse como $\sum_{i\neq j} X_iX_j$, en donde $i\neq j$ se usa para indicar que la suma incluye todos los términos en donde i no es igual a j; es decir, todos los términos que no están en la diagonal principal. De modo que hemos demostrado que $(\Sigma X_i)^2 = \Sigma X_i^2 + \sum_{i\neq j} X_iX_j$.

Observe la simetría exhibida por los términos sobre la diagonal y debajo de ella: $X_1X_2 = X_2X_1$; $X_1X_3 = X_3X_1$, etc. Cuando ocurre una simetría como esta, en vez de $\sum_{i\neq j} X_iX_j$ puede encontrar una expresión equivalente: $2\sum_{i<j} X_iX_j$. La suma de todos los términos sobre la diagonal principal es $\sum_{i<j} X_iX_j$. Debido a la simetría, la suma de los términos que están debajo de la diagonal principal es la misma. Por lo tanto, $\sum_{i\neq j} X_iX_j = 2\sum_{i<j} X_iX_j$.

Ejercicio 1.15. Escriba los términos de $(\sum_{i=1}^{4} X_i)^2 = (X_1+X_2+X_3+X_4)^2$ en forma de matriz. Considere $X_1 = 2$, $X_2 = 0$, $X_3 = 5$, $X_4 = 7$. Calcule los valores de ΣX_i^2, $2\sum_{i<j} X_i X_j$, y $(\Sigma X_i)^2$. Muestre que en este caso se verifica la igualdad $(\Sigma X_i)^2 = \Sigma X_i^2 + 2\sum_{i<j} X_i X_j$.

Un resultado importante (que utilizaremos en el capítulo III) se deriva del hecho de que

$$(\Sigma X_i)^2 = \Sigma X_i^2 + \sum_{i\neq j} X_i X_j \tag{1.8}$$

Sea $X_i = Y_i - \bar{Y}$. Substituyendo $(Y_1 - \bar{Y})$ en lugar de X_i en la ecuación (1.8) tenemos

$$\{\Sigma(Y_i - \bar{Y})\}^2 = \Sigma(Y_i - \bar{Y})^2 + \sum_{i\neq j}(Y_i - \bar{Y})(Y_j - \bar{Y})$$

Sabemos que $\{\Sigma(Y_i - \bar{Y})\}^2 = 0$, porque $\Sigma(Y_i - \bar{Y}) = 0$. Por consiguiente,

$$\Sigma(Y_i - \bar{Y})^2 + \sum_{i\neq j}(Y_i - \bar{Y})(Y_j - \bar{Y}) = 0$$

Se deduce entonces que

$$\sum_{i\neq j}(Y_i - \bar{Y})(Y_j - \bar{Y}) = -\Sigma(Y_i - \bar{Y})^2 \tag{1.9}$$

Ejercicio 1.16. Considere

$$\begin{aligned}\sum_{i\neq j}(Y_i - \bar{Y})(Y_j - \bar{Y}) &= \sum_{i\neq j}(Y_i Y_j - \bar{Y}Y_i - \bar{Y}Y_j + \bar{Y}^2)\\ &= \sum_{i\neq j} Y_i Y_j - \bar{Y}\sum_{i\neq j} Y_i - \bar{Y}\sum_{i\neq j} Y_j + \sum_{i\neq j}\bar{Y}^2\end{aligned}$$

¿Está usted de acuerdo con que $\sum_{i\neq j}\bar{Y}^2 = N(N-1)\bar{Y}^2$? Con referencia a la disposición en forma de matriz, $\bar{Y}^2$ aparece N^2 veces, pero la especificación es que $i\neq j$, de modo que no queremos contar las N veces que $\bar{Y}^2$ se encuentra en la diagonal principal. Trate de determinar los valores de $\sum_{i\neq j} X_i$ y $\sum_{i\neq j} X_j$, y entonces demuestre que $\sum_{i=j}(Y_i - \bar{Y})(Y_j - \bar{Y}) = \sum_{i\neq j} Y_i Y_j - N(N-1)\bar{Y}^2$.

Sugerencia: Considere la disposición de los datos en forma de matriz. En $\sum_{i\neq j} Y_i$, ¿cuántas veces aparece Y_1? ¿Aparece Y_2 el mismo número de veces?

1.7 SUMAS DE CUADRADOS

Por varias razones los estadísticos se interesan en los componentes de variación; es decir, la medida del monto de variación atribuible a cada una de varias fuentes. Esto involucra el cálculo de sumas de cuadrados que corresponden a las distintas fuentes de variación que son de interés. Hablaremos de un ejemplo sencillo de clasificación jerárquica y uno sencillo de clasificación cruzada.

1.7.1 CLASIFICACION JERARQUICA

Con el fin de ser algo específicos, nos referimos al ejemplo de K municipios y N_i fincas en el i-ésimo municipio. La suma de los cuadrados de las desviaciones de X_{ij} respecto a $\bar{X}_{..}$ puede dividirse en dos partes, como se ve en la siguiente ecuación:

$$\sum_{ij}^{KN_i}\sum(X_{ij}-\bar{X}_{..})^2 = \sum_{i}^{K}N_i(\bar{X}_{i.}-\bar{X}_{..})^2 + \sum_{ij}^{KN_i}\sum(X_{ij}-\bar{X}_{i.})^2 \qquad (1.10)$$

La cantidad del primer miembro de la ecuación (1.10) se llama "suma total de cuadrados". En el ejercicio 1.13, parte (9), la suma total de cuadrados era 36.

La primera cantidad del segundo miembro de la ecuación involucra los cuadrados de $(\bar{X}_{i.}-\bar{X}_{..})$, que son desviaciones de los promedios de clase (de municipios) respecto al promedio global. Se llama "suma de cuadrados interclase"; con referencia al ejemplo, se llama "suma de cuadrados intermunicipio". En el ejercio 1.13, parte (13), también se calculó, con resultado 12.

El último término de la ecuación se llama "suma de cuadrados intraclase", porque involucra las desviaciones dentro de las clases con respecto a los promedios respectivos. Ha sido presentado ya. Vea la ecuación (1.6) y la explicación que la acompaña. En el ejercicio 1.13 la suma de cuadrados

intraclase era 24, calculada en la parte (12). De modo que, en el ejercicio 1.13, tenemos la suma total de cuadrados, 36, que es igual a la suma interclase, 12, más la suma intraclase, 24. Esto verifica la ecuación (1.10).

La demostración de la ecuación (1.10) es fácil si se empieza correctamente. Escriba $X_{ij} - \bar{X}_{..} = (X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})$. Esta técnica sencilla de agregar y restar $\bar{X}_{i.}$ divide la desviación $(X_{ij} - \bar{X}_{..})$ en dos partes. La demostración sigue así:

$$\begin{aligned}
\sum_{ij}^{KN_i}\sum (X_{ij} - \bar{X}_{..})^2 &= \sum_{ij}\sum \{(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})\}^2 \\
&= \sum_{ij}\sum \{(X_{ij} - \bar{X}_{i.})^2 + 2(X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..}) + (\bar{X}_{i.} - \bar{X}_{..})^2\} \\
&= \sum_{ij}\sum (X_{ij} - \bar{X}_{i.})^2 + 2\sum_{ij}\sum (X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..}) + \sum_{ij}\sum (\bar{X}_{i.} - \bar{X}_{..})^2
\end{aligned}$$

<u>Ejercicio 1.17</u>. Demuestre que

$$\sum_{ij}^{KN_i}\sum (X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..}) = 0 \text{ y que } \sum_{ij}^{KN_i}\sum (\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i}^{K} N_i(\bar{X}_{i.} - \bar{X}_{..})^2$$

La ejecución del ejercicio 1.17 completa la demostración.

La ecuación (1.10) está escrita en una forma que muestra su significado más bien que en la forma más útil para fines de cálculo. Para fines de cálculo se usan comúnmente las siguientes relaciones:

$$\text{Total} = \sum_{ij}^{KN_i}\sum (X_{ij} - \bar{X}_{..})^2 = \sum_{ij}\sum X_{ij}^2 - N\bar{X}_{..}^2$$

$$\text{Entre clases} = \sum_{i}^{K} N_i(\bar{X}_{i.} - \bar{X}_{..})^2 = \sum_{i} N_i\bar{X}_{i.}^2 - N\bar{X}_{..}^2$$

$$\text{Dentro de clases} = \sum_{ij}^{KN_i}\sum (X_{ij} - \bar{X}_{i.})^2 = \sum_{ij}\sum X_{ij}^2 - \sum_{i} N_i\bar{X}_{i.}^2$$

$$\text{en donde } N = \sum_{i}^{K} N_i \text{ , } \bar{X}_{i.} = \frac{\sum_{j}^{N_i} X_{ij}}{N_i} \text{ y } \bar{X}_{..} = \frac{\sum_{ij}^{KN_i}\sum X_{ij}}{N}$$

Fíjese que la mayor parte de la aritmética se reduce a calcular $\sum_{i}\sum_{j}^{KN_i} X_{ij}^2$, $\sum_{i}^{K} N_i\bar{X}_{i.}^2$ y $N\bar{X}_{..}^2$. Hay alternativas que pueden usarse. Por ejemplo, se puede usar $\sum_{i}\frac{KX_{i.}^2}{N_i}$ en vez de $\sum_{i}^{K} N_i\bar{X}_{i.}^2$.

<u>Ejercicio 1.18</u>. Demuestre que

$$\sum_{i}\sum_{j}^{KN_i}(X_{ij}-\bar{X}_{i.})^2 = \sum_{i}\sum_{j}X_{ij}^2 - \sum_{i}N_i\bar{X}_{i.}^2$$

Ocurre un caso especial muy útil cuando $N_i = 2$. La suma de cuadrados intraclase resulta ser

$$\sum_{i}\sum_{j}^{K2}(X_{ij}-\bar{X}_{i.})^2 = \sum_{i}^{K}\{(X_{i1}-\bar{X}_{i.})^2 + (X_{i2}-\bar{X}_{i.})^2\}$$

Ya que $\bar{X}_{i.} = \frac{X_{i1}+X_{i2}}{2}$, es fácil demostrar que $(X_{i1}-\bar{X}_{i.})^2 = \frac{1}{4}(X_{i1}-X_{i2})^2$ y que $(X_{i2}-\bar{X}_{i.})^2 = \frac{1}{4}(X_{i1}-X_{i2})^2$. Por lo tanto, la suma de cuadrados intraclase resulta ser

$$\frac{1}{2}\sum_{i}^{K}(X_{i1}-X_{i2})^2$$

que es una forma muy conveniente para el cálculo.

1.7.2 CLASIFICACION CRUZADA

Refiérase a la matriz de la página 15 y al ejercicio 1.8. La suma total de cuadrados puede dividirse en tres partes, como se ve en la siguiente ecuación:

$$\sum_{i}\sum_{j}^{KN}(X_{ij}-\bar{X}_{..})^2 = N\sum_{i}^{K}(\bar{X}_{i.}-\bar{X}_{..})^2 + K\sum_{j}^{N}(\bar{X}_{.j}-\bar{X}_{..})^2 + \sum_{i}\sum_{j}^{KN}(X_{ij}-\bar{X}_{i.}-\bar{X}_{.j}+\bar{X}_{..})^2 \quad (1.11)$$

Refiérase al ejercicio 1.8 y encuentre la suma total de cuadrados y las tres partes. Son:

	Suma de cuadrados
Total	78
Filas	18
Columnas	54
Resto	6

Las tres partes suman el total, lo que verifica la ecuación (1.11). En el ejercicio 1.8, la suma de cuadrados llamada "resto" se calculó directamente (vea parte (10) del ejercicio 1.8). En la práctica, la suma de cuadrados "resto" generalmente se obtiene restando del total las sumas de cuadrados correspondientes a filas y a columnas.

Nuevamente, la demostración de la ecuación (1.11) no es difícil si se empieza correctamente. En este caso la desviación, $(X_{ij} - \bar{X}_{..})$, se divide en tres partes mediante la adición y substracción de $\bar{X}_{i.}$ y $\bar{X}_{.j}$:

$$(X_{ij} - \bar{X}_{..}) = (\bar{X}_{i.} - \bar{X}_{..}) + (\bar{X}_{.j} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}) \qquad (1.12)$$

<u>Ejercicio 1.19</u>. Demuestre la ecuación (1.11), elevando al cuadrado ambos miembros de la ecuación (1.12) y efectuando la suma. La demostración consiste principalmente en mostrar que las sumas de los términos que son productos (pero no cuadrados) son iguales a cero. Por ejemplo, hay que mostrar que $\sum_{i}^{K}\sum_{j}^{N}(\bar{X}_{i.} - \bar{X}_{..})(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}) = 0$.

CAPITULO II. VARIABLES ALEATORIAS Y PROBABILIDAD

2.1 VARIABLES ALEATORIAS

La palabra "aleatorio" posee una gran variedad de significados. Sin embargo, su uso en expresiones como "sucesos aleatorios", "variable aleatoria" y "muestra aleatoria" implica un procedimiento tal que la probabilidad de ocurrencia de un suceso es conocida a priori. Para seleccionar una muestra aleatoria de elementos de una población, se hace uso de tablas de números aleatorios. Hay varios modos de usar tales tablas para hacer una selección aleatoria de manera que un elemento dado tenga una probabilidad especificada de ser seleccionado.

La teoría del muestreo probabilístico se basa en el concepto de una variable aleatoria, que es una variable que, de acuerdo con el resultado del procedimiento, puede tomar cualquiera de un conjunto definido de valores. El valor de una variable aleatoria en cualquier ocasión particular es determinado por un procedimiento aleatorio de tal manera que se conoce la probabilidad de que sea igual a cada valor especificado del conjunto. Esta definición está de acuerdo con la de "muestra probabilística", que enuncia que cada elemento de la población tiene que tener una probabilidad conocida (mayor que cero) de ser seleccionado. Un propósito principal de este capítulo es presentar una introducción elemental o repaso breve de la probabilidad como base para el próximo capítulo, que tratará de los valores esperados de una variable aleatoria. Esto conduce a una base teórica del muestreo y para evaluar la exactitud de estimaciones que se hacen con los datos de una encuesta por muestreo probabilístico.

En la teoría del muestreo, generalmente empezamos con una población imaginaria de N elementos y una medida de alguna característica, X, correspondiente a cada elemento. Una representación matemática típica de

las N medidas o valores es $X_1, \ldots, X_i, \ldots, X_N$, en donde X_i es el valor de la característica X correspondiente al i-ésimo elemento. Asociada con el i-ésimo elemento hay una probabilidad, P_i, que es la probabilidad de obtener ese elemento al seleccionar aleatoriamente uno de los elementos del conjunto de N. Las P_i se llamarán "probabilidades de selección". Si cada elemento tiene la misma probabilidad de selección, $P_i = \frac{1}{N}$. No es obligatorio que las P_i sean iguales, pero vamos a especificar que cada $P_i > 0$. Al referirnos a la probabilidad de que X sea igual a X_i, usaremos $P(X_i)$ en vez de P_i.

Tenemos que darnos cuenta de la diferencia entre probabilidad de selección y probabilidad de inclusión; esta última es la probabilidad de que un elemento sea incluido en una muestra. En este capítulo, gran parte de la materia se orienta hacia las probabilidades de selección debido a su importancia en la búsqueda de valores esperados de estimaciones derivadas de muestras de varias clases.

<u>Definición 2.1</u>. Una variable aleatoria es una variable que puede ser igual a cualquier valor, X_i, en un conjunto definido, con una probabilidad de $P(X_i)$.

Cuando se selecciona aleatoriamente un elemento de una población y se mide una característica de él, el valor obtenido es una variable aleatoria. Como veremos más adelante, si se selecciona aleatoriamente una muestra de elementos de una población, la media muestral y otras cantidades calculadas de la muestra son variables aleatorias.

<u>Ilustración 2.1</u>. Uno de los ejemplos más familiares de una variable aleatoria es el número de puntos que aparecen en la cara superior de un dado después de haberlo lanzado. Este ejemplo también ilustra el concepto de probabilidad que nos interesa; es decir, la frecuencia relativa con que un suceso ocurrirá con referencia a un conjunto definido de

sucesos posibles. En el caso del dado hay seis sucesos posibles y esperamos que cada uno ocurra con la misma frecuencia, 1/6, suponiendo que el dado es lanzado un número infinito de veces, o por lo menos muchísimas veces. Implícitas en la declaración de que cada cara de un dado tiene una probabilidad de 1/6 de quedar en la parte superior hay algunas suposiciones respecto a la estructura del dado y a la naturaleza aleatoria del lanzamiento.

Las leyes aditiva y multiplicativa de probabilidad pueden expresarse de varias maneras, dependiendo del contexto en que se usen. En el muestreo, nuestro interés radica principalmente en el resultado de una sola selección aleatoria o de una serie de ellas que produce una muestra probabilística. Por lo tanto, las reglas o teoremas para la adición o la multiplicación de probabilidades se presentarán y se tratarán únicamente en el contexto del muestreo probabilístico.

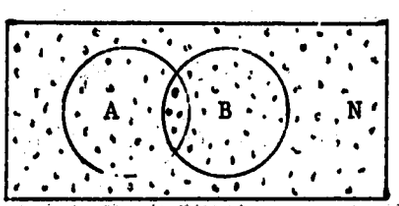
2.2 ADICION DE PROBABILIDADES

Suponga una población de N elementos y una variable X con valor X_i para el i-ésimo elemento. Es decir, tenemos un conjunto de valores de X, a saber, $X_1, \ldots, X_i, \ldots, X_N$. Sean $P_1, \ldots, P_i, \ldots, P_N$ un conjunto de probabilidades de selección, en donde P_i es la probabilidad de seleccionar el i-ésimo elemento al hacer una selección aleatoria. Especificamos que cada P_i tiene que ser mayor que cero y que $\sum_i^N P_i = 1$. Al seleccionar aleatoriamente un elemento, la probabilidad de que sea el i-ésimo elemento o el j-ésimo elemento es $P_i + P_j$. Esta regla de adición puede expresarse en forma más general. Sea P_s la suma de las probabilidades de selección correspondientes a los elementos en un subconjunto de los N elementos. Al hacer una selección aleatoria del conjunto entero, P_s es la probabilidad de que el elemento escogido pertenezca al subconjunto y

$1 - P_s$ es la probabilidad de que no pertenezca al subconjunto. Con referencia a la variable X, sea $P(X_i)$ la probabilidad de que X tome el valor X_i. Entonces $P(X_i) + P(X_j)$ representa la probabilidad de que X tome uno de los valores X_i o X_j, y $P_s(X)$ podría usarse para representar la probabilidad de que X sea igual a uno de los valores en el subconjunto.

Antes de sumar (o restar) probabilidades, debe determinarse si los sucesos son mutuamente excluyentes y si todos los sucesos posibles se han tomado en cuenta. Considere dos subconjuntos de elementos, el suconjunto A y el subconjunto B de una población de N elementos. Suponga que un elemento se selecciona aleatoriamente. ¿Cuál es la probabilidad de que el elemento pertenezca a uno de los subconjuntos A o B? Sea P(A) la probabilidad de que el elemento seleccionado sea del subconjunto A; es decir, P(A) es la suma de las probabilidades de selección correspondientes a los elementos en el subconjunto A. Se define P(B) de modo similar. Si los dos subconjuntos son mutuamente excluyentes -- lo que quiere decir que ningún elemento está en ambos subconjuntos -- la probabilidad de que el elemento seleccionado sea del subconjunto A o del subconjunto B es P(A) + P(B). Si algunos elementos están en ambos subconjuntos (vea la figura 2.1), el suceso A (el elemento seleccionado es miembro del subconjunto A) y el suceso B (el elemento seleccionado es miembro del subconjunto B) no son sucesos mutuamente excluyentes. Los elementos incluidos en ambos subconjuntos son contados una vez en P(A) y una vez en P(B). Por lo tanto, tenemos que restar P(A,B) de P(A) + P(B), siendo P(A,B) la suma de las probabilidades correspondientes a los elementos que pertenecen a ambos subconjuntos, A y B. Entonces, P(A o B) = P(A) + P(B) - P(A,B).

Figura 2.1



Para resumir, la ley aditiva de probabilidad tal como se usa arriba podría expresarse como sigue: Si A y B son subconjuntos del conjunto de todos los resultados posibles de una selección o prueba aleatoria, la probabilidad de que el resultado pertenezca al subconjunto A o al subconjunto B es igual a la probabilidad de que pertenezca a A más la probabilidad de que pertenezca a B, menos la probabilidad de que pertenezca a ambos subconjuntos, A y B.

La ley aditiva de probabilidad se extiende sin dificultad a tres o más subconjuntos. Dibuje una figura parecida a la figura 2.1, pero con tres subconjuntos, de tal manera que hayan algunos puntos comunes a los tres subconjuntos como sigue:

$$P(A \text{ o } B \text{ o } C) = P(A)+P(B)+P(C)-P(A,B)-P(A,C)-P(B,C)+P(A,B,C)$$

Como un caso para explicación adicional, suponga una población de N elementos y dos criterios para clasificarlos. Una clasificación según dos criterios podría exhibirse en la forma de la tabla 2.1:

Tabla 2.1 - Clasificación de N elementos según dos criterios

Criterio Y	Criterio X					Total
	1	...	j	...	s	
1	N_{11},P_{11}	...	N_{1j},P_{1j}	...	N_{1s},P_{1s}	$N_{1\cdot},P_{1\cdot}$
⋮	⋮		⋮		⋮	⋮
i	N_{i1},P_{i1}	...	N_{ij},P_{ij}	...	N_{is},P_{is}	$N_{i\cdot},P_{i\cdot}$
⋮	⋮		⋮		⋮	⋮
t	N_{t1},P_{t1}	...	N_{tj},P_{tj}	...	N_{ts},P_{ts}	$N_{t\cdot},P_{t\cdot}$
Total	$N_{\cdot 1},P_{\cdot 1}$	...	$N_{\cdot j},P_{\cdot j}$	...	$N_{\cdot s},P_{\cdot s}$	$N_{\cdot\cdot},P_{\cdot\cdot}$

Las columnas representan una clasificación de los elementos en términos del criterio X; las filas representan una clasificación en términos del criterio Y; N_{ij} es el número de elementos que pertenecen simultáneamente a la clase j del criterio X y a la clase i del criterio Y, mientras que P_{ij} es la

suma de las probabilidades de selección correspondientes a dichos elementos. Cada uno de los N elementos puede clasificarse en una y sólo una de las t×s casillas.

Suponga que se selecciona un elemento de la población de N. Según la ley aditiva de probabilidad podemos establecer que

$\sum_i P_{ij} = P_{\cdot j}$ es la probabilidad de que el elemento seleccionado pertenezca a la clase j del criterio X, y

$\sum_j P_{ij} = P_{i\cdot}$ es la probabilidad de que el elemento seleccionado pertenezca a la clase i del criterio Y, en donde

P_{ij} es la probabilidad de que el elemento seleccionado pertenezca simultáneamente a la clase j del criterio X y a la clase i del criterio Y.

Las probabilidades $P_{\cdot j}$ y $P_{i\cdot}$ se llaman "probabilidades marginales".

La probabilidad de que un elemento seleccionado aleatoriamente pertenezca a la clase j de X o a la clase i de Y es $P_{\cdot j} + P_{i\cdot} - P_{ij}$. (La probabilidad no es $P_{\cdot j} + P_{i\cdot}$, porque en $P_{\cdot j} + P_{i\cdot}$ hay N_{ij} elementos en la clase j de X y en la clase i de Y que se cuentan dos veces.)

Si las probabilidades de selección son iguales, $P_{ij} = \frac{N_{ij}}{N}$, $P_{\cdot j} = \frac{N_{\cdot j}}{N}$ y $P_{i\cdot} = \frac{N_{i\cdot}}{N}$.

<u>Ilustración 2.2</u>. Suponga que en una universidad hay 5.000 estudiantes de los que 1.600 están en primer año, 1.400 en segundo año y 500 que viven en el dormitorio A. De una lista de los 5.000 estudiantes se escoge aleatoriamente un estudiante. Suponiendo que cada estudiante tiene la misma probabilidad de ser seleccionado, la probabilidad de que el estudiante sea del primer año es $\frac{1600}{5000}$; que sea del segundo año es $\frac{1400}{5000}$; y que sea o del primer año o del segundo año es $\frac{1600}{5000} + \frac{1400}{5000}$. Además, la probabilidad de que el estudiante seleccionado viva en el dormitorio A es $\frac{500}{5000}$. Pero, ¿cuál es la probabilidad de que el estudiante seleccionado sea o uno de primer año o que viva en el dormitorio A? La clasificación involucra dos clasificaciones: una que se relaciona al año de

estudios y la otra a la residencia del estudiante. La información dada acerca de los 5.000 estudiantes podría tabularse como sigue:

Dormitorio	Año de estudios			Total
	Primero	Segundo	Otros	
A				500
Otro				4.500
Total	1.600	1.400	2.000	5.000

De esta tabulación es fácil ver que la respuesta a la pregunta depende de cuántos estudiantes de primer año viven en el dormitorio A. Si en la formulación del problema se hubiera dicho que 200 estudiantes de primer año viven en el dormitorio A, la respuesta habría sido $\frac{1600}{5000} + \frac{500}{5000} - \frac{200}{5000}$.

Afirmaciones sobre probabilidad tienen que hacerse e interpretarse con gran cuidado. Por ejemplo, no es correcto decir que un estudiante tiene una probabilidad de 0.1 de vivir en el dormitorio A simplemente porque 500 de los 5.000 estudiantes viven en A. Salvo que los estudiantes fueran asignados a los dormitorios mediante un procedimiento aleatorio con probabilidades conocidas, no hay base para atribuir a un estudiante una probabilidad de que viva en (haya sido asignado a) el dormitorio A. Estamos considerando el resultado de una selección aleatoria.

<u>Ejercicio 2.1</u>. Suponga que se tiene la siguiente información acerca de una población de 1.000 fincas:

600 producen maíz

500 producen soya

300 producen trigo

100 producen trigo y maíz

200 tienen una vaca por lo menos (todas las fincas que tienen vaca también producen maíz)

200 no poseen cultivos

Se selecciona aleatoriamente, con probabilidades iguales, una finca de las 1.000. ¿Cuál es la probabilidad de que la finca seleccionada

(a) produzca maíz? Respuesta: 0,6

(b) no produzca trigo?

(c) produzca maíz pero no trigo? Respuesta: 0,5

(d) produzca maíz o trigo, pero no ambos?

(e) no tenga vaca Respuesta: 0,8

(f) produzca maíz o soya?

(g) produzca maíz pero no tenga vaca? Respuesta: 0,4

(h) produzca maíz o tenga vaca, o ambas cosas?

(i) no produzca ni maíz ni trigo?

Una de las preguntas de la lista no puede contestarse.

Ejercicio 2.2. Suponga una población de 10 elementos con valores y probabilidades de selección como sigue:

Elemento	X_i	P_i	Elemento	X_i	P_i
1	2	0,05	6	11	0,15
2	7	0,10	7	2	0,20
3	12	0,08	8	8	0,05
4	0	0,02	9	6	0,05
5	8	0,20	10	3	0,10

Se selecciona aleatoriamente uno de los elementos, con probabilidad P_i. Calcule:

(a) $P(X=2)$, la probabilidad de que $X = 2$.

(b) $P(X>10)$, la probabilidad de que X sea mayor que 10.

(c) $P(X\leq 2)$, la probabilidad de que X sea menor que o igual a 2.

(d) $P(3<X<10)$, la probabilidad de que X sea mayor que 3 y menor que 10.

(e) $P(X \leq 3$ ó $X \geq 10)$, la probabilidad de que X sea menor que o igual a 3, o mayor que o igual a 10.

Nota: La respuesta a (d) y la respuesta a (e) deben sumar 1. Hasta el momento hemos tratado de la probabilidad de un suceso como resultado de una sola selección aleatoria. Cuando se hace más de una selección aleatoria simultáneamente o en sucesión, la ley multiplicativa de probabilidad es útil.

