

Denis Baranger

CONSTRUCCIÓN Y ANÁLISIS DE DATOS

**INTRODUCCIÓN AL USO DE TÉCNICAS
CUANTITATIVAS EN LA INVESTIGACIÓN SOCIAL**

Posadas, 2009

Denis Baranger

Construcción y análisis de datos : introducción al uso de técnicas cuantitativas en la investigación social /
Denis Baranger ; con colaboración de Fernanda Niño. - 3a ed. - Misiones : el autor, 2009.
100 p. ; 21x15 cm.

ISBN 978-987-05-6307-5

1. Metodología de la Investigación. I. Niño, Fernanda, colab.
CDD 001.42

Fecha de catalogación: 20/04/2009

Prefacio a la segunda edición

Este libro se propone introducir al lector al uso de algunas técnicas cuantitativas en el campo de la investigación social. No es un libro de Estadística, aún cuando buena parte de su contenido trata acerca de técnicas estadísticas. Tampoco es un libro de Metodología, entendida ésta como una disciplina estrechamente vinculada a la filosofía de la ciencia que toma por objeto el Método Científico, en singular y con mayúsculas.¹ De insistir en adscribir el texto a alguna categoría, habrá que incluirlo como un miembro más de la nutrida especie de los manuales de metodología y técnicas de la investigación social, textos que se utilizan corrientemente en las carreras de sociología, ciencia política, comunicación, trabajo social, administración, y tantas otras.

Normalmente un manual es una colección de recetas acerca de “cómo hacerlo”, por lo que se supone que no debería requerir de mayores aclaraciones. Empero en este caso no está de más formular algunas reflexiones acerca de las condiciones a las que es recomendable atenerse para su mejor uso. En este sentido, es fundamental dejar en claro que, dentro de un proceso en el que se pretenda enseñar a hacer investigación, este manual –o cualquier otro– debería ser apenas un elemento más, y sin duda no el más importante.

En efecto, el enfoque del texto es sesgado: no abarca todos los temas posibles, sino que aborda solamente lo que hace a las condiciones para hacer uso de técnicas cuantitativas. En particular, no se dice nada acerca de la denominada investigación cualitativa.² Como observa Silverman (1995), es frecuente adscribir el uso de técnicas cualitativas de investigación a un paradigma *interpretativista* en tanto que las técnicas cuantitativas, por su parte remitirían a otro paradigma, *positivista*.³ Empero, tratándose específicamente de metodología, tal vez no sea del todo conveniente concebir la cuestión en términos de paradigmas contrapuestos. En este terreno, sólo cabe hablar de la posibilidad de recurrir a distintas estrategias de investigación, cuya diferencia debe entenderse más como una cuestión de hecho que de derecho.

No desconozco que existen tradiciones disciplinares enteras, como la antropología social y cultural, que han descansado durante largo tiempo en el uso casi exclusivo de técnicas etnográficas, al punto que incluso en la actualidad tal elección continúa siendo la principal marca de su identidad. Pero lo cierto es que en la práctica de la investigación social sucede con la mayor frecuencia que se combinen técnicas cuantitativas y cualitativas. El mismísimo Lazarsfeld, a quien nadie osaría disputar el título de padre-fundador de la metodología sociológica cuantitativa, reconocía este hecho al punto de dedicarle algunos de sus textos más interesantes (cf. 1970 y 1990).

Hacia 1950 se inauguraba una clasificación de los diseños de investigación que estaba destinada a perdurar. En su manual, Selltíz *et al* (1968) distinguían diversos tipos de investigaciones, según cual fuera su objetivo principal. Así, en los estudios exploratorios, que persiguen la producción de hipótesis, prima el uso de técnicas cualitativas, como la inmersión en el campo, la observación participante, la entrevista no estructurada, etc. Por lo contrario, las técnicas cuantitativas son propias de las investigaciones descriptivas, de los estudios del tipo *survey*, que se basan en muestras con un número de casos lo suficientemente grande como para realizar inferencias estadísticas.⁴ Ahora bien, con no poca frecuencia en los estudios descriptivos, y muy especialmente en aquellos del tipo aplicado, ocurre que no se parte de un cuerpo de hipótesis previamente sistematizado, por lo que resulta casi inevitable en una primera fase *-exploratoria-*

¹ Así, para Gutiérrez y Brenes la teoría del método se refiere «a la revisión sistemática de los fundamentos lógicos y epistemológicos del quehacer científico, en cualquiera de sus ramas» (1971: 11).

² Según Denzin y Lincoln, la investigación cualitativa se caracteriza por «un énfasis en procesos o significados que no son rigurosamente examinados, o mensurados (si es que se los mide de alguna manera), en términos de cantidad, monto, intensidad o frecuencia (...) Por lo contrario, los estudios cuantitativos enfatizan la medición y el análisis de relaciones causales entre variables, y no de procesos» (1994: 4).

³ Una posición semejante es expresada por Irene Vasilachis, con la salvedad de que esta autora hace referencia a la existencia de un tercer paradigma: el *materialista histórico* (1992: 23).

⁴ El tercer tipo, irrelevante en el contexto de esta argumentación, está constituido por los estudios de comprobación de hipótesis causales –Padua (1979) hablará de estudios *explicativos*– cuya importancia puede ser grande en psicología o en psicología social, pero que es marginal para la investigación propiamente sociológica.

recurrir a técnicas cualitativas.⁵ Por otra parte, también sucede que los hallazgos producidos mediante un estudio descriptivo a partir de una muestra representativa requieren ser esclarecidos, con lo que nuevamente se hará necesario algún trabajo cualitativo, para el caso *a posteriori*, que permita comprender su significado.⁶ En suma, lo cierto es que en gran parte de la práctica de la investigación social lo cualitativo y lo cuantitativo no son términos excluyentes sino más bien complementarios y, por lo general, ambos necesarios.

Sin embargo, en este texto nos ceñimos exclusivamente al aspecto cuantitativo, no por considerar que sea la única estrategia de investigación posible,⁷ sí porque pensamos que se trata de una orientación que no se puede desconocer, y cuya especificidad demanda esclarecer una cierta lógica básica subyacente que es susceptible de ser codificada. *Nuestro objetivo es hacer más accesible un modo de pensar los datos o, mejor aún, de “pensar en datos”*, escribíamos en el prefacio a la anterior edición de este libro. Para disipar malentendidos, prefiero ahora expresar: aprender a pensar en *datos completamente estructurados* como tales. Es decir, el propósito continúa siendo simplemente enseñar a trabajar con aquel tipo de datos que se puede ordenar bajo la figura tripartita de la matriz descrita por Galtung (1966). Se trata en definitiva de desentrañar los fundamentos de lo que Blumer (1982) denominaba *análisis de variables*.⁸

La segunda razón por la cual no basta con leer este libro para aprender a investigar, tiene ver con la naturaleza misma de la materia sobre la que versa. Como cualquier oficio, el de investigador se adquiere fundamentalmente a través de la práctica realizada de la mano de una persona más experimentada, como nos lo recuerda Catalina Wainerman.⁹

Otro autor que insiste en este aspecto de la transmisión personal del *habitus* del investigador es Bourdieu (1992: 189-195), cuyo menosprecio por los manuales de metodología es conocido.¹⁰ Para él, tanto la teoría sociológica, como la epistemología y la metodología son materias que no tiene sentido enseñar separadamente, y cuya razón de ser reside, en última instancia, en su puesta en práctica en la investigación. Es evidente que, en este caso, me sería imposible seguir a Bourdieu sin caer en contradicción. Por mi parte, prefiero considerar que una cosa es afirmar la necesidad de combinar todos estos elementos en la práctica de la investigación, y otra muy distinta sostener que no son susceptibles de ser enseñados y aprendidos por separado, en un determinado nivel y con todos los *caveat* que permitan precaverse de cualquier deformación metodologista. Teoría y metodología son elementos que habrá que combinar en una síntesis que inevitablemente cada estudiante deberá realizar de un modo personal. En lo que hace a la metodología, una formación completa deberá incluir necesariamente el aprendizaje de técnicas cuantitativas tanto como cualitativas, y un fuerte componente de práctica de investigación guiada. Ahora bien, el modo y el orden en que se produzca la incorporación de todos los elementos requeridos para investigar, me parece que es materia opinable. Pero no veo por qué no sería posible en este caso, como en cualquier otro, separar analíticamente la enseñanza de los elementos que en alguna instancia -por ejemplo en una tesina- terminarán integrándose prácticamente.

En la enseñanza de esta materia existe una saludable diversidad en las secuencias curriculares establecidas en los planes de estudio de diferentes facultades. Aún cuando este texto está pensado

⁵ «Los datos cualitativos, por su riqueza descriptiva, juegan un papel fundamental en las fases exploratorias de la investigación» (Lazarsfeld, 1970: 355).

⁶ Según observa María Antonia Gallart, «La interpretación de cualquier regularidad cuantitativa, inicialmente expresada en términos estadísticos, implica elementos cualitativos no explicitados en la operacionalización de las variables» (cf. Forni *et al.*, 1992: 110).

⁷ Cf. la presentación comparada de estrategias cualitativas y cuantitativas realizada por Forni (1992).

⁸ Considero a la matriz de datos solamente como un instrumento metodológico, como un simple recurso técnico, contrariamente al punto de vista de Juan Samaja, que eleva la matriz al nivel de una categoría epistemológica. Cf. mi crítica (1994) a las ideas de Samaja y su respuesta (1994).

⁹ «No se aprende a hacer investigación en los cursos especializados de metodología y técnicas si no se hace investigación junto a un “maestro/a”, como en los gremios medievales, dentro de un proyecto de investigación dirigido por el “maestro/a”. Esto es así porque hay “algo” no codificable, difícil de transmitir del oficio de investigador» (cf. Wainerman y Sautu, 1997: 19).

¹⁰ Y que lo llevara a desistir de escribir los volúmenes que estaban previstos como continuación de *El oficio del sociólogo* (1968): «hubiera sido reconocer la división teoría/empiría que es el equivalente de la oposición, profundamente funesta, de la tradición anglo-sajona, entre *teoría* y *metodología*» (Bourdieu, 1997:55); cf. también, *Réponses*, dónde menciona «ese absurdo científico denominado “metodología”» (1992: 137).

como una introducción, para su mejor aprovechamiento es conveniente que el estudiante haya adquirido de manera previa -o en su defecto adquiriera en forma simultánea- nociones elementales de estadística, y algunos conceptos básicos de epistemología como los que se desarrollan habitualmente bajo el título de *Introducción al conocimiento científico* o en asignaturas equivalentes.

Mi propia experiencia de docencia en la Universidad Nacional de Misiones se ha basado siempre en el principio de hacer que los estudiantes lleven adelante investigaciones reales, como el mejor medio para que la enseñanza de esta materia no quede reducida a una colección de recetas abstractas y resulte eficaz. Esta modalidad, empero, no deja de plantear algunos problemas. En primer lugar, necesita ser llevada adelante bajo la forma de una asignatura de duración anual: en los límites de un cuatrimestre es materialmente imposible desarrollar todas las etapas que implica la realización de una investigación. En segundo lugar, es una alternativa muy consumidora de recursos humanos, ya que el grupo de estudiantes que trabaje conjuntamente en una investigación no puede ser demasiado grande.¹¹

En esta segunda edición, el texto ha mantenido esencialmente su estructura anterior. En lo que hace a los capítulos que ya estaban escritos, apenas me he limitado a subsanar una serie de errores, algunos tipográficos y otros conceptuales, además de sustituir ciertos ejemplos.

Hay, sin embargo, una innovación fundamental: la incorporación de un capítulo completamente nuevo dedicado al análisis de correspondencias. Ha sido todo un desafío el intento de realizar una presentación accesible de un tema que viene a cumplir la función de producir un necesario *aggiornamento*, pero cuya complejidad es notablemente mayor que el resto de los contenidos. Hemos acometido esta tarea con la colaboración invalorable de Fernanda Niño, en una experiencia de trabajo conjunto que personalmente me resultó sumamente enriquecedora.

Es posible que con el añadido de este capítulo se haga más visible lo que puede aparecer como una cierta falta de unidad, que deviene de nutrirse de vertientes aparentemente contrapuestas, como el *empirismo abstracto* lazarsfeldiano¹² y la tradición francesa de la filosofía de las ciencias. Pero esta tensión que ya estaba presente en la edición anterior, como habrá podido advertir cualquier lector atento, no es malsana. La idea es precisamente no quedar presos de las tradicionales parejas epistemológicas (Bachelard, 1986) sino contribuir a superarlas.

Varios colegas han tenido la gentileza de leer el manuscrito del nuevo capítulo 6 sobre análisis de correspondencias, y merecen nuestro reconocimiento. Las observaciones de Roberto Benencia, Mario Boleda, Juan Carlos Cernuda, Jorge Rozé, Gabriela Schiavoni y Eduardo Simonetti han sido una fuente de valiosas sugerencias para mejorarlo, aunque no siempre nos haya sido posible atenernos a ellas. Como es obvio, los errores que subsisten son de nuestra exclusiva responsabilidad.

Por último, *but not least*, deseo expresar mi agradecimiento a mis colaboradores docentes en la cátedra de “Metodología y Técnicas de la Investigación Social” de la Facultad de Humanidades y Ciencias Sociales, así como a sus estudiantes, que han resultado ser coautores de la mayoría de los ejemplos aquí presentados.

Posadas, noviembre de 1998

¹¹ Idealmente un grupo de estudiantes que trabajara en una investigación bajo la dirección de un docente no debería sobrepasar el número de doce integrantes (aunque en la práctica este límite ha llegado a duplicarse e incluso a triplicarse en ocasiones).

¹² La expresión pertenece a C. Wright Mills (1964). En efecto, el *empirismo* o *positivismo lógico* del Círculo de Viena (cf. Ayer, 1965) fue el principal referente filosófico para el pensamiento de Lazarsfeld, como señalaba acertadamente Pollak (1993: 324).

*Prefacio a la tercera edición
(versión electrónica)*

Transcurrida una década desde la preparación de la segunda edición en papel para la Editorial Universitaria de Misiones, es obvio que el texto aquí presentado en versión electrónica demandaría ser modificado en muchos de sus contenidos, para no hablar de una actualización de las referencias bibliográficas.

Empero, no es éste el partido que he adoptado para la presente versión en la que me he limitado a introducir muy ligeras correcciones que tienen que ver fundamentalmente con erratas de la anterior versión en papel. Consideraciones de tiempo tornan inviable cualquier otra alternativa, en tanto que en su estado actual todavía puede el texto continuar funcionando como un instrumento útil para la enseñanza.

Si el proceso de investigación involucra una lucha constante en contra del error, ello no es menos cierto en lo que concierne a la edición de cualquier texto. En la medida de nuestras posibilidades, iremos produciendo sucesivas versiones mejoradas. En esta tarea agradecemos desde ya los aportes que puedan realizar los lectores, solicitando nos los comuniquen por correo electrónico a la dirección <baranger@arnet.com.ar>.

Posadas, abril de 2009

Índice de contenidos

Prefacio a la segunda edición

Prefacio a la tercera edición

Índice de contenidos

Índice de figuras

Índice de tablas

CAPITULO 1: LA ESTRUCTURA DE LOS DATOS

1. Unidades de análisis, variables, valores

2. Niveles de medición

CAPITULO 2: LA MATRIZ DE DATOS

1. La forma de la matriz de datos

2. Niveles de inclusión

2.1. Objeto de investigación y unidades de análisis

2.2. La tipología de Lazarsfeld y Menzel

CAPITULO 3: NOCIONES DE MUESTREO

1. Elementos básicos de muestreo

2. El tamaño de la muestra

2.1. Muestras para verificar hipótesis de generalización

2.2. Muestras para someter a prueba hipótesis sustantivas

3. Principales tipos de muestras

3.1. Muestras no—probabilísticas

3.2. Muestras probabilísticas

CAPITULO 4: TECNICAS ELEMENTALES DE ANALISIS

1. la tabla de contingencia y el uso de los porcentajes

2. El análisis de la relación entre variables

2.1. La diferencia porcentual: una medida de la asociación

2.2. El test de χ^2 : una medida de la significación estadística

2.3. Algunos coeficientes de asociación

3. Distribuciones multivariantes conjuntas

CAPITULO 5 : INDICES Y ESCALAS

1. Espacios de propiedades e índices

Reducciones de espacios de propiedades cualitativas

El proceso de operacionalización

2. Escalas de actitudes

La escala Lickert

3. Confiabilidad y validez

Validez

Confiabilidad

Confiabilidad y validez

Referencias bibliográficas

CAPITULO 6: ANALISIS DE CORRESPONDENCIAS

1.El análisis factorial de correspondencias

2. Algunas aplicaciones del ACM

2.1 Construcción de un índice por medio del ACM

2.2 Construcción de una tipología

Índice de figuras

- Figura 1.1 Tipos de sistemas
- Figura 1.2 Ubicación de “rojo” dentro de una estructura de oposiciones
- Figura 1.3 Resumen de las características de los diferentes niveles de medición
- Figura 2.1 Tipos de investigación, según el número de variables y de unidades de análisis involucradas
- Figura 2.2 Evolución de los nuevos inscriptos y desertores en la UNaM, 1982—86.
- Figura 2.3 Representación de unidades de análisis en tres niveles
- Figura 2.4 Una tipología de colectividades
- Figura 2.5 Tipología de Lazarsfeld y Menzel
- Figura 2.6 Ejemplo de generación de una propiedad contextual -analítica
- Figura 4.1: Misiones, 1980 - Pertenencia de la población a hogares con Necesidades Básicas Insatisfechas según tipo de asentamiento
- Figura 4.2 Diagrama de dispersión de los departamentos y recta de regresión para las variables N° de núcleos del MAM y % de explotaciones de 5-50 has.
- Figura 4.3 Ejemplo de una relación no lineal entre dos variables
- Figura 5.1 Posición de las provincias argentinas en un espacio de propiedades continuo
- Figura 5.2 Ejemplo de espacio de propiedades cualitativas
- Figura 5.2.a Reducción por simplificación de dimensiones
- Figura 5.2.b Reducción numérica
- Figura 5.2.c Reducción pragmática
- Figura 5.3 Reducción de un espacio de propiedades de posición política en dos puntos del tiempo
- Figura 5.4 Etapas del proceso de operacionalización
- Figura 5.5 Ejemplo de validación constructiva
- Figura 5.6 Tipos de validez
- Figura 6.1 Histograma de los valores propios
- Figura 6.2 Posición de las ocupaciones de padres e hijos en el plano (1,2)
- Figura 6.3 Argentina, 1980-Provincias y posiciones ocupacionales
- Figura 6.4 Histograma de los valores propios
- Figura 6.5 Ubicación de los individuos y de las modalidades activas y suplementarias en el plano factorial (1,2)
- Figura 6.6 Origen social de los estudiantes
- Figura 6.7 Localización en el plano de los individuos pertenecientes a cada tipo

Índice de tablas

Tabla 1.1	Ejemplo de definición de universos en base a propiedades constantes y variables
Tabla 2.1	Ejemplo de matriz de 10 x 2
Tabla 2.2	Ejemplo de matriz de 500 x 3
Tabla 2.3	Forma general de la matriz de datos
Tabla 3.1	Tamaños de muestras al azar simple requeridos para niveles de confianza del 99% y del 95%, según valores presumibles de p y q y límites de error
Tabla 3.2	Número mínimo de unidades de análisis para un promedio de 10 casos por celda (20 casos entre paréntesis)
Tabla 3.3	Comparación de la distribución de la variable “nivel educativo” en la muestra y en la población
Tabla 3.4	Distribución en la muestra de la intención de voto según el nivel educativo
Tabla 3.5	Distribución corregida de la intención de voto según nivel educativo
Tabla 4.1	Forma lógica de la tabla de 2 x 2
Tabla 4.2	Misiones, 1980 - Población según tipo de asentamiento y pertenencia a hogares con Necesidades Básicas Insatisfechas
Tabla 4.2.a	Misiones, 1980 - Población según tipo de asentamiento
Tabla 4.2.a.1	Misiones, 1980 - Población según tipo de asentamiento (%)
Tabla 4.3.1	Misiones, 1980 - Pertenencia de la población a hogares con NBI, según tipo de asentamiento (%)
Tabla 4.1.2	Misiones, 1980 - Tipo de asentamiento de la población según pertenencia a hogares con NBI(%)
Tabla 4.3	Conformidad con el sistema de examen final según carrera
Tabla 4.3.1	Carrera según conformidad con el sistema de examen final (%)
Tabla 4.3.2	Carrera según conformidad con el examen final
Tabla 4.2.3	Misiones, 1980 - Distribución de la población por tipo de asentamiento y pertenencia a hogares con NBI
Tabla 4.4	Grado de participación política y grado de conocimiento político
Tabla 4.4.1	Participación política según conocimiento político (%)
Tabla 4.4.2	Conocimiento político según participación política (%)
Tabla 4.5	Evaluación de la situación social según NES (%)
Tabla 4.6	Valores resumida de valores críticos para χ^2
Tabla 4.7.a	Interés de estudiantes en la política según etapa de la carrera en que se encuentran
Tabla 4.7.b	Interés de estudiantes en la política según etapa de la carrera en que se encuentran
Tabla 4.8	Grado de conocimiento político según nivel económico-social
Tabla 4.9	Número de núcleos de base del MAM por departamento e indicadores agrarios seleccionados
Tabla 4.9.1	Número de núcleos del MAM según porcentaje de explotaciones medianas
Tabla 4.9.2	Ejemplo de distribución bi-variante conjunta y construcción de un coeficiente de correlación por rangos
Tabla 4.9.3	Matriz de correlaciones entre cuatro indicadores
Tabla 5.1	Ejemplo de matriz para el análisis de ítems
Tabla 6.1	Buenos Aires, 1982- Ocupación del hijo y ocupación del padre
Tabla 6.1.1	Tabla de perfiles-fila
Tabla 6.1.2	Tabla de perfiles-columna
Tabla 6.2	Coordenadas, contribuciones y cosenos cuadrados de los puntos-perfiles o modalidades
Tabla 6.3	Valores, puntajes y etiquetas de las variables seleccionadas para el IES
Tabla 6.4	Tabla de Burt para los ocho indicadores del IES
Tabla 6.5	Presentación abreviada de una tabla disyuntiva completa para 240 estudiantes y 16 modalidades del IES
Tabla 6.6	Relación entre el IESacm y el IES sumatorio
Tabla 6.7	Variables y modalidades para generar tipos sociales
Tabla 6.8	Extracción social de los estudiantes en dos establecimientos de enseñanza agropecuaria

CAPÍTULO 1

LA ESTRUCTURA DE LOS DATOS

«Otra peculiaridad de los datos científicos es que, pese a la etimología de la palabra 'dato', no son nada dado, sino que hay que producirlos, y muchas veces laboriosamente».
(Bunge; 1979)

«¿ Qué es un dato?», es la pregunta a la que quisiéramos responder en este primer capítulo. Intentaremos mostrar así como la estructura de cualquier dato es esencialmente el producto de un acto clasificatorio, acto cuya simplicidad es sólo aparente. Lejos de sernos inmediatamente “dado” -como lo sugiere engañosamente la etimología-, el dato presupone toda una serie de operaciones que se llevan a cabo en forma simultánea dentro de un sistema conformado por múltiples dimensiones. De este modo, el dato más elemental ya conlleva toda una serie de elecciones teórico-metodológicas. Y es que desde el punto de vista que se sostiene aquí, el dato es eminentemente racional: no se “ve” más que lo que ya se está dispuesto a ver.¹

Comenzaremos procediendo a un ejercicio de disección de la naturaleza del dato, mediante el cual estaremos en condiciones de aislar sus elementos constitutivos. La exposición algo dogmática de una serie de definiciones nos permitirá establecer las relaciones sistemáticas por las que se vinculan unos a otros los conceptos de ‘sistema’ y ‘unidad de análisis’, ‘propiedades’ y ‘medición’, ‘variables’ y ‘valores’. Para ello, nos valdremos de ejemplos muy sencillos y deliberadamente no sociológicos. Al término de este trabajo analítico, el dato deberá aparecer como el producto de la imbricación de dos estructuras: la que permite atribuir una propiedad a un objeto, y la que posibilita estructurar sistemáticamente conjuntos de propiedades.

1. UNIDADES DE ANÁLISIS, VARIABLES, VALORES

Según nos dice Galtung (1966) todo dato hace referencia a una estructura constituida por tres elementos: unidad de análisis, variable y valor. Cualquier dato aislado consistirá pues en...

- a) una **unidad** de análisis que
- b) en una **variable** específica presentará
- c) un determinado **valor**.

Lo que constituye el dato son estos tres elementos considerados conjuntamente con las relaciones que mantienen entre sí. Sea cual sea nuestro objeto de estudio, si queremos sostener cualquier proposición empírica acerca de él, se lo deberá concebir en términos de esa estructura tripartita. ‘Unidad’, ‘variable’ y ‘valor’ son todos términos que denotan determinados tipos de conceptos.

Una sencilla aproximación desde la lógica puede permitirnos echar algo más de luz sobre su significado, partiendo de las categorías clásicas de ‘sujeto’ y ‘predicado’. La idea más elemental de lo que es un dato puede ejemplificarse a partir de un enunciado muy simple: «ese sillón es

¹ Desde la tradición epistemológica francesa (Bachelard, 1972 y 1986) esto es una obviedad. En la vereda opuesta, el positivismo lógico (Ayer, 1965) sostenía la posibilidad de un lenguaje observacional puro dentro del cual era factible formular enunciados que permitieran dirimir acerca de la verdad-falsedad de una hipótesis. En el mundo anglosajón, hubo que esperar hasta Kuhn (1969) para que se pusiera radicalmente en cuestión la reconstrucción positivista.

Es desde este punto de vista que se puede afirmar que el conocimiento científico consiste en la identificación de sistemas de determinadas clases, en la medición de sus propiedades, y en el establecimiento de relaciones entre dichas propiedades. Por lo demás, es obvio que en ciencias sociales no nos interesan todos los sistemas, ni tampoco todas sus propiedades. De hecho, sucede que precisamente puede entenderse que lo que diferencia a las disciplinas científicas unas de otras son en gran medida los tipos de sistemas de los que se ocupan y las propiedades de dichos sistemas que toman en cuenta.⁶

Por otra parte, basándonos en determinadas propiedades podemos referirnos a **clases** de sistemas:

los **autos 0 kilómetro**,
 las **galaxias lejanas**,
 los **libros de Sociología**,
 los **alumnos de Metodología**,
 etc.

También de estas clases de sistemas podemos predicar propiedades:

«Los **autos 0 km.** son *caros* ».
 «Las **galaxias lejanas** son *difíciles de observar*»,
 «Los **libros de Sociología** son *aburridos* »
 «Los **alumnos de Metodología** son *sabios* »,
 etc.

Así como existen autos de diversos colores, también existen lápices, flores, libros y cielos de distintos colores. Todos concordaremos sin embargo en que un auto no es ni un lápiz ni una flor, ni un libro ni un cielo. Si queremos saber qué es un auto podemos recurrir al Larousse usual: « Vehículo que camina movido por un motor de explosión ». Todos coincidiremos en que un lápiz no “camina”, en que una flor no es un “vehículo”, y en que los libros y cielos no son “movidos por un motor de explosión”. Es imposible que confundamos un auto con un lápiz, porque de algún modo poseemos **definiciones** de estos objetos; sabemos que un auto se define por determinadas características: tener cuatro ruedas, un volante, asientos y una carrocería, y ser capaz de llevarnos con relativa -y cada vez menor- seguridad de un punto a otro de la superficie terrestre, rodando por los caminos. En otras palabras, el sistema ‘auto’ se caracteriza por presentar determinadas propiedades, a saber:

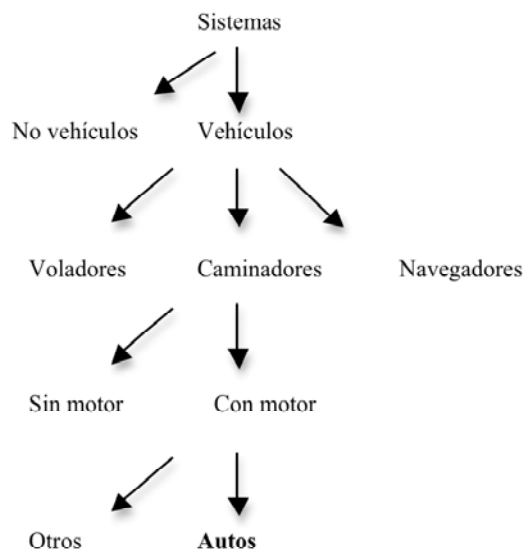
“ser *vehículo*”,
 “*caminar*”,
 “ser movido por un *motor a explosión*”,
 “*tener cuatro ruedas*”,
 etc.

Se pueden imaginar fácilmente sistemas que no satisfagan algunas de estas propiedades: una avioneta Piper o una lancha no “caminan”, una bicicleta no es “movida por un motor de explosión”, etc...

Lo que queremos remarcar, entonces, es que no solamente existen sistemas de los cuales es posible predicar propiedades, sino que los sistemas mismos son susceptibles de ser definidos en última instancia a partir de una determinada combinación de propiedades. Mucho “antes” de ser *rojo*, un ‘auto’ es **auto** en tanto presenta determinadas propiedades. Podríamos pensar que todos los sistemas se dividen en dos **clases**: “Vehículos” y “No-vehículos”; que algunos de los vehículos “caminan” mientras otros no lo hacen, y que sólo algunos de estos últimos son “movidos por un motor de explosión”:

⁶ O, para decirlo de otra manera, las ciencias se diferencian por sus “**objetos**”.

Figura 1.1: Tipos de sistemas



A partir de las nociones de “sistema” y de “propiedad”, podemos establecer los conceptos de “universo” (y, correlativamente, el de “muestra”),⁷ “unidad de análisis”, “variable”, y “constante”. Estos términos cobrarán un sentido específico cuando se los considere dentro de una estructura definida en un determinado nivel de inclusión. Así, podemos reconocer en el “**Universo**” el nivel de mayor inclusión posible, que se compone de todos los sistemas existentes, de todos los que han existido y existirán; en este nivel se incluyen también -al menos potencialmente- todas las propiedades posibles.

Ahora bien, apelando a determinadas propiedades podemos definir **sub**-universos -o, simplemente, “**universos**”- dentro de este “Universo” mayor en un número ilimitado. Por ejemplo, es posible definir el universo de los “vehículos”. Los vehículos son sistemas que permiten desplazar a otros sistemas (y también a sí mismos) de un punto a otro del espacio: se los define por poseer la propiedad de ser “medios de locomoción”.⁸ Se los puede distinguir así de todos los otros sistemas en los que no reconocemos la propiedad de funcionar como medios de locomoción. Todos los vehículos comparten esta propiedad de ser “medios de locomoción”: para esta clase de sistemas, el ser “medio de locomoción” es una **constante**. Notemos que en el nivel de inclusión inmediatamente superior -el Universo- no todos los sistemas son “medios de locomoción”: en ese nivel superior dicha propiedad funciona como una **variable**.

Podemos continuar definiendo universos de menor nivel de inclusión. Así tendremos el universo de los vehículos “caminadores” (que se diferencia de los universos de los vehículos “voladores” y “navegadores”); a su vez, dentro del universo de los vehículos caminadores, definimos el de los “movidos a motor”, y dentro de éste, el universo de los autos. La Tabla 1.1 permite apreciar cómo cada universo ⁹ está definido por una serie de propiedades constantes y se corresponde con un determinado nivel de inclusión. Todo nivel de inclusión inferior requiere para su definición a partir del nivel inmediatamente superior tomar como constante al menos una propiedad; y viceversa: al pasar de un nivel de inclusión menor a otro superior aumentará el número de propiedades variables.

⁷ Cf. *Infra*, capítulo 3.

⁸ Larousse *dixit*.

⁹ Excepto, claro está, el primero de todos: el “Universo de todos los sistemas”.

Tabla 1.1 Definición de universos en base a propiedades constantes y variables

UNIVERSOS	PROPIEDADES			
	Medio de locomoción	Medio en que se desplaza	Medio de impulsión	Número de ruedas
Sistemas	Variable	Variable	Variable	Variable
Vehículos	Constante	Variable	Variable	Variable
Vehículos caminadores	Constante	Constante	Variable	Variable
Vehículos caminadores a motor	Constante	Constante	Constante	Variable
Autos	Constante	Constante	Constante	Constante

En cualquier nivel de inclusión que se tome, las propiedades aparecerán jugando entonces un doble papel: algunas de ellas, tomadas como constantes, estarán definiendo un universo, mientras que otras funcionarán como variables y podrán ser objeto de investigación. Podemos ahora introducir la idea de que todos los sistemas que corresponden a un nivel de inclusión dado conforman un universo, y que cada uno de estos sistemas que componen dicho universo es susceptible de ser considerado como una unidad de análisis. Ello nos permite definir:

Una **unidad de análisis** es un sistema definido por presentar determinadas **propiedades**, algunas de ellas **constantes** (las que definen su pertenencia a un universo compuesto por todos los sistemas que presentan esas mismas propiedades) y otras **variables** (las que podrán ser materia de investigación dentro de ese universo).

Disponemos ahora de una primera definición de lo que es una ‘variable’, entendida como un determinado tipo de propiedad, a partir de los conceptos de ‘unidad de análisis’ y de ‘universo’. En lo que sigue, nos proponemos continuar elucidando este concepto de ‘variable’, para lo cual deberemos introducir una nueva idea: la de ‘valor’.

La siguiente definición de Galtung es un buen punto de partida para esclarecer las relaciones que median entre los tres elementos constitutivos del dato:

«Dado un conjunto de unidades, un **valor** es algo que puede predicarse de una unidad, y una **variable** es un conjunto de valores que forma una clasificación» (Galtung; 1966: I,78).

De acuerdo a lo que se ha venido exponiendo, no debe haber ninguna dificultad en reconocer de inmediato que un ‘valor’ es simplemente una propiedad. Ahora bien, hemos dicho que una variable también es una propiedad; ¿Significa esto que ‘valor’ y ‘variable’ son dos términos que denotan el mismo concepto, el de ‘propiedad’? En absoluto.

Cuando enunciamos «Este **auto** es *rojo* », ¿ En qué sentido es posible hablar de “medición”? Solamente en tanto esta propiedad ‘*rojo*’ se encuentre definida por su lugar dentro de una estructura que le confiera un sentido. Así la estructura más simple en la que podemos pensar se

basa en una relación simple de oposición:

rojo versus *no-rojo*

De este modo afirmar que el auto es rojo equivale a afirmar que dicho auto **no** es no-rojo: la afirmación de la propiedad es al mismo tiempo la negación de la no-propiedad. Podríamos introducir un término para denotar esta propiedad -el de '*rojidad*'- que denotaría una variable, entendida ahora como una característica susceptible de adoptar diferentes valores ("sí" y "no"). Cualquier unidad de análisis de la cual se pudiera predicar con sentido la rojidad presentaría en esta variable el valor "sí" o el valor "no". Así como podríamos medir la longitud de un auto («este auto mide 427 cm.»), así también mediríamos su 'rojidad':

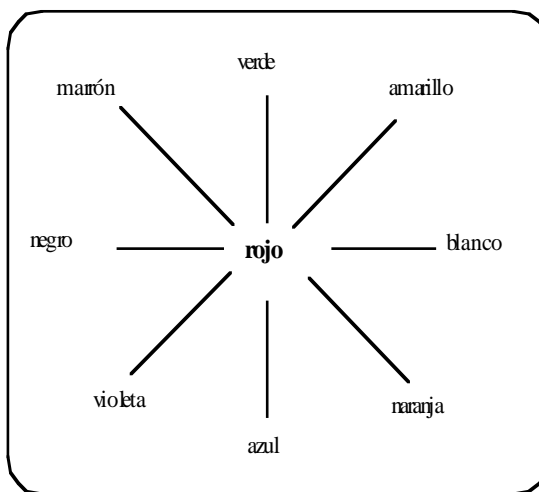
«Este **auto** presenta el valor "sí" en la variable *rojidad* »
(o, para decirlo más simplemente: «Este auto es rojo»).

De este modo, podemos establecer que:

una **variable** es un conjunto estructurado de valores¹⁰

Ahora bien, el término 'rojo' también puede definirse por su pertenencia a una estructura algo más compleja; dentro de esta estructura, el 'rojo' se definirá como tal por **no ser** ni marrón, ni verde, ni amarillo, ni blanco, ni naranja, ni azul, ni violeta, ni negro. 'Rojo' puede aparecer ahora como un valor de una variable a la que podríamos denominar '*color*', y en la que se opone a toda una serie de otras propiedades que constituyen otros tantos valores posibles de esa variable.

Figura 1.2: Ubicación de "rojo" dentro de una estructura de oposiciones



Los nombres de colores incluidos en la Figura 1.2 constituyen "un conjunto de valores que forma una **clasificación**", vale decir, lo que Galtung define como '**variable**'. 'Variable' y 'valor' aparecen pues como conceptos bien distintos, aún cuando se encuentren estrechamente relacionados. Un valor es parte de una variable, no existe sino por las relaciones que guarda con los otros valores que componen esta variable. A su vez, una variable no es otra cosa que el conjunto de los valores que la conforman y de las relaciones que éstos mantienen entre sí.

Las variables correctamente construidas deben necesariamente satisfacer los requisitos demandados a cualquier clasificación: los valores que las compongan deberán en cada caso ser *exhaustivos* -en su conjunto-, y *mutuamente excluyentes*. Así como la hemos construido, nuestra variable '*color*' es imperfecta: un auto rosado no encontraría ubicación en ningún valor; nuestra

¹⁰ Para Frege, uno de los padres de la moderna Lógica, «Una variable es un signo que, en lugar de tener un designado fijo, "recorre" un espectro de posibles designados, y la determinación de un designado particular consistirá en la asignación de un valor para esa variable» (Lungarzo, 1986: I, 36).

clasificación no es exhaustiva, y requeriría del agregado de otros valores (o **categorías**). En cambio, sí resultan mutuamente excluyentes las categorías (o valores) que la componen: cualquier persona de cuatro o más años en nuestra cultura dispone de un sistema de reglas que le permite aplicar correctamente estos valores.

2. NIVELES DE MEDICIÓN

Puesto que las variables no consisten en otra cosa que en un conjunto de valores vinculados unos a otros por relaciones de diferente tipo, una base fundamental para la clasificación de las variables es la que se encuentra en el tipo de relaciones que mantienen entre sí sus valores. Dicho de otro modo, las variables pueden ser muy diferentes en cuanto a su estructura interna. La importancia de esta distinción es crucial, porque las estructuras constitutivas de las variables determinan totalmente el tipo de operaciones a las que pueden ser sometidas.

Remitiéndonos a la clásica distinción entre ciencias fácticas y ciencias formales, recordemos que Bunge (1968) destaca el rol instrumental de las ciencias formales; es en este sentido que se puede decir que todas las ciencias fácticas recurren a la Lógica y todas tienden al uso de las Matemáticas.¹¹ La razón principal para ello es que las matemáticas pueden ser vistas como un arsenal de modelos de distinto tipo aplicables a la materia de cada una de las disciplinas que versan sobre la naturaleza o la sociedad. A partir de la formalización de determinados aspectos que hacen a los objetos reales de las ciencias, el uso de las matemáticas habilita para operar simbólicamente sobre los objetos de conocimiento, con toda la ganancia en eficiencia y precisión que ello entraña para la práctica científica.

Ahora bien, al decir de Blalock, «el empleo de un determinado modelo matemático supone que se ha alcanzado un determinado nivel de medición» (1966: 29). Es decir que la posibilidad de hacer uso de modelos matemáticos está dependiendo de que se hayan desarrollado procedimientos de medición satisfactorios. Desde este punto de vista, no es del todo errónea la idea que lo que distingue a las ciencias sociales en general de las ciencias de la naturaleza globalmente consideradas son los distintos niveles de medición en que operan predominantemente unas y otras.¹² Por supuesto, existen también diferencias -y no menores- dentro de estos dos grandes conjuntos de disciplinas. Así, la Economía se nos aparece como una ciencia social esencialmente cuantificadora, que nos habla permanentemente en **números**: horas de trabajo, toneladas de producto, valor en pesos, etc.; ha avanzado mucho más que la Sociología en dirección hacia la cuantificación.¹³

A partir del clásico trabajo de Stevens de 1951,¹⁴ es corriente considerar que existen cuatro distintos niveles de medición: nominal, ordinal, intervalar y racional.¹⁵ Algunos autores definen a la medición como el proceso de asignar números para representar cantidades,¹⁶ y excluyen de este modo el nivel nominal. Otros, a cuyo criterio nos atendremos, no consideran que la cuantificación sea una condición necesaria para poder hablar de ‘medición’. Así, por ejemplo, se puede definir ‘medida’ como «cualquier procedimiento empírico que suponga la asignación de símbolos, de los

¹¹ Puede cuestionarse este rol “instrumental” de las matemáticas con respecto a las ciencias fácticas. Así, según Althusser (1985 : 31 y ss.) el papel de las matemáticas en la Física era *constitutivo* antes que meramente instrumental: la Física es impensable sin el cálculo. Empero, aún desde esta perspectiva, las matemáticas son distintas de la Física, en tanto existen con independencia de ésta y se utilizan en otras disciplinas.

¹² Cf., por ejemplo, Torgerson (1965).

¹³ Como lo demuestra *a contrario* el estudio de las sociedades primitivas, esta posibilidad de cuantificación de la Economía está directamente ligada a una característica histórica de su objeto real: la aparición del *dinero* como medio de intercambio generalizado y la reducción de todas las mercancías, fuerza de trabajo incluida, a una base común producida por la generalización del modo capitalista de producción.

¹⁴ Reproducido en la compilación de Wainerman (1976).

¹⁵ Coombs (en Festinger y Katz, 1976) distingue seis niveles o escalas. Cf. también las críticas de Duncan (1984) a Stevens.

¹⁶ Así, Blalock define: « Por medición, nos referimos al proceso general por el que se asignan números a objetos de tal modo que queda establecido qué tipo de operaciones matemáticas pueden ser utilizadas legítimamente, dada la naturaleza de las operaciones físicas que se han usado para justificar o racionalizar dicha asignación de números a objetos» (1982: 11).

cuales los numerales son solamente un tipo, para los objetos o sucesos que concuerdan con las reglas previamente establecidas» (Sellitz *et al.*; 1965: 219). Lo cierto es que la operación de medición será posible solamente en tanto y en cuanto se pueda determinar una cierta correspondencia entre las características discernibles de nuestros objetos de estudio y las propiedades de los modelos matemáticos que les sean aplicables. Según resumen Padua y Ahman: «La medición, de hecho, corresponde a una serie de teorías conocidas como *niveles de medición*: detrás de cada uno de los distintos niveles de medición están operando una serie de principios lógicomatemáticos, que van a determinar o no el isomorfismo entre un concepto y el nivel de medición apropiado (...) se busca que el modelo matemático sea isomórfico con el concepto; esto es, que la ‘forma’ del modelo sea idéntica a la ‘forma’ del concepto» (Padua, 1979: 154).

Al hablar de *niveles* de medición se está aludiendo ya a la existencia de una jerarquía entre los modelos matemáticos aplicables a los objetos; y es que, en efecto, los hay más poderosos que otros. Presentaremos a continuación las características básicas de los cuatro niveles distinguidos por Stevens, comenzando por el más bajo de ellos.

Nivel nominal

La operación más básica y a la vez más sencilla de toda ciencia es la *clasificación*. Al clasificar, agrupamos nuestras unidades de análisis en clases mutuamente excluyentes basándonos en sus semejanzas y diferencias. De este modo obtenemos una escala nominal, que corresponde al nivel más bajo de medición, y que es a su vez una condición necesaria que deben satisfacer los niveles más elevados. El requerimiento fundamental para construir una escala nominal es que se puedan distinguir dos o más categorías significativas y que se establezcan los criterios según los cuales los individuos, grupos, objetos o respuestas serán incluidos en una o en otra categoría. Las categorías -exhaustivas y mutuamente excluyentes- no mantienen otra relación entre sí que la de ser distintas unas de otras: sobre todo no existe implicación alguna de que puedan estar representando una cantidad mayor o menor de la propiedad que se está midiendo.

En ciencias sociales son muchas las propiedades que se encuentran definidas nominalmente: sexo, estado civil, nacionalidad, religión, afiliación partidaria, etc. Es cierto que nada prohíbe la utilización de números para identificar las diferentes categorías, codificando por ejemplo ‘masculino’ como ‘1’ y ‘femenino’ como ‘2’; pero es esencial comprender que aquí los números no funcionan como tales, en el sentido de que no autorizan el empleo de las operaciones aritméticas usuales. Así, carecería de sentido sumar los códigos que presenten en la variable ‘sexo’ todos los integrantes de un grupo familiar, del mismo modo que resultaría disparatado sumar sus números de documento de identidad; pero sí en cambio la suma de sus ingresos mensuales en pesos tendría un significado claro. En el nivel nominal, si utilizamos números les hacemos cumplir la función de meros **nombres** que nos permiten identificar las categorías designándolas de modo diferente. La operación de medición queda por lo tanto limitada a referir una observación a una clase o categoría para luego poder contar cuántos casos caen dentro de cada categoría. Esto limita severamente las operaciones estadísticas utilizables que no van más allá del número de casos, el cómputo de porcentajes, el modo, el cálculo de algunos coeficientes de asociación en tablas de contingencia (Q, Phi, etc.) y de la chi-cuadrada.

Nivel ordinal

En este nivel aparece la idea de orden. No sólo se puede distinguir entre distintas categorías, sino que es además posible afirmar si una unidad de análisis posee en mayor, en igual o en menor grado que otra la propiedad que se está midiendo. Así, por ejemplo, en la escala de la jerarquía militar no solamente podemos decir que un capitán, un teniente y un sargento son diferentes entre sí, sino que dichas categorías pueden ordenarse con respecto a la **autoridad** que reviste cada uno de esos grados:

Capitán > Teniente > Sargento

Pero lo que no podemos determinar con exactitud es cuanta más autoridad tiene un capitán que un teniente, ni si la diferencia entre éstos es menor o mayor que la que media entre los grados de sargento y de teniente. « Una escala ordinal define la posición relativa de objetos o individuos con

respecto a una característica, sin implicación alguna en cuanto a la distancia entre posiciones».¹⁷ La escala de dureza de Mohs, que se aplica a los minerales, es un clásico ejemplo de escala ordinal en las ciencias de la naturaleza.

En ciencias sociales las variables medidas en el nivel ordinal son muy numerosas: los índices de nivel económico-social (NES), el prestigio ocupacional, el grado de radicalismo político o de nacionalismo, la religiosidad, etc. Así, es frecuente que los NES se ordenen de 1 a 7, sin que ello signifique que los números funcionen plenamente como tales, puesto que sólo representan relaciones del tipo ‘mayor que’ o ‘menor que’. Puede afirmarse que el NES 7 es distinto del NES 6; puede decirse también que el NES 7 es más alto que el 6; pero no podemos saber si la diferencia entre ambos es *igual* a la que separa el NES 2 del 1, ni tampoco podremos sostener que el valor 4 representa el *doble* de status socioeconómico que el 2. No obstante ello, se amplían en este nivel las medidas estadísticas utilizables; además de las aplicables a escalas nominales, se agregan la mediana, los cuartiles y percentiles, coeficientes de correlación de rangos como el de Spearman, etc.

Nivel intervalar

Cuando se alcanza a construir una escala de *intervalos iguales*, no solamente es posible diferenciar y ordenar las unidades de análisis, sino que también se puede especificar la distancia que separa a unas de otras en una determinada propiedad. Este nivel de medición requiere que se establezca algún tipo de unidad de medida que pueda ser considerado por todos como una norma común y que sea repetible, esto es, que se pueda aplicar indefinidamente a los mismos objetos produciendo los mismos resultados. Para el caso, ello es lo que ocurre en Física con las escalas de temperatura Celsius y Fahrenheit. En estas escalas, las unidades representan cantidades iguales de cambio en el volumen de una columna de mercurio bajo una determinada presión. Así, es posible afirmar por ejemplo que entre 0 C. y 10 C. existe *la misma diferencia* de temperatura que entre 10 C. y 20 C., vale decir, 10 grados Celsius en ambos casos. Pero subsiste todavía una limitación que deviene de la ausencia de un punto *cero absoluto* y que hace que las operaciones de multiplicación y de división carezcan de sentido, con lo que las proporciones resultan también carentes de significado. Tratando con una escala meramente intervalar como la de Celsius, no puede sostenerse que una temperatura de 20 C. sea el *doble* de una de 10 C. Debido a la arbitrariedad del punto cero en la escala Celsius -la temperatura de congelamiento del agua- no es posible formular este tipo de relaciones.¹⁸ Para la mayoría de las variables en ciencias sociales no existen procedimientos de medición que garanticen tan siquiera la igualdad de intervalos; Thurstone no se equivocó al denominar a su escala “de intervalos aparentemente iguales”. En este nivel se amplían sustancialmente los estadísticos aplicables, que incluyen ya a la media aritmética, la variancia, diversos modelos de regresión, y el coeficiente producto-momento de Pearson.

Nivel racional

También es llamado nivel ‘de razones’ o ‘por cocientes’: se alcanza al producir una escala de intervalos que cuenta con un punto cero absoluto. Es solamente en este nivel que se justifican afirmaciones del tipo «A pesa el doble que B», o bien «Yo gasté la mitad que vos». En Física, son numerosas las escalas definidas en este nivel: distancia en metros, intervalos de tiempo en segundos, medidas de peso, masa, velocidad, resistencia eléctrica, etc.

Empero, la distinción entre escalas de intervalos y de razones resulta irrelevante las más de las veces. Según Blalock, « esta distinción... es puramente académica, ya que es extremadamente difícil encontrar una escala legítima de intervalos que no sea al propio tiempo una escala de proporciones. Esto se debe al hecho de que, una vez establecida la magnitud de la unidad, es casi siempre posible concebir cero unidades, pese a que nunca podamos hallar un cuerpo que no posea largo o masa, u obtener una temperatura de cero absoluto. Así pues, prácticamente en todos los casos en que se dispone de una unidad, será legítimo emplear todas las operaciones de la

¹⁷ Definición de Selltitz *et al.* (1965: 219).

¹⁸ Esto se hace evidente mediante el sencillo expediente de expresar estas mismas temperaturas en grados Fahrenheit (mediante la fórmula: $F = 1,8 C + 32$) con lo que la relación no se mantiene: 68 F. no es ciertamente el “doble” de 50 F.

aritmética, incluidas las raíces cuadradas, las potencias y los logaritmos» (1966: 27).

De acuerdo a esto, a todos los efectos prácticos podemos manejarnos reteniendo la distinción entre los tres niveles que aparecen en la Figura 1.3. Habrá que tener presente siempre que los niveles superiores presentan, además de las propiedades que les son propias, todas las características que pertenecen a los niveles inferiores. Así, por ejemplo, a una variable de nivel intervalar como el ingreso mensual en pesos, se le puede aplicar -además de la media aritmética- la mediana y el modo aunque, por cierto, al precio de una gran pérdida de información.

Figura 1.3: Resumen de las características de los diferentes niveles de medición

Nivel de medición	Operación básica	Medidas y análisis estadísticos	Ejemplos de variables
<i>Intervalar</i>	Distancia entre categorías ($a + b = c$)	Media aritmética, varianza, r de Pearson, regresión, ACP	Edad en años, Ingreso en \$, Tasa de suicidios
<i>Ordinal</i>	Orden de las categorías ($a > b$)	Mediana, cuartiles, percentiles, ρ (Spearman), Gamma	NES, Nacionalismo. Religiosidad, escalas de actitudes
<i>Nominal</i>	Clasificación ($a \neq b$)	Nº de casos, Modo, %, Q , ϕ , V , χ^2 , AFC, ACM	Sexo, Situación Marital, Afiliación política, Nacionalidad, Religión

Algunos autores consideran que el subdesarrollo de las ciencias sociales obedece precisamente a la falta de medidas definidas en un nivel racional. Y es que, según Selltitz, «El descubrimiento de relaciones precisas entre características, tal como ha sido expresado en leyes numéricas, depende en buena parte de la existencia de escalas de razones» (1965: 225). Lo que se observa justamente en la historia de la ciencia es como se ha ido progresando de un nivel a otro. Como señala Stevens, «Cuando los hombres conocían la temperatura sólo mediante sensaciones, cuando las cosas eran sólo ‘más calientes’ o ‘más frías’ que otras, la temperatura pertenecía al tipo ordinal de escalas. Se convirtió en una escala de intervalo con el desarrollo de la termometría, y después que la Termodinámica utilizó la **razón** de expansión de gases para extrapolar a cero, se convirtió en una escala de razón». ¹⁹ Análogamente, la distinción entre los colores de los objetos que realizamos corrientemente en un nivel nominal, ²⁰ desde la Física puede ser considerada como una escala de intervalos: se trata de la reflexión de la luz en diferentes longitudes de onda.

Ahora bien, cabe preguntarse si las ciencias sociales, hoy subdesarrolladas, deberían llegar con el tiempo -¿cuánto más?- a mediciones intervalares de todas las propiedades fundamentales de sus objetos. ¿O, por lo contrario, es que existe una característica intrínseca a dichos objetos que limita las posibilidades de cuantificación?²¹ No existe una única respuesta a estos interrogantes. Si se atiende a la dimensión simbólica de lo social, a la importancia del significado atribuido por los actores a sus acciones, a la representación que se hacen de las relaciones sociales como un elemento que las constituye como tales, entonces puede carecer de sentido proponerse la medición

¹⁹ Cf. artículo reproducido en Wainerman, 1976: 64.

²⁰ Aunque de modo variable en culturas diferentes.

