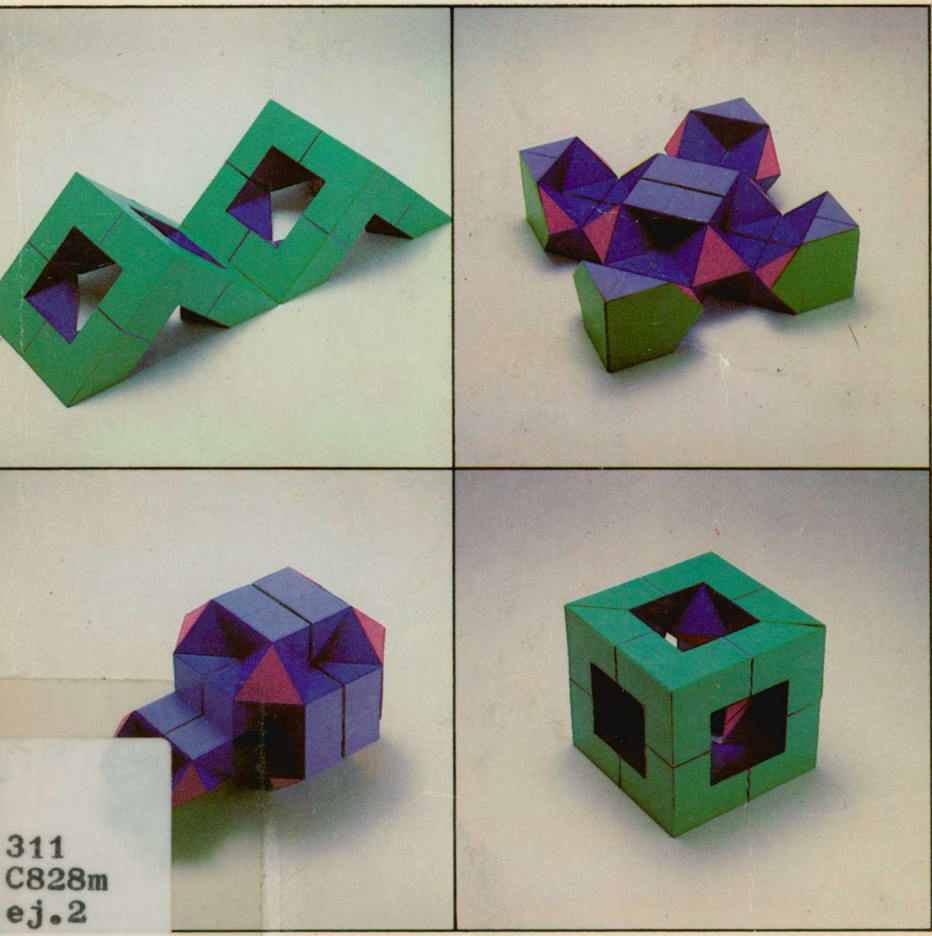
Fernando Cortés - Rosa María Rubalcava

métodos estadísticos aplicados a la investigación en ciencias sociales

análisis de asociación



311
C828m
ej.2

El Colegio de México

MÉTODOS ESTADÍSTICOS APLICADOS A LA INVESTIGACIÓN EN CIENCIAS SOCIALES

ANÁLISIS DE ASOCIACIÓN

CENTRO DE ESTUDIOS SOCIOLÓGICOS

MÉTODOS ESTADÍSTICOS APLICADOS A LA INVESTIGACIÓN EN CIENCIAS SOCIALES

ANÁLISIS DE ASOCIACIÓN

Fernando Cortés
Rosa María Rubalcava

EL COLEGIO DE MÉXICO

ÍNDICE

	Pág.
Presentación	7
Prefacio	9
1. Introducción	13
2. Asociación de variables no métricas en tablas de dos por dos	27
2.1. Distribuciones bidimensionales de frecuencias: variables dicotómicas	27
2.2. Noción de independencia estadística	35
2.3. Otras formas de medir asociación	44
2.3.1. El determinante como medida de asociación	51
2.3.2. Coeficiente Fi (Φ)	54
2.3.3. Educación y estatus socioeconómico: una aplicación de Φ	56
2.3.4. Coeficiente Q de Yule	59
2.3.5. Desarrollo social y movimiento obrero: una interpretación del caso argentino (1915-1945)	63
Ejercicios	66
3. Asociación entre tres variables dicotómicas	73
3.1. El análisis de covarianzas de Lazarsfeld	74
3.2. La ecuación de covarianzas	78
3.3. La tipología de Lazarsfeld	84
3.4. Breves comentarios sobre la aplicación de la tipología	88
3.5. El concepto de asociación y el análisis de Lazarsfeld	91
3.6. Reforma agraria y migraciones (el caso de Chile)	92
3.7. Algunas limitaciones del método para estudiar la relación entre más de dos variables dicotómicas	98
Ejercicios	100
4. Asociación en tablas de r x c	105
4.1. Asociación e independencia estadística	106
4.1.1. Asociación y análisis porcentual	106
4.1.2. Asociación y diferencias entre frecuencias observadas y esperadas. Índices derivados de ji-cuadrado	109

4.1.2.1. El coeficiente de contingencia cuadrado medio (Φ^2) 115
4.1.2.2. Coeficiente de contingencia de Tschuprow 116
4.1.2.3. Coeficiente de asociación de Cramer 118
4.1.2.4. Tres breves comentarios sobre χ^2, Φ^2, T^2 y C^2 119
4.1.2.5. Coeficiente de contingencia de Pearson 120
4.2 Asociación como el grado de correspondencia entre proposiciones empíricas y distribuciones de frecuencias 121
4.2.1. Un ejemplo. Relación entre la situación de clase y el ingreso 121
4.2.2. El coeficiente delta-ro ($\nabla\varrho$) para el caso de predicciones bidimensionales de observaciones únicas 123
4.2.3. Un caso particular de $\nabla\varrho$: el coeficiente $\varkappa$ de Cohen 133
4.2.4. Coeficiente delta-ro para el caso de predicciones bidimensionales sobre pares de observaciones 135
4.2.4.1. Tabla ordinal condensada 135
4.2.4.2. Coeficientes de asociación de Sommers y Kim 143
4.2.4.3. El coeficiente de asociación gamma (γ) 146
4.2.4.4. Consideraciones finales sobre $\nabla\varrho$ 148
Ejercicios 148
Apéndice. (Dos ejercicios resueltos) 157

Presentación

En un libro, publicado hace más de 40 años por el Fondo de Cultura Económica, don José Medina Echavarría[1] apuntaba que de todas las ciencias sociales, la Sociología había sido la más castigada por la improvisación.

La afirmación, en muchos sentidos, sigue teniendo vigencia. Y, podría agregarse, no sólo involucra a la Sociología sino a casi todas las ciencias sociales.

Una explicación posible de lo anterior es la improvisación —que de manera escueta podría definirse como la falta de rigor en el análisis, o como una escasa correspondencia entre la afirmación teórica y la evidencia empírica que haga a aquélla plausible o no— ha tenido mucho que ver con la debilidad metodológica y técnica inherente a la práctica de la investigación.

En el caso de América Latina es posible hacer la observación general de que la investigación ha tendido a privilegiar la dimensión teórica, descuidándose por tanto el componente metodológico y técnico. En otras palabras, se ha descuidado cómo hacer la investigación, cómo aproximarse al problema que se investiga, y cómo acercarse al mundo fenomenológico.

En la actualidad, y a pesar de que en la región latinoamericana se han dado pasos importantes en materia de investigación empírica, es posible comprobar que aún existe una inclinación poco sistemática y significante hacia el método y hacia la técnica.

Es por ello que el libro de Fernando Cortés y de Rosa María Rubalcava, investigadores de la Facultad Latinoamericana de Ciencias Sociales (FLACSO) y de El Colegio de México, respectivamente, tiene particular importancia, en especial en el campo de la docencia, donde se observa una marcada carencia de textos adecuados a la realidad de la investigación social en América Latina, lo que no impide que también se pueda

[1] *Sociología: ciencia y técnica*, México, Fondo de Cultura Económica, 1941. Reeditado en 1982.

utilizar como obra de consulta para aquellos investigadores que tengan que someter sus datos específicamente al análisis de asociación.

Agregaría que se trata de un trabajo serio, sistemático y didáctico para todos los interesados en conocer con cierta profundidad algo del método científico aplicado a las ciencias sociales desde una perspectiva particular: el análisis estadístico de atributos.

Los autores nos ofrecen un libro claro y sencillo, sin dejar con ello de lado el rigor en la exposición. Se trata de una búsqueda —si así quiere llamársele— en la que se pretende establecer una articulación entre el componente teórico y el proceso de investigación; que desea establecer la relación entre lo que explica, de manera tentativa, un problema u objeto de estudio (las hipótesis) y su comprobación empírica correspondiente (el mundo de los datos).

Como es obvio, una hipótesis no se comprueba exclusivamente por medio del instrumental estadístico. Depende del tipo de hipótesis formulada y de los datos implicados por ella. Sin embargo, una forma (un procedimiento) de verificación es factible mediante ese instrumental.

En esta perspectiva, los autores del libro nos ofrecen un conjunto de técnicas estadísticas en donde se privilegia la relación entre atributos o, expuesto de manera más general, entre hechos o fenómenos. Es sabido que para llegar a una explicación posible es necesario relacionar. Si no se establece una relación no es factible hablar de explicación.

En el libro que se presenta se encuentra una exposición coherente de cómo relacionar estadísticamente los fenómenos involucrados por las hipótesis teóricas. También nos enseña que dependiendo del tipo de distribución y del nivel de medición de las variables es posible utilizar distintas medidas de asociación que hacen de la prueba de una hipótesis un elemento sustantivo para el proceso de investigación y de explicación.

Esta obra tiene una ventaja adicional, desde la perspectiva de las ciencias sociales: ilustra las maneras de correlacionar variables con base en ejemplos que se desprenden de la investigación social. Ello hace que el lector no se enfrente a una serie de fórmulas abstractas. Por el contrario, el texto permite que esas fórmulas tengan un contenido concreto que se vincula a problemáticas con las que el investigador de las disciplinas sociales está familiarizado.

En síntesis: pienso que el libro de Fernando Cortés y Rosa María Rubalcava será de gran utilidad para los estudiosos de la técnica estadística y del método en las ciencias sociales y, en general, para aquellos interesados en hacer de la investigación un objeto riguroso, que aporte elementos de solución a la infinidad de problemáticas sociales que se someten a investigación.

José Luis Reyna
Julio de 1982

Prefacio

Este texto es una manifestación del interés largamente sostenido de los autores por situar las herramientas del análisis estadístico dentro del proceso de la investigación social. Se puede considerar que su eje central, es el estudio de los vínculos entre los desarrollos conceptuales y el análisis de asociación. Por lo tanto, el tratamiento de este tema va más allá de lo que normalmente se acostumbra en el ámbito de la estadística. Esta cobertura, que no es habitual para un texto como éste, probablemente no se advierta con sólo revisar el índice porque se ha plasmado de manera preferente en el enfoque y en la forma en que se han puesto los acentos en la exposición.

Nuestra preocupación por las relaciones entre teoría y estadística en el proceso de la investigación social se origina en un campo problemático que ha definido particularmente la Sociología y la Ciencia Política en América Latina. El ámbito intelectual en el que nos desenvolvemos se caracteriza, entre otras cosas, por una reflexión y problematización sistemática y constante sobre las articulaciones entre teoría, metodología y técnicas de investigación. Tal vez ello sea consecuencia, en parte, de la búsqueda de esquemas conceptuales que permitan arrojar luz sobre los peculiares problemas sociales propios de esta parte del mundo.

La sola referencia a esta reflexión no fue la única manera como se incluyeron en este trabajo algunas preocupaciones presentes en la labor de investigación de los científicos sociales latinoamericanos. En efecto, hemos incorporado algunos de los temas que han concitado el interés de nuestros colegas en los ejemplos que nos sirven para exponer el material estadístico y en los problemas que se incluyen al final de cada capítulo, a partir del segundo. No nos fue posible basar nuestra exposición del coeficiente delta-ro en ejemplos tomados de investigaciones efectuadas en la región porque no encontramos ningún estudio social que lo haya empleado.

La presentación, desde el punto de vista exclusivamente estadístico, incluye dos corrientes básicas que se engloban con el término de análisis de asociación. La primera define la asociación entre variables como

el desvío respecto a la independencia estadística, mientras que para la segunda, es el grado de correspondencia entre una definición de asociación específica y particular para cada caso, que anticipa el tipo de relación que se debe dar entre las variables y la distribución efectiva de los datos.

La primera de estas vertientes inicia sus desarrollos vinculando la fuerza con que se relacionan las variables a la diferencia entre porcentajes, continúa con el análisis de las discrepancias entre las frecuencias observadas y las esperadas y culmina con un conjunto de coeficientes que se definen en función de ji-cuadrada.

La segunda se inaugura, según nuestro parecer, con el trabajo de Goodman y Kruskal.[1] También nos parece un hito importante la contribución de Hildebrand, Laing y Rosenthal,[2] quienes dan a la noción de asociación un contenido específico y cambiante según lo requiera el tipo de proposición que interesa someter a contrastación empírica.

Aun cuando los coeficientes biserial, biserial punto y tetracórico también permiten trabajar con variables cualitativas, decidimos no incluirlos por cuanto se inscriben en el campo del análisis de correlación.

El contenido estadístico de este libro se incluye habitualmente en los programas regulares de métodos estadísticos de las escuelas de ciencias sociales de nuestras universidades y casi siempre la exposición de estos temas se sitúa inmediatamente después, o formando parte de la estadística descriptiva, antes de los cursos de inferencia estadística. En consecuencia decidimos no tratar los problemas de inferencia vinculados al análisis de asociación.

Los ejercicios son fragmentos de investigaciones empíricas, lo que lleva consigo el riesgo de desvirtuar las ideas del autor: como paliativo usamos el recurso de citar textualmente los párrafos a nuestro juicio pertinentes, sin preocuparnos por su extensión.

Las interpretaciones sobre la estructura lógica de las proposiciones a menudo no son unívocas, lo que significa que los problemas propuestos no tienen solución única. Lo importante de los ejercicios es *i*) la discusión de las proposiciones que pueden derivarse de las aseveraciones del autor y *ii*) las consecuencias de las distintas formalizaciones sobre la selección del instrumental estadístico adecuado para analizar su grado de concordancia con la información empírica.

A modo de ejemplo hemos incluido un apéndice en el que se presenta un modelo de resolución de los ejercicios. Para dos de los problemas planteados se discuten diferentes hipótesis y se procede a realizar los

[1]Leo A. Goodman y William H. Kruskal "Measures of Association for Cross Classifications", en *Journal of the American Statistical Association*, 49 (1954), pp. 732-764.

[2]David, Hildebrand, J. Laing y H. Rosenthal, *Analysis of Ordinal Data*, Sage Publications, 1977.

análisis estadísticos pertinentes. No está de más decir que este desarrollo, como se señaló, es sólo uno de los muchos posibles para el mismo problema.

Muchas personas hicieron posible este libro. Son varias las generaciones de estudiantes que tuvieron que someterse a la exposición del material y muchos los colegas que soportaron pacientemente nuestras inquisiciones, producto de lo que llamamos "lectura metodológica" de sus investigaciones. Esperamos haber recuperado debidamente las contribuciones de todos.

Debemos un reconocimiento especial a Mariquel Obregón por su excelente trabajo de mecanografía y a Enrique Pérez por habernos traducido el texto al español, a pesar de que según nosotros lo habíamos escrito en ese idioma.

Los pocos méritos que este libro pudiera tener van más allá de lo que nosotros hicimos. Las sugerencias que nos dio un evaluador, cuyo nombre desconocemos, lo enriquecieron y permitieron corregir a tiempo errores de forma y de cálculo, así como modificar y reordenar la presentación de algunos temas. También agradecemos a Humberto Muñoz, José Nun y Orlandina de Oliveira, quienes nos dieron la oportunidad de conocer sus ideas más allá de la exposición que de ellas hicieron en los trabajos que nos sirvieron para presentar el vínculo entre teoría e información y el papel que le corresponde al análisis de asociación. A Óscar Cuéllar y Ricardo Yocelevzky por su gentileza en leer y discutir las distintas versiones y por las valiosas sugerencias que nos permitieron llegar a esta forma final.

Un reconocimiento especialísimo a un amigo sin cuya intervención este libro se habría publicado hace dos años. Juan Carlos Marín (*Lito*) leyó una versión, a nuestro juicio definitiva, y sus acertadas críticas nos llevaron a rehacer totalmente el capítulo introductorio, para lo cual tuvimos que trabajar más de un año. Tratándose de la introducción puede parecer superfluo el esfuerzo realizado; sin embargo, las ideas que contiene son precisamente las que dan estructura y sustancia a los capítulos posteriores.

Quedan los mencionados exonerados de toda responsabilidad en lo que se refiere a deficiencias, errores e imprecisiones, excepto Lito, a quien reconocemos como cómplice de esta "maldad".

Fernando Cortés
Rosa María Rubalcava

México, D.F., mayo de 1984

1. Introducción

El propósito fundamental que guía este trabajo es presentar el análisis de asociación como una técnica estadística de apoyo para la investigación social. La búsqueda de este objetivo nos remite de lleno al tema de la relación entre teoría e información[1] en la investigación social. Sin embargo, éste es un tema tan vasto y complejo que supera con creces no sólo los límites de este libro, que más bien corresponde al ámbito del análisis estadístico, sino los de nuestro propio conocimiento del tema. A pesar de ello es imprescindible, para situar el instrumental estadístico que se presenta en los capítulos subsecuentes, referirnos a algunos aspectos involucrados en el vínculo entre teoría e información estadística. En consecuencia, la exposición será parcial y no podrá gozar de todo el rigor con que debiera tratarse. Nos conformaremos con no desvirtuar la imagen de esta relación ni inducir a sobresimplificaciones que puedan conducir a errores de apreciación sobre los diferentes problemas que aparecen en la investigación en ciencias sociales.

Expuesto de manera breve podríamos decir que el énfasis se pone en el estudio de la relación entre teoría y análisis de la información estadística. Interesa situarnos en aquellos puntos del proceso de la investigación que se constituyen como intersección de las operaciones teóricas y sus correspondientes operaciones empíricas ya que implican la movilización de nociones y técnicas definidas dentro del campo de la denominada estadística de atributos. Queremos destacar los vínculos que ligan en el proceso de investigación las elaboraciones propias del pensamiento teórico con los desarrollos característicos de las técnicas de análisis de la información, de modo que estemos en condiciones de orientar el estudio estadístico en función de las preocupaciones conceptuales y

[1]Entenderemos por información el resultado del proceso de transformación (traducción) de "señales" de la realidad a través de operaciones conceptuales (códigos). El aparato sensomotor del sujeto que observa la realidad desempeña el papel de filtro de las "señales" (véase Lorenz, Konrad. *La otra cara del espejo*, Plaza Janés, Madrid, 1980). En consecuencia la información es el producto de un doble recorte de la realidad, el biológico y el conceptual.

que al mismo tiempo las construcciones teóricas se basen en el material empírico acumulado en el proceso de investigación y no sólo en la consistencia interna de sus propios desarrollos.

La relación entre teoría y estadística suele darse en varios momentos a lo largo de una investigación; a veces predominan las consideraciones teóricas y otras las estadísticas. Examinemos con mayor detenimiento el tipo de relación que nos preocupa para sustentar nuestra aseveración. Si circunscribimos la discusión a las investigaciones que introducen lo empírico vía la estadística podremos encontrar una estructura que nos permita ligar las consideraciones teóricas con el manejo de la información estadística.

Ahora bien, para un texto de estadística dedicado al estudio de las relaciones entre variables no interesan todos los puntos de contacto entre teoría e información sino aquellos que permitan ligar ciertas operaciones conceptuales con el material estadístico que se expone. En el proceso de investigación la interconexión de los análisis teórico y estadístico se puede dar en distintos momentos; algunas veces las preguntas que orientan el análisis estadístico surgirán desde inquietudes conceptuales y otras veces los resultados del análisis estadístico serán los que provoquen las preguntas teóricas. Pero sea cual fuere el origen o el predominio de la relación será necesario establecer el modo lógico a través del cual se la pueda capturar.

El análisis de asociación es un instrumento estadístico concebido para analizar relaciones entre variables ligadas a las relaciones entre conceptos planteadas en el campo teórico (que no agotan todo el dominio de las preocupaciones teóricas pero son las que nos interesa destacar pues es la parte que se vincula con el análisis de asociación). En consecuencia nuestra inquietud se puede replantear en términos de encontrar las ligas entre las relaciones que involucran conceptos con las que relacionan variables. En estadística se da el nombre genérico de variable a la resultante del proceso que conecta los conceptos teóricos con los indicadores[2] o índices que se sitúan en el mundo de la información y que para los propósitos que nos animan debemos considerar como información susceptible de tratarse estadísticamente, la que no se limita sólo a las estadísticas sino que cubre también hechos históricos, acontecimientos, información periodística, etc. Por tanto, hay una conexión entre la variable estadística y el concepto que nos servirá para establecer el vínculo que nos interesa (aunque más adelante se tocará este tema,

[2] Para nosotros el indicador es una pieza fundamental en el proceso de investigación empírica que al conectarse simultáneamente con la realidad y con los conceptos posibilita la articulación teoría-realidad. Para mayores detalles véase Lazarsfeld, Paul. "De los conceptos a los índices empíricos", en Boudon y Lazarsfeld, *Metodología de las ciencias sociales*, vol. 1, LAIA, Barcelona, 1973.

no lo expondremos en profundidad porque además de ser complejo nos desviaría de nuestro propósito central).

