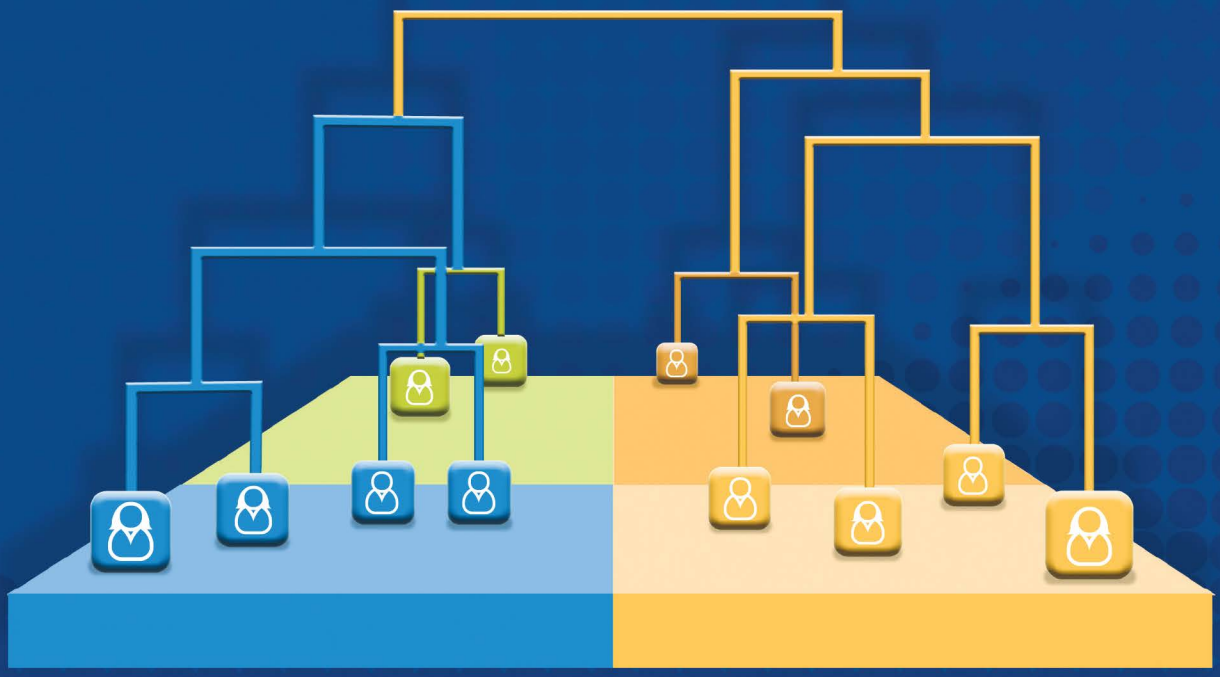
ANÁLISIS MULTIVARIADO DE DATOS

Métodos y aplicaciones

Primera edición revisada



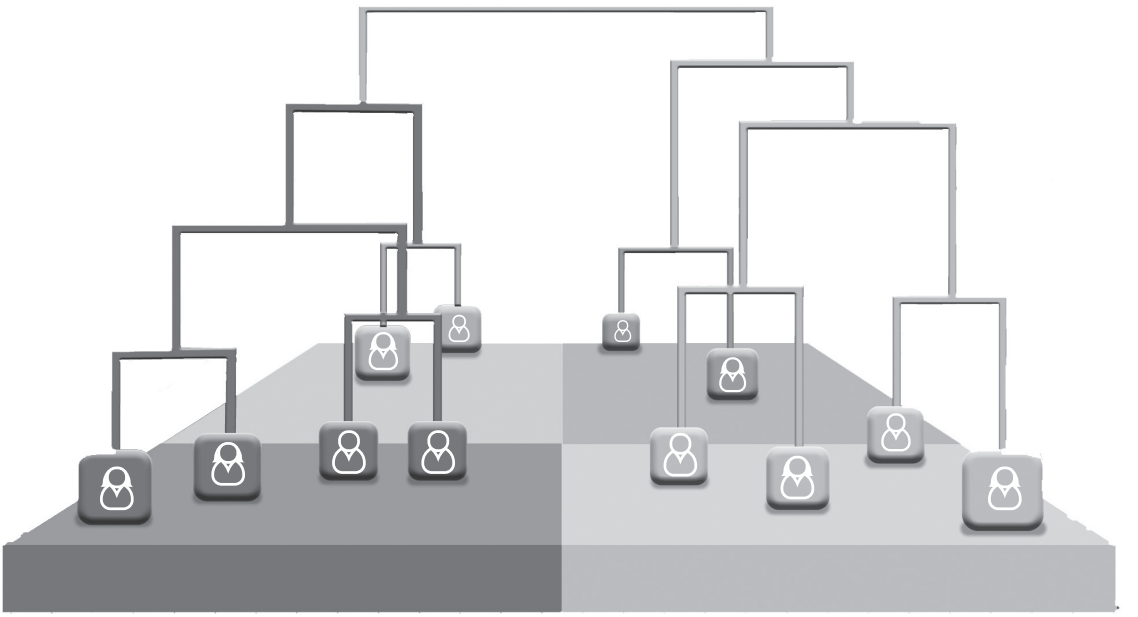
Javier Trejos Zelaya
William Castillo Elizondo
Jorge González Varela

EDITORIAL UCR

ANÁLISIS MULTIVARIADO DE DATOS

Métodos y aplicaciones

Primera edición revisada



Incluye archivos para ejercicios. Descárguelos aquí

Javier Trejos Zelaya
William Castillo Elizondo
Jorge González Varela

EDITORIAL UCR

519.535
T787a Trejos Zelaya, Javier

Análisis multivariado de datos : métodos y aplicaciones / Javier Trejos Zelaya, William Castillo Elizondo, Jorge González Varela. – Primera edición digital revisada. – San José, Costa Rica : Editorial UCR, 2021.

1 recurso en línea (xxx, 339 páginas) : ilustraciones en blanco y negro, archivo de texto, PDF, 19.5 MB.

"Incluye archivos para ejercicios"
ISBN 978-9968-02-015-2

1.ANÁLISIS MULTIVARIANTE. 2. ANÁLISIS MULTIVARIANTE – PROBLEMAS, EJERCICIOS, ETC. I. Castillo Elizondo, William, autor. II. González Varela, Jorge, autor. III. Título.

CIP/3729
CC.SIBDI.UCR

Las opciones de resaltado del texto, anotaciones o comentarios dependerán de la aplicación y dispositivo en que se realice la lectura de este libro digital.

Edición aprobada por la Comisión Editorial de la Universidad de Costa Rica
Primera edición impresa: 2014.
Primera edición digital revisada (PDF): 2021.

Editorial UCR es miembro del Sistema Editorial Universitario Centroamericano (SEDUCA), perteneciente al Consejo Superior Universitario Centroamericano (CSUCA).

Corrección filológica, revisión de pruebas, diseño, diagramación: *los autores* • Control de calidad de la versión impresa: *Grace Guzmán Aguilar* • Diseño de portada: *Floria Leiva* • Realización del PDF: *Alonso Prendas Vega* • Control de calidad de la versión digital: *Elisa Giacomin Valencia*.

Edición digital de la Editorial Universidad de Costa Rica. Fecha de creación: agosto, 2021
Universidad de Costa Rica. Ciudad Universitaria Rodrigo Facio. San José, Costa Rica.