²¹ Según Passeron, tal subdesarrollo no responde a problemas de medición, sino a una conformación epistemológica insuperable de las disciplinas que denomina “ciencias del curso del mundo histórico” en las que «Las constataciones se dan en un contexto que puede ser *designado* pero no *agotado* por un análisis finito de las variables que lo constituyen y que permitirían razonar *ceteris paribus*.» (Passeron, 1991: 25).

de ciertas variables en un nivel más elevado que el que se plantean estos mismos actores. Así, para dar un ejemplo burdo, el valor peculiar que puede revestir el color ‘rojo’ en nuestra cultura política es algo que no mantiene ninguna relación con su longitud de onda.²² Es decir, propiedades medibles en un nivel intervalar pueden funcionar socialmente en un nivel puramente nominal.²³

Por otra parte, también es frecuente que se produzca el caso inverso, cuando a propiedades nominalmente definidas se les puede atribuir un significado ordinal. Sin ir más lejos, es lo que ocurre con la variable ‘ocupación’, cuyas diferentes variedades distinguidas nominalmente son susceptibles de ser ordenadas según niveles de prestigio.²⁴ Análogamente, es corriente que en la confección de muchos índices se parta de propiedades clasificatorias a las que, a los efectos de la elaboración del índice, se les confiere un determinado valor numérico (cf. *Infra*: capítulo 5). Esto supondrá siempre ordenar las categorías de acuerdo con algún criterio, con lo que la propiedad dará lugar a la generación de una variable de nivel ordinal. Finalmente, el análisis de correspondencias (cf. capítulo 6) es otra vía por la que a partir de un conjunto de atributos es posible generar variables en un nivel intervalar.

²² Sólo pretendo ilustrar la escasa importancia de la “objetividad” de ciertas características, cuando su única función es de la ser soporte de significaciones para los actores. Una discusión en buen nivel de este problema más general puede encontrarse en la obra de Cicourel (1982: capítulo 1).

²³ Integradas a sistemas clasificatorios que se estructuran en base a oposiciones simples: “rojo/blanco”, “frío/calor”, etc. Y en este caso habrá que tomar en cuenta lo que recomendaba Abraham Moles: «La precisión del instrumento de medida debe ser del orden de precisión intrínseco de la cosa a mensurar» (1995: 130).

²⁴ Más estrictamente, deberíamos hablar en este caso de un único objeto -la ‘ocupación’- del que podemos predicar varias propiedades (calificación, prestigio, etc.).

CAPÍTULO 2

LA MATRIZ DE DATOS

Ante cualquier objeto de investigación, ya sea éste de significación teórica o de importancia meramente práctica, las decisiones metodológicas propiamente dichas tienen necesariamente lugar dentro de un cierto marco conceptual. Dentro de ese marco se debe determinar: a) el grado en que dicho objeto es susceptible de ser producido en tanto objeto científico dentro de la estructura de la matriz de datos; y b) todas las operaciones que hagan a la producción del objeto dentro de este esquema, y muy particularmente las que tengan que ver con la definición de las unidades de análisis y de las propiedades que les sean aplicables.

En este capítulo, mostraremos primero el modo en se articulan los conceptos de ‘unidad de análisis’, ‘variable’ y ‘valor’ bajo la forma de la matriz de datos, y ejemplificaremos como se puede expresar una hipótesis en ese lenguaje. Nos encontramos ahora en condiciones de aplicar estos conceptos para desmenuzar el significado empírico de cualquier proposición.¹ Todo el problema consiste precisamente en expresar en el lenguaje de los datos el contenido de un enunciado, por lo que la tarea es propiamente la de lograr una *traducción* adecuada de ese contenido.

Comenzaremos con varios ejemplos de ejercitación de esta operación de traducción para generar en forma práctica la idea de la matriz de datos. Luego nos limitaremos a glosar las partes más pertinentes del texto de Galtung, e introduciremos algunos elementos acerca de modalidades alternativas de diseño de una matriz. Finalmente, la distinción de Lazarsfeld y Menzel entre propiedades individuales y colectivas nos permitirá terminar de dibujar un cuadro somero de las distintas posibilidades abiertas en el proceso de construcción de los datos.

1. LA FORMA DE LA MATRIZ DE DATOS

Si afirmamos: «En 1980, había en la Provincia de Misiones 50.553 hogares con necesidades básicas insatisfechas»,² podemos distinguir en este enunciado:

- 1.- Una unidad de análisis ⇔ la “Provincia de **Misiones**”;
- 2.- Una variable ⇔ el “*Número de hogares con necesidades básicas insatisfechas*”;
- 3.- Un valor ⇔ “50.553”.

Estrictamente, limitándonos a esta sola unidad de análisis, puede parecer poco pertinente hablar de una ‘variable’, puesto que ésta no variaría. No obstante, el número de hogares con necesidades básicas insatisfechas pudo haber sido otro. De hecho la ‘variable’ es tal en la medida en que hace posible la comparación entre varias unidades de análisis; así, por ejemplo, para ese mismo año de 1980 la unidad de análisis Capital Federal presentaba en esta variable un valor de “67.692”.³

Este enunciado es de los más simples, desde luego, como que se refiere a una sola característica

¹ Es esencial comprender que *todo* enunciado fáctico es, por definición, susceptible de ser reducido a estos tres elementos. No queremos decir con esto que siempre haya que proceder así (existen, por ejemplo, enunciados no-fácticos con sentido); ni tampoco que ésta sea la única ni la mejor manera de obrar en todos los contextos (a un psicoanalista este esquema no le aportaría nada para interpretar el discurso de su paciente). Sí, en cambio, sostenemos que no existe investigación empírica cuantitativa que no descansa en algún tipo de matriz de datos.

² Datos de Argentina, 1984.

³ Otra posibilidad de comparación es -como se verá luego- considerar la misma unidad de análisis en distintos puntos del tiempo.

de un objeto único. Ocurre que muchas veces estamos interesados en proposiciones que se refieren a toda una clase de objetos, y en las que se establecen relaciones entre varias de sus propiedades. Por ejemplo: « En las elecciones para Gobernador del 6 de septiembre de 1987 en Posadas, las mujeres votaron por los candidatos radicales en mayor proporción que los varones». Deberíamos distinguir aquí:

1. Varias unidades de análisis

⇒ los “**electores**”⁴ del 6 de septiembre de 1987 en Posadas;

2.- Dos variables

⇒ 2.1. *Sexo* ;

⇒ 2.2. *Dirección del voto*;

3.- Los valores que conforman estas variables

⇒ 3.1. “**1**”(Varón)/“**2**”(Mujer);⁵

⇒ 3.2. “**1**”(Radical)/“**2**”(No radical).⁶

La matriz es una forma de hacer inmediatamente visible la estructura tripartita de estos datos. Así, suponiendo que se haya trabajado con una muestra de 10 electores, tendríamos:

Tabla 2.1 Ejemplo de matriz de 10 x 2

Unidades de análisis	Variables	
	1. Sexo	2. Voto
1	1	2
2	2	1
3	1	2
4	1	2
5	2	1
6	2	2
7	2	1
8	1	1
9	1	2
10	2	2

Cada fila de la matriz corresponde a una unidad de análisis (identificada por un código de 01 a 10), cada una de las dos columnas a una variable, y en las celdas figuran los valores.⁷ Entre los varones, obtendríamos una proporción de 1/5 votos por el Radicalismo; entre las mujeres habría

⁴ Operacionalizar una proposición supone contar con una definición precisa de la unidad de análisis. En este caso, se podría definir como “electores” a todas las personas inscriptas para esa fecha en el padrón electoral de Posadas, por ejemplo.

⁵ Obsérvese que hemos adjudicado en forma arbitraria los códigos numéricos “1” y “2” para simbolizar respectivamente los valores “masculino” y “femenino”.

⁶ Es evidente que en cuanto a esta segunda variable ‘*dirección del voto*’ se podrían distinguir otros tantos valores como partidos se presentaron a dichas elecciones, pero dos valores son suficientes para traducir adecuadamente el significado de la proposición que nos ocupa. Este ejemplo es ilustrativo de cómo cualquier clasificación politómica (de más de dos categorías o valores) puede reducirse a una **dicotomía** (una clasificación de dos valores).

⁷ En álgebra matricial, se define más genéricamente como ‘matriz’ a “cualquier conjunto rectangular de números” (Namboodiri, 1984: 8).

3/5 votos radicales. Como $3/5 > 1/5$, deberíamos aceptar como verdadera nuestra hipótesis.⁸

Esta proposición, aun siendo verdadera, sería todavía muy puntual. Si aspiráramos a lograr un conocimiento más general del comportamiento electoral de la población argentina, podríamos tener interés en formular una hipótesis como «En 1987 las mujeres votaron por el Radicalismo en mayor medida que los varones». En este caso las variables y los valores serían los mismos, pero sería conveniente que nuestras unidades de análisis se multiplicaran hasta abarcar electores de todos los distritos del país. En este caso, lo que era una propiedad constante de nuestras unidades de análisis “estar inscripto en el padrón electoral de Misiones” pasaría a funcionar como una variable adicional que podríamos denominar ‘Provincia’.⁹ Con una muestra de 500 electores tendríamos:

Tabla 2.2: Ejemplo de matriz de 500 x 3

Unidades de análisis	Variables		
	1. Sexo	2. Voto	3. Provincia
1	1	2	17
2	2	1	21
3	1	2	4
4	1	2	7
5	2	1	8
6	2	2	14
7	2	1	13
8	1	1	1
9	1	2	1
10	2	2	2
...
499	1	2	24
500	2	1	19

Así, por ejemplo, la unidad de análisis “006” sería una “mujer” que votó “No-radical” en “Misiones”. De un modo más general, se puede describir a cualquier matriz de datos como respondiendo a la siguiente estructura:

Tabla 2.3: Forma general de la matriz de datos

	V1	V2	V3	...	Vj	...	Vp
UA1	R11	R12	R13	...	R1j	...	R1p
UA2	R21	R22	R23	...	R2j	...	R2p
UA3	R31	R32	R33	...	R3j	...	R3p
...
UAi	Ri1	Ri2	Ri3	...	Rij	...	Rip
...
UAn	Rn1	Rn2	Rn3	...	Rnj	...	Rnp

Fuente : adaptado de Galtung (1966: I, 3).

⁸ Obsérvese que si la relación hubiera sido $3/5 > 2/5$, también se verificaría la hipótesis; en efecto, ésta se encuentra formulada con escasa precisión. Por otra parte, a los efectos pedagógicos prescindimos por el momento de toda consideración sobre la **significación estadística** del resultado (que sería nula, con una muestra de $n=10$).

⁹ Los valores de esta variable podrían corresponder a las 23 Provincias más Capital Federal, y se codificarían de ‘01’ a ‘24’.

En cualquier investigación se considerará un número finito 'n' de unidades de análisis (UUA) y un número finito 'p' de variables (V). UA_n será entonces la última (o la "enésima") unidad de análisis incluida en una investigación.¹⁰ Por su parte cada una de las variables se compondrá de un número 'r' de valores. En el esquema presentado R_{ij} es el valor que presenta la UA_i en la variable j. Esta es muy precisamente la forma en que aparecen los datos en el monitor de una computadora; y es por tanto la forma en que deberán estar dispuestos nuestros datos para estar en condiciones de ser procesados electrónicamente.¹¹

Nos encontramos ahora en condiciones de recapitular los tres principios fundamentales (Galtung; 1966: I,4-6) que debe satisfacer una matriz de datos en su construcción.

1. Principio de *comparabilidad*: la proposición ' UA_i, V_j da el valor correspondiente en R_k ' debe ser verdadera o falsa para cada i, j y k . En otras palabras, a la idea básica de que todas las unidades han de ser medidas en las mismas variables, se agrega la condición previa de que cada una de las combinaciones de una variable determinada con una unidad de análisis debe *tener sentido*: debe ser verdadero o falso que una combinación ($UA_i V_j$) presenta un valor determinado R_{jk} . Por medio de este principio se hacen comparables las variables, las unidades y los valores. La comparabilidad se obtiene cuando las tres series se ajustan las unas a las otras de tal manera que se satisfaga la condición enunciada en este primer principio. Así, si UA_i es una nación, V_j la variable 'estado civil' y R_{jk} la lista de los elementos "casados, solteros, viudos y divorciados", cualquier combinación no será ni verdadera ni falsa, sino que simplemente no tendrá sentido

2. Principio de *clasificación*: para cada variable V_j la serie de las categorías de respuestas R_{jk} debe producir una clasificación de todos los pares ($UA_i V_j$) ($i=1, \dots, n$). Para cada variable la serie de sus valores debe formar una clasificación; para cada par $UA_i V_j$ deberá haber un R_{jk} (exhaustividad), y sólo uno (exclusión mutua) en que el par pueda ser ubicado. Aplicado a un cuestionario de encuesta, este principio significa que cada interpelado deberá marcar una, y sólo una, respuesta o categoría.¹²

3. Principio de *integridad*: para cada par ($UA_i V_j$) debe hallarse empíricamente un valor R_k . En lo que hace al trabajo empírico de llenado la matriz, el *desideratum* es no dejar ninguna celda vacía. En la práctica, se debe intentar que la cantidad de celdas sin información se mantenga lo más baja posible.¹³

Desde el punto de vista de la matriz de datos, es factible considerar que las investigaciones pueden diferenciarse según el número de unidades de análisis y de variables que toman en cuenta. Se puede pensar así en investigaciones más o menos "intensivas" -según la cantidad de variables que consideren- y más o menos "extensivas" -de acuerdo al número de unidades de análisis que sean observadas. Combinando ambas características, generamos una tipología de las investigaciones posibles.

Según Galtung, «Lo ideal es la combinación [muchas, muchas] -tantas unidades y tantas dimensiones como sea posible. Sin embargo, nosotros suponemos que la palabra 'muchas' se utiliza de tal manera que esto es imposible, por falta de recursos tales como tiempo, energía, personal y dinero» (1966: I, 9). Del mismo modo la investigación del tipo [una, una] no tendría

¹⁰ Hemos modificado la simbología de Galtung. Así, hablamos de n (y no m) unidades de análisis, lo que presenta la ventaja de ser congruente con el uso más habitual que se le da al símbolo 'n' (o 'N') para designar al número de casos (o de UUA) seleccionados para una muestra o de elementos existentes en el universo. En aras a una mayor simplicidad de exposición, renunciamos también a diferenciar 'variable' de 'estímulo' (y 'objeto'), 'unidad de análisis' de 'sujeto', 'valor' de 'respuesta'.

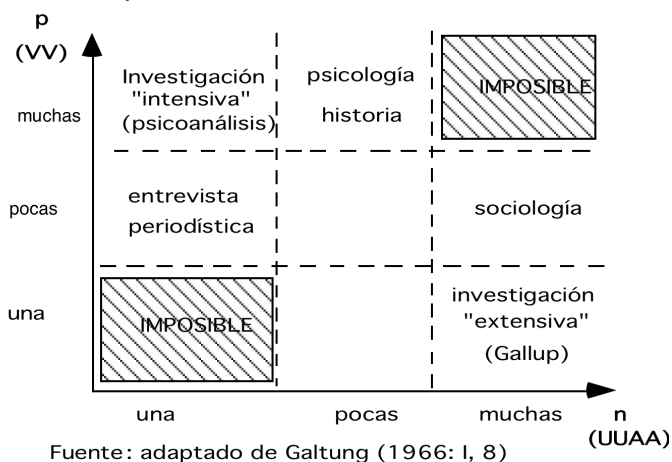
¹¹ Es también la forma más práctica de disponer los datos para su tabulación manual.

¹² En el caso en que al encuestado se le permite marcar varias respuestas, se le está presentando bajo una misma pregunta varias variables, una para cada categoría de respuesta, de tal manera que se encuentra ante dos valores posibles para cada variable: aceptación o rechazo.

¹³ «Como regla general de tipo práctico, puede decirse que un 10% es el máximo absoluto de celdas vacías admisibles en cualquier columna o cualquier fila de M y que un 5% es un máximo más aconsejable» (Galtung; 1966: I, 6).

sentido. El punteado de las líneas que delimitan las celdas intenta representar la dificultad en establecer un límite rígido entre las distintas cantidades de casos y de variables. El diagrama nos invita a pensar cómo, más allá de su ángulo inferior izquierdo, existe un vasto espacio de posibilidades para realizar investigaciones de naturaleza sumamente diversa y que en cuanto tales demandarán distintas habilidades. Las diferentes celdas representan en definitiva distintas **estrategias** de investigación.

Figura 2.1: Tipos de investigación, según el número de variable y de unidades de análisis involucradas



Si la matriz se reduce a una sola unidad -o a una única variable- se dice que "ha degenerado". En efecto, en ambos casos se termina renunciando a la posibilidad de la **comparación**, procedimiento intelectual que se encuentra en la base misma de la posibilidad del conocimiento. Como explica muy bien Galtung, existen varias razones por las que resulta conveniente evitar estas situaciones. Así, cabría preguntar "¿Por qué esta unidad y no otra?". Suponiendo en efecto que se pretendiera estudiar un sistema social, ¿Qué justificaría elegir un informante antes que otro?. No podemos sostener la creencia en el informante "puro"; la mejor prueba de ello es que si optáramos por otra unidad de análisis podríamos llegar a conclusiones muy distintas sobre nuestro objeto.¹⁴ La posibilidad misma de constatar variaciones y diferencias, y de evitar caer en estereotipos, requiere el poder comparar entre sí varias UA. Por la misma razón tampoco conviene trabajar con una sola variable: no existe la pregunta "pura".¹⁵ Disponer de varias variables permite comparar respuestas entre sí, y de este modo poder situar en un contexto la respuesta presentada en una variable, así como detectar patrones de respuesta.¹⁶

Finalmente, si el objetivo del conocimiento científico es establecer relaciones entre variables, su instrumento privilegiado habrá de ser la **correlación** -esto es, comparar valores en varias variables para un conjunto de UA. Y como se entenderá, no es posible el uso de la correlación sino a partir de un cierto número de unidades y contando por lo menos con dos variables.

La forma de la matriz de datos permite pensar con la mayor claridad la articulación entre las tres series de elementos (UA_i , V_j y R_{jk}) que concurren en la constitución del dato. Sin embargo, no siempre esta articulación se percibe nítidamente de modo inmediato. Un caso interesante se

¹⁴ «En Química o Física a menudo parece haber sido resuelto el problema de encontrar el caso puro. Si un químico desea comprobar una proposición acerca del sulfuro, puede tomar una cantidad cualquiera de sulfuro químicamente puro (siempre que su forma cristalina sea irrelevante para el experimento) y tratarlo como si fuera un puro y verdadero representante del sulfuro, S.» dice Galtung (1966: I, 10). Pero éste no es el caso en las ciencias sociales.

¹⁵ Constatación que se encuentra a la base misma de la necesidad de elaborar **índices** -esto es, variables complejas- para representar dimensiones complejas de una clase de fenómenos (cf. *infra*, capítulo 5).

¹⁶ El mismo Galtung señala excepciones a esta necesidad de contar con varias variables: «Cuando se busca la simple información acerca del contexto, o cuando se investiga una dimensión de actitud que ha ocurrido muy a menudo en el debate diario, quien investiga a través de una encuesta parece justificado al limitarse a una sola pregunta. Las encuestas de Gallup caen en una de estas categorías o en ambas, lo que las hace metodológicamente justificables desde este punto de vista» (1966: I, 10).

plantea cuando consideramos la misma unidad de análisis en distintos puntos del tiempo; “Misiones” en 1991 presentará muy probablemente un valor distinto en la variable “*Nº de hogares con NBI*” que el anotado para 1980. En rigor, cabría inquirir si “Misiones” en 1991 es la **misma** unidad de análisis que en 1980 o si se trata de dos unidades distintas; es obvio que en el fondo la respuesta a una pregunta de este tipo es materia de convención.

Interesa destacar en términos prácticos cómo se puede zanjar técnicamente esta cuestión: ¿Qué hacer cuando no se trabaja con varias UUAA sino con una sola, la que ha sido medida en el mismo conjunto de variables en distintos puntos del tiempo? Supongamos que contamos con un conjunto de datos para una UA, la Universidad Nacional de Misiones, medida en dos variables ‘Nº de Nuevos Inscriptos’ (NI) y ‘Nº de Desertores’ en cinco años sucesivos. Podríamos, por supuesto, presentar los datos bajo la forma de una matriz que hubiera degenerado en un único renglón, como en este ejemplo:

UA	NI82	NI83	NI84	NI85	NI86	De82	De83	De84	De85	De86
UNaM	858	983	1349	2358	1996	387	561	947	1530	1242

Nada nos prohíbe presentar nuestros datos bajo la forma de este vector-fila. Procediendo de este modo, estamos midiendo nuestra única UA -la UNaM- en diez variables distintas: en efecto, NI₈₂ es una variable, NI₈₃ otra, etc. Claro que es ésta una manera impráctica de presentar este conjunto de datos; no sólo esta forma es conceptualmente oscura sino que, en especial, es impropia para ser trabajada mediante cualquier programa estadístico. Obsérvese que al disponer de una sola UA, no es posible el cálculo de ninguna correlación.

Existe una alternativa mucho más eficiente; y es tratar a nuestra UA como si fueran otras tantas UUAA como medidas en diferentes puntos en el tiempo disponemos. Es decir, se asume que una UA es idéntica a sí misma, sólo en un determinado punto del continuo temporal. Debemos entonces presuponer que $UA_t \neq UA_{t+1}$. De hecho, así es cómo se resuelve técnicamente la cuestión: UNaM₈₂ y UNaM₈₃ funcionarán como dos unidades de análisis distintas. En la matriz que generemos, cada renglón corresponderá entonces a un año diferente:

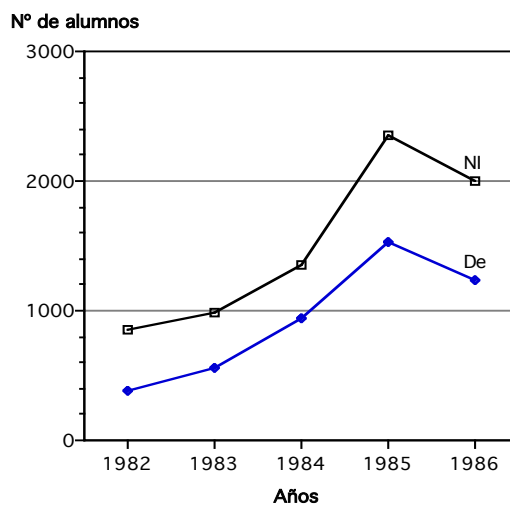
Años	Nuevos inscriptos	Desertores
1982	858	387
1983	983	561
1984	1349	947
1985	2358	1530
1986	1996	1242

Los mismos datos se han reestructurado asumiendo convencionalmente que corresponden a cinco UUAA medidas en dos variables. Bajo esta nueva presentación, salta a la vista la relación existente entre las dos variables: el crecimiento del número de desertores muestra una clara asociación con el incremento en el número de nuevos inscriptos.

Ahora bien, suponiendo que deseemos graficar estas dos series, la siguiente figura nos permite reconsiderar la cuestión desde un tercer punto de vista. Conceptualmente, se puede pensar ahora que estamos en presencia de una sola unidad de análisis -la UNaM- medida en tres variables: el año -representado en el eje horizontal-, su número de nuevos inscriptos y su número de desertores -ambas medidas en el eje vertical.¹⁷

¹⁷ Desarrollando una idea de Cattell de 1952, Rogers y Kincaid proponen la idea del “cubo de datos” (*Data-Cube*) una matriz tridimensional construida en base a (1) Unidades de Análisis (individuos o relaciones), (2) Variables, y (3) Tiempo (1981: 79-81).

Figura 2.2: Evolución de los nuevos inscriptos y desertores en la UNaM, 1982-86.



Fuente : Datos de la Secretaría G. Académica de la UNaM.

Ejemplo que debería terminar de convencernos acerca de este punto: no existe ninguna posibilidad de aplicar mecánicamente la estructura de la matriz a un conjunto de observaciones, sino que se requiere de todo un trabajo conceptual previo que es el que realmente produce -o “da”- los datos. En la sección siguiente, partiendo de la idea de ‘niveles de inclusión’, ahondaremos algo más en esta cuestión.

2. NIVELES DE INCLUSIÓN

Las variables son susceptibles de ser clasificadas de acuerdo a varios criterios, entre otros:

- a) el nivel de medición en que se inscriben: nominal, ordinal, intervalar;
- b) el papel que cumplen en el análisis: desde este punto de vista se podrá distinguir entre variables independientes, dependientes e intervinientes (cf. *infra*: capítulo 4);
- c) su grado de complejidad: se pueden reconocer así indicadores -variables simples- o índices.-variables complejas o compuestas- (cf. *infra*: capítulo 5)

Otro criterio para distinguir entre tipos de variables tiene que ver con el tipo de unidad de análisis al que se aplican; de acuerdo a éste, habrá que distinguir entre variables de individuo y variables de colectivo. Esta distinción es de fundamental importancia toda vez que en un proceso de investigación se encuentren involucrados diversos niveles de inclusión, y nos ocupará en esta sección.

2.1 Objeto de investigación y unidades de análisis

Retomemos nuestro concepto de ‘unidad de análisis’ (UA). De acuerdo a lo establecido en la primera sección, una UA es un sistema definido a partir de determinadas propiedades constantes. Lo primero que cabe aclarar es la diferencia entre el objeto de la investigación y la(s) UA con la(s) que se trabajará. Es importante destacar que no todo *objeto* científico constituye de por sí una UA. En primer lugar, existen objetos *teóricos* (por ejemplo, el modo de producción capitalista de Marx o el inconsciente de Freud) que son construcciones conceptuales que, aunque estén dirigidas en definitiva a “pensar” la realidad, en una investigación empírica no van a funcionar nunca como UA (ni tampoco como variables). En rigor estas construcciones conceptuales -explicitadas o no- se inscriben en el nivel de las condiciones previas a la construcción de cualquier dato científico. En términos de Althusser serían las que definen una *problemática* (1968), mientras que para Khun (1971) se ubicarían en el nivel del *paradigma*. Generalizando, se podría hablar para el caso de

construcciones *metaempíricas*.

Por otro lado, existen también objetos científicos -aunque ya no teóricos (en el sentido fuerte)- que tienen un correlato real empíricamente discernible.¹⁸ Para poner un ejemplo sencillo, podemos estar interesados en estudiar una comunidad local¹⁹ para conocerla y así planificar mejor las acciones de un programa de desarrollo tendiente a producir tales o cuales efectos. En este caso, el *objeto* de nuestro análisis estaría dado por ciertos procesos y características referidos a la comunidad como un todo la que, sin duda, en un determinado nivel funcionaría como una UA; sin embargo, más allá de alguna información referida a la ubicación geográfica de esta UA -la comunidad local-, a la situación de sus orígenes en el tiempo, a los hitos referentes a su “historia”, etc., nos encontraríamos muy pronto con que poco podríamos decir acerca de la misma si no dispusiéramos de datos acerca de los distintos integrantes -hogares, personas- que la componen: dentro de este estudio, se requeriría que cada uno de sus miembros funcionara a su vez como una *unidad de análisis*.²⁰ Describir a nuestro objeto ‘comunidad’ supondría entonces referirse a sus integrantes, a las acciones que realizan, y a las relaciones que mantienen entre sí.²¹

A título de ejemplo, podemos retomar la proposición que analizábamos anteriormente: «En 1980, había en la Provincia de Misiones 50.553 hogares con necesidades básicas insatisfechas»; considerábamos entonces a “Misiones” como la UA, al “Número de hogares con necesidades básicas insatisfechas” como una variable y a “50.553” como valor. ¿Pero cómo fue generado este dato, como se lo construyó? La respuesta es que, considerado en sí mismo, este dato supuso la definición de un universo “la Provincia de Misiones” que está compuesto por hogares, y desde este punto de vista dicho enunciado supuso considerar:

1. Varias UUAA ⇔ los **hogares** de la Provincia de Misiones;
2. Una variable ⇔ la “*Existencia de necesidades básicas insatisfechas*”;
3. Un valor ⇔ “*Sí*”.²²

Se puede pensar por lo tanto en la existencia de UUAA en distintos niveles de inclusión.

En la Figura 2.3, cada círculo pequeño simboliza una persona (en este caso, la unidad de menor nivel)²³, cada círculo mediano una unidad doméstica²⁴ (la unidad de segundo nivel), y el óvalo

¹⁸ Para disipar cualquier clase de equívoco, aclaremos que de atenderse a la distinción entre “objeto real” y “objeto de conocimiento” tal como la plantea Marx en el tercer apartado de la Introducción de 1857 (1971: I, 20 y ss.) los dos tipos de ‘objetos’ que distinguimos aquí formarían parte del “objeto de conocimiento”.

¹⁹ A los efectos que nos interesan, bastará con adoptar como definición de ‘comunidad local’ la de un grupo humano, de tamaño limitado, inserto dentro de una sociedad más amplia, y con una determinada base territorial.

²⁰ En realidad, hemos estado englobando en forma indiferenciada bajo esta expresión de ‘UA’ aspectos disímiles y que en tanto tales merecen ser distinguidos. Susana Torrado propone las siguientes convenciones terminológicas: «a) Unidades de análisis (UA): son las unidades “teóricamente” significativas desde la perspectiva de cada práctica o propósito analítico, por lo que su definición depende de las respectivas definiciones conceptuales.(...) b) Unidades de observación (UO): son las unidades acerca de las cuales se recogen datos a través del encuestamiento directo o indirecto». La distinción tiene su importancia porque, por lo general, la UA es operacionalizada a través de la definición de una UO. Torrado distingue también el conjunto de las Unidades de Cuenta (UC) y el de las Unidades estadísticas(UE); habitualmente, cada uno de estos conjuntos coincidirá con, o al menos incluirá, el conjunto de las UO (1983: 13-14).

²¹ El mismo Galtung incurre en algún equívoco acerca de la distinción entre ‘objeto’ y ‘UA’ cuando afirma: «A un cierto nivel en el análisis siempre se está estudiando solo una *unidad*, como cuando se hace una encuesta acerca del varón norteamericano y su conducta sexual» (1966: I, 11). De acuerdo a nuestro criterio correspondería para el caso hablar de *objeto*.

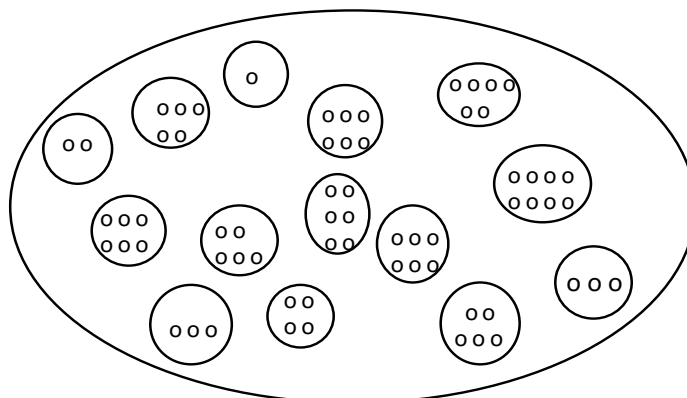
²² En el estudio sobre *La pobreza en Argentina* encontramos que los 50.553 hogares con NBI representan 39,2% de todos los hogares de Misiones. Para construir este dato, se consideraron los 128.962 hogares de la Provincia como UUAA., de los cuales 50.553 presentaban el valor “Sí” en la variable “Existencia de NBI” mientras que los restantes 78.409 presentaban el valor “No”.

²³ Como bien señalaban Lazarsfeld y Menzel, y el mismo Galtung, no es indispensable que la unidad de menor nivel coincida con personas individuales; por ejemplo, un individuo puede verse como un conjunto de roles, cada uno de los cuales constituiría la unidad de menor nivel. Una empresa puede ser considerada como un conjunto de normas, más o menos formalizadas, que se cumplen en menor o mayor medida, etc.

²⁴ Se puede definir la “unidad doméstica” como un «grupo de personas que interactúan en forma cotidiana, regular y permanente’ a fin de asegurar mancomunadamente el logro de uno o varios de los siguientes objetivos: su reproducción biológica, la preservación de su vida, el cumplimiento de todas aquellas prácticas, económicas y no económicas, indispensables para la optimización de sus condiciones materiales y no materiales de existencia»; lo que implica que

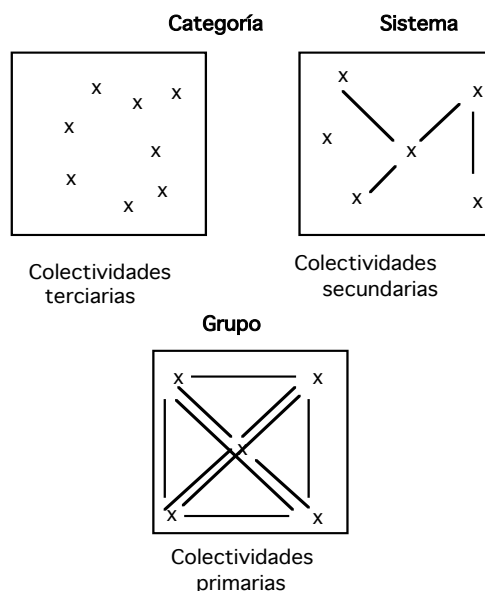
mayor representa una comunidad local (unidad de tercer nivel) integrada para el caso por catorce familias:

Figura 2.3: Unidades de análisis en tres niveles



Ahora bien, podemos creer que hay niveles de inclusión más interesantes que otros: aquellos niveles en los que se producen determinadas pautas o formas de organización, y que dan lugar a sistemas compuestos por UUAA de menor nivel que mantienen relaciones entre sí. En otras palabras, niveles que se corresponden con conjuntos estructurados de sistemas. En el límite estos sistemas pueden constituirse en verdaderos “actores” sociales colectivos, con una lógica propia. Pero no es imprescindible limitarse en el análisis a este único tipo de sistemas. El mismo Galtung reconoce que, partiendo de los individuos como unidades, pueden construirse distintos tipos de *superunidades*.

Figura 2.4: Una tipología de colectividades



Fuente: Galtung, 1966: I, 37.

Así, las ‘colectividades terciarias’ corresponden a lo que se denomina más frecuentemente ‘categoría’ (o incluso ‘agregado estadístico’), es decir, un conjunto de unidades sin estructura, conjunto que puede definirse a partir de la posesión de uno o varios atributos en común pero entre

exista *unidad de residencia*, que algunos miembros estén vinculados por relaciones de *parentesco*, que esta UD opere como una *unidad de consumo* o de *funcionamiento doméstico* (1983: 15-17).

cuyas unidades no existen relaciones. Por ejemplo, todos los votantes en la ciudad de Posadas en las elecciones del 6 de septiembre de 1987; o bien, los productores de trigo de la Argentina, etc. Un segundo tipo de unidad colectiva sería lo que Galtung denomina ‘sistema’, en este caso cada una de las unidades de primer nivel se halla en relación con al menos otra unidad: podría tratarse de los socios de un club, de los empleados de un ministerio, o de los personas pertenecientes a una Facultad de Ciencias Sociales; existe ya alguna estructura del colectivo, aún cuando sea bastante laxa. Por último, en el ‘grupo’ todas las unidades se encuentran relacionadas las unas a las otras, lo que resulta en una estructura fuertemente conectada; podrá ser el caso de un grupo familiar, de los obreros en un taller, de los miembros de un pequeño departamento en una universidad, etc.

Ciertamente es difícil pensar que cualquier tipo de colectividad ‘terciaria’ (en el sentido de Galtung) llegue a funcionar como un “actor”. No obstante ello, niveles de inclusión que definen este tipo de superunidades son absolutamente legítimos y se utilizan frecuentemente en la práctica de la investigación aplicada: éste es el caso cuando se comparan entre sí unidades geográficamente determinadas como lo son los países, regiones, provincias, departamentos, etc.²⁵ Por lo tanto, y a nuestros efectos, podemos establecer que cualquier nivel de inclusión puede determinar un conjunto de UUAA susceptible de constituirse en un objeto pertinente para la investigación.

2.2 La tipología de Lazarsfeld y Menzel

Comenzaremos ahora a abordar la tipología de propiedades individuales y colectivas que se presenta en el clásico trabajo de Lazarsfeld y Menzel. Primero, aclararemos la distinción entre colectivos y miembros, lo que nos permitirá continuar ejemplificando cómo puede traducirse en datos el significado de diversas proposiciones; luego presentaremos los diferentes tipos de propiedades incluidos en la tipología de esos autores, y finalizaremos comentando algunos problemas que se suscitan en la aplicación de la tipología.

Colectivos y miembros

Entendiendo que cualquier investigación empírica va a tender a producir la corroboración de determinadas proposiciones que versan sobre objetos reales, el punto central estará en el carácter de ‘colectivos’ o de ‘miembros’ que se les pueda adjudicar a estos objetos definidos como UUAA. Las UUAA son ‘miembros’ si se las puede considerar como incluidas dentro de unidades de mayor orden denominadas ‘colectivos’. Por lo contrario, las UUAA son ‘colectivos’ cuando es posible descomponerlas en unidades de menor nivel llamadas ‘miembros’. Los de ‘colectivo’ y ‘miembro’ son por lo tanto conceptos que se definen estructuralmente: todo miembro lo es de un colectivo, y todo colectivo lo es a su vez de sus miembros. Así en nuestro ejemplo de la Figura 2.3, cada persona sería un miembro de su colectivo: la unidad doméstica; pero a su vez cada unidad doméstica podría considerarse como miembro con relación a la unidad de mayor nivel que sería la “comunidad”.²⁶

Ahora bien, para una proposición dada las UUAA podrán ser o bien miembros o bien colectivos, o bien ambas cosas, según cuales sean los niveles de inclusión implicados. Hay que diferenciar entonces tres posibilidades:

<i>Niveles de inclusión implicados</i>	<i>Carácter de las UUAA</i>
a) Un sólo nivel	Colectivos o miembros
b) Dos niveles	Colectivos o miembros
c) Más de dos niveles	Colectivos y/o miembros

²⁵ De hecho, algunas de estas unidades, como los países y las provincias, en la medida en que no son meramente geográficas, pueden perfectamente llegar a funcionar como actores sociales. Por lo demás, cualquier proyecto social debe desarrollarse en un espacio determinado; ello lleva con frecuencia a la necesidad de fijar los límites geográficos del alcance del proyecto. Las superunidades o sub-universos así delimitados pueden carecer por completo de significado sociológico y, sin embargo, aún así no dejaran de ser relevantes tanto para la implementación del proyecto como para la evaluación de sus resultados.

²⁶ Por lo demás, nada impediría considerar varias comunidades como miembros del colectivo ‘departamento’, ni varios departamentos como miembros del colectivo ‘provincia’, etc.

En el caso a), cuando sólo está implicado un nivel de inclusión, la distinción es escasamente relevante. Por ejemplo, si se afirma:

« A mayor cantidad de años de educación formal, mayor ingreso mensual».

1. Hay un conjunto de UUAA
 - ⇒ presumiblemente **personas**;
2. Hay dos variables:
 - ⇒ 2.1. “ *Nº de años de educación formal*”;
 - ⇒ 2.2. “ *Ingreso mensual*”;
3. Con sus valores:
 - ⇒ 3.1. “0”, “1”, “2”, “3”, etc.
 - ⇒ 3.2. “Hasta 100”, 101-200, 201-300, etc.²⁷

Si nuestras UUAA son personas consideradas en sí mismas, sin que se requiera hacer referencia a ningún nivel de inclusión mayor o menor, estaremos relacionando dos variables que hacen referencia a propiedades individuales.²⁸

Podemos ejemplificar el caso b) con la proposición:

« En 1980, las viviendas rurales de Misiones se encuentran electrificadas en menor medida que las urbanas»

1. Hay un conjunto de UUAA:
 - ⇒ “**viviendas**”, que se definen por una serie de propiedades constantes: “estar en Misiones”, “en 1980”;
2. Hay dos variables:
 - ⇒ 2.1. “*Condición de electrificación*”;
 - ⇒ 2.2. “ *Tipo de zona*”;
3. Con sus valores:
 - ⇒ 3.1. “Electrificada”/ “No electrificada”;
 - ⇒ 3.2. “Rural”/”Urbana”.

Aquí las unidades se definen por una parte a partir de una propiedad que les pertenece propiamente como es la “condición de electrificación”; pero la segunda propiedad que se les aplica -ser rurales o urbanas- tiene que ver con una característica del colectivo al que pertenecen.

En el caso c) las unidades pueden ubicarse, por así decirlo, “abajo”, “arriba” o “en el medio”. Si decimos:

« En las localidades rurales más aisladas, las unidades domésticas presentan un menor ingreso monetario per cápita ».

1. Hay un conjunto de UUAA:
 - ⇒ “ **unidades domésticas**”;
2. Hay dos variables:
 - ⇒ 2.1. *Condición de aislamiento de la localidad*;
 - ⇒ 2.2. *Ingreso monetario anual per cápita* ;²⁹
3. Con sus valores:
 - ⇒ 3.1. “Aislada”/ “No aislada”;
 - ⇒ 3.2. “Hasta 200”, “201-400”, “401-800”, “801-1600”, etc.

²⁷ Los valores de la variable “Ingreso mensual” aparecen agrupados por intervalos, como si los datos fueran a ser presentados en una tabla de contingencia; es claro que nada obliga a agrupar los valores de la variable: éstos podrían ser “93”, “168”, “346”, etc. (así tendrían que presentarse los valores si, por ejemplo, se quisiera realizar un análisis de variancia para diferentes niveles de educación formal).

²⁸ También podrían tomarse departamentos, o provincias, o países, como UUAA, si se consideraran, por ejemplo, las variables “*Porcentaje de personas con 12 o más años de educación formal*” y “*Porcentaje de personas con ingresos que superan los 500 pesos por mes*”. Ciertamente, para construir el dato habría que hacer referencia a unidades de menor nivel (las personas), pero ambas variables remitirían a propiedades colectivas.

²⁹ Variable que se podría definir, por ejemplo, como la suma en pesos de todos los ingresos monetarios de una unidad doméstica durante un año, dividida por su número de miembros.

Previamente, una breve digresión acerca del significado de esta hipótesis. A primera vista podría pensarse que la proposición hace referencia a una sola variable, interpretando que su alcance se encuentra limitado al universo de las unidades domésticas pertenecientes a las localidades rurales más aisladas. Sin embargo, la lectura cuidadosa del enunciado no debe dejar dudas sobre el hecho de que la pertenencia a una localidad rural más aislada es una característica que se debe tomar aquí como una propiedad variable (y no constante). En efecto, en la segunda parte del texto de la hipótesis se hace referencia a un **menor** ingreso per cápita, con lo que se introduce la idea de que dicho ingreso “menor” lo será con respecto a otras unidades que no comparten la propiedad de pertenecer a localidades rurales aisladas. Implícitamente, entonces, el enunciado de la hipótesis está significando:

« En las localidades rurales más aisladas, las unidades domésticas presentan un menor ingreso monetario *per capita* [que en las localidades rurales menos aisladas]». ³⁰

Retomando ahora el hilo de nuestro argumento, la hipótesis ejemplifica la posibilidad de que se encuentren implicados tres niveles, siendo que en este enunciado las UUAA se ubican en el nivel intermedio. En efecto, al hablar de ‘ingreso *per capita*’ se está haciendo referencia necesariamente a los integrantes de la unidad doméstica que funciona como un colectivo con respecto a sus miembros.³¹ Pero a su vez las UUAA -esas mismas unidades domésticas- se diferencian de acuerdo a una característica del colectivo de mayor nivel -la “localidad”, que puede estar en una situación de menor o mayor aislamiento- y con relación al cual funcionan como miembros.

Otro ejemplo de proposición que implica tres niveles, pero con una ubicación diferente de las UUAA podría ser:

« Los miembros de las unidades domésticas más recientemente asentadas en los barrios marginales de Posadas, perciben remuneraciones horarias menores por su trabajo, [que los miembros de las unidades domésticas con mayor antigüedad en esos barrios] » .

1. Hay un conjunto de UUAA:

⇒ **personas que residen en barrios marginales y trabajan;**

2. Hay dos variables:

⇒ 2.1. *Antigüedad del asentamiento de la unidad doméstica* ;

⇒ 2.2. *Remuneración horaria -promedio percibida por el trabajo* ;

3. Con sus valores:

⇒ 3.1. “Hasta 6 meses”, “7-12 meses”, “13-24 meses”, etc.;

⇒ 3.2. “Menos de 1 Peso”, “1-1,49”,

“1,50-1,99”, etc.

En este caso las UUAA pertenecen al menor nivel: se trata de individuos miembros de unidades domésticas que a su vez son miembros del colectivo más amplio “barrio”.³²

Propiedades individuales y colectivas

La tipología (Lazarsfeld y Menzel, 1979: 79) está basada en las operaciones de medición involucradas en la construcción de cada tipo de propiedad, e incluye un conjunto de propiedades para caracterizar a las UUAA que funcionen como miembros, y otro conjunto para las UUAA colectivas.

³⁰Utilizamos los corchetes para representar la parte implícita del enunciado. El caso podría ser distinto si el texto rezara: «En las localidades rurales más aisladas, las unidades domésticas presentan un **bajo** ingreso monetario per cápita» (pero ello siempre y cuando se contara con una definición **absoluta** de lo que se debe entender por un “bajo” ingreso per cápita).

³¹ El “ingreso *per capita* anual” puede ser definido como la suma de todos los ingresos monetarios de los miembros de una UD, dividida por el número de miembros de esa UD.

³² Nada impediría considerar que se trata aquí de una proposición con tres variables, haciendo que la hipótesis alcanzara a los pobladores de todos los barrios de Posadas, y no solamente los marginales. La “*Remuneración horaria-promedio*” continuaría siendo la variable dependiente, mientras que como variable independiente funcionarían tanto la “*Antigüedad del asentamiento de la UD*” como el “*Tipo de barrio*” (“Marginal”/“No marginal”).

Figura 2.5: Tipología de Lazarsfeld y Menzel

PROPIEDADES	COLECTIVAS	Analítica
		Estructural
		Global
	INDIVIDUALES	Absoluta
		Relacional
		Comparativa
		Contextual

En primer lugar tenemos las propiedades *analíticas*, que “son propiedades de colectivos que se obtienen a través de operaciones matemáticas con alguna propiedad de cada miembro” (1979:85). Así, el ingreso per cápita de una unidad doméstica es una propiedad analítica de este colectivo, obtenida a partir de la suma de todos los ingresos de sus miembros individuales. Los promedios son propiedades analíticas, al igual que todas las propiedades construidas como porcentajes a partir de propiedades de miembros: el porcentaje de analfabetos en la provincia de Catamarca, el porcentaje de migrantes recientes en el Partido de La Matanza, el porcentaje de habitantes de ascendencia polaca en el Departamento de Apóstoles, etc. También las correlaciones pueden usarse para caracterizar colectivos: así, en un estudio de sociología electoral sería posible caracterizar a las diferentes provincias argentinas según el grado de correlación existente entre clase obrera y voto peronista, por ejemplo.³³

Se distinguen también las propiedades *estructurales*, que “son propiedades de colectivos que se obtienen a través de algunas operaciones con los datos sobre las relaciones de cada miembro con todos o alguno de los otros miembros” (:86). Si se está tratando con clases (colectivos) de una escuela compuestas por alumnos (miembros), se puede a partir de la información que cada alumno ha dado sobre cuál es su “compañero preferido”, construir un sociograma, una representación gráfica de estas relaciones de preferencia dentro de cada grupo. Dentro de este contexto se suele hablar de “estrella” para referirse a los individuos que reciben muchas elecciones positivas (representadas por líneas que convergen hacia él). Las distintas clases de la escuela podrán entonces diferenciarse según el número de estrellas que exista en cada una, y ésta será una propiedad estructural de cada clase. Hans Zeisel presenta el siguiente ejemplo: a todos los integrantes de un grupo se les pide que expresen su actitud hacia los demás miembros, otorgándole a cada uno un puntaje variable en una escala de 5 puntos que va desde +1 (máxima aceptación) a -1 (máximo rechazo), con un punto medio neutro (0). Zeisel plantea que para medir una propiedad estructural del grupo, su nivel general de cohesión, puede utilizarse como índice el promedio aritmético de todas las puntuaciones otorgadas (Zeisel, 1980: 105-108).

Por último, tenemos a las propiedades colectivas *globales*, las “que no se basan en información sobre las propiedades de los miembros individuales” (:87). Se trata entonces de una categoría residual de propiedades colectivas: son globales todas las propiedades que no son ni analíticas ni estructurales. Así, se ha podido distinguir entre sociedades con escritura y sociedades sin escritura. Debe insistirse en que esta propiedad global de una sociedad (“el tener escritura”) es por completo independiente del porcentaje de sus miembros que saben escribir;³⁴ así, la China clásica era una

³³ Se trataría de una propiedad analítica de cada provincia obtenida a partir de la correlación entre dos propiedades analíticas (los porcentajes de votos peronistas y de obreros en cada departamento).

³⁴ De lo contrario, se trataría de una propiedad analítica.

sociedad con escritura, aunque el dominio de ésta fuera patrimonio de un grupo reducido de mandarines. También el número de habitantes de una ciudad (o de un departamento, provincia o país) es una propiedad global.³⁵ Otro ejemplo: el hecho de que la enseñanza de la religión sea obligatoria en las escuelas es una propiedad global de la Provincia de Salta, que la diferencia de Misiones. También podría registrarse para cada uno de los países de América Latina el porcentaje de su presupuesto que es destinado a la educación, una propiedad global que no siempre se ajusta al 25% recomendado por la UNESCO. Del mismo modo el número de diarios que se editan en una provincia es una propiedad global de dicho colectivo.

Las propiedades individuales más simples son las *absolutas*, “características de los miembros que se obtienen sin hacer uso de información ni sobre el colectivo ni sobre las relaciones del miembro que se describe con otros miembros” (:90). La mayoría de las propiedades que se utilizan habitualmente para describir individuos pertenecen a esta categoría: edad, sexo, estado civil, ingreso mensual, condición de alfabetismo, ocupación, religión, etc.

En cuanto a las propiedades *relacionales*, éstas “se computan a partir de la información sobre las relaciones sustantivas entre el miembro descripto y otros miembros” (:91). Así, la popularidad medida sociométricamente es una propiedad relacional de un miembro de un grupo. En el mismo ejemplo citado de Zeisel, cada miembro puede ser caracterizado tanto por la suma de los puntajes que ha obtenido de los otros miembros (“popularidad”) como por la suma de los puntajes que ha otorgado a los otros miembros (como una medida de su “sociabilidad activa”). También el poder, en su acepción weberiana clásica de “probabilidad de encontrar obediencia”, es una propiedad eminentemente relacional. Por su parte Bourdieu, refiriéndose a las clases sociales, distinguía entre la “situación” de clase -determinada a partir de un conjunto de características económicas, sociales y culturales, consideradas intrínsecamente- y la “posición de clase” -en que se encuentra una clase o sector por su ubicación dentro de una estructura. Lo que da lugar a diferenciar entre la clase determinada como una propiedad absoluta y la clase entendida como propiedad relacional: así, «la clase superior de una ciudad pequeña presenta la mayoría de las características de las clases medias de una ciudad grande» (Bourdieu, 1982: 5); si se tomara la clase como propiedad absoluta, no existiría clase superior en una ciudad pequeña.³⁶

Por su parte, las propiedades individuales *comparativas* “caracterizan a un miembro comparando su valor en alguna propiedad (absoluta o relacional) y la distribución de esta propiedad en el colectivo total del cual es miembro” (:92). Así, por ejemplo, el hecho de que un alumno tenga un CI³⁷ de 110 es una propiedad absoluta suya, pero el que este mismo alumno sea *el más inteligente* de la clase (colectivo) es una propiedad comparativa, significando que ningún otro alumno de su clase lo supera en cociente intelectual; en cambio, con un CI de 110 un alumno de otra clase puede no ser el más inteligente si en ésta existen otros alumnos con mayor CI. Del mismo modo, el ser el hijo menor en una familia (colectivo) no implica ningún valor particular -o absoluto- en la variable “edad”, sino tener una edad menor que la de los hermanos. Otro ejemplo: considerados absolutamente, los habitantes de villas-miseria aparecen como uniformemente pobres; empero en cualquier estudio de esta población es posible plantearse una estratificación de estas familias en base a indicadores económicos dividiéndolas en nivel “alto”, “medio” y “bajo”; es claro que el nivel “alto” lo será meramente en términos comparativos.

Finalmente, están las propiedades *contextuales*, las que “describen a un miembro por la propiedad de su colectivo” (:92). Si se afirma: “los pobladores de las áreas urbanas marginales

³⁵ Podría pensarse que se trata de una propiedad analítica, obtenida contando la “existencia” de todos sus habitantes, pero como lo sugieren Lazarsfeld y Menzel (1969: 96), la existencia no es una propiedad (o en todo caso no es una propiedad *variable*); antes bien, es la condición de toda propiedad. La “existencia” se define por las propiedades constantes que definen la pertenencia de un elemento a un universo de investigación.

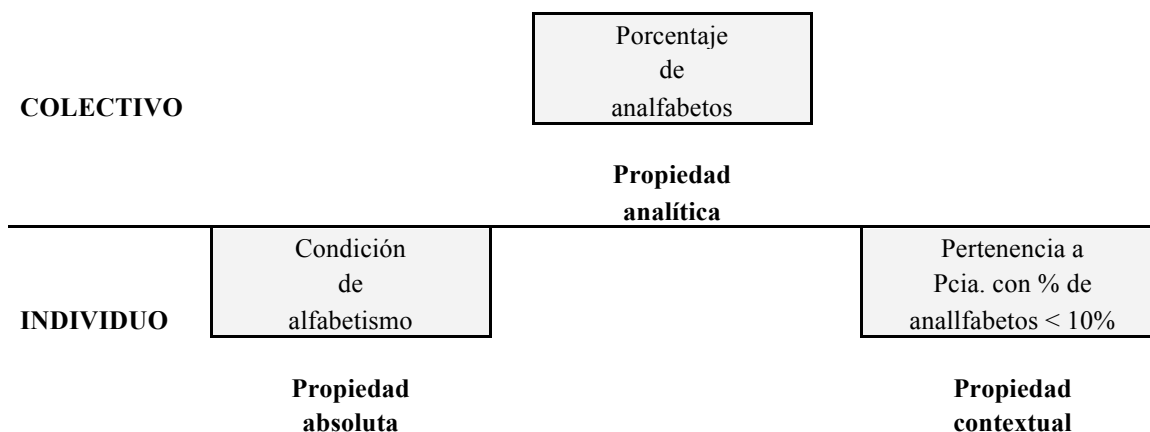
³⁶ El género es también una propiedad relacional por excelencia a la que Bourdieu trata de una manera en todo análoga a la clase. Así, hay una constancia transhistórica de «la estructura de la relación de dominación entre los hombres y las mujeres que se mantiene más allá de las diferencias *substanciales* de condición ligadas a los momentos de la historia y a las posiciones en el espacio social» (1998: 110).