En efecto, si a partir del discurso teórico habitualmente expresado en lenguaje natural podemos formalizar sus contenidos, es decir, si a partir de los desarrollos teóricos establecemos el conjunto de conexiones lógicas que les corresponden (que denominaremos proposiciones teóricas) y hacemos explícitos los pasos que ligan los conceptos con las variables estadísticas (es decir, con los índices e indicadores) entonces a cada relación entre dos o más conceptos corresponderá una relación análoga entre dos o más variables (que en este texto denominamos proposición empírica). El problema del vínculo entre la teoría y el análisis estadístico (limitado al análisis de asociación) se manifiesta de dos maneras: primero, hay una dificultad lógica para conectar las aseveraciones referidas a un cuerpo teórico, pero expresadas en lenguaje natural con lo que hemos denominado proposiciones teóricas, y segundo, hay que establecer las conexiones entre estas últimas y las proposiciones empíricas que expresan relaciones entre variables.

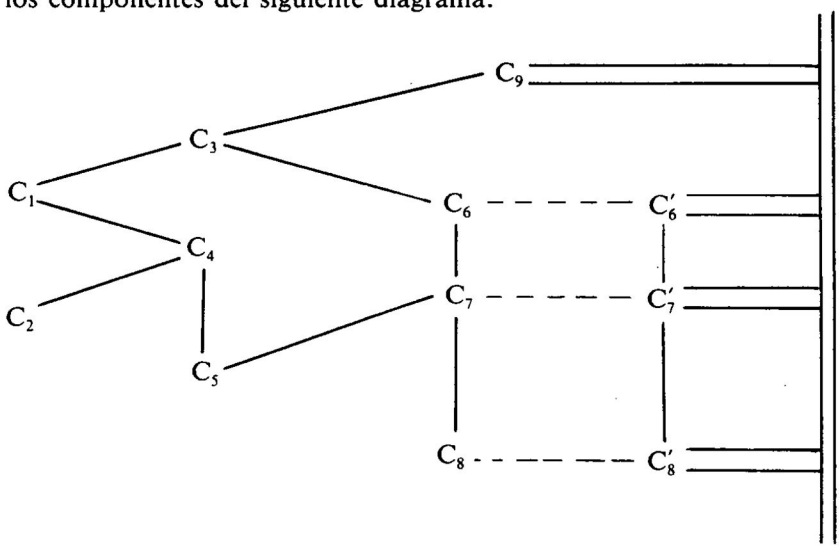
Una vez que se ha construido la proposición empírica hay que proceder a estudiar su articulación con el análisis estadístico de la información. Es exactamente este punto el que constituye el tema central de este libro.

Hemos sostenido que en el proceso de investigación es usual que haya múltiples puntos de contacto entre la teoría y el manejo estadístico de la información, lo que entra en abierta contradicción con la idea ampliamente extendida de que la investigación empieza con el desarrollo del marco teórico y culmina con el análisis de la información. En efecto, nuestra imagen del proceso es muy diferente; para nosotros la investigación empírica (es decir, el tipo de investigación que incorpora a los indicadores e índices provenientes de la realidad de manera sistemática y no a título de mero ejemplo) avanza sobre la base de la discordancia entre teoría e información. El desacuerdo entre las proposiciones teóricas y las empíricas correspondientes, así como las inconsistencias entre éstas y las observaciones, conduce a un replanteamiento de las relaciones que a veces implica la necesidad de mayores refinamientos teóricos otras un reexamen del proceso que ligó a los conceptos con sus variables y en otras oportunidades la construcción de nuevos datos estadísticos. También suele acontecer que se pongan en cuestión simultáneamente todos los procesos involucrados en la investigación. En nuestra concepción el vínculo que liga teoría y realidad es esencialmente dinámico.

El orden de la exposición de los resultados de una investigación no constituye su historia real. Es una costumbre generalizada situarlos en el punto terminal de una cadena que se inicia en el marco teórico. Se empieza por la teoría y se culmina en la presentación de la evidencia; sin embargo, el proceso de investigación está lleno de avances y retroce-

sos, de caminos sin salida, de exploración de nuevas vías y del cierre de otras. La investigación raras veces avanza linealmente; los obstáculos que hay que vencer no siempre se traducen en adelantos sino que la mayoría de las veces implican rodeos que, en el mejor de los casos, exigen un mayor grado de problematización teórico y metodológico. Es por ello que si deseáramos representarlo gráficamente deberíamos recurrir a una espiral.

Con el objeto de precisar algunas nociones centrales involucradas en la relación entre la teoría y el análisis estadístico de la información, acéptese que es posible detener la espiral de la investigación en un instante. En ese momento estarán presentes con mayor o menor énfasis los componentes del siguiente diagrama:



A la izquierda de la doble línea vertical hemos representado el campo de la teoría y a la derecha el de la realidad. En el primero denotamos con una letra C los conceptos que forman parte de la teoría. Las líneas continuas que los unen, representan las relaciones constitutivas de la teoría, que para los propósitos de este trabajo nos basta con concebir como un conjunto de conceptos teóricos relacionados. Generalmente, entre el cuerpo teórico y la realidad se encuentran los conceptos observables u objetivados simbolizados por C'.

Nos interesa remarcar algunas características de la gráfica para formarnos una idea del conjunto de pasos intermedios para conectar los conceptos teóricos con la información proveniente de la realidad:

1) En el dominio de la teoría hay una serie de conceptos que no pueden observarse o que no es posible poner en correspondencia con lo

empírico; tal es el caso de C_1, C_2, C_3, C_4 y C_5, en tanto que hay otros que sí son observables ya sea directamente (como C_9) o a través de conceptos observables relacionados con ellos (como C_6, C_7 y C_8).

2) Estos últimos sólo son susceptibles de ponerse en correspondencia con la información a través de su objetivación en conceptos observables (C'_6, C'_7, C'_8). El vínculo entre un concepto teórico y su correspondiente objetivado se representa con una línea punteada.

3) El concepto observable (que lo puede ser directamente o bien uno objetivado) se liga con la realidad a través de indicadores. A partir de este tipo de conceptos se recorta un segmento de la realidad y de la interacción surge la información. La doble línea representa, por una parte la forma como se observa la realidad de determinado concepto y por la otra los indicadores, que son el producto de la mirada conceptual sobre la realidad. La información estadística (susceptible de tratarse con el análisis de asociación) es el resultado del registro empírico de los indicadores.[3]

4) Hay que notar que la noción de indicador se refiere a un elemento que surge de la realidad y que está vinculado a un concepto observable. Sus propiedades son que existe en el mundo empírico y que es susceptible de registrarse. Adicionalmente hay que tomar en cuenta que todo indicador forma parte de una relación. La materia prima para la estadística son los indicadores, a los que se denomina variables. En primera instancia se podría pensar que ambas palabras son sinónimas, pero no sería del todo correcto. En efecto, la estadística, con su pretensión de generalidad, es decir, de servir como instrumento de cualquier ciencia sin apegarse a ninguna en particular, no considera el componente relacional del indicador. Esto quiere decir que para utilizar el conocimiento estadístico en el proceso de la investigación en general, y de la social en particular, es necesario llenar de contenido teórico a las variables. El buen uso de la estadística exige diferenciar los conceptos de las variables y establecer con claridad la naturaleza de su relación.

Hasta ahora hemos supuesto implícitamente que a cada concepto le corresponde sólo un indicador, pero en la investigación social es lo habitual que a un concepto se le asocie más de un indicador. Si éste fuera el caso, para vincular a cada concepto con una medida única habría que proceder a construir unos índices que resumieran los indicadores. Si bien dentro del arsenal de conocimiento técnico a disposición del científico social se hace una clara distinción entre indicador e índice ésta no tiene lugar en la estadística, para la cual sólo existe el término variable. Éste engloba tanto a los índices como a los indicadores.

[3]La palabra registro cubre las operaciones teórico-metodológicas e instrumentales involucradas en la observación (veáse Bunge, Mario. *La investigación científica*, Ariel, Barcelona, 1979, p. 729).

5) Una idea general que se desprende de la gráfica es que las variables sólo tienen sentido si están relacionadas con un cuerpo teórico. Los indicadores o los índices son referentes empíricos de los conceptos observables. Las variables que reflejan propiedades observables de las unidades para las cuales se registra la información sólo tienen pleno sentido cuando se llenan de contenido conceptual.

Hay que recordar que el propósito central de este libro es situar el análisis de asociación en los puntos de intersección entre las operaciones teóricas y empíricas involucradas en el proceso de la investigación social. Con el único objetivo de definir y perfilar nociones tales como variable, indicador, índice e información recurrimos al supuesto de que es posible detener el proceso de investigación y distinguir analíticamente sus componentes. Con este artificio se pretende introducir la idea de una precedencia temporal en el proceso de investigación que se originaría en el campo de la teoría. Sin embargo, esta concepción empieza a desvanecerse cuando reincorporamos la dinámica por medio de la idea de una espiral que en sus distintos niveles tendrá momentos más cercanos a lo empírico y otros más próximos a lo teórico. Además nuestra separación entre lo teórico, lo técnico y la información no es en absoluto nítida cuando se investiga. En cada acto de investigación se encuentran siempre presentes, aunque con pesos diferentes, las consideraciones teóricas, las metodológicas y las técnicas.

A lo largo de todo el texto nos preocuparemos por destacar cómo las proposiciones empíricas orientan el análisis estadístico de la información. Esto quiere decir que sólo subsidiariamente nos dedicaremos a mostrar el modo específico como se relacionan las proposiciones empíricas con las teóricas. También significa que analizaremos en abstracto el punto en que se establece el vínculo entre el mundo de la teoría y el de la información. Por abstracto queremos significar no referirlo a ningún momento preciso dentro del proceso de investigación. Enfatizamos el análisis del modo en que la proposición empírica orienta el estudio estadístico de las relaciones entre variables independientemente del lugar de la espiral en que se encuentre el proceso.

Hechas estas aclaraciones es necesario agregar una más. Para dar contenido sustantivo a los desarrollos estadísticos hemos tomado como ejemplos partes de una serie de investigaciones ya publicadas. El material está organizado de acuerdo con la lógica de la exposición de sus autores, siéndonos imposible recuperar la de la investigación. En consecuencia, los ejemplos parten del análisis de las proposiciones teóricas para establecer (sin mucha discusión) las proposiciones empíricas y de ahí en adelante se continúa examinando las consecuencias lógicas de éstas últimas sobre las distribuciones de frecuencias. El orden con que se presentan los ejemplos no se debe a que supongamos que en el proceso de investigación lo empírico esté determinado por lo teóri-

co, sino que se justifica en gran medida porque es la óptica que nos pareció razonable para incorporar el material de investigación que sirve de base a los desarrollos propiamente estadísticos y porque, además, es una manera simple de introducir al estudioso de la estadística en las complejidades del uso de esta técnica en la investigación en ciencias sociales.

Tomando en cuenta las anteriores consideraciones y con el propósito de dar contenido a las ideas expuestas veamos un par de ejemplos. Nun[4] plantea que en el volumen I de *El capital*, Marx investiga las relaciones entre capital y trabajo *en general* con el propósito de arrojar claridad sobre las leyes de funcionamiento de la forma más pura del modo de producción capitalista y así determinar por qué y cómo la acumulación capitalista produce constantemente una población obrera excedente. En el capítulo 23, Marx teoriza sobre el funcionamiento del mercado de trabajo en una economía de libre competencia y "discierne una primera tendencia al rápido *crecimiento* del proletariado industrial; dada su forma, una segunda tendencia a liquidar las diferenciaciones de la fuerza de trabajo que habían sido propias del periodo de la manufactura y a promover, por tanto, una *homogeneidad* obrera cada vez mayor en términos de calificación y de ingreso" (p. 57). Nun continúa su desarrollo señalando que con estas previsiones Marx describió acertadamente las características que asumió el mercado de trabajo en Inglaterra y los Estados Unidos en el último cuarto del siglo XIX, pero que en esa época se inicia la tendencia hacia la concentración y la centralización del capital, que también Marx pronosticó. Después de referirse a una serie de estudios, Nun nos dice que:

> El resultado de todo esto es una verdadera reversión de aquellas tendencias iniciales al crecimiento y a la homogeneización de un proletariado fundamentalmente industrial: a la vez que cae su peso relativo en el conjunto de la población económicamente activa, esta última aparece cada vez más fragmentada y diferenciada, de manera análoga a lo que sucediera en el periodo de la manufactura (p. 60).

Dice que es a raíz de estos fenómenos que han surgido una serie de estudios que, en oposición a la ortodoxia neoclásica, procuran distinguir la existencia en Estados Unidos de un mercado de trabajo primario y uno secundario. Nun dice que para varios autores este dualismo aparece ligado al corte entre los monopolistas y el resto de los empleadores y que hay suficiente evidencia de que tienen diferentes comportamientos en el mercado de trabajo, por lo menos en lo que se refiere a salarios y beneficios sociales (p. 60).

Una vez que ha marcado los cambios de tendencias en el mercado

[4]Nun, José. "La industria automotriz argentina: estudio de un caso de superpoblación flotante", en *Revista Mexicana de Sociología*, 1 (78), 1978, pp. 55-106.

de trabajo concomitantes al pasaje del capitalismo de su estadio competitivo al monopólico y que ha documentado la segmentación del mercado, Nun describe cómo el capitalismo penetró en el agro de los Estados Unidos; agrega que si bien es cierto que la mayoría de los trabajadores del sector secundario pueden definirse como un excedente respecto del mercado primario, estas ocupaciones no son supervivencias de la organización social anterior sino un emergente del nuevo estadio del desarrollo capitalista (p. 61). A continuación entra de lleno al tema central de su trabajo. Lo reproduciremos textualmente para disponer de la base que apoye nuestra lectura particular:

> Según las ramas, el uso capitalista de la maquinaria continúa sometiendo el empleo industrial a marcados vaivenes. Pero, ¿cuál es la configuración de la modalidad *flotante* de la superpoblación relativa a una industria fuertemente concentrada y dispensadora de ocupaciones predominantemente "primarias"? La disminución proporcional de la fuerza de trabajo fabril en el conjunto de los asalariados no impide que haya aumentado su volumen absoluto: así, éste creció alrededor de 13% entre 1960 y 1972. ¿Vuelven entonces a incorporarse a la industria los obreros desplazados por las fluctuaciones cíclicas o por los progresos de la mecanización? En otras palabras, ¿la característica saliente de la modalidad flotante sigue siendo la analizada por Marx, *con todas sus implicaciones potenciales en términos de solidaridades obreras, de acumulación de experiencias y de reivindicaciones a nivel de la fábrica, y, presumiblemente ahora, de cristalización de una aristocracia del trabajo en el sector monopolista*?
>
> Nuevamente, es difícil dar una respuesta comprensiva y, en todo caso, la literatura que conozco no la provee. Braverman (1974; 386-387) considera que el mecanismo de expulsión/absorción opera hoy en áreas geográficas cada vez más extendidas, de manera que la cesantía obliga a menudo al obrero a "romper los lazos con la localidad y con la comunidad". Pero también sugiere que buena parte de la mano de obra desalojada por "la rápida mecanización de la industria" *no* reingresa a plantas fabriles sino que suministra brazos baratos a los llamados "servicios", o sea, que pasa al mercado de trabajo secundario. Es muy probable que esta última generalización resulte tanto más válida cuanto menor sea el nivel de calificación del despedido; y es igualmente conocida la incidencia que tienen sobre las posibilidades de reempleo variables individuales como la edad, la educación formal, el sexo y el origen étnico (pp. 61-62).

Después de exponer dos estudios sobre el tema de la superpoblación flotante, realizados en un país central (la industria automovilística inglesa) y en uno periférico (el petróleo en Trinidad y Tobago), y mostrar las diferencias, estudia el caso de la superpoblación flotante en la industria automovilística argentina.

Al centrar el interés en las tendencias que muestra el mercado de trabajo en la fase del capitalismo monopólico deberíamos esperar un cierto grado de heterogeneidad en la fuerza de trabajo, como con-

secuencia de la reversión de las tendencias en la etapa competitiva del capitalismo, así como de los procesos de centralización, concentración y segmentación del mercado. Para observar la homogeneidad/heterogeneidad de la fuerza de trabajo Nun toma como indicador el grado de calificación que tenían al momento del despido los obreros de algunas plantas automovilísticas de Buenos Aires, usa el tipo de ocupación al momento del estudio como indicador de la segmentación del mercado de trabajo. La calificación laboral origina una variable que clasifica a los obreros despedidos según tengan o no calificación (variable dicotómica), mientras que la ocupación distingue entre ocupaciones en la industria y en los servicios. Las primeras se abren en ocupaciones dependientes en plantas industriales grandes, que proveen los puestos estables y bien remunerados, y en plantas pequeñas en las que se pagan salarios bajos y los empleos son poco estables. El primer tipo de ocupación es el característico del mercado de trabajo primario y el segundo el del secundario. Si interesase analizar por separado los procesos de segmentación del mercado de trabajo y el de homogeneidad/heterogeneidad de la fuerza de trabajo bastaría con analizar la información de cada indicador también por separado. Sin embargo, en el último párrafo de la larga cita de Nun se plantea la hipótesis de que mientras mayor sea la calificación del obrero despedido, más probable es que reingrese a las plantas fabriles. Esto quiere decir que al nivel de los conceptos se plantea una clara relación entre la heterogeneidad de la fuerza de trabajo y su acceso a los mercados; es esta relación la que permite comprender el estatus del cuadro 2 del estudio que nos sirve de ilustración.

Como por el momento sólo estamos interesados en destacar la relación entre los conceptos y sus indicadores y en señalar cómo los vínculos teóricos especifican las correspondientes asociaciones entre las variables estadísticas nos limitaremos únicamente a presentar la estructura de una tabla que se deriva de los cuadros 2 y 4 del trabajo de Nun[5] (pp. 74 y 77):

TABLA 1

Ocupación actual, por calificación en empresa de origen

	Calificación	
	No calificados	*Calificados*
Ocupaciones dependientes en:		
1. Industria		
1.1. Plantas grandes		XXXXXXXXXX
1.2. Plantas pequeñas	XXXXXXXXXX	
2. En servicios	XXXXXXXXXX	

[5]El desarrollo que presentamos del trabajo de José Nun fue ampliamente discutido con él. Agradecemos sus valiosos comentarios y sugerencias.

Del cuadro 4 del trabajo citado eliminamos la categoría "ocupaciones por cuenta propia" porque no forman parte del mercado de trabajo y por tanto no presentan interés para el tipo de proposición en estudio. Sin embargo, ello no impide su significación descriptiva. Desde el punto de vista empírico recoge una característica importante del desempleo que genera la industria automovilística argentina: aproximadamente un tercio de los trabajadores despedidos, se encontraban trabajando por cuenta propia dos años después de que los cesaron. Esta proporción es casi la misma entre los trabajadores calificados y los no calificados. En este caso tampoco se incluye a los desocupados (alrededor de 10% de los despedidos) por cuanto la hipótesis planteada lleva a examinar el proceso de reinserción laboral.

El análisis estadístico de los totales por columna permitirá formarse un juicio sobre la heterogeneidad/homogeneidad de la fuerza de trabajo. Si las observaciones se concentran básicamente en una de las categorías entonces podemos decir que la fuerza de trabajo es homogénea y la consideraremos heterógenea si se distribuyese por igual en ambas clases estadísticas. En cuanto a la segmentación del mercado de trabajo, la evidencia empírica se encuentra en las líneas. En efecto, en el caso de que la reinserción laboral fuese preferentemente en el mercado secundario las observaciones deberían concentrarse en la categoría "ocupaciones industriales dependientes en plantas pequeñas" y probablemente en las dependientes en los servicios.

La relación entre la heterogeneidad de la fuerza de trabajo y la segmentación del mercado provocaría una inserción diferencial de los desempleados. La hipótesis planteada llevaría a sostener que en el plano de los indicadores los trabajadores sin calificación, a diferencia de los que sí la tienen, tenderían a ocuparse en el mercado secundario. Probablemente habría que complementar la conjetura estableciendo que los trabajadores no calificados volverían al mercado primario siempre que hubiese demanda. Si aceptamos que las categorías estadísticas, "ocupaciones dependientes en plantas pequeñas" y "ocupaciones dependientes en los servicios" constituyen una buena aproximación al concepto de mercado de trabajo secundario entonces, y con base en la proposición empírica (es decir la relación entre los indicadores) deberíamos esperar que las frecuencias de la tabla 1 siguiesen una distribución estadística semejante a la que hemos marcado con X.

Marín[6] plantea en su estudio sobre la violencia en Argentina en el periodo de mayo de 1973 a marzo de 1976 que el Cordobazo marca el inicio de un ascenso de las luchas populares que lleva a un cuestionamiento de la violencia legalizada:

[6]Marín, Juan Carlos. "La guerra civil en Argentina", en *Cuadernos Políticos*, núm. 22, octubre-diciembre, Era, México, 1979.

El "Cordobazo" señaló algo que, hasta ese momento había pasado inadvertido para la enorme mayoría del país: el proceso de acumulación de una fuerza de enfrentamiento en los sectores populares capaz de sobrepasar las fuerzas represivas convencionales del Estado. Se trató esencialmente de una fuerza de masas cuya capacidad destructiva —de las líneas represivas de las fuerzas policiales— estuvo centrada en la convicción de que la continuidad de su lucha era justa: en su armamento moral. Esa fuerza de masas, armada moralmente, había puesto en duda el monopolio estatal de la fuerza material.