2.3 MULTIPLICACION DE PROBABILIDADES

Suponga una población de N elementos con probabilidades de selección de $P_1, \ldots, P_i, \ldots, P_N$. Cada P_i es mayor que cero y $\sum_i^N P_i = 1$. Suponga que se seleccionan dos elementos, pero que el primer elemento seleccionado se repone a la población antes de hacer la segunda selección. En este caso, el resultado de la primera selección no cambia las probabilidades de la segunda selección. Las dos selecciones (sucesos) son independientes. La probabilidad de seleccionar el i-ésimo elemento primero y el j-ésimo elemento después es P_iP_j, el producto de las probabilidades de selección P_i y P_j. Si un elemento seleccionado no es repuesto a la población antes de hacer la segunda selección, las probabilidades de selección cambian para la segunda selección. Las selecciones son dependientes.

La ley multiplicativa de probabilidad, con respecto a dos sucesos <u>independientes</u>, A y B, establece que la probabilidad conjunta de que sucedan A y B en el orden A,B es igual a la probabilidad de que suceda A multiplicada por la probabilidad de que suceda B. En forma de ecuación, $P(AB) = P(A)P(B)$. En el caso del orden B,A, $P(BA) = P(B)P(A)$, y notamos que $P(AB) = P(BA)$. Recuerde que "independencia" quiere decir que la probabilidad de ocurrencia de B no es afectada por la ocurrencia de A, y viceversa. La ley multiplicativa se extiende a cualquier número de sucesos independientes. Por lo tanto, $P(ABC) = P(A)P(B)P(C)$.

En el caso de dos sucesos dependientes, A y B, la ley multiplicativa declara que la probabilidad conjunta de que A y B sucedan en el orden A,B es igual a la probabilidad de que suceda A multiplicada por la probabilidad de que suceda B con la condición de que ya ha sucedido A. En forma de ecuación, $P(A,B) = P(A)P(B|A)$; para el orden B,A tenemos $P(BA) = P(B)P(A|B)$. La raya vertical generalmente puede traducirse como "dado" o "dado que". La notación a la izquierda de la raya se refiere al suceso bajo consideración, y la notación a la derecha, a una condición con la cual el suceso puede acontecer. $P(B|A)$ se llama "probabilidad condicional" y podría leerse "la probabilidad de B, dado A". Cuando los sucesos son independientes, $P(B|A) = P(B)$; es decir, la probabilidad condicional de que suceda B es la misma que la probabilidad no condicional de B. Extendiendo la regla multiplicativa a una serie de tres sucesos, A, B, C, que acontecen en ese orden, tenemos $P(ABC) = P(A)P(B|A)P(C|AB)$, en donde $P(C|AB)$ es la probabilidad de que suceda C, dado que ya han sucedido A y B.

2.4 MUESTREO CON REPOSICION

Cuando se saca una muestra y cada elemento seleccionado se devuelve a la población antes de hacer la siguiente selección, el método de muestreo se llama "muestreo con reposición". En este caso, el resultado de una selección no cambia las probabilidades de selección correspondientes a otra selección.

Suponga que una muestra de n elementos se selecciona con reposición. Sean $x_1, x_2, x_3, \ldots, x_n$ los valores de X en la muestra, donde x_1 es el valor de X obtenido en la primera selección, x_2 el valor obtenido en la segunda selección, etc. Fíjese que x_1 es una variable aleatoria que podría ser igual a cualquier valor en la población de valores $X_1, X_2, \ldots, X_N$,

y la probabilidad de que x_1 sea X_i es P_i. Lo mismo puede decirse de x_2, etc. Desde que las selecciones son independientes, la probabilidad de obtener una muestra de n en algún orden particular es el producto de las probabilidades de selección, a saber: $p(x_1)p(x_2)\ldots p(x_n)$, en donde $p(x_1)$ es la probabilidad P_i del elemento seleccionado en la primera tirada, $p(x_2)$ es la probabilidad P_i del elemento seleccionado en la segunda tirada, etc.

Ilustración 2.3. Como ilustración, considere una muestra de dos elementos seleccionados con probabilidades iguales y con reposición de una población de cuatro elementos. Suponga que los valores de cierta característica, X, de los cuatro elementos son X_1, X_2, X_3, y X_4. Hay 16 posibilidades:

X_1, X_1	X_2, X_1	X_3, X_1	X_4, X_1
X_1, X_2	X_2, X_2	X_3, X_2	X_4, X_2
X_1, X_3	X_2, X_3	X_3, X_3	X_4, X_3
X_1, X_4	X_2, X_4	X_3, X_4	X_4, X_4

En esta ilustración, $p(x_1)$ es igual a $\frac{1}{4}$ para cada uno de los valores (X_1, X_2, X_3, X_4) que puede tomar x_1, y lo mismo ocurre con $p(x_2)$. Por esto, cada una de las 16 posibilidades tiene una probabilidad de $(\frac{1}{4})(\frac{1}{4}) = \frac{1}{16}$.

Cada una de las 16 posibilidades es una permutación distinta que podría considerarse como una muestra. Sin embargo, como en la práctica no nos importa cuál de los dos elementos se seleccione primero, es más lógico hacer caso omiso del orden de selección. Entonces tenemos como posibles muestras y sus correspondientes probabilidades de selección las siguientes:

Muestra	Probabilidad	Muestra	Probabilidad
X_1, X_1	1/16	X_2, X_3	1/8
X_1, X_2	1/8	X_2, X_4	1/8
X_1, X_3	1/8	X_3, X_3	1/16
X_1, X_4	1/8	X_3, X_4	1/8
X_2, X_2	1/16	X_4, X_4	1/16

Note que la suma de las probabilidades es 1. Esto siempre tiene que ser el caso si todas las posibles muestras están en la lista con las probabilidades correctas. Note también que, puesto que la probabilidad (frecuencia relativa de ocurrencia) de cada muestra se conoce, el promedio de cada muestra es una variable aleatoria. En otras palabras, hay 10 posibles muestras y cualquiera de las 10 posibles medias muestrales podría haber ocurrido con la probabilidad indicada. Esto es una ilustración sencilla del hecho de que la media muestral satisface la definición de "variable aleatoria". A medida que avancemos en la teoría del muestreo, examinaremos las propiedades de la media muestral que resultan de su carácter de variable aleatoria.

Ejercicio 2.3. Con referencia a la ilustración 2.3, suponga que las probabilidades de selección son $P_1 = 1/4$, $P_2 = 1/8$, $P_3 = 3/8$, y $P_4 = 1/4$. Determine la probabilidad de cada una de las diez muestras. Recuerde que el muestreo es con reposición. Verifique sus resultados, sumando las 10 probabilidades. La suma debe ser 1. Respuesta parcial: en el caso de la muestra compuesta de los elementos 2 y 4, la probabilidad es $(\frac{1}{8})(\frac{1}{4}) + (\frac{1}{4})(\frac{1}{8}) = \frac{1}{16}$.

2.5 MUESTREO SIN REPOSICION

Cuando un elemento seleccionado no se repone a la población antes de hacer la segunda selección, el método de muestreo se llama "muestreo sin reposición". En este caso, las probabilidades de selección varían de una selección a la siguiente; es decir, las selecciones (sucesos) son dependientes.

Suponga una población de N elementos con valores de alguna característica, X, iguales a $X_1, X_2, \ldots, X_N$. Sean $P_1, \ldots, P_i, \ldots, P_N$ las probabilidades de selección en la primera selección, con cada $P_i > 0$ y

$\Sigma P_i = 1$. Suponga además que se seleccionan tres elementos, sin reposición. Sean x_1, x_2, x_3 los valores de X obtenidos en la primera, segunda y tercera selecciones aleatorias, respectivamente. ¿Cuál es la probabilidad de que $x_1 = X_5$, $x_2 = X_6$ y $x_3 = X_7$? Sea $P(X_5,X_6,X_7)$ esa probabilidad; es decir, la probabilidad de seleccionar los elementos 5, 6 y 7 en ese orden.

Según la ley multiplicativa de probabilidad para sucesos dependientes,

$$P(X_5,X_6,X_7) = P(X_5)P(X_6|X_5)P(X_7|X_5,X_6).$$

Es obvio que $P(X_5) = P_5$. Para la segunda selección, las probabilidades de selección (después de la eliminación del elemento 5) tienen que ajustarse para que sumen 1. De manera que, para la segunda selección, las probabilidades de selección son

$\frac{P_i}{1-P_5}$, $\frac{P_2}{1-P_5}$, ..., $\frac{P_N}{1-P_5}$. Es decir, $P(X_6|X_5) = \frac{P_6}{1-P_5}$.

De manera similar, $P(X_7|X_5,X_6) = \frac{P_7}{1-P_5-P_6}$. Por lo tanto,

$$P(X_5,X_6,X_7) = (P_5)(\frac{P_6}{1-P_5})(\frac{P_7}{1-P_5-P_6}) \tag{2.1}$$

Observe que $P(X_6,X_5,X_7) = (P_6)(\frac{P_5}{1-P_6})(\frac{P_7}{1-P_6-P_5})$. Por lo tanto, $P(X_5,X_6,X_7) \neq P(X_6,X_5,X_7)$ salvo que $P_5 = P_6$. En general, cada permutación de n elementos tiene una probabilidad distinta de acontecer, salvo que las P_i sean iguales. Para obtener la probabilidad exacta de seleccionar una muestra compuesta de los elementos 5, 6 y 7, sería necesario calcular la probabilidad de cada una de las seis posibles permutaciones y sumar las seis probabilidades.

Resulta oportuno aclarar que en el procedimiento efectivo de selección no es necesario calcular un nuevo conjunto de probabilidades de selección después de hacer cada selección. Haga cada selección de la misma manera en que se hizo la primera. Si se saca un elemento que ya haya sido

seleccionado, ignore el número aleatorio que resultó en la selección y siga el procedimiento de selección aleatoria hasta que se saque un elemento nuevo.

Como se ve en esta discusión muy breve, la teoría de muestreo sin reposición y con probabilidades desiguales puede ser muy compleja. Sin embargo, los textos sobre muestreo presentan maneras de evitar los problemas complejos. En realidad, es práctico y ventajoso en muchos casos usar probabilidades desiguales en el muestreo. La teoría probabilística del muestreo con probabilidades iguales y sin reposición es relativamente sencilla y se tratará con más detalle.

Ejercicio 2.4. Con una población de 4 elementos hay seis posibles muestras de dos cuando se muestrea sin reposición. Sean $P_1 = \frac{1}{4}$, $P_2 = \frac{1}{8}$, $P_3 = \frac{3}{8}$ y $P_4 = \frac{1}{4}$. Haga una lista de las seis posibles muestras y determine la probabilidad de obtener cada muestra. ¿Deben sumar 1 las probabilidades de las seis muestras? Compruebe sus resultados.

Ejercicio 2.5. Suponga que se seleccionan dos elementos de una población de 100 elementos, con reposición y con probabilidades iguales. Determine la probabilidad de que (a) el elemento número 10 no sea seleccionado; (b) el elemento número 10 sea seleccionado una sola vez; y (c) el número 10 sea seleccionado dos veces. Como verificación, las tres probabilidades deben sumar 1. ¿Por qué? Determine la probabilidad de seleccionar la combinación de los elementos 10 y 20.

Ejercicio 2.6. Refiérase al ejercicio 2.5 y cambie la especificación "con reposición" a "sin reposición". Conteste las mismas preguntas. ¿Por qué es mayor la probabilidad de lograr la combinación de los elementos 10 y 20 que la probabilidad determinada en ejercicio 2.5?

2.6 MUESTRAS ALEATORIAS SIMPLES

En la práctica, casi todas las muestras se seleccionan sin reposición. La selección de una muestra aleatoria de n elementos, con probabilidades iguales y sin reposición, de una población de N elementos se llama "muestreo aleatorio simple" ("MAS"). La selección se hace elemento a elemento; es decir, se requieren n selecciones aleatorias separadas.

Primero trataremos de la probabilidad de lograr una combinación dada de n elementos. Refiérase a la ecuación (2.1) y a la discusión que la precede. Las P_i son todas iguales a $\frac{1}{N}$ en el caso del muestreo aleatorio simple. Por lo tanto, la ecuación (2.1) en este caso se particulariza en $P(X_5,X_6,X_7) = \frac{1}{N}(\frac{1}{N-1})(\frac{1}{N-2})$. Todas las permutaciones de los tres elementos 5, 6 y 7 tienen la misma probabilidad de suceder. Hay 3! = 6 permutaciones posibles. Entonces, la probabilidad de que la muestra sea compuesta de los elementos 5, 6 y 7 es $\frac{(1)(2)(3)}{N(N-1)(N-2)}$. Cualquier combinación de tres elementos tiene la misma probabilidad de suceder. En general, todas las posibles combinaciones de n elementos tienen la misma probabilidad de selección, cuyo valor está dado por la siguiente expresión:

$$\frac{(1)(2)(3)\ldots(n)}{N(N-1)(N-2)\ldots(N-n+1)} = \frac{n!(N-n)!}{N!} \qquad (2.2)$$

Según un teorema sobre número de combinaciones, hay $\frac{N!}{n!(N-n)!}$ combinaciones (muestras) posibles de n elementos. Si cada <u>combinación</u> de n elementos tiene la misma probabilidad de ser la muestra seleccionada, la probabilidad de seleccionar una combinación especificada tiene que ser la recíproca (el inverso) del número de combinaciones. Esto está de acuerdo con la ecuación (2.2).

Una característica importante del MAS que se necesitará en el capítulo sobre valores esperados es el hecho de que el j-ésimo elemento de la población tiene la misma probabilidad que cualquier otro de ser seleccionado en

la i-ésima selección. Una expresión general de la probabilidad de qu el j-ésimo elemento de la población sea obtenido en la i-ésima selección es

$$\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)\left(\frac{N-3}{N-2}\right)\cdots\left(\frac{N-i+1}{N-i+2}\right)\left(\frac{1}{N-i+1}\right) = \frac{1}{N} \qquad (2.3)$$

Vamos a verificar la ecuación (2.3) para i = 3. La ecuación se particulariza así: $\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)\left(\frac{N-3}{N-2}\right) = \frac{1}{N}$.

La probabilidad de que el j-ésimo elemento de la población sea obtenido en la tercera selección es igual a la probabilidad de que no sea obtenido en la primera ni en la segunda selección, multiplicada por la probabilidad condicional de ser obtenido en la tercera selección, dado que no fue obtenido en la primera ni en la segunda selección. (Recuerde que el muestreo es sin reposición.) Fíjese que $\frac{N-1}{N}$ es la probabilidad de que el j-ésimo elemento no sea obtenido en la primera selección y $\frac{N-2}{N-1}$ es la probabilidad condicional de que no sea obtenido en la segunda selección. Al hacer la tercera selección, la probabilidad condicional del j-ésimo elemento es $\frac{1}{N-2}$. De manera que la probabilidad del j-ésimo elemento en la tercera selección es $\left(\frac{N-1}{N}\right)\left(\frac{N-2}{N-1}\right)\left(\frac{1}{N-2}\right) = \frac{1}{N}$. Esto verifica la ecuación (2.3) para i = 3.

En síntesis, el resultado general para cualquier tamaño de muestra es que el j-ésimo elemento en una población tiene una probabilidad igual a $\frac{1}{N}$ de ser obtenido en la i-ésima selección. Quiere decir que x_i (el valor de X obtenido en la i-ésima selección) es una variable aleatoria que tiene una probabilidad $\frac{1}{N}$ de ser igual a cualquier valor del conjunto $X_1, \ldots, X_N$.

¿Qué probabilidad tiene el j-ésimo elemento de ser incluido en una muestra de n? Acabamos de mostrar que tiene una probabilidad $\frac{1}{N}$ de ser seleccionado en la i-ésima selección. Por lo tanto, cualquier elemento dado de la población tiene n posibilidades, cada una con probabilidad $\frac{1}{N}$,

de ser incluido en una muestra. El elemento puede ser obtenido en la primera selección, o en la segunda, ..., o en la enésima selección y no puede seleccionarse dos veces porque el muestreo es sin reposición Por esta razón, las probabilidades, $\frac{1}{N}$ para cada una de las n selecciones, pueden sumarse, lo que da $\frac{n}{N}$ como la probabilidad de que cualquier elemento dado sea incluido en la muestra.

Ilustración 2.4. Suponga que se tiene una lista de 1.000 fincas que incluye algunas fincas que no son elegibles para una encuesta. No hay modo de saber de antemano si una finca incluida en la lista es inelegible. Se selecciona de la lista una muestra aleatoria simple de 200 fincas. Se visitan las 200 fincas, pero solamente las que son elegibles se incluyen en la muestra. ¿Cuál es la probabilidad que tiene una finca elegible de ser incluida en la muestra? Cada finca en la lista de 1.000 fincas tiene una probabilidad igual a $\frac{1}{5}$ de ser incluida en la muestra de 200. Todas las fincas elegibles en la muestra de 200 se incluyen en la muestra final. Por lo tanto, la respuesta es $\frac{1}{5}$.

Ejercicio 2.7. Del conjunto siguiente de 12 valores de X se va a seleccionar una muestra aleatoria simple de tres elementos: 2, 10, 5, 8, 1, 15, 7, 8, 13, 4, 6 y 2. Determine $P(\bar{x} \geq 12)$ y $P(3 < \bar{x} < 12)$. Recuerde que el número total de muestras posibles de 3 elementos puede obtenerse por aplicación de una fórmula. Puesto que cada muestra posible de tres elementos tiene la misma probabilidad de selección, puede determinar cuáles de las muestras tendrán $\bar{x} \leq 3$ ó $\bar{x} \geq 12$ sin hacer una lista de las muchas muestras posibles. Respuesta: $P(\bar{x} \geq 12) = \frac{3}{220}$; $P(\bar{x} \leq 3) = \frac{9}{220}$; $P(3 < \bar{x} < 12) = \frac{208}{220}$

2.7 ALGUNOS EJEMPLOS DEL MUESTREO ALEATORIO RESTRINGIDO

Hay muchos métodos aparte del MAS que dan a cada elemento una probabilidad igual de ser incluido en la muestra, pero algunas combinaciones de

n elementos no tienen oportunidad de ser incluidas en la muestra si no se usa MAS. Por ejemplo, se podría seleccionar cada k-ésimo elemento, empezando con un arranque aleatorio entre 1 y k inclusive. Esto se llama "muestreo sistemático". Para una muestra del 5%, k sería 20. El primer elemento para la muestra sería un número aleatorio entre 1 y 20 inclusive. Si este número es 12, entonces los números 12, 32, 52, etc. componen la muestra. Cada elemento tiene la misma probabilidad, $\frac{1}{20}$, de ser incluido en la muestra, pero hay solamente 20 combinaciones de elementos que tienen oportunidad de ser la muestra escogida. El MAS podría haber dado la misma muestra, pero es el <u>método de muestreo</u> lo que caracteriza la muestra y determina cómo se estimará el error debido al muestreo. Se puede considerar el diseño de la muestra como una cuestión de escoger un método para muestrear; es decir, de escoger restricciones que se impondrán sobre el procedimiento de selección de la muestra de tal manera que las combinaciones que tengan oportunidad de constituir la muestra escogida sean en general "mejores" que muchas de las combinaciones que pudieran resultar con MAS. A la vez, es necesario retener ciertas propiedades importantes que existen en el caso de muestras aleatorias simples. Las propiedades claves del MAS se desarrollarán en los próximos dos capítulos.

Otro método común de muestreo involucra la clasificación de todos los elementos de una población en grupos llamados "estratos", de cada uno de los cuales se selecciona una muestra. Suponga que N_i elementos pertenecen al i-ésimo estrato y que de este se selecciona una muestra aleatoria simple de n_i elementos. Esto se llama "muestreo aleatorio estratificado". Es obvio que cada elemento en el i-ésimo estrato tiene probabilidad $\frac{n_i}{N_i}$ de estar en la muestra. Si la fracción de muestreo, $\frac{n_i}{N_i}$, es la misma para todos los estratos, cada elemento de la población tiene una probabilidad igual, a saber, $\frac{n_i}{N_i}$, de estar en la muestra. Nuevamente vemos que

cada elemento de la población tiene probabilidad igual de ser escogido y de estar en la muestra escogida, pero algunas combinaciones que podrían ocurrir cuando el método es MAS no podrán suceder cuando se usa el muestreo aleatorio estratificado.

Hasta el momento, nuestra discusión se ha referido a la selección de elementos individuales, que son las unidades a que pertenecen los datos. Para los fines del muestreo, una población tiene que dividirse en partes que se llaman "unidades de muestreo". Entonces se selecciona una muestra de unidades de muestreo. Las unidades de muestreo y los elementos podrían ser idénticos. Pero muy a menudo no es práctico o no es posible usar elementos individuales como unidades de muestreo. Por ejemplo, suponga que se necesita una muestra de hogares. No existe una lista de hogares, pero hay una lista de manzanas que cubren el área que ha de ser enumerada. En este caso, se podría seleccionar una muestra de manzanas e incluir en la muestra todos los hogares de las manzanas escogidas. Las manzanas son las unidades de muestreo y los elementos son hogares. Cada elemento de la población debe pertenecer a solamente una unidad de muestreo con el fin de que la lista de unidades de muestreo tome en cuenta todos los elementos de la población sin duplicación ni omisión. Entonces, la probabilidad de seleccionar cualquier elemento dado es igual a la probabilidad de seleccionar la unidad de muestreo a que pertenece.

<u>Ilustración 2.5</u>. Suponga que una población tiene 1.800 residencias ubicadas dentro de 150 manzanas bien definidas. Hay varios planes de muestreo posibles. Podría seleccionarse una muestra aleatoria simple de 25 manzanas e incluirse en la muestra definitiva todas las residencias de las manzanas escogidas. En este caso, la fracción de muestreo es 1/6 y cada residencia tiene una probabilidad 1/6 de ser incluida en la muestra. Será esta una muestra aleatoria simple de residencias? No lo es, pero se podría

describir la muestra como una muestra aleatoria (o una muestra probabilística) y decir que cada residencia tiene una probabilidad igual de ser incluida en la muestra. Es decir, el término "muestra aleatoria simple" es aplicable a manzanas, no a residencias. Si hubiera doce residencias en cada una de las 150 manzanas, se podría seleccionar de cada manzana una muestra aleatoria simple de dos residencias. Este método, que es ejemplo de muestreo aleatorio estratificado, daría a cada residencia una probabilidad igual a 1/6 de ser incluida en la muestra.

Ilustración 2.6. Suponga que se desea una muestra de 100 adultos que viven dentro de un área especificada. No existe una lista de adultos, pero hay una lista de 4.000 residencias en el área. El plan propuesto para el muestreo consiste en seleccionar una muestra aleatoria simple de 100 residencias de la lista. Entonces el personal de campo visitará las residencias en la muestra para hacer una lista de todos los adultos que viven en cada una. Suponga que viven 220 adultos en las 100 residencias. Una muestra aleatoria simple de 100 adultos se selecciona de la lista de 220. Considere la probabilidad que tiene un adulto de la población de encontrarse en la muestra de 100 adultos.

(Debemos reconocer que la discusión que sigue hace caso omiso de problemas prácticos importantes relacionados con la definición, tal como la definición de una residencia, la definición de un adulto y la definición de "vivir en una residencia". No obstante, supongamos que las definiciones son claras, que la lista de residencias es completa, que ninguna residencia aparece en la lista más de una vez y que no existe ambigüedad respecto a la determinación de si un adulto vive o no vive en una residencia dada. Las definiciones incompletas a menudo conducen a probabilidades inexactas o a ambigüedades, lo que presenta dificultades en el análisis o en la interpretación de los resultados. Los muchos

problemas prácticos deberán explicarse en un curso aplicado sobre muestreo.)

Es obvio que la probabilidad que tiene una residencia de estar en la muestra es $\frac{1}{40}$. Entonces, cada persona en la lista de 220 tiene una probabilidad igual a $\frac{1}{40}$ de encontrarse en la lista porque, según la especificación, una persona vive en una sola residencia y la probabilidad que tiene un adulto de estar en la lista es la misma que tiene la residencia en que vive.

La segunda etapa del muestreo involucra la selección de una muestra aleatoria simple de 100 adultos de la lista de 220. La probabilidad condicional que tiene un adulto de estar en la muestra de 100 es $\frac{100}{220} = \frac{5}{11}$. Es decir, dado el hecho de que un adulto está en la lista de 220, ahora tiene la probabilidad $\frac{5}{11}$ de encontrarse en la muestra de 100.

Mantenga en mente que la probabilidad de que suceda un evento es su frecuencia relativa en ensayos repetidos. Si otra muestra fuera seleccionada de acuerdo con las especificaciones dadas, cada residencia de la población nuevamente tendría una probabilidad $\frac{1}{40}$ de estar en la muestra; pero el número de adultos en la lista no es muy probable que sea 220, de manera que la probabilidad condicional en la segunda etapa depende del número de residencias que estén en las manzanas incluidas en la muestra. ¿Tiene cada adulto la misma probabilidad de estar en la muestra? Examine el caso cuidadosamente -- su impresión inicial podría ser falsa. Cada adulto en la población tiene una probabilidad igual de estar en la lista de la primera etapa y cada adulto en la lista tiene una probabilidad igual de ser seleccionado en la segunda etapa. Pero, en términos de repetición del plan completo de muestreo, cada persona no tiene exactamente la misma probabilidad de estar en la muestra de 100. El ejercicio que sigue ayudará a aclarar la situación y es un buen ejercicio sobre probabilidades.

Ejercicio 2.8. Suponga una población de 5 residencias con los números indicados de adultos:

Residencia	Número de adultos
1	2
2	4
3	1
4	2
5	3

Se selecciona una muestra aleatoria simple compuesta de dos residencias. Luego se selecciona una muestra aleatoria simple de dos adultos de la lista combinada de adultos que viven en las dos residencias. Determine la probabilidad que tiene un adulto especificado de la residencia No.1 de encontrarse en la muestra. Respuesta: 0,19. Determine la probabilidad que tiene de estar en la muestra una adulto de la residencia No. 2. ¿Habrá relación entre la probabilidad que tiene un adulto de estar en la muestra y el número de adultos que viven en su residencia? ¿De qué manera?

Una alternativa consiste en tomar una fracción constante de los adultos en la lista en vez de un número constante de ellos. Por ejemplo, la especificación podría haber sido la de seleccionar una muestra aleatoria de $\frac{1}{2}$ de los adultos que están en la lista de la primera etapa. En este caso, con repetidas aplicaciones de las especificaciones del muestreo, la probabilidad en la segunda etapa no depende del resultado de la primera etapa, y cada adulto en la población tiene una probabilidad igual, $(\frac{1}{40})(\frac{1}{2}) = \frac{1}{80}$, de estar en la muestra. Fíjese que de acuerdo con este plan, el número de adultos en una muestra variará de muestra a muestra; de hecho, el número de adultos en la muestra es una variable aleatoria.

En algunas encuestas no es aconsejable entrevistar más de un adulto por residencia. Suponga que la primera etapa del muestreo consiste en seleccionar una muestra aleatoria simple de 100 residencias. Para la segunda etapa, considere lo siguiente: Cuando un entrevistador termina la formación de una lista de adultos en una residencia de la muestra, ha de seleccionar aleatoriamente un adulto de la lista de los que viven en la residencia, de acuerdo con un conjunto especificado de instrucciones. Entonces él entrevista al adulto seleccionado si está; si no está, regresa más tarde cuando el adulto seleccionado esté presente. ¿Qué probabilidad tiene de quedar incluido en la muestra un adulto que vive en el área? Según el teorema multiplicativo, la respuesta es $P'(D)P(A|D)$, donde $P'(D)$ es la probabilidad de que la residencia en que vive el adulto esté en la muestra y $P(A|D)$ es la probabilidad condicional de que sea seleccionado el adulto, dado que su residencia está en la muestra. Más específicamente, $P'(D) = \frac{1}{40}$ y $P(A|D) = \frac{1}{k_i}$, donde k_i es el número de adultos en la i-ésima residencia. De manera que la probabilidad que tiene un adulto de estar en la muestra, $(\frac{1}{40})(\frac{1}{k_i})$, es inversamente proporcional al número de adultos en su residencia.