³⁷ CI: Cociente Intelectual, medido a través de algún test de inteligencia.

presentan una mayor inestabilidad laboral”³⁸, el “ser poblador de un área urbana marginal” es una propiedad contextual de los individuos. Lo dicho supone que cualquier propiedad colectiva (ya sea analítica, estructural o global) es susceptible de dar lugar a una propiedad contextual al aplicarse a cualquier miembro del colectivo.

Puede partirse, por ejemplo, de una característica absoluta de los miembros (como la condición de alfabetismo) para generar una propiedad analítica (tasa de analfabetismo) que se aplique a cada uno de los miembros como propiedad contextual, sin que el significado de ésta última se confunda con la propiedad absoluta original.

Figura 2.6: Generación de una propiedad contextual-analítica



Es claro que saber leer y escribir y pertenecer a una Provincia con más de un 10% de analfabetismo son propiedades perfectamente distinguibles. Obviamente, las propiedades contextuales sólo pueden funcionar como variables cuando se analizan UUAA pertenecientes a varios colectivos: si todas las UUAA fueran miembros del mismo colectivo, se trataría de una propiedad constante en la que todas presentarían el mismo valor. De modo análogo, tampoco las propiedades comparativas tendrán sentido si pertenecen todas al mismo colectivo: de ser el caso la propiedad comparativa queda reducida a la propiedad absoluta o relacional en la que se origina.

Algunos problemas en la aplicación de la tipología

El valor de esta tipología es esencialmente heurístico: rigurosamente considerado, el conjunto de los tipos de propiedades no constituye un sistema de categorías exhaustivas y mutuamente excluyentes.³⁹ En consecuencia, es posible que se susciten dudas al aplicarla en la operacionalización de ciertos enunciados. Por ejemplo, un mismo concepto puede llegar a ser operacionalizado mediante indicadores que remiten a propiedades de diferente tipo. Si, por caso, intentamos operacionalizar el concepto de ‘industrialización’, para poder medir el “grado de industrialización” de un país (o de una región, o de una provincia), podríamos optar por varias alternativas. Así, si tomamos como indicador el “porcentaje de la PEA ocupado en la industria”, tendremos una propiedad analítica. Mientras que si elegimos el “porcentaje del PBI generado por la industria”, se tratará de una propiedad global.⁴⁰

También hay proposiciones que se pueden operacionalizar de modos alternativos: este es el caso cuando los miembros de un conjunto de colectivos son tomados como UUAA en una hipótesis que incluye una propiedad contextual. Si se dice:

³⁸ Para restablecer el significado implícito de la hipótesis se podría agregar al enunciado: “[que los pobladores de las áreas urbanas no marginales]”.

³⁹ Aún cuando es un ejercicio de los más útiles intentar aplicarla como si lo fuera.

⁴⁰ Por más que se trate de un porcentaje, el PBI industrial no se obtiene a partir de información sobre los miembros-habitantes del país.

«Los ingresos de los pobladores de barrios marginales de Resistencia son más bajos [que los ingresos de los pobladores de barrios no marginales de Resistencia]».

Parece evidente que hay aquí:

1. Un conjunto de UUAA:

⇒ individuos “**pobladores de Resistencia**”;

2. Y dos variables:

⇒ 2.1. “*Ingreso mensual*” (que remite a una propiedad ABSOLUTA);

⇒ 2.2. “*Tipo de barrio*” (que alude a una propiedad CONTEXTUAL).

Sin embargo, también podría decirse:

« En los barrios marginales de Resistencia el promedio de ingresos es menor [que en los barrios no marginales]».

Y, en este caso, habría que reconocer la existencia de:

1. Un conjunto de UUAA colectivas:

⇒ “**barrios de Resistencia**”;

2. Y dos variables:

⇒ 2.1. “*Condición de marginalidad*” (una propiedad GLOBAL);

⇒ 2.2. “*Promedio de ingresos*” (una propiedad ANALÍTICA).

Las dos interpretaciones son, por supuesto, igualmente legítimas. El principal mérito de la tipología de Lazarsfeld y Menzel es arrojar luz sobre la multiplicidad de operaciones que requiere la construcción de los datos, además de aclarar los mecanismos obrantes en determinadas falacias.

En nuestra última hipótesis, consideramos una UA colectiva “el barrio”, y definimos a la condición de marginalidad como remitiendo a una propiedad global. Sin embargo, esta caracterización merece ser discutida. En efecto, la condición de “marginal” que se le pueda otorgar a un barrio puede basarse en varias características.⁴¹ Podríamos definir como ‘marginal’ a un barrio habitado por una alta proporción de “marginales”; y se trataría entonces de una propiedad analítica.⁴² Pero también se puede sostener que la marginalidad es una característica global del barrio, con lo que continuaríamos tratando con un enunciado de dos niveles: el barrio-colectivo, y el individuo-miembro. La cuestión es: ¿ En qué se basa esta definición de la “marginalidad” como propiedad global? Ateniéndonos a la etimología de esta noción, a su sentido ecológico inicial, debemos considerar ‘marginales’ los barrios localizados en la periferia de las ciudades. Vale decir barrios que se definen por mantener una determinada relación espacial con respecto a los otros barrios de la ciudad, por una ubicación específica dentro de la estructura urbana. Ahora bien, no hay modo de incluir este tipo de definición dentro de la tipología presentada, a no ser que consideremos que se trata de una propiedad relacional, para lo cual será necesario definir la pertinencia de tres niveles de inclusión en la hipótesis. Nuestras UUAA -los barrios- deberán ahora ser consideradas como miembros integrantes del colectivo ciudad, cuando al mismo tiempo podrán continuar funcionando como colectivos con relación a sus habitantes para poder definir al

⁴¹ De acuerdo a Gino Germani, «En América Latina el término ‘marginalidad’ empezó a usarse principalmente con referencia a características ecológicas urbanas, es decir a los sectores de población segregados en áreas no incorporadas al sistema de servicios urbanos en viviendas improvisadas y sobre terrenos ocupados ilegalmente. De aquí el término se extendió a las condiciones de trabajo y al nivel de vida de este sector de la población» (1973: 12-13).

⁴² La nota anterior debería bastar para eximirnos de defendernos de cualquier acusación de circularidad en esta definición.

“promedio de ingresos” como una propiedad analítica; se trata por lo tanto de un caso en que la UA se encuentra en una posición intermedia, funcionando como colectivo respecto a un nivel inferior y como miembro respecto a un nivel de inclusión superior.

En el trabajo original de Lazarsfeld y Menzel, hay dos ejemplos de proposiciones en tres niveles. Uno de ellos está tomado de *Union Democracy*, el célebre estudio de Lipset, Trown y Coleman sobre los procesos políticos en el Sindicato Internacional de Tipógrafos⁴³ y no presenta mayores dificultades en cuanto a su estructura puramente lineal. El otro ejemplo, en cambio, ofrece mayor interés:

«Los clubes de mujeres que están divididos internamente en camarillas tienen relaciones más difíciles con otros clubes de mujeres, que aquellos que no están divididos de esta manera. Aquí los elementos⁴⁴ (“clubes de mujeres”) primero son categorizados de acuerdo con una variable estructural (división interna en camarillas), y luego se afirma algo acerca de una propiedad relacional (relaciones con los otros clubes) de cada una de las categorías estructurales» (101).

Lo que sorprende a primera vista en este ejemplo es la aparente inversión entre los niveles y las propiedades que se les aplican. Si los clubes son colectivos tendrán una estructura basada en las relaciones entre sus miembros; pero para entablar relaciones con otros clubes, habrá que considerarlos como individuos (puesto que las propiedades relacionales lo son de miembros). La propiedad estructural pertenece al colectivo, pero la propiedad relacional caracteriza también a este mismo colectivo (y no a sus miembros, como podría esperarse). Implícitamente, se requiere suponer la existencia de un tercer nivel,⁴⁵ cuya definición sólo podría producirse en términos lógicos: el conjunto de todos los Clubes de mujeres. Pero, como los autores evidencian una saludable preocupación por la sustantividad de sus reflexiones metodológicas, son renuentes a definir ese tercer nivel como un colectivo.

Ha sido señalada hace tiempo una ambigüedad en el planteo de Lazarsfeld y Menzel. Así, F. Korn escribía: «No queda claro si un colectivo es cualquier conjunto lógico...o es sólo un conjunto tal que posea la propiedad de poder ser tratado como un “actor social”» (1969:15). Para obviar este problema, F. Korn proponía que sólo se consideren como colectivos “los conjuntos que posean por lo menos una propiedad “global”, es decir, que puedan ser definidos por alguna propiedad ajena a las propiedades de sus miembros” (1969:15). Esta sería sin duda una solución, pero que no resolvería la pequeña duda que se genera en el ejemplo del Club de mujeres. Es cierto que cada uno de estos clubes puede funcionar como un “actor” y, por lo tanto, **relacionarse** con otros actores -entre los cuales habrán otros clubes femeninos. Como se ha visto, Lazarsfeld y Menzel plantean como un requisito para predicar propiedades relacionales que se trate de relaciones **sustantivas**. Pero, en cambio, no parece posible que el colectivo “conjunto de todos los clubes de mujeres” posea una estructura en algún grado significativo; parece más bien tratarse de una simple “categoría”.⁴⁶ En el ejemplo se encuentra implícita la solución al problema: es la posibilidad de que existan propiedades relacionales de los miembros, sin que se produzcan correlativamente propiedades estructurales a nivel del colectivo, y por lo tanto de que éste sea definible a partir de criterios exclusivamente lógicos.

Comparando los dos ejemplos, el nuestro y el de Lazarsfeld y Menzel, se advierte que el problema que presentan es exactamente inverso; mientras los barrios forman parte de un colectivo sociológicamente existente (una ciudad) pero no mantienen entre ellos más que relaciones espaciales (y no “sustantivas”), por su parte los Clubes pueden relacionarse entre sí como “actores”

⁴³ Hay tipógrafos (miembros) que votan a favor o en contra de su predisposición, y cuyo voto es comparado con la dirección mayoritaria del voto en su taller (colectivo-miembro), y en el local sindical (colectivo) al que pertenece el taller. Se tiene entonces una propiedad absoluta del miembro, y dos propiedades contextuales correspondientes a los colectivos de menor y mayor nivel. Se determinó así que « mayor cantidad de gente votaba en forma contraria a la propia predisposición cuando ésta estaba en conflicto con la mayoría de su taller (pero no de su local) que cuando estaba en conflicto con la mayoría de su local (pero no de su taller).» (Lazarsfeld y Menzel, 1969: 101).

⁴⁴ Es decir, las UUAA, en la terminología que utilizamos aquí.

⁴⁵ No tan “implícitamente”, ya que inmediatamente antes, se nos ha dicho que se van a abordar ahora “ejemplos de proposiciones de tres niveles” (1969: 100).

⁴⁶ Al menos no se deduce del ejemplo otra posibilidad: no se menciona ni siquiera una hipotética -aunque por cierto muy plausible- Federación de Clubes de mujeres.

sin por ello dar lugar a un colectivo cuya existencia sea algo más que lógica. Hay una manera simple de llevar remedio a esta confusión entre criterios sustantivos y lógicos, y es producir una distinción estricta entre ambos. Así, desde el punto de vista de la tipología de propiedades individuales y colectivas, debería bastar con atender a criterios puramente lógicos, sin preocuparse ni por la “existencia sociológica” del colectivo ni por el carácter “sustantivo” de las relaciones entre los miembros. Una cosa es la existencia de una posibilidad lógica; otra muy distinta es que resulte pertinente hacer uso de esa posibilidad, lo que debería resolverse exclusivamente sobre una base pragmática.

CAPÍTULO 3

NOCIONES DE MUESTREO

Cualquier investigación empírica versará sobre sistemas reales situados en el espacio y en el tiempo. El estudio de un objeto determinado es algo que puede demandar la definición de uno o varios universos. El fin que se persiga puede ser eminentemente práctico -describir qué piensan los obreros de una fábrica, o cuáles son las intenciones de voto de los habitantes de una provincia, u obtener una estimación de la tasa de desempleo en una ciudad, etc.-, o bien responder a una motivación más estrictamente científica -como cuando se trata de fundar hipótesis de un mayor grado de generalidad.

La generalidad es algo que se puede entender por lo menos en dos sentidos. Zetterberg distingue entre proposiciones teóricas y comunes basándose en cuál sea su “valor informativo” (1968: 66-67). Si se dice «A mayor educación, mayor ingreso monetario», ésta puede ser una hipótesis de un nivel de generalidad muy elevado; en el sentido de que, potencialmente, es aplicable a cualquier sociedad en la que se reconozca la existencia de algo llamado “educación” y en la que funcione alguna forma de moneda para regular el acceso de sus miembros a bienes y servicios. Sin embargo, no es ésta una hipótesis de alto valor informativo en cuanto a la variedad de fenómenos de los que da cuenta.

El problema del alcance de una hipótesis es entonces doble: por una parte, tiene que ver con sus referentes concretos -con la cantidad de sistemas o UUAA comprendidos dentro de su dominio-; por la otra, con el grado de abstracción de sus conceptos de propiedad respecto a conceptos más específicos -con la cantidad, por lo tanto, de conceptos de propiedad que subsumen. Así, cuando se dice «Toda persona tiende a realizar aquellas acciones que ayudan a mantener invariables las evaluaciones que recibe de sus iguales» se trata de una proposición de un alto nivel de generalidad en este segundo sentido. En efecto, como lo señala Zetterberg, de ésta pueden deducirse varias proposiciones más específicas.; por ejemplo, los políticos intentarán mantener su popularidad, los investigadores elegirán temas que les permitan conservar su prestigio como miembros de una comunidad científica, etc. Un concepto tan abstracto como el de “evaluaciones” estará subsumiendo varios conceptos de menor nivel (“aprobación”, “estima” y “rango”); por lo que la mera sustitución de dicho concepto en la proposición más abstracta, permitirá generar toda una serie de hipótesis específicas.

Si se admite que el desarrollo del conocimiento científico supone tender hacia la formulación de hipótesis “generales abstractas”, se entenderá que tanto el grado de generalidad como el de “abstracción” son relevantes en cuanto a determinar el alcance de una hipótesis cualquiera. Así como era posible partiendo de la idea de ‘niveles de inclusión’ determinar superunidades (“colectivos”), así también podría aplicarse la idea de inclusión a los conceptos de propiedad, obteniendo superconceptos.

El problema clásico de la inducción deviene de la imposibilidad lógica de sustentar la verdad de hipótesis generales a partir de un número de observaciones limitado y siempre finito. Ahora bien esta limitación del “número de observaciones” se refiere tanto a las UUAA como a las variables: ni estudiamos todos los elementos de un conjunto de UUAA, ni de todas las UUAA observadas consideramos todas las propiedades. Desde este punto de vista la técnica del muestreo podría considerarse como una alternativa estadística para superar el problema de la inducción. Empero esta técnica es más apta para resolver el problema en cuanto al número de las UUAA. Las razones por las que no se puede aplicar con el mismo éxito a las variables son varias; mencionemos simplemente la que deviene de la dificultad en delimitar un universo de variables. Por ende, nos limitaremos a plantear el problema del muestreo en lo referido al problema de la selección de un cierto número de UUAA. Abordaremos luego la cuestión del tamaño de la muestra, y finalmente

describiremos los diferentes tipos de muestras que se utilizan más corrientemente.

1. ELEMENTOS BÁSICOS DE MUESTREO

Ya hemos introducido el concepto de “universo” para referirnos a un conjunto de UUAAs definidas por presentar valores constantes en un conjunto de variables. Indiferentemente puede utilizarse este término o el de *población* para denotar “el conjunto de todos los casos que concuerdan con una serie determinada de especificaciones” (Selltiz *et al.*; 1968: 560). Así, cuando se habla comúnmente de “la población de Argentina”, se está haciendo referencia a un conjunto de UUAAs - “personas” - que en un momento dado (por ejemplo, el 30 de septiembre de 1980), en la variable “país de residencia” presentan el valor “Argentina”. En la jerga estadística, una población se compone de “elementos”: así, cada persona habitante de la Argentina es un *elemento* de dicha población. Contrariamente al uso corriente del término ‘población’, los elementos no tienen por qué ser personas. Podrá tratarse también de grupos, acciones, productos o cualquier tipo de objetos o sistemas que se convenga en definir como tales: cada uno de los avisos transmitidos por un canal de TV durante una semana, de los informes incluidos en un archivo, de los desplazamientos turísticos realizados por los habitantes de una región durante un período dado, de los relaciones existentes entre los miembros de una comunidad, etc., pueden ser considerados elementos de sus correspondientes poblaciones. En cualquier caso una población incluye *todos* los elementos que satisfacen las propiedades que la definen: todos los obreros de una fábrica, todos los departamentos o partidos en que se dividen las provincias argentinas, todos los municipios de Misiones, todas las noticias sobre suicidios de jubilados publicadas en diarios de Buenos Aires durante 1992, etc.

Desde este punto de vista uno de los primeros pasos en cualquier investigación empírica consiste en la determinación del universo de la investigación. Así, por ejemplo, en un trabajo realizado para la Entidad Binacional Yacyretá,¹ el universo del estudio fue definido como constituido por unidades domésticas que presentaran las siguientes características: pertenecer a la ciudad de Posadas, estar ubicadas por debajo de la cota 85, dentro de la Zona 1 de relocalización de la EBY, y cuyos miembros no fueran propietarios del terreno en que se asentaban.

Así como distinguíamos sub-universos dentro de un universo, así es posible definir poblaciones que se encuentran incluidas dentro de poblaciones mayores. En el contexto de la teoría del muestreo tales sub-poblaciones reciben el nombre de “estratos de la población”, o simplemente “estratos”. Un *estrato* se define a partir de una o más variables que permiten dividir a una población en sub-conjuntos mutuamente excluyentes. Dentro de la población de los electores de Posadas la variable “sexo” permitirá definir el estrato de los varones y el de las mujeres; dentro del universo de los establecimientos industriales de la Argentina, se puede distinguir estratos de acuerdo a la variable “número de empleados”: los que ocupan hasta 5 empleados, de 6 a 50 empleados, y de más de 50 empleados, por ejemplo. Combinando las variables “sexo” y “condición de alfabetismo”, la población de los votantes de Posadas el 14 de mayo de 1989, podría dividirse en cuatro estratos: varones alfabetos, varones analfabetos, mujeres alfabetas y mujeres analfabetas, etc.

Supongamos que deseemos conocer qué proporción de una población presenta un determinado valor en una variable dada. Si la población es pequeña, ello no ofrece mayores problemas. Por ejemplo, para conocer el porcentaje de mujeres entre los alumnos presentes en un aula, se puede rápidamente contar cuántos alumnos hay y cuántos son de sexo femenino. Se denomina **censo** a este procedimiento, que consiste en “un recuento de todos los elementos en una población y/o una especificación de las distribuciones de sus características, basados en la información obtenida para cada uno de los elementos” (Selltiz *et al.*; 1968: 561).

Si por lo contrario, la población que nos interesa es de gran tamaño, por ejemplo, los habitantes de Buenos Aires en 1993, podría resultar sumamente trabajoso obtener información sobre todos sus elementos. En este caso convendrá utilizar una **muestra**, vale decir un subconjunto de elementos

¹ “Estudio sobre los medios de subsistencia y la capacidad de pago de la población no-propietaria de Posadas a relocalizar por Yacyretá” (Informe inédito producido por la FHCS-UNaM).

de una población seleccionados para averiguar algo sobre el conjunto total de elementos que constituye esa población.² Cuando se utiliza una muestra, se lo hace en base a la creencia de que la exactitud de los resultados obtenidos de ese modo es lo más próxima posible a la que se hubiera obtenido realizando un censo desprovisto de todo tipo de error sobre la población. Por supuesto, tal creencia constituye una hipótesis cuya verdad sólo sería demostrable realizando paralelamente un censo sobre la misma población.³

Por una parte existen los *parámetros* poblacionales, es decir los valores “verdaderos” que caracterizan las distribuciones de variables en la población: por ejemplo, la media de ingresos de todos los jefes de unidades domésticas de Buenos Aires, o el porcentaje de todos los electores con intenciones de votar por el Justicialismo en las próximas elecciones; dichos parámetros son los valores que se obtendrían en una medición de todos los elementos desprovista de cualquier tipo de error.

Por la otra, se obtienen *valores muestrales*, simples *estimadores* de aquellos parámetros. Lo cierto es que nunca resulta posible asegurar la coincidencia entre los valores muestrales y los parámetros; a lo sumo, la forma de selección de la muestra podrá maximizar nuestra creencia en dicha coincidencia.

El problema del muestreo consiste entonces en cómo seleccionar una muestra de modo tal de obtener la máxima aproximación a los parámetros poblacionales compatible con las restricciones de costo y de tiempo existentes. En efecto, si se dispusiera de una cantidad ilimitada de recursos, lo que cabría sería recurrir a un censo. Pero lo cierto es que las más de las veces no es ésta la situación en que se encuentra el investigador.

A menudo se habla de la “representatividad” de la muestra. La idea de representatividad tiene que ver con la posibilidad de que la muestra sea una réplica adecuada de la población en lo que hace a las variables relevantes. Al respecto, convendrá tener en cuenta las consideraciones siguientes.

a. Una muestra en particular no es representativa en abstracto; una muestra que sea representativa para determinados propósitos, puede no serlo para otros: no existe la muestra representativa en sí, para cualquier propósito.

b. Si bien existen procedimientos para evaluar la bondad de una muestra, no es la muestra en sí la que es representativa; más bien es nuestra creencia en su representatividad la que va a depender del plan de muestreo utilizado para seleccionarla. Por *plan de muestreo* se entiende el diseño de un conjunto ordenado de operaciones que conduce a la selección de un número especificado de unidades de análisis a partir de un universo determinado.

2. EL TAMAÑO DE LA MUESTRA

Sin duda, determinar el tamaño adecuado para una muestra es una elección crucial, por todas las consecuencias negativas que puede producir una equivocación. No obstante ello, no es infrecuente que el tamaño de la muestra se determine en función de los recursos disponibles. Está claro que si la muestra es demasiado grande, se habrá realizado una inversión inútil, un derroche; en cambio, si resulta exigua, ocurrirá que no servirá a los propósitos de la investigación.⁴ Lamentablemente, y contrariamente a opiniones corrientes, no existe ninguna regla áurea del tipo “tomar un 10% de la población” para asegurarse el éxito. En algunos casos 10% resultará lastimosamente insuficiente,

² Nada impide combinar ambos procedimientos. De hecho el Censo Nacional de Población en la Argentina combina ahora un formulario corto que se aplica a todos los hogares (universo), y un formulario más extenso que sólo se aplica en una muestra de los hogares.

³ Aunque, obviamente, de ser este el caso ya no sería necesario recurrir a una muestra.

⁴ Existiría la posibilidad en ese caso de aumentar el número de unidades seleccionadas, pero ello entrañaría varias consecuencias negativas: a) el carácter de la muestra se vería desnaturalizado, en lo que hace a sus propiedades estadísticas; b) los costos de retornar al campo en busca de más datos pueden ser muy gravosos; c) se puede perder la comparabilidad entre los datos recolectados en dos períodos distintos, ya sea porque las variables objeto se hayan visto modificadas por el simple paso del tiempo, o porque el mismo proceso inicial de recolección haya conducido a una alteración en sus valores.

mientras que en otros será un total exceso.

Atendiendo a una distinción de Galtung, todas las muestras pueden clasificarse en dos grandes categorías según el tipo de hipótesis que se pretende poner a prueba. En efecto puede tratarse tanto de *hipótesis de generalización* -cuando a partir de los datos muestrales se pretende inferir, por ejemplo, el valor de un parámetro en una población dada-, o bien de *hipótesis sustantivas* -cuando se desea comprobar la existencia de determinada relación entre ciertas variables. En el primer caso los criterios pertinentes a tener en cuenta serán los habituales para cualquier tipo de inferencia estadística; en el segundo, pueden resumirse en unas breves reglas muy sencillas de comprender.

2.1 Muestras para verificar hipótesis de generalización

Existen procedimientos estadísticos que nos permiten estimar la probabilidad de que un determinado valor muestral no difiera sustancialmente del parámetro que se hubiera obtenido de haber realizado un censo. Así, si se desea averiguar qué porcentaje de los elementos de una población presenta el atributo x , es posible adoptar un plan de muestreo tal que nos garantice, por ejemplo, con un *nivel de confianza* del 95% que el parámetro no se apartará más de un 2% (*margen de error*) del valor muestral que obtengamos. Si requiriéramos mayor seguridad y precisión elegiríamos un nivel de confianza mayor (99%, por ejemplo) y un margen de error menor (1%); empero, cuanto más alto sea nuestro nivel de aspiraciones, mayor deberá ser el tamaño de la muestra, lo que supondrá un costo más elevado, puesto que deberemos seleccionar y recolectar datos sobre mayor cantidad de unidades. A los efectos prácticos, un plan de muestreo con un nivel de confianza del 99% significa que, siguiendo ese plan de muestreo, 99 veces de cada 100 no nos equivocaremos en nuestra estimación; empero, no existe nunca garantía alguna de que la muestra particular que haya resultado seleccionada no pertenezca a aquel conjunto del 1% de las muestras que nos conducirían a error.

Tabla 3.1: Tamaños de muestras al azar simple requeridos para niveles de confianza del 99% y del 95%, según valores presumibles de p y q y límites de error

Límites de error (+/- %)	Nivel de confianza de 99 %				
	Valores presumibles de p y q en %				
	10/90	20/80	30/70	40/60	50/50
1	5.968	10.609	13.924	15.914	16.577
2	1.492	2.652	3.481	3.978	4.144
3	663	1.179	1.547	1.768	1.842
4	373	663	870	995	1.036
5	239	424	557	637	663
10	60	106	139	159	166
20	15	27	35	40	41
Límites de error (+/- %)	Nivel de confianza de 95 %				
	Valores presumibles de p y q en %				
	10/90	20/80	30/70	40/60	50/50
1	3.457	6.147	8.067	9.220	9.604
2	864	1.537	2.017	2.305	2.401
3	384	683	896	1.024	1.067
4	216	384	504	576	600
5	138	246	323	369	384
10	35	61	81	92	96
20	9	15	20	23	24

Fuente: elaboración propia

La Tabla 3.1 ilustra cómo el tamaño requerido para una muestra es una función directa de tres

factores.⁵

- A El *nivel de confianza* requerido: la tabla está dividida en dos mitades. En las celdas del sector superior se presentan los tamaños de muestra requeridos para un nivel de confianza del 99%; en las del sector inferior los tamaños demandados por un menor nivel: 95%. Puede observarse cómo en el sector inferior los números son menores a sus análogos del sector superior.
- B El *margen de error*: en la primera columna se indican distintos márgenes de error para cada una de las hileras de la tabla. El menor límite de error incluido es del $\pm 1\%$; vale decir que si con ese límite producimos la estimación de que un 40% de las unidades observadas en la muestra poseen el atributo, podremos inferir que la proporción en la población estará ubicada entre el 39 y el 41%. En ambos sectores de la tabla de abajo hacia arriba crecen los tamaños requeridos: cuanto menor sea el margen de error que deseemos, mayor deberá ser el tamaño de la muestra.
- C La *variabilidad del atributo* investigado: cada columna corresponde a diferentes distribuciones del atributo en el universo, que van desde la distribución 10/90 (10% poseen el atributo, 90% no lo poseen) hasta 50/50 (una mitad posee el atributo, la otra no). En cada sector los tamaños aumentan a medida que nos desplazamos de izquierda a derecha: cuanto mayor es la variabilidad de la característica,⁶ mayor es el tamaño de muestra necesario. Así por ejemplo, mientras que para un nivel de confianza del 95% y un margen de error de $\pm 2\%$, 864 casos serían suficientes si el atributo de distribuyera en las proporciones 10/90, en el caso de que la distribución en el universo fuera 50/50 la muestra requerida sería de 2401 unidades.⁷ Este último factor supone la necesidad de anticiparse a la distribución del atributo en el universo; y de no existir elementos que permitan sustentar otra alternativa, convendrá conservadoramente suponer la distribución más desfavorable: 50/50.

En síntesis, cuanto mayores sean nuestras exigencias respecto al grado de confiabilidad y de precisión de nuestros resultados, mayor habrá de ser el tamaño de la muestra. La tabla permite también apreciar cómo juegan aquí rendimientos decrecientes. Para un atributo cuya distribución en el universo es de 20/80, 246 casos alcanzan para determinar el valor del parámetro con un nivel de 95% y un margen de error de $\pm 5\%$; si se lleva la muestra a un número de 683 unidades, el margen de error se reduce en dos puntos, pasando a ser de $\pm 3\%$. Pero, para obtener una nueva reducción de dos puntos y alcanzar un margen de $\pm 1\%$, el aumento en el tamaño deberá ser mucho mayor, requiriéndose ya 6147 casos.⁸

2.2. Muestras para verificar hipótesis sustantivas

Si el objetivo de nuestra investigación es primordialmente analítico antes que descriptivo, no estaremos tan interesados en la generalización como en la simple comprobación de la existencia de relaciones específicas entre las variables. Desde este punto de vista, el tamaño no se basará en los

⁵Para estimar el tamaño de una muestra cualquiera se podrá recurrir a la fórmula: $n = (pq)t^2/E^2$.

⁶La variabilidad máxima es 50/50, vale decir cuando $p = q$ (siendo p la proporción de los que poseen el atributo, y $q = 1 - p$). Intuitivamente, se comprende que si una población es absolutamente homogénea en cuanto a una característica dada, una muestra de uno cualquiera de los elementos de dicha población bastará para producir una estimación correcta del parámetro; en cambio si la población es heterogénea -se divide por mitades en varones y mujeres, por ejemplo- es claro que se necesitará un número mayor de observaciones para producir una estimación ajustada de la distribución de la característica en el universo.

⁷Por ende, una muestra cuyo tamaño sea suficiente para estimar un determinado parámetro, podrá no serlo para otro parámetro, si la distribución de este último es menos favorable.

⁸Los tamaños indicados en la tabla son para poblaciones infinitas. A los efectos del muestreo, si un universo está integrado por 100.000 o más elementos, se lo considera infinito. En el caso de universos finitos se aplica un factor de corrección que arrojará tamaños de muestra proporcionalmente menores cuanto más pequeño sea el universo (Henry, 1990, 120).

requerimientos demandados para producir inferencias estadísticas, sino que deberá ser tal que nos permita determinar la existencia de relaciones para el conjunto de las unidades incluidas en la muestra.

Si por ejemplo queremos estudiar la relación entre dos variables dicotómicas, ello demandará producir una tabla de contingencia en la que las unidades se clasifiquen simultáneamente en ambas variables (cf. *infra*: capítulo 4). Una tabla tal consta de cuatro celdas, y para aplicar una simple diferencia porcentual difícilmente nos contentemos con un promedio de casos por celda inferior a 10. Vale decir que *con menos de 40 casos no es posible analizar ninguna relación entre variables*; incluso, para que la tendencia sea clara, preferiremos sin duda contar con un promedio de 20 casos por celda: aún así el simple cambio de un caso de una columna o de una hilera a otra producirá una diferencia porcentual del 2,5%.

En esta perspectiva, las preguntas a formularse para determinar el tamaño de la muestra tienen que ver con el número de variables que se pretende analizar simultáneamente y con el número de valores de cada una de esas variables. Galtung presenta la Tabla 3.2.

Tabla 3.2: Número mínimo de unidades de análisis para un promedio de 10 casos por celda (20 casos entre paréntesis)

Nº de variables por tabla	Nº de valores por variable		
	2	3	4
2	40 (80)	90 (180)	160 (320)
3	80 (160)	270 (540)	640 (1280)
4	160 (320)	810 (1620)	2560 (5120)

Así, por ejemplo, para estudiar la relación entre tres variables tricotómicas se requerirá un mínimo de 270 casos para contar con un promedio de 10 observaciones por celda (o 540, si pretendemos tener 20 casos en cada combinación). Obviamente no es indispensable que todas las variables tengan el mismo número de valores, ni que deban limitarse a cuatro valores, lo que puede dar lugar a otros tamaños de muestra. Lo más práctico es entonces multiplicar el número de casos por celda por el número de valores de cada una de las variables que se desee analizar simultáneamente; si se quisiera investigar como se relacionan entre sí dos variables dicotómicas y una tercera de cinco valores, el número mínimo de casos requerido para contar con un promedio de 10 observaciones por celda sería: $(10) \times 2 \times 2 \times 5 = 200$. Por supuesto, en cualquier investigación es posible pretender analizar la relación entre varios conjuntos de variables, por lo que la pregunta que debe responderse para determinar el tamaño de la muestra es cuál es la tabla más grande que se pretende generar.

En las muestras para hipótesis sustantivas es posible de todos modos aplicar determinados test de hipótesis como la t de Student aplicada a la diferencia de medias o de proporciones, o la prueba de χ^2 . En estos casos, puede interpretarse que la hipótesis sustantiva pasa a funcionar también como una hipótesis de generalización, aunque no ya referida a una población finita y concretamente determinada, sino con relación a un universo teórico, sin límites definidos, que abarque las unidades de todo tiempo y lugar a las que sea posible aplicar las variables.

3. PRINCIPALES TIPOS DE MUESTRAS

No es nuestro propósito reseñar aquí todos los tipos posibles de muestras. Tampoco nos proponemos abordar este tema en una perspectiva estadística; ante cualquier investigación de

envergadura, convendrá recurrir a algún especialista en la cuestión.⁹ Nuestra intención es sólo describir las características de algunos diseños muestrales desde un punto de vista conceptual, para exponer algunos problemas típicos que se plantean con relación al uso de esta técnica. Desde el punto de vista de la teoría del muestreo, existen solamente dos grandes categorías de muestras. Por un lado están las muestras probabilísticas - o “al azar”-, que se caracterizan porque permiten especificar la probabilidad que tiene cada elemento de la población de resultar incluido en la muestra. El caso más sencillo es cuando todos los elementos tienen la misma probabilidad de integrar la muestra, pero no es éste el único tipo de muestreo de probabilidad. Por otra parte, hay muestras no-probabilísticas, en las no existe forma de determinar la probabilidad que tiene cada elemento de ser seleccionado, y ni siquiera la seguridad de que cada elemento tenga alguna probabilidad de ser seleccionado.

Esta distinción es fundamental, puesto que solamente los planes de muestreo probabilísticos pueden ser considerados representativos en un sentido estadístico. El muestreo probabilístico es el único que permite basarse en la teoría de las probabilidades para estimar estadísticamente el grado en que los valores muestrales pueden tender a diferir de los parámetros; una muestra probabilística permite especificar el tamaño de muestra requerido para un nivel de confiabilidad y un margen de error determinados.

3.1 Muestras no-probabilísticas

En el muestreo de no-probabilidad no hay modo alguno de evaluar estadísticamente los resultados obtenidos a partir de la muestra. No obstante ello, algunas muestras de este tipo se utilizan mucho, en razón de las ventajas comparativas que exhiben en cuanto a su comodidad y bajo costo.¹⁰

Muestra accidental (o “casual”)

Consisten simplemente en tomar los casos que “caen bajo la mano”, continuando el proceso hasta que se alcanza un cierto tamaño de la muestra. Por ejemplo, se incluirán en la muestra las primeras 100 personas que pasen por una esquina y que sean deseadas de ser entrevistadas. Si el universo es la población de una ciudad, es claro que no será ésta una muestra representativa: no todos los elementos de dicha población tienen la misma oportunidad de pasar por dicha esquina, ni menos aún coincidirán en la probabilidad de pasar por allí en el momento que se realicen las entrevistas. Pero en este caso, no existe modo alguno de evaluar los sesgos introducidos por el modo de selección de las unidades; sólo cabrá esperar que la equivocación no sea demasiado grande.

Muestra por cuotas

En el muestreo por cuotas se busca garantizar la selección de elementos pertenecientes a los diferentes estratos que conforman la población y que dichos elementos puedan ser tenidos en cuenta en las mismas proporciones en que ocurren en esa población. En todos los casos, la técnica supone sostener alguna hipótesis acerca de cuáles son las variables relevantes, las que serán utilizadas para dividir la muestra en estratos. Si, por ejemplo, se quiere conocer las intenciones de voto de una población, y se piensa que estas variarán en función del sexo, convendrá estratificar la muestra en dos estratos: “masculino y femenino”. Así, a cada entrevistador se le dará la consigna de cumplir con una “cuota” de entrevistados pertenecientes a cada estrato: tantos varones y tantas mujeres. El muestreo por cuotas requiere anticiparse a las diferencias que puedan presentar los elementos de los distintos estratos respecto al valor que exhibirán en la variable a investigar, de

⁹ O, al menos, remitirse a los capítulos pertinentes de los manuales de Blalock (1966), Padua (1979), Galtung (1966), Sellitz (1968), o Festinger y Katz (1975), o a textos específicos como los de Kalton (1987) o Henry (1991).

¹⁰ Dada la naturaleza de este manual, nos concentramos en los tipos de muestras utilizados corrientemente con fines “cuantitativos”. En una perspectiva “cualitativa” -aunque también en la fase exploratoria de investigaciones que versen sobre un gran número de casos- cabe destacar la pertinencia de otros procedimientos de enorme valor heurístico. Así, «el **muestreo teórico** se efectúa para descubrir categorías y sus propiedades y para sugerir las interrelaciones dentro de una teoría. El muestreo estadístico se hace para obtener evidencia exacta sobre la distribución de la población entre categorías a ser usadas en descripciones o verificaciones. De este modo, en cada tipo de investigación la “muestra adecuada” que deberíamos buscar (como investigadores y lectores de investigaciones) es muy diferente.» (Glaser y Strauss, 1967: 62)

modo tal de seleccionar una muestra que “represente” adecuadamente a la población, esto es que funcione como una réplica de la población en cuanto a las variables que se juzgan relevantes. En lo que respecta al estudio de preferencias, opiniones, actitudes, etc., se conoce que variables tales como el sexo, la edad, o el nivel educativo tienden a producir diferencias significativas en esas características.

Ahora bien, puede suceder que la proporción de elementos de un estrato en la muestra no sea idéntica a la existente en la población. Imaginemos que se encuesta a 100 personas obteniendo una cierta distribución por nivel educativo, la que podemos comparar con la existente en el universo:¹¹

Tabla 3.3: Comparación de la distribución de la variable “nivel educativo” en la muestra y en la población

	Nivel educativo			Total
	Bajo	Medio	Alto	
% Muestra	25	45	30	100
% población	55	30	15	100
Diferencia	-30	15	15	

La comparación muestra claramente que en la muestra los niveles alto y medio se encuentran sobrerrepresentados, mientras que el nivel bajo está subrepresentado con relación a la población.

Tabla 3.4 : Distribución en la muestra de la intención de voto según el nivel educativo

Intención de voto	Nivel educativo			Total	
	Bajo	Medio	Alto		
Justicialista	14	21	10	45	45%
No justicialista	11	24	20	55	55%
Diferencia	25	45	30	100	100%

Ahora bien, supongamos que se ha preguntado a los entrevistados por su intención de voto en las próximas elecciones para estimar qué porcentaje de votos obtendrá el Justicialismo, obteniendo estos resultados. Si se los toma sin tener en cuenta el peso distinto que tienen los diferentes niveles educativos en la muestra y en la población se concluye que hay un 45% de los entrevistados que piensan votar por el Justicialismo.

Pero puesto que el nivel bajo está subrepresentado en la muestra, y que la proporción de votos por el Justicialismo es sustancialmente mayor en ese estrato, es evidente que el porcentaje de votos con esa orientación debe ser mayor en la población. Si se conoce la distribución de la variable “nivel educativo” en la población,¹² es posible calcular cuál hubiera sido el porcentaje de votos total por el Justicialismo si los distintos estratos hubieran estado representados en la muestra en las proporciones correctas.

Para ello se recurre a simples reglas de tres: las frecuencias de la columna “bajo” se multiplican por 55/25, las de “medio” por 30/45, y las de “alto” por 15/30:

¹¹ Afortunadamente, se dispone con frecuencia de datos sobre la población a muestrear provenientes de otras fuentes. En este caso, los datos son imaginarios; pero si se tratara de muestrear la población de Posadas, por ejemplo, se podría recurrir fácilmente a los resultados de la Encuesta de Hogares para conocer la distribución de la variable en el universo.

¹² Por ejemplo, se dispone de datos censales, o de la última onda de la Encuesta Permanente de Hogares realizada en la ciudad.

Tabla 3.5 : Distribución corregida de la intención de voto según nivel educativo

Intención de voto	Nivel educativo			Total	
	Bajo	Medio	Alto		
Justicialista	31	14	5	50	50%
No justicialista	24	16	10	50	50%
Diferencia	55	30	15	100	100%

El resultado es ahora muy distinto: el marginal inferior exhibe la misma distribución de los distintos niveles educativos que se da en la población, y el porcentaje resultante de votos por el Justicialismo se eleva al 50%.

Lo que se ha hecho mediante estas sencillas operaciones aritméticas es restablecer el “peso” de los distintos estratos en la muestra, *ponderando* los resultados de acuerdo a la distribución conocida de la variable en la población. Por lo tanto, no es una condición necesaria para el muestreo por cuotas que los distintos estratos se encuentren representados en la muestra en las mismas proporciones que en la población, sino que bastará con que se cuente con un número suficiente de casos en cada estrato como para poder estimar los valores para los estratos de la población. Se denominan *autoponderadas* aquellas muestras en las cuales los estratos se encuentran representados en idénticas proporciones que en la población.

Las muestras por cuotas son mucho más eficaces que las muestras accidentales simples. De allí que su uso haya sido muy frecuente en sondeos electorales así como en la investigación de mercado en general. Sin embargo, el muestreo por cuotas sigue siendo en esencia un procedimiento accidental en el que cada estrato de la muestra es una muestra accidental del correspondiente estrato de la población.

En particular, la libertad otorgada a los entrevistadores puede dar lugar a *muestras sesgadas*. Estos son proclives a encuestar a las personas que tengan más a mano -por ejemplo, a amigos o conocidos-, a descartar a las personas que no se encuentren en su domicilio cuando los visitan, a evitar realizar entrevistas en barrios apartados, etc. Ello puede redundar en un cierto sesgo (*bias*) de la muestra, en una desviación con respecto a determinadas características reales de la población que la torne poco representativa para un objetivo determinado. Por cierto, ello dependerá del grado en que estas desviaciones se encuentren asociadas a variables relevantes con respecto al propósito de la investigación. Así, puede darse que estén subrepresentadas en la muestra las personas que habitan en barrios apartados y que este efecto carezca de importancia en sí, pero que se produzca por esta vía una subrepresentación de las personas de bajo nivel económico-social debido a que éstas tienden a residir en barrios alejados del centro urbano, por ejemplo. Si se pudiera garantizar que no se ha producido ningún efecto de esta naturaleza en la selección de las unidades, el problema estaría resuelto. Pero como se trata de una muestra accidental no existe ningún fundamento que nos permita asumir esa hipótesis.

3.2 Muestras probabilísticas

En la investigación científica, el azar no es algo que se pueda presuponer como naturalmente dado; antes bien, se busca garantizar el azar, lo cual requiere recurrir a procedimientos con frecuencia caros y laboriosos. El azar científico se persigue de manera sistemática, es algo que se construye. Así, en las muestras probabilísticas -o “aleatorias”- se requerirá que todos los elementos de una población tengan una probabilidad *conocida* de ser seleccionados.

Muestra al azar simple

Es el ejemplo más sencillo de un procedimiento de muestreo probabilístico; pero aún más que por su posibilidad de ser aplicado directamente, su interés consiste en que se encuentra a la base de todos los otros tipos de muestras aleatorias. En la muestra simple al azar, todos los elementos de la población tienen la *misma* probabilidad de resultar seleccionados, y todas las combinaciones de

elementos de para un tamaño dado de la muestra presentan también la misma probabilidad de selección. Así, si nuestro universo está constituido por 100 elementos, cada uno de éstos tendría la misma probabilidad de ser incluido en la muestra: $p = 1/100$.

La probabilidad de una muestra de un tamaño dado de ser extraída, por su parte es igual a:

$$\frac{1}{\frac{N!}{n! (N-n)!}}$$

en la que: N = tamaño de la población
n = tamaño de la muestra
! = factorial

Así, para cualquier muestra de tamaño 10, extraída de una población de 100 elementos, la probabilidad de ser seleccionada será (Padua; 1979: 66):

$$p = \frac{1}{\frac{100!}{10! (100-10)!}} = \frac{1}{124\ 303\ 388\ 140}$$

El requisito imprescindible para seleccionar una muestra al azar simple es disponer de un listado de todos los elementos que componen la población. A cada una de las unidades se le asigna un número, desde 1 hasta N. Luego, recurriendo a una tabla de números al azar, a un bolillero, a una computadora, o a cualquier otro procedimiento que garantice la producción de una serie de números al azar, se extraen en la cantidad necesaria los números correspondientes a las unidades que integrarán la muestra. Así, si se tiene una población de $N = 1000$ y se desea seleccionar una muestra de 100 unidades, se sortearán 100 números comprendidos entre el 1 y el 1000, tomando los últimos tres dígitos de los números generados.

En la práctica, el procedimiento del muestreo al azar simple se torna por demás engorroso al trabajar con poblaciones grandes. Si además, por determinadas restricciones el número de casos en la muestra es pequeño, la muestra al azar simple puede conducir a una representación absolutamente insuficiente de los estratos de menor peso relativo en la población. Es por ello que en estas condiciones se suele optar por otros diseños muestrales libres de esas desventajas. No obstante, la importancia teórica de la muestra al azar simple es fundamental, dado que todos los otros diseños probabilísticos remiten a este modelo, y a que en todos esos otros diseños el muestreo al azar simple es utilizado al menos en alguna de sus etapas.

Muestra sistemática

Se trata de un procedimiento de muestreo que, a la vez que es susceptible de simplificar notablemente la selección de las unidades, a todos los efectos prácticos presentará las mismas virtudes que el muestreo al azar simple. Aquí también se requiere como condición indispensable contar con el listado de todas las unidades, las que deberán ser numeradas correlativamente. Conociendo el número de elementos que componen la población, y habiendo determinado el número de casos necesario para la muestra, se puede establecer la fracción de muestreo dividiendo el primer número por el segundo. Luego se selecciona un solo número al azar, que corresponderá a la primera unidad seleccionada; a este número se le suma la fracción de muestreo y se obtiene la segunda unidad; luego se suma nuevamente la fracción obteniendo la unidad siguiente, y se procede así sucesivamente hasta completar la muestra.

Supongamos que queremos extraer una muestra de 100 estudiantes de un universo constituido por una Facultad de 20.000 alumnos. Nuestra fracción de muestreo será: $20000/100 = 200$. Seleccionamos un número al azar: el 4387. Nuestra muestra quedará integrada por los estudiantes que lleven los números 4387, 4587, 4787, 4987, etc. Debe quedar claro que en este diseño ni todas las unidades ni todas las combinaciones de unidades tienen las mismas probabilidades de resultar seleccionadas. Así, en nuestro ejemplo, una muestra que incluya las unidades 4387 y 4388 tiene una probabilidad cero de ser seleccionada (dada nuestra fracción de muestreo de 200).

La condición para poder utilizar este procedimiento es que la numeración de las unidades en la lista no responda a ningún orden en particular, o por lo menos que de existir un orden éste no se base en ninguna propiedad susceptible de encontrarse relacionada con las variables que se desea estudiar. En determinadas circunstancias, sin embargo, la existencia de un orden puede resultar ventajosa, pudiendo ponerse al servicio de la investigación. Imaginemos que queríamos estudiar las opiniones de los estudiantes de aquella Facultad respecto a cuestiones relacionadas con su

organización académica. Como hipótesis, no resultaría irrazonable pensar que las opiniones podrían variar según la antigüedad de los alumnos en la institución. Ahora bien, si la numeración de la lista es tal que los números más bajos corresponden a los estudiantes más antiguos y los más altos a los más recientes, un procedimiento de selección de la muestra como el expuesto garantizaría la representación en la muestra de estudiantes con diferentes grados de antigüedad en la Facultad.

Muestra estratificada

Siempre que hay razones para pensar que determinadas características de las UUAA pueden encontrarse fuertemente relacionadas con las variables que tenemos interés en investigar, convendrá asegurar la representación en la muestra de los diferentes valores que pueden asumir las unidades en dichas características. Esto es, en base a determinadas variables que por hipótesis se asumen como pertinentes se divide la población en estratos. En el diseño más simple de muestra estratificada, la única diferencia con respecto a la muestra por cuotas es el hecho de que la selección de las unidades se realiza al azar. Por cierto, esta no es una diferencia menor, puesto que obtenemos una muestra apta para realizar cualquier tipo de inferencia estadística. Pero conceptualmente ambos tipos de muestra son muy similares: así como la muestra por cuotas consiste en una serie de muestras casuales tomadas cada una dentro de un estrato diferente, así una muestra estratificada se compone de varias muestras al azar simple seleccionadas dentro de otros tantos estratos.

Por supuesto, las mismas consideraciones que se hacían acerca de la ponderación de los estratos en la muestra por cuotas se aplican también al muestreo estratificado. Existen poderosas razones por las cuales puede resultar conveniente trabajar con un diseño de muestra que no sea autoponderado. En primer lugar, una muestra autoponderada puede llevar a seleccionar un número insuficiente de casos dentro de un estrato cuyo peso dentro de la población es pequeño. Así, por ejemplo, una muestra autoponderada estratificada por nivel económico social (NES), puede llevar a contar con un magro número de unidades dentro del estrato “Clase alta”. Supongamos que la definición de “Clase alta” sea tal que comprenda a un 5% de la población. Tomando una muestra de 400 casos, sólo se contaría con 20 elementos dentro de ese estrato; un número por cierto insuficiente para poder afirmar nada acerca de los integrantes de ese estrato. De ahí que pueda resultar interesante sobremuestrear voluntariamente ese estrato, de modo de contar, por ejemplo con 80 unidades con esa característica. Por supuesto, es factible hacer jugar esta consideración también en el muestreo por cuotas.

Pero en el caso del muestreo estratificado se agrega una razón más que hace a la eficiencia del procedimiento desde un punto de vista estadístico. En efecto, se ha visto que cuanto menor es la variabilidad de una característica en una población, menor será el número de casos necesario en la muestra para contar con una estimación adecuada del parámetro. Vale decir que para un número dado de observaciones muestrales, cuanto mayor sea la homogeneidad de la población, menor será el margen de error al estimar la distribución de la variable.

Dividir la población en estratos es útil en la medida en que los estratos sean internamente homogéneos y externamente heterogéneos en cuanto a las variables de nuestro interés. Si se cuenta con una estimación previa acerca de la variabilidad del parámetro dentro de cada estrato, entonces el modo más eficiente de muestrear¹³ es aquel en que el número de casos seleccionados sea proporcional a esa variabilidad. En la práctica, ello se resume en una consigna sencilla: sobremuestrear los estratos en que la variabilidad sea mayor, submuestrear los más homogéneos. Así, si en un sondeo sobre las intenciones de voto para las elecciones presidenciales se toman las provincias argentinas como estratos, se deberá sobremuestrear aquellas en las que el resultado aparezca como más dudoso, por ejemplo. Lo expuesto debería bastar para comprender que, excepto por un pequeño ahorro de operaciones aritméticas, nada justifica muestrear a todos los estratos en la misma proporción y, por lo contrario, puede llegar a ser muy ventajoso no hacerlo.

¹³ Vale decir, el que a igual cantidad de casos produzca mayor exactitud en la estimación del parámetro, o el que nos permita obtener al mínimo costo un resultado con un nivel de confianza y un margen de error dados.

Muestra por conglomerados y en etapas múltiples

Si se quisiera hacer una muestra al azar simple -o sistemática, o aún estratificada- de la población de Buenos Aires, se tropezaría de inmediato con un obstáculo difícilmente salvable: la inexistencia de un listado de todos los habitantes de esta ciudad.¹⁴ Por cierto, nada impediría dedicar una etapa previa del muestreo a listar todos los elementos de esa población, a no ser pedestres consideraciones de costo.

Una solución podría ser recurrir a una muestra por conglomerados (*clusters*). Se trata de conglomerados de UUAA que funcionan como unidades de muestreo definidas espacialmente. Este tipo de diseño se utiliza en general para reducir los costos de la recolección de datos. En efecto, cuando el universo de la investigación resulta geográficamente extenso, es muy ventajoso, en términos económicos, tratar de concentrar la tarea en el espacio. Se reducen de este modo los costos que suponen el traslado del entrevistador de una a otra unidad.

Idealmente, los conglomerados son unidades de un mismo tamaño. Como bien observa Blalock (1966: 441), ésta es en cierto modo una estrategia opuesta a la del muestreo estratificado, aunque en ambos casos se divide a la población en grupos. Si en la muestra estratificada se seleccionan los casos *dentro* de cada estrato, asegurando así que todos los estratos estarán representados, en este otro tipo de muestra se selecciona *entre* los conglomerados; correlativamente, así como se busca que los estratos sean lo más homogéneos posible internamente, para los conglomerados cuanto más heterogéneos sean mejor será el resultado. Dividida toda la población en conglomerados, se selecciona al azar un cierto número de éstos dentro de los cuales se entrevistará a todas las unidades de análisis.