Apdo. 11501-2060 • Tel.: 2511 5310 • Fax: 2511 5257
administracion.siedin@ucr.ac.cr • www.editorial.ucr.ac.cr

Tabla de contenidos

Capítulo 1

Estadística Descriptiva

1.1 Elementos de Estadística

La Estadística trata de estudiar datos producidos en diversas situaciones. El estudio de tales datos puede ser con diversos fines, dependiendo del campo de procedencia de los datos. Debido a la dificultad de extraer a simple vista toda la información que los datos poseen, o bien las tendencias que tienen, la Estadística trata de "entender" cuáles son las estructuras que los datos encierran intrínsecamente. La Estadística consiste entonces en una serie de técnicas útiles para el análisis de los datos producidos u observados, en un cierto contexto, de manera que se puedan hacer resúmenes útiles, tanto a través de índices numéricos como gráficos.

En la actualidad, los datos son producidos en casi todas las disciplinas y actividades del ser humano: en Ciencias Sociales, Economía, Mercadotecnia, Ciencias del Comportamiento, Ciencias Médicas, Ciencias Agrícolas, Física, Meteorología, Educación, Biología, Química, etc. En general, casi cualquier actividad humana donde haya variables numéricas repetidas provenientes de la observación o de la experimentación, es susceptible de ser ayudada por las técnicas estadísticas. También es cada vez más común que esas disciplinas acudan a las técnicas multivariadas, ya que la complejidad de la información que manejan obliga a un análisis más profundo que los simples promedios y porcentajes, que no muestran las interrelaciones entre las distintas variables observadas. El desarrollo actual de la computación y los desarrollos metodológicos han permitido abordar los grandes problemas del tratamiento de datos multivariados.

En el presente capítulo se introduce la notación y terminología sobre las que reposan las técnicas multivariadas, que se verán más adelante. Se empieza definiendo lo que se entiende por individuos y variables, así como la clasificación de éstas. También se recuerdan los principales aspectos de la estadística descriptiva simple [118], aunque se supone que la mayor parte de los lectores las maneja cómodamente.

1.1.1 Individuos o unidades estadísticas

Todo estudio estadístico se hace sobre un conjunto de *individuos*, que son el objeto de observación. Estos individuos u objetos de un análisis es lo que comúnmente se llaman unidades estadísticas. Una *unidad estadística* es la entidad sobre la que se quieren obtener los datos para ser analizadas.

Al conjunto de todas las unidades estadísticas se le llama *población*. Una parte de la población se llama una *muestra*, aunque estadísticamente se buscará trabajar con una muestra *representativa* de la población es el sentido de que mantenga la misma naturaleza de la población. En esta obra no se tratan asuntos relacionados con las muestras ni con la manera de seleccionarlas. El lector interesado puede consultar [53], [103] ó [30] para una descripción de los principales métodos de muestreo.

Ejemplo 1 *Supóngase que se quiere conocer las características de los asegurados al régimen de Seguridad Social, como por ejemplo su ocupación, su sexo, su estado civil, el número de hijos que tienen, etc. Entonces los individuos u objetos de estudio son los asegurados. La población es el conjunto de todos los asegurados, pero para un estudio particular se puede extraer una muestra.*

Ejemplo 2 *Supóngase que se quiere estudiar la eficiencia de las clínicas del país. Para ello, entre otras cosas, podría considerarse el conteo del número de médicos y demás personal que tiene cada clínica, el número de personas que ha atendido en un período de tiempo (por ejemplo, en el último año), el número de habitantes que tiene la comunidad a la que atiende, etc. Entonces los individuos u objetos del análisis son las clínicas.*

En el enfoque del Análisis de Datos, generalmente no se plantea el problema de si los datos provienen de una muestra o de toda la población. Dichosamente, los resultados del Análisis en Componentes Principales —y por

ende de los métodos que derivan de él— no dependen de esta diferencia en el origen de los datos, y lo mismo ocurre con los métodos de clasificación. Desde luego, una vez obtenidos los resultados de un método multivariado, se pueden hacer ciertas pruebas de hipótesis para chequear la calidad de éstos o para facilitar la interpretación. Ahora bien, lo importante es destacar que no se hacen supuestos a priori sobre la distribución de probabilidad de las variables, ya que se trabajará directamente con los datos observados de manera que se trate de buscar el modelo a partir de los datos, sin imponer un modelo desde el principio. En el caso de trabajar con una muestra, la validez de extrapolar los resultados obtenidos a toda la población, dependerá de la representatividad de la muestra. Es decir, dependerá de si la muestra fue obtenida por algún método confiable.

Pesos de los individuos

En algunos métodos de Análisis de Datos, es importante tomar en cuenta que las unidades estadísticas o individuos pueden tener distinta importancia en un estudio. A la importancia relativa que puede tomar un individuo, se le llama *peso* o *ponderación*.

Se supondrá que los n individuos están ponderados por pesos positivos p_1, p_2, ..., p_n tales que $p_1 + \cdots + p_n = 1$. En muchas ocasiones, estos pesos serán iguales para todos los individuos, en cuyo caso $p_i = 1/n$ para todo individuo i. Salvo que se especifique lo contrario, se supondrá que los pesos son iguales. Este es el caso usual en una encuesta. Por ejemplo, en caso de que se tengan 1000 individuos y todos con la misma importancia, entonces el peso de cada uno es 1/1000.

Ejemplo 3 *Si se quiere estudiar la evolución de los porcentajes de votación obtenidos por los distintos partidos políticos, según cada provincia, y se dispone únicamente de los porcentajes de votos obtenidos por cada partido, entonces las unidades estadísticas son las provincias y la ponderación de cada provincia será el número de votantes de la misma, dividida entre el número total de individuos. Por ejemplo, supóngase que se sabe que el partido PXY obtuvo en la pasada elección 42% de los votos en Limón, 47% en Puntarenas, 52% en Cartago, etc., entonces para calcular el porcentaje obtenido en el país se sumarán los porcentajes anteriores, pero ponderados por la población respectiva. Así, si el número de votantes de Limón es 234,789, en Puntarenas es 287,376 y en Cartago es 312,545, entonces se multiplicará 42% por 234,789, 47% por 287,376, 52%*

por 312,545 y así sucesivamente. Más adelante se verá cómo calcular promedios ponderados y otros índices cuando los pesos no son iguales.

1.1.2 Las variables de la estadística

Una **variable** en estadística, es aquello que se observa o mide sobre las unidades estadísticas. Para cada individuo puede tomar un valor distinto, de ahí su nombre. En términos matemáticos, se puede definir como una función $\mathbf{x}$ del conjunto de individuos Ω a un conjunto de valores. En vista de que los valores que puede tomar $\mathbf{x}$ *varían* en ese conjunto, entonces se le da el nombre de *variable*[1] a $\mathbf{x}$.

Ejemplo 4 *Se quiere estudiar una serie de características físicas de un grupo de personas. Entonces resultará de interés medir su estatura, su peso, el perímetro del cráneo, su sexo, etc. Todas estas son variables: por ejemplo, el peso puede tomar valores diferentes para todas las personas, o bien, puede ocurrir que algunas de ellas tengan el mismo peso, pero otras lo tengan diferente. Lo importante es que no tienen el mismo valor.*

Dependiendo de la naturaleza del conjunto de posibles valores de la variable se distinguen dos tipos principales de variables: las *cuantitativas* y las *cualitativas*.

Variables cuantitativas

Una variable se llama **cuantitativa** o **numérica** cuando sus valores son números, ya sea reales o enteros.