<u>Ejercicio 2.9</u>. Suponga que hay cinco residencias y que en ellas viven 12 personas, como sigue:

Residencia	Personas
1	1, 2
2	3, 4, 5, 6
3	7, 8
4	9
5	10, 11, 12

1. Se selecciona una muestra de dos residencias, con probabilidades iguales y sin reposición. Todas las personas en las residencias seleccionadas se incluyen en la muestra. ¿Qué probabilidad tiene la persona No. 4

de ser incluida en la muestra? ¿La persona No. 9?

2. Suponga que se selecciona a una persona de la lista de 12, con probabilidades iguales. Del individuo seleccionado se obtiene información sobre dos características: su edad y el valor de la residencia en que vive. Sean $X_1, X_2, \ldots, X_{12}$ las edades de los 12 individuos e $Y_1, \ldots, Y_5$ los valores de las 5 residencias. Claro es que la probabilidad de seleccionar al i-ésimo individuo es $\frac{1}{12}$ y, por esto, $P(X_i) = \frac{1}{12}$. Determine las cinco probabilidades $P(Y_1), \ldots, P(Y_5)$. ¿Está usted de acuerdo con que $P(Y_3) = \frac{2}{12}$? Como verificación, $\Sigma P(Y_j)$ debe ser igual a 1.

3. Suponga que se selecciona una muestra de dos personas, con probabilidades iguales y <u>sin reposición</u>. Sean Y_{1j} el valor de Y_j obtenido en la primera selección , e Y_{2j} el valor de Y_j obtenido en la segunda selección. ¿Será $P(Y_{1j}) = P(Y_{2j})$? Es decir, la probabilidad de obtener Y_j en la segunda selección ¿es la misma que en la primera selección? Si la respuesta no le resulta evidente, refiérase a la sección 2.5.

<u>Ejercicio 2.10</u>. Se desea una pequeña muestra de estudiantes de tercer año matriculados en escuelas públicas de un Estado. El siguiente plan se presenta únicamente como ejercicio, sin consideración de que sea bueno o malo; se piensa obtener una muestra de matriculados en 10 clases de tercer año. Todos los estudiantes en las 10 clases serán incluidos en la muestra.

Paso 1. Seleccione una muestra aleatoria simple de 10 distritos escolares.

Paso 2. Dentro de cada uno de los 10 distritos escolares prepare una lista de escuelas públicas que enseñen al nivel del tercer año. Entonces escoja aleatoriamente una de las escuelas en la lista.

Paso 3. Para cada una de las 10 escuelas que resultaron del paso 2, haga una lista de las clases de tercer año y escoja al azar una clase. (Si hay una sola clase de tercer año en la escuela, se la incluye en la muestra.) Esto dará una muestra de 10 clases.

Describa las clases de tercer año en la población que tienen probabilidades relativamente pequeñas de ser seleccionadas. Defina la notación necesaria y escriba una expresión matemática que represente la probabilidad que tiene una clase de tercer año de estar en la muestra.

2.8 MUESTREO EN DOS ETAPAS

Por varias razones, los planes de muestreo a menudo emplean dos o más etapas de muestreo. Por ejemplo, puede seleccionarse una muestra de municipios y luego una muestra de fincas dentro de cada uno de los municipios obtenidos en la primera selección.

Las unidades usadas en la primera etapa del muestreo generalmente se llaman "unidades primarias de muestreo (UPM)". Las unidades usadas en la segunda etapa podrían llamarse "unidades secundarias de muestreo". Pero, debido a que se han hecho referencias frecuentes en este capítulo a "elementos de una población", las unidades de muestreo de la segunda etapa se llamarán "elementos".

En el caso sencillo de muestreo en dos etapas, cada elemento de la población es asociado con una y sola una unidad primaria de muestreo. Sean i el subíndice para las UPM y j el subíndice para los elementos dentro de cada UPM. De este modo, X_{ij} representa el valor de la característica X para el j-ésimo elemento de la i-ésima UPM. Sean, además

M = el número total de UPM,

m = el número de UPM seleccionadas para una muestra,

N_i = el número total de elementos en la i-ésima UPM,

n_i = el número de elementos de la i-ésima UPM incluidos en la muestra.

Entonces,

$\sum_i^M N_i = N$, el número total de elementos en la población, y

$\sum_i^m n_i = n$, el número total de elementos en la muestra.

Ahora considere la probabilidad de que un elemento sea seleccionado por un procedimiento de dos etapas: (1) seleccionar una UPM y (2) seleccionar un elemento dentro de la UPM obtenida en la etapa anterior. Sean

P_i = la probabilidad de seleccionar la i-ésima UPM,

$P_{j|i}$ = la probabilidad condicional de seleccionar el j-ésimo elemento en la i-ésima UPM, dado que ha sido seleccionada la i-ésima UPM, y

P_{ij} = la probabilidad global de seleccionar el j-ésimo elemento en la i-ésima UPM.

Entonces,

$$P_{ij} = P_i P_{j|i}$$

Si el producto de las dos probabilidades, P_i y $P_{j|i}$, es constante para cada elemento, entonces cada elemento de la población tiene la misma probabilidad de ser seleccionado. En otras palabras, dado un conjunto de probabilidades de selección, $P_1, \ldots, P_M$, para las UPM, se podría especificar que $P_{ij} = \frac{1}{N}$ (lo que implicaría $P_{j|i} = \frac{1}{NP_i}$), de modo que cada elemento de la población tenga igual probabilidad de selección.

<u>Ejercicio 2.11</u>. Refiérase a la tabla 2.1. Se seleccionará un elemento por un procedimiento de tres etapas, como sigue: (1) seleccionar una de las clases correspondientes al criterio Y (una fila) con probabilidad $\frac{N_i}{N}$; (2) dentro de la fila escogida, seleccionar una de las clases

correspondientes al criterio X (una columna) con probabilidad $\frac{N_{ij}}{N_{i.}}$; (3) dentro de la casilla resultante, seleccionar un elemento con probabilidades iguales. ¿Tendrá cada elemento en la población de N elementos la misma probabilidad de ser seleccionada? ¿Cuál es la probabilidad?

La probabilidad de que un elemento sea incluido en una muestra de dos etapas es dada por

$$P'_{ij} = P'_i P'_{j|i} \tag{2.4}$$

donde

P'_i = la probabilidad de que la i-ésima UPM esté en la muestra de UPM, y

$P'_{j|i}$ = la probabilidad condicional de que el j-ésimo elemento se encuentre en la muestra, dado que la i-ésima UPM ha sido seleccionada.

La probabilidad de inclusión, P'_{ij}, se discutirá muy brevemente en relación con tres casos importantes:

(1) Suponga que se selecciona aleatoriamente una muestra de UPM con probabilidades iguales y sin reposición. La probabilidad, P'_i, de estar en la muestra la i-ésima UPM es $f_1 = \frac{m}{M}$, en donde f_1 es la fracción de muestreo para las unidades de la primera etapa. Suponga que, en la segunda etapa de muestreo, dentro de cada una de las m UPM se selecciona una proporción constante, f_2, de sus elementos. Es decir, en la i-ésima UPM de la muestra se selecciona una muestra aleatoria simple de n_i del total de N_i elementos, con la condición $n_i = f_2 N_i$. Por lo tanto, la probabilidad condicional de que el j-ésimo elemento de la i-ésima UPM esté en la muestra es $P'_{j|i} = \frac{n_i}{N_i} = f_2$. Substituyendo en la ecuación (2.4), tenemos $P'_{ij} = f_1 f_2$, lo que demuestra que la probabilidad de inclusión de un elemento en la muestra es igual al producto de las fracciones de muestreo de las dos etapas. En este caso, P'_{ij} es constante y es la fracción global de muestreo.

Salvo que N_i sea igual para todas las UPM, el tamaño de la muestra, $n_i = f_2 N_i$, varía de UPM a UPM. También, como las UPM se seleccionan aleatoriamente, el tamaño total de la muestra, $n = \sum_i^m n_i = f_2 \sum_i^m N_i$, no es constante con respecto a repeticiones del plan de muestreo. En la práctica podrían ser muy indeseables las variaciones en los tamaños, n_i, de las muestras de las distintas UPM. Si existe información apropiada, es posible seleccionar las UPM con probabilidades que permitan hacer iguales los tamaños de las muestras, n_i, manteniendo un valor constante para las P_{ij}.

(2) Suponga que se selecciona una UPM con probabilidad $P_i = \frac{N_i}{N}$. Esto se llama comúnmente "muestreo con PPT (probabilidad proporcional al tamaño)". Suponga que dentro de la UPM se selecciona una muestra aleatoria simple de k elementos. (Si algunas N_i son menores que k, se pueden hacer consolidaciones para que toda UPM tenga N_i mayor que k.) Entonces

$$P_i' = \frac{N_i}{N}, \quad P_{j|i}' = \frac{k}{N_i} \quad \text{y} \quad P_{ij}' = \frac{N_i}{N}\left(\frac{k}{N_i}\right) = \frac{k}{N}$$

lo que quiere decir que cada elemento de la población tiene la misma probabilidad, $\frac{k}{N}$, de ser incluido en una muestra de k elementos.

La extensión de este esquema de muestreo a una muestra de m UPM podría toparse con las complicaciones indicadas en sección 2.5. No obstante, se dijo que existen medios para evitar esas complicaciones. Los textos sobre muestreo 1/ tratan en detalle este asunto, de manera que no lo incluiremos en esta monografía. La idea esencial es que uno puede seleccionar m UPM sin reposición de tal manera que la probabilidad de incluir la i-ésima UPM en la muestra sea $m(\frac{N_i}{N})$. Es decir, $P_i' = m(\frac{N_i}{N})$.

1/ Por ejemplo, Sample Survey Methods and Theory, por Hansen, Hurwitz y Madow, Tomo I, Capítulo 8. John Wiley and Sons. 1953.

Si se toma una muestra aleatoria simple de k elementos de cada una de las UPM seleccionadas, $P'_{j|i} = \frac{k}{N_i}$ y

$$P'_{ij} = m(\frac{N_i}{N})(\frac{k}{N_i}) = \frac{mk}{N} = \frac{n}{N}$$

De manera que, si los N_i se conocen exactamente para todas las M UPM en la población y si existe una lista de elementos en cada UPM, es posible seleccionar en dos etapas una muestra de n elementos de tal modo que k elementos de la muestra provengan de cada una de m UPM y cada elemento de la población tenga la misma probabilidad de estar en la muestra. En la práctica, sin embargo, se encuentra una de las dos situaciones siguientes: (a) No hay información acerca del número de elementos en las UPM , o (b) la información que existe es anticuada. No obstante, información anticuada acerca de varios elementos en la UPM puede ser muy útil. También es posible que exista una medida de tamaño que pudiera servir más eficientemente los propósitos del muestreo.

(3) Suponga que la característica Y se usa como medida de tamaño. Sea Y_i el valor de Y correspondiente a la i-ésima UPM de la población, y sea $P_i = \frac{Y_i}{Y}$, en donde $Y = \sum_i^M Y_i$. Se selecciona una muestra de m UPM de tal modo que $P'_i = m(\frac{Y_i}{Y})$ sea la probabilidad de que la i-ésima UPM sea incluida en la muestra.

Con respecto a la segunda etapa de muestreo, sea f_{2i} la fracción de muestreo para la selección de una muestra aleatoria simple dentro de la i-ésima UPM en la muestra. Es decir, $P'_{j|i} = f_{2i}$. Entonces,

$$P'_{ij} = m(\frac{Y_i}{Y})(f_{2i}) \qquad (2.5)$$

Al formular las especificaciones del muestreo, se podría escoger un valor fijo para P'_{ij} . En este contexto, P'_{ij} es la fracción global de muestreo, o proporción de la población que se incluirá en la muestra. Por

ejemplo, si se quiere una muestra del 5%, P'_{ij} sería 0,05. O, si se supiera que hay unos 50.000 elementos en la población y se quisiera una muestra de unos 2.000, se podría fijar $P'_{ij} = 0{,}04$. En otras palabras, si f es la fracción global de muestreo, tomaríamos P'_{ij} igual a f. También se toman decisiones con respecto a la medida de tamaño que se usará y el número, m, de UPM que han de seleccionarse. En la ecuación (2.5), esto deja por determinar f_{2i}. Se calcula f_{2i} para cada UPM en la muestra de la manera siguiente:

$$f_{2i} = \frac{fY}{mY_i}$$

El uso de las fracciones de muestreo f_{2i} en la segunda etapa de muestreo dará a cada elemento de la población una probabilidad igual a f de ser incluido en la muestra. Una muestra en que cada elemento de la población tiene igual probabilidad de inclusión se llama a menudo una "muestra autoponderada".

CAPITULO III. VALORES ESPERADOS DE VARIABLES ALEATORIAS

3.1 INTRODUCCION

La teoría de valores esperados de variables aleatorias se usa mucho en la teoría del muestreo; en realidad, es la base de la teoría del muestreo. Las interpretaciones de la exactitud de estimaciones derivadas de muestras probabilísticas dependen en gran parte de la teoría de valores esperados.

En el capítulo II se trató de la definición de una variable aleatoria. Es una variable que puede tomar (ser igual a) cualquiera de un conjunto definido de valores con probabilidad conocida. Sean X_i el valor de X para el i-ésimo elemento en un conjunto de N elementos y P_i la probabilidad de que sea seleccionado el i-ésimo elemento por alguna operación aleatoria de modo que se conozca de antemano P_i. ¿Cuál es el valor esperado de X?

<u>Definición 3.1</u>. El valor esperado de una variable aleatoria X es $\sum_{i=1}^{N} P_i X_i$, donde $\sum_{i=1}^{N} P_i = 1$. La notación matemática que representa el valor esperado de X es E(X). De modo que, por definición, $E(X) = \sum_{i=1}^{N} P_i X_i$.

Observe que $\Sigma P_i X_i$ es un promedio ponderado de los valores de X, con las probabilidades de selección como ponderaciones. "Valor esperado" es una expresión que significa lo mismo que "valor medio". En otras palabras, E quiere decir "el valor medio de" o "determine el valor medio de" la expresión que siga a E. Por ejemplo, $E(X^2)$ (léase "el valor esperado de X^2) se refiere al promedio de los cuadrados de los valores que toma X. Es decir, por definición,

$$E(X^2) = \sum_{i=1}^{N} P_i X_i^2$$

Si todos los N elementos tienen la misma oportunidad de ser seleccionado, todo valor de P_i tiene que ser igual a $\frac{1}{N}$ a causa del requisito de que $\Sigma P_i = 1$. En este caso, $E(X) = \sum_{i=1}^{N} (\frac{1}{N})X_i = \frac{\Sigma X_i}{N} = \bar{X}$, que es el promedio de X para todos los N elementos.

<u>Ilustración 3.1</u>. Suponga 12 elementos con los siguientes valores de X:

$X_1 = 3$	$X_5 = 5$	$X_9 = 10$
$X_2 = 9$	$X_6 = 3$	$X_{10} = 3$
$X_3 = 3$	$X_7 = 4$	$X_{11} = 8$
$X_4 = 5$	$X_8 = 3$	$X_{12} = 4$

En este conjunto, $E(X) = \frac{3 + 9 + \ldots + 4}{12} = 5$, suponiendo que cada elemento tiene la misma probabilidad de selección. O, contando el número de veces que se repite cada valor diferente de X, se obtiene una distribución de frecuencias como sigue:

X_j	N_j
3	5
4	2
5	2
8	1
9	1
10	1

donde los X_j son los valores diferentes de X, y N_j es el número de veces que aparece X_j. Hemos notado en el capítulo II que $\Sigma N_j = N$, $\Sigma N_j X_j = \Sigma X_i$ y que $\frac{\Sigma N_j X_j}{\Sigma N_j} = \frac{\Sigma X_i}{N} = \bar{X}$. Suponga que uno de los valores X_j es seleccionado aleatoriamente con probabilidad igual a P_j, donde $P_j = \frac{N_j}{\Sigma N_j} = \frac{N_j}{N}$. ¿Cuál es el valor esperado de X? Por definición, $E(X) = \Sigma P_j X_j = \Sigma \frac{N_j}{N} X_j = \frac{\Sigma N_j X_j}{N} = \bar{X}$. El estudiante puede verificar que, en esta ilustración, E(X)=5.

Note que las especificaciones de selección equivalían a la selección aleatoria, con probabilidades iguales, de uno de los 12 elementos.

Incidentalmente, una distribución de frecuencias y una distribución de probabilidades son muy parecidas. La distribución de probabilidades de X sería:

X_j	P_j
3	5/12
4	2/12
5	2/12
8	1/12
9	1/12
10	1/12

Los 12 valores, $P_i = \frac{1}{N}$, para los 12 elementos son también una distribución de probabilidades. Esta ilustración muestra dos maneras de tratar el conjunto de 12 elementos.

Al determinar valores esperados, tenga la seguridad de entender las definiciones del conjunto de valores que la variable aleatoria podría tomar y las probabilidades involucradas.

<u>Definición 3.2</u>. Cuando X es una variable aleatoria, el valor esperado de una función de X es, por definición,

$$E\{f(X)\} = \sum_{i=1}^{N} P_i\{f(X_i)\}$$

Algunos ejemplos de funciones sencillas de X son: $f(X) = aX$; $f(X) = X^2$; $f(X) = a + bX + cX^2$; $f(X) = (X - \bar{X})^2$. Para cada valor, X_i, en un conjunto definido, hay un valor correspondiente de $f(X_i)$.

<u>Ilustración 3.2</u>. Suponga que $f(X) = 2X + 3$. Con referencia al conjunto de 12 elementos analizado anteriormente, hay 12 valores de $f(X_i)$:

$$f(X_1) = 2(3) + 3 = 9$$

$$f(X_2) = 2(9) + 3 = 21$$

$$\vdots$$

$$f(X_{12}) = 2(4) + 3 = 11$$

Suponiendo que $P_i = \frac{1}{N}$, el valor esperado de $f(X) = 2X + 3$ sería

$$E(2X + 3) = \sum_{i=1}^{12} \frac{1}{N}(2X + 3) = (\frac{1}{12})(9)+(\frac{1}{12})(21)+\ldots+(\frac{1}{12})(11) = 13.$$

En términos algebraicos, para $f(X) = aX + b$, tenemos

$$E(aX + b) = \sum_{i=1}^{N} P_i(aX_i + b) = \Sigma P_i(aX_i) + \Sigma P_i b \tag{3.1}$$

Por definición, $\Sigma P_i(aX_i) = E(aX)$, y $\Sigma P_i b = E(b)$. Por lo tanto,

$$E(aX + b) = E(aX) + E(b) \tag{3.2}$$

Pues que b es constante y $\Sigma P_i = 1$, $\Sigma P_i b = b$, lo que conduce al primer teorema importante en valores esperados:

<u>Teorema 3.1</u>. El valor esperado de una constante es igual a la constante: $E(a) = a$.

Por definición, $E(aX) = \Sigma P_i(aX_i) = a\Sigma P_i X_i$. Puesto que $\Sigma P_i X_i = E(X)$, tenemos otro teorema importante:

<u>Teorema 3.2</u>. El valor esperado de una constante multiplicada por una variable es igual a la constante multiplicada por el valor esperado de la variable: $E(aX) = aE(X)$.

Aplicando estos dos teoremas a la ecuación (3.2), tenemos $E(aX + b) = aE(X) + b$. Con relación a la ilustración 3.2, tenemos $E(2X + 3) = 2E(X) + 3 = 2(5) + 3 = 13$, que es el mismo resultado obtenido en el ejemplo numérico que presentamos inmediatamente antes de la ecuación (3.1).

Ejercicio 3.1. Suponga que una variable aleatoria puede tomar cualquiera de los siguientes cuatro valores con las probabilidades indicadas:

$X_1 = 2$	$X_2 = 5$	$X_3 = 4$	$X_4 = 6$
$P_1 = 2/6$	$P_2 = 2/6$	$P_3 = 1/6$	$P_4 = 1/6$

(a) Determine $E(X)$. Respuesta: 4.

(b) Determine $E(X^2)$. Respuesta: $18\frac{1}{3}$. Note que $E(X^2) \neq \{E(X)\}^2$.

(c) Determine $E(X-\bar{X})$. Respuesta: 0. Note: Por definición,

$$E(X-\bar{X}) = \sum_{i=1}^{4} P_i(X_i - \bar{X}).$$

(d) Determine $E(X-\bar{X})^2$ * Respuesta: $2\frac{1}{3}$. Note: Por definición,

$$E(X-\bar{X})^2 = \sum_{i=1}^{4} P_i(X_i - \bar{X})^2.$$

Ejercicio 3.2. Del siguiente conjunto de tres valores de Y se ha de seleccionar un valor con una probabilidad P_i':

$Y_1 = 12$	$Y_2 = 2$	$Y_3 = 4$
$P_1' = 1/4$	$P_2' = 2/4$	$P_3' = 1/4$

(a) Determine $E(Y)$. Respuesta: $1\frac{1}{2}$.

(b) Determine $E(\frac{1}{Y})$ Respuesta: 3/16. Note: $\frac{1}{E(Y)} \neq E(\frac{1}{Y})$.

(c) Determine $E(Y-\bar{Y})^2$. Respuesta: $4\frac{3}{4}$.

3.2 VALOR ESPERADO DE LA SUMA DE DOS VARIABLES ALEATORIAS

La suma de dos o más variables aleatorias es también una variable aleatoria. Si X e Y son dos variables aleatorias, el valor esperado de X + Y es igual al valor esperado de X más el valor esperado de Y; es decir, $E(X + Y) = E(X) + E(Y)$. Dos ilustraciones numéricas ayudarán a aclarar la situación.

Ilustración 3.3. Considere las dos variables aleatorias X e Y de los ejercicios 3.1 y 3.2:

---

* $E(X-\bar{X})^2$ es una notación más conveniente que $E\{(X-\bar{X})^2\}$

$X_1 = 2$	$P_1 = 2/6$	$Y_1 = -2$	$P'_1 = 1/4$
$X_2 = 5$	$P_2 = 2/6$	$Y_2 = 2$	$P'_2 = 2/4$
$X_3 = 4$	$P_3 = 1/6$	$Y_3 = 4$	$P'_3 = 1/4$
$X_4 = 6$	$P_4 = 1/6$		

Suponga que se seleccionan un elemento del primer conjunto y un elemento del segundo conjunto, con las probabilidades indicadas arriba. ¿Cuál es el valor esperado de X + Y? La probabilidad conjunta de obtener X_i e Y_j es $P_iP'_j$, porque las dos selecciones son independientes. (Vea la página 40 para una explicación de "independencia.") Entonces por definición,

$$E(X + Y) = \sum_{i=1}^{4} \sum_{j=1}^{3} P_iP'_j(X_i + Y_j) \tag{3.3}$$

Los valores posibles de X + Y y la probabilidad de cada uno son:

X + Y	$P_iP'_j$	X + Y	$P_iP'_j$
$X_1 + Y_1 = 0$	$P_1P'_1 = 2/24$	$X_3 + Y_1 = 2$	$P_3P'_1 = 1/24$
$X_1 + Y_2 = 4$	$P_1P'_2 = 4/24$	$X_3 + Y_2 = 6$	$P_3P'_2 = 2/24$
$X_1 + Y_3 = 6$	$P_1P'_3 = 2/24$	$X_3 + Y_3 = 8$	$P_3P'_3 = 1/24$
$X_2 + Y_1 = 3$	$P_2P'_1 = 2/24$	$X_4 + Y_1 = 4$	$P_4P'_1 = 1/24$
$X_2 + Y_2 = 7$	$P_2P'_2 = 4/24$	$X_4 + Y_2 = 8$	$P_4P'_2 = 2/24$
$X_2 + Y_3 = 9$	$P_2P'_3 = 2/24$	$X_4 + Y_3 = 10$	$P_4P'_3 = 1/24$

Como verificación, la suma de las probabilidades tiene que ser 1 si todas las sumas aparecen en la lista y la probabilidad de cada una ha sido determinada correctamente. Substituyendo los valores de $X_i + Y_j$ y $P_iP'_j$ en la ecuación (3.3) obtenemos 5,5 para el valor esperado de X + Y, como sigue:

$$(\tfrac{2}{24})(0) + (\tfrac{4}{24})(4) + \ldots + (\tfrac{1}{24})(10) = 5{,}5$$

De los ejercicios 3.1 y 3.2 tenemos E(X) = 4 y E(Y) = 1,5. Por lo tanto, E(X) + E(Y) = 4 + 1,5 = 5,5, lo que verifica la afirmación anterior de que E(X + Y) = E(X) + E(Y).

Ilustración 3.4. Suponga que se selecciona una muestra aleatoria de dos elementos con reposición de la población de cuatro elementos usada en ejercicio 3.1. Sean x_1 el primer valor seleccionado y x_2 el segundo. Entonces x_1 y x_2 son variables aleatorias y su suma es una variable aleatoria. Los valores posibles de $x_1 + x_2$ y la probabilidad de cada suma, $P(x_1,x_2)$, aparecen abajo. Observe que cada orden posible de selección se trata separadamente.

x_1	x_2	$P(x_1,x_2)$	$x_1 + x_2$	x_1	x_2	$P(x_1,x_2)$	$x_1 + x_2$
X_1	X_1	4/36	4	X_3	X_1	2/36	6
X_1	X_2	4/36	7	X_3	X_2	2/36	9
X_1	X_3	2/36	6	X_3	X_3	1/36	8
X_1	X_4	2/36	8	X_3	X_4	1/36	10
X_2	X_1	4/36	7	X_4	X_1	2/36	8
X_2	X_2	4/36	10	X_4	X_2	2/36	11
X_2	X_3	2/36	9	X_4	X_3	1/36	10
X_2	X_4	2/36	11	X_4	X_4	1/36	12

Por definición, $E(x_1 + x_2)$ es

$$\frac{4}{36}(4) + \frac{4}{36}(7) + \frac{2}{36}(6) + \ldots + \frac{1}{36}(12) = 8$$

En el ejercicio 3.1 encontramos que $E(X) = 4$. Desde que x_1 es la misma variable aleatoria que X, $E(x_1) = 4$. También x_2 es la misma variable aleatoria que X, y $E(x_2) = 4$. Entonces, $E(x_1) + E(x_2) = 8$, lo que verifica que $E(x_1 + x_2) = E(x_1) + E(x_2)$.

En general, si X e Y son dos variables aleatorias, en donde X toma los valores $X_1, \ldots, X_N$, e Y toma los valores $Y_1, \ldots, Y_M$, entonces $E(X+Y) = E(X) + E(Y)$. La demostración es como sigue: Por definición,

$E(X + Y) = \sum_{ij}^{NM}\sum P_{ij}(X_i + Y_j)$, donde P_{ij} es la probabilidad de obtener la

suma $X_i + Y_j$, y $\Sigma\Sigma P_{ij} = 1$. La sumatoria doble incluye todos los valores posibles de $P_{ij}(X_i + Y_j)$. Según las reglas aplicables a las sumatorias, podemos escribir:

$$\sum_{ij}^{NM}\sum P_{ij}(X_i + Y_j) = \sum_{ij}^{NM}\sum P_{ij}X_i + \sum_{ij}^{NM}\sum P_{ij}Y_j \qquad (3.4)$$

En el primer término del segundo miembro, X_i es constante con respecto a la suma en j; en el segundo término, Y_j es constante con respecto a la suma en i. Por lo tanto, el segundo miembro de la ecuación (3.4) puede escribirse como

$$\sum_i^N X_i \sum_j^M P_{ij} = \sum_j^M Y_j \sum_i^N P_{ij}$$

Y puesto que $\sum_j^M P_{ij} = P_i$, y $\sum_i^N P_{ij} = P_j$, la ecuación (3.4) queda

$$\sum_{ij}^{NM}\sum P_{ij}(X_i + Y_j) = \sum_i^N X_i P_i + \sum_j^M Y_j P_j$$

Por definición, $\sum_i^N X_i P_i = E(X)$, y $\sum_j^M Y_j P_j = E(Y)$.

De manera que $E(X + Y) = E(X) + E(Y)$.