Desde ese momento el poder de la burguesía abandonó su desnudez y apeló a la ocupación militar —por una fuerza de guerra— del territorio, para restablecer su monopolio de los instrumentos de destrucción material. El uso policial de fuerzas de guerra fue suficiente para lograr el repliegue del movimiento de masas. La fijación de fuerzas de seguridad en los territorios de conflicto, fue suficiente para contener la fuerza de enfrentamiento de las masas populares. Como contrapartida, el enemigo quedó aislado en su espacio social (p. 51).

Marín dice que en el lapso 1969-1971 tuvo lugar un proceso de acumulación de fuerzas en los sectores populares que comprendía no sólo elementos morales, sino también materiales y que:

Para el enemigo, personificado esta vez en las fuerzas de seguridad, la necesidad de lograr imprescindiblemente el desarme moral de las fuerzas populares se constituyó en una meta prioritaria al logro de su desarme material. Para ello elaboró una estrategia militar centrada en el desarme político de las fuerzas populares; su significación total pasó prácticamente inadvertida durante todo el proceso. Comenzó por un lento repliegue político y una tendencia hacia el acuartelamiento de sus fuerzas armadas.

El "abandono" de los aparatos políticos del Estado creó una falsa imagen de neutralización de las fuerzas armadas en los procesos que comenzaron a desarrollarse a partir de mayo de 1973 (p. 51).

Enseguida sostiene, avalado por información histórica y estadística, que en 1973 se produjo un punto de inflexión en las condiciones de las luchas sociales y políticas entrándose a un periodo de lucha armada. Sostiene:

Tramo de inflexión porque a diferencia del periodo en que el poder de la burguesía se expresaba fundamental y esencialmente concentrado en el Estado, en la coacción permanente de sus diferentes aparatos burocráticos sobre la enorme mayoría de las clases populares, mediante el uso abierto de la fuerza, en el periodo constitucional que comienza aflora —encubierto hasta ese momento por la contradicción fundamental anteriormente señalada, entre el "pueblo" y el Estado— definitivamente y en forma creciente el carácter real de las luchas sociales: su carácter clasista a lo largo de todas las expresiones de la sociedad, y en particular en sus enfrentamientos. De tal modo en la frontalidad que implican los diferentes intereses sociales es que toma lugar la lucha armada; por supuesto, sagaz y hábilmente encubierta por la capaci-

> dad dominante que aún mantenían, y mantienen, las formaciones ideológicas de la burguesía en los sectores populares (p. 53).

Es a partir de estas consideraciones que Marín se preocupa por el análisis empírico de las proposiciones ideológicas dominantes. Basa su estudio en información periodística que le permite dar cuenta de 8 509 hechos armados acontecidos en Argentina de mayo de 1973 a marzo de 1976. A pesar de que somete a contrastación varias proposiciones ideológicas nosotros sólo incluiremos a título de ilustración la que subyace al cuadro 3 del mencionado estudio.

La primera proposición que analiza (proposición teórica) establece una relación entre fuerza social y tipo de violencia. El primer concepto lo objetiva en el de acción social y el segundo en acción contra el cuerpo. De manera que la proposición empírica consistente con la proposición teórica sostendría que existiría una relación entre la acción social y la acción contra el cuerpo. El concepto observable acción social lo dicotomiza en acción subversiva y antisubversiva y la acción contra el cuerpo la divide en dos categorías: las que producen bajas y las que no las producen.

La proposición ideológica concreta se plantea al nivel de los indicadores:

> la gran mayoría de los hechos que producen las fuerzas subversivas son hechos con muertos y heridos;
> en cambio, las fuerzas legales concentran la mayoría de sus hechos armados deteniendo y no matando.

La proposición empírica conduce a la construcción de la siguiente tabla:

TABLA 2

Total de hechos armados producidos en el periodo de mayo de 1973 a marzo de 1976, distinguiéndolos según produzcan bajas o no

	Hechos armados que producen bajas	*Hechos armados que no producen bajas*
Total de hechos armados producidos por la acción "subversiva"	XXXXXXXXXX	
Total de hechos armados producidos por la acción "antisubversiva"		XXXXXXXXXX

En la tabla hemos marcado con la letra X la distribución esperada de las observaciones de acuerdo con las proposiciones planteadas.

En este libro aplicaremos constantemente la lógica que hemos seguido en los ejemplos. Estableceremos las proposiciones teóricas y las relaciones con sus correspondientes proposiciones empíricas y a partir de estas últimas se derivará la distribución esperada de las observaciones que contrastada con la distribución observada, nos entregará criterios para juzgar el grado de concordancia entre la teoría y los datos.

2. Asociación de variables no métricas en tablas de dos por dos

A continuación desarrollaremos unos cuantos conceptos estadísticos para luego centrar la exposición del material en el análisis de sus vinculaciones con lo teórico.

2.1. Distribuciones bidimensionales de frecuencias: variables dicotómicas

Supongamos que tenemos dos variables (cada una indicador de un concepto) simbolizadas por X y Y y que a la primera la dicotomizamos en B y B'[1] y a la segunda en A y A'. El cruce de las dos variables dicotómicas genera una tabla de distribución de frecuencias como la 3:

TABLA 3

Distribución conjunta de frecuencias de dos variables dicotómicas

Y \ X	*B*	*B'*	*Total*
A	N(A∧B)	N(A∧B')	N(A)
A'	N(A'∧B)	N(A'∧B')	N(A')
Total	N(B)	N(B')	N

En ella debemos destacar: *i*) las distribuciones marginales de la variable X y de la variable Y, que informan sobre el número de casos en las categorías B y B' de X y sobre la cantidad de observaciones en las categorías A y A' de la variable Y. *ii*) La distribución conjunta que muestra el número de observaciones que comparten simultáneamente

[1]En B' se agrupan todas las observaciones que no se incluyeron en B. Esto debido a que en una dicotomía las dos clases (en este caso B y B') son mutuamente excluyentes y exhaustivas.

las mismas características de X y Y. Así, N(AΛB) simboliza los casos que pertenecen a las clases B de X y A de Y; N(A′ΛB) representa la cantidad de observaciones que son simúltaneamente B en X y A′ en Y. Del mismo modo se interpretan las otras frecuencias que completan la distribución bidimensional: N(AΛB′) y N(A′ΛB′). *iii*) Que N representa al total de observaciones.

De lo anterior se deriva que los componentes de una tabla bidimensional de frecuencias se pueden desglosar en tres distribuciones. Las dos marginales:

X	*Frecuencias*	*Y*	*Frecuencias*
B	N(B)	A	N(A)
B′	$\frac{N(B')}{N}$	A′	$\frac{N(A')}{N}$

y la conjunta:

X y Y	*Frecuencias*
B y A	N(AΛB)
B′ y A	N(AΛB′)
B y A′	N(A′ΛB)
B′ y A′	N(A′ΛB′)

Para ilustrar los conceptos expuestos consideremos el cuadro 1, tomado de un trabajo de Joaquín Duque y Ernesto Pastrana,[2]que se recategorizó para adecuarlo a los propósitos de nuestra exposición. En él se muestra una distribución bidimensional de frecuencias resultante de una encuesta aplicada a una muestra de la población económicamente activa del Gran Santiago en 1970.

CUADRO 1

Distribución de la población económicamente activa por rama de actividad y contexto ecológico

	Contexto ecológico		
Rama de la actividad productiva	*Poblaciones*	*Campamentos*	*Total*
Industria manufacturera	902	673	1 575
Otras	1 894	961	2 855
Total	2 796	1 634	4 430

[2]Joaquín Duque y Ernesto Pastrana, "La movilización reivindicativa urbana de los sectores populares en Chile", en *Revista Latinoamericana de Ciencias Sociales*, núm. 4, 1972, p. 286.

La variable contexto ecológico es indicador de la "movilización reivindicativa por la vivienda". En efecto, en Chile los trabajadores definidos como "no empleados"[3] tenían derecho a afiliarse al Servicio de Seguro Social y por su intermedio hacerse de una vivienda en un conjunto habitacional denominado población. Esto quiere decir que para llegar a ser vecino de una población era necesario cumplir con los requisitos impuestos por la vía institucional. Por el contrario, los campamentos se constituían por tomas de terrenos que implicaban una modalidad de presión organizada y políticamente orientada que entrañaba de hecho un rechazo a las normas de apropiación privada del espacio urbano y una acción que desbordaba las normas estatales de acceso a la vivienda. Los campamentos se constituían por canales extrainstitucionales.

La movilización reivindicativa por la vivienda reconoce sólo dos caminos, que pueden suponerse mutuamente excluyentes: el institucional y el extrainstitucional. El indicador (o variable estadística) que trata de dar cuenta de este concepto es el "contexto ecológico", que se ha dicotomizado en poblaciones y campamentos. Las primeras están en correspondencia directa con la "vía institucional" y los segundos con la "no institucional".

Nótese que en relación con el concepto "movilización reivindicativa por la vivienda", la variable "contexto ecológico" permite clasificar sin ambigüedad a la población económicamente activa (PEA) y que es exhaustiva en cuanto a la población que obtuvo éxito en su lucha por techo. En esta investigación sólo interesaba aquella parte de la PEA que se podía englobar con el calificativo de "no empleado"[4] y que se encontraba asentada en una "vivienda" propia obtenida socialmente (es decir, no por herencia ni individualmente) ya sea por vías institucionales o extrainstitucionales.

La variable "contexto ecológico" no intenta clasificar los lugares de residencia de los "no empleados" (en cuyo caso no cumpliría con el requisito de ser exhaustiva porque hay otros contextos de residencia) sino que se usa como indicador de movilización exitosa por techo; en este sentido cumple con la condición de tener categorías mutuamente excluyentes (se supone que los no empleados socialmente asentados no pueden vivir a la vez en un campamento y en una población) y exhaustivas por cuanto el camino institucional conduce a una población y la vía extrainstitucional a un campamento.

La variable "rama de la actividad productiva" es un indicador de

[3]Dadas las características de la población estadística en la cual se basó el estudio, la categoría "no empleado" engloba a todas aquellas ocupaciones que habitualmente se denominan manuales no calificadas.

[4]Se trata de aquella parte de la población de menores ingresos que no cotizaban, según la ley, a instituciones de seguridad específicas para empleados.

marginalidad. Los autores habían expuesto en un trabajo previo[5] un concepto de marginalidad basado en el tipo de actividad económica que demanda el modelo de desarrollo.[6] El argumento que permite establecer este vínculo sostiene que las distintas etapas por las que atraviesa el desarrollo de un país permiten distinguir, en cada caso, entre actividades económicas centrales y no centrales. Así por ejemplo, en la época en que el modelo de desarrollo era hacia afuera las ramas en que se basaba la dinámica económica fueron el transporte, el comercio y las actividades relacionadas con la producción del o de los bienes de exportación. En la etapa del desarrollo hacia adentro, la actividad económica que signaba el ritmo de crecimiento sería fundamentalmente la industrial.

Sobre la base de estas consideraciones, los autores plantean que las ocupaciones centrales para el desarrollo económico orientado hacia adentro serían las de la industria manufacturera. En las actividades de construcción y servicios (que conforman básicamente el grupo "otras") se llevarían a cabo labores superfluas para el modelo de desarrollo, en especial si se toma en cuenta que la PEA que se estudia es la de "no empleados" con bajos niveles de ingreso. En la categoría "otras ramas" se aglutinan vendedores ambulantes y de servicios menores, obreros de la construcción, etcétera.

Las ramas de la actividad productiva según las cuales se clasifica a la PEA indican el carácter de marginal o no marginal. Como marginales se incluye a aquellas personas con inserción laboral en la rama "otras" y como no marginales las que trabajan en la industria.

La rama de la actividad productiva que se usa como indicador (o variable estadística) de marginalidad es dicotómica en tanto cualquier miembro de la PEA labora ya sea en la industria manufacturera o en una unidad de producción clasificada en su categoría complementaria y exhaustiva por cuanto no hay ninguna persona ligada a la actividad productiva que no se pueda clasificar en una u otra categoría.

Las distribuciones marginales del cuadro de doble entrada son:

CUADRO 2

Distribuciones marginales de la PEA por contexto ecológico y rama de actividad productiva

Contexto ecológico (X)	*Frecuencia*	*%*	*Rama de actividad (Y)*	*Frecuencia*	*%*
Poblaciones	2 796	63.1	Industria manufacturera	1 575	35.6
Campamentos	1 634	36.9	Otras	2 855	64.4
Total	4 430	100.0	*Total*	4 430	100.0

[5]J. Duque y E. Pastrana, *Elementos teóricos para la interpretación de los procesos organizativo-políticos poblacionales*, ELAS-FLACSO, Santiago, diciembre de 1971.

[6]Concepto que difería del planteado por Roger Vekemans, que más bien aludía a una marginalidad de carácter cultural.

Las distribuciones marginales muestran que los "no empleados" que han tenido éxito en su lucha por la vivienda habitan preferentemente en poblaciones (63.1%) y laboran en su gran mayoría fuera de la industria manufacturera (64.4 por ciento).

Ahora bien, el cuadro 1, que se origina en el cruce del contexto ecológico con la rama de la actividad productiva, también puede interpretarse desde el punto de vista de la teoría de la marginalidad en boga por esos años. Al respecto Jorge Giusti plantea[7] que:

> Resumiendo, dividiremos la teoría del Desarrollo Social para América Latina (DESAL) en cinco aspectos principales, que son los siguientes:
>
> *Ecológico:* Desde una perspectiva ecológica o territorial, el DESAL distinguía tres tipos de vivienda, tomando en cuenta las condiciones físicas de las casas y los lugares en que se hallaban emplazadas. Estos tipos son: a) los "círculos de miseria", constituidos por los grupos de viviendas precarias que forman vecindarios donde las viviendas aparecen de manera espontánea; b) las viviendas viejas y deterioradas dentro de la ciudad y c) el vecindario planificado de origen estatal o privado.
>
> *Sociopsicológico:* sobre la base de la mencionada diferenciación ecológica, DESAL afirmaba que cada tipo de vivienda genera un grupo humano distinto. Centralizando su atención sobre los pobladores que viven en círculos de miseria (viviendas precarias), el DESAL desarrolló una serie de rasgos que caracterizarían a los pobladores. Dichos rasgos son:
>
> El vocablo poblador que se utiliza para describir a las masas urbanas marginales ilustra su reducida capacidad para actuar. Ellos simplemente "pueblan" el lugar, sólo "son", y nada más.
>
> Marginalidad significa falta de participación en los beneficios y recursos sociales y en la red de decisiones sociales.
>
> Marginalidad es falta de integración interna. La falta de participación surge de otro rasgo constituyente de marginalidad: la falta de integración interna de los grupos "marginales".
>
> El hombre "marginal" se caracteriza por su incapacidad de superar su situación actual por sí mismo. La marginalidad es tan radical que responde a la definición etimológica: constituye un problema que afecta las mismas raíces del ser del hombre: corroe la médula del potencial del hombre para el automejoramiento voluntario y racional.
>
> *Sociocultural:* entre aquellos beneficios y recursos sociales de los que carecen los pobladores, los vinculados con su trasfondo social y cultural, se encuentran: bajos niveles de vida, de salud y de vivienda y bajos niveles educacionales y culturales.
>
> *Económico:* económicamente, puede definirse al grupo humano marginal, manifestaba el DESAL, como "subproletarios", porque comparten una carencia de dos factores: ingreso y empleo. Su contribución a la productividad es mínima y resulta en salarios con los que apenas subsisten. La falta

[7]En *Organización y participación popular en Chile: el mito del hombre marginal*, Ediciones FLACSO, Buenos Aires, 1973.

de empleo estable es tal que frecuentemente no es otra cosa que desempleo disimulado.

Político: como vimos previamente, el otro aspecto de la falta de participación de la gente "marginal" se relaciona con la sociedad como una red de decisiones sociales. Es decir, que los pobladores no toman parte de las responsabilidades o tareas que deben ser emprendidas para la solución de problemas sociales en general, y para la solución de sus propios problemas en particular. Los grupos marginales, consecuentemente, no cuentan con los tipos de organización que los vincularían con la sociedad y que además, les brindarían "representación" frente a las instituciones de la sociedad participante.[8]

De las líneas anteriores se puede desprender, por una parte, que el contexto ecológico es la característica básica que determina la condición de marginalidad y, por otra, que deberíamos esperar una estrecha relación entre el lugar de residencia y la participación en la actividad económica. En buenas cuentas, según la versión de la teoría del DESAL, que nos entrega Jorge Giusti, los campamentos estaban formados por subproletarios. En cuanto a las características de los habitantes de las poblaciones no es muy explícito, pero se podría suponer que en ellas existía una cierta heterogeneidad.

De acuerdo con esta versión de la teoría de la marginalidad del DESAL se pueden especificar dos hipótesis sobre la relación entre las variables en estudio:

Si B: poblaciones , B′: campamentos,
A: industria manufacturera , A′: otras industrias,

se podría sostener la proposición empírica P_1:

$$P_1: \quad B \curvearrowright A$$
$$B' \curvearrowright A'$$

Donde $B \curvearrowright A$ significa "si B entonces A". En consecuencia, esta primera proposición establece que si una persona vive en una población entonces es no marginal, y, de acuerdo con la teoría, debería trabajar en la industria manufacturera; si habita en un campamento entonces es marginal. Es decir, sólo los marginales de la sociedad usan medios extrainstitucionales para resolver su problema habitacional en tanto que los no marginales recurren exclusivamente a las vías legales.

La otra hipótesis (P_2) que se puede sostener según esta versión de

[8]Jorge Giusti, *op. cit.*, pp. 60-62.

la teoría de la marginalidad cultural se puede expresar en los siguientes términos:

$$P_2: \quad B \rightsquigarrow (A \text{ o } A')$$
$$B' \rightsquigarrow A'$$

Esta proposición plantea que de acuerdo con la teoría deberíamos esperar que los habitantes de las poblaciones desempeñaran su actividad económica en cualquier rama productiva; esto quiere decir que un poblamiento regular estará compuesto tanto por marginales como por no marginales y que los campamentos debieran estar constituidos sólo por aquéllos.

La hipótesis establece que tanto los marginales como los no marginales usan el sistema institucional para resolver el problema de la vivienda. Pero que sólo los primeros recurren al medio extralegal.

Las dos formalizaciones lógicas son consistentes con el planteamiento de carácter general sustentado por Giusti. Comparten como proposición central que sólo los marginales recurren a caminos no institucionales para resolver el problema de la vivienda.

Esta versión de la teoría prácticamente asimila o subsume la marginalidad en el contexto ecológico. Son marginales aquellos que viven en campamentos y no lo son los que habitan en poblaciones. De aquí que sea más probable que la primera de las proposiciones lógicas refleje con más fidelidad el pensamiento del autor.

De la teoría pareciera derivarse que la variable explicativa (o independiente) sería el contexto ecológico y que la variable explicada (o dependiente) sería la rama de la actividad productiva en que se desempeña el trabajador. El hecho de que una persona viva en una población o en un campamento bastaría para establecer su carácter de marginal o no marginal y de acuerdo con su condición se esperaría que su inserción laboral fuese en aquellas ramas en que predomina el subproletariado o en la industria manufacturera.

El cruce de las variables contexto ecológico y rama de la actividad productiva es un punto de encuentro entre las dos teorías de la marginalidad. Aún cuando la perspectiva del Desal destaca lo cultural y la histórico-estructural privilegia lo económico, existe un vínculo entre ellas que justifica su coincidencia desde el punto de vista de lo empírico. En efecto, según el planteamiento de Giusti, la marginalidad se determina por el lugar de residencia (contexto ecológico) en tanto que de acuerdo con Duque y Pastrana por la inserción laboral. El primero de los conceptos de marginalidad se objetiva a través del contexto ecológico observable y el segundo por la inserción laboral.

El planteamiento de la "marginalidad económica" supone que el

grado de politización de la fuerza de trabajo y las dificultades para satisfacer la demanda de vivienda por los canales institucionales, crearon las condiciones para que los trabajadores buscaran como segundo camino el extralegal, aunque ello significara desconocer uno de los derechos básicos en que se apoya el capitalismo: la propiedad privada. Según las concepciones de la marginalidad económica los marginales y los no marginales participan por igual en la creación de un campamento; sólo los aglutinaba la imposibilidad de resolver el problema habitacional dentro del marco legal y una posición política relativamente compartida.

Esta postura sostiene que la marginalidad *no* determinaba el enrolamiento en la movilización reivindicativa por la vivienda. Al nivel de los indicadores esto significa que no hay relación entre la rama de la actividad (variable explicativa o independiente) y el contexto ecológico (variable explicada o dependiente).

De los desarrollos conceptuales presentados se deriva que la teoría de la marginalidad cultural proporciona elementos para que empíricamente se presente una estrecha relación entre el contexto ecológico y la rama de la actividad productiva, en que el primero sería la variable independiente y la segunda la dependiente. La teoría de la marginalidad económica no sólo se diferencia porque plantea la inexistencia de tal asociación, sino que invierte su sentido. Ahora la variable explicativa (independiente) sería la inserción laboral y la explicada el contexto ecológico.[9]

Una vez que se han especificado las hipótesis que derivan de ambas teorías, es decir, una vez conocidas las proposiciones teóricas y sus correspondientes proposiciones empíricas, a saber:

i) según la teoría de la marginalidad cultural[10] debe haber una relación entre los dos indicadores, de modo que se pueda predecir la inserción laboral a partir del tipo de asentamiento urbano.

$$P_1: \quad B \frown\!\!\rightarrow A$$
$$B' \frown\!\!\rightarrow A'$$

ii) y según la marginalidad económica no debiera haber relación entre las variables, por cuanto el proceso de marginalización no determina la movilización.

$$P_3: \quad A \frown\!\!\rightarrow (B \text{ o } B')$$
$$A' \frown\!\!\rightarrow (B \text{ o } B')$$

[9]Ambas teorías hacen descansar la explicación de los fenómenos que estudian en la marginalidad, pero este nombre remite a dos conceptos distintos que se objetivan de manera diferente, lo que hace cambiar el sentido del vínculo explicativo.