A menudo la técnica por conglomerados se integra en diseños de *muestra en etapas múltiples*, esto es, muestras que suponen la definición de unidades de muestreo en diferentes niveles. Así, si se quisiera medir la tasa de desempleo en la población de Buenos Aires con 14 o más años de edad, se podría estratificar la ciudad en zonas, y dentro de cada zona o estrato seleccionar al azar un cierto número de manzanas; la gran ventaja es que todas estas operaciones se realizarán en la comodidad de un escritorio si se cuenta con buena cartografía. Luego para cada una de las manzanas seleccionadas, se puede proceder en campo a un rápido relevamiento del número de unidades domésticas existentes. A partir de los listados para cada manzana, es posible seleccionar al azar unidades domésticas.¹⁵ Finalmente, en cada unidad doméstica el entrevistador podrá sortear por algún procedimiento adecuado la persona a encuestar. En este diseño multietápico, las manzanas funcionarían como unidades de muestreo de 1era etapa, las UDD serían unidades de muestreo de 2da etapa, y las UUAA -personas- serían las unidades de muestreo de última etapa.

Sin duda, las muestras por conglomerados y en etapas múltiples son muy ventajosas desde el punto de vista económico, en el sentido que con igual cantidad de recursos es posible realizar mayor número de encuestas. Sin embargo, hay que tener en cuenta que con estas técnicas, contrariamente a la del muestreo estratificado, el grado de error aumenta: esto es para obtener el mismo grado de exactitud la muestra al azar simple se bastaría con un menor número de casos.

¹⁴ Obsérvese que la utilización de la guía telefónica produciría una muestra sesgada, puesto que no todos los habitantes de Buenos Aires figuran en esa guía, y que la tenencia de teléfono, lejos de comportarse como una variable aleatoria, está asociada a otras características de las personas, como su nivel económico. A pesar de ello, las encuestas por vía telefónica son de uso frecuente (cf. Lebart, 1992).

¹⁵ En la terminología del Censo de Población se habla tradicionalmente de 'hogares' en vez de unidades domésticas.

CAPÍTULO 4

TÉCNICAS ELEMENTALES DE ANÁLISIS

En la secuencia típica del proceso de investigación, el análisis de datos se inscribe dentro de las últimas etapas. Respondiendo a un esquema previamente establecido, las observaciones han sido realizadas, luego codificadas y tabuladas. El resultado es una serie de cuadros estadísticos a los que habrá que “leer” en función de los objetivos de la investigación, tratando de destacar lo esencial de la información en ellos contenida.

En este capítulo expondremos algunas de las técnicas más elementales que son de uso frecuente en la investigación social. Comenzaremos concentrándonos en la tabla de contingencia, puesto que éste es el modo en que tradicionalmente se presentan muchos datos en las investigaciones sociales. Acordaremos una atención privilegiada a la tabla de 2×2 , en la que se plantean al nivel más simple los problemas lógicos del análisis. En este contexto, habré de referirme al uso de los porcentajes, al test de χ^2 y a algunos de los coeficientes de asociación más simples. Luego abordaremos brevemente otras formas de presentación de los datos como las distribuciones multivariantes conjuntas, lo que permitirá introducir algunas nociones sobre las relaciones entre variables pertenecientes a niveles más elevados de medición.

1. LA TABLA DE CONTINGENCIA Y EL USO DE LOS PORCENTAJES

Una tabla de contingencia es el resultado del **cruce** (o tabulación simultánea) de dos o más variables. Nos ocuparemos solamente de tablas bivariadas (o ‘bivariantes’), que también reciben los nombres de ‘clasificación cruzada’ o ‘tabulación cruzada’. Esta forma de presentación de los datos es muy típica de la investigación en ciencias sociales, que se caracteriza por un uso predominante de variables (o atributos) definidas en los niveles de medición nominal y ordinal.¹ La tabla de contingencia consiste en un cierto número de celdas en las que, como resultado de un proceso de tabulación, realizado en forma manual, mecánica o electrónica,² se han volcado las frecuencias (número de casos) correspondientes a cada combinación de valores de varias variables.

Forma lógica de la tabla de 2×2

Para analizar la forma lógica de este tipo de tablas, consideraremos la estructura más sencilla, la llamada tabla de “ 2×2 ”, o sea de dos valores por dos valores, que resulta del cruce de dos dicotomías, los atributos ‘X’ e ‘Y’ en los que hemos clasificado un conjunto de unidades de análisis

¹ Por cierto, la tabla de contingencia puede también utilizarse para volcar datos provenientes de mediciones realizadas en el nivel intervalar, pero a costa de una gran pérdida de información; como veremos, existen otras técnicas mucho más precisas, matemáticamente hablando, para el análisis de tales variables.

² En la era actual de difusión masiva de las computadoras personales, es francamente desaconsejable recurrir a modos manuales de tabulación o bien a antigüedades tales como las tarjetas tipo McBee, las máquinas clasificadoras de tarjetas tipo Hollerit, etc.

Tabla 4.1: Forma lógica de la tabla de 2 x 2

		Atributo X		Total
		No	Sí	
Atributo Y	Sí	-X Y	X Y	Y
	No	-X -Y	X -Y	-Y
Total		-X	X	n

‘n’ representa el *total* de unidades de análisis incluidas en la muestra, lo que se suele denominar ‘la frecuencia de *orden cero*’. Por su parte, ‘-X’, ‘X’, ‘Y’ y ‘-Y’ son las *frecuencias marginales* o de *primer orden*; así, por ejemplo, ‘-X’ representa el número total de casos que no presenta el atributo X, independientemente de que posean o no el atributo Y.

Por último, ‘-X Y’, ‘X Y’, ‘-X -Y’ y ‘X -Y’ representan las *frecuencias condicionales*, o de *segundo orden*; de este modo, ‘-XY’ significa el número absoluto de observaciones que combinan la ausencia del atributo X con la presencia de Y. Es importante notar que:

$$\begin{aligned}
 n &= (X) + (-X) \\
 &= (Y) + (-Y) \\
 &= (-X Y) + (X Y) + (-X -Y) + (X -Y)
 \end{aligned}$$

Aplicación del modelo a un ejemplo

Este modelo puede aplicarse para cualquier población y todo tipo de variables:

Tabla 4.2: Misiones, 1980 - Número de habitantes según tipo de asentamiento y pertenencia a hogares con NBI

Tipo de asentamiento	Hogares con NBI		Total
	No	Sí	
Urbano	194.397	96.610	291.007
Rural	122.701	166.814	289.515
Total	317.098	263.424	580.522

Fuente: elaboración propia (datos de Argentina, 1984:343).

Así, en esta tabla, las unidades de análisis son personas, y los atributos ‘X’ e ‘Y’ se traducen, respectivamente, en el hecho de pertenecer o no a un hogar con Necesidades Básicas Insatisfechas (NBI), y de residir en una zona urbana o rural (es decir, ‘no-urbana’). El título ya nos informa que se trata de la *población* de Misiones en 1980; el n corresponde por tanto a 580.522 *personas*. Se verifica efectivamente que:

$$\begin{aligned}
 580.522 &= 317.098 + 263.424 \\
 &= 291.007 + 289.515 \\
 &= 194.397 + 96.610 + 122.701 + 166.814
 \end{aligned}$$

También se cumple que: $291.007 = 194.397 + 96.610$, es decir que cada marginal es igual a la suma de las frecuencias condicionales en la hilera -o la columna - correspondiente. *El primer paso de cualquier análisis es verificar si la tabla “cierra”, vale decir, si se cumplen las relaciones aritméticas que debe satisfacer cada cifra; en caso contrario es evidente que se ha producido algún*

error en la tabulación.

¿Qué puede afirmarse en base a esta Tabla 4.2? Puede decirse que en Misiones, en el año 1980, había 580.522 personas; proposición que, a pesar de su veracidad, tiene el inconveniente de no hacer uso de **toda** la información contenida en dicha tabla.

Un modo de comenzar el análisis es describiendo los marginales. Así, se puede decir que de estos 580.522 habitantes de Misiones, 291.007 residían en zonas urbanas en tanto que 289.515 lo hacían en áreas rurales. Esto ya es más interesante, aunque la misma conclusión podría haberse derivado de una distribución de frecuencias simple:

¿Para qué sirven los porcentajes?

Vemos que hay más habitantes urbanos que rurales; exactamente, los primeros superan a los segundos en 1.492 personas. ¿Ahora bien, es ésta una diferencia importante? Depende del contexto dentro del cual ubiquemos ese número. Lo sería sin duda si comparáramos la cifra con los datos del Censo de Población de 1970, en el que los rurales aventajaban a los urbanos en 132.886 habitantes.

Tabla 4.2a: Misiones 1980, Habitantes
según tipo de asentamiento

Tipo de asentamiento	Número de habitantes	%
Urbano	291.007	50,1
Rural	289.515	49,9
Total	580.522	100,1

Fuente: Tabla 4.2

Pero, considerando intrínsecamente los datos de la Tabla 4.2a, la manera de apreciar la importancia de esas 1.492 personas de diferencia es poniéndolas en relación con el total de la población provincial. Es decir, considerar el **peso relativo** de cada grupo sobre el total de población. Se observa así que la diferencia entre la población urbana y la rural es muy escasa: 50,1 a 49,9%. Hemos calculado el porcentaje de población urbana mediante la siguiente operación:

$$\frac{291.007}{580.522} \times 100 \quad \text{o, en general} \quad Y/n \times 100$$

¿Cuándo redondear los porcentajes?

En realidad, el resultado de la operación aritmética anterior arroja la cifra de 50,12850503, la que nosotros hemos redondeado a un decimal anotando 50,1.³ ¿Por qué este redondeo? El interés de los porcentajes es indicar con la mayor claridad las dimensiones relativas de dos o más números, transformando a uno de esos números, la *base*, en la cifra 100. Es indudable que:

$$291.007/580.522 = 50/100 = 50\%$$

Matemáticamente, estas son expresiones equivalentes —o casi— pero es evidente que, en un sentido psicológico, ‘50%’ es la manera más concisa, sencilla y ventajosa de denotar la relación que nos interesa. Si se conservan muchos decimales, sólo se logra tornar más engorrosa la lectura de la tabla y se pierde la ventaja de expresar las cifras en porcentajes. Por ende se puede recomendar, siempre que sea ello posible, *como regla general, prescindir totalmente de los*

³ El neologismo ‘redondear’ significa suprimir los decimales -números a la derecha de la coma-, o conservar una limitada cantidad de éstos. El redondeo se realiza observando el decimal siguiente al que se quiere conservar; en el ejemplo, el segundo decimal es un 2 -cifra comprendida entre 0 y 4- por lo que corresponde anotar ‘50,1’, mientras que, de tratarse de un número igual o superior a 5, se anotaría ‘50,2’.

decimales. ‘50,12850503’ parece más preciso que ‘50 %’; más es ésta una precisión engañosa,⁴ y que a todos los efectos prácticos o teóricos carece absolutamente de significado.⁵

Sin embargo, en la Tabla 4.2a hemos consignado ‘50,1 %’ y no ‘50 %’. ¿Por qué ya esta primera infracción a la regla que acabamos de formular? En el caso que nos ocupa, existirían al menos dos posibles justificaciones: a) trabajando con una población de 580.522 personas, cada punto del primer decimal representa 580 individuos, una cantidad relevante para muchos propósitos; y b) porque si no conserváramos el primer decimal, obtendríamos el mismo porcentaje para ambos sectores de la población, y tal vez no nos interese producir este efecto.⁶

El otro marginal de la Tabla 4.2 podría dar lugar a un análisis en un todo análogo. Se concluiría así que un 45,4 % —o un 45 %— de la población total de Misiones vivía en 1980 en hogares con NBI.

¿Cómo se lee una tabla de contingencia?

Siempre que se considera una tabla de contingencia es recomendable comenzar el análisis por las distribuciones univariadas de los marginales, para luego pasar al examen de las frecuencias condicionales, que nos permitirá aprehender el sentido peculiar de cada cruce de variables.

En un paso ulterior podríamos entonces hacer una lectura de cada una de las cifras contenidas en las celdas de la Tabla 4.2:

- ① 194.397 personas vivían en hogares sin NBI en áreas urbanas;
- ② 96.610 lo hacían en hogares con NBI en áreas urbanas;
- ③ 122.701 pertenecían a hogares sin NBI en áreas rurales;
- ④ 166.814 habitaban en áreas rurales en hogares con NBI.

Todas estas proposiciones son *verdaderas*, en el sentido de que traducen con exactitud el significado de cada cifra; pero consideradas en conjunto constituyen una lectura puramente *redundante* de la información contenida en la Tabla 4.2, y no agregan nada a lo que ésta ya está mostrando por sí misma. En general, cuando se analizan tabulaciones bi-variadas, el interés debe focalizarse en determinar si existe alguna *relación* entre las dos variables. En otros términos, partimos siempre de una *hipótesis*, más o menos explícita, acerca de la existencia o no de una relación entre las dos variables.

Modos alternativos de análisis

Hay básicamente dos modos de abordar el análisis de una tabla. De acuerdo a una distinción establecida por Zelditch (1959), se la puede analizar de manera asimétrica o simétrica. En el modo asimétrico, el interés está puesto en observar el efecto de una de las variables sobre la otra. Por lo contrario, en el análisis simétrico no se presupone que una variable funja como “**causa**” de la otra. Abordaremos sucesivamente estas dos alternativas, teniendo siempre presente que una tabla no es intrínsecamente simétrica o asimétrica, sino que esta distinción se limita al modo en que se decide encarar el análisis en función de los objetivos del investigador.

⁴ Dada la imprecisión de nuestros instrumentos de medición.

⁵ Según Galtung, « La presentación de los porcentajes con 1 o incluso con 2 decimales no tiene sentido a menos que 1) la **calidad** de la recolección sea tan buena que tenga sentido decir que el 70,1% y no el 70,2% dicen ‘sí’, etc.; 2) el **propósito** de la recolección de datos sea tal, que sea diferente para la interpretación que el 70,1% y no el 70,2% diga ‘sí’, etc. En general sugerimos que los porcentajes deben presentarse sin ningún decimal, para evitar una impresión de exactitud que es a menudo completamente espuria» (1966: II, 231). Ya Bachelard decía que « El exceso de precisión, en el reino de la cantidad, se corresponde muy exactamente con el exceso de pintoresco, en el reino de la cualidad», y veía en tales excesos las marcas de un espíritu no científico (1972: 212).

⁶ En una visión *diacrónica*, ‘50,1%’ podría tener el valor de significar la inversión de una tendencia: de un predominio histórico de la población rural, se pasa a una preponderancia de los habitantes urbanos. En cambio, desde un punto de vista *sincrónico*, podría convenir escribir ‘50%’, destacando la semejanza cuantitativa entre ambos sectores.

¿Qué es analizar una tabla asimétricamente?

El caso asimétrico: se plantea siempre que se elige considerar que una variable -la variable *independiente*-incide sobre la distribución de la otra -la variable *dependiente*. Hay tablas en las que cualquiera de las dos variables puede fungir como “causa” de la otra. Aunque también suele ocurrir que se “imponga”, por así decirlo, el análisis en una determinada dirección.

La “regla de causa y efecto” o “primera regla de Zeisel” se aplica, en palabras de su autor, «siempre que uno de los dos factores del cuadro dimensional pueda considerarse como causa de la distribución del otro factor. La regla es que *los porcentajes deben computarse en el sentido del factor causal*» (Zeisel, 1962: 37).

¿ Puede esta regla aplicarse a nuestra Tabla 4.2? Responder positivamente a esta pregunta supondrá considerar, por ejemplo, que el tipo de asentamiento de la población “determina” o “condiciona” una probabilidad diferencial de pertenecer a un hogar con NBI. Esta hipótesis es plausible, si se tiene en cuenta que, por lo general, en nuestros países subdesarrollados el nivel de vida de las poblaciones rurales es inferior al de los habitantes urbanos.

¿ Cómo se lee el título de una tabla?

El título ya es merecedor de algunas observaciones, en tanto ejemplifica un cierto código cuyas reglas debemos conocer si queremos comprender acabadamente el significado de la Tabla 4.2.1:

- ① Las dos variables se encuentran claramente identificadas; se trata, respectivamente, de ‘Pertenencia de la población a hogares con Necesidades Básicas Insatisfechas’ (que en el encabezamiento de las columnas figura como ‘Hogares con NBI’, y cuyos valores son ‘Sí’ y ‘No’) y de ‘Tipo de asentamiento’ (con los valores ‘Urbano’ y ‘Rural’).

Tabla 4.2.1: Misiones, 1980 - Pertenencia de la población a hogares con NBI según tipo de asentamiento (%)

<i>Tipo de asentamiento</i>	<i>Hogares con NBI</i>		<i>Total (= 100%)</i>
	<i>No</i>	<i>Sí</i>	
<i>Urbano</i>	66,8	33,2	291.007
<i>Rural</i>	42,4	57,6	289.515
<i>Total</i>	54,6	45,4	580.522

Fuente: Tabla 4.2

- ② Entre los nombres de las dos variables se intercala la preposición ‘según’; no hubiera sido incorrecto utilizar alguna otra preposición como ‘por’ o ‘de acuerdo’, pero cabe atender al orden en que se introducen los nombres de las variables. Tomando el ‘tipo de asentamiento’ como variable independiente, ésta es introducida a continuación de la proposición ‘según’; en efecto, la presentación de los datos en la Tabla 4.2.1 apunta a destacar esta idea: **según** sea su tipo de asentamiento tenderán las personas a diferir en cuanto al valor mantenido en la variable dependiente.

- ③ El título finaliza con la expresión ‘(%)’; ello nos indica que las cifras consignadas en las celdas son porcentajes, y no frecuencias absolutas.⁷
- ④ En el encabezamiento de la última columna aparece la expresión ‘**Total (100%)**’. Esto quisiera expresar: a) que en dicha columna las cifras **no** son porcentajes sino frecuencias absolutas; y b) que las cifras absolutas de la columna fueron tomadas como base para calcular los porcentajes de las celdas.⁸

¿Cómo se lee una cifra porcentual?

Procedamos ahora a la lectura de la Tabla 4.2.1. Habiendo tomado ‘Tipo de asentamiento’ como variable independiente, hemos en consecuencia calculado los porcentajes “en el sentido de esta variable, nuestro “factor causal”. Ello quiere decir que *las bases para el cálculo porcentual están dadas por el total de casos para cada valor de la variable independiente*.

En la celda superior izquierda de la tabla observamos ‘66,8’, y sabemos -por el título - que la cifra corresponde a un porcentaje. La lectura correcta de esta cifra tiene lugar en dos pasos, cada uno de los cuáles supone responder a una pregunta.

① Lo primero que debemos inquirir es: “¿ 66,8% **de qué?** (o ¿de quiénes?)”. La única respuesta correcta es: “del 100% constituido por los 291.007 habitantes urbanos”; es decir, buscamos primero en la tabla dónde está el 100% —en la primera hilera—, y dirigimos luego nuestra vista hacia el encabezamiento de dicha hilera leyendo: ‘Urbano’. Cumplimentado este primer paso, estaremos en condiciones de preguntarnos con éxito...

② ...“ ¿ **Qué sucede** con este 66,8%?”, y podremos responder: “viven en hogares sin NBI”. A esta segunda pregunta respondimos simplemente dirigiendo nuestra atención hacia el encabezamiento de la columna: ‘No’.

Así, el significado de la primera celda puede expresarse: «De todos los habitantes urbanos de Misiones, hay un 66,8% que pertenece a hogares sin NBI».

Igualmente correcto sería escribir:

« Un 66,8% de la población urbana vive en hogares sin NBI».

Es obvio, que existe una posibilidad cierta de optar por diferentes redacciones; pero lo fundamental es que la expresión literaria respete el significado de la cifra. Se puede pensar en dos grandes tipos de problemas que se plantean en la lectura de los porcentajes.

❶ Hablaremos de problemas “*lógicos*” cuando se produce una *falsa* lectura de la cifra porcentual. Estos errores devienen de una *confusión acerca de la base* sobre la cual está calculado el porcentaje. En cualquier tabla de doble entrada, existen potencialmente tres bases sobre las cuales es posible calcular los porcentajes, a saber,

— El total de la hilera: ‘291.007’, en este caso;

⁷ Igualmente claro sería omitir el signo de porcentaje en el título y consignarlo a continuación de cada cifra: 66,8%, 33,2%, etc.

⁸ Esta última convención, además de ser tan arbitraria como las anteriores, está lejos de ser universalmente reconocida. Otra manera habitual de proceder es hacer figurar en la última columna para todas las hileras la expresión ‘100,0’; pero aparte de ser ésta una información redundante (en el primer renglón es obvio que $33,2 + 66,8 = 100,0$), este procedimiento tiene el inconveniente de que hace desaparecer toda referencia a las frecuencias absolutas que fueron tomadas como base; en cambio, mientras éstas continúen apareciendo siempre será posible reconstruir las frecuencias absolutas correspondientes a las celdas: por ejemplo, $291.000 \times 0,332 = 96.614 \approx 96.610$ (la pequeña diferencia deviene del redondeo del porcentaje). Otra posibilidad es anotar entre paréntesis las bases de los porcentajes: ‘100,0 (291.007)’. También es posible duplicar cada cifra del cuadro consignando siempre las frecuencias absolutas y relativas -estas últimas de preferencia con algún recurso tipográfico distinto-, aunque esta práctica tiende a restarle nitidez a los datos.

- El total de la columna, '317.098'; y
- El "total total", el 'n': '580.522'.

Se comete un *error lógico* cuando un porcentaje es leído sobre una base que no fue la utilizada para calcularlo. Así, si se lee « Un 66,8% de los habitantes de Misiones son urbanos y viven en hogares sin NBI», la expresión lingüística da a entender que el porcentaje fue calculado sobre el total de la población provincial, con lo cual *el enunciado pasa a expresar una proposición falsa* (el porcentaje que correspondería a dicha expresión lingüística no sería '66,8' sino '33,5').

Igualmente erróneo sería escribir «En Misiones, un 66,8 % de las personas pertenecientes a hogares sin NBI residen en asentamientos urbanos». La construcción de esta frase supone que el 66,8% fue calculado sobre el total de personas pertenecientes a hogares sin NBI, con lo que el enunciado es también falso (para esta redacción, el porcentaje correcto sería '61,3'). Por ende, hay una sola manera de generar enunciados verdaderos; es eligiendo una construcción lingüística que dé cuenta sin ambigüedad alguna del modo en que la cifra porcentual ha sido efectivamente calculada.⁹

❷ Pero también se presentan problemas **pragmáticos**. Sucede que diferentes redacciones son susceptibles de comunicar distintos significados. Comparemos los siguientes enunciados:

a.-«**Más de** dos tercios de los habitantes urbanos viven en hogares que no presentan NBI »;

b.-« **Solamente** un 66,8% de los habitantes urbanos pertenece a hogares sin NBI ».

Tanto 'a' como 'b' expresan correctamente el porcentaje, desde una perspectiva puramente lógica. Sin embargo, es evidente que ambos enunciados no tienen el mismo significado: ciertamente 'a' trasunta una visión de la situación más optimista que 'b'. Sucede que, como lo explicara hace tiempo el lingüista Roman Jakobson, no es posible denotar sin connotar: las operaciones de selección y combinación que están necesariamente en obra en la producción de todo discurso introducen en él una dimensión ideológica.¹⁰ Podemos probar de eliminar los adverbios en nuestros enunciados 'a' y 'b', con lo que obtenemos expresiones cuyo valor lingüístico es muy similar:

a1.-« Dos tercios de los habitantes urbanos viven en hogares que no presentan NBI »;

b1.-« Un 66,8% de los habitantes urbanos pertenece a hogares sin NBI ».

Aparentemente habríamos eliminado así toda valoración, permaneciendo sólo la fría cifra. Pero esto es creer que el significado de un enunciado individual sólo depende de su contenido intrínseco. Lo cierto es que este enunciado se inserta en un contexto más amplio, el discurso al que pertenece, cuyo significado global concurre a producir, pero que a la vez determina grandemente su propia significación. Lo expresado abona la idea de que estos problemas que hacen a la pragmática del discurso son inevitables. A lo sumo puede intentarse limitarlos controlando en alguna medida la adverbiación y la adjetivación.

⁹ Puesto que siempre existen tres alternativas para el cálculo de los porcentajes, es un hecho tan lamentable cuanto inevitable que para leer un porcentaje siempre existan dos posibilidades de equivocarse y sólo una de acertar...

¹⁰ Verón caracterizó a la ideología « como un **nivel de significación** de todo discurso transmitido en situaciones sociales concretas, referido al hecho inevitable de que, por su propia naturaleza, todo mensaje transmitido en la comunicación social posee una dimensión connotativa» (Verón, 1972: 309).

¿Cómo se lee un conjunto de porcentajes?

Podemos ahora leer el conjunto de las cifras de la Tabla 4.2.1, que traducimos en la siguiente serie de enunciados:

1. - Un 66,8% de los habitantes urbanos pertenece a hogares sin NBI;
2. - Un 33,2% de los habitantes urbanos pertenece a hogares con NBI;
3. - Un 42,4% de los habitantes rurales pertenece a hogares sin NBI;
4. - Un 57,6% de los habitantes rurales pertenece a hogares con NBI;
5. - Un 54,6% de todos los habitantes pertenece a hogares sin NBI; y
6. - Un 45,4% de todos los habitantes pertenece a hogares con NBI.

Todos estos enunciados son verdaderos. Empero, su mera enumeración no constituye una “buena” lectura de la Tabla 4.2.1. En efecto, este conjunto de enunciados a) es en gran medida redundante; y, sobre todo, b) no apunta a destacar lo fundamental, esto es, la relación entre las variables que postula nuestra hipótesis y que es la única razón por la que los datos han sido presentados como se lo ha hecho, calculando los porcentajes en una dirección determinada.

En cuanto a la redundancia, debe resultar claro que el contenido del enunciado 2 ya está incluido —implícitamente— en el enunciado 1: si un 66,8% de los habitantes urbanos pertenece a hogares sin NBI, y nos encontramos tratando con una variable dicotómica, ello implica que *necesariamente* hay un 33,2% de los habitantes urbanos en hogares con NBI.¹¹ Y viceversa: si es verdadero el enunciado 2, necesariamente lo será también el 1. Es evidente que la misma relación se da para los pares de enunciados 3 y 4, y 5 y 6.

Aunque ello no resulte tan obvio, también son redundantes en cierto modo los porcentajes correspondientes a la hilera del total. Así, el que 45,4% de todos los habitantes pertenezcan a hogares con NBI no es más que el resultado de un promedio ponderado entre el 33,2% de urbanos y el 57,6% de rurales que presentan esta característica. Es ésta una propiedad interesante de los porcentajes marginales: necesariamente su valor se ubicará dentro de un rango limitado por los valores porcentuales consignados en las celdas correspondientes; en este caso, el porcentual del marginal deberá ser superior a 33,2 e inferior a 57,6; el que se encuentre “más cerca” de una u otra de estas cifras dependerá sólo del peso relativo de ambos grupos (el ‘rural’ y el ‘urbano’) sobre la población total.¹² Es por esta razón que frecuentemente se omite la presentación de los porcentajes marginales.¹³

En suma, si intentamos reducir al mínimo la redundancia en la lectura de la tabla, podemos considerar que lo esencial de la información está contenido en los enunciados 2 y 4 (o, indiferentemente, en los 1 y 3). De este modo, destacaremos el sentido fundamental que queremos prestarle a los datos: en estas dos cifras -33,2% y 57,6%-¹⁴ está resumido lo que la tabla significa para nosotros. Comparando estos dos porcentajes, nuestra lectura pone en evidencia la relación estocástica entre las dos variables postulada por nuestra hipótesis:

«Mientras que en la población urbana hay un 33,2% de habitantes en hogares con NBI, entre los pobladores rurales este porcentaje asciende al 57,6%».

Se corrobora por lo tanto la existencia de una probabilidad diferencial de pertenecer a un hogar con NBI en función del tipo de asentamiento de la población.

Una alternativa interesante para presentar esta información puede ser mediante un gráfico de columnas. Se ve claramente cómo ambas poblaciones son de tamaños similares y cómo la

¹¹ 33,2 es el complemento necesario para alcanzar al 100,0%; en efecto, $100,0 - 66,8 = 33,2$.

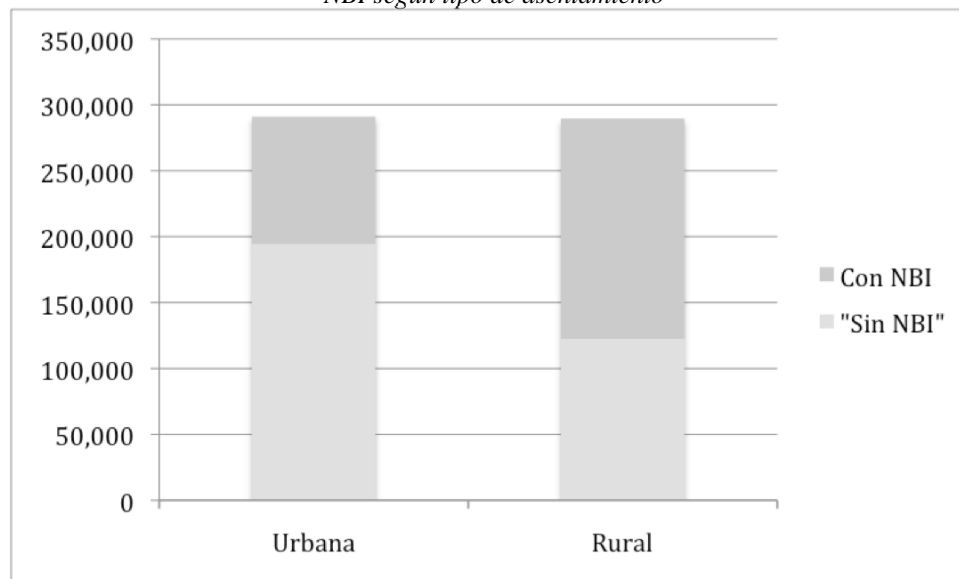
¹² En el caso particular de la tabla que nos ocupa, ambos grupos tienen aproximadamente el mismo peso, por lo que podría calcularse: $(33,2 + 57,6) / 2 = 45,4\%$.

¹³ Sin embargo, cuando se trabaja con tablas de mayores dimensiones -y no basadas en dicotomías-, el porcentaje marginal puede funcionar como un punto de referencia útil que facilita la atribución de un significado a los porcentajes en las celdas. Incluso en nuestra misma Tabla 4.1.1 podría decirse que, frente a un 45,4% del total de la población que pertenece a hogares con NBI, el 33,2% de los ‘urbanos’ es **comparativamente bajo**.

¹⁴ O, alternativamente, en el par: 66,8% y 42,4%.

proporción de personas pertenecientes a hogares con NBI es mucho mayor en el campo:¹⁵

Figura 4.1: Misiones, 1980 - Pertenencia de la población a hogares con NBI según tipo de asentamiento



Fuente: Tabla 4.2.1

Primera regla de Zeisel

Nos encontramos ahora en condiciones de completar nuestra formulación de la primera regla de Zeisel, referida al caso del análisis asimétrico:

« LOS PORCENTAJES SE CALCULAN EN EL SENTIDO DE LA VARIABLE INDEPENDIENTE, Y SE COMPARAN EN EL SENTIDO DE LA VARIABLE DEPENDIENTE. »¹⁶

En efecto, esto es todo lo que hemos hecho: hemos computado los porcentajes en el sentido horizontal (en este caso), y los hemos leído en el sentido vertical.¹⁷ Lo fundamental es calcular los porcentajes en la dirección adecuada, esto es, en el sentido de la variable a la que asignaremos el rol de ‘independiente’ en nuestro análisis.

¿Cuál es la variable independiente?

Ahora bien, cuál sea la variable independiente, es una cuestión que está enteramente supeditada a los objetivos de nuestro análisis. Así como en la Tabla 4.2.1 se eligió el ‘Tipo de asentamiento’, así puede tomarse también como independiente la variable ‘Pertenencia a hogares con NBI’:

¿ En qué sentido puede pensarse que el hecho de pertenecer o no a un hogar con NBI “determine” el tipo de asentamiento de las personas? ¿ Puede sostenerse una brumosa hipótesis según la cual los “pobres” preferirían residir en áreas rurales?.

¹⁵ Los gráficos tienen un gran poder de comunicación y son un recurso al que se puede apelar en la presentación de informes de investigación. Es posible que muchas personas perciban mejor una relación expresada visualmente, aunque se pierda algo de precisión con respecto a los datos numéricos.

¹⁶ Esta fórmula pertenece a Zelditch (1959). Galtung dice: «sacar siempre los porcentajes perpendicularmente a la dirección de la comparación» (1966: II, 233).

¹⁷ Algunos autores sostienen la conveniencia de presentar siempre la variable independiente en el encabezamiento del cuadro y la variable dependiente en las hileras, con lo que los porcentajes se computarían siempre en el sentido vertical. Hay un único fundamento razonable para esta práctica, y es el mantener la analogía con el tratamiento de variables cuantitativas graficadas en un diagrama de ejes cartesianos, en el que la x -variable independiente- aparece siempre en la abscisa.

Tabla 4.2.2: Misiones, 1980 - Distribución de la población por tipo de asentamiento y pertenencia a hogares con NBI

Tipo de asentamiento	Hogares con NBI		Total
	No	Sí	
	61	37	
Rural	39	63	50
Total	(317.098)	(263.424)	(580.522)

Fuente: Tabla 4.2

La Tabla 4.2.2 nos permite pensar la diferencia que media entre las expresiones ‘variable independiente’ y ‘causa’. Es evidente que en este caso no tiene demasiado sentido pensar en la pobreza como “causa” del tipo de asentamiento.¹⁸ Pero es perfectamente posible leer:

« En las zonas rurales de la Provincia se concentra el 63% de las personas pertenecientes a hogares con NBI, frente a sólo un 39% de las que pertenecen a hogares no carenciados ».

Se trata de una presentación de los datos de la Tabla 4.2 que tiende a destacar cómo la pobreza se concentra mayoritariamente en las áreas rurales de Misiones. No solamente la Tabla 4.2.2. es tan “verdadera” como la 4.2.1, sino que ambas son igualmente válidas. Aún cuando la Tabla 4.2.1 indujera en nosotros un mayor sentimiento de satisfacción, esta segunda interpretación no sería menos legítima por ello.¹⁹

Segunda regla de Zeisel

Existe sin embargo una limitación al sentido en que es lícito computar los porcentajes, cuando se trabaja con datos muestrales. No siempre las muestras tienen la virtud de ser autoponderadas. Por diversas razones, puede ocurrir que una muestra no sea representativa de la población en algún sentido; hemos visto en el capítulo anterior que el caso es frecuente al utilizar diseños de muestra estratificados o por cuotas.

Imaginemos que queremos investigar acerca de la conformidad de un grupo de estudiantes de las carreras de Trabajo Social y de Turismo con el sistema de promoción por examen final; a tales efectos, seleccionamos una muestra de 40 alumnos de cada carrera, a sabiendas de que los totales de alumnos eran de 160 para Trabajo Social y de 80 para Turismo.

En casos semejantes sólo cabe calcular los porcentajes en el sentido en que se lo ha hecho en el ejemplo de la Tabla 4.3, y se podrá concluir que la proporción de disconformes con el sistema de aprobación por examen final es más elevada entre los alumnos de Turismo (82%) que entre los de Trabajo Social (60%).

¹⁸ Estrictamente, para poder hablar de una relación causal entre las variables X e Y se requiere contar con evidencia de tres tipos: a) variación concomitante de X e Y; b) precedencia temporal de X con respecto a Y; y c) eliminación de otros posibles determinantes de Y (cf. Sellitz *et al.*, 1968: 100 y ss.).

¹⁹ En los estudios de mercado es usual distinguir entre dos tipos de porcentajes, según la dirección en la que han sido calculados. Así, el porcentaje ‘de penetración’, también denominado ‘cuota de mercado’ se calcula sobre el total de integrantes de una categoría –de edad, sexo, nivel educativo, etc.– e indica cuántos de este total consumen el producto (o manifiestan su intención de votar por un candidato, si se trata de *marketing* político); en cambio, el porcentaje ‘de composición’ indica sobre el total de consumidores del producto (o de votantes del candidato), qué proporción corresponde a una categoría en particular (cf. Antoine; 1993: 33 y ss.). Por analogía, mientras que la Tabla 4.2.1 estaría indicando una mayor penetración de la pobreza en áreas rurales (57,6% de los rurales son pobres), la tabla 4.2.2 mostraría el peso mayoritario de los habitantes rurales en la composición de la población con NBI (63% de los pobres son rurales).

Tabla 4.3: Conformidad con el sistema de examen final según carrera

Conformidad con el examen final	Carrera		Total
	Turismo	Trabajo Social	
Sí	7 18%	16 40%	23 29%
	33 82%	24 60%	57 71%
No	40	40	80
Total	100%	100%	100%

Fuente: elaboración propia.

En general, la segunda regla de Zeisel -que no es más que una limitación a la primera - afirma:

« CUANDO UN CONJUNTO DE MARGINALES NO ES REPRESENTATIVO DE LA POBLACION, LOS PORCENTAJES DEBEN COMPUTARSE EN LA DIRECCION EN QUE LA MUESTRA NO ES REPRESENTATIVA ».²⁰

En efecto, en nuestra muestra la relación entre los alumnos de las dos carreras es de 1:1 (40 en cada una), en tanto sabemos que en la población la relación real es de 1:2 (hay el doble de alumnos en Trabajo Social). Como nuestra muestra no es representativa por carrera, los porcentajes sólo pueden calcularse en esa dirección: sobre el total de alumnos de cada carrera.

¿ Qué ocurriría si calculáramos directamente los porcentajes en el sentido horizontal? Concluiríamos -erróneamente - que del total de los estudiantes que se manifiestan conformes con el sistema de examen final hay un 70% que pertenece a la carrera de Trabajo Social:

Tabla 4.3.1: Carrera según conformidad con el sistema de examen final (%)

Conformidad con el examen final	Carrera		Total
	Turismo	Trabajo Social	
Sí	30	70	100
No	58	42	100

(n = 80)

Fuente: Tabla 4.3

Es verdad que *en la muestra* se da este 70%; pero ello ocurre debido a un factor arbitrario que es el tamaño relativo de la muestra en ambas carreras. Como en la muestra la carrera de Trabajo Social se encuentra subrepresentada con relación a su peso real en la población, y sus estudiantes son más conformistas que los de Turismo, en la población deberá ser mayor el porcentaje de conformes concentrados en aquella carrera.

Supongamos que de haber trabajado con el universo, se hubiera obtenido las mismas proporciones de conformistas en ambas carreras que las registradas en la Tabla 4.3. Los resultados serían los presentados en la Tabla 4.3.2.²¹

²⁰ La expresión de la regla pertenece a Zelditch (1959).

²¹ Para construir la Tabla 4.3.2, simplemente multiplicamos por 2 las frecuencias absolutas correspondientes a los estudiantes de Turismo, y por 4 las de Trabajo Social.

Tabla 4.3.2: Carrera según conformidad con el sistema de examen final

Conformidad con el examen final	Carrera		Total
	Turismo	Trabajo Social	
Sí	14 18%	64 82%	78 100%
No	66 41%	96 59%	162 100%
Total	80 33%	160 67%	240 100%

Fuente: elaboración propia.

Se observa que hay en realidad un 82% de los conformistas que pertenecen a Trabajo Social y que, por lo tanto, el 70% que arrojaba la Tabla 4.3.1 no podía ser tomado como una estimación válida de la proporción existente en la población. Al no ser representativa la muestra en cuanto al peso relativo de ambas carreras, el cómputo *directo* de los porcentajes sólo se puede realizar como se lo hizo en la Tabla 4.3.

Si se desea calcular los porcentajes en la otra dirección, no se lo puede hacer directamente, sino que es indispensable recurrir a algún sistema de ponderación de las frecuencias análogo al utilizado en la Tabla 4.3.2.

¿Y el modo simétrico?

Las dos reglas de Zeisel sintetizan lo esencial para el tratamiento asimétrico de tablas de contingencia. El análisis simétrico de estas tablas reviste comparativamente un interés menor. En este caso se computarán los porcentajes correspondientes a todas las frecuencias condicionales y marginales sobre la misma base del total de casos.

En el análisis asimétrico, el cálculo de los porcentajes sobre columnas -o sobre hileras- permite lograr una estandarización de las frecuencias condicionales que quedan así liberadas de los efectos de las diferencias marginales. Esto nos permitía en la Tabla 4.3 comparar un 82% de disconformes en Turismo con un 60% en Trabajo Social, aún cuando en la población hubiera el doble de Trabajadores Sociales.

En cambio, si los porcentajes se calculan todos sobre el 'n' del cuadro, no se logra ninguna estandarización, ya que las diferencias marginales continúan pesando sobre las frecuencias condicionales. En este sentido, debe resultar evidente la necesidad de que la muestra sea representativa en todos los sentidos, si se desea analizar simétricamente una tabla compuesta a partir de observaciones muestrales. Así, no cabría someter la Tabla 4.3 a un tratamiento simétrico, por la misma razón que tampoco resultaba lícito el cómputo horizontal de los porcentajes.

Pero, sobre todo, el análisis simétrico no es apto para examinar la existencia de una relación de dependencia entre las dos variables; optamos por este tipo de análisis cuando **no** interesa indagar acerca del presunto "efecto" de una variable sobre la otra. Así la Tabla 4.2 podría también ser analizada simétricamente.

Tabla 4.2.3: Misiones, 1980 - Distribución de la población por tipo de asentamiento y pertenencia a hogares con NBI

Tipo de asentamiento	Hogares con NBI		Total
	No	Sí	
Urbano	33,5	16,6	49,9
Rural	21,1	28,8	
Total	54,6	45,4	(580.522)

Fuente: Tabla 4.2.

Leeremos así que, de los 580.522 habitantes de la Provincia, hay un 33,5% que pertenece a hogares urbanos sin NBI, seguido por un 28,8% de rurales con NBI, 21,1% rurales sin NBI y 16,6 de urbanos con NBI. En esta forma de presentación de los datos, ya no se visualiza con la misma claridad el efecto de una variable sobre la otra, lo que no implica que ésta deje de ser una interpretación tan legítima como las anteriores. Simplemente, habrá variado nuestro propósito. Es posible, por ejemplo, que tengamos un interés especial en saber que un 28,8% de la población de Misiones pertenece a hogares rurales con NBI, para comparar esa cifra con el 1,8% que se registra para la misma categoría de población en la Provincia de Buenos Aires, más urbanizada y menos pobre, o con el 30,0% de la vecina Corrientes, más urbanizada y más pobre.

La Tabla 4.2 se basa en datos censales. Muchas investigaciones realizadas por muestreo pueden no perseguir el objetivo de determinar la existencia de una relación entre dos variables, sino proponerse la simple estimación de la proporción de una población que reúne determinadas características. De ser el caso, el tratamiento simétrico de los datos obtenidos por muestra permite obtener estimaciones de las proporciones de personas dentro de cada categoría de la población. Pero, si por lo contrario el objetivo es establecer una relación de dependencia entre dos variables, convendrá tratar la tabla asimétricamente.

2. EL ANÁLISIS DE LA RELACIÓN ENTRE VARIABLES

Cuando observamos mediante el tratamiento asimétrico de una tabla que una de las variables aparece determinando o afectando a la otra, podemos decir que ambas variables están *asociadas*.²² La medida de asociación más frecuentemente utilizada es, por lejos, la diferencia porcentual. Por otra parte, cuando se trata con muestras se plantea el problema adicional de determinar la significación estadística que se le puede prestar a una asociación entre variables. Abordaremos sucesivamente estos aspectos, para presentar luego algunos coeficientes de asociación.

2.1 La diferencia porcentual: una medida de la asociación

Por su simplicidad de cálculo y por la claridad de su significado, la diferencia porcentual es sin duda la medida de asociación más popular. En esencia consiste en una sistematización de la primera regla de Zeisel.

Consideremos el siguiente ejemplo, cuyos datos provienen de una muestra de 121 estudiantes,²³ partiendo de la hipótesis de que el grado de conocimiento político condiciona el grado de

²² El concepto de 'asociación' se usará para describir la existencia de una relación entre variables en una tabla de contingencia; para distribuciones multivariantes se hablará de 'correlación'.

²³ Los datos provienen del estudio "Participación política del estudiante de la FHCS-UNaM" (inédito) realizado por alumnos de Antropología Social en 1984.

participación política.²⁴

Tabla 4.4: Grado de participación política y grado de conocimiento político

Participación política	Conocimiento político		Total
	Bajo	Alto	
Alto	6	13	19
Bajo	59	43	102
Total	65	56	121

Fuente: elaboración propia.

Si lo que se quiere es comprobar el efecto del conocimiento sobre la participación, los porcentajes se deben computar en el sentido de la variable ‘conocimiento’, o sea verticalmente:

Tabla 4.4.1: Grado de participación política según grado de conocimiento político (%)

Participación política	Conocimiento político		Dif. %
	Bajo	Alto	
Alto	9	23	14
Bajo	91	77	-14
Total	100	100	(n = 121)

Fuente: elaboración propia.

Hemos simplemente aplicado la regla según la cual, los porcentajes se computan en dirección de la variable independiente y se comparan en la otra dirección. Salvo que ahora hacemos aparecer en la última columna la diferencia porcentual:

LA DIFERENCIA PORCENTUAL SE CALCULA EN LA DIRECCION EN QUE SE REALIZA LA COMPARACION

Mientras que en los alumnos de Bajo conocimiento sólo hay un 9% con alta participación, entre los de Alto conocimiento hay un 23%: es decir, hay un 14% más de alta participación política.²⁵

²⁴ Como habrá de verse a la brevedad, es igualmente plausible sostener la hipótesis de que «A mayor grado de participación política, mayor conocimiento». En términos de Zetterberg, éste es un ejemplo de relación **reversible e interdependiente** entre las variables (1968: 59 y ss.).

²⁵ Igualmente podríamos haber comparado los porcentajes de Baja participación, encontrando que en los alumnos de Alto conocimiento hay un 14% menos. Tratando con variables dicotómicas, y considerando que por definición los porcentajes deben sumar 100%, no puede sorprendernos que las diferencias porcentuales sean en ambos renglones

Ahora bien, imaginemos que quisiéramos en cambio determinar el efecto de la participación sobre el conocimiento:

Tabla 4.4.2: Grado de participación política y grado de conocimiento político

<i>Participación política</i>	<i>Conocimiento político</i>		<i>Total</i>
	<i>Bajo</i>	<i>Alto</i>	
<i>Alto</i>	32	68	100
<i>Bajo</i>	58	42	100
<i>Dif. %</i>	-26	26	(n = 121)

Fuente: elaboración propia.

Entre los altamente participativos hay un 26% más con alto conocimiento. Este 26% es *también* una medida de la asociación entre las variables, y tan válida como la anterior, aunque a todas luces diferente. Según sea nuestro interés, podremos optar por una u otra cifra; pero lo que muestra el ejemplo es que la diferencia porcentual no nos brinda una medida *general* de la asociación en la tabla. Sucede que los porcentajes son sensibles a los cambios en las distribuciones marginales, y que precisamente en la Tabla 4.4 estos marginales difieren en forma notable (19 y 102 para ‘participación’, 56 y 65 para ‘conocimiento’).

Cualquier uso de la diferencia porcentual como indicador resumen de la asociación en una tabla implica una gran pérdida de información. Además, este problema se magnifica al trabajar con tablas de formato mayor al 2 x 2; cuanto más elevado sea el número de valores de cada variable, se multiplicará la cantidad de diferencias porcentuales computables, y resultará aún más discutible la elección de una de las tantas diferencias posibles como medida resumen de la asociación en la tabla.

Tabla 4.5: Evaluación de la situación social según NES (%)

<i>Evaluación de la situación social</i>	<i>Nivel económico-social</i>			<i>Total</i>
	<i>Bajo</i>	<i>Medio</i>	<i>Alto</i>	
<i>Favorable</i>	35	38	47	41
<i>Neutra</i>	18	32	29	27
<i>Desfavorable</i>	45	30	24	32
<i>Total (100%)</i>	(40)	(73)	(49)	(162)

Fuente: elaboración propia.

idénticas aunque de signo contrario: necesariamente, todo aumento del porcentaje en una categoría debe implicar una disminución en la otra.

La Tabla 4.5 permite ilustrar este problema.²⁶ Puesto que el NES ha sido tomado como variable independiente, lo lógico es leer el cuadro comparando entre sí los porcentajes de cada hilera. Evidentemente, el sentido general de la tabla es que cuanto menor es el NES, más negativa resulta la evaluación de la situación: ello surge nítidamente de la comparación de los porcentajes de la última hilera. Pero es claro que no existe una única diferencia porcentual, sino nueve posibilidades distintas de cómputo de diferencias; solamente en esa última hilera, sería posible comparar 45 con 30, 30 con 24, o 45 con 24; y va de suyo que ninguna de estas diferencias es más “verdadera” que las otras.

Aún tomando en consideración estos defectos, no cabe menospreciar a la diferencia porcentual como instrumento del análisis. De hecho, en su práctica cotidiana el investigador la aplicará casi instintivamente, para tener una medida rápida de la asociación. Por lo demás, al trabajar con muestras la cuestión no radica simplemente en determinar el grado en que dos variables están asociadas, sino que se plantea un problema adicional.

¿Es esta relación estadísticamente significativa?

En la Tabla 4.4.1. la hipótesis inicial parecía corroborarse. Se observaba en efecto una diferencia positiva del 14% en cuanto a la participación de los estudiantes que contaban con una mayor grado de conocimiento político. Sin embargo, esta relación se verifica en una muestra, constituida por 121 estudiantes que eran sólo una parte de la totalidad de los estudiantes de la FHCS-UNaM en 1984. La muestra con la que trabajamos es solamente una de las tantas muestras que se hubieran podido extraer del universo de la investigación. Tal vez el azar haya sido la razón de que apareciera en la muestra este 14% más, cuando en realidad esta relación no se daba en el universo. La cuestión es: ¿Podemos considerar a esa diferencia del 14% lo suficientemente importante como para asumir que representa una diferencia existente realmente en el universo?. Cuando nos formulamos este tipo de preguntas, estamos inquiriendo si la relación es *estadísticamente significativa*.

2.2 El test de χ^2 : una medida de la significación estadística

El test de χ^2 (chi-cuadrado) es una de las respuestas posibles a esta pregunta. Dicho test es una de las pruebas de significación estadística más populares y se basa en una medida de cuánto se apartan las frecuencias condicionales observadas en la muestra de lo que serían las frecuencias esperables si no existiera ninguna relación entre las variables.

Retornemos a los datos de la Tabla 4.4 para considerar exclusivamente las frecuencias marginales:

	B	A	
A			19
B			102
	65	56	121

Examinando sólo los marginales no se puede decir nada acerca de la relación entre las variables. En cierto sentido, debemos pensar que estos marginales son lo que son. O, más exactamente, dados estos marginales, no podríamos tener **cualquier** frecuencia dentro de la tabla.²⁷ Sin embargo, dentro de los límites establecidos por los marginales, es evidente que se puede imaginar muy diversas distribuciones de las frecuencias condicionales, y que estas distribuciones podrán ser muy diferentes en lo que hace a la relación entre las dos variables. Así, podríamos obtener:

²⁶ Los datos están tomados de un estudio (inédito) realizado por alumnos de la carrera de Antropología Social, en ocasión de las elecciones del 6 de septiembre de 1987 en Posadas.

²⁷ Así, por ejemplo, es claro que en ninguna de las dos celdas superiores podría haber una frecuencia mayor que ‘19’ (las frecuencias han de ser necesariamente números positivos).

		B	A		
A		0	19	19	
B		65	37	102	
		65	56	121	

a. Máxima relación posible

En esta alternativa la totalidad de los estudiantes con Bajo conocimiento tienen Baja participación. O bien, podríamos encontrarnos con esta distribución:

		B	A		
b. Ausencia total de relación	A	10	9	19	
	B	55	47	102	
		65	56	121	

Si computáramos los porcentajes de Alta participación, observaríamos que éstos son prácticamente idénticos para ambos niveles de conocimiento.²⁸

¿Qué entendemos por ‘ausencia de relación’?

Introducimos ahora una simbología nueva para representar la tabla de 2 x 2, que es la que se utiliza corrientemente para las fórmulas estadísticas:

a	b	a+b
c	d	c+d
a+c	b+d	n

Con la tabla b. ejemplificamos un caso de nula asociación entre las variables. En efecto, en esa tabla se comprueba que

$$10/65 = 9/56 = 19/121.^{29}$$

Vale decir que, generalizando, la asociación en una tabla de 2 x 2 será nula cuando:

$$\frac{a}{a+c} = \frac{b}{b+d} = \frac{a+b}{n}$$

Es claro que si el porcentaje en una celda es igual al de su marginal, el porcentaje en la otra celda también habrá de ser idéntico. Por lo tanto, podemos decir que la asociación es nula siempre que

$$\frac{a}{a+c} = \frac{a+b}{n}$$

O, expresado de otra forma,³⁰ no habrá asociación entre las variables cuando se cumpla que

$$f_e(a) = \frac{(a+b) \cdot (a+c)}{n}$$

Dicho de otro modo, cuando la frecuencia ‘a’ sea igual al producto de sus marginales (‘a + b’ y ‘a + c’) dividido por el total de casos en la tabla (‘n’), diremos que la asociación es nula. Esta es la *frecuencia esperada* en ‘a’ de no existir relación entre las variables. La podemos simbolizar como ‘f_e(a)’. De manera absolutamente análoga podemos definir, las frecuencias esperadas en todas las celdas:

²⁸ De hecho, así hemos procedido para elaborar esta distribución hipotética de las frecuencias condicionales.