Ejemplo 5 *Son variables cuantitativas el peso, la edad y la estatura de una persona, la temperatura de una habitación, el número de camas de un hospital.*

Entre las variables cuantitativas se distinguen dos tipos:

[1] Debe observarse que este término estadístico, que proviene del lenguaje de la probabilidad *variable aleatoria*, no debe confundirse con el término usual en matemática de variable, en cuyo contexto se usa ese término para designar a los elementos del dominio de la función (en este caso, a los individuos), mientras que en la definición estadística se usa el término variable para designar a la función misma.

- Las variables **continuas** son aquéllas que pueden tomar como valores cualquier número real, es decir, un valor con decimales. Siempre es importante plantearse el asunto de las unidades de medida de una variable continua, ya que en algunos casos pueden influir en los resultados de un análisis. Por ejemplo, son variables continuas el peso, la estatura, la temperatura, un porcentaje. Puede observarse que las variables cuantitativas continuas tienen una *unidad de medida*. Es decir, se miden en alguna unidad que permita tener una idea de qué tanto posee un individuo de la característica representada por la variable.

  Ejemplo 6 *El peso*[2] *puede medirse en kilogramos, libras, miligramos, toneladas, etc. El uso de una unidad dependerá de la naturaleza de las unidades estadísticas. Por ejemplo, si se trata de personas, entonces el peso se mediría en kilógramos o libras; si se trata de pastillas contra el dolor de cabeza y se quiere medir el peso de la aspirina contenida, entonces éste se puede medir en miligramos; si se trata de exportaciones de café, entonces éstas se pueden medir en toneladas. Se debe notar que en algunos análisis que se estudiarán más adelante, las unidades de medida pueden tener influencia en los resultados, por lo que se recomienda tener cuidado en la escogencia de las mismas.*

- Las variables **discretas**, también llamadas de **conteo**, son aquéllas que sólo pueden tomar valores discretos, es decir, números enteros positivos. Son variables discretas, por ejemplo, el número de estudiantes en un aula, el número de hijos de un familia, la edad (dada en años cumplidos). Las variables discretas tienen por lo general los números enteros positivos como unidad de medida (aunque por ejemplo la edad tiene unidad de medida el número de años).

Variables cualitativas

Si la variable puede tomar su valor solamente en un conjunto finito de posibilidades, tales que todas ellas significan una *cualidad* o *atributo*, entonces se llama una **variable cualitativa** o **categórica**.

Ejemplo 7 *El sexo de una persona es una variable cualitativa, pues un individuo solo puede tener dos cualidades para esta variable: masculino o femenino.*

[2]Para efecto didácticos, a lo largo de este texto se usa el concepto de peso y masa como sinónimos.

Ejemplo 8 *El estado civil de una persona es una variable cualitativa, pues un individuo solo puede tener una de seis cualidades: soltero, casado, divorciado, separado, viudo, en unión libre. Es decir, hay un conjunto finito de posibilidades, todas ellas excluyentes entre sí.*

Las posibles cualidades que tiene una variable cualitativa, se llaman las **modalidades** de la variable. Algunos autores las llaman también **categorías** o **atributos**.

Se distinguen tres tipos de variables cualitativas:

- Si las modalidades están ordenadas, entonces la variable se llama **ordinal**. Por ejemplo, la variable *nivel de estudios* es ordinal, ya que sus modalidades están ordenadas según la duración de los estudios: las modalidades podrían ser por ejemplo sin educación, educación primaria incompleta o completa, educación secundaria incompleta o completa, técnica, universitaria.

- Si las modalidades no están ordenadas, entonces la variable se llama **nominal**. Por ejemplo, el sexo o el estado civil son variables nominales ya que las modalidades de estas variables no tienen un orden lógico.

- Un caso especial de variable cualitativa nominal es cuando se tienen solo dos modalidades que reflejan la presencia o la ausencia de una cualidad; este tipo de variables se llaman **binarias**, **dicotómicas** o de **presencia–ausencia**. Por ejemplo, si un paciente tiene o no tiene una determinada enfermedad.

Codificación de variables cualitativas. Para las variables cualitativas, un aspecto de suma importancia es el de la *codificación*. Por ejemplo, para la variable sexo, se puede pensar en codificar la modalidad "femenino" como 1 y la modalidad "masculino" como 0. Sin embargo, es claro que tal escogencia es totalmente arbitraria, ya que perfectamente se pudo haber escogido 1 para femenino y 2 para masculino, o cualquier otra cosa, siempre que el código asignado a cada modalidad sea *diferente* con el fin de no crear ninguna ambigüedad. La codificación es en general necesaria en vista de que la mayoría de los programas de computación manipulan información numérica. Por lo tanto, este es un asunto al que hay que prestarle la mayor importancia en el momento de elaborar un cuestionario y de tabularlo.

Ejemplo 9 *En la práctica, la variable sexo puede aparecer codificada de la siguiente manera:*

Estudiante	Sexo
Ana	1
Juan	2
Pedro	2
Carmen	1
Luis	2

o bien

Estudiante	Sexo
Ana	1
Juan	0
Pedro	0
Carmen	1
Luis	0

Hay muchos programas estadísticos que necesitan hacer una *codificación disyuntiva completa*, esto es, poner una columna completa para cada modalidad.

Ejemplo 10 *El ejemplo anterior (9) quedaría codificado en forma disyuntiva completa como sigue:*

	Sexo	
Estudiante	Feme.	Masc.
Ana	1	0
Juan	0	1
Pedro	0	1
Carmen	1	0
Luis	0	1

Las columnas de la tabla del ejemplo 10, se llaman las *indicatrices* o *indicadoras* de cada modalidad: un 1 indica que el individuo correspondiente posee la modalidad y un 0 que no la posee. Es claro que las modalidades de la variable cualitativa definen una partición sobre el conjunto de individuos. La partición en el ejemplo 10 sería {Ana,Carmen}, {Juan,Pedro,Luis}.

Como para toda partición, las clases tienen asociada una función característica: esta función es precisamente la indicatriz de la modalidad.

Discretización de una variable cuantitativa. Por otro lado, es claro que cualquier variable cuantitativa puede "codificarse" como variable cualitativa, estableciendo niveles en el rango de la variable cuantitativa. Este proceso se conoce como discretización de una variable cuantitativa.

Ejemplo 11 *Para ciertos análisis, puede ser más útil manipular el salario como variable cualitativa que como cuantitativa: se puede entonces pensar en establecer categorías de salario, como muy bajo (menos de 200 dólares), bajo (entre 200 y menos de 500 dólares), medio (entre 500 y menos de 2 000 dólares), alto (entre 2 000 y menos de 4 000 dólares) y muy alto (más de 4 000 dólares).*

Al hacer una codificación como la anterior, se pierde la estructura algebraica de $\mathbb{R}$ pero se mantiene la de orden y quizás se gana en síntesis. La utilidad de una codificación como ésta sólo se verá a la luz de los objetivos del estudio y las herramientas de que se disponga.

Algunos autores llaman a las variables *escalas de medida* y las clasifican en cuatro tipos: nominal, ordinal, intervalo y razón. Los dos primeros grupos corresponden a las variables cualitativas que se han llamado de la misma manera. Las variables tipo intervalo son variables cuantitativas tales que no existe un cero absoluto y no existen operaciones de multiplicación y división, además la distancia entre los números de la escala es igual; un ejemplo es la temperatura. Las variables de tipo razón son las variables cuantitativas que cuentan con un cero absoluto que representa la ausencia total de medida y se puede realizar cualquier operación aritmética o lógica; algunos ejemplos son peso, estatura o salario. En este texto no se usará esta clasificación de las variables.