[10]Por el momento nos limitaremos a trabajar sólo con la proposición P_1; más adelante incluiremos P_2.

¿Cómo podemos validarlas? ¿Con cuál de las dos teorías concuerdan los datos? ¿Cuál de las conceptualizaciones entra en contradicción con la evidencia empírica?

Hasta este punto hemos examinado la relación entre las proposiciones teóricas y las proposiciones empíricas en el plano lógico. Interesa ahora (para estar en condiciones de responder a las preguntas formuladas) estudiar los vínculos que permiten unir las proposiciones empíricas con la distribución de los datos. Para este propósito será necesario dedicarse a exponer algunos principios básicos que permiten el análisis estadístico de una tabla de frecuencias de doble entrada en el caso particular en que las dos variables son dicotómicas.

2.2. Noción de independencia estadística

Tomemos la proposición empírica P_1 postulada por la teoría de la marginalidad cultural y examinemos cuál es la distribución esperada de los datos. De acuerdo con P_1 los datos del cuadro 1 se deberían ubicar *preferentemente* en las casillas que forman la diagonal principal de la tabla (la que va de izquierda a derecha). Decimos que *preferentemente*, a pesar de que la teoría sostiene que *exactamente* porque por razones de azar (o si se quiere por otros factores no considerados en el análisis, cuyo peso explicativo es insignificante) estamos dispuestos a aceptar *algunas* observaciones en la diagonal secundaria sin que ello invalide la proposición empírica y en consecuencia la teórica.

Los términos *preferentemente* y *algunas* tienen la suficiente vaguedad como para admitir que con una misma distribución bidimensional de datos un investigador decida rechazar la proposición empírica y por ende la teórica, mientras que otro no. ¿Cómo establecer un criterio que permita decidir hasta qué punto las desviaciones se pueden tolerar, de modo que no se rechacen las proposiciones?

La respuesta a esta pregunta debe buscarse por dos vías no excluyentes: una, establecer una medición de la fuerza con que se asocian las variables; otra, descartar que la asociación observada se debe al azar. Este último camino cae dentro del ámbito de la inferencia estadística y por tanto no se tratará en este libro. En adelante nos preocuparemos por la manera como procede la estadística para medir el grado de relación entre dos variables dicotómicas.

La idea de asociación se define por oposición a la de independencia. Las dos teorías de marginalidad se contradicen porque una plantea que debe existir una relación entre las variables en tanto que la otra sostiene que deben ser independientes.

En este apartado examinaremos el concepto de independencia estadística y diferentes maneras de objetivarlo, así como la forma de medir la fuerza o el grado de relación de dos variables.

A primera vista, los datos del cuadro 1 parecieran aportar alguna evidencia en favor de la teoría de la marginalidad cultural ya que en las poblaciones hay más trabajadores industriales (902) que en los campamentos (673). Decimos que alguna evidencia porque 673 trabajadores industriales que habitan en campamentos parece un desvío demasiado grande. Sin embargo, ese mayor número sólo es aparente por cuanto los totales marginales son diferentes. De los 2 796 habitantes de poblaciones 902 trabajan en la industria, mientras que de las 1 634 personas que viven en campamentos 673 laboran en la manufactura. La comparación de los valores absolutos carece de sentido por cuanto se ubican en totales de distinto tamaño. Para eliminar el efecto del tamaño diferencial, las frecuencias absolutas se expresan en proporciones (que multiplicadas por 100 constituyen porcentajes) y posteriormente se procede a examinar la diferencia entre ellas.

Al calcular los porcentajes[11] con respecto al total de cada columna del cuadro 1, obtenemos:

CUADRO 3

Distribución porcentual de la PEA por rama de actividad económica dentro de cada contexto ecológico

	Contextos ecológicos		
Rama de actividad	*Poblaciones*	*Campamentos*	*Total*
Industria manufacturera	32.3	41.2	35.6
Otras	67.7	58.8	64.4
Total	100.0	100.0	100.0

Al controlar el tamaño de las poblaciones y campamentos aparece una imagen totalmente distinta de la que obtuvimos con las frecuencias absolutas. El cuadro 3 nos dice que la situación es más bien la opuesta a la que plantea la proposición empírica de la marginalidad cultural. En efecto, el porcentaje de trabajadores de la industria manufacturera es mayor en los campamentos (41.2%) que en las poblaciones (32.3%). Este resultado permite rechazar la hipótesis sostenida por la teoría de la marginalidad cultural por cuanto las observaciones no sólo no se agrupan en la diagonal principal, sino que además se comportan de manera opuesta a la hipotetizada. ¿También rechaza la proposición planteada por la teoría de la marginalidad económica?

[11] De aquí en adelante no nos preocuparemos por distinguir con demasiada precisión entre proporción y porcentaje.

Sabemos que plantear que la movilización reivindicativa urbana no se relaciona con la marginalidad (proposición teórica) origina la proposición empírica:

$$P_3: \quad A \frown (B \text{ o } B')$$
$$A' \frown (B \text{ o } B')$$

El examen del vínculo entre la proposición empírica y la distribución de los datos requiere que se controle el tamaño cuando las frecuencias marginales son distintas, lo que equivale a decir que se deben calcular porcentajes.

De acuerdo con la proposición empírica, los trabajadores de la industria manufacturera (no marginales) y los de las otras ramas (marginales) se deben distribuir por igual entre poblaciones y campamentos. A diferencia de la proposición P_1 de la teoría de la marginalidad cultural que predicaba que las observaciones debían concentrarse en la diagonal principal del cuadro, la de la marginalidad económica P_3, nos dice que deben distribuirse de la misma manera dentro de cada rama en ambos contextos ecológicos. La proporción de marginales en las poblaciones debiera ser similar a la de no marginales.

En consecuencia, la marginalidad y la movilización reivindicativa urbana serán independientes si comprobamos que no existe relación en el plano de los indicadores (rama de la actividad productiva y contexto ecológico). Para saber si hay o no concomitancia debemos comparar el porcentaje de marginales que viven en poblaciones con el correspondiente a no marginales que habitan en el mismo tipo de asentamiento.[12] Las variables serán independientes si ambos porcentajes son estadísticamente iguales.

CUADRO 4

Distribución porcentual de la PEA por contexto ecológico dentro de cada rama de actividad económica

	Contextos ecológicos		
Rama de actividad	*Poblaciones*	*Campamentos*	*Total*
Industria manufacturera	57.3	42.7	100.0
Otras	66.3	33.7	100.0
Total	63.1	36.9	100.0

[12]En una tabla de dos por dos basta con comparar esos porcentajes para juzgar acerca de la relación entre las variables, por cuanto los restantes son el complemento a cien. Es decir, si conocemos el porcentaje de trabajadores que viven en poblaciones, la fracción complementaria habita en campamentos.

Pareciera que la hipótesis de independencia que sostiene la teoría de la marginalidad económica muestra un cierto grado de adecuación con los datos. Los porcentajes de trabajadores que han logrado su asentamiento urbano a través de alguna vía institucional es similar entre los no marginales (57.3%) y los marginales (66.3%). Estos resultados llevan a Duque y Pastrana a sostener que:

> Analizando las características ocupacionales de la fuerza de trabajo de las poblaciones y campamentos, puede descubrise el carácter ideológico de las interpretaciones que categorizan estos contextos, bien sea como "marginales" o bien como formados por el "lumpen proletariado". Por el contrario, los antecedentes confirman la heterogeneidad de las situaciones de clase de los sectores populares urbanos, donde a la presencia del proletariado se suma la concurrencia del subproletariado, integrado por los trabajadores de la construcción, de los servicios personales de baja calificación, apreciables proporciones de trabajadores independientes de las más variadas actividades (artesanales, comerciales, paraproductivas industriales).[13]

Si bien podemos sostener que hay adecuación entre la proposición empírica y la distribución de los datos aún no estamos en condiciones de generar medidas estadísticas que permitan saber: *i*) si hay o no relación entre las dos variables dicotómicas y *ii*) la fuerza o grado de la relación. Para lograr este propósito es necesario establecer el concepto de independencia estadística. Antes de entrar a su desarrollo, debemos dejar constancia de un procedimiento que se encuentra implícito en los cálculos que ya hemos realizado.

Los porcentajes del cuadro 3 se obtuvieron de los totales de las columnas y los del cuadro 4 de los totales por línea. En ambos casos el cálculo se hizo tomando en cuenta la argumentación teórica: los porcentajes del cuadro 3 nos indican la distribución relativa de los trabajadores que se desempeñan en la industria manufacturera o en otras industrias en las poblaciones y en los campamentos. En el cuadro 4 señalan el peso relativo de los habitantes de cada contexto ecológico dentro de la PEA industrial y de las otras ramas. Sabemos que según los desarrollos de la teoría de la marginalidad cultural la variable independiente es el contexto ecológico y sus frecuencias marginales son los totales por columna; para la teoría de la marginalidad económica la variable independiente es la rama de la actividad productiva y sus frecuencias marginales se encuentran en las líneas. En consecuencia, siguiendo el hilo del argumento teórico, que se expresa a través de la proposición empírica, concluimos que los porcentajes los calculamos basándonos en el número de observaciones por categoría de la variable independiente. *Es una convención ampliamente aceptada que los porcentajes se calcu-*

[13]Duque y Pastrana, "La movilización...", *op. cit.*, p. 287.

len sobre las frecuencias marginales, de la variable independiente. Su justificación deriva de su coherencia lógica con la proposición empírica. Decimos que es una convención porque la comparación de porcentajes entrega el mismo resultado independientemente de que se calculen respecto a los totales de las correspondientes líneas o columnas. Obsérvese que las discrepancias porcentuales de los cuadros 3 y 4 son las mismas.

Considérese la siguiente tabla genérica:

TABLA 4

Distribución porcentual conjunta de dos variables dicotómicas

$Y \backslash X$	B	B'	*Total*
A	P(A/B)	P(A/B')	P(A)
A'	P(A'/B)	P(A'/B')	P(A')
Total	100	100	100

Se supone que X es la variable independiente y Y la dependiente. De acuerdo con la convención establecida, P(A/B) y P(A'/B) simbolizan los porcentajes de observaciones que tienen la característica A y A', respectivamente, en el total de la columna B. Del mismo modo P(A/B') y P(A'/B') indican los porcentajes que poseen las características A y A' en el total de B'. P(A) y P(A') son los porcentajes de observaciones que en el total tienen el atributo A y A' respectivamente (compárense estas definiciones con las dadas a propósito de los símbolos incluidos en la tabla 3).

Las variables X y Y serán estadísticamente independientes si se cumple que:

$$P(A/B) = P(A/B') \tag{1}$$

Las variables X y Y son estadísticamente independientes si el porcentaje de observaciones que poseen el atributo A (PEA ocupada en la industria manufacturera) es el mismo en B (poblaciones) que en no B (campamentos). Como se puede apreciar todos los desarrollos de esta sección han tenido como base la noción de independencia estadística.

Es fácil demostrar que si en una tabla de dos por dos se cumple la

igualdad de porcentajes que supone la independencia estadística entonces también se cumple que:[14]

$$P(A/B) = P(A/B') = P(A) \tag{2}$$

Es decir, ambos porcentajes son iguales al porcentaje marginal de observaciones que tienen la característica A. Del mismo modo:

$$P(A'/B) = P(A'/B') = P(A') \tag{3}$$

Las triples igualdades (2) y (3) permiten derivar que:[15]

[14]Tomemos la expresión:

$$P(A) = P(A/B)P(B) + P(A/B')P(B') = \frac{N(A \wedge B)}{N(B)} \frac{N(B)}{N} + \frac{N(A \wedge B')}{N(B')} \cdot \frac{N(B')}{N} = \frac{N(A)}{N}$$

y supongamos que $P(A/B) = P(A/B')$,

es decir que:

$$\frac{N(A \wedge B)}{N(B)} = \frac{N(A \wedge B')}{N(B')},$$

entonces,

$$\frac{N(A \wedge B)}{N(B)} \left[\frac{N(B) + N(B')}{N} \right] = \frac{N(A)}{N}.$$

Por tanto,

$$\frac{N(A \wedge B)}{N(B)} = \frac{N(A)}{N} \quad \text{y} \quad P(A/B) = P(A).$$

[15]Así por ejemplo si tomamos la expresión (2):

$$P(A/B) = P(A)$$

y la expresamos en términos de frecuencias, tenemos que:

$$\frac{N(A \wedge B)/N}{N(B)/N} = \frac{N(A)}{N};$$

luego,

$$N(A \wedge B) = \frac{N(A)N(B)}{N}$$

$$(4)\quad \begin{cases} 4.1. \quad N(A \wedge B) = \dfrac{N(A)N(B)}{N} \\ 4.2. \quad N(A' \wedge B) = \dfrac{N(A')N(B)}{N} \\ 4.3. \quad N(A \wedge B') = \dfrac{N(A)N(B')}{N} \\ 4.4. \quad N(A' \wedge B') = \dfrac{N(A')N(B')}{N} \end{cases}$$

La igualdad (1) y las ecuaciones (4) muestran dos maneras diferentes de expresar el principio de independencia estadística. La primera nos dice que las variables X y Y son estadísticamente independientes si la proporción de observaciones que son A es la misma en B y B′. Este criterio lo aplicamos en los cuadros 3 y 4 para estudiar el vínculo entre el contexto ecológico y la rama de actividad productiva. Las segundas (las ecuaciones 4) nos dicen que si las dos variables son estadísticamente independientes entonces cada frecuencia de la distribución conjunta debiera coincidir con el producto de sus correspondientes marginales, dividido entre el total de observaciones.

Como las igualdades (4) se obtuvieron bajo el supuesto de independencia estadística, se concluye que deben interpretarse como las frecuencias esperadas en caso de que no hubiese relación entre X y Y. Su aplicación sistemática genera una *distribución bidimensional de frecuencias esperadas* cuya construcción descansa en el supuesto de ausencia de relación entre las variables.

Tenemos las distribuciones conjuntas observada y esperada de X y Y, que se traducen en dos tipos de frecuencias conceptualmente diferentes (aunque a veces coincidan numéricamente) para cada casilla, lo que hace necesario introducir una notación que permita distinguirlas. De aquí en adelante usaremos N como un operador[16] que, aplicado sobre un atributo, da como resultado frecuencias observadas. Así, por ejemplo, N(A∧B) simboliza al número de casos que tienen simultáneamente las características A y B. Para representar las frecuencias esperadas usaremos E como un operador, de manera que al actuar sobre un atributo específico nos da el número de observaciones que deberíamos esperar bajo el supuesto de independencia estadística. Como las ecuaciones (4) sirven para calcular las frecuencias esperadas es necesario reescribirlas de acuerdo con esta nueva simbología.

[16]También representa al total de observaciones.

$$(5)\quad \begin{cases} 5.1.\ E(A\Lambda B) = \dfrac{N(A)N(B)}{N} \\ 5.2.\ E(A'\Lambda B) = \dfrac{N(A')N(B)}{N} \\ 5.3.\ E(A\Lambda B') = \dfrac{N(A)N(B')}{N} \\ 5.4.\ E(A'\Lambda B') = \dfrac{N(A')N(B')}{N} \end{cases}$$

En la tabla 5 se muestra la distribución de frecuencias esperadas conjuntas de las dos variables dicotómicas.

TABLA 5

Distribución esperada conjunta de dos variables dicotómicas

$Y \backslash X$	B	B'	*Total*
A	E(AΛB)	E(AΛB′)	E(A)
A′	E(A′ΛB)	E(A′ΛB′)	E(A′)
Total	E(B)	E(B′)	N

No es difícil demostrar que:

$$N(B) = E(B)\ ;\ N(B') = E(B')\ ;\ N(A) = E(A)\ \text{y}\ N(A') = E(A')$$

y, por tanto, el total de observaciones es igual en ambas distribuciones.

Al aplicar las ecuaciones (5) a los datos del cuadro 2 tenemos:

$$E(A\Lambda B) = \frac{1\,575 \times 2\,796}{4\,430} = 994\ ;$$

$$E(A\Lambda B') = \frac{1\,575 \times 1\,634}{4\,430} = 581$$

$$E(A'\Lambda B) = \frac{2\,855 \times 2\,796}{4\,430} = 1\,802\ ;$$

$$E(A'\Lambda B') = \frac{2\,855 \times 1\,634}{4\,430} = 1\,053$$

Con estas frecuencias podemos construir el cuadro 5.

CUADRO 5

Distribución esperada* de la PEA por rama de actividad productiva y contexto ecológico

	Contexto ecológico		
Rama de actividad productiva	*Poblaciones*	*Campamentos*	*Total*
Industria manufacturera	994	581	1 575
Otras	1 802	1 053	2 855
Total	2 796	1 634	4 430

*De acuerdo con el supuesto de independencia estadística.

Con la comparación del cuadro de frecuencias esperadas con el de frecuencias observadas (cuadro 2) queremos destacar que:

i) las distribuciones marginales son iguales y por lo tanto el número total de observaciones es el mismo.

ii) ambas distribuciones conjuntas son, en términos generales, bastante parecidas. Como la distribución observada se desvía levemente de la esperada, se puede afirmar que hay una ligazón débil entre ambas variables o, en otros términos, que los datos parecen sustentar la hipótesis de la marginalidad económica, no así la de la cultural. Esta conclusión es similar a la que habíamos obtenido con la comparación de porcentajes, y no podía ser de otro modo por cuanto los dos procedimientos son equivalentes desde el punto de vista matemático.

Es usual que cuando el análisis se efectúa por la vía de las frecuencias esperadas no se presenten dos cuadros sino sólo uno, poniéndolas entre paréntesis, inmediatamente abajo de su correspondiente frecuencia observada.

CUADRO 6

Distribuciones observada y esperada de la PEA por rama de actividad productiva y contexto ecológico

	Contexto ecológico		
Rama de actividad productiva	*Poblaciones*	*Campamentos*	*Total*
Industria manufacturera	902	673	1 575
	(994)	(581)	
Otras	1 894	961	2 855
	(1 802)	(1 053)	
Total	2 796	1 634	4 430

En las marginales y en el total no se anotan las frecuencias esperadas porque son iguales a las observadas. Esta tabla presenta de manera sintética las distribuciones de los cuadros 1 y 5 y permite apreciar a simple vista los desvíos de la distribución conjunta observada respecto de la independencia estadística.

Para saber si dos variables dicotómicas son o no estadísticamente independientes se puede recurrir a dos procedimientos. El primero consiste en comparar los porcentajes y en caso de que su diferencia sea pequeña o nula se dice que las variables son estadísticamente independientes. El segundo consiste en examinar las discrepancias entre las frecuencias observadas y esperadas; en caso de que coincidan o difieran poco, se dice que no hay asociación entre X y Y. Como se ve las dos vías de análisis son equivalentes desde los puntos de vista matemático y estadístico lo que necesariamente conduce a los mismos resultados sustantivos.

Hemos descrito dos maneras de analizar una distribución conjunta de dos variables dicotómicas que nos permiten apreciar si están o no relacionadas, pero no nos dan pie para construir un juicio acerca del grado o la fuerza del vínculo que las une.

2.3. Otras formas de medir asociación

El concepto de independencia estadística formalizado en la ecuación (1) contiene la idea básica para definir asociación y abordar el problema de su medición. La fuerza de la relación entre dos variables se define por su lejanía respecto a la independencia estadística. A mayor distancia mayor asociación. En lo que respecta a los porcentajes, esto quiere decir que la asociación entre X y Y será mayor mientras más grande sea la diferencia entre P(A/B) y P(A/B′).[17]

De acuerdo con esta idea se podría definir como indicador del grado o de la fuerza de la relación entre las variables la diferencia:

$$D = P(A/B) - P(A/B') \tag{6}$$

En el caso de independencia estadística D = 0, por cuanto se cumple que P(A/B) = P(A/B′). Cuando P(A/B) > P(A/B′) el valor de D será positivo; entonces se dice que la relación entre X y Y es directa o positiva. Si P(A/B) < P(A/B′) entonces D será menor que cero y la asociación entre las variables se califica como inversa o negativa.

[17]Esta idea difiere del significado corriente que se adscribe a la palabra asociación. En su uso ordinario se dice que A y B están asociados si ambos atributos aparecen juntos con alguna frecuencia. En esta acepción A y B están asociados sólo si aparecen conjuntamente en un número mayor de casos que el esperado si ellos fuesen independientes.

Se tiene el máximo de relación positiva directa (la fuerza de la asociación es máxima) cuando:

$$P(A/B) = 100 \qquad \text{y} \qquad P(A/B') = 0$$

lo que implica que:

$$D = 100$$

El caso de máxima relación inversa se produce cuando

$$P(A/B) = 0 \qquad \text{y} \qquad P(A/B') = 100$$

y:

$$D = -100$$

Tenemos, entonces, que la discrepancia (D) puede asumir valores entre 100 y −100 con los siguientes significados:

$$(7) \qquad D = \begin{cases} 100\text{: máxima relación positiva} \\ 0\text{: independencia estadística} \\ -100\text{: máxima relación negativa} \end{cases}$$

Por medio de la ecuación (6) hemos construido una escala que permite calibrar el grado o la fuerza de la relación a partir de una simple diferencia entre porcentajes. Así, por ejemplo, al calcular D con los datos del cuadro 3 se obtiene: $D = 32.3 - 41.2 = -8.9$, valor que denota una baja relación inversa, lo que apoya la hipótesis de la marginalidad económica. Este valor, mucho más cercano a 0 que a 100, permite afirmar que el contexto ecológico determina débilmente, o no determina,[18] la marginalidad.