²⁹ 9/56 es ligeramente superior a 10/65, pero ello obedece sólo a la imposibilidad de que aparezcan fracciones de estudiantes en las celdas dado que las frecuencias condicionales deben ser números enteros.

³⁰ Simple transformación de la ecuación, multiplicando ambos términos por la expresión ‘(a + c)’.

$$f_e(b) = \frac{(a + b) \cdot (b + d)}{n}$$

$$f_e(c) = \frac{(a + c) \cdot (c + d)}{n}$$

$$f_e(d) = \frac{(b + d) \cdot (c + d)}{n}$$

$f_e(a)$, $f_e(b)$, $f_e(c)$ y $f_e(d)$ se denominan ‘frecuencias esperadas’ en el sentido de que, dado un conjunto de marginales, son las frecuencias que esperaríamos hallar de no existir relación alguna entre las dos variables.

El test de χ^2 consiste simplemente en medir cuánto se desvían las frecuencias observadas respecto a las esperadas, debiendo entenderse que el conjunto de las frecuencias esperadas configura sólo un *modelo* posible de no-asociación basado en la idea de independencia estadística.

¿Cómo se calcula χ^2 ?

Para cualquier tabla de h hileras por c columnas, la fórmula para calcular χ^2 es la siguiente:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Para los datos de la Tabla 4.4, el cálculo de χ^2 se hará:

celda	f_o	f_e	$(f_o - f_e)$	$(f_o - f_e)^2$	$(f_o - f_e)^2 / f_e$
a	6	10,21	- 4,21	17,72	1,736
b	13	8,79	4,21	17,72	2,016
c	59	54,79	4,21	17,72	0,323
d	43	47,21	- 4,21	17,72	0,375
Σ	121	121,00	0,0		4,450

1) en la columna ‘ f_o ’ consignamos las frecuencias observadas en cada celda;

2) bajo ‘ f_e ’, las frecuencias esperadas (para su cálculo, conservamos dos decimales); observar la suma: son los mismos 121 estudiantes distribuidos ahora según un modelo de independencia estadística de las variables;

3) en ‘ $(f_o - f_e)$ ’, se consigna para cada celda la diferencia entre la frecuencia esperada y la observada; es una particularidad de la tabla de 2×2 que estos valores sean iguales para todas las celdas, con excepción de su signo; como es lógico (se trata siempre de los mismos 121 estudiantes) al tomar en cuenta los signos de las diferencias, su suma se hace igual a cero;

4) en $(f_o - f_e)^2$, elevamos al cuadrado las diferencias de la columna anterior; es un modo interesante de solucionar el problema de los signos, evitando que las diferencias se anulen;

5) estandarizamos cada una de las diferencias $(f_o - f_e)^2$ dividiendo a cada una de ellas por la frecuencia esperada correspondiente; es evidente, por ejemplo, que una diferencia de 4,21 es mucho más importante en la celda ‘b’ que en la ‘c’: en ‘b’ f_e es un 70% mayor que f_o , en tanto que en ‘c’ f_e sólo supera en un 10% a f_o ;

6) por último, sumamos todos los valores de la última columna, obteniendo el valor que arroja χ^2 para la Tabla 4.4:

$$\chi^2 = 4,450$$

¿Cómo interpretar un valor de χ^2 ?

Este valor sólo cobra sentido cuando se lo compara con el **valor crítico** correspondiente de la Tabla de χ^2 , la que presentamos en una versión resumida en la página siguiente. El valor crítico depende del nivel de significación con el que deseemos trabajar y del número de grados de libertad de nuestra tabla.

Tabla 4.6: Tabla resumida de valores críticos para χ^2

gl	Probabilidad				
	.10	.05	.02	.01	.001
1	2.706	3.841	5.412	6.635	10.827
2	4.605	5.991	7.824	9.210	13.815
3	6.251	7.815	9.837	11.341	16.268
4	7.779	9.488	11.668	13.277	18.465
5	9.236	11.070	13.388	15.086	20.517
6	10.645	12.592	15.033	16.812	22.457
7	12.017	14.067	16.622	18.475	24.322
8	13.362	15.507	18.168	20.090	26.125
9	14.684	16.919	19.679	21.666	27.877
10	15.987	18.307	21.161	23.209	29.588

A) El **nivel de significación** se refiere a la probabilidad de equivocarnos que estemos dispuestos a aceptar; si, por ejemplo, elegimos un nivel de significación de .05, ello equivale a considerar aceptable un riesgo del 5%.³¹ El riesgo consiste en la probabilidad de que dos variables que no están asociadas en la población aparezcan relacionadas en nuestros datos muestrales. Así, un nivel de significación de .05 equivale a aceptar que, realizando infinitas extracciones de una muestra de tamaño n - el tamaño de nuestra muestra - a partir de la población, en un 5% de esas infinitas muestras posibles las variables aparecerían asociadas aún cuando no lo estuvieran en el universo.³² Por lo general, los niveles más frecuentemente utilizados son el .05 y el .01.

B) El número de **grados de libertad** ('gl', en nuestra tabla),³³ para una tabla de contingencia de 'h' hileras por 'c' columnas, se obtiene mediante la simple fórmula:

$$g.l. = (h-1) \cdot (c-1)$$

En una tabla de contingencia, los grados de libertad pueden interpretarse intuitivamente de la manera siguiente: dado un conjunto de marginales, los grados de libertad representan el número mínimo de celdas que es suficiente llenar para que queden determinadas todas las frecuencias condicionales restantes. Para toda tabla de 2×2 tendremos:

$$(2-1) \cdot (2-1) = 1 \text{ gl}$$

Lo que significa que una vez que se conoce la frecuencia condicional para una de las cuatro

³¹ Obsérvese que los números de esta tabla están escritos de acuerdo a la notación inglesa en la que el punto '.' cumple el papel de la coma y viceversa; así, '.01' equivale en notación española a '0,01'. Presentamos la tabla con la notación inglesa porque es la que se encontrará en los manuales de estadística, incluso en su traducción castellana. Hemos resumido la tabla limitándola a niveles de significación mayores que .20 y a 10 grados de libertad: para tablas que superen el formato 4×4 -o 6×3 -, remitirse a una tabla más completa en cualquier manual de estadística.

³² Estrictamente hablando, en éste al igual que en todos los casos de inducción estadística, se debe razonar en términos del rechazo o aceptación de la **hipótesis nula**. En el ejemplo, la hipótesis nula sostendría: «No existen diferencias en el grado de participación según el grado de conocimiento». En efecto, la verdad de la conclusión no garantiza lógicamente la verdad de las premisas mientras que, en cambio, de la falsedad de la conclusión sí es posible **deducir** la falsedad de las premisas. En Estadística, la "falacia de afirmar el consecuente", que consiste en no descartar una hipótesis falsa, recibe el nombre de 'Error de Tipo II'; en tanto que el riesgo contrario -el de descartar una hipótesis cierta, que deviene de la introducción del elemento de la probabilidad- se denomina 'Error de tipo I' (cf. Blalock, 1966: 106 y ss.).

³³ En las tablas de los manuales se encontrará 'df', del inglés '*degree of freedom*'.

celdas, se puede calcular las frecuencias de las otras restantes mediante simples sumas y restas. Si se tratara de una tabla de 3 x 3, tendríamos

$$(3-1) \cdot (3-1) = 4 \text{ gl, etc.}$$

Supongamos que para nuestro ejemplo de la Tabla 4.4 consideramos adecuado un nivel de .01. Buscamos en la cabecera de la Tabla 4.6 ('Probabilidad') la columna '.01' y en 'gl' la hilera 1. El valor crítico que corresponde a la combinación de $p=.01$ y $gl=1$, resulta ser: 6,635. Como el valor de χ^2 para nuestra tabla (4,450) es menor, deberemos concluir que la asociación entre las dos variables **no** es significativa al nivel de .01. O, si se quiere, hay más de un 1% de chances de que la relación de las dos variables en la tabla sea fruto del mero azar implícito en la selección de los casos de nuestra muestra y que, por ende, no esté reflejando una relación real existente en la población. Si, en cambio, nos fijáramos un nivel de .05, el valor crítico sería de 3,841 (ver Tabla), menor que el registrado para la Tabla 4.4, y deberíamos concluir que existe una relación estadísticamente significativa en ese nivel.

Resumiendo, para utilizar χ^2 los *pasos a seguir* serán:

- 1) Fijarnos un nivel de significación;³⁴
- 2) Determinar el número de grados de libertad de la tabla a analizar;
- 3) Calcular χ^2 para dicha tabla;
- 4) Comparar el valor de χ^2 en esta tabla con el valor crítico 'VC' en la Tabla del χ^2 :

si $\chi^2 > VC$, se concluye que la relación es estadísticamente significativa.

Si en cambio se da el caso contrario ($\chi^2 < VC$), habrá que concluir que la relación no es estadísticamente significativa al nivel que se había fijado. No hay por ello razón para desesperar; antes bien, «los fracasos en encontrar relaciones significativas allí donde se espera hallarlas son siempre interesantes y con frecuencia pueden ser mucho más productivos que el logro del resultado previsto. Lo inesperado es lo que más nos obliga a pensar» (Erikson y Nosanchuk, 1979: 253).

Desde un punto de vista metodológico ortodoxo, la manera correcta de proceder es: *primero* fijarse un nivel de significación, y *después* constatar si la relación es estadísticamente significativa en ese nivel. Empero, lo cierto es que en la práctica del procesamiento de encuestas la difusión del PC y de los *software* estadísticos que calculan automáticamente el valor de χ^2 para cualquier tabla, no ha hecho más que acentuar la tendencia a determinar primero el valor de χ^2 para los datos, para luego constatar el nivel de significación alcanzado.

Así, en cualquier paquete estadístico el comando '*contingency table*' (o equivalente) arrojará en la pantalla, en general previamente a la tabla misma, una serie de valores semejante a ésta:

$$< \text{DF} = 1 ; \chi^2 = 4,445 ; p = .035 >$$

Vale decir, el programa brinda una probabilidad *exacta*, '.035', indicando así que hay un 3,5% de probabilidades de que la asociación observada obedezca al puro azar. Y éste será el resultado consignado en el informe: «la relación entre las variables A y B es significativa al nivel de $p=.035$ ».

Lo cierto es que la rígida distinción entre lo exploratorio y lo confirmatorio, entre el 'contexto del descubrimiento' y el 'contexto de la prueba' -distinción que Popper popularizara inspirándose en Reichenbach- es más el precepto de una epistemología que se autoerige en normativa que un principio efectivamente puesto en obra en la investigación empírica.

¿Cuándo se puede usar χ^2 ?

Como cualquier test de significación, χ^2 requiere para ser utilizado el cumplimiento de ciertas *condiciones*:

- 1) Los datos deben consistir en *mediciones independientes de casos seleccionados al azar*. Así, la utilización del test para datos provenientes de una muestra no-probabilística constituye un

³⁴ La elección de un nivel de significación es una decisión esencialmente arbitraria, ya que no puede deducirse ni de la teoría sustantiva ni de la estadística: «Puesto que es imposible establecer los costos de una decisión incorrecta con respecto a las teorías científicas, la elección del nivel de significación se realiza sobre una base subjetiva y arbitraria, incluso si el nivel elegido es alguno de los convencionales, .05 o .01» (Henkel, 1976: 77-78).

sin sentido. Idealmente, debería tratarse de una muestra al azar simple; o al menos de aproximaciones razonables, como muestras al azar simples con bajas tasas de no-respuesta, o bien extraídas de listados incompletos pero que contengan al menos algunos casos de todos los tipos que sean relevantes para el estudio, etc.

2) El número de casos debe ser lo suficientemente grande, lo que se expresa en una relación determinada entre el tamaño de las frecuencias esperadas y el número de celdas de la tabla. A los efectos prácticos, aplicar las siguientes reglas: en tablas de 2×2 , las frecuencias *esperadas* no podrán ser inferiores a 10; en tablas de mayor formato, para un nivel de .05 bastará con un mínimo de 6 casos por celda, y algo más para niveles de significación más elevados (cf. Henkel, 1976: 77-81).

¿Para qué sirve y para qué no sirve χ^2 ?

El test de χ^2 puede ser de utilidad para determinar si una relación entre dos variables es *estadísticamente* significativa. Conviene tener siempre presente que la significación estadística no es sinónimo ni de ‘relevancia teórica’, ni de ‘importancia práctica’. De hecho, el nivel de significación se encuentra dependiendo en relación directa de dos factores: la *fuerza* de la relación entre las variables, y el tamaño de la muestra. Se entiende entonces cómo en muestras muy grandes χ^2 puede producir un valor estadísticamente significativo aunque la relación entre las variables sea muy débil.

Por esta misma razón χ^2 no permite afirmar nada acerca de la fuerza de la relación. Un sencillo ejemplo ilustra este punto. Imaginemos que en vez de contar con una muestra de 121 estudiantes hubiéramos seleccionado 1210 alumnos para nuestra investigación; y que nos encontráramos con la misma diferencia del 14% entre los alumnos de Bajo y Alto grado de conocimiento político. Obtendríamos:

	B	A	
A	60	130	190
B	590	630	1020
	650	560	1210

El cálculo de χ^2 arrojaría entonces un valor de 44,5. Vale decir que, al multiplicar por 10 todas las frecuencias, el valor de χ^2 queda a su vez multiplicado por 10. Cuando trabajábamos con 121 casos, la relación entre las variables no era estadísticamente significativa al nivel de .01; ahora observamos que supera con creces el valor crítico para .001 (10.827). Y sin embargo, en ambos casos encontramos la misma diferencia porcentual de 14%. En suma, resulta fácil concluir que χ^2 *no sirve para medir la fuerza de la asociación*, porque su valor varía en función de n .

2.3. Algunos coeficientes de asociación

Ya hemos observado, al presentar la diferencia porcentual como medida de la asociación en una tabla, las razones por las que ésta no resultaba adecuada como medida resumen única de la asociación. Otro modo de medir la fuerza de la relación entre dos variables es utilizando un coeficiente de asociación.

¿Qué es un coeficiente de asociación?

A nuestros efectos, bastará con decir que es una fórmula construida de una manera tal que, al aplicarla a un conjunto de datos, arroja un valor único que puede ser interpretado como una medida del grado en que dos variables se encuentran relacionadas. Estadísticamente, se pretende en general que un coeficiente de asociación esté construido de tal modo que produzca siempre un valor comprendido entre 1 y 0. Esto facilitará considerablemente la interpretación del resultado: ‘0’ indicará la ausencia total de relación, mientras que ‘1’ estará mostrando asociación perfecta entre las variables. Un coeficiente de asociación mide entonces la fuerza o el grado de relación existente entre dos variables.

Coefficientes derivados de χ^2

Existen varios coeficientes de asociación que se derivan de χ^2 . Así, puesto que se sabe que χ^2 varía

directamente en función de n , la solución más simple parece ser la de estandarizar el valor de x^2 dividiéndolo por el número de casos en la tabla. Se obtiene así ϕ^2 (fi-cuadrado):³⁵

$$\phi^2 = \frac{x^2}{n}$$

Con los datos de la Tabla 4.4 obtendríamos:

$$\phi^2 = \frac{4,45}{121} = 0,037$$

Queda claro que ϕ^2 se independiza de n ; si se hubiera trabajado con 1210 estudiantes su valor hubiera sido el mismo. Para cualquier tabla del formato 2×2 , o de $h \times 2$, o de $2 \times c$, ϕ^2 variará entre 0 y 1. Pero, lamentablemente, esta interesante propiedad se pierde cuando ambas dimensiones de la tabla presentan más de dos valores: el límite máximo resulta en estos casos mayor que 1.

Para solucionar este inconveniente se idearon otros coeficientes de asociación que son simples funciones de n ; así, en 1919, Tschuprov definía:

$$T^2 = \frac{\phi^2}{\sqrt{(h-1)(c-1)}}$$

Sin embargo, aunque el límite superior de T es 1, este valor sólo puede alcanzarse para tablas con igual número de hileras que de columnas (en las que $h=c$); si el formato no es cuadrado, el valor de T será siempre inferior a la unidad.

Ya en 1946, el coeficiente V de Cramer superó este defecto:

$$V^2 = \frac{\phi^2}{\text{Min}(h-1; c-1)}$$

‘Min ($h - 1$; $c - 1$)’ significa aquí que se tomará o bien ‘ $h - 1$ ’, o bien ‘ $c - 1$ ’, el que resulte menor de estos dos números. Así, para una tabla de 6×4 , se tendría $\text{Min} (h - 1; c - 1) = \text{Min} (6-1; 4-1) = 3$, y por lo tanto:

$$V^2 = \frac{\phi^2}{3}$$

Comoquiera que V es susceptible de alcanzar el valor ‘1’ para tablas de cualquier formato, esto la hace una medida preferible a T .

Coeficientes para la tabla de 2×2

En lo que hace a la tabla de 2×2 , una rápida inspección de las fórmulas de los coeficientes presentados debería bastar para comprender que, *en este caso particular*:

$$\phi^2 = T^2 = V^2$$

Para este formato de tabla hay dos coeficientes de asociación de uso muy frecuente, que son la Q de Yule ³⁶ y ϕ . Remitiéndonos a la notación habitual para esta tabla, sus fórmulas son las

³⁵ La letra griega ‘ ϕ ’ se pronuncia ‘Fi’.

³⁶ Introducida por el estadístico británico G.U. Yule ya en 1900, y que a veces aparece denominada como ‘ Q de Kendall’.

siguientes:

$$Q = \frac{bc - ad}{bc + ad} \quad \phi = \frac{bc - ad}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$$

Aplicando estas fórmulas a los datos de la Tabla 4.4 , se obtiene:

$$Q = \frac{(13 \times 59) - (6 \times 43)}{(13 \times 59) + (6 \times 43)} = \frac{767 - 258}{767 + 258} = \frac{509}{1025} = 0,50$$

$$\phi = \frac{(13 \times 59) - (6 \times 43)}{\sqrt{19 \times 102 \times 65 \times 56}} = \frac{767 - 258}{\sqrt{7054320}} = \frac{509}{2656} = 0,19$$

¿Cómo interpretar el valor de un coeficiente?

Indicando ‘1,00’ (o ‘-1,00’) una correlación perfecta, se suele hablar de una ‘correlación ‘muy fuerte’ para valores superiores a 0,80, de una ‘fuerte’ correlación entre 0,60 y 0,80, ‘moderada’ de 0,40 a 0,60, y ‘débil’ de 0,20 a 0,40 . Por debajo de 0,20, probablemente no exista correlación, y se trate simplemente de un producto del azar -al menos que se cuente con una muestra muy grande- (Fitz-Gibbon, 1987: 82). Si aceptamos este modo de interpretar los valores de nuestros coeficientes, nos encontramos con que, en tanto Q indicaría una asociación moderada entre las variables, de acuerdo a ϕ no existiría ni siquiera una asociación débil. El valor de Q es mucho mayor que el de ϕ . De hecho, esto ocurrirá toda vez que calculemos ambos coeficientes; es una **ley**: para cualquier tabla de 2 x 2 se cumple siempre que $Q > \phi$.³⁷

Ello no debe sorprendernos , puesto que se trata de dos coeficientes diferentes. Mientras que ϕ se deriva de χ^2 -como se ha visto-³⁸ y puede entenderse como un caso especial del coeficiente producto-momento de Pearson, por su parte Q no es más que una aplicación del coeficiente γ (Gamma) de Goodman y Kruskal al caso particular de la tabla de 2 x 2 (Reynolds, 1977).

¿Cuál elegir: Q o ϕ ?

Hay que evitar caer en la tendencia espontánea a utilizar sistemáticamente Q debido a que arroja un valor mayor: Q sería “más gratificante”, nos permitiría “encontrar” relaciones entre las variables más frecuentemente que ϕ . Tampoco hay que caer en el pudor inverso: utilizar ϕ para estar seguro de que “realmente” existe asociación entre las variables. Lo cierto es que la elección deberá basarse en el modelo de asociación “perfecta” que sea el más adecuado para una hipótesis determinada.

En efecto, existen dos modos alternativos de definir la asociación perfecta en una tabla de 2 x 2. Para esclarecer este punto, consideremos en qué condiciones los coeficientes Q y ϕ van a alcanzar el valor ‘1,00’. Supongamos que hemos cruzado los atributos X e Y:

Ejemplo de asociación completa				
		Atributo X		
		B	A	
Atributo Y	A	20	50	70
	B	30	0	30
Q= 1,00		50	50	100

Para que el coeficiente Q alcance su valor máximo, basta con que no se registre ninguna frecuencia en cualquiera de las cuatro celdas. Como en este ejemplo hay 0 casos en la celda d, ello es suficiente para que Q indique una asociación perfecta.

³⁷ Excepto en un caso: los valores de Q y ϕ serán iguales (‘1,00’) cuando se dé una asociación perfecta entre las variables.

³⁸ Se trata, como se puede sospechar, de la raíz cuadrada de ϕ^2 .

Ejemplo de asociación absoluta				
		Atributo X		
		B	A	
Atributo Y	A	0	70	70
	B	30	0	30
$\phi = 1,00$		30	70	100

Este segundo ejemplo permite ver cómo ϕ es más exigente que Q . Para alcanzar el valor 1,00, ϕ demanda que una de las diagonales registre 0 frecuencias en ambas celdas (como ocurre con a y d, en el cuadro). Además, para que esta condición pueda cumplirse, es necesario que ambos marginales sean simétricos (en ambos atributos, observamos 70 casos con el valor 'sí', y 30 con 'no').

Se debe distinguir, entonces entre dos modelos de 'asociación':

1) Existe ASOCIACION *COMPLETA* cuando todos los casos que presentan un valor en una variable tienen un mismo valor en la segunda variable, pero sin que todos los que tienen este valor en la segunda tengan el mismo valor en la primera (dicho de otro modo: « todos los X son Y, pero no todos los Y son X»; o bien: «todos los X son Y, pero no todos los no-X son no-Y»).

2) Hablaremos de ASOCIACION *ABSOLUTA* cuando la relación es bi-unívoca, es decir cuando todos los casos con un valor en una variable presentan el mismo valor en la otra, y viceversa («Todos los X son Y, y todos los Y son X»).³⁹

Si se toma como paradigma de medida de la asociación al coeficiente de producto-momento de Pearson, se deberá elegir ϕ en todos los casos.⁴⁰ Pero también cabe pensar que de las teorías sociológicas existentes es posible deducir hipótesis de dos tipos: a) las que aseveran que debe haber una celda nula, postulando asociación completa entre las variables; y b) las que afirman la existencia de dos celdas nulas, y que predicen asociación absoluta.

Desde este punto de vista, Fernando Cortés desarrolla un ejemplo acerca de la relación entre la clase social paterna y la ocupación del hijo en contextos societales diferentes. Así, en una sociedad tradicional, cabría esperar teóricamente que todos los hijos de clase alta se desempeñaran en ocupaciones no manuales, y que todos los insertos en ocupaciones no manuales provinieran de padres de clase alta. De ser esta la hipótesis, ϕ sería el coeficiente más adecuado, puesto que se postula que las dos variables están ligadas por una relación de implicación recíproca. Por lo contrario, en una sociedad "moderna", con un margen cierto de movilidad social, probablemente no existieran muchos hijos de clase alta en ocupaciones manuales, pero los no manuales no se reclutarían sólo en la clase alta; si se sostiene esta hipótesis, convendrá utilizar la Q de Yule. En suma, desde esta perspectiva, el coeficiente más apropiado se elegirá en función del tipo de hipótesis que se esté investigando.

¿Qué hacer con el signo de ϕ y Q ?

Un último punto a desarrollar con relación a estos coeficientes se refiere al signo positivo o negativo que pueden asumir los valores que arrojen. De hecho, ambos presentan un rango de variación que va de +1 a -1, pasando por 0. Tanto +1 como -1 indican asociación perfecta, mientras que 0 supone la ausencia total de relación entre las variables. Se habla entonces de 'asociación negativa' o 'positiva', según sea negativo o positivo el valor resultante. Lo cierto es que estos

³⁹ Las expresiones de 'asociación completa' y 'absoluta' pertenecen a F. Cortés. Boudon muestra la relación estos modelos de asociación con las nociones lógicas de 'implicación simple' [si X entonces Y] y de 'implicación recíproca' [si x entonces Y, y si Y entonces X] (cf. Boudon, 1971: cap. 2). Por su parte, Mora y Araujo hablaba, más gráficamente, de relaciones 'curvilineales' y 'lineales', que resultan en distribuciones de frecuencias de tipo 'rinconal' y 'diagonal' (1965:5).

⁴⁰ Blalock parece inclinarse por esta alternativa, con la siguiente salvedad: si los marginales de una variable dependiente continua se han determinado arbitrariamente -por ejemplo, dicotomizando desde la mediana-, y como resultado de ello ambos marginales no son simétricos, ϕ no tiene posibilidad alguna de alcanzar el valor 1 y es conveniente emplear Q (1966: 257).

coeficientes pueden aplicarse a variables definidas en cualquier nivel de medición.⁴¹ Ahora bien, tratando con variables de nivel nominal deberá prescindirse totalmente del signo en la interpretación del resultado; ello debido precisamente a la inexistencia de un orden intrínseco de sus categorías, por lo que la mera permutación de las hileras (o de las columnas, pero no de ambas) redundaría en el cambio de signo del valor arrojado por el coeficiente. Si en cambio se trata de dos variables ordinales, el signo tendrá un significado claro del tipo «A más X, más Y», por cuanto indicará la dirección de la relación; en este caso habrá entonces que precaverse de que la ordenación de las categorías de ambas variables sea congruente. En efecto, supongamos que cruzamos dos variables con los valores ‘Alto’ y ‘Bajo’:

	B	A		A	B
A	n	N	A	n	N
B	N	n	B	N	n

Asociación positiva

Asociación negativa

En la tabla de la izquierda, hay asociación positiva (‘N’ simboliza las frecuencias mayores dentro de la fila o de la columna respectiva); pero en el cuadro de la derecha, en cambio, la asociación es negativa, por más que las frecuencias parezcan distribuidas de modo similar; sucede que en este último caso se ha alterado el **orden** de disposición de los valores de la variable-columna.

Dicotomizar abusivamente es riesgoso

Si se tiene presente que entre dos puntos cualesquiera siempre es posible trazar una recta, se comprenderá cómo los resultados basados en la relación entre dos dicotomías pueden ser engañosos. Por supuesto, esta salvaguarda no es aplicable cuando se trabaja con auténticas dicotomías; si, por ejemplo, queremos relacionar con el sexo el haber votado o no en una elección, es evidente que no existe ninguna alternativa fuera de la tabla de 2 x 2. Pero distinto es el caso cuando se trata de variables de un nivel superior al nominal que han sido **reducidas** a dicotomías.⁴² Ya sea por contar con un número reducido de observaciones en la muestra, o por que se espera encontrar de este modo alguna relación, lo cierto es que con frecuencia el investigador se ve llevado a dicotomizar sus variables.

Un sencillo ejemplo permite ilustrar convenientemente los riesgos que entraña el abuso de la dicotomización. Supongamos que se quiere determinar si el interés en la política que manifiestan estudiantes universitarios depende en alguna medida de la etapa en que se encuentran de su carrera.

Tabla 4.7a : Interés de estudiantes en la política según etapa de la carrera en que se encuentran

Interés en la política	Etapa en la carrera		Total
	Inicial	Superior	
Alto	85	85	170
Bajo	65	65	130
Total	150	150	300

Fuente: elaboración propia.

En la Tabla 4.7a se cuenta con datos imaginarios para una muestra de 300 estudiantes. La variable independiente aparece dicotomizada, y se observa que en ambas categorías el porcentaje

⁴¹ Aunque, como se verá, para variables intervalares su uso entrañará una gran pérdida de información.

⁴² Puesto que en los niveles superiores de medición son aplicables todas las propiedades propias de los niveles inferiores, incluido el nominal, cualquier variable puede reducirse a una dicotomía, definida en términos de ausencia/presencia de una característica.

de estudiantes interesados en la política es idéntico (57%): se concluye que no existe relación alguna entre las variables.

La Tabla 4.7b presenta los mismos datos, con la única diferencia de que la variable independiente ha sido tricotomizada: mientras que en las nuevas categorías “Inicial” y “Superior” hay un 40% de estudiantes politizados, en la categoría intermedia el porcentaje es del 90%: se deberá concluir que existe un efecto de la fase en los estudios sobre el interés por la política. Sucede que este efecto -el aumento del interés de la fase inicial a la media, seguido de un descenso de la fase media a la superior- se encontraba totalmente disimulado por la dicotomización de la variable que repartía por igual los 100 estudiantes ‘intermedios’ entre las dos categorías extremas. Por ende, siempre que sea posible, es conveniente no dicotomizar las variables ordinales y conservar cuanto menos tres valores.⁴³

*Tabla 4.7.b : Interés de estudiantes en la política
según etapa de la carrera en que se encuentran*

<i>Interés en la política</i>	<i>Etapa en la carrera</i>			<i>Total</i>
	<i>Inicial</i>	<i>Media</i>	<i>Superior</i>	
<i>Alto</i>	40	90	40	170
<i>Bajo</i>	60	10	60	130
<i>Total</i>	100	100	100	300

Fuente: elaboración propia.

El coeficiente Gamma de Goodman y Kruskal

Para el caso general de la tabla de $h \times c$, existen muchos coeficientes, además del V de Cramer.⁴⁴ Uno de los más conocidos es el Gamma (γ) de Goodman y Kruskal, que se utiliza para tablas basadas en la combinación de dos variables ordinales.

La fórmula para el cálculo de γ es la siguiente:

$$\frac{n_i - n_d}{n_i + n_d}$$

En la que n_i significa el número de pares de casos con iguales órdenes en ambas variables; mientras que n_d representa la cantidad de comparaciones en las que los órdenes resultan discordantes. Por cierto, es necesario comprender con toda claridad qué se entiende por órdenes iguales y órdenes distintos:

⁴³ Más en general, Hyman demuestra lúcidamente los peligros que entraña considerar una variable independiente de nivel ordinal o superior sin tomar en cuenta la totalidad del recorrido de dicha variable. Si los valores en una parte del recorrido no producen diferencias, ello no significa necesariamente que no exista relación con la variable que la hipótesis postula como dependiente: la relación puede darse con los valores correspondientes a la parte no incluida (ya sea posteriores o anteriores) del recorrido de la variable (cf. Hyman, 1971: 236 y ss.).

⁴⁴ Una exposición sistemática sobre todos los coeficientes para tablas de $h \times c$ está más allá de los objetivos de este trabajo de carácter introductorio. El lector interesado podrá consultar a Galtung (1966: II), y Zelditch (1959), así como los trabajos más recientes de Hildebrand, Laing y Rosenthal (1977), Liebetrau (1985), y Leach (1979).

	B	M	A
A	a	b	c
M	d	e	f
B	g	h	i

Comparando dos casos cualesquiera ubicados en distintas celdas pueden darse tres situaciones: (1) que sus órdenes sean iguales (cualquier caso en g será más bajo en ambas variables que un caso de e, b, c o f); (2) que estén parcialmente empatados (casos ubicados en una misma hilera –como a y b, o en una misma columna –e y h); (3) que sus órdenes sean distintos (así, comparando un caso de a con otro de e, f, h, o bien i, se da que el primero es más alto que los segundos en la variable-hilera, pero más bajo que éstos en la variable-columna).

Por razones de simplicidad, nos basamos en la tabla de 3 x 3, aunque está claro que este coeficiente puede aplicarse a cualquier tabla de h x c. Gamma razona comparando para cada par de casos los órdenes mantenidos en ambas variables. El procedimiento para el cálculo se especifica a continuación.

a. cálculo de los órdenes iguales: n_i

Para calcular los pares de casos que mantienen igual orden en ambas variables, se multiplica la celda rayada (g) por las punteadas (b+c+e+f); luego se continúa multiplicando las frecuencias en cada celda por las celdas ubicadas hacia arriba y hacia la derecha: d (b+c), h (f+c), y e (c).

	B	M	A
A	a	b	c
M	d	e	f
B	g	h	i

	B	M	A
A	a	b	c
M	d	e	f
B	g	h	i

b. cálculo de los órdenes distintos: n_d

Para computar los pares de casos que tienen órdenes distintos en ambas variables, se multiplica la celda rayada (a) por las punteadas (e+h+f+i); luego, se continúa multiplicando las frecuencias en cada celda por las ubicadas hacia abajo y hacia la derecha.

Podemos ahora aplicar γ a un ejemplo.

Tabla 4.8: Grado de conocimiento político según Nivel económico-social

Grado de conocimiento político	Nivel económico-social			Total
	Bajo	Medio	Alto	
Alto	1	13	18	32
Medio	15	49	26	90
Bajo	24	11	5	40
Total	40	73	49	162

Fuente: elaboración propia.

El cuadro se basa en datos inéditos de una encuesta realizada a electores de Posadas en el año 1987. Cada encuestado fue clasificado simultáneamente en un índice de conocimiento político y en un índice de nivel económico-social. La acumulación de frecuencias en la diagonal principal parecería sustentar una relación del tipo « A mayor NES, mayor conocimiento político». El coeficiente V de Cramer arroja una asociación débil de 0,37.

Desarrollemos el cálculo de n_i :

En la última hilera tenemos: $24 (49 + 13 + 26 + 18) = 2544$

Siempre en la última hilera: $11 (26 + 18.) = 484$
 Luego, en la segunda hilera:⁴⁵ $15 (13 + 18) = 465$
 Por último: $49 (18) = 882$
 La suma de las cuatro cantidades es $\Sigma n_i = 4375$

Con análogo procedimiento calculamos los n_d :

$1 (49 + 26 + 11 + 5) = 91$
 $13 (26 + 5) = 403$
 $15 (11 + 5) = 240$
 $49 (5) = 245$
 $\Sigma n_d = 979$

Tenemos entonces: $\gamma = \frac{4375 - 979}{4375 + 979} = 0,63$

Gamma es tan “rinconal” como Q;⁴⁶ γ está basado en una definición de asociación completa y no absoluta, por lo que en cualquiera de estas situaciones alcanzará su máximo de 1 (o de -1). Por otra parte, se trata de una medida simétrica de asociación; esto es, no presupone que alguna de las dos variables sea la independiente.⁴⁷ Gamma se presta a una interpretación probabilística relativamente simple: tomando dos casos cualesquiera, el valor de γ significa la probabilidad de que si un caso presenta un valor más alto que el otro en una primera variable, también lo supere en la segunda variable. Al tratarse de variables ordinales, se debe tomar siempre en cuenta el signo; en nuestro ejemplo, se puede concluir que hay un 63% de probabilidades de que si un elector es más alto que otro en conocimiento político, también resulte más alto en su nivel económico-social (o viceversa). Si se quisiera, en cambio, determinar el efecto que tiene el NES sobre el grado de conocimiento político, convendría recurrir a alguna otra medida de carácter asimétrico.⁴⁸

En suma, lo importante es tener en cuenta siempre que cualquier coeficiente responde a un particular modelo de asociación; por ello, constantemente se están elaborando nuevos coeficientes y reinterpretando los modelos ya existentes. Para elegir un coeficiente de asociación se deberá por lo tanto tener en cuenta: a) el nivel de medición de las variables; b) si se quiere analizar la tabla de modo simétrico o asimétrico; y c) el modelo de asociación requerido por la hipótesis.

3. DISTRIBUCIONES MULTIVARIANTES CONJUNTAS

La tabla de contingencia, siendo una forma típica en que se presentan los datos en la investigación social, está lejos de agotar todas las posibilidades. La distribución multivariante conjunta, de la que nos ocuparemos en este apartado, es un modo alternativo de desplegar los datos, cuya utilidad es evidente cuando se trabaja con un número reducido de UUAA y con variables de un nivel superior al nominal.

En un conocido trabajo de Leopoldo J. Bartolomé se sustenta la hipótesis de que «los colonos misioneros que conformaron la base de sustentación del MAM (Movimiento Agrario Misionero) pueden ser gruesamente caracterizados como integrantes de una pequeña burguesía rural de propietarios de explotaciones medianas» (Bartolomé, 1982: 47). Los datos presentados en la siguiente tabla tienden a abonar dicha hipótesis.

En la columna (a) se consigna el número de núcleos del MAM para cada departamento de Misiones; ese número puede ser considerado como un indicador válido del grado de implantación alcanzado por el movimiento en cada área. Por su parte, las columnas (b), (c) y (d) se refieren a

⁴⁵ La última celda de la tercera hilera (la que indica '5') no tiene otras celdas ubicadas hacia arriba y hacia la derecha, por ende hay que continuar con la segunda hilera. Lo mismo se aplica para este cálculo a las celdas de la primera hilera.

⁴⁶ Como ya se ha dicho, Gamma es Q, cuando se lo aplica a una tabla de 2 x 2. En este caso, la única celda arriba y a la derecha de c es b; la única abajo y a la derecha de a, es d; por lo tanto hay que computar bc - ad.

⁴⁷ Si se requiere, por las características de la hipótesis, una medida simétrica basada en la idea de asociación absoluta, se puede recurrir a la t (tau) de Kendall.

⁴⁸ Por ejemplo, el d (delta) de Sommers.

indicadores agrarios contruidos en base a datos censales.

Tabla 4.9: Número de núcleos de base del MAM por departamento e indicadores agrarios seleccionados

<i>Departamentos</i>	<i>Nº de núcleos en 1973</i>	<i>% explot. de 5 a 50 has. en Depto.</i>	<i>% explot. de 5 a 50 has. en Provincia</i>	<i>% mano de obra familiar</i>
	<i>(a)</i>	<i>(b)</i>	<i>(c)</i>	<i>(d)</i>
<i>Oberá</i>	44	85,6	17,7	78,4
<i>Caingúas</i>	37	88,4	18,9	86,9
<i>Leandro N. Alem</i>	20	87,1	11,3	89,3
<i>Apóstoles</i>	17	65,1	4,6	s.d.
<i>Montecarlo</i>	16	66,5	3,0	50,8
<i>San Ignacio</i>	15	63,8	7,1	66,8
<i>L.G. San Martín</i>	12	75,8	5,9	82,9
<i>Veint. de Mayo</i>	12	79,0	6,7	91,5
<i>Guaraní</i>	10	64,0	4,0	87,8
<i>San Javier</i>	8	80,5	4,9	86,4
<i>San Pedro</i>	6	60,3	1,5	83,2
<i>Eldorado</i>	5	70,1	6,1	57,2
<i>Iguazú*</i>	3	55,5	1,4	40,0
<i>Candelaria</i>	0	67,0	2,9	79,2
<i>Concepción</i>	0	48,5	1,5	79,3
<i>General Belgrano</i>	0	63,6	1,7	84,9
<i>Capital</i>	0	21,0	0,8	68,1

* Los tres núcleos que aparecen en Iguazú pertenecen en realidad al área de influencia de Eldorado.

(a) Informe de prensa del MAM (El Territorio, 26-x-74, p.9).

(b), (c) y (d) Datos del Censo Nacional Agropecuario de 1969.

Fuente: Bartolomé, L., 1982: 33.

La distribución multivariante tiene exactamente la misma forma de la matriz de datos. Cada renglón corresponde a una UA -un departamento de Misiones, para el caso- y en las columnas aparecen diferentes variables. Cuando el investigador elige esta forma para presentar sus datos,⁴⁹ es porque desea destacar ciertas relaciones entre las variables. En la tabla se comprueba a simple vista una tendencia bastante clara: cuanto mayor es el número de núcleos del MAM, mayores son los porcentajes en las columnas restantes.

Aunque en verdad es preferible invertir los términos: en el análisis, el número de núcleos del MAM debe funcionar como variable dependiente, puesto que la hipótesis busca explicar por qué el MAM se desarrolla más en ciertas áreas que en otras, debido a diferencias en la estructura agraria.

Por otra parte, no todos los indicadores agrarios presentados son igualmente interesantes para relacionarlos con el número de núcleos del MAM. Consideremos la columna (b): ¿Podría tomarse el % de explotaciones medianas calculado sobre el total *departamental* como variable independiente? Esta elección parecería escasamente coherente. ¿Cómo relacionar un porcentaje calculado sobre el total de explotaciones de cada departamento, con la distribución de todos los núcleos de la *Provincia* por departamento?

Desde un punto de vista lógico, hay una falta de homogeneidad entre ambos indicadores. Un departamento podría muy bien aparecer en la columna (b) con un elevado porcentaje de explotaciones medianas sin que necesariamente fuera esperable que presentara un gran número de núcleos del MAM, dato este último que estaría dependiendo a la vez de la cantidad de

⁴⁹ Y no simplemente para trabajarlos.

explotaciones agrícolas de todo tamaño existente en ese departamento. El mismo razonamiento es aplicable a la columna (d), en la que el peso relativo de la mano de obra familiar en las explotaciones está calculado sobre los totales departamentales.

Si la hipótesis del autor fuera verdadera, cabría esperar que cuanto mayor fuera el número absoluto de explotaciones agrícolas medianas, mayor debería ser la cantidad de núcleos en la columna (a); nótese, en cambio, que en este razonamiento es por completo irrelevante que las cantidades en (c) sean frecuencias absolutas o porcentajes sobre el total provincial. Por ende, lo más coherente, es ver en qué medida (a) está dependiendo de (c).

¿Por qué no utilizar la tabla de contingencia ?

Ciertamente con estos datos es posible elaborar una tabla de contingencia. Así, tomando las columnas (a) y (c) obtenemos la siguiente tabla de 2 x 2 :

Tabla 4.9.1: Número de núcleos del MAM según porcentaje de explotaciones medianas

Nº de núcleos	% de expl. medianas		
	Bajo	Alto	
Alto	2	7	9
Bajo	6	2	8
	8	9	17

Fuente: Tabla 4.9

Para construir esta tabla ambas variables han sido dicotomizadas por la mediana. Los departamentos “Altos” son, respectivamente, los que cuentan con un número de diez o más núcleos del MAM, y los que concentran un porcentaje de las explotaciones medianas provinciales superior al 4,5%. Una simple inspección de la tabla revela una relación de forma diagonal; calculando, se obtiene un ϕ de 0,53, que denota una clara asociación positiva entre ambas variables.

Este proceder conlleva el grave inconveniente de la pérdida de información que acarrea. En la tabla de contingencia, entre los departamentos “Altos” en número de núcleos se ubicarán tanto Guaraní (10) como Oberá (44), cuando en el segundo la cantidad más que cuadruplica la del primero. Además, observando estos mismos departamentos en las columna (c) de la tabla de distribución multivariante se percibe rápidamente que una relación numérica muy similar se observa también para esta otra variable: “17,7” representa un poco más que el cuádruple de “4,6”. Nuestra inquietud debería acrecentarse al notar que mientras Oberá es “Alto-Alto”, Guaraní es en cambio “Bajo-Alto”: ¿no deberían ambos departamentos ubicarse en la misma diagonal principal de la tabla de contingencia?. Se podría cuestionar entonces el criterio utilizado para dicotomizar las variables, aunque lo cierto es que con 17 casos resultaría casi absurdo proceder de otro modo.⁵⁰

En suma, la pérdida de información que supone reagrupar en grandes conjuntos a las UUAA no se limita a una posible arbitrariedad o a una falta de precisión de sus categorías. Ocurre que se pierde también información referida a las relaciones existentes entre los valores: desaparece la íntima relación cuya existencia podíamos intuir entre los valores originales de ambas variables.

El coeficiente ρ de Spearman

En la Tabla 4.9.2, partiendo de una distribución bivalente conjunta, se desarrolla un procedimiento

⁵⁰ En ausencia de un criterio teórico, hay sólo dos formas de “cortar” una variable: por el lado de los valores, o por el lado de las UUAA (como se lo ha hecho en este caso). Para dicotomizar el número de núcleos a partir de los valores, se hubiera debido promediar los valores extremos de la variable para fijar el punto de corte: así, $0 + 44 = 22$, con lo cual sólo Oberá y Cainguás aparecerían como “altos” en la variable dicotomizada.

que permite construir una medida de la relación entre las variables que supera parcialmente estos defectos.

Tabla 4.9.2: Ejemplo de distribución bivalente conjunta y construcción de un coeficiente de correlación por rangos

Departamentos	Nº núcleos en 1973 (I)	% explot. 5-50 has. (II)	Rango en (I) (III)	Rango en (II) (IV)	d (V)	d ² (VI)
Oberá	44	17,7	1	2	-1	1
Caingúas	37	18,9	2	1	1	1
Leandro N. Alem	20	11,3	3	3	0	0
Apóstoles	17	4,6	4	9	-5	25
Montecarlo	16	3,0	5	11	-6	36
San Ignacio	15	7,1	6	4	2	4
L.G. San Martín	12	5,9	7,5	7	0,5	0,25
Veint. de Mayo	12	6,7	7,5	5	2,5	5,25
Guaraní	10	4,0	9	10	-1	1
San Javier	8	4,9	10	8	2	4
San Pedro	6	1,5	11	14,5	3,5	12,25
Eldorado	5	6,1	12	6	6	36
Iguazú	3	1,4	13	16	-3	9
Candelaria	0	2,9	15,5	12	3,5	12,25
Concepción	0	1,5	15,5	14,5	1	1
General Belgrano	0	1,7	15,5	13	2,5	6,25
Capital	0	0,8	15,5	17	-1,5	2,25
Total					0	157,5

Fuente: Tabla 4.9

Las columnas (I) y (II) reproducen las columnas (a) y (c) de la tabla original. En las columnas (III) y (IV) se consignan los órdenes respectivos de los valores correspondientes a las dos variables. Así, en la columna (III) Oberá presenta el número '1', significando que ocupa el primer rango en cuanto a número de núcleos; pero al mismo departamento en la columna (IV) le corresponde el número '2', puesto que es superado por Caingúas en cuanto al porcentaje de explotaciones medianas de la Provincia que concentra.

En la columna (I) se producen empates: Libertador G. San Martín y 25 de Mayo tienen ambos 12 núcleos del MAM; la cuestión se resuelve adjudicándoles el mismo rango promedio en la columna (III).⁵¹

En la columna (V) se registra la diferencia de rangos, el resultado de restar de (III) el rango en (IV); como es obvio (se trata siempre de los mismos 17 departamentos) todo rango ganado por un departamento debe ser perdido por otro, por lo que todas las diferencias al ser sumadas se anulan entre sí, totalizando 0. Como ya se ha visto, una manera práctica de superar este inconveniente es elevar las diferencias al cuadrado, lo que se hace en la última columna. Se obtiene así: $\sum d^2 = 157,50$.

Si los rangos de los departamentos coincidieran perfectamente, las diferencias entre los rangos serían de 0 en todos los casos y claramente la sumatoria de las diferencias al cuadrado también totalizaría 0. Un coeficiente adecuado debería arrojar un valor de 1 en esa circunstancia, indicando una correlación positiva máxima. Si en cambio los respectivos órdenes fueran perfectamente inversos -esto es, al primer departamento en la primera variable le correspondiera el rango 1 y en la

⁵¹ Los rangos que les corresponden son el 7 y el 8; se suman estos valores y se los divide por el número de departamentos empatados -2, en este caso-, obteniendo el rango promedio de 7,5. El mismo temperamento se adopta para los cuatro últimos departamentos de la columna (I), adjudicándoles a cada uno un rango promedio de 15,5 en la columna (III).

segunda el rango 17, al que le siguiera los rangos 2 y 16, y al último los rangos 17 y 1- se estaría ante una correlación negativa máxima, la que un coeficiente de correlación debería representar con el valor de -1. Si finalmente no hubiera ninguna relación entre los dos órdenes el valor debería ser de 0.

Desde 1904 existe un coeficiente que satisface estos requisitos, el ρ (rho) de Spearman cuya fórmula es la siguiente:

$$\rho = 1 - \frac{6\sum d^2}{N(N^2 - 1)}$$

Aplicando esta fórmula a nuestro ejemplo, se obtiene:

$$\rho = 1 - \frac{6 \times 157,50}{17(17^2 - 1)} = 1 - 0,193 = 0,807$$

El valor arrojado por ρ de 0,807 está indicando una asociación mucho más fuerte entre las dos variables que la denotada anteriormente por ϕ , lo que nos ilustra sobre la conveniencia de no trabajar los datos con procedimientos propios de niveles de medición inferiores a los que éstos admiten.

El coeficiente ρ de Spearman ha gozado durante mucho tiempo de una abusiva popularidad, sin duda atribuible a la simplicidad de su cómputo. Sin embargo, es aún un instrumento bastante imperfecto, puesto que está planteado para variables definidas en un nivel ordinal.

Ya se ha visto que al trabajar con UUAA colectivas, como países, provincias o departamentos, no es infrecuente que nuestras mediciones den lugar a variables intervalares. De hecho, esto es precisamente lo que ocurre en nuestro ejemplo: tanto el número de núcleos del MAM como el porcentaje de explotaciones medianas están definidas en ese nivel. En estas condiciones, utilizar el rho de Spearman sigue entrañando una gran pérdida de información, por la simple razón de que no toma en cuenta las distancias que separan a unas unidades de otras en ambas variables. Así, en la primera columna de la tabla vemos que Cainguás tiene 17 núcleos más que Alem, mientras que en San Pedro sólo existe un núcleo más que en Eldorado. Pero, al considerar las distancias como simples diferencias de orden, estas distancias quedan igualadas, puesto que ambos pares de Departamentos ocupan rangos contiguos.

El coeficiente r de Pearson

Por esta razón, una mejor manera de rendir justicia a estos datos será utilizando el coeficiente de producto-momento de Pearson. Como es sabido,⁵² este coeficiente mide la cantidad de dispersión de las observaciones en torno a la recta de mínimos cuadrados,⁵³ recta que responde a una ecuación del tipo:

$$y = a + bx$$

Así, la ecuación de la recta de mínimos cuadrados será para este ejemplo:

$$y = -0,59633 + 2,1449 x$$

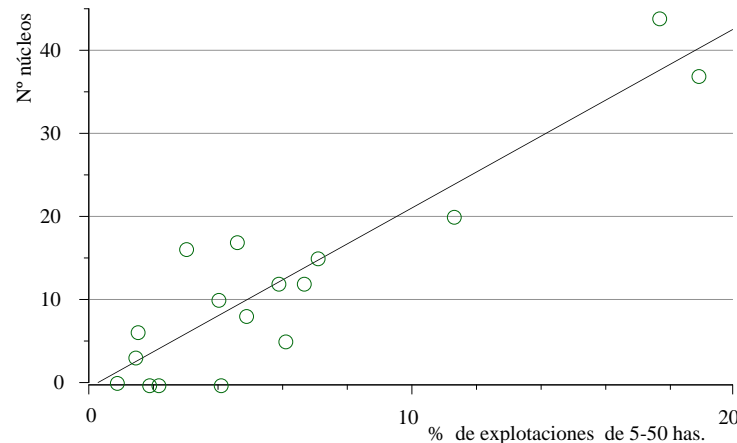
Adjudicando en esta ecuación cualquier valor a x , se obtiene el valor previsto para Y .

En la Figura siguiente se presenta un diagrama de dispersión: cada punto simboliza la ubicación de un departamento de Misiones en el espacio que determinan las variables consideradas.

⁵² No nos extenderemos aquí sobre el tema de regresión y correlación; sugerimos remitirse a cualquier manual de estadística elemental.

⁵³ La recta de mínimos cuadrados es la que minimiza los cuadrados de las distancias con respecto a los puntos en que se ubican las UUAA dentro del espacio conformado por dos variables.

Figura 4.2: Diagrama de dispersión de los departamentos y recta de regresión para las variables N° de núcleos del MAM y % de explotaciones de 5-50 has.



Si todas las observaciones se ubicaran sobre la recta, se estaría en presencia de una correlación lineal perfecta, y el r de Pearson arrojaría un valor de 1 (o de -1, en caso de tratarse de una correlación negativa). Aplicando la fórmula del coeficiente producto-momento se obtiene para los datos de la Tabla 4.9:

$$r = 0,927$$

Un valor aún mayor que el de ρ y muy próximo a 1, indicando así una correlación muy elevada entre las variables. Esta relación puede interpretarse como un elemento de prueba muy fuerte en favor de la hipótesis.

Una diferencia entre este coeficiente y los otros que hemos presentado es la interpretación estadística rigurosa que puede darse de sus valores. En efecto, si se eleva el resultado al cuadrado se obtiene:

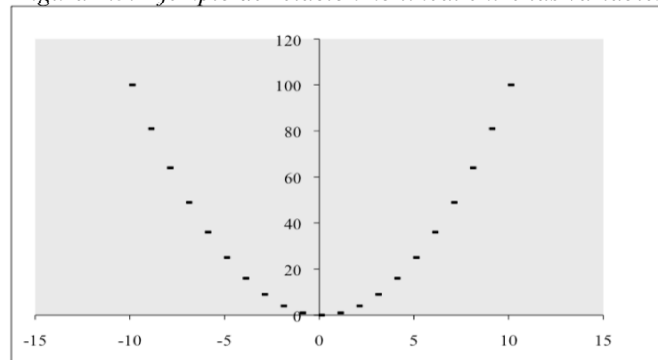
$$r^2 = 0,859$$

r^2 tiene un significado matemático preciso: es la proporción de la variancia de Y explicada por X .

Examinar siempre primero el diagrama de dispersión

El diagrama de dispersión es un modo particularmente útil de presentar los datos para visualizar rápidamente si existe alguna relación entre las variables y la forma que ésta toma. En particular, será un modo de evitar errores en la interpretación de bajos valores del coeficiente r de Pearson. Como se ha visto, dicho coeficiente mide el grado en que se puede establecer la existencia de una relación lineal entre dos variables. Sin embargo, es claro que pueden existir relaciones no lineales entre variables.

Figura 4.3: Ejemplo de relación no lineal entre las variables



En la situación de la Figura 4.3, la recta de regresión toma la forma de una paralela a la abscisa, y el valor de r será 0, indicando que la relación *lineal* entre las variables es nula. Sin embargo, las

UUAA (representadas por puntos) describen una relación que responde visiblemente a la ecuación de una parábola; esto es, un tipo de relación no monotónica entre las variables.⁵⁴

¿Es legítimo inferir relaciones de un nivel a otro?