1.2 Tablas de datos

Para hacer un análisis de datos, generalmente se disponen los datos en arreglos rectangulares en forma de matriz, llamados tablas de datos. En ellas, las filas y columnas describen a individuos o variables, según sea el caso. A continuación se presentan los principales tipos de tablas de datos.

1.2.1 Tablas de individuos $\times$ variables

En las tablas de individuos por variables, los individuos se asocian con las filas y las variables con las columnas. Esto es, cada fila representa a un individuo y cada columna representa a una variable.

Ejemplo 12 *Considérese que se han observado 7 variables cuantitativas sobre un grupo de 10 estudiantes. Las primeras cinco variables son las notas*

obtenidas por los estudiantes en cinco materias: Matemáticas (Mate), Ciencias (Cien), Español (Espa), Historia (Hist) y Educación Física (EdFi), todas ellas en escala de 0 a 10, y las otras dos variables son el peso del estudiante (medido en libras) y la estatura (medida en centímetros). Los datos se presentan en la Tabla 1.1.

Estudiante	Mate	Cien	Espa	Hist	EdFi	Peso (lbs.)	Estatura (cms.)
Lucía	7.0	6.5	9.2	8.6	8.0	126	162
Pedro	7.5	9.4	7.3	7.0	7.0	140	168
Inés	7.6	9.2	8.0	8.0	7.5	130	169
Luis	5.0	6.5	6.5	7.0	9.0	150	172
Andrés	6.0	6.0	7.8	8.9	7.3	142	165
Ana	7.8	9.6	7.7	8.0	6.5	128	165
Carlos	6.3	6.4	8.2	9.0	7.2	144	170
José	7.9	9.7	7.5	8.0	6.0	134	165
Sonia	6.0	6.0	6.5	5.5	8.7	135	170
María	6.8	7.2	8.7	9.0	7.0	128	166

Tabla 1.1: Tabla de datos de las notas escolares con peso y estatura.

Ejemplo 13 *Al realizar una encuesta, normalmente se disponen los datos en una tabla de individuos × variables. Considérese que en una encuesta se ha recogido información como el nombre, el sexo, la edad, el estado civil, el número de hijos, el ingreso mensual bruto, etc. Entonces la tabla de datos tendría una forma como la mostrada en la Tabla 1.2.*

Número de encuestado	Sexo	Edad (años)	Estado civil	Número de hijos	Ingreso mensual (colones)	...
001	M	34	Casado	1	356.000	...
002	F	24	Soltera	0	188.000	...
003	F	52	Divorciada	2	141.000	...
004	M	46	Soltero	0	170.000	...
005	F	38	Casada	3	592.000	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabla 1.2: Parte de una tabla de datos proveniente de una encuesta.

Generalmente, en una tabla de datos como la Tabla 1.2 se codifican las variables cualitativas, de modo que la tabla puede ser almacenada en una base de

datos o archivo numérico, para lo cual el usuario debe señalar los códigos asociados a cada modalidad. Dependiendo del software estadístico que se use, el mismo permitirá cierta forma de hacer esta codificación.

Supóngase que se tienen n individuos descritos por p variables. Se denota $\mathbf{X}$ a una tabla de datos de filas $\times$ columnas, entonces $\mathbf{X}$ es un matriz que tiene n filas y p columnas.

La tabla de datos se puede ver como sigue:

$$\mathbf{X} = \begin{array}{c|cccccc} & \mathbf{x}^1 & \mathbf{x}^2 & \cdots & \mathbf{x}^j & \cdots & \mathbf{x}^p \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & & & & \vdots \\ \mathbf{x}_i & x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & & & & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{array}$$

En general, en la notación x_{ij} para la entrada (i, j) el primer subíndice denota al individuo i y el segundo subíndice a la variable j. Por otro lado, se distinguirá a los individuos de las variables poniendo subíndice a los individuos y superíndice a las variables: $\mathbf{x}_i$ denota al individuo i y $\mathbf{x}^j$ denota a la variable j.

Este tipo de tablas serán usadas más adelante, en técnicas como el Análisis en Componentes Principales, y la Clasificación Automática. En algunos casos especiales, también se podría usar el Análisis Factorial de Correspondencias.

1.2.2 Tablas de variables $\times$ variables

Se trata de tablas en que tanto las filas como las columnas describen a variables, o a modalidades de éstas en el caso cualitativo. Es el caso de las tablas de contingencia que se analizan en Análisis Factorial de Correspondencias o las tablas de Burt para Análisis de Correspondencias Múltiples.

Sean $\mathbf{x}$ y $\mathbf{y}$ dos variables cualitativas que poseen respectivamente las modalidades $\mathbf{x}^1$, $\mathbf{x}^2$, ..., $\mathbf{x}^p$ y $\mathbf{y}^1$, $\mathbf{y}^2$, ..., $\mathbf{y}^q$. Se puede entonces construir la tabla estadística siguiente, que resulta de cruzar las variables $\mathbf{x}$ y $\mathbf{y}$: la entrada (j, k) de la tabla representa el número de individuos que poseen simultáneamente las

modalidades $\mathbf{x}^j$ y $\mathbf{y}^k$. Ese número se denotará x_{jk}. Una tabla de datos construida de esta forma se llama *tabla de contingencia* o *tabla cruzada*, también conocida como tabla de doble entrada. El análisis de este tipo de tablas se introduce en la sección 1.3, página 17, y se desarrolla en el capítulo 4 sobre Análisis Factorial de Correspondencias.

Ejemplo 14 *En una encuesta se ha preguntado por el nivel de estudios de un conjunto de 1 200 presonas, así como por su nivel de ingresos. Los ingresos han sido codificados de la siguiente forma:*

salario muy bajo:	*menos de 200 dólares mesuales*
salario bajo:	*entre 200 y menos de 500 dólares mensuales*
salario medio:	*entre 500 y menos de 2 000 mesuales*
salario alto:	*entre 2 000 y menos de 4 000 mensuales*
salario muy alto:	*4 000 mensuales o más.*

Entonces los datos se han dispuesto en una tabla tal que cada casilla contiene el número de personas entrevistadas con determinado nivel de estudios y determinado nivel de salario. La tabla de contingencia obtenida se muestra en la Tabla 1.3.

Nivel de estudios	Nivel de salario				
	Bajo	Medio	Alto	Muy alto	Total
Ninguno	200	21	2	0	223
Primario	217	45	5	6	273
Técnico	156	105	46	32	339
Secundario	73	93	24	2	192
Universitario	6	86	52	29	173
Total	652	350	129	69	1200

Tabla 1.3: Tabla de contingencia que cruza el nivel de salario con el nivel de estudios.

Ejemplo 15 *Una tabla, o matriz, de Burt cruza varias variables cualitativas. Considere el caso de una encuesta efectuada en el año 2004 en el Instituto Tecnológico de Costa Rica (ITCR), acerca del servicio del comedor. Para ilustrar este ejemplo, se consideran solamente 4 de las variables medidas sobre 147 encuestados:*

Edad: *se dispone de la edad en años cumplidos de los encuestados, y se ha recodificado en 4 intervalos:*

- *1: de 25 años o menos,*
- *2: entre 26 y 35 años,*
- *3: entre 36 y 45 años,*
- *4: de 46 años o más.*

Sexo: *el género de cada encuestado se ha codificado en femenino (f) o masculino (m).*

Ocupación: *esta variable tiene 3 categorías:*

- *E: estudiante,*
- *D: docente,*
- *A: administrativo.*

Rapidez: *Se pregunta la opinión acerca de la rapidez del servicio de la Soda-Comedor del ITCR, con 4 modalidades (nadie repondió "muy buena"):*

- *MM: muy mala,*
- *M: mala,*
- *R: regular,*
- *B: buena.*

La tabla de Burt generada, para estas 4 variables, se muestra en la Tabla 1.4. Más adelante, se retomará este ejemplo para ilustrar algunas de las técnicas de análisis multivariado.