Otra vía para objetivar el concepto de independencia estadística nos la proporcionan las ecuaciones (5). En efecto, la fuerza con que se relacionan las variables se puede medir a través de la discrepancia entre las frecuencias observadas y las esperadas, de acuerdo con el supuesto de independencia estadística. Este camino consiste en establecer una función que varíe con el grado de concordancia entre la distribución efectiva de las observaciones y la que debería haberse observado (esperada) si X y Y no estuviesen relacionadas. Así tendríamos:

[18]Esta imprecisión tal vez podría resolverse a través de una prueba de hipótesis, tema estadístico que queda fuera de los límites de este libro.

$$(8)\quad \begin{cases} 8.1.\ D_{11} = N(A\Lambda B) - E(A\Lambda B) = N(A\Lambda B) - \dfrac{N(A)N(B)}{N} \\ 8.2.\ D_{12} = N(A\Lambda B') - E(A\Lambda B') = N(A\Lambda B') - \dfrac{N(A)N(B')}{N} \\ 8.3.\ D_{21} = N(A'\Lambda B) - E(A'\Lambda B) = N(A'\Lambda B) - \dfrac{N(A')N(B)}{N} \\ 8.4.\ D_{22} = N(A'\Lambda B') - E(A'\Lambda B') = N(A'\Lambda B') - \dfrac{N(A')N(B')}{N} \end{cases}$$

Donde D simboliza las diferencias entre frecuencias observadas y esperadas, y los subíndices permiten identificar cada casilla de la tabla. D_{11} denota la diferencia entre las frecuencias de la casilla correspondiente a las primeras línea y columna; D_{12} a la discrepancia en la casilla de la primera línea y la segunda columna, y así sucesivamente.

Los dos caminos para obtener un indicador del grado de asociación entre variables dicotómicas (el análisis de porcentajes y la comparación de frecuencias observadas y esperadas) se definen por contraste con la independencia estadística. Mientras más grande es la discrepancia mayor es la fuerza de la relación y viceversa. Definir la asociación por oposición a la independencia produce resultados claramente interpretables sólo en caso de que no haya relación entre X y Y. Esta aseveración es doble por cuanto sostiene, por una parte, que si se lleva a cabo un análisis de porcentajes entonces se tendrá que $D = 0$; en cambio, si se sigue la vía de las frecuencias esperadas se tendrá:

$$D_{11} = D_{12} = D_{21} = D_{22} = 0$$

Por otra parte, afirma que la lejanía respecto de la independencia no es suficiente para establecer el grado de aproximación con la asociación perfecta. Como no se ha definido conceptualmente lo que se entenderá por asociación perfecta es, lógicamente, imposible interpretar el valor de D de acuerdo con esta perspectiva. A primera vista pareciera haber una respuesta inmediata, por cuanto sabemos que el rango de variación de D va de −100 a 100 para los casos de relación inversa y directa máximas, respectivamente.

Los valores límite de D establecen una posibilidad estadística que no necesariamente se encuentra ligada con cualquier proposición empírica.[19] Ésta puede ser de tal naturaleza que si los datos se comportan de acuerdo con la teoría, entonces D no debe alcanzar ninguno de los

[19]Recuérdese que la proposición empírica define el tipo de asociación. Cuando la tabla observada coincide con ella, la asociación es por definición la máxima posible.

valores extremos. Puede darse el caso de que no exista una correspondencia entre las bondades estadísticas de la medida y los requerimientos que surgen desde el campo de la teoría. La inadecuación entre el abanico de posibilidades que entrega el conocimiento técnico (usualmente como respuesta a problemas planteados dentro de su propio ámbito) y las necesidades que derivan del proceso de investigación social, delimita un campo problemático de carácter general que limitaremos al caso particular del análisis de asociación. En adelante nos preocuparemos por denifir el grado de asociación entre las variables en función de las conexiones lógicas que unen las proposiciones empíricas con la distribución de los datos. Con este propósito retomemos la discusión sobre la marginalidad.

En el caso de la teoría de la marginalidad cultural habíamos concluido que no estaba formulada con la precisión suficiente para permitirnos distinguir entre:

P_1: B ⌒→ A
B′ ⌒→ A′

y

P_2: B ⌒→ A
B′ ⌒→ (A o A′)

La primera proposición (P_1) sostiene que si sabemos que una observación tiene el atributo B en la variable X, entonces deberá presentar el A en la variable Y, y si es B′ en X deberá ser A′ en Y. Si las variables se relacionaron de acuerdo con P_1, entonces la distribución bidimensional de frecuencias deberá ser tal que todas las observaciones que comparten el atributo B deberán ser A y que todas aquellas que presenten la característica A también deberán ser B. Por tanto, todas las observaciones deberán ubicarse en la diagonal principal, como se observa en la distribución hipotética del cuadro 7.

CUADRO 7

Distribución bidimensional de frecuencias (hipotética) para el caso en que P_1 es válida

Y \ X	*B*	*B′*	*Total*
A	50 (100%)	0	50
A′	0	150 (100%)	150
Total	50	150	200

Al aplicar (6) a los datos de este cuadro, tenemos que:

$$D = 100 - 0 = 100$$

Si la proposición empírica hubiese sido la inversa y los casos se hubieran distribuido de acuerdo con ella, D sería igual a:

$$D = 0 - 100 = -100$$

Estos resultados nos llevan a la conclusión de que si seguimos el camino de las diferencias porcentuales para examinar la consistencia entre la proposición empírica (del tipo P_1) y la distribución de las observaciones, entonces D podrá asumir los valores situados dentro del rango de variación limitado por ±100. Ya sabíamos que cuando D asume el valor nulo las variables son estadísticamente independientes.

En el caso de que la segunda proposición (P_2) sea válida se derivan las siguientes consecuencias en la distribución de las observaciones: *i*) si un caso es B en X, entonces deberá presentar A en Y y *ii*) si es B′ en X entonces puede ser A o A′ en Y, sin que esto contradiga la teoría que originó la proposición empírica. Toda observación que tenga el atributo B en X, deberá tener también el atributo A en Y, pero no todo A será necesariamente B. Para que esta proposición sea corroborada por los datos no se requiere que las observaciones se desplieguen sólo por la diagonal principal, sino que también aparezcan en la casilla de la primera línea y en la segunda columna.

CUADRO 8

Distribución de frecuencias (hipotética) para el caso en que P_2 es válida

Y \ *X*	*B*	*B′*	*Total*
A	50 (100%)	30 (60%)	80
A′	0	20 (40%)	20
Total	50	50	100

A pesar de que la distribución de frecuencias de las variables X y Y se ajusta a la proposición P_2, la diferencia porcentual dista bastante de alcanzar el valor máximo:

$$D = 100 - 60 = 40$$

Este ejemplo muestra que aun cuando la proposición empírica prediga adecuadamente la distribución de las observaciones, el valor de D se puede encontrar muy lejos del valor máximo posible. Es fácil ver que, en general, cuando la proposición teórica da lugar a una proposición empírica del tipo de P_2, el valor de D que corresponde a un ajuste perfecto con la distribución de los casos no necesariamente alcanzará el valor máximo. En general, a cada distribución bidimensional de frecuencias corresponderá un D máximo posible diferente, lo que implica que para emitir un juicio sobre el grado de relación entre dos variables será necesario confrontarlo con el valor específico de D. De este modo se podrá calibrar el valor particular que asumió la discrepancia.

La indeterminación de los límites de D, bajo las condiciones impuestas por P_2, es una consecuencia de su definición en términos de lejanía o proximidad respecto de la independencia estadística. El sentido que tenga un valor de D dependerá de que sea posible definir conceptualmente un valor límite a la asociación, de manera que se pueda juzgar en cada caso concreto el grado de lejanía (cercanía) respecto de la independencia y simultáneamente de cercanía (lejanía) a la asociación, tal como se definió. Ahora bien, en este trabajo entenderemos por asociación el grado de concordancia entre una proposición empírica y su correspondiente distribución observada de frecuencias de X y Y.

Como conclusión parcial podríamos decir que cuando el análisis se efectúa recurriendo a la comparación de las diferencias de porcentajes (es decir cuando se calcula D) es posible formarse una idea del grado de asociación entre las variables dicotómicas X y Y si la proposición empírica es del tipo P_1. En este caso D puede asumir todos los valores entre ±100 con sus correspondientes interpretaciones sistematizadas en (7). Pero si la hipótesis es del tipo P_2 la asignación de contenido a un D específico no será tan simple por cuanto sus límites variarán para cada distribución de frecuencias observada. Se podrá tener una idea de la lejanía respecto de la independencia pero no de la cercanía respecto de la asociación, a menos que se proceda a calcular en cada caso el valor máximo (mínimo si la relación es inversa) de D.

El otro camino para medir el grado de asociación entre las variables X y Y (comparar las frecuencias observadas con las esperadas) surge, al igual que el análisis de porcentajes, por contraste con la independencia estadística: el valor observado de cada casilla de la tabla se compara con el que se obtendría (esperado) si las variables fuesen independientes. Es decir, esta estrategia analítica tampoco se basa en una idea clara de la asociación que permita originar límites de un intervalo cuyos extremos correspondan a una idea bien definida de asociación entre las variables. En la próxima sección nos dedicaremos a examinar la relación

entre las proposiciones empíricas y la distribución de frecuencias conjunta; así que no abundaremos en mayores detalles.

Nos limitaremos por el momento a destacar algunas propiedades estadísticas de las discrepancias entre las frecuencias observadas y esperadas.

En primer lugar, hay que llamar la atención sobre el hecho de que en el cuadro 6 las diferencias entre las frecuencias observadas y esperadas son iguales en valor absoluto (92) en las cuatro casillas. Esto no es casual; corresponde a una propiedad fácil de entender en las tablas de dos por dos. Como las distribuciones marginales por líneas y columnas son las mismas para ambas frecuencias, entonces el valor de una discrepancia en una casilla necesariamente originará otra de la misma magnitud, pero de signo contrario. Por tanto, para saber cuánto se desvía la distribución de frecuencias observada de la distribución generada bajo el supuesto de independencia, bastará con calcular la discrepancia de una sola casilla; lo habitual es que se elija la de la primera línea y la primera columna (casilla 1,1).

También hay que notar que el valor que asume la diferencia entre la frecuencia observada y la esperada en la casilla (1,1), es diferente en los cuadros 7 y 8 a pesar de que en ambos casos la distribución de frecuencias es consistente con la proposición empírica. Para las observaciones del cuadro 7 tenemos:

$$D_{11} = 50 - \frac{50 \times 50}{200} = 37.5$$

y para las del cuadro 8:

$$D_{11} = 50 - \frac{50 \times 80}{100} = 10$$

Si D midiese la asociación como el grado de correspondencia entre la distribución que esperaríamos en virtud de los planteamientos teóricos y las frecuencias efectivamente observadas entonces ambos valores de D_{11} aunque numéricamente diferentes deberían corresponder a valores máximos. Esto nos lleva a concluir que de nuevo estamos en presencia de una escala cuyos valores extremos no están definidos, de modo que no podemos calibrar un valor particular de D_{11} obtenido en una tabla cualquiera. Sabemos que $D_{11} = 0$ significa que las variables son estadísticamente independientes y que en la medida en que nos alejemos del valor nulo aumentará el grado de asociación, sea cual sea el significado que se dé a esta palabra.

En adelante el propósito central de nuestros desarrollos será dar sentido al concepto de asociación.

2.3.1. El determinante como medida de asociación

El determinante de una tabla de dos por dos presenta un interés limitado desde el punto de vista de su aplicación. Su importancia se sitúa en el dominio teórico de la estadística, ya que es el puente para vincular los coeficientes de asociación de uso habitual con la medida de la fuerza de la relación, que descansa en las diferencias entre las frecuencias observadas y esperadas. En las secciones siguientes se estudiará la relación funcional entre el determinante y los coeficientes Q de Yule y Φ. En este apartado nos dedicaremos a estudiarlo como medida de asociación y a establecer su vínculo con D_{11}.

Es relativamente simple demostrar[20] que:

$$(9) \qquad ND_{11} = N(A\Lambda B)N(A'\Lambda B') - N(A'\Lambda B)N(A\Lambda B')$$

Si llamamos Δ al miembro de la derecha:

$$(10) \qquad \Delta = N(A\Lambda B)N(A'\Lambda B') - N(A'\Lambda B)N(A\Lambda B')$$

Donde Δ simboliza al determinante de la tabla de dos por dos y la ecuación (9) muestra que la discrepancia entre las frecuencias observada y esperada de la casilla (1,1) multiplicada por el total de observaciones (N) es igual al valor de Δ.

El primer componente del lado derecho de la ecuación (10) es el producto del número de casos en las casillas (1,1) y (2,2) y se puede interpretar como la densidad, peso o importancia de la diagonal principal. Tam-

[20]De acuerdo con la ecuación (8):

$$D_{11} = N(A\Lambda B) - E(A\Lambda B) = N(A\Lambda B) - \frac{N(A)N(B)}{N}$$

$$ND_{11} = N[N(A\Lambda B)] - N(A)N(B)$$

Al escribir el total N y las frecuencias marginales en términos de las frecuencias conjuntas:

$$ND_{11} = [N(A\Lambda B) + N(A'\Lambda B) + N(A\Lambda B') + N(A'\Lambda B')]\,[N(A\Lambda B)] - [N(A\Lambda B)+N(A\Lambda B')]\,[N(A\Lambda B) + N(A'\Lambda B)]$$

Realizando los productos:

$$ND_{11} = [N(A\Lambda B)]^2 + N(A'\Lambda B)N(A\Lambda B) + N(A\Lambda B')N(A\Lambda B) + N(A'\Lambda B')N(A\Lambda B) - N(A\Lambda B)^2 - N(A\Lambda B')N(A\Lambda B) - N(A\Lambda B)\,N(A'\Lambda B) - N(A\Lambda B')\,N(A'\Lambda B)$$

y al cancelar términos llegamos a:

$$ND_{11} = N(A'\Lambda B')N(A\Lambda B) - N(A'\Lambda B)N(A\Lambda B') = \Delta$$

bién puede pensarse como la resultante del grado de asociación directa entre las variables ya que mientras mayor sea la fuerza de esta relación, mayor será la cantidad de observaciones ubicadas en la diagonal principal. Al contrario, mientras mayor sea el peso de las casillas (1,1) y (2,2) mayor será la fuerza de la relación positiva.

El segundo producto que compone la igualdad (10) admite una conceptualización similar. Muestra la densidad, peso o importancia de la diagonal secundaria y por lo tanto se puede interpretar como indicador del grado de relación inversa entre X y Y. El número de observaciones en la diagonal secundaria es una resultante de la fuerza de la asociación negativa.

De acuerdo con esta manera de conceptualizar los componentes de (10), Δ es un índice de asociación que compara (por diferencia) los pesos de las diagonales de la tabla o, en otros términos, que mide el grado de relación en una distribución bidimensional de frecuencias, como la resultante de la asociación positiva menos la negativa.

Ya sabemos que si las variables X y Y son estadísticamente independientes, entonces $D_{11} = 0$. Al sustituir este valor en las igualdades (9) y (10), se concluye que Δ también será nulo.

Examinemos los valores límite que puede alcanzar el determinante. Si $N(A'\Lambda B) = 0$, o $N(A\Lambda B') = 0$; o $N(A'\Lambda B) = N(A\Lambda B') = 0$ entonces tenemos que:

$$\Delta_{máx} = N(A\Lambda B)N(A'\Lambda B')$$

y si:

$$N(A\Lambda B) = 0;\ o\ N(A'\Lambda B') = 0;\ o\ N(A\Lambda B) = N(A'\Lambda B') = 0$$

entonces se tiene:

$$\Delta_{mín} = -N(A'\Lambda B)N(A\Lambda B')$$

Por tanto, el rango de variación para el determinante será:

$$\Delta_{mín} \leq \Delta \leq \Delta_{máx}.$$

Un valor alto de Δ significará que la relación directa entre las variables es fuerte; un valor pequeño, estrecha asociación inversa. Para calibrar un valor de Δ correspondiente a una tabla específica se deberá calcular cada vez su máximo y su mínimo.

Es importante notar que este coeficiente es sensible a transformacio-

nes proporcionales. Si se amplifican[21] todas las frecuencias de la tabla por una constante K mayor que la unidad, entonces:

$$\Delta_2 = K^2\, N(A\Lambda B)N(A'\Lambda B') - K^2\, N(A'\Lambda B)N(A\Lambda B')$$

$$\Delta_2 = K^2[N(A\Lambda B)N(A'\Lambda B') - N(A'\Lambda B)N(A\Lambda B')]$$

$$\Delta_2 = K^2\Delta_1$$

donde Δ_1 y Δ_2 simbolizan, respectivamente, el determinante antes y después de aplicar K. La amplificación de todas las frecuencias de la tabla por la misma constante no modifica el grado de relación entre las variables; sin embargo, el *valor de este coeficiente aumenta* con el cuadrado de la constante. Por supuesto que $\Delta_{máx}$ y $\Delta_{mín}$ también varían con el cuadro de la constante, por lo que la posición absoluta del Δ calculado, se altera, no así su ubicación relativa. Por tanto, al aplicar el determinante a dos distribuciones de frecuencias, la comparación de sus valores sólo tendrá sentido si previamente se los localiza dentro de sus correspondientes intervalos de variación.

Para evitar la sensibilidad de Δ a las transformaciones proporcionales se puede recurrir a normalizar las distribuciones pasando todas las frecuencias a porcentajes respecto al total. Esto equivale a definir un nuevo determinante (Δ') relacionado con el anterior (Δ) a través de:

$$(11) \qquad \Delta' = \frac{\Delta}{N^2} = \frac{N(A\Lambda B)}{N}\,\frac{N(A'\Lambda B')}{N} - \frac{N(A'\Lambda B)}{N}\,\frac{N(A\Lambda B')}{N}$$

cuyos límites son:

$$(12) \qquad \Delta'_{máx} = \frac{N(A\Lambda B)}{N} \cdot \frac{N(A'\Lambda B')}{N}$$

y

$$(13) \qquad \Delta'_{mín} = -\frac{N(A'\Lambda B)}{N} \cdot \frac{N(A\Lambda B')}{N}$$

Al aplicar una constante de proporcionalidad K a la distribución conjunta, tanto los numeradores como los denominadores de (11), (12) y (13) se verán afectados por K^2 y por consiguiente no se modificará el valor de Δ'.

[21] El argumento también es válido si se aplica una reducción.

A pesar de que hemos controlado el efecto de las transformaciones proporcionales, los límites del determinante siguen siendo variables de una distribución a otra, lo que implica que la comparación directa de los Δ' carece de sentido. En las secciones siguientes presentaremos dos coeficientes de asociación que corresponden a diferentes maneras de resolver el problema de los límites del determinante.

Antes de concluir este apartado queremos recordar que tanto Δ como Δ' por ser funciones de D_{11} nos proveen de una definición de asociación sólo como desvío con respecto a la independencia estadística, lo que no nos permite disponer de una interpretación razonable para juzgar el grado de relación entre X y Y.

2.3.2. Coeficiente Fi (Φ)

Este coeficiente cierra el recorrido del determinante al intervalo ± 1, normalizándolo por la raíz cuadrada del producto de las frecuencias marginales:

(14)

$$\Phi = \frac{\Delta}{\sqrt{N(A)N(A')N(B)N(B')}} = \frac{N(A\wedge B)N(A'\wedge B')-N(A\wedge B')N(A'\wedge B)}{\sqrt{N(A)N(A')N(B)N(B')}}$$

Cuando X y Y son estadísticamente independientes $\Delta = 0$ y $\Phi = 0$. Si la distribución de frecuencias es tal que todas las observaciones se ubican en la diagonal principal tenemos:

TABLA 6

Distribución bidimensional de frecuencias que origina un Φ máximo

$Y \backslash X$	B	B'	*Total*
A	$N(A\wedge B)$	0	$N(A)$
A'	0	$N(A'\wedge B')$	$N(A')$
Total	$N(B)$	$N(B')$	N

de manera que se cumple que: $N(A\wedge B) = N(A) = N(B)$

$N(A'\wedge B') = N(A') = N(B')$ y $N(A\wedge B') = N(A'\wedge B) = 0$

Si se remplazan estas igualdades en la ecuación (14) tenemos:

$$\Phi = \frac{N(A)N(A') - 0}{\sqrt{N(A)N(A')N(A)N(A')}} = 1$$

De donde se concluye que cuando las frecuencias de la diagonal secundaria son nulas, el coeficiente Φ alcanza su valor máximo.

Supongamos ahora que las frecuencias de la diagonal principal son nulas:

TABLA 7

Distribución bidimensional de frecuencias que origina un Φ mínimo

$Y \setminus X$	B	B'	*Total*
A	0	$N(A \Lambda B')$	N(A)
A′	$N(A' \Lambda B)$	0	N(A′)
Total	N(B)	N(B′)	N

Las frecuencias de esta tabla satisfacen las siguientes relaciones:

$$N(A' \Lambda B) = N(B) = N(A'); \; N(A \Lambda B') = N(A) = N(B')$$

y

$$N(A \Lambda B) = N(A' \Lambda B') = 0$$

Al sustituir estos valores en (14) se tiene:

$$\Phi = \frac{0 - N(A)N(A')}{\sqrt{N(A)N(A')N(A)N(A')}} = -1$$

En consecuencia Φ asume su valor más pequeño cuando los casos se ubican sobre la diagonal secundaria y en la principal sólo hay valores nulos.