Si la demostración no le resulta evidente, escriba los valores de $P_{ij}(X_i + Y_j)$ en forma de matriz. Entonces, siga los pasos de la suma en la demostración.

Es resultado dado arriba se extiende a cualquier número de variables aleatorias; es decir, el valor esperado de una suma de variables aleatorias es la suma de los valores esperados de cada una. En efecto, hay un teorema muy importante que es aplicable a una combinación lineal de variables aleatorias.

<u>Teorema 3.3</u>. Sea $u = a_1u_1 + \ldots + a_ku_k$, donde $u_1, \ldots, u_k$ son variables aleatorias y $a_1, \ldots, a_k$ son constantes. Entonces

$E(u) = a_1E(u_1) + \ldots + a_kE(u_k)$, o, empleando sumatoria,

$$E(u) = E\sum_i^k a_i u_i = \sum_i^k a_i E(u_i)$$

La generalidad del teorema 3.3 es impresionante. Por ejemplo, con respecto al muestreo de una población $X_1, \ldots, X_N$, u_1 podría ser el valor de X obtenido en la primera selección, u_2 el valor obtenido en la segunda selección, etc. Las constantes podrían ser ponderaciones. Entonces, en este caso, u sería un promedio ponderado de las medidas tomadas de la muestra. O, suponga que $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ son promedios de una muestra aleatoria para k grupos distintos de edades. Los promedios son variables aleatorias y el teorema podría aplicarse a cualquier combinación lineal de promedios. En realidad, u_i podría ser cualquier función de variables aleatorias. Es decir, la única condición que sirve de base para el teorema es que u_i sea una variable aleatoria.

Ilustración 3.5. Suponga que deseamos encontrar el valor esperado de $(X + Y)^2$, donde X e Y son variables aleatorias. Antes de poder aplicar el teorema 3.3, tenemos que elevar $(X + Y)$ al cuadrado. Entonces, $E(X + Y)^2 = E(X^2 + 2XY + Y^2)$. La aplicación del teorema 3.3 da $E(X+Y)^2 = E(X^2) + 2E(XY) + E(Y^2)$.

Ilustración 3.6. Ahora mostraremos que

$$E(X-\bar{X})(Y-\bar{Y}) = E(XY) - \bar{X}\bar{Y}, \text{ en donde } E(X) = \bar{X}, \text{ y } E(Y) = \bar{Y}.$$

Pues que $(X-\bar{X})(Y-\bar{Y}) = XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}$, tenemos

$$E(X-\bar{X})(Y-\bar{Y}) = E(XY-\bar{X}Y-X\bar{Y}+\bar{X}\bar{Y})$$

y la aplicación del teorema 3.3 nos da

$$E(X-\bar{X})(Y-\bar{Y}) = E(XY) - E(\bar{X}Y) - E(X\bar{Y}) + E(\bar{X}\bar{Y})$$

Y, ya que $\bar{X}$ e $\bar{Y}$ son constantes, $E(\bar{X}Y) = \bar{X}E(Y) = \bar{X}\bar{Y}$, $E(X\bar{Y}) = \bar{X}\bar{Y}$ y $E(\bar{X}\bar{Y}) = \bar{X}\bar{Y}$. Entonces, $E(X-\bar{X})(Y-\bar{Y}) = E(XY) - \bar{X}\bar{Y}$.

Ejercicio 3.3. Suponga que $E(X) = 6$ y $E(Y) = 4$. Determine

(a) $E(2X + 4Y)$ Respuesta: 28

(b) $\{E(2X)\}^2$ Respuesta: 144

(c) $\sqrt{E(Y)}$ Respuesta: 2

(d) $E(5Y-X)$ Respuesta: 14

Ejercicio 3.4. Demuestre lo siguiente, suponiendo que $E(X) = \bar{X}$ y $E(Y) = \bar{Y}$:

(a) $E(X-\bar{X}) = 0$

(b) $E(aX-bY) + cE(Y) = a\bar{X} + (c-b)\bar{Y}$

(c) $E\{a(X-\bar{X}) + b(Y-\bar{Y})\} = 0$

(d) $E(X+a)^2 = E(X^2) + 2a\bar{X} + a^2$

(e) $E(X-\bar{X})^2 = E(X^2) - \bar{X}^2$

(f) $E(aX+bY) = 0$ para cualesquiera valores de a y b si $E(X) = 0$ y $E(Y) = 0$.

3.3 VALOR ESPERADO DE UN ESTIMADOR

El teorema 3.3 se utilizará ahora para determinar el valor esperado del promedio de una muestra aleatoria simple de n elementos seleccionados sin reposición de una población de N elementos. El término "muestra aleatoria simple" implica selección sin reposición y con probabilidades iguales. La media muestral es

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}$$

donde x_i es el valor de X correspondiente al i-ésimo elemento en la muestra. Sin pérdida de generalidad, podemos considerar el subíndice de x como correspondiente a la i-ésima selección; es decir, x_1 es el valor de X obtenido en la primera selección, x_2 el valor obtenido en la segunda, etc. Desde que cada x_i es una variable aleatoria, $\bar{x}$ es una combinación lineal de variables aleatorias. Por lo tanto, es aplicable el teorema 3.3:

$$E(\bar{x}) = \frac{1}{n}\{E(x_1) + \ldots + E(x_n)\}$$

En el capítulo anterior, sección 2.6, establecimos que cualquier elemento dado de la población tenía una probabilidad $\frac{1}{N}$ de ser seleccionado en la i-ésima selección. Esto quiere decir que x_i es una variable aleatoria que tiene probabilidad $\frac{1}{N}$ de tomar cualquier valor incluido en el conjunto de la población $X_1, \ldots, X_N$. Por lo tanto,

$$E(x_1) = E(x_2) = \ldots = E(x_n) = \bar{X} \quad y$$

$$E(\bar{x}) = \frac{\bar{X} + \ldots + \bar{X}}{n} = \bar{X}$$

El hecho de que $E(\bar{x}) = \bar{X}$ es una de las muy importantes propiedades del promedio de una muestra aleatoria simple. Incidentalmente, $E(\bar{x}) = \bar{X}$ tanto en el muestreo con reposición como en el muestreo sin reposición.

<u>Definición 3.3</u>. Un parámetro es una cantidad calculada a partir de todos los valores de la población. El total de X, el promedio de X, la proporción de elementos para los cuales $X_i < A$, o cualquier otra cantidad calculada de medidas que incluyen todos los elementos de la población, es un parámetro. El valor numérico de un parámetro es generalmente desconocido, pero existe por definición.

<u>Definición 3.4</u>. Un estimador es una fórmula o regla matemática para lograr una estimación de una muestra. La fórmula para determinar la media muestral, $\bar{x} = \frac{\Sigma x_i}{n}$, es un ejemplo sencillo de un estimador. Suministra una estimación del parámetro $\bar{X} = \frac{\Sigma X_i}{N}$.

<u>Definición 3.5</u>. Un estimador es insesgado cuando su valor esperado es igual al parámetro del cual es una estimación. En el ejemplo citado arriba, $\bar{x}$ es una estimación insesgada de $\bar{X}$, porque $E(\bar{x}) = \bar{X}$.

<u>Ejercicio 3.5</u>. Suponga una población de cuatro elementos que tienen los siguientes valores de X: $X_1 = 2$; $X_2 = 5$; $X_3 = 4$; $X_4 = 6$. Demuestre que, para muestras aleatorias simples de tamaño 2, $N\bar{x}$ es un estimador insesgado del total de la población, $\Sigma X_i = 17$. Haga una lista de todas las

muestras posibles de tamaño 2 y calcule $N\bar{x}$ para cada muestra. Esto producirá el conjunto de valores que puede tomar la variable aleatoria $N\bar{x}$. Considere la probabilidad de cada uno de los valores posibles de $N\bar{x}$ y muestre aritméticamente que $E(N\bar{x}) = 17$.

Una muestra de elementos de una población no siempre se selecciona con probabilidades de selección iguales. El muestreo con probabilidades desiguales es complicado cuando el muestreo es sin reposición, de manera que nos limitaremos al muestreo con reposición.

Ilustración 3.7. El conjunto de cuatro elementos y las probabilidades asociadas que usamos en el ejercicio 3.1 servirán como ejemplo de estimación insesgada cuando se seleccionan muestras de dos elementos con probabilidades desiguales y con reposición. Nuestro estimador del total de la población, 2 + 5 + 4 + 6 = 17, será

$$x' = \frac{\sum_{i=1}^{n} \frac{x_i}{P_i}}{n}$$

El estimador x' es una variable aleatoria. Sigue una lista del conjunto de valores que puede tomar x' y las probabilidades respectivas:

Muestras posibles	x'_j	P_j	Muestras posibles	x'_j	P_j
x_1x_1	6	4/36	x_2x_3	19,5	4/36
x_1x_2	10,5	8/36	x_2x_4	25,5	4/36
x_1x_3	15	4/36	x_3x_3	24	1/36
x_1x_4	21	4/36	x_3x_4	30	2/36
x_2x_2	15	4/36	x_4x_4	36	1/36

Ejercicio 3.6. Verifique los valores citados de x'_j y P_j y determine el valor esperado de x'. Por definición, $E(x') = \Sigma P_j x'_j$. Su respuesta debe ser 17, porque x' es un estimador insesgado del total de la población.

Con el fin de considerar un caso general de muestreo con reposición y con probabilidades desiguales, suponga que en la población X toma los valores $X_1, \ldots, X_j, \ldots, X_N$ y que las probabilidades de selección son respectivamente $P_1, \ldots, P_j, \ldots, P_N$. Sea x_i el valor de X correspondiente al i-ésimo elemento en una muestra de n elementos y p_i la probabilidad de selección de ese elemento. Entonces, un estimador insesgado del total de la población es

$$x' = \sum_{i=1}^{n} \frac{x_i}{p_i}$$

Ahora demostraremos que $E(x') = \sum_{j=1}^{N} X_j$. Para facilitar la comparación de x' con u en el teorema 3.3, x' puede escribirse como sigue:

$$x' = \frac{1}{n}\left(\frac{x_1}{p_1}\right) + \ldots + \frac{1}{n}\left(\frac{x_n}{p_n}\right)$$

Ahora está claro que $a_i = \frac{1}{n}$ y que $u_i = \frac{x_i}{p_i}$. Por lo tanto,

$$E(x') = \frac{1}{n}\left\{E\left(\frac{x_1}{p_1}\right) + \ldots + E\left(\frac{x_n}{p_n}\right)\right\} \tag{3.5}$$

La cantidad $\frac{x_1}{p_1}$, que es el resultado de la primera selección aleatoria de la población, es una variable aleatoria que podría tomar cualquiera de los siguientes valores: $\frac{X_1}{P_1}, \ldots, \frac{X_N}{P_N}$. La probabilidad de que $\frac{x_1}{p_1}$ tome el valor $\frac{X_j}{P_j}$ es P_j. Entonces, por definición,

$$E\left(\frac{x_1}{p_1}\right) = \sum_j^N P_j\left(\frac{X_j}{P_j}\right) = \sum_j^N X_j$$

Ya que el muestreo es con reposición, es obvio que cualquier $\frac{x_i}{p_i}$ es la misma variable aleatoria que $\frac{x_1}{p_1}$. Por lo tanto, la ecuación (3.5) resulta

$$E(x') = \frac{1}{n}\left(\sum_j^N X_j + \ldots + \sum_j^N X_j\right)$$

Ya que hay n términos en la suma, se tiene

$$E(x') = \sum_{j}^{N} X_j$$

Ejercicio 3.7. Como corolario, demuestre que el valor esperado de $\frac{x'}{n}$ es igual al promedio de la población.

Ya debe estar familiarizándose con la idea de que un estimador construido con los datos de una muestra probabilística es una variable aleatoria. Las personas responsables del diseño y la selección de muestras y de la producción de estimaciones derivadas de muestras se interesan en el conjunto de valores que puede tomar un estimador basado en una muestra y con las probabilidades asociadas con ellos.

Definición 3.6. La distribución de un estimador generado por el muestreo probabilístico es la distribución en el muestreo del estimador.

Los valores de x'_j y P_j en la ilustración numérica 3.7 son un ejemplo de una distribución en el muestreo. Los estadísticos se interesan principalmente en tres características de una distribución en el muestreo: (1) el promedio (centro) de la distribución con relación al valor del parámetro que se estima, (2) una medida de la variación de valores posibles de un estimador con respecto al promedio de la distribución, y (3) la forma de la distribución. Acabamos de tratar de la primera. Cuando el valor esperado de un estimador es igual al parámetro que se estima, sabemos que el promedio de la distribución en el muestreo es igual al parámetro que se estima. Pero, en la práctica, generalmente no se conocen los valores de los parámetros. Para juzgar la exactitud de un estimador, necesitamos información acerca de las tres características de la distribución en el muestreo. Tratemos ahora de la medida más generalmente aceptada de la variación de una variable aleatoria.

3.4 VARIANCIA DE UNA VARIABLE ALEATORIA

La variancia de una variable aleatoria, X, es el promedio de los cuadrados de los valores de las desviaciones de X respecto a su media; es decir, es el valor medio de $(X-\bar{X})^2$. La raíz cuadrada de la variancia es la desviación estándar de la variable.

Definición 3.7. El términos de valores esperados, la variancia de una variable aleatoria, X, es $E(X-\bar{X})^2$, donde $E(X) = \bar{X}$. Puesto que X es una variable aleatoria, $(X-\bar{X})^2$ es una variable aleatoria y, por definición de valor esperado,

$$E(X-\bar{X})^2 = \sum_i^N P_i(X_i - \bar{X})^2$$

En el caso de que $P_i = \frac{1}{N}$, tenemos la fórmula más conocida de la variancia:

$$E(X-\bar{X})^2 = \frac{\sum_i^N (X_i - \bar{X})^2}{N} = \sigma_X^2$$

Los varios símbolos que representan la variancia incluyen σ^2, σ_X^2, V^2, S^2, Var(X) y V(X). La variancia a menudo se define como

$$\frac{\Sigma(X_i - \bar{X})^2}{N-1}$$

Esto se explicará en la sección 3.7.

3.4.1 VARIANCIA DE LA SUMA DE DOS VARIABLES ALEATORIAS INDEPENDIENTES

Dos variables aleatorias, X e Y, son independientes si la probabilidad conjunta, P_{ij}, de obtener X_i e Y_j es igual a $(P_i)(P_j)$, en donde P_i es la probabilidad de seleccionar X_i del conjunto de valores de X, y P_j es la probabilidad de seleccionar Y_j del conjunto de valores de Y. La variancia de la suma de dos variables aleatorias independientes es la suma de la variancia de la primera y la de la segunda. Es decir,

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$

Ilustración 3.8. En la ilustración 3.3, X e Y eran independientes. Hicimos una lista de todos los valores posibles de $X_i + Y_j$ y la probabilidad de cada uno. De esa lista podemos calcular fácilmente la variancia de X + Y. Por definición,

$$\sigma^2_{X+Y} - E\{(X+Y)-(\bar{X}+\bar{Y})\}^2 = \sum_{ij}\sum P_i P_j\{(X_i + Y_j)-(\bar{X}+\bar{Y})\}^2 \qquad (3.6)$$

Substituyendo en la ecuación (3.6), tenemos

$$\sigma^2_{X+Y} = \frac{2}{24}(0-5,5)^2 + \frac{4}{24}(4-5,5)^2 + \ldots + \frac{1}{24}(10-5,5)^2 = \frac{85}{12}$$

Las variancias de X y de Y se calculan como sigue:

$$\sigma^2_X = E(X-\bar{X})^2 = \frac{2}{3}(2-4)^2 + \frac{2}{6}(5-4)^2 + \frac{1}{6}(4-4)^2 + \frac{1}{6}(6-4)^2 = \frac{7}{3}$$

$$\sigma^2_Y = E(Y-\bar{Y})^2 = \frac{1}{4}(-2-1,5)^2 + \frac{2}{4}(2-1,5)^2 + \frac{1}{4}(4-1,5)^2 = \frac{19}{4}$$

Ahora tenemos $\sigma^2_X + \sigma^2_Y = \frac{7}{3} + \frac{19}{4} = \frac{85}{12}$, lo que verifica la declaración hecha: la variancia de la suma de dos variables aleatorias independientes es la suma de las variancias.

Ejercicio 3.8. Demuestre que $E\{(X+Y)-(\bar{X}+\bar{Y})\}^2 = E(X+Y)^2 - (\bar{X}+\bar{Y})^2$. Después calcule la variancia de X + Y para el caso de la ilustración 3.3, usando la fórmula $\sigma^2_{X+Y} = E(X+Y)^2 - (\bar{X}+\bar{Y})^2$. El resultado debe ser igual al resultado obtenido en la ilustración 3.8.

Ejercicio 3.9. Refiérase a la ilustración 3.3 y a la lista de posibles valores de X+Y y de las probabilidades correspondientes. En vez de $X_i + Y_j$, haga una lista de los productos $(X_i - \bar{X})(Y_j - \bar{Y})$ y muestre que $E(X_i - \bar{X})(Y_j - \bar{Y}) = 0$.

Ejercicio 3.10. Calcule $E(X-\bar{X})(Y-\bar{Y})$ para el ejemplo numérico usado en la ilustración 3.3, usando la fórmula $E(XY) - \bar{X}\bar{Y}$ que se derivó en la ilustración 3.6.

3.4.2 VARIANCIA DE LA SUMA DE DOS VARIABLES ALEATORIAS DEPENDIENTES

La variancia de variables aleatorias dependientes involucra la

covariancia, que se define como sigue:

<u>Definición 3.8</u>. La covariancia de dos variables aleatorias, X e Y, es $E(X-\bar{X})(Y-\bar{Y})$, donde $E(X) = \bar{X}$, y $E(Y) = \bar{Y}$. Por la definición de valor esperado,

$$E(X-\bar{X})(Y-\bar{Y}) = \sum_{ij}\sum P_{ij}(X_i - \bar{X})(Y_j - \bar{Y})$$

donde la doble sumatoria cubre todos los valores posibles de X y de Y.

Los símbolos comúnmente usados para representar la covariancia son σ_{XY}, S_{XY} y $Cov(X,Y)$.

Puesto que $(X+Y)-(\bar{X}+\bar{Y}) = (X-\bar{X})+(Y-\bar{Y})$, podemos derivar una fórmula para la variancia de X+Y como sigue:

$$\begin{aligned}\sigma^2_{X+Y} &= E\{(X+Y)-(\bar{X}+\bar{Y})\}^2\\ &= E\{(X-\bar{X})+(Y-\bar{Y})\}^2\\ &= E\{(X-\bar{X})^2 + (Y-\bar{Y})^2 + 2(X-\bar{X})(Y-\bar{Y})\}\end{aligned}$$

Entonces, según el teorema 3.3,

$$\sigma^2_{X+Y} = E(X-\bar{X})^2 + E(Y-\bar{Y})^2 + 2E(X-\bar{X})(Y-\bar{Y})$$

y por definición obtenemos

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}$$

A veces se usa σ_{XX} en vez de σ^2_X para representar la variancia. De modo que

$$\sigma^2_{X+Y} = \sigma_{XX} + \sigma_{YY} + 2\sigma_{XY}$$

En el caso de dos variables aleatorias independientes, $P_{ij} = P_iP_j$. Por lo tanto, $E(X-\bar{X})(Y-\bar{Y}) = \sum_{ij}\sum P_iP_j(X_i - \bar{X})(Y_j - \bar{Y})$.

Escriba en detalle, si es necesario, la expresión siguiente para comprobar que es correcta:

$$\sum_{ij}\sum P_iP_j(X_i - \bar{X})(Y_j - \bar{Y}) = \sum_i P_i(X_i - \bar{X})\sum_j P_j(Y_j - \bar{Y}) = 0 \qquad (3.7)$$

Esta ecuación establece que la covariancia σ_{YY} es cero cuando X e Y son

independientes. Fíjese que, en la ecuación (3.7), $\sum_i P_i(X_i - \bar{X}) = E(X-\bar{X})$, y $\sum_j P_j(Y_j - \bar{Y}) = E(Y-\bar{Y})$, lo que, en el caso de variables aleatorias independientes, demuestra que $E(X-\bar{X})(Y-\bar{Y}) = E(X-\bar{X})E(Y-\bar{Y})$. Al trabajar con variables aleatorias independientes, a menudo se encuentra muy útil el siguiente teorema importante:

Teorema 3.4. El valor esperado del producto de variables aleatorias independientes, $u_1, u_2, \ldots, u_k$, es el producto de sus valores esperados:

$$E(u_1 u_2 \ldots u_k) = E(u_1)E(u_2)\ldots E(u_k)$$

3.5 VARIANCIA DE UN ESTIMADOR

La varianza de un estimador obtenido de una muestra probabilística depende del método de muestreo. Derivaremos la fórmula para la varianza de $\bar{x}$, el promedio de una muestra aleatoria de n elementos seleccionados con probabilidades iguales, con reposición y sin ella. Después, se derivará la varianza de un estimador del total de la población para el caso de muestreo con reposición y probabilidades desiguales de selección.

3.5.1 PROBABILIDADES DE SELECCION IGUALES

La varianza de $\bar{x}$, el promedio de una muestra aleatoria de n elementos seleccionados con probabilidades iguales y con reposición de una población de N elementos, es

$$Var(\bar{x}) = \frac{\sigma_X^2}{n}, \text{ donde } \sigma_X^2 = \frac{\sum_i^N (X_i - \bar{X})^2}{N}.$$

La demostración es la siguiente:

Por definición, $Var(\bar{x}) = E\{\bar{x}-E(\bar{x})\}^2$. Hemos demostrado que $E(\bar{x}) = \bar{X}$. Por lo tanto, $Var(\bar{x}) = E(\bar{x} - \bar{X})^2$. Por substitución y manipulación algebraica, obtenemos

$$\mathrm{Var}(\bar{x}) = E\left(\frac{x_1 + \ldots + x_n}{n} - \bar{X}\right)^2$$

$$= E\left(\frac{(x_1 - \bar{X}) + \ldots + (x_n - \bar{X})}{n}\right)^2$$

$$= \frac{1}{n^2} E\{ \sum_{i=1}^{n} (x_i - \bar{X})^2 + \sum_{i \neq j} \sum (x_i - \bar{X})(x_j - \bar{X}) \}$$

Aplicando el teorema 3.3, obtenemos

$$\mathrm{Var}\ (\bar{x}) = \frac{1}{n^2} \{ \sum_{i=1}^{n} E(x_i - \bar{X})^2 + \sum_{i \neq j} \sum E(x_i - \bar{X})(x_j - \bar{X}) \} \qquad (3.8)$$

La ecuación (3.8) se puede escribir en forma detallada como sigue:

$$\mathrm{Var}(\bar{x}) = \frac{1}{n^2}\{E(x_1-\bar{X})^2 + E(x_2-\bar{X})^2 + \ldots + E(x_1-\bar{X})(x_2-\bar{X}) + E(x_1-\bar{X})(x_3-\bar{X}) + \ldots\} .$$

Ya que el muestreo es con reposición, x_i y x_j son independientes y las covariancias son nulas. Por ejemplo, $E\{(x_1 - \bar{X})(x_2 - \bar{X})\} = E(x_1 - \bar{X})E(x_2 - \bar{X})$ y sabemos que $E(x_1 - \bar{X})$ y $E(x_2 - \bar{X})$ son iguales a cero. Ahora, considere $E(x_1 - \bar{X})^2$. Ya hemos mostrado que x_1 es una variable aleatoria que puede tomar cualquiera de los valores $X_1, \ldots, X_N$ de la población con probabilidades iguales. Por lo tanto,

$$E(x_1 - \bar{X})^2 = \frac{\sum_{j}^{N} (X_j - \bar{X})^2}{N} = \sigma_X^2$$

El mismo argumento es aplicable a x_2, x_3, etc. Por lo tanto, la ecuación (3.8) se reduce a $\mathrm{Var}(\bar{x}) = \frac{1}{n^2}(\sigma_X^2 + \ldots + \sigma_X^2) = \frac{\sigma_X^2}{n}$.

El procedimiento matemático para determinar la variancia de $\bar{x}$ cuando el muestreo es <u>sin</u> reposición es el mismo que en el caso de muestreo <u>con</u> reposición, hasta la ecuación (3.8) inclusive. Las covariancias que aparecen en la ecuación (3.8) no son nulas, porque x_i y x_j no son independientes. Por ejemplo, en la primera selección un elemento tiene probabilidad $\frac{1}{N}$ de ser seleccionado, pero en la segunda selección la probabilidad es condicionada por el hecho de que el elemento escogido en la primera selección no fue repuesto. Considere la primera covariancia en la ecuación (3.8).

Para determinar $E(x_1 - \bar{X})(x_2 - \bar{X})$ tenemos que considerar el conjunto de valores que podría tomar $(x_1 - \bar{X})(x_2 - \bar{X})$. Nos ayudará considerar la siguiente matriz:

$$\begin{matrix} (X_1 - \bar{X})^2 & (X_1 - \bar{X})(X_2 - \bar{X}) & \dots & (X_1 - \bar{X})(X_N - \bar{X}) \\ (X_2 - \bar{X})(X_1 - \bar{X}) & (X_2 - \bar{X})^2 & \dots & (X_2 - \bar{X})(X_N - \bar{X}) \\ \vdots & \vdots & & \vdots \\ (X_N - \bar{X})(X_1 - \bar{X}) & (X_N - \bar{X})(X_2 - \bar{X}) & \dots & (X_N - \bar{X})^2 \end{matrix}$$

La variable aleatoria $(x_1 - \bar{X})(x_2 - \bar{X})$ tiene igual probabilidad de ser cualquiera de los productos en la matriz, excepto los términos cuadrados de la diagonal principal. Hay $N(N-1)$ de tales productos. Por lo tanto,

$$E(x_1 - \bar{X})(x_2 - \bar{X}) = \frac{\sum_{i \neq j}^{N} \sum^{N} (X_i - \bar{X})(X_j - \bar{X})}{N(N-1)}$$

Según la ecuación (1.9) del capítulo I,

$$\sum_{i \neq j}^{N} \sum^{N} (X_i - \bar{X})(X_j - \bar{X}) = - \sum_{i}^{N} (X_i - \bar{X})^2$$

Por lo tanto,

$$E(x_1 - \bar{X})(x_2 - \bar{X}) = - \frac{\sum_{i}^{N} (X_i - \bar{X})^2}{N(N-1)} = - \frac{\sigma_X^2}{N-1}$$

La misma evaluación es aplicable a todas las demás covariancias de la ecuación (3.8). Hay $n(n-1)$ términos de este tipo en la ecuación (3.8) y el valor de cada uno es $- \frac{\sigma_X^2}{N-1}$. Por esto, la ecuación (3.8) se reduce a

$$Var(\bar{x}) = \frac{1}{n^2}\{\sum_{i}^{n} E(x_i - \bar{X})^2 - n(n-1)\frac{\sigma_X^2}{N-1}\}$$

Reconociendo que $E(x_i - \bar{X})^2 = \sigma_X^2$, y después de algunas operaciones algebraicas fáciles, obtenemos el siguiente resultado:

$$Var\ (\bar{x}) = \frac{N-n}{N-1}(\frac{\sigma_X^2}{n}) \qquad (3.9)$$

El factor $\frac{N-n}{N-1}$ se llama "factor de corrección para poblaciones finitas", porque no se usa al tratar de poblaciones infinitas o al muestrear con reposición, lo que es equivalente a muestrear una población infinita.

En el caso de dos características, X e Y, de elementos en la misma muestra aleatoria simple, la covariancia de $\bar{x}$ e $\bar{y}$ es dada por una fórmula análoga a la ecuación (3.9), a saber,

$$\mathrm{Cov}(\bar{x},\bar{y}) = \frac{N-n}{N-1}\left(\frac{\sigma_{XY}}{n}\right) \tag{3.10}$$

3.5.2 PROBABILIDADES DE SELECCION DESIGUALES

En la sección 3.3 demostramos que $x' = \frac{\sum_i^n \frac{x_i}{p_i}}{n}$ es un estimador insesgado del total de la población. Esto era para el caso de muestreo con reposición y con probabilidades de selección desiguales. Ahora procederemos a determinar la variancia de x'.