1.2.3 Tablas de individuos $\times$ individuos

Se trata de tablas que tienen tanto por filas como por columnas a individuos. Un caso típico es una tabla de distancias: en la entrada (i, i') de la matriz se tiene la distancia calculada entre el individuo i y el individuo i', denotada $d(\mathbf{x}_i, \mathbf{x}_{i'})$. Estas tablas son muy usadas en Clasificación Automática y en Escalamiento Multidimensional.

	Edad				Sexo		Ocupación			Rapidez			
	1	2	3	4	F	M	E	D	A	MM	M	R	B
< 25	93	0	0	0	36	57	92	0	1	0	9	42	42
26-35	0	18	0	0	6	12	4	6	8	0	0	9	9
36-45	0	0	24	0	13	11	0	7	17	1	0	12	11
> 45	0	0	0	12	2	10	0	6	6	0	1	4	7
F	36	6	13	2	57	0	37	3	17	1	4	24	28
M	57	12	11	10	0	90	59	16	15	0	6	43	41
E	92	4	0	0	37	59	96	0	0	0	9	43	44
D	0	6	7	6	3	16	0	19	0	0	0	13	6
A	1	8	17	6	17	15	0	0	32	1	1	11	19
MM	0	0	1	0	1	0	0	0	1	1	0	0	0
M	9	0	0	1	4	6	9	0	1	0	10	0	0
R	42	9	12	4	24	43	43	13	11	0	0	67	0
B	42	9	11	7	28	41	44	6	19	0	0	0	69

Tabla 1.4: Tabla de Burt entre 4 de las variables acerca del servicio de comedor del ITCR.

Ejemplo 16 *Considérese la Tabla 1.5 de datos que muestra la distancia en línea recta (en kilómetros) entre algunas ciudades de Costa Rica: San José (S.J.), Alajuela (Ala.), Cartago (Car.), Heredia (Her.), Puntarenas (Pun.), Limón (Lim.), Liberia (Lib.) y Golfito (Gol.). Es un ejemplo típico de tabla de individuos $\times$ individuos, donde los individuos son las ciudades.*

	S.J.	Ala.	Car.	Her.	Pun.	Lim.	Lib.	Gol.
San José	0	18.0	18.0	9.0	82.5	114.0	168.0	172.5
Alajuela	18.0	0	36.0	10.5	67.5	127.5	150.0	184.5
Cartago	18.0	36.0	0	25.5	99.0	97.5	186.0	157.5
Heredia	9.0	10.5	25.5	0	78.0	118.5	160.5	181.5
Puntarenas	82.5	67.5	99.0	78.0	0	195.0	97.5	232.5
Limón	114.0	127.5	97.5	118.5	195.0	0	271.5	150.0
Liberia	168.0	150.0	186.0	160.5	97.5	271.5	0	330.0
Golfito	172.5	184.5	157.5	181.5	232.5	150.0	330.0	0

Tabla 1.5: Tabla de datos con la distancia entre algunas ciudades.

Ejemplo 17 *Se dispone de una matriz de datos donde 10 estudiantes de sexto grado han calificado la afinidad que tienen por cada uno de sus compañeros. Por filas se tienen las notas que asignan los estudiantes, entre 1 y 5, y por columnas*

las notas que les son asignadas por sus compañeros. Una matriz de este tipo es llamada una sociomatriz.El grupo está formado por 5 mujeres y 5 varones. En la diagonal, se han colocado las notas máximas, para la calificación de un estudiante a sí mismo. Los datos se presentan en la Tabla 1.6.

	Iren	Flor	Beat	Silv	Hele	Anto	Migu	Fede	Este	Dieg
Irene	5	4	5	2	3	2	2	2	3	2
Flor	5	5	4	3	4	3	3	3	4	3
Beatriz	4	5	5	2	3	3	3	4	3	3
Silvia	2	4	5	5	5	2	3	3	4	3
Helena	3	4	4	5	5	1	2	2	2	1
Antonio	1	3	1	2	1	5	5	2	3	2
Miguel	2	4	3	2	2	5	5	2	3	3
Federico	3	4	4	3	3	3	3	5	4	4
Esteban	2	5	3	3	3	4	4	4	5	3
Diego	2	4	3	3	2	4	3	3	5	5

Tabla 1.6: Tabla de datos: sociomatriz en que 10 estudiantes de sexto grado califican la afinidad hacia cada uno de sus compañeros.

1.3 Análisis estadísticos univariados y bivariados

Siempre que se haga un análisis de datos, es imprescindible tener un conocimiento profundo del comportamiento individual de cada variable. Aún si el objetivo es hacer un análisis multivariado, esta etapa previa de profundización es indispensable.

Para ello, se han definido varios índices que miden este comportamiento y se han diseñado varias técnicas, en su mayoría con apoyo gráfico, para tener una mejor visión de lo que mide o explica cada variable. Se puede decir que lo que se quiere es un resumen numérico y un resumen gráfico de la variable. A este tipo de análisis se le llama análisis de una variable o **análisis univariado**.

Según sea la naturaleza de las variables hay diferentes tipos de análisis univariados que se pueden hacer. A continuación se recordarán rápidamente los principales índices y gráficos univariados y bivariados. El lector interesado puede consultar más ampliamente sobre este punto en [47] o bien [118].

Análisis estadístico univariado. Si la variable a analizar es cuantitativa, se medirán su *tendencia central* y su *dispersión*. Entre las primeras se cuentan la

media, la mediana y la media de los valores extremos. Los cuartilos y percentiles permiten tener una idea del comportamiento de una variable según su orden. Entre las principales medidas de dispersión están la desviación estándar (y su cuadrado, la varianza), la desviación media, la desviación mediana, la desviación cuartil y la extensión. El coeficiente de variación es el cociente de la desviación estándar entre la media.

Si $x_1, x_2, \ldots, x_n$ son las observaciones de la variable cuantitativa $\mathbf{x}$ y $p_1, p_2, \ldots, p_n$ son las ponderaciones[3] de los individuos, es usual denotar su media por $\bar{\mathbf{x}}$:

$$\bar{\mathbf{x}} = \sum_{i=1}^{n} p_i x_i, \tag{1.1}$$

su desviación estándar $\sigma_{\mathbf{x}}$:

$$\sigma_x = \sqrt{\sum_{i=1}^{n} p_i (x_i - \bar{\mathbf{x}})^2} = \sqrt{\sum_{i=1}^{n} p_i x_i^2 - \bar{\mathbf{x}}^2} \tag{1.2}$$

y su varianza $\operatorname{var}(\mathbf{x})$:

$$\operatorname{var}(\mathbf{x}) = \sigma_x^2. \tag{1.3}$$

Si la variable a analizar es cualitativa o binaria, se calculan las frecuencias (absolutas y relativas) de cada modalidad, y en caso de ser ordinal la variable, es usual calcular también las frecuencias acumuladas.