Tenemos entonces que:

$$-1 \leq \Phi \leq 1$$

donde el valor 1 indica que entre X y Y hay máxima asociación directa, el −1 que hay máxima relación inversa y cuando es igual a 0 que no hay asociación.[22]

¿Pero qué significa que hay máxima relación directa entre X y Y? Lo único que sabemos es que se trata del valor más alejado de la independencia estadística, pero no se sabe muy bien a qué se aproxima por

[22]Para evitar complejidades matemáticas en la determinación de los límites del coeficiente Φ se optó por el camino de la comprobación en lugar de la demostración rigurosa. Se seguirá el mismo procedimiento cuando se exponga el coeficiente Q.

cuanto no se ha definido conceptualmente el sentido de los límites del coeficiente.

Con base en la distribución de frecuencias de la tabla 6, se puede decir que el coeficiente Φ mide, en una escala que va de 0 a 1, el grado de aproximación entre la proposición empírica:

P_1: B ⟶ A
B′ ⟶ A′

y la distribución efectiva de los datos. Valores de Φ positivos y cercanos a cero indican no sólo independencia estadística sino también lejanía respecto a P_1. Los valores cercanos a 1 indican alejamiento de la independencia, pero también un alto grado de adecuación entre P_1 y la distribución bidimensional de los datos.

Un examen similar, pero esta vez aplicado a la tabla 7, nos permite afirmar que los valores de Φ entre 0 y -1 indican el grado de adecuación entre los datos y la proposición empírica:

P_1: B ⟶ A′
B′ ⟶ A

De manera análoga al caso de la relación directa, los valores de Φ se ubican no sólo respecto al punto de independencia estadística ($\Phi = 0$) sino también en función del máximo grado de adecuación entre P_1 y la distribución bidimensional de frecuencias ($\Phi = -1$).

En conclusión tenemos que:

$\Phi =$
- 1 Si las observaciones se distribuyen según: P_1: B ⟶ A; B′ ⟶ A′
- 0 Si X y Y son estadísticamente independientes
- -1 Si los casos se distribuyen de acuerdo con: P_1: B ⟶ A′; B′ ⟶ A

2.3.3. Educación y estatus socioeconómico: una aplicación de Φ

Para ilustrar el uso del coeficiente Φ tomaremos un ejemplo desarrollado por Raymond Boudon, quien dice:

> Consideremos el siguiente problema. En la perspectiva generalmente aceptada de que las sociedades industriales son ampliamente meritocráticas, el nivel educativo alcanzado es uno de los principales determinantes del estatus. A partir de esto, es grande la tentación de concluir que el logro educativo debería ser también el principal factor de la movilidad social. En otras palabras, si el logro educativo es realmente un poderoso determinante del estatus, la probabilidad de que un individuo tenga mayor estatus que su padre debería ser mayor, en la medida que su nivel educativo es mayor. Inversamente, la probabilidad de que un individuo caiga a un estatus más bajo que el de su padre debería ser mayor, en tanto menor sea su nivel educativo.[23]

Boudon mide el logro educacional a través de la variable logro educacional del hijo con relación al del padre (X), También el estatus socioeconómico vincula al del hijo en relación con el del padre (Y). Es evidente que ambas variables son tricotómicas, pero para los propósitos del ejemplo las hemos dicotomizado en:

$$X = \begin{cases} B & \text{Mayor nivel educacional del hijo en relación con el del padre} \\ B' & \text{Igual o menor nivel educacional del hijo en relación con el del padre} \end{cases}$$

$$Y = \begin{cases} A & \text{Mayor estatus socioeconómico del hijo con relación al del padre} \\ A' & \text{Igual o menor estatus socioeconómico del hijo con relación al del padre} \end{cases}$$

Ahora bien, del planteo de Boudon se puede deducir que la proposición empírica que se deriva del párrafo citado es del tipo:

$$B \longrightarrow A$$
$$B' \longrightarrow A'$$

Es decir, en las sociedades industriales modernas si los hijos alcanzan un nivel educativo mayor que el del padre, entonces su estatus socioeconómico también será mayor. Si, por el contrario, logran un grado de estudio, menor o igual, dicho estatus debería ser igual o menor que el del padre.

El cuadro que nos servirá para examinar el grado de concordancia entre la proposición empírica y la distribución de los datos se tomó de un estudio realizado por Centers en Estados Unidos:[24]

[23]Raymond Boudon, *Education, Oportunity & Social Inequality: Changing prospects in western society*, John Wiley & Sons, Nueva York, 1974, p. 3.

[24]Robert Centers, "Education and Occupational mobility", *American Sociological Review*, 14, 1949, pp. 143-144. Cita tomada de R. Boudon, *op. cit.*, p. 3.

CUADRO 9

Estatus socioeconómico del hijo con respecto al del padre como función de su nivel educativo en relación con el del padre*

	Nivel educacional del hijo en relación con el del padre		
Y \ X	*Mayor*	*Igual o menor*	*Total*
Estatus socioeconómico del hijo en relación con el del padre — Mayor	134	30	164
Igual o menor	157	95	252
Total	291	125	416

*En este cuadro presentamos las dos variables dicotomizadas, pero en el original están tricotomizadas. Se optó por formar una sola categoría con el igual y el menor porque de esa manera se respeta cabalmente la estructura de los datos de la tabla de Centers.

Iniciaremos el análisis con el examen de las distribuciones marginales para luego establecer el grado de relación entre X y Y. La variable X muestra el resultado de la expansión del sistema educativo en la sociedad industrial del siglo XX: la escolaridad ha aumentado continuamente, lo que conduce a que una alta proporción (70%) de los hijos tenga un nivel de educación mayor que el de sus padres. Un fenómeno totalmente opuesto ocurre con el estatus socioeconómico; pareciera que el desarrollo de los Estados Unidos proporciona o pone a disposición de la sociedad una cantidad decreciente de ubicaciones sociales mejores para los hijos que para los padres (39.4%). Tenemos entonces que el proceso de desarrollo educativo capacita (supuestamente) para ingresar a las mejores posiciones sociales, pero éstas son cada vez más escasas. El cuadro 9 nos muestra que hay 291 hijos con mayor nivel educativo que el de sus padres, pero sólo 164 puestos sociales que se ubican por encima del que desempeñan o han desempeñado los padres.

El análisis de marginales nos alerta en el sentido de que no puede haber una correspondencia perfecta entre la proposición empírica y la distribución de los datos al no cumplirse el requisito aritmético de que $N(A) = N(B)$ y $N(A') = N(B')$.

El cálculo del coeficiente Φ da el siguiente resultado:

$$\Phi = \frac{134 \times 95 - 157 \times 30}{\sqrt{164 \times 252 \times 125 \times 291}} = \frac{12\,730 - 4\,710}{38\,772.49} = \frac{8\,020}{38\,772.49}$$

$$\Phi = 0.21$$

Valor que denota un bajo grado de relación que lleva a concluir que el logro educativo no parece ser, al nivel de la sociedad global, el principal factor de movilidad social.

Para terminar con este ejemplo sería importante destacar que la conclusión que se obtuvo no invalida la meritocracia. En efecto, aun cuando la educación sea la palanca con la que cada individuo alcanza un mayor estatus socioeconómico, la restricción que impone el volumen de estatus mayores para los hijos hace que algunos que han obtenido un mayor nivel educativo sólo tengan acceso a posiciones sociales menores o iguales que las de sus padres. Este resultado nos muestra que a pesar de que existe una relación en lo individual entre educación y estatus socioeconómico, ésta no se manifiesta al nivel de lo social.[25]

2.3.4. Coeficiente Q de Yule

Este coeficiente de asociación aplicable a tablas de dos por dos proviene de otra opción para cerrar el recorrido del determinante al intervalo ±1:

$$(15) \qquad Q = \frac{\Delta}{N(A\wedge B)N(A'\wedge B') + N(A\wedge B')N(A'\wedge B)} =$$

$$= \frac{N(A\wedge B)N(A'\wedge B') - N(A'\wedge B)N(A\wedge B')}{N(A\wedge B)N(A'\wedge B') + N(A'\wedge B)N(A\wedge B')}$$

Sabemos que si X y Y son estadísticamente independientes entonces $\Delta = 0$ y por tanto $Q = 0$.

Para que Q alcance el valor máximo basta con que una frecuencia de la diagonal secundaria sea nula:

TABLA 8

Distribuciones bidimensionales de frecuencias que originan un Q máximo

$Y \backslash X$	B	B'	*Total*
A	$N(A\wedge B)$	0	$N(A)$
A′	$N(A'\wedge B)$	$N(A'\wedge B')$	$N(A')$
Total	$N(B)$	$N(B')$	N

(a)

$Y \backslash X$	B	B'	*Total*
A	$N(A\wedge B)$	$N(A\wedge B')$	$N(A)$
A′	0	$N(A'\wedge B')$	$N(A')$
Total	$N(B)$	$N(B')$	N

(b)

En la tabla 8(a) se cumple que $N(A\wedge B) = N(A)$, $N(A\wedge B') = 0$; y $N(A'\wedge B') = N(B')$. Sustituyendo en (15) obtenemos:

$$Q = \frac{N(A)\,N(B') - 0}{N(A)\,N(B') + 0} = 1$$

[25]Boudon analiza ampliamente esta situación en su libro, *Efectos perversos y el orden social*, Premia Editora de Libros, México, 1980.

y en la tabla 8(b) se tiene que $N(A \wedge B) = N(B)$, $N(A' \wedge B) = 0$; y $N(A' \wedge B') = N(A')$. Al remplazar estas igualdades en (15) se llega a:

$$Q = \frac{N(B)\,N(A') - 0}{N(B)\,N(A') + 0} = 1$$

Este índice de asociación alcanza su valor máximo si $N(A \wedge B') = 0$ o $N(A' \wedge B) = 0$. Es evidente que también será igual a uno si $N(A \wedge B') = N(A' \wedge B) = 0$.

El coeficiente Q de Yule asume su valor mínimo cuando una frecuencia de la diagonal principal es nula:

TABLA 9

Distribuciones bidimensionales de frecuencias que originan un Q mínimo

$Y \backslash X$	B	B'	*Total*
A	0	$N(A \wedge B')$	N(A)
A′	$N(A' \wedge B)$	$N(A' \wedge B')$	N(A′)
Total	N(B)	N(B′)	N

(a)

$Y \backslash X$	B	B'	*Total*
A	$N(A \wedge B)$	$N(A \wedge B')$	N(A)
A′	$N(A' \wedge B)$	0	N(A′)
Total	N(B)	N(B′)	N

(b)

En la tabla 9(a) se cumple que $N(A \wedge B') = N(A)$; $N(A \wedge B) = 0$, y que $N(A' \wedge B) = N(B)$. Al sustituir estas igualdades en (15) se llega a:

$$Q = \frac{0 - N(B)\,N(A)}{0 + N(B)\,N(A)} = -1$$

De la tabla 9(b) se desprende que $N(A' \wedge B) = N(A')$; $N(A' \wedge B') = 0$; y que $N(A \wedge B') = N(B')$. Al sustituir en la ecuación (15) se tiene:

$$Q = \frac{0 - N(A')\,N(B')}{0 + N(A')\,N(B')} = -1$$

Se obtiene el valor mínimo de Q si $N(A \wedge B) = 0$ o $N(A' \wedge B') = 0$. Si ambas frecuencias son nulas entonces también $Q = -1$.

El rango de variación de la medida de asociación Q es:

$$-1 \leqslant Q \leqslant 1$$

El valor 1 se obtiene cuando hay máxima relación directa entre X y Y. El −1 cuando hay máxima relación inversa y es cero en los casos en que X y Y son estadísticamente independientes.

Como Q corresponde a una normalización particular de Δ también mide la asociación como lejanía a la independencia y deja indeterminada la definición.

Las distribuciones de frecuencias de la tabla 8 nos permiten sostener que el coeficiente Q de Yule mide, en una escala limitada por 0 y 1, el grado de aproximación entre la distribución de los datos y las proposiciones empíricas.

P_2: B⟶(A o A′) B′⟶A′ P_2': B⟶A B′⟶(A o A′)

Los valores de Q positivos y cercanos a cero señalan simultáneamente cercanía respecto de la independencia y lejanía respecto de P_2 o P_2'. Los valores cercanos a la unidad nos dicen que hay un buen grado de ajuste entre P_2 o P_2' y la distribución de las observaciones. Hay que destacar que un valor específico del coeficiente de Yule no permite distinguir entre ambas proposiciones empíricas (que normalmente corresponderían a proposiciones teóricas diferentes), por lo que será necesario examinar la conformación de la distribución bidimensional de frecuencias para estar en condiciones de emitir un juicio acerca del grado de bondad de ajuste entre P_2 o P_2' (según sea el caso) y la información estadística.

CUADRO 10

Dos distribuciones bidimensionales (hipotéticas) diferentes que originan un mismo valor de Q

Y \ X	*B*	*B′*	*Total*	*Y \ X*	*B*	*B′*	*Total*
A	50	5	55	A	50	20	70
A′	20	10	30	A′	5	10	15
Total	70	15	85	*Total*	55	30	85
	(a)				(b)		

El cuadro 10 nos muestra dos distribuciones con un mismo valor para Q.

$$Q = \frac{50 \times 10 - 5 \times 20}{50 \times 10 + 5 \times 20} = \frac{400}{600} = 0.667$$

y

$$Q = \frac{50 \times 10 - 20 \times 5}{50 \times 10 + 20 \times 5} = \frac{400}{600} = 0.667$$

Sin embargo, la configuración del cuadro 10(a) tiende a mostrar un mayor grado de concordancia con P_2 que con P'_2. La distribución del cuadro 10(b) parece reflejar un mayor nivel de adecuación con P'_2 que con P_2.

En el caso de que las variables X y Y estén positivamente asociadas tenemos dos proposiciones empíricas cuyo grado de acuerdo con una distribución de frecuencias no se puede determinar sólo a partir del valor de Q por cuanto es incapaz de distinguir entre P_2 *y* P'_2. *Por tanto, para juzgar la bondad de ajuste entre una proposición empírica y una distribución de frecuencias específica hay que considerar la configuración de los casos.*

Un análisis similar al que hemos realizado a partir de la tabla 8, pero esta vez realizado sobre la tabla 9 nos permite sostener que los valores de Q entre −1 y 0, miden el grado de armonía entre las observaciones y alguna de las siguientes proposiciones empíricas:

P''_2: B⟶A′ B′⟶(A o A′)

P'''_2: B⟶(A o A′) B′⟶A

Para distinguir entre P''_2 y P'''_2 a partir de un valor negativo de Q es necesario realizar una inspección de la distribución en las cuatro casillas.

En conclusión tenemos que:

$$Q = \begin{cases} 1 \text{ Si las observaciones se distribuyen según:} \\ \quad P_2: \ B \longrightarrow (A \text{ o } A') \text{ o } P'_2: \ B \longrightarrow A \\ \quad\quad B' \longrightarrow A' \qquad\qquad B' \longrightarrow (A \text{ o } A') \\ 0 \text{ Si X y Y son estadísticamente independientes.} \\ 1 \text{ Si los casos se distribuyen de acuerdo con:} \\ \quad P''_2: \ B \longrightarrow A' \quad \text{o } P'''_2: \ B \longrightarrow (A \text{ o } A') \\ \quad\quad B' \longrightarrow (A \text{ o } A') \qquad B' \longrightarrow A \end{cases}$$

Para distinguir entre las dos proposiciones que corresponden al tramo positivo o negativo del recorrido de Q, es necesario examinar la configuración de los casos en las casillas.

2.3.5. Desarrollo social y movimiento obrero: una interpretación del caso argentino (1915-1945)

Para ilustrar el uso del coeficiente Q de Yule tomaremos como ejemplo una investigación que analiza el carácter de la lucha obrera en función del tipo de desarrollo socioeconómico:

> Los sectores estratégicos o fundamentales para los diferentes modelos de acumulación son los que adquieren mayor peso político en tanto organización sindical. Hay una relación concreta entre los sectores clave del modelo de acumulación y desarrollo impuesto por las clases dominantes, la división internacional del trabajo y los niveles de organización, conciencia, práctica política y poder de dichos sectores sindicales. Ello hace que generalmente sean vanguardia de las luchas sindicales y que, por otra parte, tengan mayor poder de negociación con respecto al aparato de dominación estatal.
>
> Analizamos ahora esta relación en el modelo de acumulación agroexportador implementado desde 1915 hasta 1930 para corroborar la hipótesis.[26]

Un poco más adelante dice:

> Todo el sistema de transporte terrestre se construía en función del proyecto exportador. El ferrocarril, que era el medio de transporte más eficaz del momento, se desarrolló construyendo todas las redes hacia las ciudades puerto, sin conexiones internas.[27]

También plantea que:

> Intentaremos demostrar cómo el sector de transportes que constituía el sector económico estratégico del modelo de acumulación del proceso político económico, es el que se convierte en vanguardia, en términos políticos. O por lo menos es el que posee mayor poder político y de movilización de sus agremiados.
>
> Vamos a corroborar técnicamente que este sector fue el más combativo; por otra parte, que sus luchas no fueron específicamente salariales, sino que en su mayoría se entablaron por motivos de organización y de mejoras en las condiciones de trabajo.[28]

Sostiene que a partir de 1930 se produce una reorientación del modelo de acumulación y que:

> Hay una clara relación entre el proceso de industrialización y el detrimento de la ganancia en el sector transportes. Otro será el sector económicamente

[26] Ana Jaramillo, "Movimiento obrero y acumulación de capital: el caso argentino", en *Revista Mexicana de Ciencias Políticas y Sociales*, núm. 89, julio-septiembre, 1977, p. 181.

[27] *Ibidem*, p. 181.

[28] *Ibidem*, p. 182.

estratégico en el nuevo modelo de acumulación, y el capital británico tratará de vender el sector que hasta ahora había sido la columna vertebral de su dominación.

Este cambio que se produce en la organización y estrategia del capital tiene repercusiones en la organización y estrategia del trabajo.

El sector de vanguardia dejará de ser el sector transportes, dejará de ser el más conflictivo y la industrialización llamará a una nueva distribución ocupacional que trae aparejadas dos consecuencias fundamentales: el proceso de industrialización, con la incorporación masiva de fuerza de trabajo a la industria y, por otra parte, el proceso de urbanización como consecuencia lógica.

Los sectores de sustitución de importaciones en los cuales se basaba la industrialización, pasan a ser los más conflictivos (...) Analizando entonces el periodo 1915-1930 y el posterior 1930-1945, explicaremos los sectores estratégicos de la economía y su relación con el proceso de acumulación y el proceso político sindical. Por otra parte, los analizamos a partir de su forma específica de lucha: el conflicto, la huelga.[29]

La autora sostiene que debe haber una relación entre "hegemonía económica" y "proceso político sindical". El análisis histórico del tipo de desarrollo que ha seguido Argentina le permite concluir que hay un punto de inflexión al inicio de la cuarta década de este siglo, cuando se pasó de un modelo orientado hacia afuera a otro volcado hacia el mercado interno; por tanto, para que la distribución de las observaciones concuerde con la proposición debe haber un cambio en los sectores específicos de donde provienen los obreros que conforman la vanguardia.

El concepto hegemonía económica se objetiva distinguiendo las ramas de la actividad en que se basa el modelo de acumulación de aquellas que son accesorias y dentro de esos grupos elige una de cada tipo.[30] Según este criterio define como variable explicativa (X) las ramas de la actividad productiva y supone que ella sólo puede asumir dos valores: ferrocarriles y vestido. El primero es el sector hegemónico indiscutido en el modelo agroexportador y sobre el segundo descansa el desarrollo del modelo industrializador.

Por otra parte en el texto se sostiene que el proceso político sindical se objetiva a través de la lucha: el conflicto, la huelga. La variable explicada o dependiente (Y), tipo de lucha, la clasifica según su naturaleza en salarial y organizacional. Esta dicotomización lleva implícitos los supuestos de que es posible calificar un conflicto o una huelga en sólo una de ambas categorías y que no hay otras causas que provoquen su estallido.

[29] *Ibidem*, p. 186

[30] El procedimiento es discutible por cuanto no trabaja con una dicotomía.

Tenemos entonces que la variable independiente es "ramas de la actividad productiva", con ferrocarriles y vestido, y que la variable explicada es "conflictos laborales", dicotomizada en salariales y organizacionales.

La proposición teórica dice que los conflictos desarrollados por las organizaciones obreras de las actividades hegemónicas obedecen fundamentalmente a motivos de organización. No hay una declaración explícita acerca del carácter de los conflictos en las actividades no hegemónicas, pero a juzgar por el tratamiento de los datos, el planteo es que en estos sectores se pueden presentar ambos tipos de conflictos sin que ello invalide la proposición.