Retornando a la hipótesis inicial, cabe aquí una reflexión a propósito de los límites a la posibilidad de producir inferencias legítimas a partir de una correlación entre variables como la hallada. El muy elevado valor del r de Pearson nos dice que, al nivel de los departamentos, las dos variables muestran una correlación casi perfecta. Empero, no es posible deducir de ello que esta misma asociación se produzca al nivel de los colonos individuales, lo que supondría incurrir en la falacia ecológica.⁵⁵ Es una situación recurrente en la práctica de la investigación el no disponer de datos en el nivel en que se plantea la hipótesis.

En el estudio sobre el MAM, para disponer de esos datos se podría haber realizado una encuesta por muestreo a colonos de Misiones, inquirendo acerca de su pertenencia o no al MAM así como sobre el tamaño de sus explotaciones. El resultado sería una tabla de la siguiente forma:

		¿ Opera una explotación mediana?	
		No	Sí
¿ Perteneces al MAM?	Sí	n	N
	No	N	n

Otra alternativa pudiera haber sido disponer de datos sobre la distribución de las superficies de las explotaciones de los afiliados al MAM para compararla con la distribución de las superficies de todas las explotaciones de la Provincia obtenida a partir de fuentes secundarias.

Cualquiera de estas alternativas presenta sin embargo inconvenientes por su costo o por la imposibilidad material de ser llevada a cabo.⁵⁶ Existiendo en cambio datos secundarios, es recomendable encarar su utilización aún cuando no nos proporcionen el mismo grado de certeza en nuestras conclusiones. La correlación verificada a nivel de los departamentos es un elemento de juicio más que, conjuntamente con la ausencia de hipótesis alternativas que den cuenta de esta correlación, hace plausible sostener la existencia de la misma relación en el nivel de los actores.

La matriz de correlaciones

Además de la tabla de contingencia y de la distribución multivariante conjunta, nos referiremos brevemente a una última forma de presentar los datos. En la matriz de correlaciones, ha desaparecido toda referencia visible a las UUAA y a los valores originales de las variables. Los elementos de esta matriz son los nombres de las variables, que se repiten en las hileras y en las columnas, y en las celdas se consigna los valores que arroja la correlación entre cada par de variables.

⁵⁴ Una relación monotónica es aquella en la cual los incrementos en X están asociados siempre o bien con incrementos en Y, o bien con disminuciones en Y, todo a lo largo del recorrido de X. En las relaciones no monotónicas, en cambio, esto no se cumple; así, en la Figura XIX, los incrementos en X producen primero incrementos en Y, pero a partir del valor '10' en X pasan a producir disminuciones en Y. Existen por lo tanto, tres tipos de relaciones entre variables: lineales, monotónicas no lineales, y no monotónicas (cf. Hartwig y Dearing, 1979: 48-50).

⁵⁵ La falacia ecológica es un caso particular de lo que Galtung denomina la "falacia del nivel equivocado", que consiste en la transferencia ilegítima de relaciones verificadas en un nivel hacia otro.

⁵⁶ Así, en los años en que fue realizada esta investigación, las condiciones políticas hacían imposible pensar en realizar una encuesta sobre estos temas.

Tabla 4.9.3: Matriz de correlaciones entre cuatro indicadores

	Nº de núcleos en 1973	% explot. de 5 a 50 has. en Depto.	% explot. de 5 a 50 has. en Provincia	% mano de obra familiar
Nº de núcleos	1,000	0,656	0,927	0,200
% 5-50 (Depto.)	0,656	1,000	0,722	0,397
% 5-50 (Prov.)	0,927	0,722	1,000	0,309
% mdo familiar	0,200	0,397	0,309	1,000

Fuente: Tabla 4.9

En la Tabla 4.9.3 se presenta una matriz de correlaciones entre las cuatro variables incluidas en la tabla original. Los valores son los que arroja r de Pearson. En la diagonal, los valores son siempre 1,000 significando ello que la correlación de cada variable consigo misma es perfecta.⁵⁷

La matriz de correlaciones permite una visualización en conjunto de las relaciones entre varios pares de variables, y es de múltiples aplicaciones en la investigación social. Como veremos, en la construcción de índices y escalas es posible utilizarla para seleccionar los indicadores o ítems más apropiados; asimismo, esta matriz es un paso intermedio para el desarrollo de varias técnicas de análisis más complejas.

Las técnicas que hemos desarrollado hasta aquí son las más elementales. Hoy en día, el análisis de datos requiere de instrumentos más poderosos. Sobre todo, ya se trate de tablas de contingencia o de distribuciones multivariantes, es esencial poder acometer el análisis simultáneo de las relaciones entre más de dos variables.

Para las tablas de contingencia, el análisis multivariable encuentra sus fundamentos en el clásico trabajo de Lazarfeld (1968). En lo que hace a variables intervalares, existe una variedad de técnicas que se adaptan a diferentes situaciones. Así, por ejemplo, para analizar la relación entre una variable intercalar dependiente, y varias variables independientes de nivel nominal y/o ordinal, se puede recurrir al análisis de variancia (Iversen y Norpoth, 1976); si en cambio tanto las variables independientes como la dependiente son intervalares, el análisis de regresión múltiple puede ser una alternativa adecuada. Recomendamos la lectura de los buenos manuales⁵⁸ para familiarizarse con estas técnicas un poco más complejas. Por último, y siempre dentro de las técnicas estadísticas clásicas, también se puede recurrir al análisis factorial (Kim y Mueller; 1978a y 1978b) para una multiplicidad de propósitos. En el capítulo 6 nos ocuparemos de un tipo particular de técnicas factoriales más recientes que resultan especialmente útiles para la investigación sociológicamente orientada.

En cuanto al uso de todas estas técnicas, se ve hoy en día muy simplificado por la difusión de la computadoras personales,⁵⁹ y la existencia en el mercado de *software* cada vez más poderosos. A esta altura, cualquier lego en computación tiene a su disposición una variedad de paquetes

⁵⁷ Por lo demás, la diagonal separa la tabla en dos sectores simétricos: cuando se correlaciona, por ejemplo, “% mano de obra familiar”(hilera) con “Nº de núcleos” (columna), se obtiene el mismo valor de 0,200 que al correlacionar Nº de núcleos” (hilera) con “% mano de obra familiar”(columna); por esta razón es frecuente representar solamente uno de los sectores de la matriz.

⁵⁸ Aunque algo *démodée*, la obra de Blalock (1966) sigue siendo un modelo de rigor metodológico. Alternativas más adaptadas a las condiciones actuales de la investigación pueden ser los textos de Lewis-Beck (1995), de Sirkin (1995) y de Wright (1997).

⁵⁹ Nos referimos a las denominadas computadoras ‘PC’ (del inglés, *Personal Computer*) de uso cotidiano, ya sea que trabajen con Windows, con Unix o con el sistema Macintosh, las cuales, para las aplicaciones mencionadas, han sustituido por completo a las computadoras de mayor tamaño (*mainframe*).

estadísticos que incluyen todas las técnicas mencionadas y muchas otras, y que son por lo general de muy sencilla utilización.⁶⁰

⁶⁰ El *software* consiste en la parte “blanda” de la computación, los programas con instrucciones para realizar distintos tipos de operaciones: procesamiento de textos, graficadores, hojas electrónicas, etc. El *hardware* está constituido por la parte “dura”: la computadora en sí, los soportes físicos de la información (discos rígidos, diskettes, zip-disks, etc.). En lo que hace al *software*, hoy en día una hoja de cálculo común (Excel) alcanza para la mayoría de las aplicaciones que se requieren corrientemente. Sino, se puede recurrir a paquetes estadísticos varios. Uno de los más conocidos es el SAS (el *must*, aunque con fama de poco amigable). Aunque, sin duda, el programa más popular en ciencias sociales ha sido y sigue siendo el SPSS (*Statistical Package for the Social Sciences*-Paquete Estadístico para las Ciencias Sociales); ya en 1979 el manual de Padua dedicaba un capítulo a reseñar las técnicas incluidas en el SPSS (en aquella época era un programa para *mainframe*, pero desde hace años se encuentra disponible en versiones para PC).

CAPÍTULO 5

ÍNDICES Y ESCALAS

En este capítulo nos ocuparemos de los problemas que presenta la construcción de índices y de escalas. La cuestión consiste precisamente en «cómo un concepto, proveniente del lenguaje común o de la reflexión teórica sobre la realidad social, puede ser traducido en una medida».¹ Se trata de uno de los aspectos fundamentales del “proceso de operacionalización”, vale decir, de las operaciones típicas que se requieren para generar un instrumento de medición partiendo de un concepto o de una noción.²

A partir del concepto de ‘espacio de propiedades’, expondremos las distintas alternativas de reducción y pasaremos a describir las etapas del proceso de operacionalización, mostrando como un índice consiste esencialmente en la reducción de un espacio.

Luego, presentaremos brevemente algunas técnicas para la construcción de escalas destinadas a la medición de actitudes, definiendo a las escalas como tipos particulares de índices. Por último, abordaremos el tema de la confiabilidad y validez de los instrumentos de medición, haciendo referencia a los distintos procedimientos utilizables para evaluar su bondad.

1. ESPACIOS DE PROPIEDADES E ÍNDICES

El concepto de ‘espacio de propiedades’, tal como fue desarrollado por Lazarsfeld (1993) y Alan Barton (1969), es uno de los instrumentos analíticos más poderosos para la investigación social. Sus aplicaciones son múltiples, aunque aquí nos centraremos en la idea de *reducción* de un espacio de propiedades.³ Según Carlos Borsotti, «la propiedad-espacio se define como el conjunto de posiciones que puede ocupar una unidad de análisis en una variable o en un sistema de varias variables».⁴ En este sentido, toda variable considerada individualmente ya configura por sí misma una propiedad-espacio unidimensional.

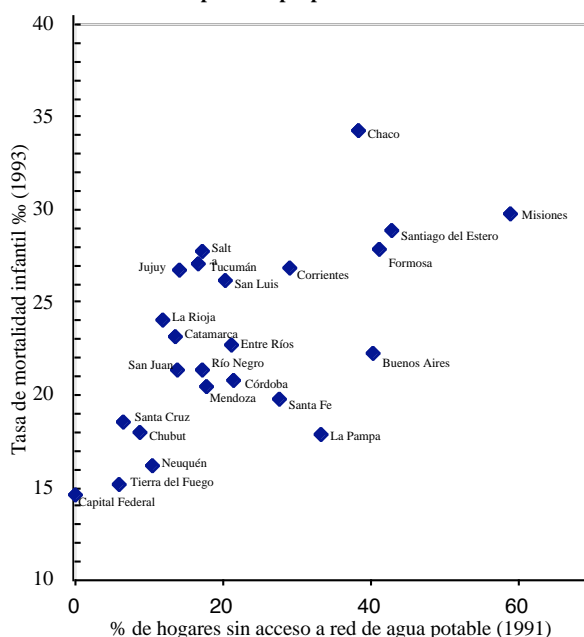
¹ Palabras del Prefacio de Jean Stoetzel al primer volumen de la obra de Boudon y Lazarsfeld que -no por azar- se titula “El vocabulario de las ciencias sociales”.

² Concebido en su mayor amplitud, el “proceso de operacionalización” designa el trabajo de producción de la matriz de datos definiendo sus tres elementos constitutivos: UUAA, variables y valores. Juan Samaja propone agregar a la estructura de la matriz un cuarto elemento: los indicadores definidos «como los procedimientos aplicados a dimensiones relevantes de la variable para efectuar su medición» (cf. Samaja, 1992: 17). Sin embargo, tanto lo que Samaja denomina ‘variables’ como lo que llama ‘indicadores’ desempeñan exactamente el mismo papel en la estructura de la matriz de datos: ambos figuran como columnas en dicha matriz. Como se verá, ocurre precisamente que las columnas-variable son generadas a partir de operaciones sobre las columnas-indicador. La cuestión de la selección de los indicadores, de cómo se mide, es tan ajena a la forma de la matriz como el problema de seleccionar una muestra de UUAA.

³ No nos ocuparemos aquí del concepto de ‘substrucción’, operación inversa a la de reducción y de innegable trascendencia para el trabajo teórico.

⁴ Cf. Padua, 1979: 239.

Figura 5.1: Posición de las provincias argentinas en un espacio de propiedades continuo



Fuente: datos del Programa Argentino para el Desarrollo Humano.

Si tomamos, por caso, las provincias argentinas como unidades de análisis, cada una de ellas es susceptible de ser ubicada en una línea horizontal que simbolice el continuo de todos los valores posibles que pueden asumir en una variable cualquiera de nivel intervalar -por ejemplo, el porcentaje de hogares sin acceso a la red de agua potable sobre el total de hogares de cada provincia. Del mismo modo, estas mismas unidades pueden ubicarse en otro continuo, la tasa de mortalidad infantil,⁵ siendo que el conjunto de valores comprendidos entre el 0 y el 1000 ‰ constituye una nueva propiedad-espacio. Ahora bien, por múltiples razones podemos tener interés en ubicar nuestras unidades de análisis en espacios multidimensionales, es decir espacios en los que cada unidad aparecerá ubicada simultáneamente en dos o más dimensiones o variables.

En la Figura 5.1 se observa el resultado de distribuir las provincias en un espacio de propiedades conformado por las dos variables mencionadas: obtenemos un clásico diagrama de ejes cartesianos.⁶

Sin embargo, esta forma de representación tan satisfactoria sólo es posible si estamos tratando con variables de nivel intervalar, basadas típicamente en propiedades analíticas de colectivos (las provincias de nuestro ejemplo). Cuando lidiamos con unidades de análisis de menor nivel, y en especial tratándose de individuos, nuestras mediciones alcanzan raramente dicho nivel y debemos conformarnos con trabajar con atributos y órdenes. ¿Qué sucede entonces cuando se combinan dos propiedades *cualitativas*? El espacio de propiedades resultante ya no consiste en un plano continuo, no se resuelve en un par de ejes cartesianos, sino que toma la figura de un conjunto ordenado de celdas, cada una de las cuales representa una determinada combinación de valores o categorías.

Reducciones de espacios de propiedades cualitativas

Supongamos que estamos realizando una investigación en el campo de la Sociología del Trabajo y que, inquiriendo a un conjunto de obreros acerca de su satisfacción con la tarea que realizan, hemos llegado a diferenciarlos -mediante el uso de algún tipo de índice o de escala- según su satisfacción con la tarea sea Alta (A), Media (M) o Baja (B). Al mismo tiempo, hemos preguntado

⁵ La tasa de mortalidad infantil «se obtiene dividiendo las defunciones de menores de un año por el total de nacimientos vivos -ponderados por mil- ocurridos en un área y en un año determinados (Boleda; 1993: 73).

⁶ Agregando un tercer eje, z, es posible producir la representación de un espacio tridimensional. Empero, los espacios de más de tres dimensiones no son representables gráficamente.

a los mismos obreros si se consideraban satisfechos con el salario que percibían, obteniendo una clasificación análoga en Alta, Media y Baja satisfacción con la remuneración. Si ahora nos interesa ver cómo se distribuyen estas unidades (los obreros) simultáneamente en las dos clasificaciones, la combinación de ambas variables dará lugar a un espacio de propiedades de nueve (3×3) **variantes**. No por sencillo resulta menos esencial captar que al conformar ese conjunto de nueve variantes lo que hicimos fue generar una nueva clasificación: cada obrero encuestado deberá necesariamente encontrar su ubicación en una celda y solamente en una celda. Para decirlo con toda claridad, lo que ha ocurrido es que, combinando dos variables (cuyos valores originales eran, en ambos casos, A, M y B), *hemos creado una nueva variable*, con los valores AA, AM, AB, MA, MM, MB, BA, BM y BB.⁷

Figura 5.2: Ejemplo de espacio de propiedades cualitativas

Satisfacción con el salario	Satisfacción con la tarea		
	Baja	Media	Alta
Alta	AB	AM	AA
Media	MB	MM	MA
Baja	BB	BM	BA

Ahora bien, puede ocurrir que nos resulte impráctico manejarnos en el análisis con estos nueve valores de nuestra nueva variable o bien que, a los efectos de nuestra investigación, la relación que medie entre esta variable y **otras** variables no se destaque con la suficiente nitidez al trabajarla en un nivel de desagregación que implica tanto detalle. Hay muchos motivos por los cuales puede resultar conveniente **reducir** el número de valores que conforma el espacio de propiedades. Por ejemplo, podemos estar trabajando con una muestra que, por su escaso tamaño, imponga limitaciones a la magnitud de las tabulaciones cruzadas que tenga sentido encarar;⁸ también puede ocurrir que no todas las posiciones -o valores- de la nueva variable revistan la misma significación ni sean igualmente importantes con respecto a las hipótesis que tenemos en mente. Desde este punto de vista, lograremos una mayor claridad conceptual refundiendo varias posiciones de nuestro espacio de propiedades original en una nueva categoría. En suma, existen motivos tanto empíricos como teóricos que con frecuencia nos llevarán a encontrar ventajas en proceder a una **reducción** de un espacio de propiedades. Esto supone contar con criterios determinados en base a los cuales se pueda reagrupar el conjunto de valores que constituye el espacio de propiedades original en un número menor de categorías. A continuación, pasamos a describir tres tipos distintos de reducción señalados por Barton.⁹

Reducción por simplificación de dimensiones

Sin duda, el caso más obvio. Es sabido que cualquier variable es susceptible de ser reducida a una dicotomía. Si entonces reducimos el número de valores de una o más de las dimensiones que conforman un espacio de propiedades, se reducirá necesariamente el número de celdas de dicho espacio.

⁷ Dado que una variable es un conjunto de valores tal que forma una clasificación (cf. *supra*: capítulo 1), las combinaciones de variables son a su vez variables.

⁸ Hemos visto en el capítulo 3 como Galtung sugería la necesidad de contar con un número de casos que contemple como mínimo un promedio de 10 unidades por celda. En nuestro ejemplo, la nueva variable cuenta con nueve posiciones, por lo que para cruzarla aunque más no sea con otra variable dicotómica, requeriríamos un tamaño mínimo de la muestra de $9 \times 2 \times 10 = 180$ unidades de análisis.

⁹ Barton hace referencia a un cuarto tipo: la reducción *funcional*; ésta es la que está obrando en el análisis de escalograma de Guttman (cf. *infra*: sección 2 de este capítulo) y en las distintas técnicas de análisis factorial (algunas de las cuales serán abordadas en el capítulo 6).

Figura 5.2.a: Reducción por simplificación de dimensiones

Satisfacción con el salario	Satisfacción con la tarea		
	Baja	Media	Alta
Alta	AB		AA
Media	BB		BA
Baja			

En nuestro ejemplo, hemos reagrupado en ambas dimensiones los valores “Media” y “Baja” bajo el rótulo arbitrario de “Baja” (tal vez, con mayor exactitud, pudiéramos haber denominado a esta nueva categoría “No Alta”). De un espacio de propiedades de nueve variantes pasamos así a uno de cuatro; las combinaciones originales MM, MB, BM y BB dan ahora lugar a una sola posición: BB.

Reducción numérica

Cuando es sostenible pensar que las diferentes dimensiones que conforman un espacio de propiedades se relacionan todas con una misma característica subyacente, es posible practicar este tipo de reducción.

Figura 5.2.b: Reducción numérica

Satisfacción con el salario	Satisfacción con la tarea		
	Baja	Media	Alta
Alta	2	3	4
Media	1	2	3
Baja	0	1	2

Así, la satisfacción con la tarea realizada y la satisfacción con el salario son factibles de ser subsumidas una idea más general, que englobaría a ambas, la de ‘satisfacción con el trabajo’. En efecto, parece lógico suponer que cuanto más satisfecho se sienta un obrero con la tarea que realiza, mayor será la satisfacción con su trabajo (y viceversa), y que lo mismo ocurre respecto a la satisfacción con su salario.

En estas condiciones, y suponiendo que ambas dimensiones tengan la misma relevancia en cuanto a la satisfacción con el trabajo, se pueden asignar idénticos puntajes numéricos a las posiciones análogas en cada dimensión. En nuestro ejemplo, asignamos 2 puntos a ‘Alta satisfacción’, 1 a ‘Media’ y 0 a ‘Baja’. Un obrero con alta satisfacción tanto respecto a su tarea como a su remuneración obtiene así un puntaje -máximo- de 4 puntos; uno con baja satisfacción en ambas dimensiones obtendría el mínimo de 0 puntos. El resultado de este procedimiento es la generación de una nueva variable que cuenta con cinco posiciones -o valores- ordenados de 0 a 4, y que son interpretables como la expresión de grados crecientes de satisfacción con el trabajo.

Nada impediría por lo demás reducir nuevamente este conjunto resultante de categorías. Así, podríamos definir:

4 o 3 puntos	⇒ Alta satisfacción con el trabajo
2 puntos	⇒ Mediana satisfacción con el trabajo
1 o 0 puntos	⇒ Baja satisfacción con el trabajo

En definitiva, mediante este procedimiento de reducción numérica hemos generado una nueva variable, a la que convendremos en denominar ‘índice’.

Reducción ‘pragmática’

Si observamos nuestra reducción numérica con mayor detalle, podemos plantearnos la siguiente cuestión. Considerando qué significa tener una ‘mediana satisfacción con el trabajo’, vemos que los obreros que obtengan un puntaje de ‘2’ pertenecerán necesariamente a alguna de estas tres combinaciones:

2 (salario) + 0 (tarea)

1 (salario) + 1 (tarea)

0 (salario) + 2 (tarea)

¿ Ahora bien, es legítimo englobar en una misma categoría -el valor ‘2’- a personas cuyas situaciones son tan diferentes? Puede que sí o puede que no: ello depende de cuáles sean nuestros objetivos, de lo que nos propongamos lograr al llevar adelante esta reducción.¹⁰ Podría ocurrir que nos interesara conservar la especificidad de alguna(s) de las categorías del espacio original, por revestir un significado especial para la investigación que estamos desarrollando. Así, si tuviéramos la idea -o hipótesis- de que la insatisfacción que deviene de una baja remuneración es distinta de la que produce una tarea monótona o de escaso prestigio, y que, por ejemplo, ambos tipos de insatisfacción bajo determinadas condiciones tenderían a estar asociados a diversos correlatos ideológico-políticos, sería absolutamente legítimo proponer una reducción pragmática como la siguiente:

Figura 5.2.c: Reducción pragmática

Satisfacción con el salario	Satisfacción con la tarea		
	Baja	Media	Alta
Alta	Insatisfacción profesional	Mediana satisfacción	Alta satisfacción
Media	Baja satisfacción		Insatisfacción económica
Baja			

En el ejemplo, los que tenían un puntaje alto (4 o 3) en la reducción numérica siguen agrupados bajo el rótulo de ‘Alta satisfacción’; lo mismo ocurre con los de bajo puntaje (1 o 0). Pero aparecen dos nuevas categorías: “Insatisfacción profesional” (en la celda superior izquierda) e “Insatisfacción económica” (en la inferior derecha). Observemos sin embargo que el conjunto de los valores producidos por esta reducción pragmática ya sólo se presenta en forma *parcialmente ordenada*. En efecto, se puede convenir en que:

Alta satisfacción > Media satisfacción > Baja satisfacción

¿Pero qué ocurre con las categorías restantes de insatisfacción “profesional” y “económica”? Ya no encuentran ubicación en esa secuencia. En consecuencia, podemos decir que mientras cualquier reducción numérica deviene en una variable ordinal, nuestra reducción pragmática produce una variable que se inscribe en el nivel nominal de medición.

En lo que se podría discrepar parcialmente con Barton, es en la denominación de ‘pragmática’¹¹ que le da a la última forma de reducción que hemos examinado. En efecto, si algo debe quedar claro es que cualquier reducción que hagamos será siempre “pragmática”, en el sentido de que deberá ser *útil* a nuestros objetivos. En el ejemplo que dimos, tal vez cupiera hablar de un tipo de reducción “mixta” (se combina una reducción numérica básica con algún criterio adicional). Sin embargo, no toda reducción “pragmática” parte necesariamente de una reducción numérica previa. El siguiente es un ejemplo clásico de Barton que permite ilustrar este punto a partir de datos de un estudio de panel:

¹⁰ Dice Barton: «Sin embargo, para algunos propósitos, cualquier sistema de indicadores ponderados resulta muy rígido y agrupa las categorías de modo muy arbitrario» (en Korn, 1969: 64).

¹¹ Denominación que conservaremos, no obstante, en aras a la tradición.

Figura 5.3: Reducción de un espacio de propiedades de posición política en dos puntos del tiempo

<i>Voto efectivo</i>	<i>Intención de voto</i>		
	Demócrata	No sabe	Republicano
Demócrata	Constancia	Cristalización	Cambio
No sabe	Abandono	Constancia nocristalizada	Abandono
Republicano	Cambio	Cristalización	Constancia

Fuente: Korn, 1969: 66.

En definitiva, la de ‘reducción pragmática’ es una categoría residual, a la que habrán de ir a parar todas las formas que no procedan ni “por simplificación de dimensiones” ni “numéricamente”.

El proceso de operacionalización

Ya hemos mencionado que el resultado de una reducción numérica es un índice. Cuando realizamos esa reducción de nuestro espacio original, tal vez no fuéramos conscientes de estar generando un índice. Disponíamos de dos variables a las que combinamos en un espacio de propiedades, con lo que aparentemente -por una suerte de generación espontánea- de los mismos datos surgía el tal índice. Lo primero que debemos destacar es esta idea elemental: *un índice consiste esencialmente en la reducción de un espacio de propiedades*. No existe por lo demás ninguna magia en la generación del índice; es nuestra capacidad imaginativa la que nos permite introducir una idea más general -la de ‘satisfacción con el trabajo’- para subsumir en ella las dos variables originales considerándolas como aspectos o dimensiones parciales de la misma. Sin esta idea, nuestra reducción carecería totalmente de sentido.

Nos encontramos ahora en condiciones de plantear el problema de la construcción de índices “al derecho”, por así decirlo. Hay un texto célebre de Paul Lazarsfeld,¹² en el que aborda las etapas del proceso de operacionalización. Este proceso consiste muy precisamente en la construcción de un índice que permita *traducir* un concepto -de menor o mayor grado de abstracción- en una variable mensurable. Describiremos a continuación las cuatro etapas que componen este proceso.

☛ La primera etapa corresponde a la *representación del concepto en imágenes*. «El concepto, en el momento en que cobra cuerpo, no es más que una entidad concebida en términos vagos, que otorga sentido a las relaciones observadas entre los fenómenos».¹³ De acuerdo a Francis Korn, llegaríamos aquí desde la noción primigenia hasta el nivel de una definición *nominal*¹⁴ del concepto.¹⁵

☛ La fase siguiente consiste en analizar los componentes más importantes de esta noción primera. Se trata de discernir los aspectos o *dimensiones* principales inherentes al concepto. Estas pueden ser deducidas analíticamente del concepto general que las engloba, o bien ser producidas más empíricamente a partir de la estructura de sus correlaciones.¹⁶ Así, en nuestro ejemplo, tratándose de operacionalizar la noción de ‘satisfacción con el trabajo’, cabría reconocer al menos una dimensión “económica” y otra de tipo “profesional”. En otra terminología, podríamos decir

¹² Publicado originalmente en *Daedalus*, y reproducido posteriormente en infinidad de antologías, entre otras en las de Boudon/Lazarsfeld (1965) y de Korn (1969).

¹³ Cf. Boudon y Lazarsfeld, 1965: 28.

¹⁴ Cf. Korn, 1969: 10-11.

¹⁵ En rigor, el punto de partida puede consistir tanto en una noción -proveniente del sentido común, tal vez con una gran carga ideológica, etc.-, como en un concepto -inserto dentro de un campo teórico, con significado “constitutivo” (cf. Torgerson, 1965).

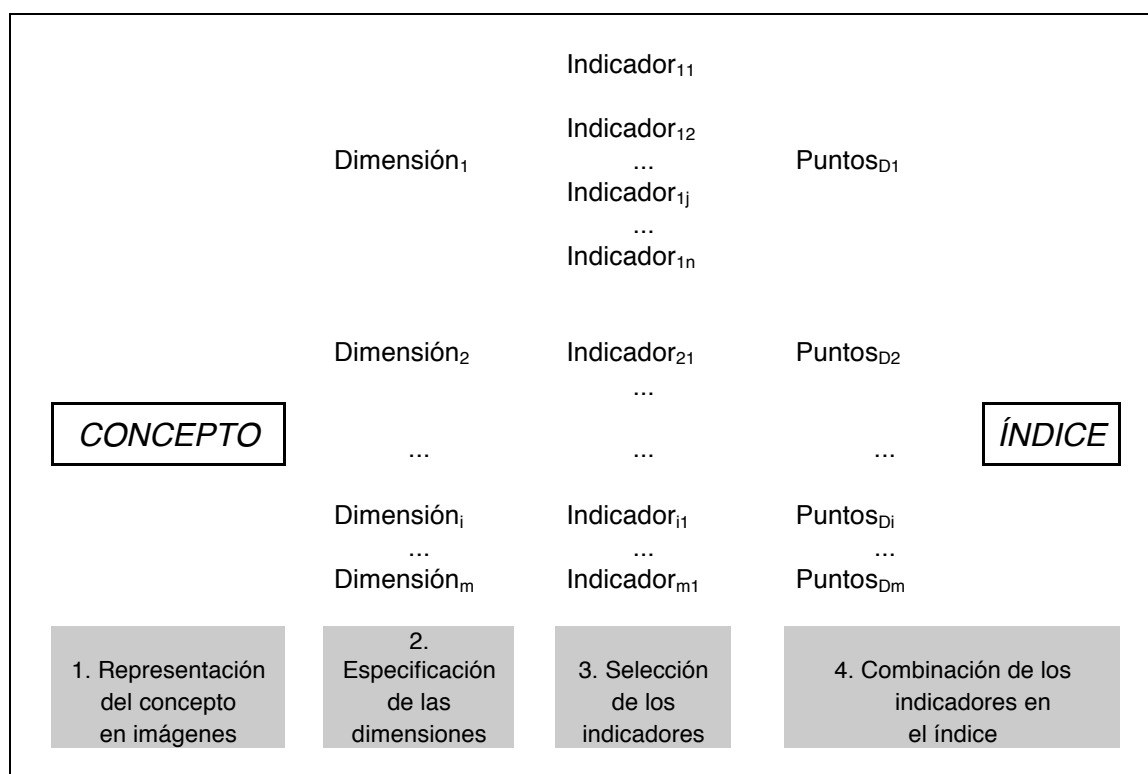
¹⁶ Este último proceder, más inductivo, será abordado en el capítulo 6.

que esta etapa consiste en la producción de una definición “*real*” del concepto.¹⁷

☛ La tercera etapa supone proceder a la *selección de los indicadores*. Así, deberíamos contar con un grupo de preguntas (o ítems) dirigidos a medir la satisfacción con el salario, y con otro grupo de indicadores destinados a representar la satisfacción con la tarea. La necesidad de combinar varios indicadores (o ítems, o preguntas) deviene de que la relación que vincula cada indicador a la dimensión correspondiente -y por ende al concepto- es una relación esencialmente probabilística y no de certidumbre.¹⁸

☛ Por último, los indicadores seleccionados deben combinarse, siguiendo reglas determinadas, en alguna clase de *índice*. Un procedimiento sencillo de combinación de los indicadores es adjudicar un cierto número de puntos a cada respuesta (o valor, o categoría) dentro de cada indicador, con lo que cada unidad de análisis presentará un puntaje total -su puntaje en el índice-, consistente en la suma de los puntajes obtenidos en cada indicador. Es a través de estas dos últimas etapas que se arriba a una definición *operacional* del concepto.

Figura 5.4: Etapas del proceso de operacionalización, según Lazarsfeld



En todo este procedimiento, es de notar que con frecuencia la reducción del espacio de propiedades será implícita, puesto que éste no requiere haber sido diagramado como tal. Hay algunos puntos de esta presentación que merecen destacarse. Si bien Lazarsfeld parece inclinarse por la alternativa de que el índice se genere a partir de los propios datos,¹⁹ tampoco excluye la posibilidad de un proceder más teórico. Por otra parte, en lo que hace a las dimensiones del concepto, es importante señalar que el proceso de especificarlas es esencialmente analítico y que

¹⁷ Cf. Korn, 1969: 10-11. La expresión ‘definición real’ es sumamente discutible (cf. Cohen y Nagel, 1967); por lo demás, esta cuestión no es ajena a la muy antigua “querrela de los universales”, a la que más modernamente se alude como “nominalismo versus realismo” (cf. Galtung, 1966: Cap. I).

¹⁸ Dice Lazarsfeld: «Cada indicador posee un carácter específico y no se lo debe jamás considerar como plenamente representativo para la clasificación que se busca» (cf. Boudon y Lazarsfeld, 1965: 33).

¹⁹ «¿“Dados” por quién?», preguntaría Pierre Bourdieu.

tiende a descansar en el criterio subjetivo del investigador;²⁰ no existe, en efecto, regla alguna objetiva que nos permita determinar cuál es el número de dimensiones requerido para operacionalizar un concepto determinado. Por último, en lo que se refiere a los indicadores, se los deberá seleccionar primordialmente atendiendo a su validez interna (cf. *infra*), vale decir de acuerdo a su congruencia con el concepto a medir. En este sentido, debe quedar claro que tampoco el número de los indicadores requeridos se encuentra determinado *a priori*. Lazarsfeld habla de un **universo** hipotético de indicadores posibles, con lo que la cuestión de seleccionarlos pasa a convertirse en un problema de *muestreo* de indicadores. Al enunciar su *principio de la intercambiabilidad de los indicadores*, Lazarsfeld hace gala de un pragmatismo extremo. Su razonamiento parte de que no existe la medición pura en ciencias sociales, sino que nuestras medidas serán siempre y necesariamente imperfectas. Así, por ejemplo, si se toman dos muestras distintas de indicadores de una misma característica subyacente, la correlación entre los dos índices generados de ese modo, con ser elevada, no alcanzará nunca a ser perfecta. No obstante, si antes que una clasificación perfecta de nuestras unidades de análisis nos interesa la relación entre la característica que investigamos y alguna otra variable externa, ambos índices conducirán a resultados prácticamente equivalentes.

En síntesis, cuantos más indicadores seleccionemos para un índice determinado, mejor será nuestra medición de la característica subyacente.²¹ Un *indicador* es una característica o respuesta manifiesta que se encuentra ya definida en términos de lo que hay que observar. La medida compleja que se obtiene combinando los valores obtenidos por un individuo en cada uno de los indicadores propuestos para la medición de una variable se denomina *índice*. «La diferencia entre un índice y un indicador es entonces de grado. Un índice es un complejo de indicadores de dimensiones de una variable, y constituye por lo tanto el indicador total de una variable compleja (Korn, 1969: 12).»

2. ESCALAS DE ACTITUDES

Si un índice no es más que una forma de reducir un espacio de propiedades, podemos decir del mismo modo que las escalas son una clase particular de índices,²² cuya diferencia específica consiste en la utilización de una serie de procedimientos objetivos para la selección de los indicadores (o ítems), lo que permite controlar mediante técnicas estandarizadas si éstos se inscriben efectivamente dentro de una dimensión común.

Las escalas fueron desarrolladas originalmente con miras a la medición de actitudes, aun cuando puedan ser aplicables a otra clase de objetos. En la *Enciclopedia Internacional de las Ciencias Sociales* (I,15) podemos leer esta definición de Milton Rokeach: Se entiende por ‘actitud’ «una organización relativamente duradera de creencias en torno a un objeto o situación, que predispone a reaccionar preferentemente de una manera determinada.»

Hay varios tipos de escalas de actitudes, aunque todos ellos se orientan a conocer la *dirección*

²⁰ Criterio que *a posteriori* podrá ser compartido por otros miembros de la comunidad científica, pero esta es otra cuestión.

²¹ Aunque juega aquí una suerte de “ley de rendimientos decrecientes”. Según H. W. Smith, «Los primeros ítems que se agregan contribuyen en mucho mayor medida a la confiabilidad que los últimos añadidos. Así, luego de un cierto punto (i.e., 8 a 10 ítems) cualquier ítem adicional contribuirá casi insignificativamente a aumentar la confiabilidad...¿Por qué se mantiene esta relación entre la longitud de la escala y la confiabilidad? Cualquier medida tiene fuentes idiosincráticas de error. Por ejemplo, el uso del ingreso como una medida del status socioeconómico presenta problemas idiosincráticos como el de incluir en la clase alta miembros de la Mafia con un salario de 100.000 US. Con cada ítem agregado a la escala puede esperarse que decrezcan los errores ítem-específicos, puesto que normalmente se espera que se cancelen unos a otros. Esto es, puede esperarse que cada medida incluya factores irrelevantes para lo que el investigador está midiendo, pero como estos componentes irrelevantes son presumiblemente distintos para cada ítem, y como todos los ítems comparten un foco común, el conjunto total de medidas proveerá normalmente una medición más pura del fenómeno en estudio que cualquier medición aislada». (cf. Smith, 1975: 60-61).

²² Según Samuel Stouffer, «Se dice que existe escala cuando a partir de una distribución de frecuencias multivariada de un universo de atributos, es posible derivar una variable cuantitativa con la cual caracterizar los objetos de un modo tal que cada atributo sea una función simple de aquella variable cuantitativa» (citado en Padua, 1979: 160).

(positiva o negativa) y la *intensidad* (mayor o menor)²³ de las actitudes de personas individuales hacia un determinado objeto utilizando un conjunto de ítems. Cada ítem consiste en un juicio, en una frase que funciona como indicador de la variable que se pretende medir. Así, el enunciado «Las mujeres no deberían meterse en política» puede funcionar como un ítem dirigido a medir la actitud hacia la participación de las mujeres en la política; o bien «las representaciones estudiantiles deberían participar en las decisiones sobre los planes de estudio», sería un ítem que mediría la actitud hacia la participación de los estudiantes en el gobierno universitario. De acuerdo al modo en que es construida la escala y al tipo de análisis al que da lugar, es corriente distinguir entre tres tipos de escalas:

- Centradas en el *sujeto*: se persigue ubicar al sujeto en el continuo actitudinal (típicamente, es el caso de las escalas Lickert);
- Centradas en el *estímulo*: se basan en adjudicar un valor a cada estímulo (o ítem), por lo que la estrategia consiste en localizar los ítems en el continuo (son las de Thurstone, entre las que se cuenta su “escala de intervalos aparentemente iguales”);
- Centradas en las *respuestas*: aquí el propósito es escalar tanto a los sujetos como a los ítems, lo que da lugar a una escala acumulativa (es el caso del análisis de escalograma de Guttman).

Lo que diferencia estos tres tipos de escalas no son los ítems considerados en sí mismos -éstos pueden ser los mismos-, sino la relación lógica que guardan entre ellos. Por lo demás, las tres formas son semejantes en cuanto suponen en principio el uso de jueces, miden en el nivel ordinal y se plantean como unidimensionales.²⁴

La escala Lickert

Nos extenderemos exclusivamente sobre ésta, por ser el tipo más simple en cuanto a su construcción y en cuanto a su flexibilidad, que la hace adaptable a propósitos variados. La escala Lickert consiste en un mero índice sumatorio. Se solicita la reacción del sujeto a una serie de ítems, cada uno de los cuales expresa tentativamente la actitud subyacente, y a los que el sujeto responde en términos de su grado de acuerdo-desacuerdo. Para esta escala, los ítems no deben escalonarse a lo largo del continuo, sino que el método de construcción supone seleccionar ítems que sean definidamente favorables o desfavorables con relación al objeto que se quiere medir. Sumando los puntajes obtenidos por un sujeto en cada ítem, le corresponderá un puntaje global que estará expresando su posición en el continuo. Los pasos a seguir en la construcción de la escala son los detallados a continuación.

1. La construcción de los ítems: se debe construir entre 30 y 50 ítems. Se trata esencialmente de un proceso imaginativo, construyendo ítems que resulten satisfactorios en términos de su validez facial o aparente por su relación con la característica que se pretende medir.²⁵ Para la construcción de cada ítem, Edwards (1957) ha formulado una serie de criterios:²⁶

- a. evitar ítems que apunten al pasado en vez del presente;
- b. evitar ítems que dan demasiada información sobre hechos, o aquellos que pueden ser interpretados como tales;
- c. evitar los ítems ambiguos;
- d. evitar los ítems irrelevantes con respecto a la actitud a medir;

²³ Puesto que una actitud neutral (ni positiva, ni negativa) se asocia generalmente a una baja intensidad, estas dos propiedades pueden combinarse en una única escala que mide simultáneamente dirección e intensidad.

²⁴ Sobre las escalas tipo Thurstone y Guttman, el lector podrá remitirse al trabajo bastante completo de Padua y Ahman (cf. Padua, 1979). En lo que respecta específicamente al análisis de escalograma de Guttman, se encontrará un excelente ejemplo completamente desarrollado en Wainerman (1976). Para un repertorio bastante completo de índices y escalas, cf. Miller (1991).

²⁵ «Los ítems son contruidos a partir de juicios que expresan una relación postulada a nivel de la teoría sustantiva, y de observaciones empíricas de grupos o sujetos que pertenecen a grupos o asociaciones que manifiestan la propiedad que se quiere medir. Los ítems así pueden ser extraídos de libros, publicaciones y artículos que tratan teóricamente sobre el objeto que se quiere medir. También puede el investigador acudir al análisis de contenido sobre discursos o manifiestos de individuos y asociaciones (por ejemplo, si se trata de medir radicalismo/conservadorismo, una fuente muy rica para la formulación de ítems son las declaraciones de grupos de interés: empresarios, grupos políticos de izquierda y de derecha, etc.)» (cf. Padua, 1979: 170).

²⁶ Los hemos reproducido a partir del texto de Padua (1979).

- e. los ítems deben expresar actitudes favorables o desfavorables con respecto a la actitud, y no expresar distintos puntos del continuo;
- f. evitar los ítems con los cuales todos o casi nadie concuerda;
- g. los ítems deben ser formulados en lenguaje simple, claro y directo;
- h. preferentemente, no exceder las 20 palabras por ítem;
- i. un ítem debe contener sólo una frase lógica;
- j. evitar el uso de palabras como 'todos', 'siempre', 'nadie', etc.;
- k. de ser posible, los ítems deben ser formulados con frases simples, y no compuestas;
- l. usar palabras que el entrevistado pueda comprender;
- m. evitar las negaciones, en particular las dobles negaciones;
- n. combinar los ítems formulados positiva y negativamente en una proporción de 50%-50%.

Aunque a veces obvias y algo redundantes, estas recomendaciones constituyen una buena base para evaluar críticamente cada ítem. Sin embargo, la experiencia muestra que no toman en cuenta un aspecto esencial como es el grado de competencia lingüística de la población a encuestar. Es un error presumir que cualquier persona está en condiciones de responder coherentemente a una batería de ítems. En sectores con muy bajo nivel de educación formal, este tipo de ítems simplemente no funciona, o bien demanda la reformulación sistemática de los enunciados por parte del encuestador, con lo que las respuestas pierden comparabilidad ya que lo son a estímulos diversos.

Cada ítem consiste en un juicio (o enunciado) sobre el cual el juez -o el sujeto- deberá expresar su grado de acuerdo o desacuerdo, para lo que se le ofrece un número de alternativas que puede ser de 3, 4, 5, 6 o 7.

Ejemplos de alternativas de respuesta

<i>3 alternativas</i>	<i>5 alternativas</i>	<i>6 alternativas</i>
<input type="checkbox"/> De acuerdo	<input type="checkbox"/> Muy de acuerdo	<input type="checkbox"/> Muy de acuerdo
<input type="checkbox"/> Ni de acuerdo ni en desacuerdo	<input type="checkbox"/> De acuerdo	<input type="checkbox"/> De acuerdo
<input type="checkbox"/> En desacuerdo	<input type="checkbox"/> Ni de acuerdo ni en desacuerdo	<input type="checkbox"/> Más bien de acuerdo
	<input type="checkbox"/> En desacuerdo	<input type="checkbox"/> Más bien en desacuerdo
	<input type="checkbox"/> Muy en desacuerdo	<input type="checkbox"/> En desacuerdo
		<input type="checkbox"/> Muy en desacuerdo

Algunos autores recomiendan incluir la alternativa "Sin opinión". Así, es sabido que muchas más personas contestarán "No sé" si esta alternativa les es expresamente ofrecida que si no lo es. El problema de no separar los "sin opinión" es que pueden terminar confundándose con una actitud neutra, cuando se trata de casos distintos.²⁷

Es recomendable que los ítems sean precedidos de un breve texto del tipo:
"Las siguientes afirmaciones son opiniones con respecto a las cuales algunas personas están de acuerdo y otras en desacuerdo. Indique por favor (marcando con una x en el casillero correspondiente) *la alternativa que más se asemeja a su opinión*".

2. Administración de los ítems a una muestra: una vez contruidos los ítems, se los debe aplicar a la población-objeto. Si se desea validar previamente la escala puede aplicarsela primero a una muestra de 50 a 100 personas que actuarán como jueces (idealmente estas personas deberán ser seleccionadas al azar de una población que presente características similares a la que se desea estudiar).

3. Asignación de puntajes a los ítems: aquí comienza propiamente hablando el análisis de la

²⁷ Se ha señalado, además, cómo « el tamaño de este efecto es parcialmente una función del modo en que la opción "sin opinión" es ofrecida», siendo el **efecto de filtro** mayor cuando la opción es dada mediante una pregunta previa y no sólo como una alternativa de respuesta (cf. Converse y Presser, 1986: 35).

escala. Cada ítem habrá sido clasificado *a priori* en positivo o negativo. Un ítem *positivo* es aquel que denotará una actitud favorable al objeto por parte de la persona que concuerde con él; por ejemplo, el ítem A6: «Los niños con SIDA deben ir a la misma escuela que los chicos sanos». En cambio, si alguien manifiesta su acuerdo con el enunciado del A3 -«Se debe prohibir el casamiento a las personas que estén enfermas de SIDA»- nos inclinaremos a pensar que su actitud hacia los portadores de SIDA es más bien desfavorable, por lo que éste será considerado un ítem *negativo*.²⁸

Si, por ejemplo, se ha optado por tres alternativas de respuesta, en un ítem positivo se asignarán 2 puntos a la persona que esté “de acuerdo”, 1 a la respuesta neutra, y 0 a “en desacuerdo”; los ítems negativos serán en cambio puntuados de modo inverso (0 para “de acuerdo”, 2 para “en desacuerdo”). La ventaja de contar con ítems positivos y negativos es controlar los efectos del *response-set*, es decir mantener controlados a los respondentes que tienden a contestar automáticamente por sí (*yeah-sayers*) o por no.

4. Asignación de puntajes totales para cada sujeto: se suman los puntajes obtenidos por cada encuestado. Asumiendo la hipótesis de que los ítems han sido bien elegidos -esto es, que están expresando efectivamente la actitud subyacente- se espera que las personas con actitudes favorables presenten puntajes totales elevados, mientras que las desfavorables obtendrán bajos puntajes. Así, por ejemplo, si la versión original de la escala consta de 30 ítems puntuados desde 2 hasta 0, los puntajes totales de cada sujeto podrán variar entre 0 y 60 puntos ($60 = 30 \times 2$).

Tabla 5.1: Ejemplo de matriz para el análisis de ítems

	Jueces	A1	A3	A6	A10	A12	A24	A27	Puntos
	1	2	2	2	2	2	2	0	12
	2	2	2	2	2	2	2	0	12
	3	2	2	2	2	2	2	0	12
	4	2	1	1	2	2	2	0	12
	5	2	2	2	2	2	2	0	12
	6	2	1	2	2	2	2	0	11
	7	2	1	2	2	2	2	0	11
	8	2	2	2	2	2	0	0	10
	9	2	0	0	2	2	2	2	10
	10	2	2	2	2	0	2	0	10
	11	2	2	2	2	0	2	0	10
	12	2	2	2	2	0	2	0	10
	13	2	2	0	0	2	2	0	8
	14	2	2	0	2	2	0	0	8
	15	2	0	0	1	2	2	1	8
	16	2	1	1	1	1	1	0	7
	17	2	0	0	0	2	2	1	7
	18	2	0	0	0	2	2	0	6
	19	2	0	0	0	2	0	1	5
	20	2	0	0	0	2	0	1	5
	Mediaa	2	1,8	1,8	2	2	2	0,4	
	Mediab	2	0,2	0,2	0,2	1,8	1	0,6	
	Difer.	0	1,6	1,6	1,8	0,2	1	-0,2	
	valor t	0	5,657	5,657	9	1	2,236	-0,426	
	valor p	>0,9999	0,0005	0,0005	>0,0001	0,3466	0,0558	0,6811	

5. Análisis de los ítems: se ordenan todos los sujetos según su puntaje, de mayor a menor; hecho lo cual se descarta el 50% central de la distribución para continuar trabajando solamente con los cuartiles 1 y 4. Así, sobre los veinte encuestados de la Tabla 5.1,²⁹ nos quedaríamos sólo con

²⁸ Estos ejemplos están tomados de una investigación realizada durante 1992 por estudiantes de Antropología Social de la UNaM.

²⁹ Esta tabla fue confeccionada con fines puramente ilustrativos: son muy pocos sujetos e ítems. En realidad, se trabajó con una muestra de 140 personas de Posadas.

diez: cinco constituirán el grupo “Alto” (los 5 que presentaran los más altos puntajes) y cinco el grupo “Bajo”. Para cada uno de estos dos grupos se calcula el puntaje promedio en cada ítem, pudiendo computarse luego la diferencia entre las medias. Esta es la base para analizar el poder discriminatorio de cada ítem: utilizando, por ejemplo, la *t* de Student, se pueden eliminar los ítems en los cuales la diferencia de las medias del grupo alto y del bajo no arrojan valores de *t* significativos. El procedimiento está basado en asumir la hipótesis de que en su conjunto los ítems están relacionados con la actitud que se persigue medir, de modo que se pueda eliminar a los ítems cuyos valores de respuesta no sean consistentes con los puntajes totales.

Así, en el ejemplo, se concluye fácilmente que el ítem A1 («Conviene que antes de casarse las personas se hagan una prueba para saber si están enfermas de SIDA») no discrimina: todos los jueces han respondido del mismo modo, por lo que la media es idéntica en ambos grupos y la diferencia es 0. Otros ítems malos son el A12 («Los trabajadores del hospital tienen derecho a negarse a atender a los enfermos de SIDA») y el A27 («El SIDA es un producto del libertinaje sexual»), en los que las diferencias entre las medias son mínimas y, consecuentemente, los valores de *t* no alcanzan un nivel de significación aceptable. De los ítems incluidos en la Tabla el mejor es el A10 («Una empresa debe tener derecho a despedir a un empleado enfermo de SIDA»): la diferencia entre las medias es de 1,8 -sobre un máximo posible de 2-, y el valor de *t* -“9”- es significativo al nivel de 0,0001.

6. La versión final de la escala: en definitiva, se conservan para la versión final únicamente los ítems cuya relación con el conjunto de la escala original se haya comprobado mediante algún test de significación, y se descarta por completo el resto. La cantidad de ítems seleccionados según su poder discriminatorio (como mínimo, seis) constituirá la escala final. Esta se podrá aplicar entonces a la población-objeto, en el caso de haber trabajado con una muestra de jueces, o sino simplemente se computará el puntaje definitivo de los sujetos en los ítems que integren la versión final de la escala.

El puntaje de cada individuo en la escala será igual a la suma de los puntajes parciales obtenidos en cada ítem. Así, si se trata de una escala de 10 ítems puntuados de 0 a 2, los puntajes de los individuos estarán comprendidos en el rango de 0 a 20 puntos. Estos puntajes extremos expresan claramente actitudes desfavorables y favorables, respectivamente. Pero la interpretación de los puntajes individuales que caigan entre los dos valores extremos es problemática, y no puede hacerse en términos absolutos. Dado que la escala mide en un nivel ordinal, y que la localización del punto neutro no es conocida, la interpretación de los puntajes intermedios sólo es posible en términos relativos al grupo estudiado.³⁰

3. CONFIABILIDAD Y VALIDEZ

El objetivo de las técnicas de recolección de datos (cuestionarios, tests, escalas, etc.) es obtener evidencia comprobada que sea significativa para responder a las preguntas que formulamos en nuestra investigación. Hemos visto que para ello debemos producir definiciones operacionales que nos permitan medir los conceptos en los que estamos interesados.

Cuando se trata de evaluar la bondad de cualquier instrumento de medición, el problema básico es definir cuáles deberán ser consideradas como *verdaderas* diferencias en la característica que se está midiendo distinguiéndolas de las variaciones que devienen de meros errores en la medición. Así, cuando me pregunto: “¿Qué es lo que mide mi instrumento?; “¿Son relevantes los datos que me proporciona para la característica que busco medir?”; o bien “¿Las diferencias que resultan de la aplicación de mi instrumento representan verdaderas diferencias en la característica a medir, o

³⁰ «No existe evidencia que sugiera que el punto neutro se corresponda con el punto medio de los valores escalares». dicen McIver y Carmines (1986: 28). Así, de una persona con 10 puntos en nuestro ejemplo no se puede afirmar que necesariamente tenga una actitud neutra hacia los enfermos de SIDA.

están reflejando la incidencia de otros factores?», estoy inquiriendo por la validez.³¹

Por otra parte, la confiabilidad se refiere a la consistencia entre medidas independientes del mismo fenómeno. En este caso, la pregunta es si el mismo instrumento utilizado en diferentes ocasiones y/o por diferentes investigadores producirá los mismos resultados.³² Validez y confiabilidad aparecen por lo tanto como dos requisitos que deberá satisfacer cualquier instrumento de medición satisfactorio. Abordaremos sucesivamente estos dos aspectos, para luego concluir acerca de la relación que mantienen entre sí.

3.1 Validez

Desde Louis Guttman en adelante, ha sido usual dividir la cuestión en validez ‘interna’ y ‘externa’. Según Zetterberg: «La diferencia principal es que la primera expresa una relación ‘lógica’, mientras que la segunda expresa una relación ‘empírica’. En otras palabras, la validez interna se puede apreciar sin estudios empíricos, mientras que la determinación de la validez externa consiste en poner a prueba una hipótesis» (1968: 92).