Por definición, $\mathrm{Var}(x') = E\{x' - E(x')\}^2$. Sea $X = \sum_i^N X_i$. Entonces, de $E(x') = X$ se deduce que

$$\mathrm{Var}(x') = E\left(\frac{\frac{x_1}{p_1}+\ldots+\frac{x_n}{p_n}}{n} - X\right)^2 = \frac{1}{n^2}E\left\{\left(\frac{x_1}{p_1} - X\right)+\ldots+\left(\frac{x_n}{p_n} - X\right)\right\}^2$$

$$= \frac{1}{n^2}E\left\{\Sigma\left(\frac{x_i}{p_i} - X\right)^2 + \sum_{i\neq k}\Sigma\left(\frac{x_i}{p_i} - X\right)\left(\frac{x_k}{p_k} - X\right)\right\}$$

Aplicando el teorema 3.3, se tiene

$$\mathrm{Var}(x') = \frac{1}{n^2}\left\{\Sigma E\left(\frac{x_i}{p_i} - \bar{X}\right)^2 + \sum_{i\neq j}\Sigma E\left(\frac{x_i}{p_i} - X\right)\left(\frac{x_k}{p_k} - X\right)\right\} \tag{3.11}$$

Note la similitud entre las ecuaciones (3.8) y (3.11) y que los pasos que conducen a estas dos ecuaciones son los mismos. De nuevo, ya que el muestreo es con reposición, las covariancias de la ecuación (3.11)

son nulas. Por lo tanto, la ecuación (3.11) se reduce a

$$Var(x') = \frac{1}{n^2}\{\sum_i^n E(\frac{x_i}{p_i} - X)^2\}$$

Por definición, $E(\frac{x_i}{p_i} - X)^2 = \sum_i^N P_i(\frac{X_i}{P_i} - X)^2$. Por lo tanto,

$$Var(x') = \frac{\sum_i^N P_i(\frac{X_i}{P_i} - X)^2}{n} \qquad (3.12)$$

<u>Ejercicio 3.11</u>. (a) Refiérase al ejercicio 3.1 y calcule la varianza de x' para muestras de tamaño dos (es decir, $n = 2$), usando la ecuación (3.12).

(b) Luego pase a la ilustración 3.7 y calcule la variancia de x', usando los valores numéricos de x'. No olvide que los valores de x' tienen probabilidades desiguales. Según la definición 3.7, la variancia de x' es $\sum_j^{10} P_j(x'_j - X)^2$, donde $X = E(x')$, x'_j es uno de los 10 posibles valores de x' y P_j es la probabilidad de x'_j.

3.6 VARIANCIA DE UNA COMBINACION LINEAL

Antes de presentar un teorema general sobre la variancia de una combinación lineal de variables aleatorias, se darán unas pocas relaciones claves entre variancias y covariancias. En las siguientes ecuaciones, X e Y son variables aleatorias y a, b, c y d son constantes:

$$Var(X+a) = Var(X)$$

$$Var\ (aX) = a^2Var(X)$$

$$Var(aX+b) = a^2Var(X)$$

$$Cov(X+a,Y+b) = Cov(X,Y)$$

$$Cov(aX,bY) = abCov(X,Y)$$

$$Cov(aX+b,cY+d) = acCov(X,Y)$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

$$Var(X+Y+a) = Var(X+Y)$$

$$Var(aX+bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$$

Ilustración 3.9. Las relaciones anteriores se verifican fácilmente utilizando la teoría de valores esperados. Por ejemplo,

$$\begin{aligned} Var(aX+b) &= E\{aX+b-E(aX+b)\}^2 \\ &= E\{aX+b-E(aX)-E(b)\}^2 \\ &= E\{aX-aE(X)\}^2 \\ &= E\{a(X-\bar{X})\}^2 \\ &= a^2E(X-\bar{X})^2 \\ &= a^2Var(X) \end{aligned}$$

Ejercicio 3.12. Tal como se hizo en la ilustración 3.9, utilice la teoría de valores esperados para demostrar que $Cov(aX+b,cY+d) = acCov(X,Y)$.

Tal como en el teorema 3.3, sea $u = a_1u_1 + \ldots + a_ku_k$, donde $a_1, \ldots, a_k$ son constantes y $u_1, \ldots, u_k$ son variables aleatorias. Por definición, la variancia de u es

$$Var(u) = E\{u-E(u)\}^2$$

Por substitución,

$$\begin{aligned} Var(u) &= E\{a_1u_1 + \ldots + a_ku_k - E(a_1u_1 + \ldots + a_ku_k)\}^2 \\ &= E\{a_1(u_1-\bar{u}_1) + \ldots + a_k(u_k-\bar{u}_k)\}^2, \text{ donde } E(u_i) = \bar{u}_i . \end{aligned}$$

Elevando al cuadrado la cantidad dentro de { } y considerando los valores esperados de los términos resultantes, llegamos al teorema siguiente.

Teorema 3.5. La variancia de u, una combinación lineal de variables aleatorias, es dada por la siguiente ecuación:

$$Var(u) = \sum_i^k a_i^2\sigma_i^2 + \sum_{i\neq j}\sum a_ia_j\sigma_{ij}$$

donde σ_i^2 es la variancia de u_i, y σ_{ij} es la covariancia entre u_i y u_j.

Los teoremas 3.3 y 3.5 son muy útiles, porque muchos estimadores basados en muestras probabilísticas son combinaciones lineales de variables aleatorias.

Ilustración 3.10. Suponga que de una muestra aleatoria simple se han obtenido datos correspondientes a dos características, X e Y, siendo los valores en la muestra $x_1, \ldots, x_n$, e $y_1, \ldots, y_n$. ¿Cuál es la variancia de $\bar{x} - \bar{y}$? Con base en la teoría y resultados que han sido presentados, se puede proceder inmediatamente a escribir la respuesta. Por el teorema 3.5 sabemos que $Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y}) - 2Cov(\bar{x},\bar{y})$. De las especificaciones del muestreo, conocemos las variancias de $\bar{x}$ e $\bar{y}$ y la covariancia. Vea las ecuaciones (3.9) y (3.10). De modo que se obtiene fácilmente el resultado:

$$Var(\bar{x} - \bar{y}) = \left(\frac{N-n}{N-1}\right)\left(\frac{1}{n}\right)(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}) \qquad (3.13)$$

Algunos lectores podrían tener curiosidad acerca de la relación entre covariancia y correlación. Por definición, la correlación entre X e Y es

$$r_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

Por esto, se podría poner $r_{XY}\sigma_X\sigma_Y$ en vez de σ_{XY} en la ecuación (3.13).

Ejercicio 3.13. Suponga que en una publicación estadística se encuentra que el rendimiento del maíz es de 87 quintales por hectárea en el Estado A y 83 es el rendimiento estimado para el Estado B. Los errores estándares estimados son 1,5 y 2,0 quintales. Usted se interesa en el error estándar de la diferencia en rendimiento entre los dos Estados y quiere saber la magnitud de la diferencia estimada en relación con su error estándar. Determine el error estándar de la diferencia. Puede aceptar que las dos estimaciones de rendimiento son independientes, porque la selección de la muestra en un Estado fue enteramente independiente

de la otra. Respuesta: 2,25.

Ilustración 3.11. Sin duda los estudiantes familiarizados con el muestreo ya habrán reconocido la aplicación de los teoremas 3.3 y 3.5 a varios planes de muestreo y métodos de estimación. Por ejemplo, en el caso del muestreo aleatorio estratificado, un estimador del total de la población es

$$x' = N_1\bar{x}_1 + \ldots + N_k\bar{x}_k = \Sigma N_i\bar{x}_i$$

donde N_i es el número de unidades de muestreo de la población que se encuentren en el i-ésimo estrato, y $\bar{x}_i$ es el promedio por unidad de muestreo de la característica X correspondiente a una muestra de n_i unidades de muestreo del i-ésimo estrato. Según el teorema 3.3,

$$E(x') = E\Sigma N_i\bar{x}_i = \Sigma N_i E(\bar{x}_i) \ .$$

Si el muestreo es tal que $E(\bar{x}_i) = \bar{X}_i$ en todos los estratos, x' es un estimador insesgado del total de la población. Según el teorema 3.5,

$$\mathrm{Var}(x') = N_1^2\mathrm{Var}(\bar{x}_1) + \ldots + N_k^2\mathrm{Var}(\bar{x}_k) \qquad (3.14)$$

No hay términos de covariancia en la ecuación (3.14), porque la selección de la muestra en un estrato es independiente de la hecha en otro estrato. Suponiendo una muestra aleatoria simple en cada estrato, la ecuación (3.14) se particulariza así:

$$\mathrm{Var}(x') = N_1^2\left(\frac{N_1-n_1}{N_1-1}\right)\frac{\sigma_1^2}{n_1} + \ldots + N_k^2\left(\frac{N_k-n_k}{N_k-1}\right)\frac{\sigma_k^2}{n_k}$$

en donde σ_i^2 es la variancia de X entre unidades de muestreo dentro del i-ésimo estrato.

Ilustración 3.12. Suponga que $x_1', \ldots, x_k'$ son estimadores insesgados independientes de la misma cantidad, T. Es decir, $E(x_i') = T$. Sea σ_i^2 la variancia de x_i'.

Considere un promedio ponderado de los estimadores; a saber,

$$x' = w_1 x_1' + \ldots + w_k x_k' \tag{3.15}$$

donde $\Sigma w_i = 1$. Entonces

$$E(x') = w_1 E(x_1') + \ldots + w_k E(x_k') = T \tag{3.16}$$

Es decir, para cualquier conjunto de ponderaciones con $\Sigma w_i = 1$, el valor esperado de x' es T. ¿Cómo deberán escogerse las ponderaciones?

La variancia de x' es

$$Var(x') = w_1^2\sigma_1^2 + \ldots + w_k^2\sigma_k^2$$

Si ponderamos igualmente los estimadores, $w_i = \frac{1}{k}$ y la variancia de x' es

$$Var(x') = \frac{1}{k}\left(\frac{\Sigma\sigma_i^2}{k}\right) \tag{3.17}$$

lo que es la variancia media dividida por k. Sin embargo, es razonable dar mayor ponderación a los estimadores que tengan variancia baja. Con la utilización del cálculo diferencial podemos determinar las ponderaciones que minimicen la variancia de x'. Las ponderaciones óptimas son inversamente proporcionales a las variancias de los estimadores. Es decir, $w_i \propto \frac{1}{\sigma_i^2}$.

Como ejemplo, suponga que se tienen dos estimadores insesgados independientes de la misma cantidad que se originan en dos muestras distintas. La ponderación óptima de los dos estimadores sería

$$\frac{(\frac{1}{\sigma_1^2})x_1' + (\frac{1}{\sigma_2^2})x_2'}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}$$

Como otro ejemplo, suponga que $x_1', \ldots, x_k'$ son los valores de X en una muestra de k unidades de muestreo seleccionadas con probabilidades iguales y con reposición. En este caso, cada x_i' es un estimador insesgado de $\bar{X}$. Si tomamos $w_i = \frac{1}{k}$, entonces x' es $\bar{x}$, el promedio simple de los valores en la muestra. Fíjese que, como es de esperarse, la ecuación (3.16) se reduce a $E(\bar{x}) = \bar{X}$. Además, puesto que cada estimador, x_i', es

la misma variable aleatoria que podría tomar cualquier valor del conjunto $X_1, \ldots, X_N$, claro está que todos los σ_i^2 tienen que ser iguales a $\sigma^2 = \frac{\Sigma(X_i - \bar{X})^2}{N}$. Por lo tanto, la ecuación (3.17) se reduce a $\frac{\sigma^2}{n}$, lo que está de acuerdo con la primera parte de la sección 3.5.1.

Ejercicio 3.14. Si se iguala x_i' de la ecuación (3.15) con $\frac{x_i}{p_i}$ de la sección 3.5.2 y se toma $x_i = \frac{1}{n}$, y $k = n$, entonces x' de la ecuación (3.15) resulta ser lo mismo que $x' = \frac{\Sigma \frac{x_i}{p_i}}{n}$ de la sección 3.5.2. Muestre que, en este caso, la ecuación (3.17) se vuelve la misma que la ecuación (3.12).

3.7 ESTIMACION DE VARIANCIAS

Todas las fórmulas de variancias que hemos presentado en las secciones anteriores han involucrado cálculos sobre un conjunto de valores de una población. En la práctica, tenemos datos correspondientes a una muestra únicamente. De manera que tenemos que considerar métodos para estimar variancias, empleando los datos obtenidos de muestras.

3.7.1 MUESTREO ALEATORIO SIMPLE

En la sección 3.5.1 encontramos que la variancia del promedio de una muestra aleatoria simple es

$$\mathrm{Var}(\bar{x}) = \frac{N-n}{N-1}\left(\frac{\sigma_X^2}{n}\right) \tag{3.18}$$

donde $\sigma_X^2 = \frac{\sum_i^N (X_i - \bar{X})^2}{N}$.

Como un estimador de σ_X^2, la fórmula $\frac{\sum_i^n (x_i - \bar{x})^2}{n}$ parece ser muy apropiada. Sin embargo, al muestrear poblaciones finitas, es costumbre definir variancia entre unidades de la población como sigue:

$S^2 = \frac{\sum_i^N (X_i - \bar{X})^2}{N-1}$ y utilizar $s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$ como un estimador de S^2. Una explicación de esto aparecerá cuando encontremos el valor esperado de s^2 como sigue:

La fórmula para s^2 puede escribirse en una forma más conveniente para encontrar $E(s^2)$, así:

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1} = \frac{\Sigma x_i^2 - n\bar{x}^2}{n-1}$$

y $$E(s^2) = \frac{1}{n-1}\{\sum_i^n E(x_i^2) - nE(\bar{x}^2)\}$$

Hemos demostrado anteriormente que x_i es una variable aleatoria que tiene probabilidades iguales de tomar cualquier valor en el conjunto $X_1, \ldots, X_N$.

Por lo tanto, $E(x_i^2) = \frac{\sum_i^N X_i^2}{N}$, y $\sum_i E(x_i^2) = \frac{n\Sigma X_i^2}{N}$. Entonces,

$$E(s^2) = \frac{n}{n-1}\{\frac{\Sigma X_i^2}{N} - E(\bar{x}^2)\} \qquad (3.19)$$

Sabemos, por definición, que $\sigma^2_{\bar{x}} = E(\bar{x} - \bar{X})^2$, y es fácil demostrar que $E(\bar{x} - \bar{X})^2 = E(\bar{x}^2) - \bar{X}^2$. Por lo tanto, $E(\bar{x}^2) = \sigma^2_{\bar{x}} + \bar{X}^2$. Por substitución en la ecuación (3.19) obtenemos

$$E(s^2) = \frac{n}{n-1}(\frac{\Sigma X_i^2}{N} - \bar{X}^2 - \sigma^2_{\bar{x}})$$

Por definición, $\sigma^2_{\bar{x}} = \frac{\Sigma(X_i - \bar{X})^2}{N} = \frac{\Sigma X_i^2}{N} - \bar{X}^2$, y ya que el método especificado de muestreo era MAS, $\sigma^2_{\bar{x}} = \frac{N-n}{N-1}(\frac{\sigma^2_X}{n})$, por lo que tenemos $E(s^2) = \frac{n}{n-1}\{\sigma^2_X - (\frac{N-n}{N-1})\frac{\sigma^2_X}{n}\}$, lo que después de simplificación resulta ser

$$E(s^2) = \frac{N}{N-1}\sigma_X^2$$

Note que, de las definiciones dadas arriba para σ_X^2 y S^2, se deduce que $S^2 = \frac{N}{N-1}\sigma_X^2$. Por lo tanto, $E(s^2) = S^2$. Puesto que s^2 es un estimador insesgado de S^2, ahora pondremos $\frac{N-1}{N}S^2$ en lugar de σ_X^2 en la ecuación (3.18), lo que da

$$Var(\bar{x}) = \frac{N-n}{N}(\frac{S^2}{n}) \qquad (3.20)$$

Las fórmulas (3.18) y (3.20) dan resultados idénticos para $Var(\bar{x})$, y ambas están de acuerdo con $E(\bar{x} - \bar{X})^2$ como definición de variancia. Hemos demostrado que s^2 es un estimador insesgado de S^2. Poniendo s^2 en lugar de S^2 en la ecuación (3.20), tenemos

$$Var(\bar{x}) = \frac{N-n}{N}(\frac{s^2}{n}) \qquad (3.21)$$

como un estimador de la variancia de $\bar{x}$. Con respecto a la ecuación (3.18), $\frac{N-1}{N}s^2$ es un estimador insesgado de σ_X^2. Cuando $\frac{N-1}{N}s^2$ se pone en lugar de σ_X^2, se obtiene la ecuación (3.21).

Puesto que, en la ecuación (3.20), $\frac{N-n}{N}$ es exactamente 1 menos la fracción de muestreo, y s^2 es un estimador insesgado de S^2, hay algo de ventaja en usar la ecuación (3.20) y $S^2 = \frac{\Sigma(X_i - \bar{X})^2}{N-1}$ como una definición de variancia entre unidades de muestreo en la población.

<u>Ejercicio 3.15</u>. Para una pequeña población de 4 elementos, suponga que los valores de X son $X_1 = 2$, $X_2 = 5$, $X_3 = 3$ y $X_4 = 6$. Considere muestras aleatorias simples de tamaño 2. Hay seis muestras posibles.

(a) Para cada una de las muestras calcule $\bar{x}$ y s^2. Es decir, encuentre la distribución en el muestreo de $\bar{x}$ y la distribución en el muestreo de s^2.

(b) Calcule S^2; luego determine $Var(\bar{x})$, usando la ecuación (3.20).

(c) Calcule la variancia entre los seis valores de $\bar{x}$ y compare el resultado con el valor obtenido en (b) para $Var(\bar{x})$. Los resultados deben ser iguales.

(d) De la distribución en el muestreo de s^2 calcule $E(s^2)$ y verifique que $E(s^2) = S^2$.

3.7.2 PROBABILIDADES DE SELECCION DESIGUALES

En la sección 3.5.2 derivamos una fórmula para la variancia del estimador x', donde

$$x' = \frac{\Sigma \frac{x_i}{p_i}}{n} \tag{3.22}$$

El muestreo se hizo con probabilidades de selección desiguales y con reposición. Encontramos que la variancia de x' está dada por

$$\mathrm{Var}(x') = \frac{\sum_i^N P_i \left(\frac{X_i}{P_i} - X\right)^2}{n} \tag{3.23}$$

Para estimar $\mathrm{Var}(x')$ con base en una muestra, uno podría inclinarse, en primer lugar, a usar una fórmula del mismo tipo que la ecuación (3.23), pero no resulta conveniente. La ecuación (3.23) es un promedio ponderado de los cuadrados de desviaciones $\left(\frac{X_i}{P_i} - X\right)^2$ que refleja las probabilidades desiguales de selección. Si se aplicara el mismo sistema de ponderaciones en una fórmula para estimar la variancia sobre la base de una muestra, se estaría en efecto aplicando dos veces las ponderaciones: primero en el procedimiento mismo de selección y luego a los datos de la muestra. Las probabilidades desiguales de selección ya están incorporadas en la muestra misma.

Tal como se hizo en algunas de las discusiones anteriores, mire al estimador como sigue:

$$x' = \frac{\frac{x_1}{p_1} + \ldots + \frac{x_n}{p_n}}{n} = \frac{x_1' + \ldots + x_n'}{n}, \text{ donde } x_i' = \frac{x_i}{p_i}.$$

Cada x_i' es un estimador independiente insesgado del total de la población.

Puesto que cada valor de x_i' recibe una ponderación igual en la determinación de x', parece que la fórmula siguiente para estimar $Var(x')$ podría servir:

$$Var(x') = \frac{s^2}{n} \tag{3.24}$$

donde $s^2 = \dfrac{\sum_i^n (x_i' - x')^2}{n-1}$. Siguiendo un método parecido al usado en la sección 3.7.1, es posible demostrar que

$$E(s^2) = \sum_i^{N'} P_i \left(\frac{X_i}{P_i} - X\right)^2$$

Es decir, la ecuación (3.24) provee un estimador insesgado de la $Var(x')$, cuyo valor está dado en la ecuación (3.23). La demostración se deja como ejercicio.

<u>Ejercicio 3.16</u>. Hacemos referencia al ejercicio 3.1, a la ilustración 3.7 y al ejercicio 3.11. En la ilustración 3.7, la distribución en el muestreo de x' (vea la ecuación (3.22)) se da para muestras de tamaño 2 de la población de 4 elementos que fue usada en el ejercicio 3.1.

(a) Calcule $Var(x') = \frac{s^2}{n}$ (ecuación (3.24)) para cada una de las 10 muestras posibles.

(b) Calcule el valor esperado de $Var(x')$ y compárelo con el resultado obtenido en el ejercicio 3.11. Deben ser iguales. Recuerde, al determinar el valor esperado de $Var(x')$, que los x' no ocurren con frecuencias iguales.

3.8 RAZON ENTRE DOS VARIABLES ALEATORIAS

En la teoría y en la práctica del muestreo se encuentran a menudo estimadores que son razones entre variables aleatorias. Se señaló anteriormente que $E(\frac{u}{w}) \neq \frac{E(u)}{E(w)}$, donde u y w son variables aleatorias. Ahora se presentarán fórmulas para el valor esperado de una razón y para la

variancia de una razón, pero sin la derivación. Las fórmulas son aproximaciones:

$$E\left(\frac{u}{w}\right) \doteq \frac{\bar{u}}{\bar{w}} + \frac{\bar{u}}{\bar{w}}\left(\frac{\sigma_w^2}{\bar{w}^2} - \frac{\rho_{uw}\sigma_u\sigma_w}{\bar{u}\bar{w}}\right) \tag{3.25}$$

$$Var\left(\frac{u}{w}\right) \doteq \left(\frac{\bar{u}}{\bar{w}}\right)^2\left(\frac{\sigma_u^2}{\bar{u}^2} + \frac{\sigma_w^2}{\bar{w}^2} - \frac{2\rho_{uw}\sigma_u\sigma_w}{\bar{u}\bar{w}}\right) \tag{3.26}$$

donde $\bar{u} = E(u)$

$\bar{w} = E(w)$

$\sigma_u^2 = E(u - \bar{u})^2$

$\sigma_w^2 = E(w - \bar{w})^2$

y $\rho_{uw} = \frac{\sigma_{uw}}{\sigma_u\sigma_w}$, donde $\sigma_{uw} = E(u-\bar{u})(w-\bar{w})$

Para una discusión de las condiciones bajo las que las ecuaciones (3.25) y (3.26) son buenas aproximaciones, refiérase a Hansen, Hurwitz y Madow (Sample Survey Methods and Theory, Tomo I, Cap. 4, John Wiley & Sons, 1953). Las condiciones generalmente son satisfechas con respecto a estimadores derivados de encuestas por muestreo. Como regla empírica, la fórmula para la variancia generalmente se acepta como satisfactoria si el coeficiente de variación de la variable del denominador es menos de 0,1; es decir, si $\frac{\sigma_w}{\bar{w}} < 0{,}1$. En otras palabras, esta condición requiere que el coeficiente de variación del denominador sea menos del 10 por ciento. Un coeficiente de variación más grande podría ser tolerable antes de tener que preocuparse por la aproximación de la ecuación (3.26). La condición $\frac{\sigma_w}{\bar{w}} < 0{,}1$ es más estricta que lo necesario para considerar como insignificante el sesgo de una razón. Con pocas excepciones, en la práctica se hace caso omiso del sesgo de una razón. Algo de la lógica que sostiene esta práctica aparecerá en la ilustración que sigue. En resumen, las condiciones en las cuales las ecuaciones (3.25) y (3.26) no son buenas

aproximaciones son tales que es muy probable que la razón tenga valor dudoso, debido a variancia grande.

Si u y w son combinaciones lineales de variables aleatorias, la teoría presentada en secciones anteriores es aplicable a u y a w. Suponiendo que u y w son estimadores derivados de una muestra, para estimar $Var(\frac{u}{w})$ tome en cuenta el diseño de la muestra y substituya en la ecuación (3.26) estimaciones de $\bar{u}$, $\bar{w}$, σ_u^2, σ_w^2 y ρ_{uw}. Ignore la ecuación (3.25), salvo que haya motivo para creer que el sesgo de la razón podría ser importante en relación con su error estándar.

Es de interés notar la similitud entre Var(u-w) y $Var(\frac{u}{w})$. Según el teorema 3.5,

$$Var\ (u-w) = \sigma_u^2 + \sigma_w^2 - 2\rho_{uw}\sigma_u\sigma_w$$

Por definición, la variancia relativa de un estimador es la variancia del estimador dividida por el cuadrado de su valor esperado. De modo que, en términos de la variancia relativa de una razón, la ecuación (3.26) puede escribirse

$$Var\ Rel(\frac{u}{w}) = \frac{\sigma_u^2}{\bar{u}^2} + \frac{\sigma_w^2}{\bar{w}^2} - 2\rho_{uw}\frac{\sigma_u\sigma_w}{\bar{u}\bar{w}}$$

La similitud es una ayuda para recordar la fórmula para la $Var(\frac{u}{w})$.

<u>Ilustración 3.23</u>. Supongamos que tenemos una muestra aleatoria simple de n elementos de una población de N. Sean $\bar{x}$ e $\bar{y}$ los promedios de la muestra referentes a las características X e Y. Entonces, $u = \bar{x}$ y $w = \bar{y}$:

$$\sigma_u^2 = \frac{N-n}{N}(\frac{S_X^2}{n}) \qquad y \qquad \sigma_w^2 = \frac{N-n}{N}(\frac{S_Y^2}{n})$$

Fíjese que la condición discutida anteriormente, $\frac{\sigma_w}{\bar{w}} < 0{,}1$, se satisface si la muestra es suficientemente grande para asegurar que

$$\frac{N-n}{N}(\frac{S_Y^2}{n\bar{Y}^2}) < 0{,}1^2$$

Substituyendo en la ecuación (3.26), obtenemos lo siguiente como la varianza de la razón:

$$\mathrm{Var}\left(\frac{\bar{x}}{\bar{y}}\right) = \left(\frac{N-n}{N}\right)\left(\frac{1}{n}\right)\frac{\bar{X}^2}{\bar{Y}^2}\left(\frac{S_X^2}{\bar{X}^2} + \frac{S_Y^2}{\bar{Y}^2} - \frac{2\rho_{XY}S_XS_Y}{\bar{X}\bar{Y}}\right)$$

Es sesgo de $\frac{\bar{x}}{\bar{y}}$ como estimador de $\frac{\bar{X}}{\bar{Y}}$ está dado por el segundo término de la ecuación (3.25). En el caso de esta ilustración, se tiene

$$\left(\frac{N-n}{N}\right)\left(\frac{1}{n}\right)\frac{\bar{X}}{\bar{Y}}\left(\frac{S_Y^2}{\bar{Y}^2} - \frac{\rho_{XY}\sigma_X\sigma_Y}{\bar{X}\bar{Y}}\right)$$

A medida que aumenta el tamaño de la muestra, el sesgo disminuye como $\frac{1}{n}$, mientras que el error estándar de la razón disminuye a una tasa menor, a saber, $\frac{1}{\sqrt{n}}$. Como consecuencia, no tenemos que preocuparnos por la posibilidad de que el sesgo se vuelva importante en relación con el error de muestreo a medida que aumente el tamaño de la muestra. Una posible excepción ocurre cuando se combinan varias razones. Un ejemplo es el muestreo aleatorio estratificado cuando se tienen muchos estratos y se hacen estimaciones de la razón separadamente para cada estrato. Este asunto es tratado en los textos sobre muestreo.

3.9 ESPERANZA CONDICIONAL

La teoría de esperanza condicional y variancia condicional de una variable aleatoria es una parte muy importante de la teoría del muestreo, especialmente de la teoría del muestreo en etapas múltiples. Trataremos de la teoría con respecto al muestreo en dos etapas.