En función de la hipótesis (proposición teórica) que se maneja y definiendo:

$$X \begin{cases} B & \text{Ferrocarriles} \\ B' & \text{Vestido} \end{cases}$$

$$Y \begin{cases} A & \text{Salario} \\ A' & \text{Organización} \end{cases}$$

entonces la proposición empírica válida para el periodo 1915-1930 será:

$$B \frown\!\!\rightarrow A'$$

y

$$B' \frown\!\!\rightarrow (A \text{ o } A')$$

Según hemos visto en la sección anterior el examen del grado de adecuación entre esta proposición empírica y la distribución de las observaciones se debe realizar a través del coeficiente Q de Yule. En el estudio de referencia encontramos que:

CUADRO 11

Distribución del número de huelgas, según causas (tipo de lucha) por ramas de la actividad productiva (hegemonía sectorial) en Argentina, 1919

		Rama de actividad productiva		
	$Y \backslash X$	*Ferrocarriles*	*Vestido*	*Total*
Causas	Salario	2	53	55
	Organización	28	18	46
	Total	30	71	101

El análisis de las marginales permite afirmar que la distribución de conflictos por rama es altamente favorable al sector no hegemónico (a vestido corresponde un 70%) y la distribución por causas es más o menos equilibrada entre salario y organización (55 y 46 por ciento), respectivamente.

Para examinar la relación entre hegemonía económica y proceso político sindical se estudia el grado y la forma de la asociación entre sus indicadores. Lo primero se obtiene con el cálculo de un coeficiente que nos proporcione una medida de la adecuación entre la proposición empírica y la distribución de las observaciones y lo segundo con el análisis de la conformación interna del cuadro.

La proposición empírica nos dice que la asociación entre X y Y debe medirse a través de Q y su valor debe ser negativo:

$$Q = \frac{2 \times 18 - 53 \times 28}{2 \times 18 + 53 \times 28} = -0.95$$

Los datos y la proposición empírica tienen un alto grado de correspondencia. La relación es inversa y se aproxima bastante al valor límite inferior −1. La conformación de las frecuencias conjuntas también se mueve de acuerdo con la proposición empírica;[31] los conflictos en los ferrocarriles son básicamente de índole organizacional.

Habría sido interesante examinar un cuadro con las mismas variables una vez producido el tránsito a un modelo de acumulación basado en el desarrollo del mercado interno, puesto que en tal caso la relación debería haber cambiado a directa y fuerte. Sin embargo, la carencia de información fue insoslayable, de modo que el estudio debió encaminarse por otros derroteros.

Ejercicios[32]

A continuación presentamos seis problemas tomados de diferentes textos. En varios tuvimos que realizar algunos cambios con el propósito de ajustarlos al caso particular de las tablas de dos por dos, lo que quizá haya desvirtuado los argumentos de los autores. Esto se puede evitar con la lectura de los trabajos originales; para ello incluimos las referencias precisas al final de esta sección.

En los primeros cinco ejercicios hay que intentar establecer la proposición teórica, siempre que sea posible,[33] y su correspondiente proposición empírica y proceder al análisis estadístico adecuado.

[31]No debe olvidarse que a un valor del coeficiente Q de Yule le pueden corresponder dos proposiciones empíricas.

[32]Véase la bibliografía al final de los ejercicios.

[33]En ocasiones los autores dan el mismo nombre al concepto y a la variable estadística, de modo que no resulta fácil distinguir entre los dos tipos de proposiciones.

1. Jorge N. Solomonoff, en su estudio sobre la ideología del movimiento obrero argentino nos dice que:

> En el secular conflicto entre débiles y poderosos, poseedores y desposeídos, los trabajadores siempre soportaron el peso de las fuerzas represivas propias de todo sistema de dominación. Pero la visualización del conjunto de las relaciones sociales como un sistema cuyas partes juegan un papel preciso dentro del todo, implica un proceso de generalización, lo que a su vez requiere de una serie de mediaciones (p. 265).

Después de sostener que el papel ideológico de la religión decreció sistemáticamente durante el siglo XIX, plantea que las masas movilizadas, de las cuales surgía el nuevo proletariado urbano, eran más receptivas que los estratos populares para absorber puntos de vista racionales del mundo.

En seguida, señala que:

> Dentro del campo de la organización obrera, las *élites* dirigentes disponían de mayor grado de libertad para la formulación de definiciones de la situación político-social como marco de orientación teleológica y el establecimiento de metas para la acción. En estos sistemas valorativos y metas propuestas para la acción obrera en la época, distinguimos dos líneas principales: una considera al Estado, a partir de su carácter represivo, esencialmente un elemento conservador del sistema de desigualdades económicas y políticas. De donde resulta que el establecimiento de una sociedad igualitaria requiere como condición la eliminación del aparato estatal. Es decir, la posición anarquista. La propuesta socialista espera en cambio que las clases asalariadas se apoderen del Estado, por vía de un partido que las representa, a fin de utilizarlo como instrumento de sus propias reivindicaciones. Dentro de esta conceptualización se confieren al Estado cualidades que le permitirían actuar como árbitro entre los contrapuestos intereses sectoriales y su utilización por los trabajadores como agente de una distribución equitativa de la riqueza social (pp. 265-266).

A partir de los documentos en que se consignan los acuerdos y las resoluciones de los congresos obreros argentinos celebrados entre 1890 y 1915 intenta capturar:

> el grado en que las definiciones teleológicas propuestas por las distintas élites (dirigentes gremiales) proveían de expresión simbólica y pautas de acción a las aspiraciones de los estratos asalariados (p. 266).

Después de comprobar que en la totalidad de los documentos anarcosindicalistas (15) y socialistas (7) examinados se conceptúa a la sociedad global dividida en clases opuestas, afirma que:

entre la visualización de la sociedad global como una estructura de grupos o de clases en conflicto y la definición del Estado, y por tanto, las expectativas respecto a las posibilidades de su utilización como medio por las distintas fuerzas sociales no existe, como hemos visto, una relación biunívoca ni independiente de los sistemas ideológicos (p. 267).

Ítems	*Total*	*Anarcosindical*	*Socialista*
El Estado como instrumento de dominación de la clase trabajadora	16	15	1
El Estado como mediador en los conflictos entre el capital y el trabajo	6	—	6

2. A partir de una muestra estadística de los estudiantes de Berkeley (1950), S.M. Lipset realizó un estudio que originó el artículo "La formación de opiniones en una situación de crisis". Dentro del tema central —posición ideológica y formación de opiniones—, se interesó por estudiar las vinculaciones entre la ideología de los estudiantes y sus opiniones, favorables o desfavorables, sobre que los profesores "comunistas" impartan clases. De este estudio hemos seleccionado el siguiente cuadro:

	Ideología de los estudiantes		
Política de empleo a profesores comunistas	*Liberales*	*Conservadores*	*Total*
En favor	130	76	206
En contra	82	102	184
Total	212	178	390

3. Ricardo Cinta, en su estudio "Burguesía nacional y desarrollo", define su posición teórico-metodológica en los siguientes términos:

> El análisis integrado del desarrollo busca rescatar, de los conceptos de desarrollo y subdesarrollo, la especificidad estructural que define la relación entre centro-periferia entre los países avanzados y los de desarrollo tardío y las formas políticas de dominación, interna y externa, que posibilitan y condicionan el desenvolvimiento de las áreas atrasadas. El concepto de *dependencia*, que alude directamente a estas condiciones de existencia y funcionamiento del sistema económico y del sistema político, mostrando las vinculaciones entre ambos, tanto en lo que se refiere al plano interno de los países como al externo, se constituye así, en la *categoría analítica* central del marco interpretativo (p. 168).

Después de analizar el proceso de industrialización mexicano, con lo cual comprueba que después de la segunda guerra mundial se plantea

un nuevo tipo de vínculo con el sistema capitalista internacional y, en particular, con la economía estadunidense (p. 174) pasa a estudiar la conformación de la estructura industrial y de las élites empresariales. Plantea que:

> El grado de integración y consolidación de la industria que se fue logrando en el país a partir del proceso sustitutivo condujo, de manera paralela, a la formación, dentro del ámbito de las clases productoras, de ciertas élites empresariales cuyas posiciones dentro del sistema industrial las coloca en una situación dominante. Los límites de dichas élites empresariales en el control del proceso económico pueden establecerse a partir del análisis de su participación dentro de la estructura productiva y de los grados de control que en ella le corresponden (p. 183).

Dentro de esta óptica presenta un cuadro que se refiere a la propiedad nacional o extranjera de las empresas del sector I (productor de bienes de capital) que se ubican entre las 200 más grandes del país.

Distribución de las mayores empresas productoras de bienes de capital (México, 1965)

	Extranjeras	*Nacionales*	*Total*
Primeras 50 más grandes	8	3	11
Desde la 51 a la 200	8	8	16
Total	16	11	27

4. En el artículo "Migración y marginalidad ocupacional en la ciudad de México", Humberto Muñoz, Orlandina de Oliveira y Claudio Stern, analizan las relaciones entre los procesos migratorios y la marginalidad ocupacional en la ciudad de México. Estudian hasta qué punto la condición migratoria es un elemento que condiciona la inserción laboral. Los datos que analizan los obtuvieron de una encuesta por muestreo realizada en el área metropolitana de la ciudad de México de noviembre de 1969 a febrero de 1970. Para mayores detalles sobre la forma de elaboración de las categorías utilizadas, se puede consultar el apéndice I del artículo. A continuación incluimos el cuadro 1 del trabajo citado:

Distribución de la población ocupada y remunerada
de 21 a 60 años de edad por condición de nativo o no nativo,
y por grupos ocupacionales marginales y no marginales (porcentajes)

	Condición	
Grupos ocupacionales	*Nativo*	*No nativo*
Marginales	28.1	15.8
No marginales	71.9	84.2
Total	100.0	100.0

5. En el trabajo de Nilda Sito, "Estructura ocupacional, desarrollo y sindicalismo en los países latinoamericanos", se nos entrega la siguiente distribución bidimensional de frecuencias de 15 países de la región:

	Estructura		
Sindicalismo	*Tradicional*	*Moderna*	*Total*
Bajo	7	0	7
Alto	2	6	8
Total	9	6	15

La autora sostiene que las características sectoriales del patrón de desarrollo en América Latina se captan mejor por medio de la estructura ocupacional que por la participación de cada sector en el producto. El concepto estructura lo objetiva en la distribución sectorial de la población económicamente activa y sobre esta base procede a clasificar a cada país en la categoría tradicional o moderna.

Por otra parte, considera que la organización sindical es el mejor indicador de la existencia de movimientos obreros agrupados en torno a los intereses reivindicativos y se mide como el porcentaje de obreros sindicalizados sobre el total de la población económicamente activa en 1960.

6. En la obra de Emile Durkheim, *El suicidio*, leemos textualmente:

> Suiza es de muy interesante estudio desde este mismo punto de vista, pues como se encuentran en ella poblaciones francesas y alemanas se puede observar separadamente la influencia del culto sobre cada una de las dos razas, y esta influencia es la misma en una y otra. Los cantones católicos producen cuatro o cinco veces menos suicidios que los de los protestantes, cualquiera que sea su nacionalidad.

Cantones franceses		*Cantones alemanes*		*Conjunto de cantones de todas las nacionalidades*	
Católicos	83 suicidios por millón de habitantes	Católicos	87 suicidios	Católicos	86.7 suicidios
				Mixtos	212 suicidios
Protestantes	453 suicidios por millón de habitantes	Protestantes	293 suicidios	Protestantes	326.3 suicidios

La acción del culto es tan poderosa que domina a las demás. Por otra parte, en numerosos casos se ha podido determinar directamente el número de suicidios por millón de habitantes de cada población confesional; he aquí las cifras encontradas por distintos investigadores:

Suicidios en los diferentes países
por un millón de sujetos de cada confesión

	Protestantes	*Católicos*	*Judíos*	*Investigadores*
Austria (1852-1859)	79.5	51.3	20.7	Wagner
Prusia (1849-1855)	159.9	49.6	46.4	Wagner
Prusia (1869-1872)	187	69	96	Morselli
Prusia (1890)	240	100	180	Prinzing
Baden (1852-1862)	139	117	87	Legoyt
Baden (1870-1874)	171	136.7	124	Morselli
Baden (1878-1888)	242	170	210	Prinzing
Baviera (1844-1856)	135.4	49.1	105.9	Morselli
Baviera (1844-1891)	224	94	193	Prinzing
Wurtemburg (1846-1860)	113.5	79.9	65.6	Wagner
Wurtemburg (1873-1876)	190	120	60	Durkheim
Wurtemburg (1881-1890)	170	119	142	Durkheim

Establezca las diferencias entre ambos procedimientos de validación. Además, emita un juicio acerca del grado de adecuación de la información numérica presentada para contrastar, en cada caso, la hipótesis sustentada.

REFERENCIAS BIBLIOGRÁFICAS DE LOS EJERCICIOS

Cinta, Ricardo. "Burguesía nacional y desarrollo", en *El perfil de México en 1980*, vol. 3, Siglo XXI, México, 4a. edición, 1976.

Durkheim, Emile. *El suicidio*, UNAM, México, 1a. ed., 1974.

Lipset, S.M. "La formación de opiniones en una situación de crisis", en Boudon y Lazarsfeld, *Metodología de las Ciencias Sociales*, II: *Análisis empírico de la causalidad*, Editorial Laia, Barcelona, 1974.

Muñoz, Humberto, Orlandina de Oliveira y Claudio Stern. "Migración y marginalidad ocupacional en la ciudad de México", en *El Perfil de México en 1980*, vol. 3, Siglo XXI, México, 4a. ed., 1976.

Sito, Nilda. "Estructura ocupacional, desarrollo y sindicalismo en los países latinoamericanos", en *Revista Latinoamericana de Sociología*, Centro de Investigaciones Sociales, Instituto Torcuato Di Tella, Buenos Aires, 1971.

Solomonoff, Jorge N. *Ideologías del movimiento obrero y conflicto social*, Editorial Proyección, Buenos Aires, 1971.

3. Asociación entre tres variables dicotómicas

En sentido estricto la proposición teórica de la investigación sobre las relaciones entre desarrollo social y movimiento obrero en Argentina, citada en el capítulo anterior implica la consideración simultánea de tres conceptos: hegemonía económica, proceso político sindical y modelo de acumulación. Este último concepto se objetiva por medio de una variable dicotómica cuyo punto de corte se ubica alrededor de 1930, distinguiéndose, de este modo, entre el modelo agroexportador y el industrial.

Desde el punto de vista teórico general, el autor sostiene que la vinculación entre la hegemonía económica y el proceso político sindical es invariable ante el modelo de acumulación y que las actividades económicas que le son centrales variarían según los distintos periodos por los que atraviese.

El problema que tenemos entre manos consiste, entonces, en analizar la relación entre tres variables dicotómicas: a X y Y, ya definidas, se agrega Z, con las categorías C, modelo agroexportador y C′, modelo industrial.

En el ejemplo que usamos para ilustrar el coeficiente Φ vimos que no había relación entre el nivel educativo y el estatus socioeconómico. ¿Se deberá esta carencia de asociación a la acción de un tercer factor, por ejemplo, el estatus ya alcanzado por el padre?

Estos dos casos sirven como ejemplo de un sinnúmero de situaciones en que hay que considerar simultáneamente más de dos variables. En este capítulo veremos las complejidades analíticas que surgen al tratar estas situaciones. En las primeras secciones presentaremos el análisis de covarianzas de Lazarsfeld apegándonos a su propia exposición[1] y en los últimos apartados incluiremos algunos comentarios acerca del

[1]Pensamos que la exposición de Paul Lazarsfeld es de una claridad tal que resulta pedagógicamente insuperable. Por ello hemos optado por seguir sus desarrollos en "La interpretación de las relaciones estadísticas como propiedad de investigación", en R. Boudon y P. Lazarsfeld, *Metodología de las Ciencias Sociales*, vol. II, Editorial Laia, Barcelona, 1974.

uso de esta técnica cuando se incorporan consideraciones sobre las relaciones teóricas y sus correspondientes proposiciones empíricas. En consecuencia, la lógica que ha guiado la exposición experimentará un quiebre momentáneo pero se retomará más adelante. En lo inmediato, esto quiere decir que en lo que sigue no se explicitarán las proposiciones empíricas.

3.1. El análisis de covarianzas de Lazarsfeld[2]

Lazarsfeld empieza su exposición mostrando las relaciones entre la edad y el tipo de audición radiofónica:

CUADRO 12

Proporciones de oyentes de radio según tipos de programas radiales, por categorías de edad

		Edad	
	Y_1, Y_2 y $Y_3 \backslash X$	Jóvenes	Viejos
Tipos de programas	Religiosos	17%	26%
	Tribunas políticas	34%	45%
	Música clásica	30%	29%
	Total de casos	1 000	1 300

Es conveniente notar que el cuadro 12 constituye una síntesis de tres cuadros, uno de los cuales es el siguiente:

CUADRO 13

Distribución de los oyentes de radio según la audición de programas religiosos, por categorías de edad

		Edad		
	$Y_1 \backslash X$	*Jóvenes*	*Viejos*	*Total*
Escuchan programas religiosos	Sí	170	338	508
	No	830	962	1 792
	Total	1 000	1 300	2 300

[2]Debido a que es interesante mostrar las complicaciones que surgen en el estudio de las relaciones cuando se incluye una tercera variable, no desviaremos nuestra atención hacia el análisis de las distribuciones marginales. Nos podemos tomar esta licencia porque se trata de un trabajo de naturaleza eminentemente pedagógica; en una investigación difícilmente podríamos justificar la omisión de las marginales.

Del mismo modo se podría generar otro para edad y tribunas políticas (Y_2) y un tercero para edad y programas de música clásica (Y_3).

Del cuadro 12 se puede derivar que: *i*) habría un cierto grado de relación entre edad y audición de programas religiosos: los viejos tienden a sintonizar este tipo de programas en mayor proporción que los jóvenes (26% y 17% respectivamente). *ii*) Parece que la edad permite explicar las preferencias por los programas políticos. Los viejos muestran una marcada inclinación por este tipo de programas radiales (45%) en contraposición con los jóvenes (34%). *iii*) Por último, la audición de música clásica no varía con la edad; más bien es uniforme: la sintonizan el mismo porcentaje de jóvenes (30%) y de viejos (29%), de manera que se puede concluir que ambas variables son independientes.

Sin embargo, sabemos que en las sociedades industriales ha tenido lugar un proceso de expansión en el sistema educativo lo que debiera reflejarse en que las variables edad y nivel de instrucción estarían relacionadas:

CUADRO 14

Distribución de los oyentes de programas radiales según nivel de instrucción por categorías de edad

		Edad		
	Z \ X	*Jóvenes*	*Viejos*	*Total*
Nivel de instrucción	Superior	600 (60%)	400 (30.8%)	1 000
	Inferior	400	900	1 300
	Total	1 000	1 300	2 300

Este cuadro muestra, como esperábamos, un cierto grado de asociación entre la edad y el nivel de instrucción, que es mayor en los jóvenes que en los viejos. Las relaciones entre la edad y el tipo de programas radiales que se escuchan están afectadas por el nivel de instrucción; por tanto para medir las asociaciones que nos interesan debemos construir cuadros que permitan controlar su efecto. En este tipo de análisis el control se lleva a cabo por medio del principio de "igual presencia" de la tercera variable. Se estudia la relación entre X y Y para un mismo valor de Z; por ejemplo, se analiza la asociación entre la edad y la audición de programas religiosos (Y_1) para todos aquellos que tienen un nivel de instrucción superior (C) y por separado para las personas que poseen un nivel de instrucción inferior (C').

CUADRO 15

Distribución de los oyentes de radio según la audición de programas religiosos, por nivel de instrucción y categoría de edad

		Nivel de instrucción superior			*Nivel de instrucción inferior*		
		Jóvenes	*Viejos*	*Total*	*Jóvenes*	*Viejos*	*Total*
Escuchan programas religiosos	Sí	55 (9.2%)	*45 (11.5%)	100	115 (28.8%)	*285 (31.7%)	400
	No	545	355	900	285	615	900
	Total	600	400 (a)	1 000	400	900 (b)	1 300

*La suma de los viejos con nivel de instrucción superior e inferior que escuchan programas religiosos debería coincidir con el total que se presenta en el cuadro 13 (338). La diferencia de ocho casos se encuentra en el texto de Lazarsfeld y lo hemos transcrito sin modificaciones. Esta anomalía se compensa con los ocho casos de más que aparecen en los viejos que no escuchan programas religiosos. El mismo problema aparece en el cuadro siguiente.

La relación inversa que habíamos detectado entre la edad y la sintonía de programas religiosos prácticamente desaparece al efectuar el control por medio del nivel de instrucción. La subpoblación formada por las personas con un nivel de instrucción superior escucha menos emisiones radiales religiosas que la que posee un nivel inferior. Sin embargo, en los dos casos tenemos una situación próxima a la independencia estadística, lo que origina, en parte, la débil asociación que habíamos comprobado entre X y Y.

Sabemos que hay un cierto grado de relación entre la edad y la audición de tribunas políticas. Veamos qué sucede al incorporar al análisis el nivel de instrucción.

CUADRO 16

Distribución de los oyentes de radio según la audición de programas políticos, nivel de instrucción y categoría de edad*

		Nivel de instrucción superior			*Nivel de instrución inferior*		
		Jóvenes	*Viejos*	*Total*	*Jóvenes*	*Viejos*	*Total*
Escuchan programas políticos	Sí	240 (40%)	*220 (55%)	460	100 (25%)	*360 (40%)	460
	No	360	180	540	300	540	840
	Total	600	400 (a)	1 000	400	900 (b)	1 300

*Véase la nota del cuadro 15.

El cuadro 16 muestra que al controlar por educación se destaca la relación original. Si bien el porcentaje de personas que sintonizan tribu-

nas políticas es más alto en el nivel de instrucción superior que en el inferior, en los dos casos la proporción de viejos que escuchan programas políticos es superior en 15% a la de jóvenes.

Para los programas de música clásica tenemos una situación diferente.