Los principales gráficos asociados a una variable cuantitativa son generalmente los histogramas (que pasan por la escogencia de un número de clases en las que se distribuyen las observaciones, construyéndose una distribución de frecuencias), las cajas de dispersión (o *boxplot*) y los diagramas tallo–hoja. Estos gráficos permiten tener una idea de la dispersión y distribución de los datos [118].

En el caso de una variable cualitativa, hay una serie de gráficos que se usan, dependiendo de los intereses en la descripción, como los gráficos de barras, de bastones o circulares, que representan proporcionalmente a las frecuencias.

[3] Recuérdese que en el caso usual, se tiene $p_i = 1/n$.

Análisis estadístico bivariado. El análisis bivariado consiste en el estudio de las relaciones entre parejas de variables, y también forma parte de la descripción simple de una tabla de datos.

En el caso de tener dos variables cuantitativas, se suele hacer el diagrama de dispersión, el cual grafica en ejes de abscisas y de ordenadas a las dos variables, y permite ver la asociación entre ellas. El **coeficiente de correlación lineal** —también llamado coeficiente de correlación de Pearson— denotado r, es una cuantificación de la relación entre dos variables cuantitativas $\mathbf{x}$ y $\mathbf{y}$:

$$r(\mathbf{x},\mathbf{y}) = \frac{\text{cov}\,(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}},$$

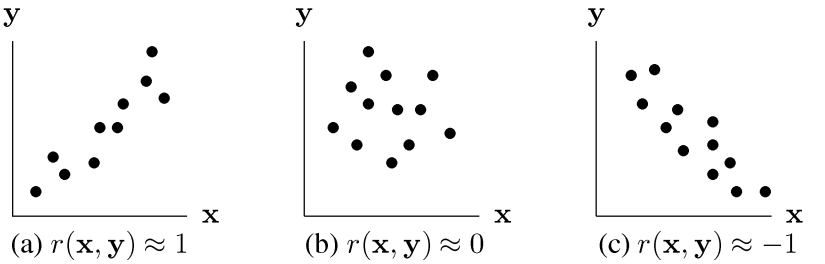
donde la covarianza es:

$$\text{cov}\,(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} p_i(x_i-\bar{\mathbf{x}})(y_i-\bar{\mathbf{y}}) = \sum_{i=1}^{n} p_i x_i y_i - \bar{\mathbf{x}}\bar{\mathbf{y}}.$$

Posteriormente se probará que $-1 \leq r(\mathbf{x},\mathbf{y}) \leq 1$. El coeficiente de correlación lineal se interpreta así:

- Si $r(\mathbf{x},\mathbf{y}) \approx 1$: hay una fuerte correlación directa, esto significa que a valores grandes de $\mathbf{x}$ corresponden valores grandes de $\mathbf{y}$ y que a pequeños valores de $\mathbf{x}$ corresponden pequeños valores de $\mathbf{y}$. Es decir, las variables tienen un comportamiento similar sobre todos los individuos. Lo anterior se puede ilustrar en el diagrama de dispersión que se muestra en la Figura 1.1(a).
- Si $r(\mathbf{x},\mathbf{y}) \approx 0$: no hay correlación, esto significa que a valores grandes de $\mathbf{x}$ corresponden tanto valores grandes como pequeños de $\mathbf{y}$, y que a valores pequeños de $\mathbf{x}$ también corresponden valores grandes como pequeños de $\mathbf{y}$. Es decir, el comportamiento de las variables no tiene ninguna relación entre sí. Ver la Figura 1.1(b).
- Si $r(\mathbf{x},\mathbf{y}) \approx -1$: hay una fuerte correlación inversa, lo que significa que a valores grandes de $\mathbf{x}$ corresponden valores pequeños de $\mathbf{y}$, y a valores pequeños de $\mathbf{x}$ correponden valores grandes de $\mathbf{y}$. Es decir, las variables tienen un comportamiento opuesto una de la otra. Ver la Figura 1.1(c).

Más adelante se hará la interpretación geométrica del coeficiente de correlación lineal, como un coseno, lo que es muy importante para el desarrollo posterior del análisis multivariado.



(a) $r(\mathbf{x},\mathbf{y}) \approx 1$ (b) $r(\mathbf{x},\mathbf{y}) \approx 0$ (c) $r(\mathbf{x},\mathbf{y}) \approx -1$

Figura 1.1: El coeficiente de correlación lineal muestra el tipo de relación entre dos variables cuantitativas

Si las dos variables son cualitativas, entonces se suele estudiar la independencia entre las modalidades de las dos variables mediante un índice de asociación, que usualmente es el índice de chi–cuadrado (denotado χ^2); existe una técnica factorial para el análisis del tipo de dependencia entre las modalidades, que es el Análisis Factorial de Correspondencias, que además provee gráficos de fácil lectura. Esta técnica será presentada en el capítulo 4.

Sean $\mathbf{x}$ y $\mathbf{y}$ dos variables cualitativas que poseen respectivamente las modalidades $\mathbf{x}^1$, $\mathbf{x}^2$, …, $\mathbf{x}^p$ y $\mathbf{y}^1$, $\mathbf{y}^2$, …, $\mathbf{y}^q$. Denotando x_{jk} la entrada (j,k) de la tabla de contingencia, entonces se escribe, según sea el caso, el total de la modalidad $\mathbf{x}^j$:

$$x_{j\cdot} = \sum_{k=1}^{q} x_{jk},$$

el total de la modalidad $\mathbf{y}^k$:

$$x_{\cdot k} = \sum_{j=1}^{p} x_{jk},$$

y el tamaño de la muestra o de la población total:

$$x_{\cdot\cdot} = n = \sum_{j=1}^{p} x_{j\cdot} = \sum_{k=1}^{q} x_{\cdot k} = \sum_{j=1}^{p}\sum_{k=1}^{q} x_{jk}.$$

Denótese x'_{jk} la cantidad $\dfrac{x_{j\cdot}x_{\cdot k}}{x_{\cdot\cdot}}$. Se dice que las variables $\mathbf{x}$ y $\mathbf{y}$ son *independientes* si $x_{jk} = x'_{jk}$, para todo j y todo k. Véase que la igualdad anterior es equivalente a:

$$\frac{x_{jk}}{x_{\cdot k}} = \frac{x_{j\cdot}}{x_{\cdot\cdot}} \tag{1.4}$$

para todo $j = 1, 2, \ldots, p$ y todo $k = 1, 2, \ldots, q$. Esto quiere decir que las variables son independientes si la proporción de individuos que poseen simultáneamente $\mathbf{x}^j$ y $\mathbf{y}^k$, entre los que poseen $\mathbf{y}^k$, es la misma proporción de aquéllos que poseen $\mathbf{x}^j$ en la población total.