La notación que se usará en esta sección y en la próxima es:

M es el número de UPM (unidades primarias de muestreo) en la población.

m es el número de UPM en la muestra.

N_i es el número total de elementos en la i-ésima UPM.

$N = \sum_{i}^{M} N_i$ es el número total de elementos en la población.

n_i es el número de elementos en la muestra tomados de la i-ésima UPM.

$N = \sum_{i}^{m} n_i$ es el número total de elementos en la muestra.

$\bar{n} = \frac{n}{m}$

X_{ij} es el valor de X correspondiente al j-ésimo elemento en la i-ésima UPM. Se refiere a un elemento en la población; es decir, j = 1, ..., N_i; i = 1, ..., M.

x_{ij} es el valor de X correspondiente al j-ésimo elemento en la muestra de la i-ésima UPM en la muestra; es decir, los subíndices i y j se refieren al conjunto de UPM y elementos en la muestra.

$X_{i\cdot} = \sum_{j}^{N_i} X_{ij}$ es el total de la población correspondiente a la i-ésima UPM.

$\bar{X}_{i\cdot} = \frac{X_{i\cdot}}{N_i}$ es el promedio de X correspondiente a todos los elementos en la i-ésima UPM.

$\bar{X}_{\cdot\cdot} = \frac{\sum_{i}\sum_{j}^{N_i} X_{ij}}{N} = \frac{\sum_{i}^{M} X_{i\cdot}}{N}$ es el promedio de todos los N elementos.

$\bar{X}_{\cdot} = \frac{\sum_{i}^{M} X_{i\cdot}}{M}$ es el promedio de los totales de las UPM. Note bien la diferencia entre $\bar{X}_{\cdot\cdot}$ y $\bar{X}_{\cdot}$

$x_{i\cdot} = \sum_{j}^{n_i} x_{ij}$ es el total de la muestra correspondiente a la i-ésima UPM en la muestra.

$\bar{x}_{i\cdot} = \frac{x_{i\cdot}}{n_i}$ es el promedio de los n_i elementos en la muestra de la i-ésima UPM.

$\bar{x}_{..} = \dfrac{\sum_{i}\sum_{j}^{mn_i} x_{ij}}{n}$ es el promedio de todos los elementos en la muestra.

Tome el caso de muestreo aleatorio simple con probabilidades de selección iguales y sin reposición, en ambas etapas. Considere la muestra de n_i elementos de la i-ésima UPM. Sabemos, de la sección 3.3, que $\bar{x}_{i.}$ es un estimador insesgado del promedio de la UPM, $\bar{X}_{i.}$; es decir, $E(\bar{x}_{i.}) = \bar{X}_{i.}$ y para una i fija (una UPM especificada), $EN_i\bar{x}_{i.} = N_iE(\bar{x}_{i.}) = N_i\bar{X}_i = X_{i.}$. Pero, debido a la primera etapa del muestreo, $EN_i\bar{x}_i$ tiene que tratarse como una variable aleatoria, de manera que es necesario considerar el valor esperado de un valor esperado.

Primero, considere a X como una variable aleatoria, en el contexto de muestreo de una sola etapa, que podría ser igual a cualquiera de los valores X_{ij} en la población de $N = \sum_{i}^{M} N_i$ elementos. Sea P(ij) la probabilidad de seleccionar el j-ésimo elemento en la i-ésima UPM; es decir, P(ij) será la probabilidad de que X sea igual a X_{ij}. Por definición,

$$E(X) = \sum_{i}\sum_{j}^{MN_i} P(ij)X_{ij} \tag{3.27}$$

Ahora considere la selección de un elemento como un procedimiento de dos etapas: (1) selección de una UPM con probabilidad P(i), y (2) selección de un elemento de la UPM seleccionada, con probabilidad $P(j|i)$. En palabras, $P(j|i)$ es la probabilidad de seleccionar el j-ésimo elemento en la i-ésima UPM, dado que la i-ésima UPM ya ha sido seleccionada. Así pues, $P(ij) = P(i)P(j|i)$. Por substitución, la ecuación (3.27) queda

$$E(X) = \sum_{i}\sum_{j}^{MN_i} P(i)P(j|i)X_{ij} \quad \text{o sea}$$

$$E(X) = \sum_{i}^{M} P(i)\sum_{j}^{N_i} P(j|i)X_{ij} \tag{3.28}$$

Por definición, $\sum_{j}^{N_i} P(j|i)X_{ij}$ es el valor esperado de X para un valor fijo

de i. Se llama "esperanza condicional".

Sea $E_2(X|i) = \sum_j^{N_i} P(j|i)X_{ij}$, donde $E_2(X|i)$ es la notación que usaremos para la esperanza condicional. Repitiendo, $E_2(X|i)$ quiere decir "el valor esperado de X para una i fija". El subíndice 2 indica que la esperanza condicional se refiere a la segunda etapa de muestreo. E_1 y E_2 se referirán a la primera etapa y a la segunda etapa, respectivamente.

Substituyendo $E_2(X|i)$ en la ecuación (3.28), obtenemos

$$E(X) = \sum_i^M P(i)E_2(X|i) \tag{3.29}$$

Hay un valor de $E_2(X|i)$ para cada una de las M UPM. Puesto que $E_2(X|i)$ es una variable aleatoria con probabilidades P(i), el segundo miembro de la ecuación (3.28) es, por definición, el valor esperado de $E_2(X|i)$. Esto conduce al siguiente teorema:

<u>Teorema 3.6</u>. $E(X) = E_1E_2(X|i)$

Suponga que $P(j|i) = \frac{1}{N_i}$ y $P(i) = \frac{1}{M}$. Entonces,

$$E_2(X|i) = \sum_j^{N_i}\left(\frac{1}{N_i}\right)X_{ij} = \bar{X}_{i.}$$

y $$E(X) = E_1(\bar{X}_{i.}) = \sum_i^M\left(\frac{1}{M}\right)(\bar{X}_{i.}) = \frac{\Sigma\bar{X}_{i.}}{M}$$

En este caso, E(X) es un promedio no ponderado de los promedios de las UPM. Es importante notar que, si P(i) y P(j|i) se escogen de tal manera que P(ij) sea constante, cada elemento tiene la misma probabilidad de selección. Este punto se tratará después.

El teorema 3.3 trata del valor esperado de una combinación lineal de variables aleatorias. Hay un teorema correspondiente para la esperanza condicional. Suponga que la combinación lineal es

$$U = a_1u_1 + \ldots + a_ku_k = \sum_{t=1}^k a_tu_t$$

donde $a_1, \ldots, a_k$ son constantes y $u_1, \ldots, u_k$ son variables aleatorias. Sea $E(U|c_i)$ el valor esperado de U bajo una condición especificada c_i, donde c_i es una de las condiciones del conjunto de M condiciones que podrían ocurrir. El teorema sobre esperanza condicional entonces puede expresarse simbólicamente como sigue:

<u>Teorema 3.7</u>. $E(U|c_i) = a_1E(u_1|c_i) + \ldots + a_kE(u_k|c_i)$

o como $$E(U|c_i) = \sum_t^k a_tE(u_t|c_i) .$$

Compare los teoremas 3.7 y 3.3 y fíjese que el 3.7 es parecido al 3.3, excepto que se aplica la esperanza condicional. Suponga que c es un suceso aleatorio y que su probabilidad es P(i). Entonces $E(U|c_i)$ es una variable aleatoria y, por definición, el valor esperado de $E(U|c_i)$ es $\sum_i^M P(i)E(U|c_i)$, que es E(U). Así pues, tenemos el siguiente teorema:

<u>Teorema 3.8</u>. El valor esperado de U es el valor esperado del valor esperado condicional de U, lo que se escribe simbólicamente como sigue:

$$E(U) = EE(U|c_i) \tag{3.30}$$

Substituyendo en la ecuación (3.30) el valor de $E(U|c_i)$ en el teorema 3.7, tenemos

$$E(U) = E\{a_1E(u_1|c_i)+\ldots+a_kE(u_k|c_i)\} = E\{\sum_t^M a_tE(u_t|c_i)\} \tag{3.31}$$

<u>Ilustración 3.14</u>. Suponga un muestreo en dos etapas con muestreo aleatorio simple en ambas etapas. Sea x', definido como sigue, el estimador del total de la población:

$$x' = \frac{M}{m}\sum_i^m \frac{N_i}{n_i}\sum_i^{n_i} x_{ij} \tag{3.32}$$

<u>Ejercicio 3.17</u>. Examine el estimador x' de la ecuación (3.32). Exprésselo en otras formas que pudieran ayudar a mostrar su estructura lógica. Por ejemplo, para una i fija, ¿qué es $\frac{N_i}{n_i}\sum_j^{n_i} x_{ij}$? ¿Le parece una

manera razonable de estimar el total de la población?

Con el fin de exhibir x' como una combinación lineal de variables aleatorias, es conveniente expresarla en la forma siguiente:

$$x' = \{\frac{M}{m}(\frac{N_1}{n_1})x_{11}+\ldots+\frac{M}{m}(\frac{N_1}{n_1})x_{1n_1}\}+\ldots+\{\frac{M}{m}(\frac{N_m}{n_m})x_{m1}+\ldots+\frac{M}{m}(\frac{N_m}{n_m})x_{mn_m}\} \quad (3.33)$$

Supongamos que queremos encontrar el valor esperado de x' para determinar si es igual al total de la población. Según el teorema 3.8,

$$E(x') = E_1E_2(x'|i) \quad (3.34)$$

$$E(x') = E_1E_2\{(\frac{M}{m}\sum_i^m \frac{N_i}{n_i}\sum_j^{n_i} x_{ij})|i\} \quad (3.35)$$

Las ecuaciones (3.34) y (3.35) se obtienen simplemente por la substitución de x' como la variable aleatoria en la ecuación (3.30). La c_i ahora se refiere a cualquiera de las m UPM en la muestra. Primero, tenemos que solucionar la esperanza condicional, $E_2(x'|i)$. Puesto que $\frac{M}{m}$ y $\frac{N_i}{n_i}$ son constantes con respecto a la esperanza condicional, podemos escribir (usando el teorema 3.7):

$$E_2(x'|i) = \frac{M}{m}\sum_i^m \frac{N_i}{n_i}\sum_j^{n_i} E_2(x_{ij}|i) \quad (3.36)$$

Sabemos que, para cualquier UPM dada en la muestra, x_{ij} es un elemento en una muestra aleatoria simple de la UPM y, según la sección 3.3, su valor esperado es el promedio de la UPM, $\bar{X}_{i.}$. Es decir,

$$E_2(x_{ij}|i) = \bar{X}_{i.}$$

$$\sum_j^{n_i} E_2(x_{ij}|i) = n_i\bar{X}_{i.} \quad (3.37)$$

Substituyendo el resultado de la ecuación (3.37) en la (3.36) nos da

$$E_2(x'|i) = \frac{M}{m}\sum_i^m N_i\bar{X}_{i.} \quad (3.38)$$

El próximo paso es encontrar el valor esperado de $E_2(x'|i)$. En la ecuación (3.38), tanto N_i como $\bar{X}_{i\cdot}$ son variables aleatorias asociadas con la primera etapa de muestreo. Así es que tomaremos $X_{i\cdot} = N_i\bar{X}_{i\cdot}$ como la variable aleatoria, lo que nos da, en lugar de la ecuación (3.38),

$$E_2(x'|i) = \frac{M}{m}\sum_i^m X_{i\cdot}$$

Por lo tanto,

$$E(x') = E_1\left(\frac{M}{m}\sum_i^m X_{i\cdot}\right)$$

Por el teorema 3.3, tenemos

$$E_1\left(\frac{M}{m}\sum_i^m X_{i\cdot}\right) = \frac{M}{m}\sum_i^m E_1(X_{i\cdot})$$

Puesto que

$$\sum_i^m E_1(X_{i\cdot}) = m\left(\frac{\sum_i^M X_{i\cdot}}{M}\right),$$

Entonces,

$$E_1\left(\frac{M}{m}\sum_i^m X_{i\cdot}\right) = \sum_i^M X_{i\cdot}$$

Por lo tanto, $E(x') = \sum_i^M X_{i\cdot} = X_{\cdot\cdot}$. Esto demuestra que x' es un estimador insesgado del total de la población.

3.10 VARIANCIA CONDICIONAL

"Variancia condicional" se refiere a la variancia de una variable sujeta a una condición o limitación especificada. Está relacionada con probabilidad condicional y esperanza condicional.

En la determinación de la variancia de x' (vea la ecuación (3.32) o la (3.33)) se usará el siguiente teorema importante:

<u>Teorema 3.9</u>. La variancia de x' está dada por

$$V(x') = V_1E_2(x'|i) + E_1V_2(x'|i)$$

donde V_1 es la variancia correspondiente a la primera etapa del muestreo y V_2 es la variancia "condicional" correspondiente a la segunda etapa.

Hemos discutido $E_2(x'|i)$ y notado que hay un valor de $E_2(x'|i)$ para cada UPM en la población. Entonces, $V_1E_2(x'|i)$ es simplemente la variancia de los M valores de $E_2(x'|i)$.

En el teorema 3.9, la variancia condicional $V_2(x'|i)$ es, por definición,

$$V_2(x'|i) = E_2[\{x' - E_2(x'|i)\}^2|i]$$

Para entender $V_2(x'|i)$, piense en x' como una combinación lineal de variables aleatorias (vea la ecuación (3.33)). Considere la variancia de x' cuando i se mantiene constante. Todos los términos (variables aleatorias) en la combinación lineal ahora son constantes, salvo las que se originan en el muestreo dentro de la i-ésima UPM. Por lo tanto, $V_2(x'|i)$ está asociada con la variación entre elementos en la i-ésima UPM. $V_2(x'|i)$ es una variable aleatoria con M valores en el conjunto, uno para cada UPM. Entonces, $E_1V_2(x'|i)$ es, por definición,

$$E_1V_2(x'|i) = \sum_i^M P(i)V_2(x'|i)$$

Es decir, $E_1V_2(x'|i)$ es un promedio de M valores de $V_2(x'|i)$ ponderados por P(i), la probabilidad de la i-ésima UPM de ser incluida en la muestra.

Se darán tres ilustraciones de la aplicación del teorema 3.9. En cada caso habrá cinco pasos en la determinación de la variancia de x':

Paso 1 - Determine $E_2(x'|i)$

Paso 2 - Determine $V_1E_2(x'|i)$

Paso 3 - Determine $V_2(x'|i)$

Paso 4 - Determine $E_1V_2(x'|i)$

Paso 5 - Combine los resultados de los pasos 2 y 4.

<u>Ilustración 3.15</u>. Esta es una ilustración sencilla, escogida porque

el resultado ya apareció en párrafos anteriores y porque no involucra una combinación lineal de variables aleatorias. Suponga que x' en el teorema 3.9 es simplemente la variable aleatoria X, donde X tiene la misma probabilidad de tomar cualquiera de los valores X_{ij} en el conjunto de $N = \sum_i^M N_i$. Sabemos que la variancia de X puede expresarse como sigue:

$$V(x') = \frac{1}{N}\sum_{ij}^{MN_i}\sum (X_{ij} - \bar{X}_{..})^2 \tag{3.39}$$

En el caso de muestreo en dos etapas, un método equivalente de seleccionar un valor de X es seleccionar primero una UPM y luego seleccionar un elemento dentro de la UPM, con la condición $P(ij) = P(i)P(j|i) = \frac{1}{N}$. Esta condición se satisface si se toma $P(i) = \frac{N_i}{N}$, y $P(j|i) = \frac{1}{N_i}$. Ahora queremos encontrar V(X) mediante el uso del teorema 3.9 y verificar el resultado con la ecuación (3.39).

Paso 1. De las especificaciones para la selección aleatoria, sabemos que $E_2(x'|i) = \bar{X}_{i.}$. Entonces,

Paso 2. $V_1E_2(x'|i) = V_1(\bar{X}_{i.})$. Sabemos que $\bar{X}_{i.}$ es una variable aleatoria que tiene probabilidad $\frac{N_i}{N}$ de tomar el i-ésimo valor en el conjunto $\bar{X}_1, \ldots, \bar{X}_M$. Entonces, por definición de la variancia de una variable aleatoria,

$$V_1E(x'|i) = \sum_i^M \frac{N_i}{N}(\bar{X}_{i.} - \bar{X}_{..})^2 \tag{3.40}$$

donde

$$\bar{X}_{..} = \sum_i^M \frac{N_i}{N}\bar{X}_{i.} = \frac{\sum_i^M X_{i.}}{N}$$

Paso 3. Por definición,

$$V_2(x'|i) = \sum_j^{N_i} \frac{1}{N_i}(X_{ij} - \bar{X}_{i.})^2$$

Paso 4. Puesto que cada valor de $V_2(x'|i)$ tiene probabilidad $\frac{N_i}{N}$,

$$E_1V_2(x'|i) = \sum_i^M \frac{N_i}{N} \sum_j^{N_i} \frac{1}{N_i}(X_{ij} - \bar{X}_{i.})^2 \tag{3.41}$$

Paso 5. De las ecuaciones (3.40) y (3.41) obtenemos

$$V(x') = \frac{1}{N}\{\sum_i^M N_i(\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_i^M \sum_j^{N_i}(X_{ij} - \bar{X}_{i.})^2\} \tag{3.42}$$

El hecho de que las ecuaciones (3.42) y (3.39) coinciden es verificado haciendo uso de la ecuación (1.10) del capítulo I.

<u>Ilustración 3.16</u>. Determine la variancia del estimador x' dado por la ecuación (3.32), suponiendo muestreo aleatorio simple en ambas etapas de muestreo.

Paso 1. Es aplicable el teorema 3.7. Es decir,

$$E_2(x'|i) = \sum_{ij}^{mn_i}\sum E_2\{\frac{M}{m}(\frac{N_i}{n_i})x_{ij}|i\}$$

lo que quiere decir "sume los valores esperados condicionales de cada uno de los n términos de la ecuación (3.33)".

Con respecto a cualquiera de los términos de la ecuación (3.33), la esperanza condicional es

$$E_2\{\frac{M}{m}(\frac{N_i}{n_i})x_{ij}|i\} = \frac{M}{m}(\frac{N_i}{n_i})E_2(x_{ij}|i) = \frac{M}{m}(\frac{N_i}{n_i})\bar{X}_{i.} = \frac{M}{m}(\frac{X_{i.}}{n_i}).$$

Por lo tanto,

$$E_2(x'|i) = \sum_{ij}^{mn_i}\sum \frac{M}{m}(\frac{X_{i.}}{n_i}) \tag{3.43}$$

Con referencia a (3.43) y sumando con respecto a j, tenemos

$$\sum_j^{n_i} \frac{M}{m}(\frac{X_{i.}}{n_i}) = \frac{M}{m}X_{i.} .$$

Por esto, la ecuación (3.43) se reduce a

$$E_2(x''|i) = \frac{M}{m}\sum_i^m X_{i\cdot} \tag{3.44}$$

Paso 2. Determine $V_1E_2(x''|i)$. Esto es sencillo, porque $\frac{\sum_i^m X_{i\cdot}}{m}$ en la ecuación (3.44) es el promedio de una muestra aleatoria de tamaño m del conjunto de totales de las UPM, $X_1, \ldots, X_M$. Por lo tanto,

$$V_1E_2(x''|i) = M^2\left(\frac{M-m}{M-1}\right)\frac{\sigma^2_{b_1}}{m} \tag{3.45}$$

donde

$$\sigma^2_{b_1} = \frac{\sum_i^M (X_{i\cdot} - \bar{X}_\cdot)^2}{M} \quad y \quad \bar{X}_\cdot = \frac{\sum_i^M X_{i\cdot}}{M}$$

En el subíndice de σ^2 la "b" representa la variancia entre UPM, y el subsubíndice "1" distingue esta variancia de las variancias entre UPM en las ilustraciones posteriores.

Paso 3. La determinación de $V_2(x''|i)$ es más complicada, porque hay que derivar la variancia condicional de una combinación lineal de variables aleatorias. Sin embargo, esto es análogo al uso del teorema 3.5 para encontrar la variancia de una combinación lineal de variables aleatorias. El teorema 3.5 es aplicable, excepto que $V(u|i)$ reemplaza a $V(u)$, y variancia condicional y covariancia condicional reemplazan a las variancias y covariancias en la fórmula para $V(u)$. A medida que se deriva la solución, note que la estrategia consiste en transformar el problema de tal modo que sea posible usar resultados obtenidos anteriormente.

Examine el estimador x'' de la ecuación (3.33) y determine si existen covariancias. Un elemento seleccionado de una UPM es independiente de un elemento seleccionado de otra; pero dentro de una UPM la situación es la misma que teníamos al encontrar la variancia del promedio de una muestra aleatoria simple. Esto nos sugiere escribir x'' en términos de $\bar{x}_{i\cdot}$, porque los $\bar{x}_{i\cdot}$ son independientes. De manera que empezaremos con

$$x' = \frac{M}{m}\sum_i^m N_i\bar{x}_{i\cdot}$$

Entonces,

$$V_2(x'|i) = V_2\{(\frac{M}{m}\sum_i^m N_i\bar{x}_{i\cdot})|i\}$$

Ya que los $\bar{x}_{i\cdot}$ son independientes,

$$V_2(x'|i) = \frac{M^2}{m^2}\sum_i^m V_2(N_i\bar{x}_{i\cdot}|i)$$

y, ya que N_i es constante con respecto a la variancia condicional,

$$V_2(x'|i) = \frac{M^2}{m^2}\sum_i^m N_i^2 V_2(\bar{x}_{i\cdot}|i) \tag{3.46}$$

Puesto que el muestreo dentro de cada UPM es muestreo aleatorio simple,

$$V_2(\bar{x}_{i\cdot}|i) = (\frac{N_i - n_i}{N_i - 1})\frac{\sigma_i^2}{n_i} \tag{3.47}$$

donde

$$\sigma_i^2 = \sum_j^{N_i}\frac{1}{N_i}(X_{ij} - \bar{X}_{i\cdot})^2$$

Paso 4. Después de substituir la ecuación (3.47) en la ecuación (3.46) y aplicar el teorema 3.3, tenemos

$$E_1V_2(x'|i) = \frac{M^2}{m^2}\sum_i^m E_1\{N_i^2(\frac{N_i - n_i}{N_i - 1})\frac{\sigma_i^2}{n_i}\}$$

Puesto que la primera etapa del muestreo era muestreo aleatorio simple, y cada UPM tenía la misma probabilidad de estar en la muestra,

$$E_1\{N_i^2(\frac{N_i - n_i}{N_i - 1})\frac{\sigma_i^2}{n_i}\} = \frac{1}{M}\sum_i^M N_i^2(\frac{N_i - n_i}{N_i - 1})\frac{\sigma_i^2}{n_i}$$

Por lo tanto,

$$E_1V_2(x'|i) = \frac{M}{m}\sum_i^M N_i^2(\frac{N_i - n_i}{N_i - 1})\frac{\sigma_i^2}{n_i} \tag{3.48}$$

Paso 5. La combinación de las ecuaciones (3.48) y (3.45) da el resultado:

$$V(x') = M^2\left(\frac{M-m}{M-1}\right)\frac{\sigma_{b_1}^2}{m} + \frac{M}{m}\sum_i^M N_i^2\left(\frac{N_i - n_i}{N_i - 1}\right)\frac{\sigma_i^2}{n_i} \qquad (3.49)$$

<u>Ilustración 3.17</u>. Las especificaciones del muestreo son: (1) en la primera etapa seleccione m UPM con reposición y con probabilidades $P(i) = \frac{N_i}{N}$ y (2) en la segunda etapa seleccione una muestra aleatoria simple de $\bar{n}$ elementos de cada una de las m UPM seleccionadas en la primera etapa. Esto dará una muestra de $n = m\bar{n}$ elementos. Encuentre la variancia del estimador del total de la población.

Es necesario cambiar el estimador, porque las UPM no fueron seleccionadas con probabilidades iguales. Los valores en la muestra tienen que ser ponderados por las recíprocas (los inversos) de sus respectivas probabilidades de selección, para que el estimador resulte insesgado.

Sean

$P'(ij)$ la probabilidad de que elemento ij esté en la muestra,

$P'(i)$ la probabilidad de que la i-ésima UPM esté en la muestra (vea la nota que sigue), y

$P'(j|i)$ la probabilidad condicional de ser seleccionado el j-ésimo elemento, dado que la i-ésima UPM ya está en la muestra.

Nota: (Según las especificaciones del muestreo, $P'(i) = m\frac{N_i}{N}$, lo que indica que la i-ésima UPM puede aparecer en una muestra más de una vez, y se cuenta cada vez que aparezca. Es decir, si la i-ésima UPM es seleccionada más de una vez, se selecciona una muestra de $\bar{n}$ dentro de la i-ésima UPM cada vez que ella sea seleccionada.)

Por substitución,

$$P'(ij) = \left(m\frac{N_i}{N}\right)\frac{\bar{n}}{N_i} = \frac{m\bar{n}}{N} = \frac{n}{N} \qquad (3.50)$$

La ecuación (3.50) quiere decir que cada elemento tiene la misma probabilidad de estar en la muestra. Por consiguiente, el estimador es muy simple:

$$x' = \frac{N}{m\bar{n}} \sum_{i}^{m} \sum_{j}^{\bar{n}} x_{ij} \tag{3.51}$$

<u>Ejercicio 3.18</u>. Demuestre que x' de la ecuación (3.51) es un estimador insesgado del total de la población.

Al buscar $V(x')$, nuestro primer paso fue determinar $E_2(x'|i)$.

Paso 1. Por definición,

$$E_2(x'|i) = E_2\{(\frac{N}{m\bar{n}} \sum_{i}^{m} \sum_{j}^{\bar{n}} x_{ij})|i\}$$

Puesto que i es constante con respecto a E_2,

$$E_2(x'|i) = \frac{N}{m\bar{n}} \sum_{i}^{m} \sum_{j}^{\bar{n}} E_2(x_{ij}|i) \tag{3.52}$$

Se deja como ejercicio el pasaje de la ecuación (3.52) al resultado siguiente:

$$E_2(x'|i) = \frac{N}{m} \sum_{i}^{m} \bar{X}_{i.} \tag{3.53}$$

Paso 2. De la ecuación (3.53) tenemos

$$V_1E_2(x'|i) = V_1(\frac{N}{m} \sum_{i}^{m} \bar{X}_{i.})$$

Ya que los $\bar{X}_{i.}$ son independientes,

$$V_1E_2(x'|i) = \frac{N^2}{m^2} \sum_{i}^{m} V_1(\bar{X}_{i.})$$

A causa de que la primera etapa de muestreo es con probabilidades proporcionales a los N_i y con reposición,

$$V_1(\bar{X}_{i.}) = \sum_{i}^{M} \frac{N_i}{N}(\bar{X}_{i.} - \bar{X}_{..})^2 \tag{3.54}$$

Sea

$$V_1(\bar{X}_{i.}) = \sigma^2_{b_2}$$

Entonces,

$$V_1E_2(x'|i) = \frac{N^2}{m^2}(m\sigma^2_{b_2}) = \frac{N^2}{m}\sigma^2_{b_2} \tag{3.55}$$

Ejercicio 3.19. Demuestre que $E(\bar{X}_{i.}) = \bar{X}_{..}$, lo que indica que es apropiado usar $\bar{X}_{..}$ en la ecuación (3.54).