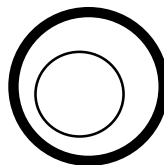
Validez interna

Para referirse a este tipo, suelen emplearse de manera indistinta las expresiones ‘validez lógica’ o ‘facial’,³³ así como la de ‘validez de contenido’.³⁴

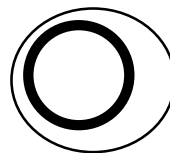
Aunque es evidente que la validez interna no es suficiente para evaluar la calidad de un instrumento, no deja de ser una condición necesaria, que debe ser satisfecha previamente a cualquier otro análisis de validez.³⁵ Al decir de Goode y Hatt: «Se refiere al análisis, teórico o de ‘sentido común’... y se emplea casi siempre debido a que dimana de la definición precisa del continuo» (1969: 292).

Retornemos al caso que presentábamos acerca de una medida de la idea de ‘satisfacción con el trabajo’. Justamente Zetterberg ha ilustrado este tema recurriendo a ese mismo ejemplo (1968: 92 y s.). Hablar de ‘validez perfecta’ implicaría que hubiera una coincidencia tal que el indicador tuviera exactamente el mismo alcance de contenido que la definición nominal del concepto. Suponiendo que simbolizáramos el alcance de la definición nominal del concepto por un círculo de trazo grueso y el del indicador por uno de trazo fino, es posible representar gráficamente los siguientes problemas típicos de validez:

a. La definición contiene el indicador, pero también otras cosas distintas: «¿ Está satisfecho con la ventilación de su lugar de trabajo?» (es obvio que la ‘satisfacción’ debe implicar algo más que la ventilación



b. El indicador contiene la definición, pero también otras cosas distintas: «¿ Le gusta vivir en Posadas?» (además de su trabajo en Posadas, a la persona puede gustarle, la ciudad, el clima, etc.).



³¹ Es en este sentido que Goode y Hatt definen sintéticamente: «Una escala tiene validez cuando verdaderamente mide lo que afirma que puede medir» (1969: 292).

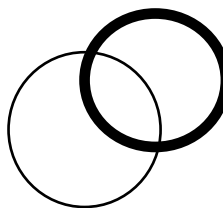
³² «Cabe confiar en una escala cuando produzca constantemente los mismos resultados al aplicarla a una misma muestra» (cf. Goode y Hatt, 1969: 289).

³³ Cf. Goode y Hatt, 1969: 292.

³⁴ Cf. Carmines y Zeller, 1979: 20.

³⁵ De lo contrario, se incurriría en el riesgo cierto de no saber -literalmente- de **qué** se está hablando.

c. El indicador y la definición coinciden parcialmente: «¿Le gustan sus amigos y conocidos aquí en Posadas?» (la satisfacción con los compañeros de trabajo puede pertenecer a la idea de 'satisfacción con el trabajo', no así la satisfacción con otros amigos).



Es para minimizar este tipo de errores que es necesario el análisis del concepto especificando sus principales dimensiones. Según Cronbach y Meehl: «la validez de contenido se establece demostrando que los distintos elementos del test equivalen a muestras de un universo que interesa al investigador. Este tipo de validez se suele establecer por lo general deductivamente, definiendo un universo de elementos y extrayendo sistemáticamente muestras del mismo para elaborar el test».³⁶

Para Carmines y Zeller, la validez de contenido -de mayor uso en Psicología y Educación que en Sociología y Ciencia Política- «depende del grado en que una medida empírica refleja un dominio específico de contenido. Por ejemplo, un test sobre las operaciones aritméticas carecería de validez de contenido si los problemas del test se focalizaran sólo en la suma, dejando de lado la resta, la multiplicación y la división» (1979:20). Un ejemplo en Sociología es la medida de Seeman de alienación.³⁷ Sobre la base de un examen cuidadoso de la literatura, Seeman concluyó que la 'alienación' podía ser concebida en términos de cinco dimensiones: falta de poder, falta de normas, carencia de sentido, aislamiento social y auto-extrañamiento. A su vez, cada una de estas dimensiones podría subdividirse: la 'falta de poder' podría ser considerada en sus aspectos político, social y económico.

Sin embargo, «es imposible especificar exactamente cuantos ítems han de ser desarrollados para cualquier dominio de contenido» (Carmines y Zeller, 1979: 21). En realidad, el proceso es muy complejo, tratándose de los conceptos abstractos que se encuentran típicamente en las ciencias sociales. Es difícil encontrar algún concepto teórico sobre el que exista consenso acerca del dominio de contenido pertinente para el fenómeno. Los conceptos simplemente no han sido descriptos con la exactitud requerida, por lo que no existe además modo alguno de muestrear el contenido. «Y, sin una muestra probabilística del contenido, es imposible asegurar la representatividad de los ítems particulares» (Carmines y Zeller, 1979: 22). Por estas razones, en la evaluación de las medidas en ciencias sociales no es suficiente la simple validez de contenido.

Validez externa

Por ende, debe intentarse "salir" del concepto en búsqueda de otras fuentes de validación. Deberán encontrarse criterios externos: lo que en inglés se denomina *criterion-related validity*.³⁸ El problema de la validez es que uno no conoce la verdadera posición de la persona en la variable que se desea medir, lo que hace que la estimación de la validez de una medida sea siempre problemática. Así, dicen Selltitz *et al.*: «A falta de un conocimiento directo de la verdadera posición de un individuo con respecto a la variable que se está midiendo, la validez de un instrumento se mide por el grado en que sus resultados son compatibles con otras evidencias significativas» (1968:181).

En este sentido, la validez *pragmática* será la que se obtenga al poder contestar afirmativamente a la pregunta: "¿Es útil este instrumento de medida?" Bastará entonces con correlacionar nuestro instrumento con un criterio externo: cuanto más alto el valor resultante del coeficiente de correlación que se utilice, mayor será la validez del instrumento; y ello, independientemente de cualquier otro tipo de consideración: no importa por qué razón nuestro instrumento es válido, será suficiente con que funcione. La validez pragmática se presta fácilmente a un uso ateórico y

³⁶ Cf. "Validez de los conceptos en los tests psicológicos", en Wainerman (Ed.), 1976:158.

³⁷ Citado por Carmines y Zeller, 1979: 20-21.

³⁸ Y que en la traducción castellana del manual de Selltitz *et al.* aparece como "validez pragmática".

empiricista.

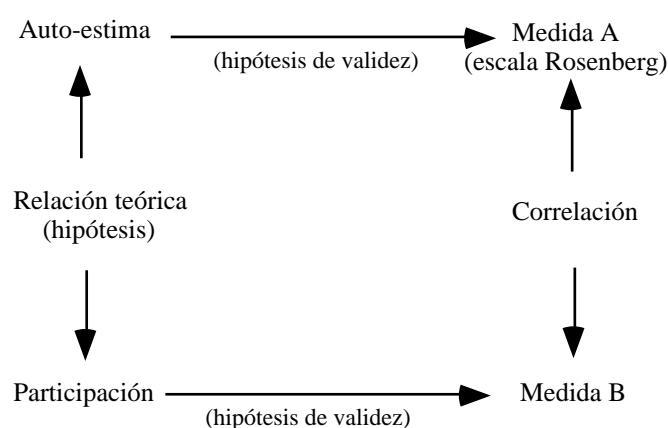
Técnicamente, se suele diferenciar entre dos tipos de validez pragmática. Si el criterio existe en el presente, se habla de validez *concurrente*. Por ejemplo, una respuesta verbal de un sujeto que afirma haber votado podría compararse con la información del padrón electoral. Por otro lado, la validez *predictiva* se refiere a un criterio futuro que se correlacionará con nuestro instrumento; los tests utilizados en la selección de personal para distintas ocupaciones son un típico ejemplo de validez predictiva. Ahora bien, la validación de criterio depende tanto de la medida del criterio externo como de calidad de nuestro propio instrumento de medición. Lo fundamental es que exista un criterio válido y razonable con el que las puntuaciones de nuestro test puedan compararse. Así, por ejemplo, si se tratara de un test que midiera la necesidad de atención psiquiátrica, el criterio externo podría estar dado por el diagnóstico de uno o varios psiquiatras. Pero, si existe realmente un criterio externo perfectamente seguro, cabe preguntarse cuál puede ser la utilidad de validar un nuevo instrumento alternativo; la ventaja podrá consistir en un costo menor, en una mayor rapidez de aplicación de la nueva medida, en la no necesidad de recurrir a un especialista, etc.

Sin embargo, en la práctica ocurre con frecuencia que no existe ningún criterio externo comprobado. Para la mayoría de las mediciones en ciencias sociales, la validación pragmática -concurrente o predictiva- no es aplicable debido a la ausencia de una variable-criterio relevante. Y cuanto más abstracto sea el concepto, más difícil resultará descubrir un criterio adecuado para evaluar la validez de su medición.

Validez constructiva

Según Cronbach y Meehl, ésta es una alternativa en el caso de la inexistencia de criterios disponibles totalmente válidos. Para ellos, la validez constructiva -también denominada “validez de conceptos”- «debe emplearse en todos aquellos casos en que no resulta aceptable ningún criterio o universo de contenidos para definir adecuadamente la cualidad que ha de medirse». ³⁹ Por su parte, Carmines y Zeller entienden que la validez constructiva se refiere al «grado en que una medida particular se relaciona con otras medidas consistentes con hipótesis derivadas teóricamente acerca de los conceptos (o construcciones) que están siendo medidos» (1979: 23).

Figura 5.5: Ejemplo de validación constructiva



Si, por ejemplo, se quiere evaluar la validez de una medida de la auto-estima -la Escala Rosenberg- y se cuenta con una hipótesis teórica acerca de la relación de la auto-estima con la participación en actividades escolares, se administra la escala a un grupo de estudiantes y se mide su participación, correlacionando luego ambas medidas. Si la correlación es positiva y elevada, se cuenta con un elemento de juicio que asienta la validez constructiva de la Escala Rosenberg. ⁴⁰

«La validez constructiva implica tres pasos distintos. Primero, debe especificarse la relación teórica entre los propios conceptos. Segundo, debe examinarse la relación empírica entre las medidas de los conceptos. Finalmente, la evidencia empírica debe ser interpretada en términos de

³⁹ En Wainerman (Ed.), 1976: 159.

⁴⁰ La escala Rosenberg está integrada por 10 ítems de este tipo:

SE1: « Siento que tengo muchas buenas cualidades.»;

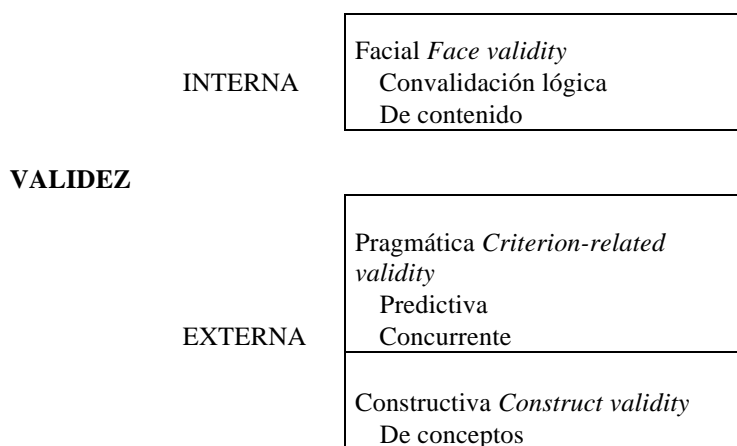
SE2: « Quisiera tener más respeto por mí mismo.»;

SE3: « Siento que soy una persona valiosa, al menos tanto como muchas otras.»;...

SE10: « Globalmente, me siento satisfecho conmigo mismo.» (cf. McIver y Carmines, 1982:32)

cómo clarifica la validez constructiva de la medida en cuestión»⁴¹. Este tipo de validación requiere la existencia de una red de conexiones teóricas que vincule el concepto a otras construcciones; pero tampoco es necesaria una teoría acabada completamente sistematizada: basta con que se puedan establecer varias hipótesis teóricamente derivadas que involucren al concepto. La validez constructiva será tanto mayor cuanto más numerosas sean las predicciones exitosas sobre diferentes variables relacionadas. Ahora bien, ¿Qué sucede si la predicción es errada?. Serán posibles varias interpretaciones. La interpretación más típica de la evidencia negativa es la falta de validez constructiva de la medida: el indicador no mide lo que pretende medir. El problema es que, lamentablemente, con igual rigor podría deducirse la incorrección del marco teórico, o bien todavía que es la medida de 'participación' la que carece de validez, con lo que se vuelve a plantear la necesidad de variables-criterio externas correctamente medidas y que tengan además relevancia teórica.⁴²

Figura 5.6: Tipos de validez



3.2 Confiabilidad

El otro requisito que debe cumplimentar todo instrumento de medición es la confiabilidad. Para Carmines y Zeller: «la confiabilidad se refiere al grado en que un experimento, un test, o cualquier instrumento de medida produce el mismo resultado en ensayos repetidos» (1979: 10). Así, por ejemplo, un metro elástico que se estirara en mayor o menor grado al aplicarlo no sería un instrumento confiable para medir la longitud. Básicamente, hay cuatro técnicas para estimar la confiabilidad: test-retest, formas alternativas, división por mitades e intercorrelación media. Las presentaremos sintéticamente.

Test-retest

Consiste en aplicar dos veces el mismo instrumento de medición a una misma población en ocasiones diferentes. Se supone que cuanto mayor sea la coincidencia entre ambas medidas, mayor será la confiabilidad. El problema es que la primera aplicación del instrumento puede afectar las respuestas a la segunda aplicación; es sabido que existe en los sujetos una tendencia a aparecer como consistentes, por lo que su memoria puede convertirse en fuente de una confiabilidad espúrea. Esta tendencia se agrava cuanto más breve es el lapso que media entre ambas mediciones.

⁴¹ Cf. Carmines y Zeller, 1979: 23.

⁴² Simple aplicación de la ley de De Morgan: la negación de una conjunción de hipótesis equivale a la disyunción de sus negaciones. Como explica M. Mora y Araujo: «En la validación externa elegimos algunas hipótesis para utilizarlas como hipótesis auxiliares, y esto se justifica simplemente en términos de una decisión teórica que considera a dichas hipótesis una condición necesaria de la validez del índice (estrictamente hablando, la falsedad de dichas hipótesis es una condición suficiente de la invalidez del índice). Un índice de prestigio ocupacional que no guarde ninguna relación con el nivel educacional sería difícilmente aceptable, por ejemplo» (cf. 1971: 26).

Pero, si en cambio se alarga ese período, aumenta la probabilidad de que la confiabilidad sea espúreamente baja, debido a la posibilidad de que se haya producido un cambio *real* en el fenómeno.⁴³ Así, la actitud de una persona hacia la pena de muerte podría variar por haber, por ejemplo, asistido a una ejecución entre las dos mediciones.

Formas alternativas (o múltiples)

Es una técnica muy utilizada en Educación para estimar la confiabilidad de todo tipo de tests. En este caso, se trata de dos instrumentos distintos que se aplican a una misma población, recomendándose un lapso de unas dos semanas entre ambas medidas. Se elimina el problema de una alta confiabilidad espúrea que pudiera ser inducida por la memoria de los sujetos (puesto que ambos tests **son** diferentes). Pero subsiste la imposibilidad de distinguir entre el cambio real y una baja confiabilidad; por esta razón, los resultados de esta técnica son más fáciles de interpretar si el fenómeno investigado es relativamente estable en el tiempo (por ejemplo, un test de aritmética). El principal problema radica en producir dos formas efectivamente *equivalentes* del mismo test; así, una baja correlación entre ambas formas puede no estar significando una baja confiabilidad sino deberse a la no equivalencia de las formas alternativas.

División por mitades

La ventaja de esta modalidad es que no requiere dos aplicaciones separadas en el tiempo; a la misma población se le aplica en una sola ocasión un test dividido en dos mitades. «El conjunto total de ítems se divide en dos mitades y los puntajes en cada mitad se correlacionan para obtener una estimación de la confiabilidad. Las mitades pueden considerarse aproximaciones a formas alternativas» (Carmines y Zeller, 1979: 41). Sin embargo, surge alguna indeterminación en la estimación de la confiabilidad, debido a las múltiples combinaciones para la división de los ítems en mitades. Típicamente, los ítems se dividen en pares e impares, o bien se los divide al azar. De hecho, para una escala de 10 ítems, las posibilidades distintas de división son 125; y ocurre que distintas divisiones producen correlaciones ligeramente diferentes entre las dos mitades.

Intercorrelación media

Se le denomina también ‘método de la consistencia interna’, y permite evitar las variaciones que obedecen a diferencias en la división por mitades. Este procedimiento proporciona el índice más estable de confiabilidad, estimada a partir de la correlación entre todos los ítems. Se utiliza el α de Cronbach:

$$\alpha = N r / [1 + r (N - 1)]$$

En el que N es igual al número de ítems, y r representa la correlación media interítem, que es la suma de las correlaciones entre cada par de ítems dividida por el número de pares de ítems. Alpha varía entre 0 y 1; cuanto mayor sea el número de ítems y mayor la correlación media interítem, mayores será, alpha y la confiabilidad del instrumento. Vale decir que un agregado de ítems a la escala incrementará su confiabilidad siempre y cuando no produzca una reducción de la intercorrelación media. Pero, como ya hemos visto, «el agregado de ítems indefinidamente tiene un impacto progresivamente menor sobre la confiabilidad» (Carmines y Zeller, 1979: 46).

3.3 Confiabilidad y validez

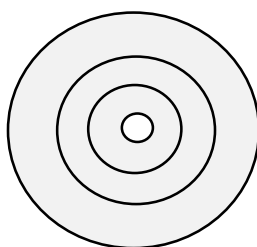
Es usual clasificar los errores no-muestrales⁴⁴ en errores *constantes* (sistemáticos, o de sesgo) y de *azar* (o variables). Este es un buen punto de partida para establecer la relación entre confiabilidad y validez. Según se puede leer en Selltitz: «Un error constante es el introducido en la medida por algún factor que sistemáticamente afecta a la característica objeto de medida o al propio proceso de medida... El error de azar es debido a aquellos aspectos pasajeros de la persona, de la situación de

⁴³ “ Por esta razón, se dice que los procedimientos de test-retest miden la confiabilidad en términos de *estabilidad* de los hallazgos” (cf. Smith, 1975: 59).

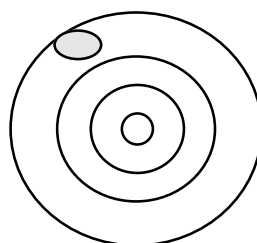
⁴⁴ Esto es, los que devienen del instrumento de medida y su aplicación, y del procesamiento de los datos, y no de la selección de las unidades de análisis

medida, del procedimiento de medida, etc., que parecen variar de una medición a otra, aunque la característica que se intenta medir no haya cambiado» (1965: 179-180). Podría pensarse que la validez se relaciona con el error constante en tanto que la confiabilidad depende del error de azar. Sin embargo, es preciso tener en cuenta que la validez del instrumento depende totalmente de su confiabilidad, mientras que ésta es independiente de aquella. Podemos ejemplificar esto sencillamente imaginando que estamos tirando al blanco con un rifle.⁴⁵ El rifle representa nuestro instrumento; el centro del blanco, la característica que pretendemos medir. Se pueden dar las tres situaciones que ejemplificamos en el diagrama.

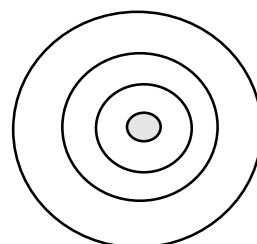
a) en el primer caso, los disparos se dispersan al azar: nuestro rifle carece de validez tanto como de confiabilidad;



b) aquí, los disparos se concentran sistemáticamente en el sector superior izquierdo del blanco: nuestro instrumento es confiable, aunque no válido;



c) los disparos se concentran sistemáticamente en el centro del blanco: el instrumento es confiable y, por añadidura, también válido.



El ejemplo permite apreciar cómo la confiabilidad es una condición necesaria para la validez. Como dice H. Zetterberg: «Confiabilidad es la medida en que un indicador nos provee de registros no ambiguos. La confiabilidad es un prerequisite necesario de la validez. Los instrumentos no confiables carecen siempre de validez» (1968: 98).⁴⁶ Incluso, se puede demostrar que la validez de una medida no puede superar nunca la raíz cuadrada de su coeficiente de confiabilidad; por lo que si, por ejemplo, la estimación de la confiabilidad de nuestro instrumento es de .64, será imposible que se pueda correlacionar con otra variable con un coeficiente superior a .80 (cf. Smith, 1975: 78).

Si la validez se refiere al hecho de que nuestro instrumento mida realmente lo que pretendemos medir, puede perfectamente ocurrir que nuestra medida sea confiable sin ser válida: en cuyo caso estaríamos midiendo **otra** cosa, diferente de la característica que nos interesa. Pero si en cambio nuestro instrumento no es confiable, entonces no estará midiendo *ninguna* cosa: la falta de confiabilidad es una condición suficiente de invalidez.

⁴⁵ El ejemplo está tomado de Nuria Cortada de Kohan (1968: cap. 3).

⁴⁶ Lo mismo puede leerse en Sellitz: «cualquier falta de fiabilidad de un test disminuye su validez» (1968: 205).

CAPÍTULO 6

INTRODUCCIÓN AL ANÁLISIS DE CORRESPONDENCIAS

Denis Baranger y Fernanda Niño

En base a la obra fundamental de Jean-Paul Benzécri (1973) tuvo su origen la denominada ‘escuela francesa de análisis de datos’ (*analyse des données*), que ha desarrollado una serie de métodos de análisis multivariados como el análisis factorial de las correspondencias (AFC) y el análisis de correspondencias múltiples (ACM), y renovado otros como el análisis en componentes principales (ACP) y las técnicas de clasificación automática.¹

En particular el análisis de correspondencias es una técnica cuyo uso se ha ido popularizando en los últimos años. En este proceso tuvo mucho que ver la utilización que hizo Pierre Bourdieu del ACM en *La distinction* (1979) y en su producción posterior.² Aunque sin llegar al extremo de lo que sucede en Francia, donde es habitual que las revistas de actualidad publiquen planos factoriales para ilustrar los resultados de encuestas, la técnica del ACM se ha ido abriendo un camino y su uso se ha ido generalizando en otros países tanto en el campo de la investigación básica en ciencias sociales como en los estudios de *marketing* y de opinión pública.

No nos referiremos aquí a los fundamentos del análisis factorial de correspondencias (AFC), el que puede ser considerado como una generalización del análisis en componentes principales (ACP) adaptado al procesamiento de datos cualitativos.³ Nuestro propósito en este capítulo no es reemplazar a una indispensable formación específica en esta materia;⁴ más bien nos proponemos realizar una presentación que ilustre acerca de las potencialidades de las técnicas multivariadas de análisis de datos, brindando al lego algunos elementos para una mejor interpretación de un modo de presentación de los resultados de la investigación al que se está viendo crecientemente confrontado.

Examinaremos aquí algunos ejemplos de usos de estas técnicas multivariadas. Comenzaremos con el análisis de correspondencias a partir de la presentación de dos ejemplos de tablas de contingencia. Luego abordaremos el ACM mostrando cómo puede ser visto simplemente como una forma más de reducción de un espacio de propiedades basado en variables cualitativas, y cómo es posible utilizarlo como una alternativa al procedimiento numérico corrientemente utilizado en la construcción de índices.⁵ Por último ejemplificaremos cómo se puede combinar el ACM con técnicas de clasificación automática en el análisis de encuestas.

1. EL ANÁLISIS FACTORIAL DE CORRESPONDENCIAS

Si bien el Análisis Factorial de Correspondencias (AFC) es un método especialmente adecuado para explorar tablas de grandes dimensiones, comenzaremos analizando una tabla de 4x4 cuyo significado es relativamente simple de aprehender, pero que servirá para introducir los principios

¹ Todos éstas constituyen, según Lebart, “métodos de estadística descriptiva multidimensional”, los que comprenden dos grandes familias: los métodos factoriales y los métodos de clasificación; ambas familias de técnicas son corrientemente utilizadas de modo complementario en el análisis de datos de encuesta (Lebart; 1989: 23).

² Tanto en la revista que dirige Bourdieu, *Actes de la Recherche*, como en las obras en que da cuenta de sus investigaciones empíricas y en aquellas provenientes de su grupo de colaboradores es recurrente la presentación de resultados mediante el uso de planos factoriales.

³ Cf. Tenenhaus (1994: 145).

⁴ Durante los últimos años el Programa PRESTA <<http://www.ulb.ac.be/assoc/presta/>>, desarrollado desde la U. Libre de Bruselas con el apoyo de la Comunidad Europea, ha encarado esta tarea a nivel de América del Sur.

⁵ Esta aplicación del ACM se inscribe dentro de la alternativa que Lazarsfeld (1993) y Barton (1969) denominaban ‘reducción funcional de un espacio de propiedades’.

básicos de esta técnica.

Tabla 6.1: Buenos Aires, 1982 - Ocupación del hijo y ocupación del padre

Ocupación del padre	Ocupación del hijo				Total
	Manual Baja	Manual Alta	No manual Baja	No manual Alta	
No manual Alta	2	5	14	72	93
No manual Baja	18	28	119	56	221
Manual Alta	23	58	50	17	148
Manual Baja	69	51	52	18	190
Total	112	142	235	163	652

Fuente: datos adaptados de Jorrat, 1997: 102.

Para determinar la existencia de una situación de independencia estadística entre dos variables, disponemos del coeficiente de asociación φ^2 (fi cuadrado). Empero φ^2 , como cualquier coeficiente, es una medida sintética que no da cuenta de la estructura de la relación entre las variables. Si, en cambio, estamos interesados en analizar esa estructura, será necesario descomponer esa relación y describir su *forma*, identificando cuáles son los valores de cada variable que se asocian más entre sí. En la práctica, éste es el tipo de análisis que presenta un mayor interés.

En el ejemplo, el φ^2 podría ser alto, tanto porque los hijos tendieran a mantener el mismo tipo de ocupación de sus padres, como porque la cambiaran en una determinada dirección.⁶ Por otro lado, también podría ocurrir que el valor del φ^2 fuera bajo, y que aún así existieran ciertas asociaciones entre modalidades, y que éstas –aún siendo débiles- resultaran de interés para el problema estudiado.

En la Tabla 6.1 el estudio de la relación entre estas dos variables podría traducirse en las siguientes preguntas: ¿es idéntica la distribución en ocupaciones de los hijos de padres de diferentes categorías ocupacionales (o hay una tendencia de los hijos a repetir las ocupaciones de sus padres)?⁷ Y por otra parte: ¿para los distintos grupos de hijos difiere la repartición de sus padres en las diferentes ocupaciones?

En una tabla de dimensiones reducidas, es posible analizar estas cuestiones mediante el cálculo de porcentajes en fila o en columna, según cuál sea la pregunta a la que se busque responder, y comparando las filas o las columnas entre ellas o con respecto al marginal (cf. *supra*, capítulo 4).

Sabemos que si existiera *independencia* entre las ocupaciones de padres e hijos, debería observarse similitud en la repartición proporcional de las ocupaciones de los hijos (el perfil ocupacional de los hijos), para cada grupo de padres, pero también coincidencia en el perfil ocupacional de los padres para cada categoría ocupacional de los hijos. En efecto, si recordamos que los marginales constituyen promedios ponderados de las distribuciones en el interior de las filas (o de las columnas), entonces en caso de independencia, esos *perfiles* deberían coincidir con la distribución en el total de las observaciones (perfil marginal). Así el análisis de una tabla se reduce a analizar en qué medida las distribuciones proporcionales en fila (o en columna) se diferencian del *perfil medio*. Consideremos la comparación de los perfiles fila:

⁶ Para la tabla 6.1, el coeficiente V de Cramer –derivado de fi cuadrado- arroja un valor de 0,38; en cambio, gamma, que saca partido del carácter ordinal de ambas variables nos da 0,59, indicando ya una asociación positiva.

⁷ Para un análisis detallado de estas cuestiones –que escapan a nuestro objetivo en esta presentación- recomendamos la lectura del artículo de Jorrat (1997), en el que además se introduce toda una serie de técnicas específicamente adaptadas al estudio de problemas de movilidad.

Tabla 6.1.1: Tabla de perfiles-fila

Ocupación del padre	Ocupación del hijo				Total
	Manual Baja	Manual Alta	No manual Baja	No manual Alta	
No manual Alta	0,022	0,054	0,151	0,774	1,000
No manual Baja	0,081	0,127	0,538	0,253	1,000
Manual Alta	0,155	0,392	0,338	0,115	1,000
Manual Baja	0,363	0,268	0,274	0,095	1,000
Total	0,172	0,218	0,360	0,250	1,000

Fuente: Tabla 6.1

En las celdas de la diagonal, que corresponden a los hijos cuya categoría ocupacional coincide con la de su padre, las proporciones son siempre mayores a la proporción marginal correspondiente, lo que pone en evidencia que las ocupaciones de padres e hijos tienden a coincidir.⁸ Sin embargo, no todos los hijos pertenecen a la misma categoría ocupacional de su padre.

Si en la tabla se comparan las filas entre sí, se observa que la distribución proporcional de las ocupaciones de los hijos (perfil) para cada una de las categorías de ocupación de los padres, difiere del perfil ocupacional de los hijos en la población.

De la misma manera, podemos comparar las columnas para determinar si entre los individuos con diferentes posiciones ocupacionales, se observan diferencias en la ocupación de sus padres (ver Tabla 6.1.2). Así, entre los hijos que tienen ocupaciones manuales hay mayor frecuencia de padres con ocupaciones también manuales. En general, los hijos con ocupaciones manuales presentan un perfil ocupacional de sus padres muy distinto al de los hijos con ocupaciones no manuales.

Tabla 6.1.2: Tabla de perfiles-columna

Ocupación del padre	Ocupación del hijo				Total
	Manual Baja	Manual Alta	No manual Baja	No manual Alta	
No manual Alta	0,018	0,035	0,060	0,442	0,143
No manual Baja	0,161	0,197	0,506	0,344	0,339
Manual Alta	0,205	0,408	0,213	0,104	0,227
Manual Baja	0,616	0,359	0,221	0,110	0,291
Total	1,000	1,000	1,000	1,000	1,000

Fuente: Tabla 6.1

Tratándose de una tabla de pequeño tamaño, el análisis mediante el procedimiento de comparar porcentajes no presenta mayores dificultades. Distinto es el caso cuando la tabla es de grandes dimensiones, situación en la que el AFC se revelará de la mayor utilidad. Lo que se logra mediante

⁸ Como lo expresa Jorrat: «...la concentración de los casos en la diagonal principal es un indicador de la tendencia a la heredad y autoreclutamiento ocupacional» (1997: 105).

el AFC es transformar la tabla de datos original en un gráfico que permite *visualizar* las asociaciones existentes entre las modalidades.⁹ En el gráfico producido por el AFC cada perfil estará representado por un punto, de manera tal que será posible comparar la similitud entre los perfiles-fila (o columna)¹⁰ y también la diferencia de cada uno de éstos con respecto al perfil global, de un modo en todo análogo al procedimiento utilizado en el análisis numérico.

Si la relación entre las variables puede estudiarse como la diferencia entre los perfiles fila/columna con respecto a los marginales fila/columna (los perfiles medios), es posible representar gráficamente esa dispersión mostrando a cada uno de los perfiles como un punto en un espacio cuyo origen de coordenadas está dado por el perfil medio. De este modo la diferencia entre perfiles se traduce en una distancia entre los puntos correspondientes; asimismo, la diferencia entre un perfil-fila (o columna) con respecto al perfil medio se traduce en la distancia entre cada punto perfil y el origen del sistema de coordenadas.

Las coordenadas de cada punto en el espacio de representación quedan determinadas por sus frecuencias relativas en la tabla de puntos-perfiles. En nuestro ejemplo, los puntos-perfiles *ocupación del padre*, tendrán cuatro coordenadas y por lo tanto deberán representarse en un espacio de cuatro dimensiones. Así, las coordenadas para el punto correspondiente a los padres con ocupación manual alta serían: 0,155; 0,392; 0,338; y 0,115. De la misma manera se representarían los puntos *ocupación del hijo*. Con este criterio de representación, la proximidad de los puntos se deberá interpretar como similitud entre perfiles: así, en nuestro ejemplo, esperamos que el punto 'Padre no manual alta' se sitúe muy lejos del punto 'Padre manual baja'.

Para obtener un indicador de la diferencia entre perfiles o de la variabilidad de los puntos con respecto al perfil medio, se puede sumar las distancias de todos los puntos al centro de gravedad (origen del sistema), pero si se quiere que esa medida de la dispersión tome en cuenta la parte de la población representada en cada punto, entonces las distancias deben ponderarse por el peso relativo de cada categoría en la población.¹¹ Así se obtiene un coeficiente indicador de la dispersión de los perfiles con respecto al perfil medio y que por lo tanto mide la relación entre las variables. Se puede demostrar que esa cantidad conocida como *inercia de la nube de puntos*¹² es igual al φ^2 de la tabla.

Empero, se plantea el problema de la imposibilidad de *visualizar* las distancias entre los perfiles (filas/columnas) en los espacios originales, dado que estamos tratando con espacios de más de tres dimensiones. De lo que se trata entonces, es de generar un nuevo espacio de representación que permita esa visualización, a partir de sucesivas representaciones gráficas planas, de modo tal que estas visualizaciones en el plano conserven lo más posible las distancias originales entre los puntos.¹³

Para construir ese nuevo espacio de representación, se busca un primer eje tal que las coordenadas de cada punto perfil en él traduzcan lo mejor posible la distancia de los individuos en el espacio original; luego, un segundo eje que dé cuenta de las diferencias que el primero no pudo resumir, y así sucesivamente.

De esta manera el nuevo espacio de representación –que supone un cambio de coordenadas– si bien sigue siendo *multidimensional*, permite establecer una jerarquía entre los ejes, ya que se construye de manera tal que el primer eje resume la mayor parte de la inercia y por lo tanto es la mejor aproximación a la nube de puntos original, que el segundo eje da cuenta de la mayor parte

⁹ Según lo define Bry, «el análisis factorial es un principio geométrico que permite convertir automáticamente una gran tabla de datos en imágenes sintéticas que hacen visibles sus principales estructuras» (1995: 3).

¹⁰ En realidad se trata de dos representaciones gráficas –una para cada tabla de perfiles (fila y columna)– que permitirán lecturas por comparación semejantes a las realizadas sobre las tablas.

¹¹ En realidad, se hace la suma ponderada de los *cuadrados* de las distancias de cada punto al centro de gravedad.

¹² «El término inercia (o más específicamente 'momento de inercia') es tomado de la mecánica (...) La inercia tiene una interpretación geométrica como una medida de la dispersión de los perfiles en el espacio multidimensional. A mayor inercia, mayor dispersión de los perfiles» (Greenacre; 1994: 12).

¹³ Como señala J. Rovin: «De acuerdo al objetivo del análisis de correspondencias, las primeras dimensiones representan habitualmente la mayor parte de la variabilidad de la matriz de datos original. Si este es el caso, la posición de las proyecciones de los puntos-perfiles en un subespacio de menores dimensiones es una buena aproximación de la posición de los perfiles en el espacio original multidimensional» (en Greenacre y Blasius; 1994: 210).

posible de la variabilidad *residual*, etc.

Se descompone así la *inercia total* o *fi cuadrado de la tabla*, en una suma de variancias proyectadas sobre cada eje, con la particularidad de que esas variancias decrecen del primero al último eje. Con esta forma de construcción del nuevo espacio, el análisis de una nube de puntos en más de tres dimensiones, puede reducirse habitualmente a las dos o tres primeras, y por lo tanto a uno o dos planos, obteniendo una buena aproximación de la nube de puntos original.

Aunque es frecuente que un AFC se resuma en un único gráfico cuando se trata de presentar los resultados de una investigación, lo cierto es que el plano factorial en sí mismo es sólo una parte del producto del AFC, el que además comprende toda una serie de tablas y de medidas numéricas que resultan indispensables para la correcta interpretación de esos resultados.

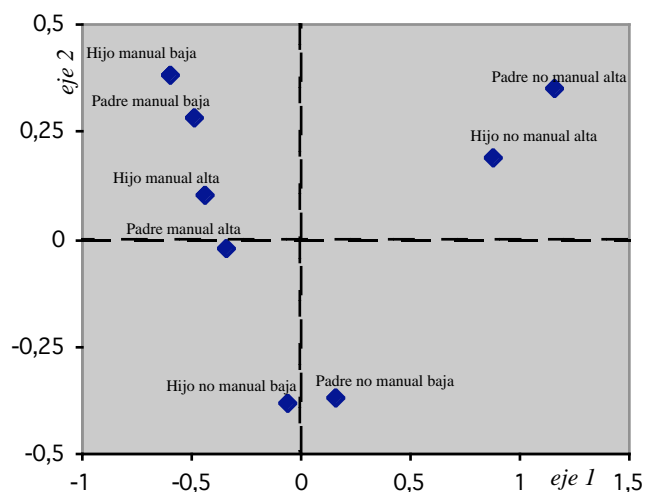
Figura 6.1: Histograma de los valores propios

Factor Nº	Valor propio	%	% acum.	
1	.2978	71.24	71.24	
2	.0870	20.82	92.06	
3	0.332	7.94	100.0	

Suma de los valores propios = 0,4180

En la Figura 6.1 se presenta la descomposición de la inercia, o φ^2 de la tabla, en tres ejes o factores. A cada eje le corresponde un valor propio que indica la cantidad de inercia o variabilidad de la que da cuenta. Así, el primer eje, con un valor propio, o *eigenvalue*,¹⁴ de 0,2978, resume un 71,24% de la inercia total, mientras que el segundo eje da cuenta de un 20,82%. De este manera el plano conformado por los dos primeros factores concentra un 92,06% de la inercia, y por lo tanto constituye una buena aproximación gráfica de la tabla original. La suma de los valores propios - 0,4180- es igual a φ^2 .

Figura 6.2: Posición de las ocupaciones de padres e hijos en el plano (1,2)



En una lectura inmediata del plano, teniendo en cuenta que el origen de las coordenadas representa el perfil medio, las categorías más diferentes de ese perfil medio aparecerán más

¹⁴ Este es el término de uso habitual en la literatura anglosajona sobre análisis factorial, referido a la raíz característica, que es una propiedad matemática de una matriz (cf. Kim y Mueller; 1978).

alejadas, esto es, con coordenadas más altas –positivas o negativas-. Así en el primer eje se oponen las ocupaciones manuales y las no manuales altas tanto para los hijos como para los padres, modalidades que se alejan en direcciones opuestas respecto al centro, lo que traduce una gran diferencia en sus perfiles. Por otro lado, las categorías manuales altas y bajas aparecen con coordenadas próximas porque sus perfiles son semejantes. Así, en el primer factor -eje horizontal- los perfiles de padres e hijos se ordenan consistentemente desde las ocupaciones manuales bajas hasta las no manuales altas, mientras que el segundo eje aparece oponiendo las ocupaciones bajas no manuales al resto.

Como regla general:

La proximidad entre dos puntos-perfiles -entre dos categorías de una misma variable- indica una similitud entre sus perfiles. En cambio, la proximidad entre elementos de diferentes variables indica asociación entre las modalidades.

Así, por ejemplo, para la variable ‘ocupación del hijo’ la proximidad entre las categorías *Ocupación manual baja* y *Ocupación manual alta* se explica porque son semejantes en sus perfiles, ambos caracterizados por un excedente de padres con ocupación manual baja: mientras en el total de hijos hay un 29.1% con padres de ocupación manual baja, en esas categorías los porcentajes son de 61,6 y 35,9 respectivamente (cf. tabla 6.1.2).

En el primer eje las modalidades de ocupación no manual baja de padres e hijos se ubican prácticamente en el origen, lo que significa que en ese eje no están representadas sus diferencias respecto al perfil medio. Es el segundo eje el que da cuenta de las diferencias de esos perfiles con respecto a sus correspondientes marginales, y que marca la diferencia entre estos perfiles y los restantes.

Tabla 6.2: Coordenadas, contribuciones y cosenos cuadrados de los puntos-perfiles o modalidades

Puntos- perfiles	Dist.	COORDENADAS			CONTRIBUCIONES			COSENOS CUADRADOS		
		1	2	3	1	2	3	1	2	3
HNoA	.81	.88	.19	.01	64.6	10.3	.1	.96	.04	.00
HNoB	.15	-.06	-.38	-.07	.4	58.4	5.2	.02	.95	.03
HMaA	.30	-.44	.10	.31	14.4	2.7	61.1	.66	.04	.31
HMaB	.57	-.60	.38	-.25	20.6	28.7	33.6	.63	.23	.11
PNoA	1.48	1.16	.35	.04	64.7	20.3	.7	.91	.08	.00
PNoB	.17	.16	-.37	-.09	2.9	54.2	9.0	.15	.80	.05
PMaA	.22	-.34	-.02	.32	8.7	.1	68.5	.53	.00	.47
PMaB	.34	-.49	.28	-.16	23.6	25.5	21.8	.71	.22	.07

Fuente: Tabla 6.1.

Dentro del producto de un AFC se incluye generalmente una tabla en la que figuran distintos indicadores que permiten afinar la interpretación de los resultados visibles en el plano factorial. Así la *distancia al origen* (‘dist.’) está indicando para cada punto cuanto se aleja globalmente del perfil medio dentro del espacio original. Las *coordenadas* en los tres ejes indican la posición de cada punto en cada uno de los factores, y cuando asumen valores extremos es porque se trata de un perfil muy diferente al del perfil medio.

Las *contribuciones* nos informan para cada punto acerca de su aporte a la inercia resumida en cada eje, y por lo tanto sobre la intervención del punto en cuestión en la variabilidad explicada por ese eje. Este aporte está dependiendo tanto de la diferencia de ese perfil al marginal o perfil medio, como de su peso relativo en la población total. De esta manera, una contribución importante de un elemento puede devenir de una gran diferencia respecto al perfil medio, aunque su peso no sea tan grande. Es lo que ocurre con la modalidad ‘PNoA’,¹⁵ la que más contribuye al primer factor (64,7%), porque su coordenada en el primer eje es muy alta. Este primer eje está *construido*

¹⁵ Es corriente recurrir al uso de etiquetas abreviadas para representar a las modalidades tanto en los planos factoriales como en las tablas. Así, ‘HNoA’ deberá leerse: ‘hijos con ocupación no manual alta’, etc.

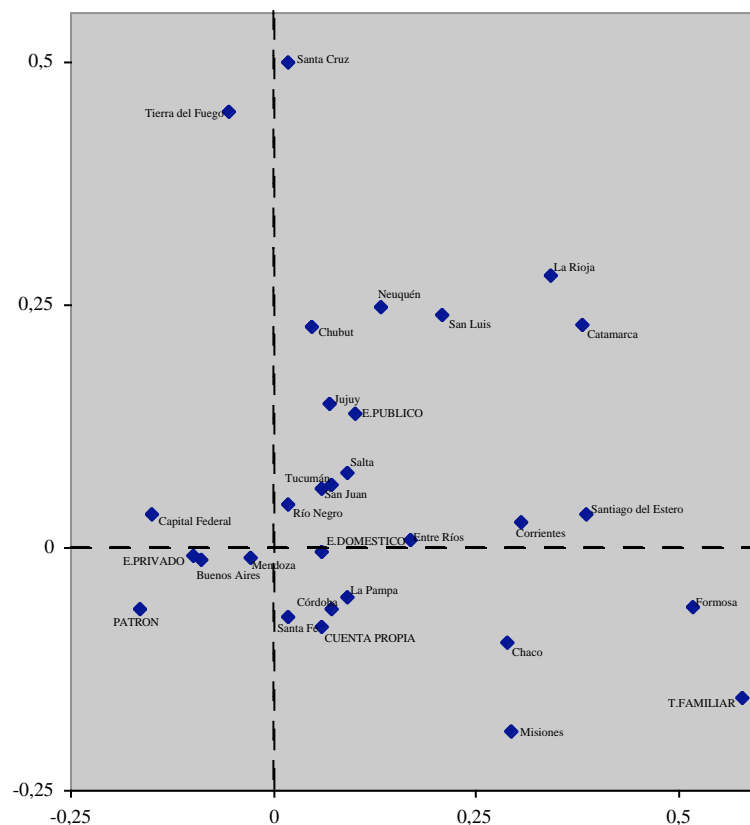
especialmente por ese punto y el punto ‘PMaB’ (contribución 23.6%).

La parte de la tabla bajo el título *cosenos cuadrados*, indica para cada punto su calidad de representación en todos los ejes –expresada en porcentajes–, lo que significa en qué medida el eje traduce la diferencia de ese punto-perfil con respecto al perfil medio. Por otro lado, como un punto estará totalmente representado en todos los ejes, la suma de esos índices a lo largo de todos los ejes da 100. Por ejemplo, la diferencia al perfil medio del punto HNoA está casi totalmente representada en el primer eje (96%), y totalmente representada en el primer plano (96% + 4% = 100%). En cambio en el primer eje la diferencia al perfil medio del punto HNoB no está bien traducida, puesto que su coseno cuadrado es de 2%; sin embargo sus características son traducidas por el segundo eje, donde su calidad de representación es del 95%.

El AFC es esencialmente una técnica de exploración de los datos que nos permite llegar a los mismos resultados que el análisis numérico convencional. Empero, sus ventajas son más evidentes cuando se trabaja sobre tablas de grandes dimensiones, al permitir una visualización de conjunto de la estructura de los datos.

El siguiente ejemplo basado en datos del Censo de 1980 nos permitirá ilustrar rápidamente este punto. Se trata de una tabla de contingencia en la que están dispuestas en filas las 24 provincias argentinas, y en columnas seis categorías ocupacionales: ‘patrón o socio’, ‘cuenta propia’, ‘trabajador familiar sin remuneración’, ‘empleado privado’, empleado público’ y ‘empleado en servicio doméstico’.

Figura 6.3: Argentina, 1980 - Provincias y posiciones ocupacionales



Inmediatamente salta a la vista como las distintas regiones del país se ubican en distintas porciones del plano: la región nordeste en la zona inferior derecha, el noroeste en la superior derecha, etc. Este ordenamiento es producto de las posiciones similares que ocupan en los dos primeros factores las provincias de una misma región. El primer eje opone un polo “moderno” representado por Capital Federal y la Provincia de Buenos Aires a otro “tradicional” o “subdesarrollado”. En lo que hace a las posiciones ocupacionales, este factor diferencia los distritos con mayor proporción de “empleados en el sector privado” –significando una mayor desarrollo de las relaciones de producción propiamente capitalistas– de aquellos en los que predominan formas familiares de producción evidenciadas en la mayor participación de la categoría “trabajador familiar sin remuneración” –Misiones, Chaco y Formosa.

En el segundo eje es muy buena la calidad de la representación de las provincias patagónicas: Santa Cruz (.97 de coseno cuadrado), Chubut (.90), Tierra del Fuego (.87), Neuquén (.77). Estas cuatro provincias en conjunto contribuyen en un 38% a ese segundo eje. En lo que hace a las posiciones ocupacionales el eje discrimina a los distritos con mayor peso del empleo público – región patagónica y noroeste- de aquellos donde hay predominio de los “cuenta propia”.

Algunas provincias están mal representadas en el plano (1,2). Así, La Pampa, Tucumán y Mendoza en realidad sólo aparecerían bien representadas en el eje 3, y Río Negro en el 4 (coseno cuadrado = .55). En estos casos el gráfico no está mostrando de modo adecuado cual es la distancia que mantienen respecto al perfil medio, y en consecuencia no se puede afirmar, por ejemplo, que Río Negro y San Juan presenten perfiles similares.

2. ALGUNAS APLICACIONES DEL ACM

Empero, la popularidad creciente del análisis de correspondencias no obedece tanto al AFC propiamente dicho, sino a la utilidad que presta el ACM en el análisis de encuestas.¹⁶ Para explicar la naturaleza de esta técnica comenzaremos retomando el problema de construir un índice a partir de un conjunto de indicadores.

2.1 Construcción de un índice por medio del ACM

Los datos se refieren a la población de dos establecimientos de enseñanza técnica agropecuaria ubicados en una localidad de la zona centro de la Provincia de Misiones.¹⁷ En esta población nos proponemos distinguir estratos que reflejen las diferencias existentes en el nivel de vida de sus miembros. Para ello, disponemos de una serie de indicadores referidos a los artefactos poseídos y a los servicios a los que tienen acceso las unidades domésticas a las que pertenecen los alumnos.¹⁸ Mediante estos indicadores es posible construir un índice de equipamiento doméstico y acceso a servicios (IES).

Tabla 6.3: Valores, puntajes y modalidades de las variables seleccionadas para el IES

	<i>Variables</i>	<i>Valores</i>	<i>Puntos Índice</i>	<i>Modalidades ACM</i>
	<i>Conexión de electricidad</i>	Tiene	1	LUZ1
		No tiene	0	LUZ0
	<i>Teléfono</i>	Tiene	1	TEL1
		No tiene	0	TEL0
	<i>TV Color</i>	Tiene	1	TVC1
		No tiene	0	TVC0
	<i>Videocasetera</i>	Tiene	1	VID1
		No tiene	0	VID0
	<i>Canal de cable</i>	Tiene	1	CAB1
		No tiene	0	CAB0
	<i>Heladera</i>	Tiene	1	HEL1
		No tiene	0	HEL0
	<i>Cocina a Gas</i>	Tiene	1	GAS1
		No tiene	0	GAS0
	<i>Lavarropas</i>	Tiene	1	LAV1
		No tiene	0	LAV0

¹⁶ El ACM puede ser presentado como una simple generalización del AFC (cf. Crivisqui, 1993: 207).

¹⁷ El relevamiento consistió en un censo de alumnos que tuvo lugar en 1996 y estuvo a cargo de estudiantes de la Licenciatura en Trabajo Social con la coordinación de María Rosa Fogeler.

¹⁸ Muchos de estos indicadores son tomados en cuenta habitualmente para la construcción de índices de nivel económico-social, aunque para la elaboración de esta medida es usual combinarlos con otros indicadores referidos a las dimensiones ocupacional y cultural.

La combinación de estas ocho variables dicotómicas genera un espacio de propiedades que se podría diagramar en una tabla de $2^8 = 256$ celdas. Empero, conforme a la lógica del procedimiento explicitado en el capítulo 5, el IES como índice sumatorio puede obtenerse por el simple expediente de asignar un puntaje a cada uno de los valores de las variables incluidas en la Tabla 6.3, y de sumar estos puntajes para cada unidad de análisis.¹⁹ Así, una unidad de análisis, en vez de ser descripta por los valores que presenta en cada uno de los ocho indicadores puede ser caracterizada mediante un único valor numérico comprendido entre 0 y 8, que está resumiendo su posición en el espacio de propiedades. A su vez ese puntaje puede ser recodificado en cuatro categorías; cortando ‘por el lado de la variable’, damos lugar a una variable de nivel ordinal:

IES muy bajo	0-2 puntos
IES bajo	3-4 puntos
IES medio	5-6 puntos
IES medio alto	7-8 puntos

De este modo el puntaje IES constituye una medida que sintetiza las ocho dimensiones del espacio de propiedades original en una sola. Ahora bien, en cualquier técnica de análisis factorial el propósito es básicamente análogo: se trata siempre de resumir un conjunto de variables en un nuevo conjunto de variables más reducido.²⁰ El objetivo es reducir el espacio de propiedades generando variables-resumen, los denominados *factores*, que permitan poner en evidencia las diferencias entre las unidades de análisis de acuerdo a las combinaciones de características que presentan.²¹

En este sentido, el ACM permite visualizar similitudes y diferencias en los individuos -las unidades de análisis- a través de representaciones gráficas construidas a partir de un análisis de la estructura que surge de la interrelación entre las características observadas. Esta estructura llevará a identificar grupos de individuos que presentan características semejantes.

Para proceder a construir nuestro índice por medio del ACM partimos de las mismas variables con las *modalidades*²² cuyas etiquetas identificadoras están incluidas en la última columna de la Tabla 6.3. El análisis de correspondencias permitirá generar gráficos en los que se localizará cada una de las unidades de análisis (en nuestro ejemplo, los 240 estudiantes censados) en el espacio de las modalidades, de manera que individuos con características semejantes aparecerán próximos los unos de los otros en ese espacio. Simultáneamente cada una de las modalidades se localizará en el espacio de los individuos, de modo tal que las modalidades asociadas presentarán coordenadas similares. Ambos espacios, el de los individuos y el de las modalidades, constituyen lecturas complementarias del mismo conjunto de datos. De este modo, es posible asignar a cada individuo sus coordenadas factoriales, las que constituirán variables numéricas que podrán ser utilizadas para la construcción de un índice.

Evidentemente el conjunto de las modalidades que intervienen de modo activo en la generación del espacio factorial ha de ser homogéneo, para que el análisis pueda tener un significado interpretable. En nuestro ejemplo, las dieciseis modalidades de las ocho variables del IES son tomadas como activas, y su homogeneidad deviene de que todas pueden considerarse como

¹⁹ En la práctica bastará, en el programa en el que estemos operando, con generar una nueva columna de la matriz -el puntaje IES- mediante una simple fórmula que consista en sumar las columnas de los puntajes correspondientes a cada variable-indicador.

²⁰ Para una sugerente introducción, no técnica, al análisis factorial clásico, ver Gould (1997, cap. 6).

²¹ Así, de querer llevar a cabo el procedimiento numérico de un modo más ortodoxo, tratándose de variables intervalares, se puede aplicar un análisis de componentes principales (ACP) sobre las variables-indicadores previamente estandarizadas (es decir, sometidas a una transformación lineal de modo que todas presenten una distribución semejante con una media aritmética de cero y una varianza igual a uno). En ese caso, ante una estructura claramente unidimensional, la coordenada de cada unidad de análisis en el primer factor se tomará como su puntaje en el índice. Cf., por ejemplo, Weller y Romney (1990: 26-31).