Una manera de caracterizar las diferencias entre los x_{jk} y los x'_{jk} es mediante la cantidad χ^2 (léase **chi-cuadrado**):

$$\chi^2(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{(x_{jk} - x'_{jk})^2}{x'_{jk}}$$

es decir

$$\chi^2(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{1}{x_{..}} \frac{(x_{..}x_{jk} - x_{j.}x_{.k})^2}{x_{j.}x_{.k}}. \tag{1.5}$$

Esta cantidad representa la diferencia entre el producto de las frecuencias relativas teóricas $\frac{x'_{jk}}{x_{..}} = \frac{x_{j.}}{x_{..}} \times \frac{x_{.k}}{x_{..}}$ y las frecuencias relativas observadas $\frac{x_{jk}}{x_{..}}$. La asociación entre $\mathbf{x}$ y $\mathbf{y}$ será mayor conforme $\chi^2(\mathbf{x}, \mathbf{y})$ sea grande, y $\chi^2(\mathbf{x}, \mathbf{y})$ será cercano a cero en el caso de independencia entre $\mathbf{x}$ y $\mathbf{y}$. Se puede probar que (ver ejercicio 4 al final del capítulo):

$$\chi^2(\mathbf{x}, \mathbf{y}) = x_{..} \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{x_{jk}^2}{x_{j.}x_{.k}} - x_{..}$$

También es muy usado el coeficiente de contingencia de Pearson, denotado Φ^2:

$$\Phi^2(\mathbf{x}, \mathbf{y}) = \frac{\chi^2(\mathbf{x}, \mathbf{y})}{x_{..}}$$

y el T^2 de Chuprov:

$$T^2(\mathbf{x}, \mathbf{y}) = \frac{\Phi^2(\mathbf{x}, \mathbf{y})}{\sqrt{(p-1)(q-1)}} = \frac{\chi^2(\mathbf{x}, \mathbf{y})}{x_{..}\sqrt{(p-1)(q-1)}}.$$

El Φ^2 elimina el efecto del tamaño de una muestra y sirve para comparar dos tablas de contingencia de las mismas dimensiones, mientras que el T^2 elimina además el efecto del número de modalidades, por lo que puede servir para comparar cualquier par de tablas de contingencia. Además, el T^2 está comprendido entre 0 y 1 (ver ejercicio 5).

Ejemplo 18 *Considérense los datos de la Tabla de contingencia 1.3 de la página 11, que cruza el nivel de salario con el nivel de estudios en un conjunto de 1200 entrevistados. Para calcular el índice de chi-cuadrado, se calculan los términos* $\frac{x_{jk}^2}{x_{j\cdot}.x_{\cdot k}}$*, que son:*

Nivel de estudios	Nivel de salario			
	Bajo	Medio	Alto	Muy alto
Ninguno	0.27511073	0.00565022	0.00013904	0
Primario	0.26455088	0.02119309	0.00070988	0.00191113
Técnico	0.11010369	0.09292035	0.04838672	0.04377752
Secundario	0.04256933	0.12870535	0.02325581	0.00030193
Universitario	0.00031916	0.12214698	0.12116323	0.07045321

Al calcular la suma de todos esos términos, multiplicarla por $n = x_{\cdot\cdot} = 1200$ *y restarle* $n = x_{\cdot\cdot}$*, se obtiene que* $\chi^2 = 448.04$*. Además,* $\Phi^2 = 448.04/1200 = 0.3734$ *y* $T^2 = 0.3734/(4 \times 3) = 0.0311$.

Supóngase ahora que se tiene una variable cuantitativa $\mathbf{x}$ y una variable cualitativa $\mathbf{y}$ con modalidades $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q$. Si bien es cierto que se podría analizar la asociación entre $\mathbf{x}$ y $\mathbf{y}$ discretizando $\mathbf{x}$, por ejemplo construyendo un histograma de $\mathbf{x}$ y calculando luego el índice de χ^2, es preferible no perder la información de continuidad que posee la variable cuantitativa.

El **cociente de correlación** mide la intensidad de la asociación entre $\mathbf{x}$ y $\mathbf{y}$, calculando la dispersión que tiene $\mathbf{x}$ restringido en cada una de las modalidades de $\mathbf{y}$.

Ejemplo 19 *Supóngase que se tienen las siguientes variables, observadas sobre 20 individuos,* $\mathbf{x}$*: salario en dólares y* $\mathbf{y}$*: nivel de estudios, esta última con las modalidades primario, secundario, técnico y universitario. Agrupadas según las modalidades de* $\mathbf{y}$*, las observaciones son*

Nivel de estudios $\mathbf{y}$	Salario en dólares $\mathbf{x}$	Media de $\mathbf{x}$ $\bar{\mathbf{x}}_j$
Primario	\$267 \$503 \$208 \$198 \$250 \$263	\$281.50
Secundario	\$845 \$471 \$310 \$830	\$614.00
Técnico	\$759 \$1200 \$810 \$650	\$854.75
Universitario	\$1500 \$1113 \$2300 \$900 \$2100 \$1621	\$1589.00

Separando los datos según las modalidades de $\mathbf{y}$ *y calculando la media para cada uno de los grupos, se tiene que el salario medio para las personas con nivel educativo primario es $281.50, para los de nivel secundario es $614.00, para los de nivel técnico es $854.75 y para los de nivel universitario es $1589.00. La media total es $854.90. Se puede ver que hay diferencias grandes entre los salarios medios para cada modalidad de* $\mathbf{y}$ *y que los técnicos tienen un salario promedio muy parecido al de la media total, aún si ninguno de ellos tiene realmente un salario medio.*

El cálculo del cociente de correlación se basa en una comparación de las medias de $\mathbf{x}$ para cada una de las modalidades de $\mathbf{y}$. Sean $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_q$ las medias de $\mathbf{x}$ para cada una de las q modalidades de $\mathbf{y}$. Entonces el cociente de correlación entre $\mathbf{x}$ y $\mathbf{y}$ es:

$$\begin{aligned} \eta(\mathbf{x}, \mathbf{y}) &= \frac{\text{var}\left\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_q\right\}}{\text{var}\left(\mathbf{x}\right)} \\ &= \frac{\sum_{j=1}^{q} \mu_j \bar{\mathbf{x}}_j^2 - \bar{\mathbf{x}}^2}{\sum_{i=1}^{n} p_i x_i^2 - \bar{\mathbf{x}}^2} \end{aligned}$$

donde $\text{var}\left(\mathbf{x}\right)$ es la varianza de $\mathbf{x}$ y donde $\text{var}\left\{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_q\right\}$ es la varianza de las medias, la cual debe ser calculada de manera *ponderada* de acuerdo con el total de cada modalidad de $\mathbf{y}$; esto es, siendo μ_j la suma de los pesos de los individuos que tienen la modalidad j de $\mathbf{y}$:

$$\mu_j = \sum_{y_i = j} p_i.$$

Se probará como ejercicio que el cociente de correlación está entre 0 y 1. Cuando $\eta(\mathbf{x}, \mathbf{y})$ está cercano a 1 hay una fuerte asociación entre $\mathbf{x}$ y $\mathbf{y}$, y cuando está cercano a 0 hay muy poca asociación. Una desventaja que tiene $\eta(\mathbf{x}, \mathbf{y})$, en el caso en que $\mathbf{y}$ sea una variable cualitativa ordinal, es que no indica cuando una relación puede ser fuerte pero negativa entre $\mathbf{x}$ y $\mathbf{y}$.

Ejemplo 20 *Considerando los datos del ejemplo anterior (19), la varianza de las medias es 271,913.68, la cual se calcula así:*

$$\begin{aligned} 271,913.68 &= 0.3 \times (281.50 - 854.90)^2 + 0.2 \times (614.00 - 854.90)^2 + \\ &\quad +0.2 \times (854.75 - 854.90)^2 + 0.3 \times (1589.00 - 854.90)^2. \end{aligned}$$

La varianza total es 368,165.59 por lo que el cociente de correlación es

$$\eta(\mathbf{x}, \mathbf{y}) = \frac{271,913.68}{368,165.59} = 0.73856,$$

lo cual se puede interpretar como que la asociación entre el salario y el nivel educativo es del 73.85% para ese grupo de individuos.