Paso 3. Para determinar $V_2(x'|i)$, primero escriba el estimador como

$$x' = \frac{N}{m}\sum_i^m \bar{x}_{i.} \tag{3.56}$$

Entonces, ya que los $\bar{x}_{i.}$ son independientes,

$$V_2(x'|i) = \frac{N^2}{m^2}\sum_i^m V_2(\bar{x}_{i.})$$

y

$$V_2(\bar{x}_{i.}) = \frac{N_i - \bar{n}}{N_i - 1}\cdot\left(\frac{\sigma_i^2}{\bar{n}}\right)$$

donde

$$\sigma_i^2 = \sum_j^{N_i} \frac{1}{N_i}(X_{ij} - \bar{X}_{i.})^2$$

Por lo tanto,

$$V_2(x'|i) = \frac{N^2}{m^2}\sum_i^m \frac{N_i - \bar{n}}{N_i - 1}\left(\frac{\sigma_i^2}{\bar{n}}\right)$$

Paso 4. $E_1V_2(x'|i) = \frac{N^2}{m^2}\left(\frac{1}{\bar{n}}\right)\sum_i^m E_1\left(\frac{N_i - \bar{n}}{N_i - 1}\sigma_i^2\right)$

Puesto que la probabilidad de $V_2(x'|i)$ es $\frac{N_i}{N}$,

$$E_1V_2(x'|i) = \frac{N^2}{m^2}\left(\frac{1}{\bar{n}}\right)\sum_i^m\left\{\sum_i^M \frac{N_i}{N}\left(\frac{N_i - \bar{n}}{N_i - 1}\right)\sigma_i^2\right\}$$

que se reduce a

$$E_1V_2(x'|i) = \frac{N^2}{m\bar{n}}\sum_i^M \frac{N_i}{N}\left(\frac{N_i - \bar{n}}{N_i - 1}\right)\sigma_i^2 \tag{3.57}$$

Paso 5. Al combinar (3.55) 7 (3.57) tenemos el resultado:

$$V(x') = N^2\left\{\frac{\sigma_{b_2}^2}{m} + \frac{1}{m\bar{n}}\sum\frac{N_i}{N}\left(\frac{N_i - \bar{n}}{N_i - 1}\right)\sigma_i^2\right\} \tag{3.58}$$

CAPITULO IV. LA DISTRIBUCION DE UN ESTIMADOR

4.1 PROPIEDADES DE MUESTRAS ALEATORIAS SIMPLES

La distribución de un estimador es una base principal para juzgar la exactitud de una estimación basada en una encuesta por muestreo. Se refiere a todas las posibles estimaciones que podrían ocurrir en la repetición continua de un plan de muestreo y estimador (método de estimación) prescritos. Gracias a la teoría y la comprobación empírica de la teoría, no es necesario generar físicamente la distribución de un estimador mediante la selección de muchas muestras y la preparación de una estimación de cada una. Sin embargo, con el fin de poseer una distribución tangible de un estimador que sirva de base para discusión, hemos preparado una ilustración.

Ilustración 4.1. Considere muestras aleatorias simples de 4 tomadas de una población supuesta de 8 elementos. Hay $\frac{N!}{n!(N-n)!} = \frac{8!}{4!4!} = 70$ muestras posibles. En la tabla 4.1 se indican los valores de todas las 70 muestras de cuatro elementos. Primero se ordenaron las 70 muestras con el fin de facilitar el apunte correcto de todas. Se calculó el promedio, $\bar{x}$, de cada muestra y las muestras se ordenaron de acuerdo con los valores de $\bar{x}$ para su presentación en la tabla 4.1. La distribución de $\bar{x}$ la forman los 70 valores de $\bar{x}$ mostrados en la tabla, y cada uno de los 70 valores de $\bar{x}$ tiene la misma probabilidad de ser la estimación de una muestra aleatoria simple. Estos 70 valores se han arreglado como una distribución de frecuencias en la tabla 4.2.

Como fue explicado anteriormente, una de las propiedades del muestreo aleatorio simple es la de que la media muestral es un estimador insesgado del promedio de la población; es decir, $E(\bar{x}) = \bar{X}$. Esto quiere decir que la distribución de $\bar{x}$ está centrada en $\bar{X}$. Si la teoría es

Tabla 4.1 - Muestras de 4 elementos de una población de 8 *

No. de la muestra	Valores de x_i	$\bar{x}$	s^2	No. de la muestra	Valores de x_i	$\bar{x}$	s^2
1c	2,1,6,4	3,25	4,917	36e	1,6,8,9	6,00	12,667
2	2,1,4,7	3,50	7,000	37e	1,4,8,11	6,00	19,333
3	2,1,4,8	3,75	9,583	38e	2,6,8,9	6,25	9,583
4	2,1,6,7	4,00	8,667	39e	2,4,8,11	6,25	16,250
5	2,1,4,9	4,00	12,667	40e	1,6,7,11	6,25	16,917
6	2,1,6,8	4,25	10,917	41e	1,4,11,9	6,25	20,917
7	2,1,6,9	4,50	13,667	42	1,7,8,9	6,25	12,917
8	2,1,4,11	4,50	20,333	43ce	6,4,7,8	6,25	2,917
9ce	2,1,7,8	4,50	12,333	44e	2,6,7,11	6,50	13,667
10	1,6,4,7	4,50	7,000	45e	2,4,11,9	6,50	17,667
11e	2,1,7,9	4,75	14,917	46	2,7,8,9	6,50	9,667
12	2,6,4,7	4,75	4,917	47e	1,6,8,11	6,50	17,667
13	1,6,4,8	4,75	8,917	48e	6,4,7,9	6,50	4,333
14	2,1,6,11	5,00	20,667	49e	2,6,8,11	6,75	14,250
15e	2,1,8,9	5,00	16,667	50e	1,6,11,9	6,75	18,917
16	2,6,4,8	5,00	6,667	51	1,7,8,11	6,75	17,583
17	1,6,4,9	5,00	11,337	52e	6,4,8,9	6,75	4,917
18e	1,4,7,8	5,00	10,000	53e	2,6,11,9	7,00	15,333
19e	2,1,7,11	5,25	21,583	54	2,7,8,11	7,00	14,000
20	2,6,4,9	5,25	8,917	55	1,7,11,9	7,00	18,667
21e	2,4,7,8	5,25	7,583	56e	6,4,7,11	7,00	8,667
22e	1,4,7,9	5,25	12,250	57	4,7,8,9	7,00	4,667
23e	2,1,8,11	5,50	23,000	58	2,7,11,9	7,25	14,917
24e	2,4,7,9	5,50	9,667	59	1,8,11,9	7,25	18,917
25	1,6,4,11	5,50	17,667	60e	6,4,8,11	7,25	8,917
26e	1,6,7,8	5,50	9,667	61	2,8,11,9	7,50	15,000
27e	1,4,8,9	5,50	13,667	62ce	6,4,11,9	7,50	9,667
28ce	2,1,11,9	5,75	24,917	63	6,7,8,9	7,50	1,667
29	2,6,4,11	5,75	14,917	64	4,7,8,11	7,50	8,333
30e	2,6,7,8	5,75	6,917	65	4,7,11,9	7,75	8,917
31e	2,4,8,9	5,75	10,917	66	6,7,8,11	8,00	4,667
32e	1,6,7,9	5,75	11,583	67	4,8,11,9	8,00	8,667
33e	1,4,7,11	5,75	18,250	68	6,7,11,9	8,25	4,917
34e	2,6,7,9	6,00	8,667	69	6,8,11,9	8,50	4,333
35e	2,4,7,11	6,00	15,333	70c	7,8,11,9	8,75	2,917

* Los valores de X para la población de 8 elementos son: $X_1 = 2$, $X_2 = 1$, $X_3 = 6$, $X_4 = 4$, $X_5 = 7$, $X_6 = 8$, $X_7 = 11$, $X_8 = 9$; $\bar{X} = 6{,}00$; y

$$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{N-1} = 12$$

Tabla 4.2 - Distribución en el muestreo de $\bar{x}$

$\bar{x}$	Frecuencia relativa de $\bar{x}$		
	Muestreo aleatorio simple (Ilustración 4.1)	Muestreo por conglomerados (Ilustración 4.2)	Muestreo aleatorio estratificado (Ilustración 4.2)
3,25	1	1	
3,50	1		
3,75	1		
4,00	2		
4,25	1		
4,50	4	1	1
4,75	3		1
5,00	5		2
5,25	4		3
5,50	5		4
5,75	6	1	5
6,00	4		4
6,25	6	1	5
6,50	5		4
6,75	4		3
7,00	5		2
7,25	3		1
7,50	4	1	1
7,75	1		
8,00	2		
8,25	1		
8,50	1		
8,75	1	1	
Total	70	6	36
Valor esperado de $\bar{x}$.....	6,00	6,00	6,00
Variancia de $\bar{x}$...	1,50	3,29	0,49

correcta, el promedio de $\bar{x}$ para las 70 muestras, todas con igual probabilidad de aparecer, debe ser igual al promedio de la población, 6,00. El promedio de las 70 muestras sí es igual a 6,00.

De la teoría de valores esperados, también sabemos que la variancia de $\bar{x}$ está dada por

$$S^2_{\bar{x}} = \frac{N-n}{N}\left(\frac{S^2}{n}\right) \qquad (4.1)$$

donde

$$S^2 = \frac{\sum_i^N (X_i - \bar{X})^2}{N-1}$$

Con referencia a la ilustración 4.1 y a la tabla 4.1, $S^2 = 12{,}00$, y $S^2_{\bar{x}} = \frac{8-4}{8}\left(\frac{12}{4}\right) = 1{,}5$. La fórmula (4.1) puede verificarse mediante el cálculo de la variancia entre los 70 valores de $\bar{x}$ como sigue:

$$\frac{(3{,}25 - 6{,}00)^2 + (3{,}50 - 6{,}00)^2 + \ldots + (8{,}75 - 6{,}00)^2}{70} = 1{,}5$$

Puesto que S^2 es un parámetro de la población, generalmente es desconocido. Afortunadamente, como dijimos en el capítulo III, $E(s^2) = S^2$, donde

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$$

En la tabla 4.1 se incluye el valor de s^2 para cada una de las 70 muestras; el promedio de los 70 valores de s^2 es igual a S^2. El hecho de que $E(s^2) = S^2$ es otra propiedad importante de las muestras aleatorias simples. En la práctica, s^2 se usa como un estimador de S^2. Es decir,

$$s^2_{\bar{x}} = \frac{N-n}{N}\left(\frac{s^2}{n}\right)$$

es un estimador insesgado de la variancia de $\bar{x}$.

En síntesis, acabamos de verificar tres propiedades importantes de las muestras aleatorias simples:

(1) $E(\bar{x}) = \bar{X}$

(2) $S_{\bar{x}} = \sqrt{\frac{N-n}{N}} \left(\frac{S}{\sqrt{n}}\right)$

(3) $E(s^2) = S^2$

El error estándar de $\bar{x}$, $S_{\bar{x}}$, es una medida de cuánto varía $\bar{x}$ con respecto a $\bar{X}$ en muestreos repetidos. Incidentalmente, note que la ecuación (4.1) muestra cómo se relaciona la variancia de $\bar{x}$ con el tamaño de la muestra. Ahora tenemos que considerar la forma o configuración de la distribución de $\bar{x}$.

Definición 4.1. La distribución de un estimador a menudo se llama "distribución en el muestreo". Se refiere a la distribución de todos los valores posibles de un estimador que pudieran ocurrir bajo un plan prescrito de muestreo.

4.2 FORMA DE LA DISTRIBUCION EN EL MUESTREO

En relación con el muestreo aleatorio hay gran cantidad de literatura sobre la distribución de estimadores, pero no trataremos de hacer una revisión de ella. En la práctica, la distribución generalmente se acepta como normal (vea la figura 4.1), salvo que sea "pequeño" el tamaño de la muestra. La teoría y los ensayos empíricos demuestran que la distribución de un estimador se acerca rápidamente a la distribución normal a medida que aumenta el tamaño de la muestra. La proximidad que tiene la distribución de un estimador a la distribución normal depende de (1) la distribución de X (es decir, la forma de la distribución de frecuencias de X en la población que se muestrea), (2) la forma del estimador, (3) el diseño de la muestra y (4) el tamaño de la muestra. No es posible suministrar unas pocas guías sencillas y precisas para decidir cuándo el grado de aproximación es suficientemente bueno. En la práctica

generalmente es cuestión de trabajar como si la distribución del estimador fuera normal, pero tener en mente la posibilidad de que la distribución puede diferir considerablemente de la normal cuando la muestra es muy pequeña y la distribución de la población es muy asimétrica.*

Es muy afortunado que la distribución en el muestreo sea aproximadamente normal, porque esto da una base para afirmaciones probabilísticas acerca de la precisión de un estimador. En cuanto a notación, x' será la expresión general para representar cualquier estimador y $\sigma_{x'}$ será el error estándar de x'.

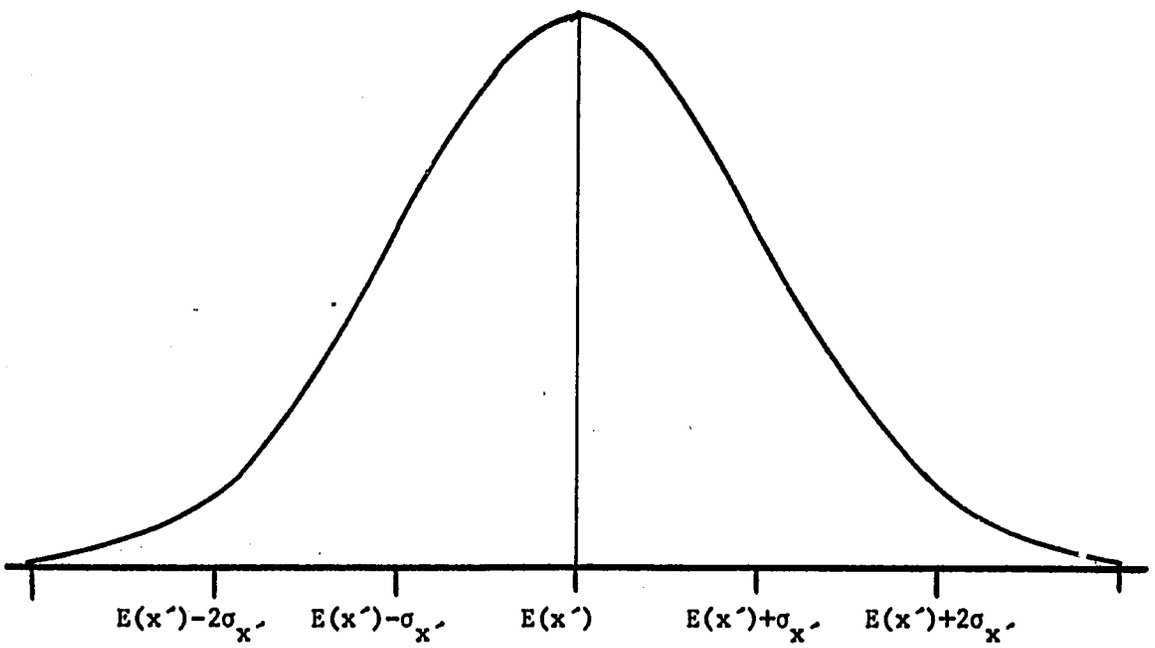


Fig. 4.1. Distribución (normal) de un estimador

La figura 4.1 es una representación gráfica de la distribución en el muestreo de un estimador. Es la distribución normal. En la ecuación matemática para la distribución normal de una variable hay dos parámetros: el valor medio y el error estándar de la variable.

---

* Para una buena discusión de la distribución de un estimador basado en una muestra, vea el Tomo I, Cap. 1 de Sample Survey Methods and Theory por Hansen, Hurwitz y Madow. John Wiley & Sons. 1953.

Suponga que x' es un estimador basado en una muestra probabilística. Las características de la distribución en el muestreo de x' quedan especificadas por tres cosas: (1) el valor esperado de x', $E(x')$, que es el promedio de la distribución, (2) el error estándar de x', $\sigma_{x'}$, y (3) la suposición de que la distribución es normal. Si x' tiene distribución normal, dos tercios de los valores que x' podría tomar se encuentran entre los límites $\{E(x') - \sigma_{x'}\}$ y $\{E(x') + \sigma_{x'}\}$; el 95 por ciento de los posibles valores de x' se encuentran entre $\{(x') - 2\sigma_{x'}\}$ y $\{E(x') + 2\sigma_{x'}\}$; y el 99,7 por ciento de las estimaciones no se apartarán más de $3\sigma_{x'}$ de $E(x')$.

<u>Ejercicio 4.1</u>. Refiriéndose a la ilustración 4.1, calcule $E(\bar{x}) - \sigma_{\bar{x}}$ y $E(\bar{x}) + \sigma_{\bar{x}}$. En la tabla 4.2 encuentre la proporción de los 70 valores de $\bar{x}$ que están entre $E(\bar{x}) - \sigma_{\bar{x}}$ y $E(\bar{x}) + \sigma_{\bar{x}}$. ¿Cómo compara esta proporción con la esperada, suponiendo que la distribución en el muestreo de $\bar{x}$ es normal? No se espera que la aproximación a la normal sea muy cercana, debido al tamaño pequeño de la población y de la muestra. También calcule $E(\bar{x}) - 2\sigma_{\bar{x}}$ y $E(\bar{x}) + 2\sigma_{\bar{x}}$ y determine la proporción de los 70 valores de $\bar{x}$ que se encuentran entre estos dos límites.

4.3 DISEÑO DE MUESTRAS

Hay muchos métodos para el diseño y selección de muestras y para hacer estimaciones basadas en muestras. Cada método de muestreo y cada estimador tiene una distribución en el muestreo. Puesto que la distribución en el muestreo se supone normal, la comparación entre métodos alternativos se hace en términos de $E(x')$ y $\sigma_{x'}$ (o $\sigma^2_{x'}$).

En el caso del muestreo aleatorio simple, hemos visto que, para una muestra de n, <u>toda posible combinación de n elementos</u> tiene la misma probabilidad de ser la muestra seleccionada. Algunas de estas posibles

combinaciones (muestras) son mucho mejores que otras. Es posible introducir restricciones en el muestreo para que ciertas combinaciones no puedan aparecer, o para que ciertas combinaciones tengan mayor probabilidad de ocurrir que otras. Esto se puede hacer sin instroducir sesgo en el estimador $\bar{x}$ y sin perder la base para la estimación de $\sigma_{\bar{x}}$. La discusión de diseños particulares de muestras no es propósito principal de este capítulo; sin embargo, se usarán unas pocas ilustraciones sencillas para introducir el tema de diseño y para ayudar a desarrollar conceptos de variación en el muestreo.

Ilustración 4.2. Suponga que la población de 8 elementos usada en la tabla 4.1 se arregla de modo que se componga de cuatro unidades de muestreo como sigue:

Unidad de muestreo	Elementos	Valores de X	Total de la unidad
1	1 y 2	$X_1 = 2$, $X_2 = 1$	3
2	3 y 4	$X_3 = 6$, $X_4 = 4$	10
3	5 y 6	$X_5 = 7$, $X_6 = 8$	15
4	7 y 8	$X_7 = 11$, $X_8 = 9$	20

Para fines de muestreo, la población ahora se compone de cuatro unidades de muestreo en vez de ocho elementos. Si seleccionamos una muestra aleatoria simple de dos unidades de muestreo de la población de cuatro unidades de muestreo, es obvio que es aplicable la teoría de muestreo para el muestreo aleatorio simple. Esta ilustración destaca la importancia de hacer una clara distinción entre las unidades de muestreo y los elementos a los que corresponden las mediciones. Una unidad de muestreo corresponde a una selección aleatoria y la variación entre unidades de muestreo (selecciones aleatorias) es lo que determina el error de muestreo de un estimador. Cuando las unidades de muestreo se componen de más de un elemento, el muestreo comúnmente se llama "muestreo por

conglomerados", porque los elementos en una unidad de muestreo generalmente se encuentran juntos geográficamente. En el caso de una muestra aleatoria simple de 2 unidades de muestreo, la variancia de $\bar{x}_c$ (donde $\bar{x}_c$ es la media muestral por unidad de muestreo) es

$$S^2_{\bar{x}_c} = \frac{N-n}{N}\left(\frac{S^2_c}{n}\right) = 13,17$$

donde

$$N = 4,\ n = 2 \text{ y } S^2_c = \frac{(3-12)^2+(10-12)^2+(15-12)^2+(20-12)^2}{3} = \frac{158}{3}$$

En vez del promedio por unidad de muestreo, probablemente estaremos más interesados en el promedio por elemento, que es $\bar{x} = \frac{\bar{x}_c}{2}$, porque hay dos elementos en cada unidad de muestreo. La variancia de $\bar{x}$ es la cuarta parte de la variancia de $\bar{x}_c$. Entonces, la variancia de $\bar{x}$ es $\frac{13,17}{4} = 3,29$.

Hay solamente seis muestras aleatorias posibles, como sigue:

Unidades de muestreo	Media muestral por unidad de muestreo	s^2_c
1 y 2	6,5	24,5
1 y 3	9,0	72,0
1 y 4	11,5	144,5
2 y 3	12,5	12,5
2 y 4	15,0	50,0
3 y 4	17,5	12,5

donde $s^2_c = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1}$ y x_i es el total de la i-ésima unidad de muestreo. Tenga la seguridad de notar que s^2_c (que es la estimación de S^2_c hecha de la muestra) es la variancia entre unidades de muestreo en la muestra y no la variancia entre elementos individuales en la muestra. De la lista de seis muestras es fácil verificar que $\bar{x}_c$ es un estimador insesgado del promedio de la población por unidad de muestreo y que s^2_c es un estimador insesgado de 13,17, la variancia entre las 4 unidades de muestreo en la

población. También la variancia entre los seis valores de $\bar{x}$ es 13,17, lo que está de acuerdo con la fórmula.

Las seis muestras conglomeradas posibles se encuentran entre las 70 muestras de la tabla 4.1. Los números que las identifican en la lista son 1, 9, 28, 43, 62 y 70. Una "c" sigue a cada uno de estos números. La distribución en el muestreo para las seis muestras aparece en la tabla 4.2, lo que permite una comparación con la distribución correspondiente al muestreo aleatorio simple. La comparación muestra que la selección aleatoria entre estas seis es menos deseable que la selección aleatoria entre las 70. Por ejemplo, la probabilidad de ocurrencia de uno de los promedios extremos (3,25 ó 8,75) es $\frac{1}{3}$ en el caso de muestreo por conglomerados y de sólo $\frac{1}{35}$ cuando se selecciona una muestra aleatoria simple de cuatro elementos. En esta ilustración, la restricción en el muestreo (la conglomeración o agrupamiento de elementos) aumentó la variancia de la muestra, de 1,5 hasta 3,29.

Es importante notar que la variancia media entre elementos dentro de los cuatro conglomerados o grupos es solamente 1,25. (Los estudiantes deben calcular la variancia de cada grupo para verificar este resultado.) Este valor es mucho menor que 12,00, la variancia entre los 8 elementos de la población. En realidad, la variancia entre elementos dentro de conglomerados es usualmente menor que la variancia entre todos los elementos de la población, porque los conglomerados (unidades de muestreo) usualmente se componen de elementos que están juntos, por lo que generalmente tienen una tendencia de ser parecidos.

<u>Ejercicio 4.2</u>. En la ilustración 4.2, si la variancia media entre elementos dentro de conglomerados hubiera sido mayor que 12,00, la variancia en el caso de muestreo por conglomerados habría sido menor que la variancia de una muestra aleatoria simple de elementos. Repita lo que se

hizo en la ilustración 4.2, usando como unidades de muestreo los conglomerados formados por los elementos 1 y 6, 2 y 5, 3 y 8, y 4 y 7. Estudie los resultados.

Ilustración 4.3. Quizá el método más común de muestreo es el que se basa en la asignación de unidades de muestreo de una población a grupos llamados "estratos", de cada uno de los cuales se saca una muestra aleatoria simple. Suponga que la población usada en la ilustración 4.1 se divide entre dos estratos, como sigue:

Estrato 1: $X_1 = 2$, $X_2 = 1$, $X_3 = 6$, $X_4 = 4$

Estrato 2: $X_5 = 7$, $X_6 = 8$, $X_7 = 11$, $X_8 = 9$

El plan de muestreo consiste en seleccionar de cada estrato una muestra aleatoria simple de dos elementos. Hay 36 muestras posibles de 4, dos de cada estrato. Estas 36 muestras se identifican en la tabla 4.1 con una "e" después del número de la muestra para facilitar su comparación con las 70 muestras aleatorias simples y con las seis muestras de conglomerados. Vea también la tabla 4.2.

Considere la variancia de $\bar{x}$. Podemos escribir

$$\bar{x} = \frac{\bar{x}_1 + \bar{x}_2}{2}$$

en donde $\bar{x}_1$ es la media muestral del estrato 1 y $\bar{x}_2$ es la media muestral correspondiente al estrato 2. Según el teorema 3.5,

$$S^2_{\bar{x}} = (\tfrac{1}{4})(S^2_{\bar{x}_1} + S^2_{\bar{x}_2} + 2S_{\bar{x}_1\bar{x}_2})$$

Sabemos que la covariancia, $S_{\bar{x}_1\bar{x}_2}$, es cero, porque el muestreo de un estrato es independiente del muestreo del otro estrato. Y, ya que la muestra dentro de cada estrato es una muestra aleatoria simple,

$$S^2_{\bar{x}_1} = \frac{N_1 - n_1}{N_1}\left(\frac{S^2_1}{n_1}\right)$$

donde
$$S_1^2 = \frac{\sum_i^{N_1} (X_{1i} - \bar{X}_{1.})^2}{N_1 - 1}$$

El subíndice "1" se refiere al estrato 1. $S^2_{\bar{x}_2}$ es de la misma forma que $S^2_{\bar{x}_1}$. Por lo tanto,

$$S^2_{\bar{x}} = \frac{1}{4}\{\frac{N_1 - n_1}{N_1}(\frac{S_1^2}{n_1}) + \frac{N_2 - n_2}{N_2}(\frac{S_2^2}{n_2})\}$$

Puesto que

$$\frac{N_1 - n_1}{N_1} = \frac{N_2 - n_2}{N_2} = \frac{1}{2} \text{ y } n_1 = n_2 = 2,$$

$$S^2_{\bar{x}} = \frac{1}{8}(\frac{S_1^2 + S_2^2}{2}) = \frac{1}{8}(\frac{4,92 + 2,92}{2} = 0,49$$

La variancia, 0,49, es comparable con 1,5 en la ilustración 4.1 y con 3,29 en la ilustración 4.2.

En la ilustración 4.2, las unidades de muestreo eran grupos de dos elementos y la variancia entre estos grupos (unidades de muestreo) apareció en la fórmula para la variancia de $\bar{x}$. En la ilustración 4.3, cada elemento era una unidad de muestreo, pero el procedimiento de selección (aleatorización) fue restringido a tomar un estrato (subconjunto) a la vez, de manera que la variancia fue determinada por la variabilidad dentro de los estratos. A medida que estudie planes de muestreo, fórmese imágenes mentales de la variación de la que depende el error de muestreo. Con la experiencia y los conocimientos acumulados acerca de los tipos de variación que aparecen en distintas poblaciones, uno se hace experto en juzgar la eficiencia de planes alternativos de muestreo en relación con los objetivos específicos de una encuesta.

Si la población y las muestras en las ilustraciones citadas hubieran sido más grandes, las distribuciones de la tabla 4.2 habrían sido aproximadamente normales. Así es que, puesto que la forma de la distribución

de un estimador derivado de una encuesta por muestreo probabilístico se acepta como normal, solo dos atributos tienen que ser evaluados; a saber, su valor esperado y su variancia.

En las ilustraciones dadas, se supusieron condiciones ideales. Tales condiciones no existen en el mundo real, de manera que la teoría tiene que extenderse para que se cuadre más exactamente con las condiciones existentes. Hay varias fuentes de error o variación que han de evaluarse. La naturaleza de la relación entre la teoría y la práctica es un factor principal entre los que gobiernan la tasa de progreso en el mejoramiento de la exactitud de los resultados de las encuestas.

Ahora extenderemos los conceptos de error hacia planteamientos más prácticos.

4.4 ERROR DE RESPUESTA

Hasta ahora hemos hablado del muestreo bajo la suposición implícita de que se obtuvieran medidas de todos los n elementos en una muestra y que la medida de cada elemento sea sin error. Ninguna de las dos suposiciones cuadra exactamente con el mundo real. Existen errores de varias clases. Por ejemplo, en una encuesta de fincas, se define "finca", pero la aplicación de la definición involucra cierto grado de ambigüedad con respecto a si ciertas explotaciones satisfacen o no la definición. O dos personas podrán tener interés en la misma parcela, dando lugar a la posibilidad de que la parcela se cuente dos veces (incluida como parte de dos fincas), o podrá omitirse del todo.