²² Se utiliza el término *modalidad*, como sinónimo de ‘categoría’, para referirse a los valores de una variable cualitativa.

indicadores del nivel de vida.²³

Si las variables están relacionadas unas con otras, la información que aporte cada una será redundante, y por lo tanto será posible reducir el espacio de propiedades. Así, por ejemplo, en una situación en la que todos los que tuvieran teléfono estuvieran conectados a un canal de cable, y viceversa, con una sola de estas dos variables se tendría la misma información que con las dos, llegándose a la misma clasificación de los individuos; y en ese caso la variabilidad entre los individuos estaría explicada por una sola variable.²⁴ Sin embargo, en la práctica la situación es bastante más compleja, ya que la variabilidad depende de más de dos variables, y además éstas no se encuentran perfectamente correlacionadas.

La llamada ‘tabla de Burt’ permite analizar las relaciones que mantienen todas las variables entre sí,²⁵ y consiste en una presentación conjunta de todas las tablas de contingencia generables a partir de un conjunto de variables tomadas de a dos. En cada línea como en cada columna de esta tabla figuran las dieciseis modalidades que componen las ocho variables consideradas, y en la diagonal se ve la relación de cada variable consigo misma (vale decir su distribución de frecuencias). La tabla de Burt resume la información que se quiere analizar y sobre ella operan los algoritmos de cálculo del ACM.²⁶

Tabla 6.4: Tabla de Burt para los ocho indicadores del IES

LUZ1	206	0	35	171	155	51	58	148	43	163	190	16	156	50	175	31
LUZ0	0	34	0	34	3	31	0	34	0	34	13	21	12	22	0	34
TEL1	35	0	35	0	31	4	17	18	26	9	35	0	35	0	35	0
TEL0	171	34	0	205	127	78	41	164	17	188	168	37	133	72	140	65
TVC1	155	3	31	127	158	0	57	101	43	115	152	6	128	30	139	19
TVC0	51	31	4	78	0	82	1	81	0	82	51	31	40	42	36	46
VID1	58	0	17	41	57	1	58	0	21	37	56	2	50	8	55	3
VID0	148	34	18	164	101	81	0	182	22	160	147	35	118	64	120	62
CAB1	43	0	26	17	43	0	21	22	43	0	43	0	43	0	43	0
CAB0	163	34	9	188	115	82	37	160	0	197	160	37	125	72	132	65
HEL1	190	13	35	168	152	51	56	147	43	160	203	0	156	47	168	35
HEL0	16	21	0	37	6	31	2	35	0	37	0	37	12	25	7	30
GAS1	156	12	35	133	128	40	50	118	43	125	156	12	168	0	145	23
GAS0	50	22	0	72	30	42	8	64	0	72	47	25	0	72	30	42
LAV1	175	0	35	140	139	36	55	120	43	132	168	7	145	30	175	0
LAV0	31	34	0	65	19	46	3	62	0	65	35	30	23	42	0	65
	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
	LUZ		TEL		TVC		VID		CAB		HEL		GAS		LAV	

El estudio de la asociación entre las variables conduce a analizar la asociación entre sus modalidades, donde cada modalidad representa la clase de los individuos que la poseen. Cada línea de la tabla de Burt permite estudiar la relación entre una modalidad y el conjunto de las modalidades restantes. En un plano factorial dos modalidades estarán más próximas entre sí cuanto más se comporten en forma similar en relación a las restantes modalidades. Así, ‘CAB1’ y ‘TEL1’

²³ Nótese que la homogeneidad demandada es exactamente la misma que requiere el procedimiento numérico, en el cual los indicadores también deben ser seleccionados en base a su significado común. Del mismo modo que la edad o el sexo no pueden ser tomados como indicadores para la construcción de un IES, tampoco tendría sentido que sus modalidades intervinieran activamente en la generación de este espacio factorial.

²⁴ Según lo expresaba Lazarsfeld: «En el caso de la reducción funcional, algunas combinaciones son eliminadas en vista de las relaciones existentes entre las variables» (1993: 161)

²⁵ Obsérvese que, si se tratara de variables numéricas, como es el caso en un ACP, este papel de la tabla de Burt sería suplido simplemente por la matriz de correlaciones (cf. *supra*: capítulo 4).

²⁶ La tabla de Burt resulta generada automáticamente a partir de la matriz de datos original -también denominada ‘tabla de códigos condensados’- por cualquier *software* adecuado. La utilización práctica de estas técnicas, dado el volumen de los cálculos demandados, sólo ha sido posible con el advenimiento de la computación. En Francia, dos programas muy conocidos son el SPAD y el ADDAD. También estas técnicas han sido incorporadas -con éxito dispar- a los principales paquetes estadísticos: el BMDP en 1988, el SAS y el SPSS en 1990 (cf. Greenacre y Blasius, 1994).

representan clases de individuos que se comportan de manera similar respecto al resto de las modalidades. El análisis de correspondencias va a producir una representación geométrica de estas asociaciones.

El ACM puede realizarse igualmente sobre una tabla ‘lógica’ o ‘booleana’, también denominada ‘disyuntiva completa’.²⁷

Tabla 6.5: Versión abreviada de la tabla disyuntiva completa para 240 estudiantes y 16 modalidades del IES

UA	Luz		Tel		TVC		Vid		Cab		Hel		Gas		Lav		
	sí	no	sí	no	sí	no	sí	no	sí	no	sí	no	sí	no	sí	no	
1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	8
2	1	0	0	1	1	0	0	1	0	1	1	0	1	0	0	1	8
3	1	0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	8
4	1	0	0	1	1	0	0	1	0	1	1	0	0	1	0	1	8
5	1	0	0	1	1	0	0	1	0	1	1	0	0	1	0	1	8
.	8
i	8
.	8
239	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	8
240	1	0	1	0	1	0	0	1	1	0	1	0	1	0	1	0	8
Total	206	34	35	205	158	82	58	182	43	197	203	37	168	72	175	65	

En una tabla booleana cada renglón corresponde a una unidad de análisis. A diferencia de la matriz de datos, cada columna se corresponde con una modalidad (y no ya con una variable). Para cada modalidad su presencia-ausencia en un individuo es simbolizada respectivamente por un ‘1’ o por un ‘0’. Así la suma de todos los números de cada renglón es siempre la misma y resulta igual al número de las variables (ocho, en este ejemplo). En cambio la suma de cada columna nos informa acerca del número de unidades de análisis que presentan una modalidad en particular.

La tabla disyuntiva puede ser vista simultáneamente como un conjunto de columnas y como un conjunto de hileras. Así las unidades de análisis serán más o menos parecidas unas a otras de acuerdo al perfil que presenten: por ejemplo, el individuo 1 y el 239 presentan perfiles opuestos, son totalmente disímiles, mientras que el 4 y el 5 tienen perfiles idénticos.

Una representación geométrica adecuada de la variabilidad observada entre los individuos debe posicionarlos de tal modo que aparezcan más próximos en el espacio cuanto más semejantes sean en las modalidades que presentan (así, el 1 y el 239 deberán aparecer muy alejados el uno del otro, mientras que el 4 y el 5 deberán ocupar el mismo punto).

Ahora bien, es posible realizar el mismo análisis con las columnas. Tratándose de variables dicotómicas, como es el caso, los perfiles de sus modalidades son exactamente opuestos (hay una correlación negativa perfecta entre ambas modalidades). Igualmente en este caso es posible observar columnas que son semejantes, significando con ello que las modalidades en cuestión tienden a estar presentes -o ausentes- en los mismos individuos: es lo que ocurre en nuestro ejemplo con ‘Tel sí’ y ‘Cab sí’. De este modo, también las modalidades pueden ser consideradas en su perfil como puntos localizables en el espacio de los individuos.²⁸ Estos puntos aparecerán más próximos unos a otros cuanto más semejantes sean las modalidades entre sí, esto es, cuanto mayor

²⁷ «El origen de la terminología ‘tabla disyuntiva completa’ es el siguiente: el conjunto de los valores x_{jk} de un mismo individuo para las modalidades de una misma variable, comporta el valor 1 una vez (completa) y sólo una vez (disyuntiva)» (Escofier y Pagès; 1990: 48)

²⁸ El ACM hace jugar a las columnas y los renglones de la tabla el mismo papel, en el sentido de que no otorga preeminencia en el análisis a uno de estos elementos sobre el otro. Desde este punto de vista, en los términos en que se planteaba la estructura del dato en el capítulo 1, es indiferente afirmar «el auto es rojo» o «el rojo es auto» (cf. Fénelon; 1981). Es necesario pasar de una visión en la que el énfasis está puesto exclusivamente en las variables y en sus relaciones, para las cuáles las unidades de análisis funcionan apenas como soportes, a un punto de vista en el que las variables pueden jugar igualmente el rol de soportes con relación a los individuos.









sea el número de individuos en que coincidan en su presencia-ausencia.

En el caso del ACM, al igual que en el AFC, la técnica va a operar generando un espacio de representación que permita visualizar adecuadamente las distancias entre modalidades y entre individuos en los espacios originales. Ello supone generar nuevas variables -los *factores*- que sinteticen el conjunto de las variables originales; tales nuevas variables deberán mantener la más alta correlación posible con las variables originales.

El nuevo sistema de ejes -factores- se obtiene a partir de una operación matemática llamada 'diagonalización de una matriz', que se realiza sobre la tabla de Burt o la tabla lógica. Así, es posible localizar los individuos y las modalidades en el nuevo sistema de coordenadas, en el que el primer eje da cuenta de la mayor proporción de la inercia susceptible de ser resumida, el segundo eje explica la mayor parte posible de la variabilidad residual, etc.

Al igual que en caso del AFC, el ACM no se limita a la presentación de un plano factorial para dar cuenta de los resultados de una investigación, sino que incluye también todas las denominadas 'ayudas a la interpretación', esto es, las tablas en que se presentan medidas numéricas indispensables para una correcta interpretación de los datos.

Figura 6.4: Histograma de los valores propios

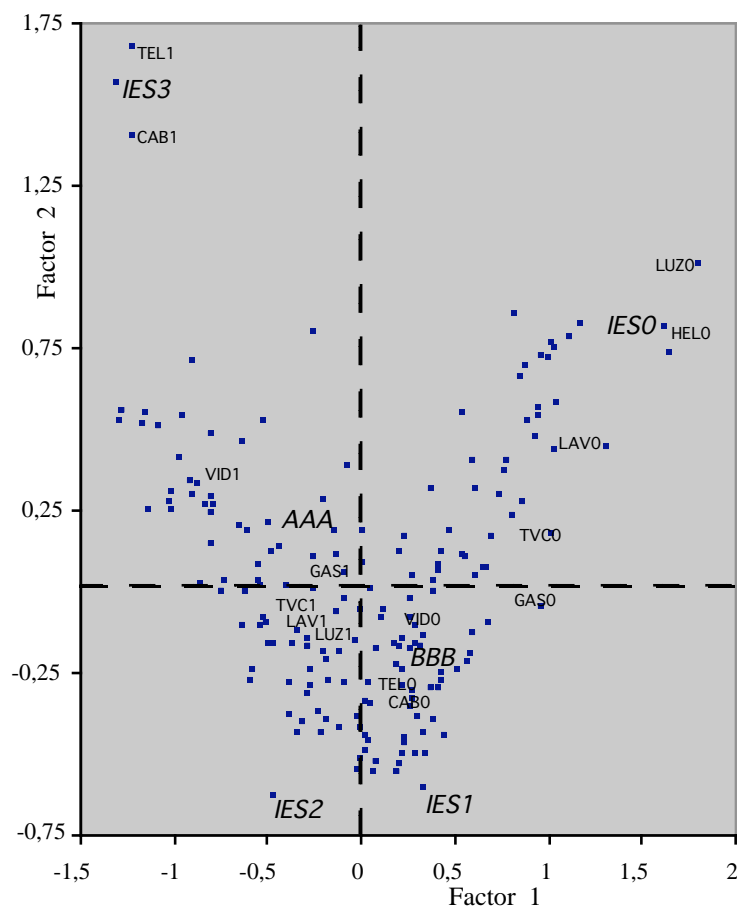
Factor Nº	Valor propio	%	% acum.	
1	.4237	42.37	42.37	
2	.1656	16.56	58.94	
3	.1092	10.92	69.86	
4	.0860	8.60	78.46	
5	.0690	6.90	85.36	
6	.0636	6.36	91.72	
7	.0452	4.52	96.24	
8	.0376	3.76	100.00	

Del análisis del histograma surge que el primer factor con un valor propio de .4237 concentra un 42,37 % de la inercia -o la variancia- original, en tanto que el segundo factor sólo representa un 16,56%. Esta baja pronunciada de la inercia que exhibe el segundo factor en comparación con el primero es compatible con la idea de una estructura básicamente unidimensional. Por otra parte, desde el tercer factor en adelante la disminución de la inercia de los sucesivos factores tiene lugar muy paulatinamente. En suma, en la combinación de los factores 1 y 2 que define el primer plano factorial se concentra un 58,94% de la inercia original, por lo que pragmáticamente podemos decidir limitar nuestro análisis a estos dos factores.²⁹

En el ACM la inercia total es una cantidad fija independiente de cual sea la estructura de los datos.³⁰ Esta cantidad es igual a $(k/p)-1$, en la que k es el número total de modalidades y p el número de variables. En este caso particular, con 16 modalidades que corresponden a 8 variables, la inercia total es igual a 1, lo que ocurrirá siempre, cualquiera sea el número de variables, si éstas son todas dicotómicas. Asimismo, tratándose de variables dicotómicas el número máximo de factores, también es igual a p (en general el número de ejes es igual a $k - p$).

²⁹ La decisión de cuántos factores hay que tomar en cuenta en un ACM no es algo que se pueda resumir en un conjunto de reglas preestablecidas.

³⁰ «La inercia total depende únicamente del número de variables y de modalidades y no de las relaciones entre las variables» (Lebart, Morineau, Piron; 1995: 120). Por ende, en ACM, el tamaño absoluto del valor propio por sí mismo no indica nada, puesto que puede corresponder a una inercia de cualquier magnitud.

Figura 6.5: Ubicación de los individuos y de las modalidades activas y suplementarias en el plano factorial (1,2)

Tanto los individuos como las modalidades aparecen representados como puntos en el plano factorial determinado por la combinación de los dos primeros factores (en realidad, los puntos están localizados en el espacio de ocho factores, y por ende lo que se observa en el plano 1,2 es simplemente la mejor proyección posible de aquella localización en un plano).

En este plano factorial, el primer factor está representado en el eje horizontal y el segundo en la ordenada. Los puntos representan la proyección de los 240 alumnos en este espacio, mientras que las modalidades, también simbolizadas por puntos, están identificadas cada una por su etiqueta. Si dos individuos aparecen con coordenadas similares en ambos ejes es porque presentan globalmente las mismas características. Si dos modalidades aparecen próximas en el plano, es por encontrarse asociadas en general con las mismas modalidades de las restantes variables (como lo indica la tabla de Burt) y porque se encuentran presentes globalmente en el mismo conjunto de individuos (lo que se hace visible en la tabla lógica).

Al considerar la proyección en el eje 1 horizontal de las dieciseis modalidades de la Tabla 6.1 que corresponden a las ocho variables consideradas para el índice, se observa que éstas se ordenan de izquierda a derecha desde 'TEL1' hasta 'LUZ0' y 'HELO'. En los extremos se ubican aquellas modalidades (clases de individuos) más diferentes del total de la población (perfil medio); así, este eje opone los que poseen los bienes y servicios menos elementales (cable, teléfono y video), al grupo de los que no acceden a los equipamientos y servicios básicos (luz y heladera),³¹ con lo que la coordenada en el primer factor resulta efectivamente interpretable en términos del nivel de equipamiento doméstico de las familias de los estudiantes. Por su parte, el segundo eje da cuenta

³¹ El que los individuos más ricos aparezcan en el sector izquierdo del plano, es un efecto producido arbitrariamente por el algoritmo actuante en la reducción del espacio original: el signo -negativo o positivo- de los valores en un eje carece por completo de importancia, y en este caso habrá que interpretar que los valores positivos más altos corresponden a niveles mayores de pobreza.

de una diferencia residual entre los individuos, separando las situaciones extremas en uno u otro sentido de las intermedias.

De este modo se distinguen distintas zonas en el plano: en el cuadrante superior izquierdo están los individuos en situación más favorable; a medida que nos desplazamos hacia abajo y hacia la derecha aparecen individuos en situaciones menos favorables; finalmente, ya del lado derecho se ubican individuos cada vez más pobres cuanto más elevada es su coordenada en el segundo eje.³²

En todas estas técnicas de análisis multivariado, tanto los individuos como las modalidades pueden jugar dos papeles: activo e ilustrativo. Los individuos y las modalidades **activos** son los que intervienen en la generación de los factores. En cambio, las modalidades e individuos **ilustrativos** (a veces también denominados ‘suplementarios’) no contribuyen a la construcción del espacio factorial y simplemente resultan proyectados en éste.³³ La proyección de estas características suplementarias o ilustrativas tiene por objetivo estudiar su relación con los factores.

Al proyectar las modalidades de la variable ‘Escuela’ se observa que mientras que la Escuela A –‘AAA’– está del lado izquierdo del diagrama, la B se localiza en la zona correspondiente a los individuos con menor IES. Esto parece indicar una diferencia en la extracción social del alumnado de estas instituciones: en promedio los alumnos de la Escuela A tienden a presentar un más alto nivel de equipamientos y servicios que los de la B.

En base a las coordenadas de los individuos en el primer factor hemos generado mediante un procedimiento de clasificación ascendente jerárquica una variable ordinal, distinguiendo cuatro niveles de IES. Al proyectar también en el plano las modalidades de esta variable observamos que se despliegan siguiendo la forma de una parábola o de una herradura. Mientras el primer eje ordena las modalidades desde la más baja (‘IES0’) a la más alta (‘IES3’), el segundo factor opone las modalidades extremas ubicadas en la zona superior, a las no-extremas (‘medio’ y ‘bajo’) del IES en la parte inferior,³⁴ como era esperable dado que estas modalidades se construyeron en base el ACM anteriormente realizado.

Finalmente, podemos preguntarnos qué relación existe entre este IES generado mediante el ACM y el que habíamos construido antes por el procedimiento sumatorio más habitual.

Tabla 6.6: Relación entre el IESacm y el IES sumatorio

IESacm	IES sumatorio				Total
	Muy bajo	Bajo	Medio	Medio-alto	
IES3	0	0	0	39	39
IES2	0	0	93	0	93
IES1	0	61	0	0	61
IES0	47	0	0	0	47
Total	47	61	93	39	240

Como se puede observar, hay una correlación positiva perfecta ($V = 1$) entre los resultados de las dos técnicas.³⁵ ¡Por algo el procedimiento sumatorio convencional -tan sencillo como robusto- ha sido utilizado con éxito durante tanto tiempo!

Es importante aclarar cómo ambas formas de reducción –la numérica y la factorial (o *funcional*)– resultan conceptualmente diferenciables. Según Lazarsfeld, «en la reducción funcional se da una *relación real* entre dos de los atributos que reduce el número de combinaciones» (1993: 160; *itálicas nuestras*), y se trabaja a partir de esas relaciones comprobadas. En cambio la reducción numérica, cuando su propósito es clasificar a las unidades de análisis en una escala

³² Así, es posible entender que lo que está representado en el plano factorial es el *espacio social* en el sentido de Bourdieu: «las distancias espaciales sobre el papel equivalen a las distancias sociales» (Bourdieu; 1997: 30).

³³ Se calcula las coordenadas de los individuos ilustrativos como media ponderada de las coordenadas de las modalidades que presentan, y análogamente las coordenadas de las modalidades suplementarias como media ponderada de las coordenadas de los individuos que las presentan.

³⁴ Se trata de una configuración típica a propósito de la cual se suele hablar de “efecto Guttman”. Según explican Escofier y Pagès, «Cuando un efecto de escala es muy fuerte, influencia varios ejes de acuerdo a la siguiente propiedad: el factor de rango s es una función polinómica de grado s del primero» (1990: 240).

³⁵ La correlación entre los puntajes del IES sumatorio, tomados como una variable intervalar, y las coordenadas en el Factor 1 arroja $r = 0,996$.

unidimensional (como es el caso, por ejemplo, en un índice de nivel económico-social), se basa en una suerte de apuesta implícita acerca de la existencia de esas relaciones.

2.2 Construcción de una tipología

Para finalizar presentamos un ejemplo más en el que exponemos los resultados de la aplicación de un ACM en combinación con el uso de un procedimiento de clasificación para la construcción de una tipología.

Entre los objetivos de la investigación realizada se planteaba determinar si los dos establecimientos de enseñanza secundaria diferían en cuanto al origen social de su alumnado. Teniendo en cuenta la información relevada, nos propusimos generar una tipología de los estudiantes dividiéndolos en clases que resultaran ser lo más homogéneas posibles internamente a la vez que lo más heterogéneas unas con respecto a las otras. Lo que esperábamos era que dichos tipos se diferenciaran unos de otros en sus conductas, expectativas y opiniones.³⁶ Las variables que consideramos significativas a este efecto fueron las siguientes:

Tabla 6.7: Variables y modalidades para generar tipos sociales

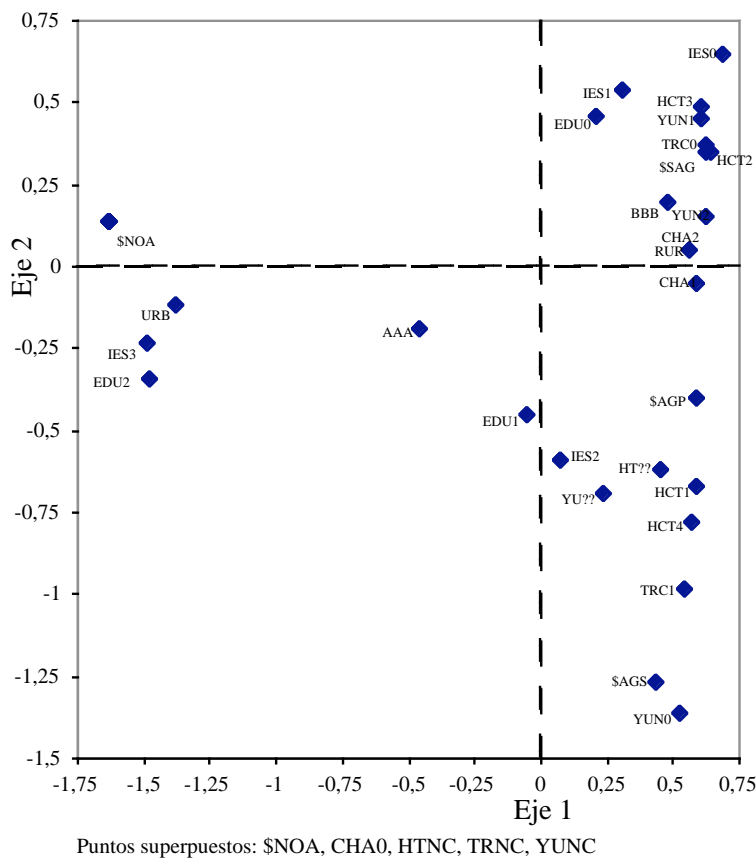
Variables	Valores	Etiquetas ACM
Residencia habitual	Urbana	URB
	Rural	RUR
Tenencia de chacra	Sí	CHSI
	No	CHNO
Tenencia de tractor	Sí	TRSI
	No	TRNO
	No corresponde	TRNC
Nº de yuntas de bueyes	No tiene	YUN0
	Una yunta	YUN1
	Dos o más yuntas	YUN2
	No corresponde	YUNC
	Sin dato	YU??
Nº de hectáreas	Sin chacra	HTNC
	< de 15 has.	HCT1
	15-<30 has.	HCT2
	30-<50	HCT3
	50 o más	HCT4
	Sin dato	HC??
Fuente de ingreso	Sólo agrícola	\$SAg
	Agrícola principal	\$AgP
	Agrícola Secundario	\$AgS
	No agrícola	\$NoA
Nivel de educación	No completó Primaria	EDU0
	Primaria Completa	EDU1
	Secundaria completa (y +)	EDU2
I. Equipamiento doméstico	Muy bajo	IES0
	Bajo	IES1
	Medio	IES2
	Medio-alto	IES3

Nuevamente se trata de realizar la reducción de un espacio de propiedades mediante un ACM, con la salvedad que ya no se presume que se trata de una estructura unidimensional como era el caso con el IES. Empero, hay todavía una homogeneidad de significado: todas estas variables se refieren a características que pueden ser consideradas como pertinentes al proponernos generar

³⁶ En términos de Bourdieu (1980), se puede pensar que cada una de estas clases así construidas tenderá a agrupar a individuos con similares *habitus*.

tipos que puedan distinguirse unos de otros por su ubicación en el espacio social.³⁷

Figura 6.6: Origen social de los estudiantes



El plano factorial correspondiente a este ACM muestra como las modalidades se localizan con cierta coherencia y definen regiones diferenciadas. El primer eje (con un valor propio de 0,7794 correspondiente a un 31% de la inercia) da cuenta por construcción de la mayor variabilidad en los datos, poniendo en evidencia una estructura que opone a dos grupos cuya existencia se hubiera podido anticipar en base a las asociaciones entre las variables observables en una tabla de Burt: así, hacia la izquierda del plano se localizan las modalidades asociadas a la no tenencia de chacra, como la ausencia de tractor o de yunta, el ingreso no agrícola,³⁸ y la residencia urbana; mientras que hacia a la derecha se ubican las características ligadas a la actividad agrícola.

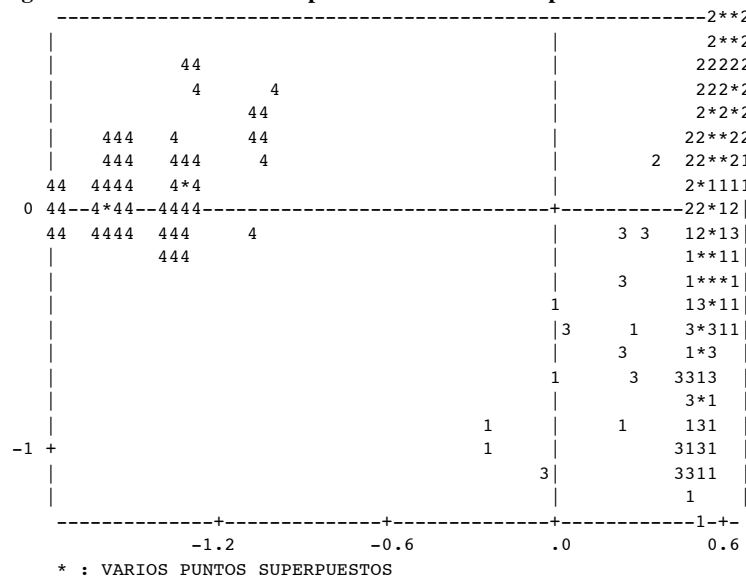
El segundo eje (con un valor propio de 0,22) marca una diferenciación dentro de los individuos rurales. En el extremo superior aparecen las modalidades vinculadas a la explotación agrícola familiar en pequeña escala: no se tiene tractor, se cuenta tan sólo con una yunta, el ingreso es sólo agrícola, la superficie es de 15 a 50 hectáreas; además tanto el IES como el nivel de educación resultan ser bajos. En cambio, se ubican en la parte inferior del plano tanto los colonos medios (tractor, más de 50 has., etc.) como las familias para las cuales los ingresos agrícolas son secundarios; ambos grupos presentan un nivel intermedio en relación a la educación y al IES.

³⁷ Según Lebart: «Presentar los resultados de un estudio tipológico (ya se trate de análisis factorial o de clasificación) sólo tiene sentido si la lista de las preguntas llamadas activas está claramente especificada (...) El conjunto de las preguntas activas debe satisfacer un criterio de homogeneidad cuyos fundamentos son bastante intuitivos. Es necesario que los cálculos de las distancias tengan un sentido, de modo que los reagrupamientos puestos ulteriormente en evidencia también tengan un sentido: sería por ejemplo muy inapropiado mezclar variables objetivas y variables de opinión, porque los reagrupamientos obtenidos serían ambivalentes y por lo tanto difíciles de ser interpretados» (Lebart; 1986: 4-5).

³⁸ Como la asociación entre estas modalidades es perfecta, sus coordenadas son idénticas, con lo que los puntos aparecen exactamente superpuestos.

Una clasificación automática en base a los factores producidos por el ACM³⁹ arroja los resultados que se observan en la figura 6.7, en la que cada individuo en vez de estar representado por un punto, lo está por un número que identifica la clase a la que pertenece.

Figura 6.7: Localización en el plano de los individuos pertenecientes a cada tipo



En el sector izquierdo, la clase 4 abarca a un 27% de las familias; es la de los “*no agricultores*”⁴⁰: los ingresos son exclusivamente no agrícolas, y el 89 % vive en la ciudad. Una mitad tiene el máximo nivel de IES (“medio-alto”), y además esta categoría concentra un 77% de los jefes de familia con instrucción superior a la primaria.

En las tres clases restantes el 100% de las familias tienen chacra. Así, la clase 1 es la de los “*colonos medios*”, que son un 25% del total. Un 85% de ellos cuenta con tractor, un 51% tiene al menos dos yuntas; y un 59% opera explotaciones de más de 50 hectáreas. El ingreso es sólo agrícola para un 69% de ellos. Por otra parte, 58% alcanza un nivel de IES medio.

La Clase 3 (15% de la población) agrupa a los “*agricultores part-time*”. Aunque entre éstos un 56% tienen menos de 15 hectáreas, todos los que tienen esta superficie caen en esta clase. En un 69% se trata de familias para las cuales la actividad agrícola provee sólo de un ingreso secundario. El 92 % carece de tractor y el 47% no tiene ni una yunta. El 66% de los jefes cuenta con primaria completa.

Por último, en el sector superior derecho se ubican los miembros de la clase 2, la que totaliza un 34% de los casos. Se trata de los “*colonos pobres*”: en un 99% son rurales, en un 93% el ingreso es solamente agrícola, y el 98% no tiene tractor; en esta clase un 77% de los jefes no completó la escuela primaria.

³⁹ Este modo de proceder que combina los resultados de un análisis factorial con técnicas de clasificación automática es típico de la escuela francesa. El método de clasificación ascendente jerárquica compara a los individuos a través de sus coordenadas factoriales y los agrupa de tal manera que las clases sean lo más homogéneas *dentro* de ellas y lo más heterogéneas *entre* ellas. En realidad lo que se obtiene es una jerarquía de particiones, a partir de cuyo análisis se decide el número de clases que interesan. Estos métodos se utilizan frecuentemente como complemento del análisis factorial ya que permiten distinguir grupos de individuos similares más allá de los primeros factores que se consideran en el análisis factorial. Así, los agrupamientos resultan de comparaciones sobre todas las dimensiones en que se descompuso la inercia, o bien sobre una aproximación de ese espacio total, dejando de lado los últimos ejes que en general dan cuenta de variaciones aleatorias.

⁴⁰ Según recomienda Lebart, es conveniente que las etiquetas elegidas sean lo más neutras posibles (1986:7); de hecho, las clases inevitablemente incluirán algunos individuos para los cuales las etiquetas no resultarán del todo adaptadas.

Tabla 6.8: Extracción social de los estudiantes en dos establecimientos de enseñanza agropecuaria (%)

	Colonos medios	Colonos pobres	Agricultores part-time	No agricultores	Total (100%)
Escuela A	25	16	13	46	(123)
Escuela B	24	52	17	7	(117)

$$\chi^2 = 57,00; p < 0,0001; V = 0,44.$$

Fuente: Encuesta a establecimientos de enseñanza agropecuaria, 1996.

Basándonos en la tipología que hemos construido podemos sintetizar sin dificultad las diferencias entre las dos escuelas en lo que hace a la extracción social de sus estudiantes. De este modo el establecimiento A exhibe casi una mitad (46%) de sus estudiantes que son hijos de no agricultores, mientras que la escuela B registra 52% de alumnos procedentes de familias de colonos pobres, en tanto que en ambos institutos hay la misma proporción -una cuarta parte de los estudiantes- que proviene de hogares de colonos medios.

Este ha sido solamente un último ejemplo en este capítulo donde apenas nos hemos propuesto introducir algunas aplicaciones del ACM, a costa de una simplificación que muchos podrán juzgar excesiva. En la presentación que hemos hecho del tema nos hemos preocupado por enfatizar la continuidad entre este nuevo enfoque del análisis de encuestas y el más tradicional: se trata de un conjunto de herramientas alternativo para lograr resultados que no deberían ser sustancialmente distintos de los alcanzados por otros vías. En definitiva, el uso provechoso que se pueda hacer de estas técnicas va a depender tanto de la característica del problema específico que se aborde, como del estilo de trabajo con el cual el investigador se encuentre más a gusto. Aspectos ambos que se encuentran obviamente ligados entre sí, además de relacionarse, claro está, con las preferencias teóricas y las posturas epistemológicas. Por nuestra parte, si supiéramos que este capítulo ha servido al menos para despertar el interés del lector, estimulándolo a perfeccionarse en esta temática,⁴¹ nos consideraríamos plenamente satisfechos.

⁴¹ La bibliografía es frondosa, y contantemente se van agregando nuevos aportes. Algunos textos en inglés, todavía no tan numerosos, son los de Blasius y Greenacre (1998), Greenacre y Blasius (1994), Weller y Romney (1990), y Phillips (1998). Pero las referencias más importantes todavía siguen siendo producidas en la lengua francesa. Así, existe una revista especializada -*Cahiers de l'analyse des données*- donde se han publicado regularmente artículos a cargo de Benzécri y de sus discípulos. El poco convencional texto de Fénelon (1981) contiene sin duda muchos hallazgos de gran valor pedagógico para el neófito. En cuanto a la obra de Escofier y Pagès (1990), puede considerarse como un manual clásico de introducción a las técnicas factoriales francesas. El libro de Lebart, Morineau y Piron (1995) es sumamente recomendable, ya que logra una presentación elegante y concisa de todas las técnicas, y a la vez brinda muy útiles indicaciones acerca de su complementariedad con la versión de la estadística predominante en los países anglosajones. En castellano, aparte de la traducción del texto de Escofier y Pagès, se puede encontrar una presentación rigurosa y accesible del análisis de correspondencias simples y múltiples en el libro de Crivisqui (1993).

Referencias bibliográficas

- ACOSTA L.A., JORRAT J.R., 1992, *Prestigio ocupacional en la Argentina*, Buenos Aires, Facultad de Ciencias Sociales.
- ALTHUSSER L., 1968, *Lire le Capital*. París, Maspéro. 2e. Ed.
- ALTHUSSER L., 1985, *Curso de filosofía para científicos*, Barcelona, Planeta-Agostini.
- ANDERSON T.R. y ZELDITCH, M., 1968, *A Basic Course in Statistics with Sociological Applications*, Nueva York, Holt, Rinehart and Winston Inc.
- ANTOINE J., 1993, *El sondeo, una herramienta de marketing*, Buenos Aires, Deusto.
- ARGENTINA Rep., 1984, *La pobreza en la Argentina*, Buenos Aires, Indec.
- AYER A. J., 1965, *El positivismo lógico*, México, FCE.
- BACHELARD G., 1972, *La formation de l'esprit scientifique*, París, J. Vrin.
- BACHELARD G., 1986, *Le rationalisme appliqué*. París, PUF.
- BARANGER D., 1994, "Sobre la idea de la matriz de datos", Comentario bibliográfico de Epistemología y Metodología de J. Samaja, *Desarrollo Económico*, 33, 132, 605-11.
- BARTOLOME L.J., 1982, "Base social e ideología en las movilizaciones agraristas en Misiones entre 1971 y 1975", *Desarrollo Económico*, Buenos Aires, 22, 85, 25-56.
- BARTON A.H., 1969, "El concepto de 'espacio de propiedades'", en F. Korn (Ed.).
- BEARDEN W. O., NETEMEYER R. G., MOBLEY, M. F., 1993, *Handbook of Marketing Scales. Multi-Item Measures for Marketing and Consumer Behaviour Research*, Newbury Park, Londres y Nueva Delhi, Sage.
- BENSON O., 1974, *El laboratorio de ciencia política*, Buenos Aires, Amorrortu.
- BENZECRI J.P. et al., 1973, *L'analyse des données, Tome 1: La taxinomie, Tome 2: L'analyse des correspondances*, París, Dunod.
- BERNARD H.R., 1995, *Research Methods in Anthropology. Qualitative and Quantitative Approachs*, Walnut Creek (CA, Altamira Press, 2nd. Edition.
- BLALOCK H.M. Jr., 1966 , *Estadística Social*, México, FCE.
- BLALOCK H.M. Jr., 1982 , *Conceptualization and Measurement in the Social Sciences*, Beverly Hills, Sage.
- BLASIUS J., GREENACRE M., 1998, *Visualization of Categorical Data*, San Diego y Boston, Academic Press.
- BLUM A., 1991, *Matématiques et statistique appliquées aux sciences sociales*, París, Dunod.
- BOUDON R., 1971, *Les mathématiques en sociologie*, París, PUF.

- BOUDON R., LAZARSFELD P., 1965, *Le vocabulaire des Sciences Sociales*, La Haya, Mouton.
- BOURDIEU P., 1969, "Condición de clase y posición de clase", en J. Szabón (comp.), *Estructuralismo y sociología*, Buenos Aires, Nueva Visión, 71-100.
- BOURDIEU P., 1979, *La distinction. Critique sociale du jugement*, París, Minuit.
- BOURDIEU P., 1997, *Capital cultural, escuela y capital social*, México, Siglo XXI.
- BOURDIEU P., 1998, *La domination masculine*, París, Seuil.
- BOURDIEU P. (con la colaboración de L.D. Wacquant), 1992, *Réponses. Pour une anthropologie réflexive*, París, Seuil.
- BOURDIEU P., PASSERON J.-P., CHAMBOREDON J.-C., 1968, *Le métier de sociologue*. París, Mouton/Bordas.
- BRAITHWAITE R.B., 1963, *La explicación científica*, Madrid, Tecnos.
- BROWN H.I., 1983, *La nueva filosofía de la ciencia*, Madrid, Tecnos.
- BRY X., *Analyses factorielles simples*, París, Economica.
- BUNGE M., 1968, *La ciencia, su método y su filosofía*, Buenos Aires, Siglo Veinte.
- BUNGE M., 1979, *La investigación científica*, Barcelona, Ariel, 6a.Ed..
- CARMINES E. G., ZELLER R.A., 1979, *Reliability and Validity Assessment*, Beverly Hills, Sage.
- CHALMERS A. F., 1987, *¿ Qué es esa cosa llamada ciencia?*, Madrid, Siglo XXI.
- CHAMPAGNE P. et al., 1990, *Initiation à la pratique sociologique*, París, Dunod.
- CIBOIS Ph., 1990, *L'analyse des données en sociologie*, París, PUF.
- CICOUREL A.V., 1982, *El método y la medida en Sociología*, Madrid, Editora Nacional.
- COHEN M., NAGEL E., 1968, *Introducción a la lógica y al método científico*, Buenos Aires, Amorrortu (2 vols.).
- CONVERSE J., 1987, *Survey Research iun the United States. Roots & Emergence, 1890-1960*, Berkeley, University of Calkifornia Press.
- CONVERSE J.M., PRESSER S., 1986, *Survey Questions. Handcrafting the Standardized Questionnaire*, Beverly Hills, Sage.
- COPI I.M., 1967, *Introducción a la lógica*, Buenos Aires, Eudeba, 5a. Ed.
- CORTADA de KOHAN N., 1968, *Manual para la construcción de tests objetivos de rendimiento*, Buenos Aires, Paidós.
- CORTES F., s.f., "Algunos coeficientes de asociación en tablas de 2 por 2 y sus relaciones", México, Flasco.
- CRIVISQUI E. M. 1993, *Análisis factorial de las correspondencias. Un instrumento de investigación en ciencias sociales*, Asunción, Université Libre de Bruxelles/Universidad Católica de Asunción.

- DENZIN N.K., LINCOLN Y.S., 1994, "Introduction: Entering the Field of Qualitative Research", en DENZIN N.K., LINCOLN Y.S., *Handbook of Qualitative Research*, Thousand Oaks, Londres y Nueva Delhi, Sage, 1-17.
- DIAZ E., HELER M., 1987, *El conocimiento científico. Hacia una visión crítica de la ciencia*, Buenos Aires, Eudeba.
- DROESBEKE J.-J., LAVALLEE P., 1996, "La no respuesta en las encuestas", *Methodologica*, 4, 1-39.
- DUCROT O., TODOROV T., 1978, *Diccionario enciclopédico de las ciencias del lenguaje*, México, Siglo XXI.
- DUNCAN O.D., 1984, *Notes on Social Measurement. Historical & Critical*, New York, Russell Sage Foundation.
- ERICKSON B.H., NOSANCHUK T.A., 1979, *Understanding Data*, Milton Keynes, The Open University Press.
- ESCOFIER B., PAGES J., 1990, *Analyses factorielles simples et multiples. Objectifs, méthodes et interprétation*, París, Dunod (hay trad. castellana, Univ. del País Vasco, Bilbao, 1992) .
- ESCOFIER B., PAGES J., 1997, *Initiation aux traitements statistiques. Méthodes, méthodologie*, Rennes, Presses universitaires de Rennes.
- FENELON J.-P., 1981, *Qu'est-ce que l'analyse des données?*, París, Lefonen.
- FESTINGER L., KATZ D.(Eds.), 1975, *Los métodos de investigación en las ciencias sociales*, Buenos Aires, Paidós.
- FITZ-GIBBON C.T., MORRIS L.L., 1987, *How to Analyse Data*, Beverly Hills y Londres, Sage.
- FORNI, F.H.; GALLART, M.A., VASILACHIS I., 1992, *Métodos cualitativos II. La práctica de la investigación*, Buenos Aires, Cedral.
- FOWLER F.J., 1993, *Survey Research Methods*, Newbury Park (CA), Sage.
- FREIXA I BLANXART M. et al, 1992, *Análisis exploratorio de datos, nuevas técnicas estadísticas*, Barcelona, PPU.
- GAETA R., ROBLES N., 1985, *Nociones de epistemología*, Buenos Aires, Eudeba.
- GALTUNG J., 1966, *Teoría y métodos de la investigación social*, Buenos Aires, Eudeba (2 vols.).
- GARCIA FERRANDO M., 1995, *Socioestadística. Introducción a la estadística en sociología*, Madrid, Alianza.
- GERMANI G., 1973, *El concepto de marginalidad- Significado, raíces históricas y cuestiones teóricas, con particular referencia a la marginalidad urbana*, Buenos Aires, Nueva Visión.
- GLASER B.G., STRAUSS A.L., 1967, *The Discovery of Grounded Theory. Strategies for Qualitative Research*, Nueva York, Aldine.
- GOODE W., HATT P., 1969, *Métodos de investigación social*, México, Trillas.
- GOULD S.J., 1997, *La falsa medida del hombre*, Barcelona, Drakontos.

- GREENACRE M., BLASIUS J., 1994, *Correspondence Analysis in the Social Sciences*, Londres y San Diego, Academic Press.
- GUIBOURG R. A., GHIGLIANI A.M., GUARINONI R.V., 1985, *Introducción al conocimiento científico*, Buenos Aires, Eudeba.
- GUTIERREZ C., BRENES A. (Eds.), 1971, *Teoría del método en las ciencias sociales*, San José de Costa Rica, Educa.
- HARTWIG F., DEARING B. E., 1979, *Exploratory Data Analysis*, Beverly Hills y Londres, Sage.
- HENKEL R.E., 1976, *Tests of Significance*, Beverly Hills and London, Sage.
- HENRY G. T., 1990, *Practical Sampling*, Newbury Park, London and New Delhi, Sage.
- HILDEBRAND D.K., LAING J.D., ROSENTHAL H., 1977, *Analysis of Ordinal Data*, Beverly Hills y Londres, Sage.
- HYMAN H., 1971, *Diseño y análisis de las encuestas sociales*, Buenos Aires, Amorrortu.
- IVERSEN G.R., NORPOTH H., 1976, *Analysis of Variance*, Beverly Hills, Sage.
- JORRAT J.R., "En la huella de los padres: movilidad ocupacional en el Buenos Aires de 1980", *Desarrollo Económico*, 37, 145, 91-115
- KALTON G., 1987, *Introduction to Survey Sampling*, Beverly Hills y Londres, Sage.
- KEREN G., LEWIS Ch., 1993, *A Handbook for Data Analysis in the Behavioral Sciences. Methodological Issues*, Hillsdale (N.J.), Lawrence Erlbaum.
- KIM J.-O., MUELLER Ch.W., 1978, *Introduction to Factor Analysis. What It Is and How To Do It*, Beverly Hills, Sage.
- KIM J.-O., MUELLER Ch.W., 1978, *Factor Analysis. Statistical Methods and Practical Issues*, Beverly Hills, Sage.
- KORN F. (Ed.), 1969, *Conceptos y variables en la investigación social*, Buenos Aires , Nueva Visión.
- KORN F., 1995, "Encuestas de opinión: ¿Diagnósticos, pronósticos o pasatiempos?", *Desarrollo Económico* , 34, 136: 609-13.
- KUHN T., 1971, *La estructura de las revoluciones científicas*, México, FCE.
- LAZARSFELD P., 1968, "El análisis de relaciones estadísticas", en M. Mora y Araujo (Ed.), 1968, 27-37.
- LAZARSFELD P., 1970 , "Quelques fonctions de l'analyse qualitative en sociologie", en P.Lazarsfeld, *Philosophie des sciences sociales*, París, Gallimard, 318-360 [original en inglés en *Frankfurter Beiträge zur Soziologie*, 1955, I, 321-351].
- LAZARSFELD P., 1993, "Some Remarks on Typological Procedures in Social Research", en P.Lazarsfeld, *On Social Research and its Language*, Chicago y Londres, The University of Chicago Press, 158-171 [versión resumida por Ann Pasanella del original alemán, *Zeitschrift für Sozialforschung*.6, 1937, 99-106].
- LAZARSFELD P., MENZEL H., 1969, "Sobre la relación entre propiedades individuales y

- colectivas", en F. Korn (Ed.), *Conceptos y variables en la Investigación social*, Buenos Aires, Nueva Visión, 77-103.
- LEACH C., 1979, *Introduction to Statistics - A Nonparametric Approach for the Social Sciences*, Nueva York y Toronto, John Wiley & Sons.
- LEBART L., 1986, "Qui pense quoi? Evolution et structure des opinions en France de 1978 à 1984", *Consommation*, 4,3-22.
- LEBART L., 1989, "Stratégies du traitement des données d'enquête", *La Revue du Modulad*, 3, 21-29.
- LEBART L., 1990, "Les questions ouvertes, outils de contrôle, d'évaluation, de valorisation", *Mots*, 23,76-89.
- LEBART L., MORINEAU A., PIRON M., 1995, *Statistique exploratoire multidimensionnelle*, París, Dunod.
- LEBART L.(Ed.), 1992, *La qualité de l'information dans les enquêtes*, París, Dunod.
- LEWIS-BECK M.S., 1995, *Data Analysis An Introduction*, Thousand Oaks (CA), Sage.
- LIEBETRAU A.M., 1983, *Measures of association*, Beverly Hills y Londres, Sage.
- LUNGARZO C., 1986, *Lógica y lenguajes formales*, Buenos Aires, Cedral, (2 Vols.).
- MARX K., 1971 *Elementos fundamentales para la crítica de la economía política (borrador) 1857-1858*, Buenos Aires, Siglo XXI (2 Vols).
- MCIVER J.P., CARMINES E. G., 1981, *Unidimensional Scaling*, Beverly Hills, Sage.
- MILLER D.C., 1991, *Handbook of Research Design and Social Measurement*, Newbury Park (Ca), Sage.
- MILLS C.Wright, 1964, *La imaginación sociológica*, México, FCE.
- MOLES A.A., 1995, *Les sciences de l'imprécis*, París, Seuil.
- MORA Y ARAUJO M. (Ed.), 1968, *El análisis de datos en la investigación social*, Buenos Aires, Nueva Visión.
- MORA Y ARAUJO M. (Ed.), *Medición y construcción de índices*, Buenos Aires, Nueva Visión.
- MORA Y ARAUJO M., 1965, "Recomendaciones para la lectura de cuadros". Buenos Aires, Servicio de Documentación de Sociología.
- NAGEL E., 1968, *La estructura de la ciencia*, Buenos Aires, Paidós.
- NAMBOODIRI K., 1984, *Matrix Algebra-An Introduction*, Beverly Hills y Londres, Sage.,
- PADUA J., (Ed.), 1979, *Técnicas de investigación aplicadas a las ciencias sociales*, México, FCE.
- PASSERON J.-P., 1991, *Le raisonnement sociologique. L'espace non-poppérien du raisonnement naturel*, París, Nathan.
- POLLAK M., 1993, "Paul F. Lazarsfeld fondateur d'une multinationale scientifique". en M. Pollak, *Une identité blessée*, París, Métailié, 319-360 (original de 1979 en *Actes*, nº25).

- POPPER K., 1967, *La lógica de la investigación científica*, Madrid, Tecnos.
- REYNOLDS H.T., 1977, *Analysis of Nominal Data*, Beverly Hills y Londres, Sage.
- RICHARD-ZAPPELLA J., 1990a, "Variations interrogatives dans la question de sondage", *Mots*, 23, 24-38.
- RICHARD-ZAPPELLA J., 1990b, "Mobilisation de l'opinion publique par les sondages", *Mots*, 23, 60-75.
- ROBERT S., 1978, *La logique, son histoire, ses fondements*, Québec, Editions Le Préambule.
- ROGERS E.M., KINCAID D.L., 1981, *Communication Networks- Toward a New Paradigm for Research*, Nueva York y Londres, The Free Press.
- ROISTACHER R.C., 1979, "Acquisition and management of social network data", en P.W. Holland, S. Leinhardt, *Perspectives on Social Network Research*, Nueva York, Academic Press.
- SAMAJA J., 1994, *Epistemología y metodología. Elementos para una teoría de la investigación científica*, Buenos Aires, Eudeba, 2a. edición aumentada.
- SCHRODT Ph.A., 1985, *Microcomputer Methods for Social Scientists*, Beverly Hills y Londres, Sage.
- SCHUMAN H., 1966, "The Random Probe, a Technique for Evaluating the Validity of Closed Questions", *American Sociological Review*, 21, 218-222.
- SCHUMAN H., PRESSER S., 1979, "The Assesment of 'No Opinion' in Attitude Surveys", en K.F. Schuessler, *Sociological Methodology*, San Francisco, Jossey-Bass, 241-275.
- SCHUMAN H., PRESSER S., 1996, *Questions & Answers in Attitude Surveys. Experiments on Question Form, Wording, and Context*, Thousand Oaks, Sage.
- SCHUSTER F.G., 1982, *Explicación y predicción*, Buenos Aires, CLACSO.
- SCHUSTER F.G., 1992, *El método en las ciencias sociales*, Buenos Aires, Cedral.
- SELLTIZ, Claire et al., 1968, *Métodos de investigación en las relaciones sociales*, Madrid, Rialp.
- SHROEDER L. D., SJOQUIST D.L., STEPHAN P., 1986, *Understanding Regression Analysis. An Introductory Guide*, Beverly Hills, Sage.
- SILVERMAN D., 1995, *Interpreting Qualitative Data. Methods for Analysing Talk, Text and Interaction*, London, Thousand Oaks, New Delhi: Sage.
- SIRKIN R.M., 1995, *Statistics for the Social Sciences*, Thousand Oaks, Sage.
- SMITH, H. W., 1975, *Strategies of Social Research*, Englewood Cliffs, N.J., Prentice-Hall.
- SOROKIN, Pitrim A., 1964, *Achaques y manías de la sociología moderna y ciencias afines*, Madrid, Aguilar.
- SPECTOR P.E., 1992, *Summated Rating Scale Construction. An introduction*, Newbury Park, Londres y Nueva Delhi, Sage.
- STEWART D.W., SHAMDASANI P.N., 1990, *Focus Groups. Theory and Practice*, Newbury Park (CA) y Londres, Sage.

- TENENHAUS M., *Méthodes statistiques en gestion*, París, Dunod.
- TORGERSON W., 1965, *Theory and Methods of Scaling*. Nueva York, Wiley and Sons..
- TORRADO S., 1983, *La familia como unidad de análisis en censos y encuestas de hogares*, Buenos Aires, CEUR.
- TRAUGOTT M.W., LAVRAKAS P.J., 1997, *Encuestas, guía para electores*, México, Siglo XXI.
- VASILACHIS I., 1992, *Métodos cualitativos I. Los problemas teórico-epistemológicos*, Buenos Aires, Cedral.
- VERON E., 1972, *Conducta, estructura y comunicación*, Buenos Aires, Tiempo Contemporáneo.
- VOGT W.P., 1993, *Dictionnary of Statistics and Methodology, A Nontechnical Guide for the Social Sciences*, Newbury Park, London and Nerw Delhi, Sage.
- WAINERMAN C, SAUTU R., 1997, *La trastienda de la investigación*, Buenos Aires, Editorial de Belgrano.
- WAINERMAN C. (Ed.), 1976, *Escalas de medición en ciencias sociales*, Buenos Aires, Nueva Visión.
- WILLER D., 1969, *La sociología científica*, Buenos Aires, Amorrortu.
- WRIGHT D.B., 1997, *Understanding Statistics. An Introduction for the Social Sciences*, Londres, Sage.
- ZEISEL H., 1962, *Dígalo con números*, México, FCE.
- ZELDITCH M. Jr., 1959, *A Basic Course in Sociological Statistics*, Nueva York, Holt.
- ZETTERBERG H., 1968, *Teoría y verificación en Sociología*, Buenos Aires, Nueva Visión.