Debe observarse que η no identifica el sentido de la relación entre las variables $\mathbf{x}$ y $\mathbf{y}$, como lo hace r para el caso cuantitativo.

Ejercicios

1. Sea $\mathbf{x}$ una variable cuantitativa y sea a una constante.
   (a) Se define la nueva variable $\mathbf{y} = \mathbf{x}+a$, entendiéndose que la constante a se suma a cada observación de la variable $\mathbf{x}$. Pruebe que $\bar{\mathbf{y}} = \bar{\mathbf{x}}+a$ y que $\operatorname{var}\mathbf{y} = \operatorname{var}\mathbf{x}$.
   (b) Se define la nueva variable $\mathbf{y} = a\mathbf{x}$, donde la constante a se multiplica a cada observación de la variable $\mathbf{x}$. ¿Cuánto es $\bar{\mathbf{y}}$ y $\operatorname{var}\mathbf{y}$?

2. Sea $\mathbf{x}$ una variable cuantitativa.
   (a) Considérese la operación de centraje, que se define como $\mathbf{y} = \mathbf{x} - \bar{\mathbf{x}}$. Pruebe que la media de $\mathbf{y}$ es cero.
   (b) Considérese la operación de estandarización, que se define como $\mathbf{y}' = \frac{\mathbf{x}}{\sigma_x}$. Pruebe que la desviación estándar y la varianza de $\mathbf{y}'$ son iguales a 1.
   (c) Concluya que $\mathbf{z} = \frac{\mathbf{x}-\bar{\mathbf{x}}}{\sigma_x}$ tiene media cero y varianza uno.

3. Sean $\mathbf{x}, \mathbf{y}$ dos variables cuantitativas y sean a, b dos constantes. ¿Qué relación hay entre $\operatorname{cov}(a\mathbf{x}, b\mathbf{y})$ y $\operatorname{cov}(\mathbf{x}, \mathbf{y})$? ¿Qué relación hay entre $r(a\mathbf{x}, b\mathbf{y})$ y $r(\mathbf{x}, \mathbf{y})$? Indique cuanto es $\operatorname{cov}(\mathbf{x}+a, \mathbf{y})$.

4. Demuestre que, dadas dos variables cualitativas $\mathbf{x}$ y $\mathbf{y}$ con p y q modalidades, respectivamente, el índice de chi–cuadrado tiene la siguiente propiedad:
$$\chi^2(\mathbf{x}, \mathbf{y}) = x_{..} \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{x_{jk}^2}{x_{j.} x_{.k}} - x_{..}$$

5. Sean $\mathbf{x}$ y $\mathbf{y}$ dos variables cualitativas. Pruebe que el T^2 de Chuprov cumple: $0 \leq T^2(\mathbf{x}, \mathbf{y}) \leq 1$.

6. Considere la siguiente tabla de datos con dos variables cualitativas:

sexo	*profesión*
masculino	oficinista
masculino	obrero
femenino	obrera
femenino	artista
femenino	artista
femenino	oficinista
masculino	artista
masculino	oficinista
femenino	artista

Haga una codificación disyuntiva completa de la tabla de datos.
Con las matrices asociadas a la codificación anterior, construya la tabla de contingencia que cruza a las dos variables cualitativas.

7. Sean $\mathbf{x}$ una variable cuantitativa y $\mathbf{y}$ una variable cualitativa con q modalidades. Se denota $\bar{\mathbf{x}}_j$ a la media de $\mathbf{x}$ calculada sobre los individuos que presentan la modalidad j de $\mathbf{y}$, para todo $j = 1, \ldots, q$. Es decir,

$$\bar{\mathbf{x}}_j = \frac{1}{\mu_j} \sum_{y_i = j} p_i \mathbf{x}_i.$$

   (a) Pruebe que la varianza de $\mathbf{x}$ se descompone como:

$$\text{var}\,(\mathbf{x}) = \sum_{j=1}^{q} \mu_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})^2 + \sum_{j=1}^{q} \sum_{\mathbf{y}_i = j} p_i (\mathbf{x}_i - \bar{\mathbf{x}}_j)^2$$

   donde $\mu_j = \sum_{\mathbf{y}_i = j} p_i$ es la suma de los pesos de los individuos que presentan la modalidad j de $\mathbf{y}$. El primer término de esta suma se llama la varianza inter–clases, y el segundo término se llama la varianza intra–clases.

   (b) Pruebe que $0 \leq \eta(\mathbf{x}, \mathbf{y}) \leq 1$.

Acerca de los Autores

Javier Trejos Zelaya estudió Matemática en la Universidad de Costa Rica, donde obtuvo la Licenciatura. Hizo sus estudios doctorales en la Universidad Paul Sabatier, Francia. Es investigador en la Universidad de Costa Rica desde 1993, habiendo publicado sobre la relación entre los métodos de análisis de datos y la optimización combinatoria, particularmente en clasificación automática, escalamiento multidimensional y regresión. Dirigió el Centro de Investigación en Matemática Pura y Aplicada (CIMPA), fue coordinador del Espacio Universitario de Estudios Avanzados y fue decano de la Facultad de Ciencias (2012-2020).

William Castillo Elizondo es Licenciado y Máster en Matemática por la Universidad de Costa Rica. Fue director de la Escuela de Matemática, cofundador y director del CIMPA. Participó activamente desde 1985 en proyectos de investigación teórica y aplicada sobre diversos métodos de análisis de datos multivariados, y contribuyó a la divulgación de dichos métodos organizando congresos e impartiendo cursos y conferencias en varias universidades. Además es coautor de numerosos artículos científicos publicados en revistas nacionales e internacionales.

Jorge González Varela estudió Profesorado en Matemáticas en la Universidad de Chile, posteriormente Licenciatura y Maestría en Matemática en la Universidad de Costa Rica. Fue Profesor en la Universidad de Santiago de Chile hasta 1973. Entre 1974 y 2006 trabajó en la Universidad de Costa Rica, en la cual participó en varios proyectos de investigación en el área de análisis de datos multivariados. Algunos de estos resultados se implementaron en paquetes computacionales que se utilizan por estudiantes y empresas. Es coautor de varias publicaciones en revistas nacionales e internacionales. Falleció en Chile en 2017.

El presente libro presenta los fundamentos matemáticos y las ideas intuitivas de los métodos más importantes de análisis multivariado de datos: el análisis de componentes principales, el análisis de correspondencias, la clasificación automática, el análisis discriminante y el análisis de tablas múltiples a través del método *statis*. Se desarrollan numerosos ejemplos, tanto con datos reales como con datos simulados, y al final de cada capítulo incluye ejercicios teóricos y prácticos. Se expone ampliamente la interpretación de los resultados para facilitar la práctica de los métodos expuestos.

Cada capítulo está organizado de manera que inicialmente se presenta el objetivo del método, enseguida se desarrolla este con sus propiedades de la mano de un ejemplo de ilustración, dejándose las demostraciones de algunas propiedades teóricas para el final del capítulo, cuando no son esenciales para la comprensión del método. Además, cada capítulo contiene varios ejemplos completos de aplicación junto con sus resultados e interpretaciones. Al final se presentan ejercicios tanto teóricos como prácticos para que el lector pueda ejercitarse en la comprensión de la teoría y la práctica.

Se espera que el presente libro sirva como material de referencia y estudio para investigadores y estudiantes que necesiten la herramienta del análisis multivariado.

EDITORIAL UCR