En parte con el propósito de enfatizar que el error en una estimación es más que una cuestión de muestreo, los estadísticos a menudo clasifican las muchas fuentes de error entre dos clases generales: (1) errores de muestreo, que son los asociados con el hecho de que se tienen medidas

de una muestra de elementos en vez de medidas de todos los elementos de la población y (2) errores ajenos al muestreo - los que ocurren ya sea que se aplique muestreo o no. Los modelos matemáticos de error pueden ser muy complejos cuando incluyen un término para cada una de las muchas fuentes de error e intentan representar exactamente el mundo real. Pero no siempre son necesarios modelos de error complicados; su uso depende de los propósitos.

Para fines de explicación, usaremos dos modelos muy simplificados de error de respuesta. Esto introducirá el tema de error de respuesta y dará algunos indicios sobre la naturaleza del efecto del error de respuesta sobre la distribución de un estimador. Por simplicidad, supondremos que se obtiene una medida para cada elemento en una muestra aleatoria y que no existe ninguna ambigüedad respecto a la identidad o definición de un elemento. Así estaremos considerando el error de muestreo y el error de respuesta simultáneamente.

Ilustración 4.4. Sean $T_1, \ldots, T_N$ los "valores verdaderos" de alguna variable para los N elementos de una población. La mención de valores verdaderos origina varias preguntas sobre la naturaleza de un valor verdadero. Por ejemplo, ¿cuál es su peso verdadero? ¿Cómo definiría el peso verdadero de una persona? No entraremos en el problema de definir valores verdaderos; nos limitaremos a suponer que existen valores verdaderos de acuerdo con alguna definición práctica. Al intentar averiguar T_i, podría obtenerse algún valor distinto. Denotemos con X_i el valor realmente obtenido. La diferencia, $e_i = X_i - T_i$, es el error de respuesta correspondiente al i-ésimo elemento. Si la variable es, por ejemplo, el peso de una persona, el peso X_i observado para la i-ésima persona depende de cómo y cuándo se realice la medición. Sin embargo, por simplificar, suponga que el valor obtenido sea siempre X_i, sin importar las condiciones

en que se realice la medición. En otras palabras, suponga que el error de respuesta, e_i, del i-ésimo elemento sea constante. En este caso hipotético, realmente estamos muestreando un conjunto de valores poblacionales, $X_1, \ldots, X_N$, en vez de un conjunto de valores verdaderos, $T_1, \ldots, T_N$.

En las condiciones citadas la teoría de muestreo es aplicable exactamente al conjunto de valores poblacionales $X_1, \ldots, X_N$. Si se selecciona una muestra aleatoria simple de elementos y se obtienen medidas de todos los elementos en la muestra, entonces $E(\bar{x}) = \bar{X}$. Es decir, si el propósito del muestreo es estimar $\bar{X}$, el estimador es insesgado; pero si el propósito es estimar

$$\bar{T} = \frac{\sum_i^N T_i}{N}$$

el estimador es sesgado, salvo que $\bar{T}$ sea igual a $\bar{X}$. La cantidad $\bar{X} - \bar{T}$ se denomina "sesgo de respuesta".

El error de respuesta, $e_i = X_i - T_i$ puede escribirse así:

$$X_i = T_i + e_i \tag{4.2}$$

Entonces, en el caso de una muestra aleatoria simple,

$$\bar{x} = \frac{\sum^n x_i}{n} = \frac{\sum^n (t_i + e_i)}{n}$$

Por lo tanto,

$$\bar{x} = \bar{t} + \bar{e}$$

De la teoría de valores esperados tenemos

$$E(\bar{x}) = E(\bar{t}) + E(\bar{e})$$

Puesto que $E(\bar{x}) = \bar{X}$ y $E(\bar{t}) = \bar{T}$, síguese que

$$\bar{X} = \bar{T} + E(\bar{e})$$

De modo que $\bar{x}$ es un estimador sesgado de $\bar{T}$ salvo que $E(\bar{e}) = 0$, donde

$E(\bar{e}) = \frac{\sum^{N} e_i}{N}$. Es decir, $E(\bar{e})$ es el promedio de los errores de respuesta, e_i, de toda la población.

En el caso de muestreo aleatorio simple, la varianza de $\bar{x}$ es

$$S^2_{\bar{x}} = \frac{N-n}{N}\left(\frac{S^2_X}{n}\right), \text{ donde } S^2_X = \frac{\sum_i^N (X_i - \bar{X})^2}{N-1}$$

¿Qué efecto tiene el error de respuesta sobre la variancia de X y la de x? Ya hemos dicho que el valor observado del i-ésimo elemento es igual a su valor verdadero más un error de respuesta; es decir, $X_i = T_i + e_i$. Suponiendo muestreo aleatorio, T_i y e_i son variables aleatorias. Podemos usar el teorema 3.5 del capítulo III y escribir

$$S^2_X = S^2_T + S^2_e + 2S_{T,e} \qquad (4.3)$$

donde S^2_X es la variancia de X, S^2_T es la variancia de T, S^2_e es la variancia de e y $S_{T,e}$ es la covariancia entre T y e. Los términos del segundo miembro de la ecuación (4.3) no pueden ser evaluados sin datos correspondientes a X_i y T_i, pero la ecuación sí muestra cómo el error de respuesta afecta la variancia de X y, en consecuencia, la de $\bar{x}$.

Como un ejemplo numérico, suponga una población de 5 elementos y los siguientes valores de T y de X:

	T_i	X_i	e_i
	23	26	3
	13	12	-1
	17	23	6
	25	25	0
	7	9	2
Promedio	17	19	2

Quizá los estudiantes quieran verificar los siguientes resultados, especialmente para la variancia de e y la covariancia entre T y e:

S_X^2 62,5 $\qquad S_T^2 = 54,0 \qquad S_e^2 = 7,5 \qquad S_{T,e} = 0,5$

Como verificación de la ecuación (4.3), tenemos

$$62,5 = 54,0 + 7,5 + 2(0,5)$$

De datos de una muestra aleatoria simple se calcularía

$s_{\bar{x}}^2 = \dfrac{\sum_i^n (x_i - \bar{x})^2}{n-1}$ y se usaría $\dfrac{N-n}{N}\left(\dfrac{s_x^2}{n}\right)$ como una estimación de la variancia de $\bar{x}$. ¿Es claro que s_x^2 es un estimador insesgado de S_X^2 en vez de S_T^2 y que el impacto de variación en e_i está incluido en s_x^2?

Para resumir, el error de respuesta causó un sesgo en $\bar{x}$ como una estimación de $\bar{T}$ que era igual a $\bar{X} - \bar{T}$. Además, era una fuente de variación incluida en el error estándar de $\bar{x}$. Para evaluar el sesgo y la variancia que sean atribuibles al error de respuesta, es necesario tener información sobre X_i y T_i.

<u>Ilustración 4.5</u>. En este caso, suponemos que el error de respuesta para un elemento dado no es constante. Es decir, si un elemento se midiera en varias ocasiones, los valores observados para el i-ésimo elemento podrían variar aunque el valor verdadero, T_i, quedara sin cambio. Sea el modelo de error

$$X_{ij} = T_i + e'_{ij}$$

donde X_{ij} es el valor observado de X para el i-ésimo elemento cuando la observación se toma en una ocasión particular, j;

T_i es el valor verdadero de X para el i-ésimo elemento; y

e'_{ij} es el error de respuesta para el i-ésimo elemento en una ocasión particular, j.

Suponga, para cualquier elemento dado, que el error de respuesta, e'_{ij}, es una variable aleatoria. Podemos tomar $e'_{ij} = \bar{e}_i + e_{ij}$, donde $\bar{e}_i$ es el valor medio de e_{ij} para una i fija; es decir, $\bar{e}_i = E(e'_{ij}|i)$. Esto

separa el error de respuesta en dos componentes: uno constante, $\bar{e}_i$, y uno variable, e_{ij}. Por definición, el valor esperado de e_{ij} es cero, para cualquier elemento dado. Eso es, $E(e_{ij}|i) = 0$.

Substituyendo $\bar{e}_i + e_{ij}$ por e'_{ij}, el modelo se vuelve

$$X_{ij} = T_i + \bar{e}_i + e_{ij} \qquad (4.4)$$

El modelo, ecuación (4.4), ahora está en forma adecuada para compararse con el modelo en la ilustración 4.4. En la ecuación (4.4), $\bar{e}_i$, tal como e_i en la ecuación (4.2), es constante para un elemento dado. Por esto, los dos modelos son parecidos, excepto el término adicional, e_{ij}, en la ecuación (4.4), que permite considerar la posibilidad de que el error de respuesta para el i-ésimo elemento no sea constante.

Suponga una muestra aleatoria simple de n elementos y una observación para cada elemento. Del modelo dado por la ecuación (4.4) ya podemos escribir la media muestral como sigue:

$$\bar{x} = \frac{\sum_i t_i}{n} + \frac{\sum_i \bar{e}_i}{n} + \frac{\sum_i e_{ij}}{n}$$

La suma con respecto a j no es necesaria, porque hay una sola observación para cada elemento en la muestra. En las condiciones especificadas, el valor esperado de $\bar{x}$ puede expresarse como sigue:

$$E(\bar{x}) = \bar{T} + \bar{e}$$

donde

$$\bar{T} = \frac{\sum_i^N T_i}{N} \quad y \quad \bar{e} = \frac{\sum_i^N \bar{e}_i}{N}$$

La variancia de $\bar{x}$ es complicada, salvo que se supongan ciertas condiciones adicionales. Suponga que todos los términos de covariancia sean iguales a cero. También, suponga que la variancia condicional de e_{ij} sea constante para todos los valores de i; es decir, que sea $V(e_{ij}|i) = S_e^2$.

Entonces la variancia de $\bar{x}$ es

$$S^2_{\bar{x}} = \frac{N-n}{N}\left(\frac{S^2_T}{n}\right) + \frac{N-n}{N}\left(\frac{S^2_{\bar{e}}}{n}\right) + \frac{S^2_e}{n}$$

donde

$$S^2_T = \frac{\sum_i^N (T_i - \bar{T})^2}{N-1}, \quad S^2_{\bar{e}} = \frac{\sum_i^N (\bar{e}_i - \bar{e})^2}{N-1}$$

y S^2_e es la variancia condicional de e_{ij}; es decir, $V(e_{ij}|i)$. En este modelo, la variancia de $\bar{x}$ no disminuye a cero a medida que $n \to N$. Sin embargo, suponiendo que N sea grande, la variancia de $\bar{x}$, que se vuelve $\frac{S^2_e}{N}$, probablemente sea despreciable.

<u>Definición 4.2</u>. <u>Error cuadrático medio</u>. En términos de la teoría de valores esperados, el error cuadrático medio de un estimador x' es $E(x' - T)^2$, donde T es el valor verdadero que se desea estimar. Es fácil demostrar que $E(x' - T)^2 = \{E(x') - T\}^2 + E\{x' - E(x')\}^2$. Así pues, el error cuadrático medio, ECM, puede expresarse como:

$$ECM = B^2 + \sigma^2_{x'} \tag{4.5}$$

donde

$$B = E(x') - T \tag{4.6}$$

y

$$\sigma^2_{x'} = E\{x' - E(x')\}^2 \tag{4.7}$$

<u>Definición 4.3</u>. <u>Sesgo</u>. En la ecuación (4.5), B (por la palabra inglesa <u>bias</u>) es el sesgo de x' como estimador de T.

<u>Definición 4.4</u>. <u>Precisión</u>. La precisión de un estimador es su error estándar; es decir, el valor de $\sigma_{x'}$ que se obtiene de la ecuación (4.7).

La precisión es una medida de la capacidad de repetición de resultados. Conceptualmente, es una medida de la dispersión de estimaciones que se generarían por muchas repeticiones de los mismos procedimientos de muestreo y de estimación, en las mismas condiciones. Con respecto a la distribución en el muestreo, es una medida de la dispersión de las estimaciones alrededor del centro de la distribución y no incluye ninguna

indicación de dónde se encuentra el centro de la distribución en relación con el valor verdadero.

En las ilustraciones 4.1, 4.2 y 4.3 se supuso implícitamente que el valor verdadero era $\bar{X}$; es decir, T era igual a $\bar{X}$. Por lo tanto, B era cero y el error cuadrático medio de x' era lo mismo que la variancia de x'. En las ilustraciones 4.4 y 4.5, el campo se ensanchó un poco con la introducción del error de respuesta y con el examen teórico del efecto del error de respuesta sobre $E(x')$ y $\sigma_{x'}$. En la práctica, muchos factores pueden influir en la distribución en el muestreo de x'. Esto es, los datos en una muestra están sujetos a errores que pueden ser atribuidos a varias fuentes.

Con los datos de una muestra se calcula una estimación x' y una estimación de la variancia de x'. ¿Cómo interpretar los resultados? En las ilustraciones 4.4 y 4.5 descubrimos que el error de respuesta puede dividirse entre sesgo y variancia. El error de cualquier fuente puede ser, por lo menos conceptualmente, dividido entre sesgo y variancia. Una estimación de una muestra está sujeta a la influencia combinada del sesgo y la variancia que correspondan a cada una de las varias fuentes de error. Cuando se utilizan los datos de una muestra para calcular una estimación de la variancia de x', la estimación es una combinación de variancias que podrían ser identificadas con varias fuentes. Asimismo, la diferencia entre $E(x')$ y T es una combinación de sesgos que podrían identificarse con varias fuentes.

La figura 4.2 muestra la distribución en el muestreo de x' correspondiente a cada uno de cuatro casos diferentes: A, sin sesgo y con error estándar bajo; B, sin sesgo y con error estándar grande; C, con sesgo grande y error estándar bajo; y C, sesgo grande y error estándar grande. La exactitud de un estimador a veces se define como la raíz cuadrada del

error cuadrático medio del estimador. Según esta definición, podríamos describir los estimadores que tienen las cuatro distribuciones en el muestreo de la figura 4.2 como sigue: En el caso A, el estimador es preciso y exacto; en B, el estimador no tiene precisión y por eso carece de exactitud; en C, el estimador es preciso pero inexacto a causa del sesgo; y en D, el estimador es inexacto a causa de sesgo y baja precisión.

Infortunadamente, generalmente no es posible determinar exactamente la magnitud del sesgo en una estimación, ni la de un componente particular del sesgo. Sin embargo, a menudo hay evidencia de la magnitud del sesgo, basada en la experiencia general, en el conocimiento de la perfección relativa de los procedimientos de la encuesta y en investigaciones especiales. El autor acepta el punto de visto de que el error cuadrático medio es un concepto apropiado de la exactitud. En ese contexto, lo importante es la magnitud del ECM y el valor de B en relación con $\sigma_{\bar{x}}$. Ese punto de vista es importante, porque no es posible estar seguro de que B es igual a cero. Nuestra meta debe ser preparar las especificaciones de la encuesta y llevar a cabo las operaciones de la misma de tal modo que sea pequeño B en relación con $\sigma_{\bar{x}}$. O se podría decir que queremos que el ECM sea mínimo para un costo dado de la encuesta. Las maneras de obtener evidencia sobre la magnitud del sesgo es un tema mayor y no está dentro del alcance de esta monografía.

Es importante saber algo acerca de la magnitud del sesgo, B, en relación con el error estándar, $\sigma_{\bar{x}}$. El error estándar depende principalmente del diseño de la muestra y su tamaño. En el caso de muchas poblaciones encontradas en las encuestas, a medida que aumenta el tamaño de la muestra, el error estándar se vuelve pequeño en comparación con el sesgo. De hecho, el sesgo podría ser más grande que el error estándar aún en el

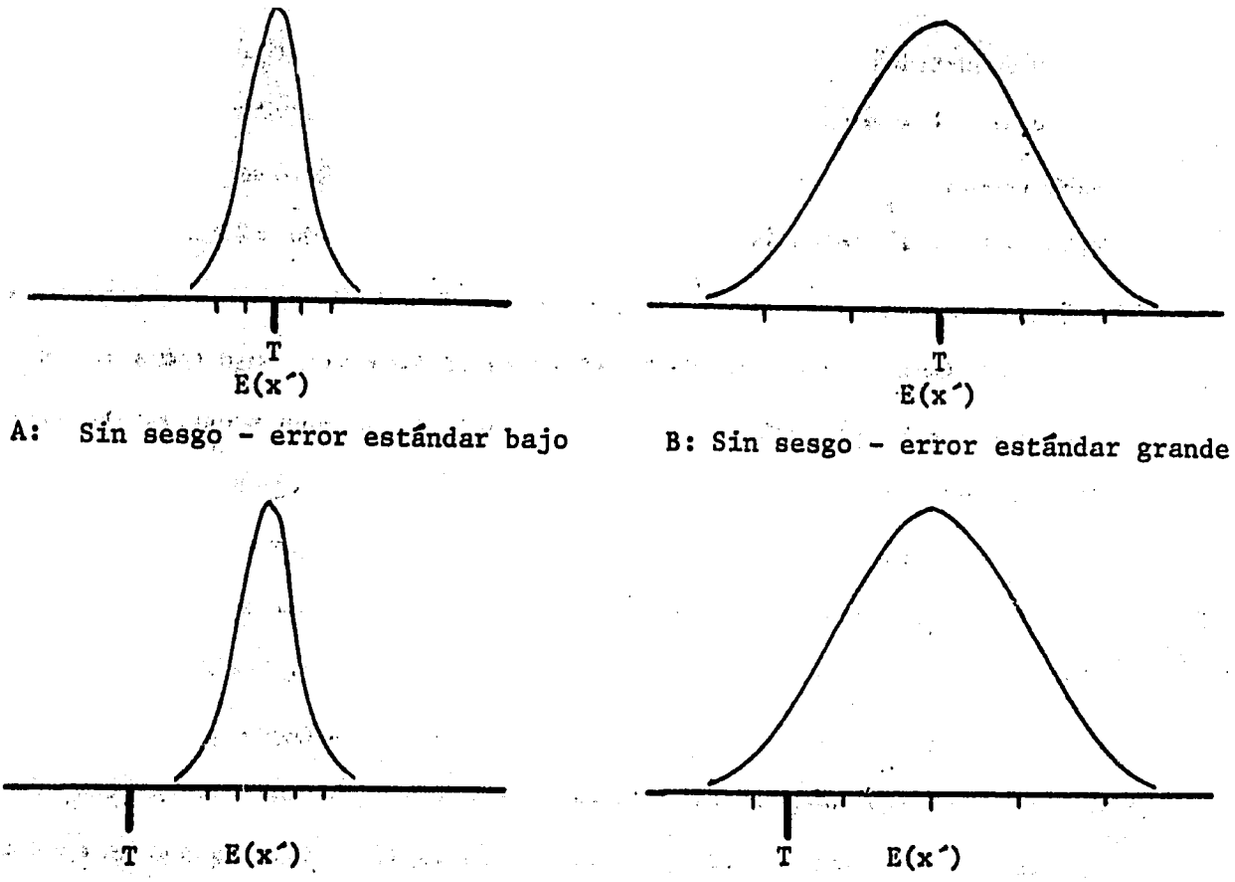


Figura 4.2. Ejemplos de cuatro distribuciones en el muestreo

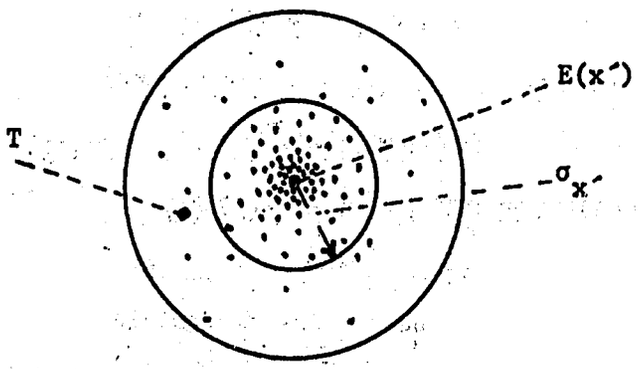


Figura 4.3. Distribución en el muestreo
(cada puntito representa una estimación)

caso de muestras de tamaño mediano -- por ejemplo, unos pocos centenares de casos, según las circunstancias. El punto importante es que, si el error cuadrático medio ha de ser pequeño, tanto B como $\sigma_{x'}$ tienen que ser pequeños. Los métodos para reducir B son muy distintos de los necesarios para reducir $\sigma_{x'}$. La mayor preocupación respecto al error ajeno al muestreo es el sesgo y no el efecto sobre la variancia. En el diseño y selección de muestras y en todos los procedimientos de la encuesta, se hace un esfuerzo en evitar sesgos que originan en el muestreo. Sin embargo, en los trabajos de encuestas es necesario tener siempre en mente la existencia de sesgos potenciales y procurar minimizar tanto los sesgos como el error aleatorio (es decir, $\sigma_{x'}$).

Los comentarios anteriores son aplicables en parte a los censos también. Los resultados de ambos tienen error cuadrático medio. Ambos son encuestas con respecto a la utilización de los resultados. Es decir, ocurren inferencias inciertas en el uso de los resultados de un censo lo mismo que en el uso de una muestra. La única diferencia es que en un censo se procura obtener una medida para cada uno de los N elementos, pero el poner n = N no reduce a cero el ECM. En realidad, no se tiene seguridad absoluta de que disminuya siempre el ECM cuando el tamaño de la muestra aumenta, porque a medida que disminuye la variancia puede aumentar el sesgo. Esto puede suceder especialmente cuando la población es grande y las preguntas del cuestionario son de tal naturaleza que es difícil obtener contestaciones sencillas y exactas. En el caso de una muestra grande o de un censo, podría ser más difícil controlar el error, en comparación con una muestra pequeña. Así es posible que un censo sea menos exacto (tenga un ECM mayor) que una muestra en que las fuentes de error sean mejor controladas. Mucho depende de la clase de información que se recolecte.

4.5 SESGO Y ERROR ESTANDAR

Los vocablos "sesgo", "sesgado" e "insesgado" poseen una gran variedad de significados entre distintos individuos. Como resultado, existe mucha confusión, especialmente porque los términos a menudo se usan con soltura. Técnicamente, parece lógico definir el sesgo de un estimador como igual a B en la ecuación (4.6), o sea, la diferencia entre el valor esperado del estimador y el valor verdadero. Pero, salvo en los casos hipotéticos o de ejemplos construidos, no se conocen valores numéricos para $E(x')$ y para T. Por eso, definir un estimador insesgado por la condición $B = E(x') - T = 0$ tiene muy poco o ningún valor práctico, salvo que se desee aceptar que el valor verdadero es igual a $E(x')$. Desde el punto de vista del muestreo, hay condiciones que dan una base racional para aceptar $E(x')$ como el valor verdadero. No obstante, sin importar la manera en que se defina el valor verdadero, se necesita una interpretación buena y práctica de $E(x')$.

Ha sido práctica común entre los estadísticos de encuestas llamar insesgada una estimación cuando está basada en métodos insesgados de muestreo y estimación. Por ejemplo, en la ilustración 4.4, $\bar{x}$ se llamaría una estimación insesgada, porque el método de muestreo y de estimación es insesgado. En otras palabras, puesto que $\bar{x}$ es una estimación insesgada de $\bar{X}$, se puede interpretar $\bar{x}$ como estimación insesgada del resultado que se habría obtenido si se hubieran medido todos los elementos de la población.

En la ilustración 4.5, el valor esperado de $\bar{x}$ es más difícil de describir. Sin embargo, con referencia al método de muestreo y de estimación, $\bar{x}$ es "insesgada" y podría ser calificada como "estimación insesgada" aunque $E(\bar{x})$ no sea igual a $\bar{T}$.

Una declaración que dice simplemente "la estimación es insesgada"

es incompleta y puede ser engañosa, especialmente si uno no está familiarizado con el contexto y con los conceptos de sesgo. Llamar insesgada una estimación equivale a decir que la estimación es una estimación insesgada de su valor esperado. Sin importar cómo se defina "sesgo" o cómo se use el término, $E(x')$ es el promedio de la distribución en el muestreo de x', y este concepto de $E(x')$ es muy importante, porque $E(x')$ aparece en el error estándar, $\sigma_{x'}$, de x' y también en B. Vea las ecuaciones (4.6) y (4.7).

Como un concepto o visualización sencilla del error de una estimación derivada de una encuesta, al autor le gusta la comparación de una estimación a un tiro a un blanco con arma de fuego o arco y flecha. Considere la repetición múltiple de una encuesta, usando el mismo plan de muestreo, pero escogiendo una muestra distinta cada vez. Este procedimiento generaría una distribución en el muestreo o un gran número de estimaciones; cada estimación correspondería a un tiro al blanco.

En la figura 4.3 cada puntito representa la estimación correspondiente a una muestra. El centro del grupo de puntos es marcado $E(x')$, porque corresponde al valor esperado del estimador. Alrededor del punto $E(x')$ hay un círculo que contiene dos tercios de los puntos. El radio de este círculo corresponde a $\sigma_{x'}$, el error estándar de la estimación. El círculo exterior tiene un radio de dos veces el error estándar y contiene el 95 por ciento de los puntos. El centro del blanco es marcado "T".

En la práctica, generalmente tenemos una sola estimación, x', y una estimación, $s_{x'}$, del error estándar de x'. Con referencia a la figura 4.3, esto quiere decir un punto y una estimación del radio del círculo alrededor de $E(x')$ que contendría dos tercios de las estimaciones que se obtendrían en muestreos repetidos. No sabemos el valor de T ni el valor de $E(x')$. Cuando hacemos una afirmación respecto al error estándar de x'

estamos expresando un grado de confianza respecto a la proximidad en que se encuentra la estimación derivada de la encuesta en relación con $E(x')$; es decir, cuán próximo se encontrará probablemente uno de los puntos en la figura 4.3 al punto desconocido, $E(x')$. Un juicio acerca de la distancia a que se encuentre $E(x')$ de T depende de cómo se define T y de una evaluación de la magnitud de los sesgos asociados con varias fuentes de error.

Infortunadamente, no es fácil hacer una afirmación corta, rigurosa, completa e interpretativa acerca del error estándar de x'. Si el error estándar estimado de x' es el 3 por ciento, se podría declarar simplemente ese hecho sin hacer interpretación. No es de mucha ayuda decir, por ejemplo, que la probabilidad es 2/3 de que la estimación esté dentro del 3% de su valor esperado, porque una persona conocedora de los conceptos ya sabe eso y probablemente la afirmación no sería de mucha ayuda a la persona que no conozca los conceptos. Suponga que uno diga: "el error estándar de x' quiere decir que hay una probabilidad 2/3 de que la estimación esté dentro del 3% del valor que se hubiera obtenido de un censo tomado en exactamente las mismas condiciones". Esto es un buen tipo de declaración, pero cuando uno repara en detalles más refinados, se necesita interpretar "un censo tomado en exactamente las mismas condiciones" - especialmente en vista del hecho de que es imposible tomar un censo en idénticas condiciones.

Para resumir, considere una encuesta como un sistema o procedimiento completamente definido que incluye todos los detalles que podrían afectar una estimación, como el método de muestreo, el método de estimación, las palabras usadas en el cuestionario, los procedimientos de entrevista, la selección, capacitación y supervisión de los entrevistadores, y la redacción, corrección y procesamiento de los datos. Conceptualmente, el muestreo

se repite muchas veces, manteniendo constantes todas las especificaciones y condiciones. Esto generaría una distribución en el muestreo como la ilustrada en las figuras 4.2 ó 4.3. Tenemos que darnos cuenta de que un cambio en cualquiera de las especificaciones o condiciones de la encuesta, por más leve que parezca el cambio, tiene la potencialidad de cambiar la distribución en el muestreo, especialmente el valor esperado de x'. Cambios en los planes de la encuesta, aunque no se cambien las definiciones de los parámetros que se estiman, a menudo resultan en diferencias que son mayores que el error aleatorio atribuible al muestreo.

La buena planificación y administración demandan una evaluación de errores de todas las fuentes y un empeño en equilibrar precisión y costos, con el fin de que el ECM caiga dentro de límites aceptables y a un costo mínimo.