CUADRO 17

Distribución de los oyentes de radio según la audición de programas de música clásica, por nivel de instrucción y categoría de edad

		Nivel de instrucción superior			*Nivel de instrucción inferior*		
		Jóvenes	Viejos	Total	Jóvenes	Viejos	Total
Escuchan programas de música clásica	Sí	192 (32%)	208 (52%)	400	112 (28%)	171 (19%)	283
	No	408	192	600	288	729	1 017
	Total	600	400 (a)	1 000	400	900 (b)	1 300

Del cuadro 12 derivamos que no hay relación entre la edad y la audición de programas de música clásica. Sin embargo, al controlar por la educación vemos que en el nivel de instrucción superior la proporción de viejos que sintonizan ese tipo de emisiones es bastante mayor que la de jóvenes, y que los porcentajes se invierten en el nivel de instrucción inferior. La carencia de asociación entre X y Y_3 se debe, en gran parte, a las relaciones de sentido opuesto que se anulan entre sí.

El ejemplo que hemos desarrollado muestra que la introducción de una tercera variable puede llevar a cualquiera de tres situaciones diferentes:

i) La proporción de viejos que sintonizan programas religiosos es superior a la de jóvenes, pero al controlar por nivel de instrucción la relación desaparece. Se comprueba una asociación entre X y Y_1, que desaparece al controlar por Z.

ii) Es mayor el porcentaje de viejos que de jóvenes que escuchan tribunas políticas. Al introducir la educación aumenta el nivel de asociación. Al examinar la relación entre X y Y_2 se concluye que las variables están ligadas, pero al introducir Z el ligamen se acentúa.

iii) La proporción de viejos y jóvenes que oyen música clásica es la misma. Sin embargo, cuando en el análisis se incluye la instrucción se tiene que en el nivel superior el porcentaje de viejos que sintonizan música clásica es mayor que el de jóvenes y que en el nivel inferior, a éstos corresponde proporción mayor. Teníamos independencia estadística entre X y Y_3, pero al introducir Z aparece una relación inversa en C (nivel de instrucción superior) y una asociación directa en C′ (nivel de instrucción inferior).

En resumen cuando dos variables están relacionadas y se incluye una tercera, el vínculo puede desaparecer, permanecer o hacerse más fuerte. Si las variables son independientes, la inclusión de una tercera puede hacer que aparezca una relación.

El interés de estos tres ejemplos reside más en su forma que en su contenido. Las tres situaciones no son sino casos particulares de una estructura general que se estudiará en la sección siguiente.

3.2. La ecuación de covarianzas

El camino analítico usado en los ejemplos del apartado anterior (estudiar la relación entre dos variables a la luz de una tercera), es bastante frecuente en el proceso de investigación. Esta forma de análisis Lazarsfeld la sistematizó en una ecuación. La fórmula que propuso permite reconstituir la relación entre las variables X y Y cuando se considera Z como tercera variable, dicotomizada en C y C′. Esta última usualmente se denomina variable test o variable control:

$$(16) \qquad (X\ Y) = (X\ Y;\ C) \oplus (X\ Y;\ C') \oplus (X\ Z)\ (Y\ Z)$$

Los términos de esta ecuación[3] tienen los siguientes significados:

(X Y) simboliza la asociación entre las variables explicativa y explicada, es decir, la relación original. Por ejemplo, expresa la relación entre la edad (X) y la audición de programas religiosos (Y_1).

(X Y; C) denota la asociación entre X y Y en la subpoblación definida por C. En el ejemplo se trataría de la relación entre edad (X) y sintonía de programas religiosos (Y_1) en el nivel de instrucción superior (C).

(X Y; C′) expresa la relación entre X y Y en la subpoblación C′. Representaría el grado de asociación entre la edad (X) y la audición de programas religiosos (Y_1) en el nivel de instrucción inferior (C′).

(X Z) simboliza la relación entre las variables explicativa y control. Mostraría la ligazón entre edad (X) y educación (Z) y

(Y Z) denota el vínculo entre la variable test y la explicada. Sería la relación entre la sintonía de programas religiosos (Y_1) y la escolaridad (Z).

[3]Quien se interese por la demostración de esta fórmula, puede consultar Paul Lazarsfeld, "El álgebra de los sistemas dicotómicos", en Boudon y Lazarsfeld, *op. cit.* En sentido estricto (X Y) simboliza el determinante de la tabla cruzada, pero la relación de Δ con los coeficientes de asociación y con las diferencias porcentuales permite generalizar su uso para cualquier medida de asociación que tenga una relación conocida con el determinante.

⊕ Representa una suma cualitativa. Para que la agregación de los componentes de la ecuación de covarianzas arroje exactamente el mismo resultado en ambos miembros de la igualdad, habría que incluir un juego de ponderaciones. Su exclusión no afecta los propósitos usuales del análisis.

Cuando en el análisis consideramos tres variables dicotómicas se generan cinco tablas cruzadas. Tres que dan cuenta de todas las relaciones posibles entre las variables tomadas de dos en dos (X con Y; X con Z; Y con Z) y dos que muestran el vínculo entre X y Y en las subpoblaciones C y C′.

El juego de tablas resultante es el siguiente:

TABLA 10

Distribuciones de frecuencias para estudiar la relación entre X y Y cuando se introduce la variable Z

	(X Y)		
Y \ X	*B*	*B′*	*Total*
A	N(AΛB)	N(AΛB′)	N(A)
A′	N(A′ΛB)	N(A′ΛB′)	N(A′)
Total	N(B)	N(B′)	N

(a)

	(X Z)		
Z \ X	*B*	*B′*	*Total*
C	N(BΛC)	N(B′ΛC)	N(C)
C′	N(BΛC′)	N(B′ΛC′)	N(C′)
Total	N(B)	N(B′)	N

(b)

	(Y Z)		
Y \ Z	*C*	*C′*	*Total*
A	N(AΛC)	N(AΛC′)	N(A)
A′	N(A′ΛC)	N(A′ΛC′)	N(A′)
Total	N(C)	N(C′)	N

(c)

	(X Y; C) C		
Y \ X	*B*	*B′*	*Total*
A	N(AΛBΛC)	N(AΛB′ΛC)	N(AΛC)
A′	N(A′ΛBΛC)	N(A′ΛB′ΛC)	N(A′ΛC)
Total	N(BΛC)	N(B′ΛC)	N(C)

(d)

(X Y; C′) C′

Y \ X	*B*	*B′*	*Total*
A	N(AΛBΛC′)	N(AΛB′ΛC′)	N(AΛC′)
A′	N(A′ΛBΛC′)	N(A′ΛB′ΛC′)	N(A′ΛC′)
Total	N(BΛC′)	N(B′ΛC′)	N(C′)

(e)

La tabla 10(a) es producto del cruce entre las variables originales; 10(b) y 10(c) muestran las distribuciones bidimensionales de frecuencias provenientes del cruce de la tercera variable (Z) con las variables explicativa y explicada. A estas dos últimas distribuciones se les conoce como *marginales* porque sus frecuencias internas se encuentran en los márgenes de las tablas 10(d) y 10(e). Estas dos últimas muestran la distribución

conjunta de X y Y en las subpoblaciones C y C′ y permiten calcular las denominadas *parciales* o *condicionales*.

Al cuantificar la fuerza de la relación en cada tabla se origina un componente de la ecuación (16) que hemos indicado en la parte superior de las tablas.

El uso de esta ecuación puede ayudar a entender algunos resultados paradójicos, como sería el de la relación entre el número de cigüeñas (X) y el de nacimientos (Y), que queda explicada de inmediato al introducir el tipo de comunidad (Z) dicotomizado en rural (C) y urbano (C′). La asociación desaparece en las subpoblaciones rural y urbana, es decir $(X\,Y;\,C) = (X\,Y;\,C') = 0$. Además, las variables explicativa y explicada tienen una relación no nula con la variable control. En efecto, se sabe que en las comunidades rurales hay más nacimientos y más cigüeñas que en las urbanas; por tanto, al existir un ligamen entre el contexto geográfico (Z) y los nacimientos (Y) y entre el primero y las cigüeñas (X), se concluye que las relaciones marginales no son nulas: $(X\,Z) \neq 0$ y $(Y\,Z) \neq 0$. La combinación de asociaciones parciales iguales a cero y marginales diferentes de cero permite explicar la relación original (X Y) por las marginales; por el vínculo de la variable control con X y Y.

La ecuación de Lazarsfeld también puede servir para encontrar explicaciones que enriquezcan la comprensión de algunas asociaciones. Por ejemplo, podríamos desarrollar el análisis de la relación entre la edad y la audición de programas religiosos usando como tercera variable el nivel de instrucción. Para cuantificar la ecuación (16) usaremos el determinante como medida de asociación. Para realizar el cálculo podremos recurrir a aplicar Δ sobre cuadros normalizados, o, lo que es equivalente, a utilizar la fórmula para Δ′ de la ecuación (11) sobre cuadros no normalizados.

Al aplicar Δ′ sobre la información del cuadro 13 tenemos:

$$\Delta'_{xy} = -0.022$$

y del cuadro 15 (a) y (b) se llega a:

$$\Delta'_{xy\cdot c} = -0.005 \qquad \text{y} \qquad \Delta'_{xy\cdot c'} = -0.006$$

El determinante para el cuadro 14 es:

$$\Delta'_{xz} = 0.072$$

Para calcular la relación marginal entre las variables control (Z) y explicada (Y) es necesario que antes se obtenga la distribución conjunta. A partir del cuadro 15 se deriva:

CUADRO 18

Distribución de los oyentes de radio, según la audición de programas religiosos por nivel de instrucción

		Nivel de instrucción		
	Y \ Z	*Superior*	*Inferior*	*Total*
Escuchan programas religiosos	Sí	100	400	500
	No	900	900	1 800
	Total	1 000	1 300	2 300

De este cuadro se obtiene que:

$$\Delta'_{yz} = -0.051$$

Al usar el determinante para medir el grado de relación entre la edad, la sintonización de programas religiosos y la educación, la ecuación (16) toma la siguiente forma particular:

$$-0.022 = -0.005 \oplus -0.006 \oplus (0.072)\,(-0.051)$$

La asociación entre la edad y la audición de programas religiosos se explica, básicamente, por las relaciones marginales (nivel de instrucción con edad y nivel de instrucción con audición de programas religiosos), puesto que al controlar por nivel de instrucción las relaciones parciales tienden a desaparecer.

Desde el punto de vista analítico, de las formas particulares que asume (16) dos revisten especial interés porque ambas definen situaciones extremas contra las cuales conviene comparar cada caso concreto.

Si la tercera variable Z no está relacionada con X, entonces (16) asume la forma:[4]

$$(17) \qquad (X\ Y) = (X\ Y;\ C) \oplus (X\ Y;\ C') \oplus (0)\,(Y\ Z)$$

Cuando en un análisis específico la ecuación de Lazarsfeld toma esta forma entonces la relación entre X y Y se debe a los vínculos que existen entre ambas variables en las subpoblaciones C y C'.

[4]La consideración de la variable Z en el análisis será pertinente si tiene una relación con Y, por lo que no se incluye explícitamente el caso (Y Z) = 0. Si la tercera variable no se vinculase con X ni con Y entonces sería superfluo considerarla. Desde el punto de vista estrictamente matemático en esta ecuación (X Z) = (Y Z) = 0 es equivalente de (Y Z) = 0.

Tenemos una situación enteramente diferente si:

$$(18) \qquad (X\ Y) = 0 \oplus 0 \oplus (X\ Z)\ (Y\ Z)$$

en cuyo caso la asociación entre X y Y se origina en los vínculos de ambas variables con Z. Es decir, al efectuar el control por medio de la tercera variable se desvanece la relación original en cada una de las subpoblaciones. La relación se debe a las asociaciones marginales. Esta forma particular de la ecuación de covarianzas permite construir una explicación vía relaciones parciales para el ejemplo de los nacimientos y las cigüeñas y también para el caso de la audición de programas religiosos.

Tenemos, entonces, que la ecuación de covarianzas de Lazarsfeld puede asumir dos formas polares, una en que se hacen cero las marginales y la relación se debe a las parciales y otra en que se anulan las parciales y el vínculo entre X y Y se debe a las marginales. Es conveniente bautizar uno y otro polos para simplificar la comunicación. Cuando la ecuación asume la forma (17) se acostumbra decir que el análisis se hace por *elaboración de parciales* o condicionales y cuando asume la estructura de (18), que se hace por *elaboración de marginales.*

Conviene destacar que estas son las dos formas ideales que puede asumir la ecuación; por tanto, en situaciones prácticas rara vez aparecen nítidamente. Las buenas aproximaciones al caso puro de un tipo u otro de elaboración se presentan de manera bastante similar a la ecuación que calculamos para el ejemplo de la audición de programas religiosos. Pero en la mayoría de los casos prácticos se presenta una situación intermedia que resulta de una mezcla de elaboración por medio de parciales y marginales.

La ecuación de Lazarsfeld expresa en términos matemáticos una estructura de carácter lógico de validez general: el análisis de la relación entre dos variables a la luz de una tercera. Que la mayoría de las veces se aplique a estudios empíricos no invalida su uso estrictamente lógico en el campo de algunas teorizaciones (de las proposiciones teóricas). En palabras de Lazarsfeld:

> Así pues, hemos reducido a su armazón lógico un procedimiento de investigación familiar. Consiste en analizar la relación entre dos variables mediante la introducción de una tercera variable. Hasta ahora lo hemos aplicado a ejemplos derivados de estudios empíricos, pero sería fácil mostrar que ciertas teorías tienen una estructura lógica idéntica (...) Durkheim sugiere que el suicidio es menos frecuente entre los católicos que entre los protestantes porque la comunidad católica está más cohesionada.[5]

[5]Paul Lazarsfeld, "La interpretación...", *op. cit.*, p. 31.

Además sostiene que la validez de su ecuación rebasa los límites impuestos por el número de variables y por el carácter dicotómico de cada una de ellas:

> Para simplificar hemos restringido el análisis al caso de tres variables. La introducción de variables suplementarias no entraña ninguna modificación en la lógica del razonamiento, como se puede ver fácilmente. Por otra parte, nosotros tenemos en cuenta las variables dicotómicas. La introducción de variables que pudieran tomar un número de valores superior a dos complicaría el argumento sin cambiar la forma del mismo.[6]

Para completar su modelo, agrega una serie de consideraciones respecto a la ubicación temporal de las variables que, combinadas con las formas que puede asumir la ecuación, completan una tipología contra la cual se contrasta cada caso concreto que se analiza.

Sostiene que en cualquier investigación hay que intentar determinar la secuencia temporal de las variables. En ocasiones la identificación de la precedencia cronológica es directa o evidente, como en la relación entre la duración del matrimonio (X) y la armonía conyugal (Y), pero en otras la identificación del orden temporal podría requerir un breve análisis:

> Se ha observado una relación inversa entre la importancia del ingreso y la frecuencia de pertenencia a organizaciones. Lo más verosímil es suponer una prioridad al estatus económico.[7]

En otras ocasiones no es posible determinar la secuencia temporal de las variables, en cuyo caso la validez de la tipología que presentaremos se reduciría a las dos situaciones polares que derivan de la ecuación. La postura de Lazarsfeld respecto a la importancia del tiempo es inequívoca:

> De hecho siempre es teóricamente posible determinar el orden cronológico de las variables; el progreso de la investigación empírica consiste precisamente, entre otras cosas, en determinar la anterioridad entre variables. Evidentemente sería interesante discutir los procedimientos de análisis que pueden utilizarse en el caso en que el orden temporal es dudoso, pero ello nos alejaría de nuestro tema. De ahora en adelante, admitiremos que el orden temporal es conocido.[8]

[6] *Ibidem*, p. 32.
[7] *Ibidem*, p. 32.
[8] *Ibidem*, p. 33.

3.3. La tipología de Lazarsfeld

Al presentar su tipología Lazarsfeld plantea:

> Desde este momento estamos en condiciones de presentar el punto esencial de nuestra argumentación: a partir de dos variables primitivas y de una variable test se pueden efectuar cuatro operaciones fundamentales. El hecho de que estas operaciones se realicen sobre datos empíricos o que correspondan a análisis teóricos no tiene ninguna importancia. Si se analiza la relación entre dos variables a la luz de una tercera, el análisis corresponderá siempre a una de estas cuatro operaciones o a una combinación de las mismas.[9]

Las cuatro operaciones fundamentales que constituyen la tipología surgen de la combinación de las dos formas polares que puede asumir la ecuación de covarianzas y de la ubicación temporal de la tercera variable (Z) con relación a X y Y. Si se supone que X es cronológicamente anterior a Y y se considera que Z sólo puede tomar dos posiciones, (posterior a X y anterior a Y —intermedia—, o anterior a X), entonces tenemos que las dos posiciones de la tercera variable (intermedia o anterior a X), cruzadas con las dos formas límite que puede tomar la ecuación, originan:

TABLA 11

Los tipos puros de elaboración

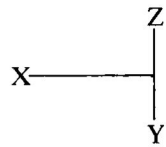
	Forma de la ecuación	
Posición de la variable control	$(X\ Z) = 0$ $(X\ Y;\ Z) \neq 0$	$(X\ Z) \neq 0$ $(X\ Y;\ Z) = 0$
Anterior	P.A.	M.A.
Intermedia	P.I.	M.I.

La casilla identificada con P.A. (parcial anterior) corresponde al caso en que la tercera variable (Z) es anterior a X y en que la ecuación asume la forma representada por la igualdad (17). Como el producto de las marginales es igual a cero, la asociación entre X y Y se debe a la relación (X Y) en las subpoblaciones C y C'. Con una gráfica se puede llegar a una expresión sintética de las condiciones que definen P.A. (parcial anterior).

[9] *Ibidem*, p. 34.

GRÁFICA 1

Relaciones entre las variables en el tipo P.A.



En la gráfica 1 se supone que el tiempo es creciente en la medida en que nos movemos desde la parte superior a la inferior de la página.

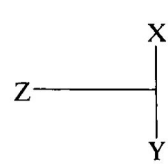
Como ejemplo del tipo P.A., tendríamos el efecto de la propaganda de una película según el grado de instrucción de los espectadores. Podríamos esperar que en la medida en que el nivel de instrucción (Z) sea menor, el efecto (Y) de la propaganda (X), sería mayor: en esta ilustración tenemos que la marginal (X Z) sería igual a cero (carencia de asociación entre instrucción y propaganda) y que la relación parcial en la subpoblación de mayor educación (X Y; C) sería menor que la observada en la subpoblación de menor educación (X Y; C′); por tanto el vínculo entre X y Y se debe a las asociaciones condicionales. En cuanto al orden temporal, tendríamos que la instrucción (Z) es anterior a la propaganda (X). Esta estructura de relación entre las variables es bastante similar a las situaciones en que tenemos una disposición (nivel de instrucción), un estímulo (propaganda) y una respuesta (efecto de la propaganda).

Es usual que una de las asociaciones parciales que dan cuenta del ligamen tentre X y Y sea mayor que la otra y que a partir de este hecho surjan preguntas acerca de los factores que conducen a dicha diferencia y por esta vía se abran nuevas puertas para continuar la investigación.

El tipo P.I. (parcial intermedia) comparte con el P.A., la misma estructura de la ecuación; la diferencia se encuentra en que la tercera variable (Z) se ubica entre X y Y, como se puede apreciar en la gráfica 2.

GRÁFICA 2

Relaciones entre las variables en el tipo P.I.



Para este caso tendríamos como ejemplo:

> Que al estudiar la relación entre el éxito profesional y el ambiente, liberal o autoritario, de la educación se ha podido observar que los jóvenes educados en una atmósfera liberal tienen menos éxito en una situación profesional de estructura autoritaria.[10]

Tenemos que el tipo de educación (X, dicotomizada en liberal y autoritaria) y el éxito profesional (Y) se relacionan en virtud de que los jóvenes educados en una atmósfera liberal tienen menos éxito si el lugar de trabajo es autoritario (X Y; Z) que si es liberal (X Y; Z'). Es decir, ambas asociaciones condicionales son no nulas y diferentes. Se supone que el tipo de educación recibida no se relaciona con la atmósfera de trabajo (X Z) = 0, o, dicho de manera diferente, que los mecanismos de selección de personal no discriminan en función del tipo de ambiente educativo en que se formaron los candidatos a los diferentes puestos. Además, tenemos que la educación (X) es anterior a la variable lugar de trabajo (Z). De esta manera se concluye que el ejemplo es del tipo P.I.

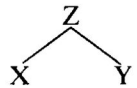
Los tipos de elaboración P.A. y P.I., tienen una relación de equivalencia. Para pasar de uno a otro basta con intercambiar Z con X. En función de esto Lazarsfeld plantea que:

> Tenemos pues de hecho un tipo único de elaboración para el cual el vocablo *especificación* es sin duda el más apropiado. La consideración de la anterioridad temporal entre X y Z introduce dos subtipos de especificación.[11]

El tipo M.A. (marginal anterior) se genera cuando Z es anterior a X y la ecuación de covarianzas asume la forma expresada por (18). El vínculo entre X y Y se origina en las relaciones marginales, es decir, en el producto de (X Z) y (Y Z).

GRÁFICA 3

Relaciones entre las variables en el tipo M.A.



Esta estructura de relación entre las variables permite entender las relaciones espurias, ilusorias o equívocas. La cantidad de cigüeñas (X) se relaciona con el número de nacimientos (Y). Pero al introducir el tipo de localidad (Z) dicotomizado en rural (C) y urbano (C') se tiene

[10] *Ibidem*, p. 35.
[11] *Ibidem*, p. 35.

que (X Y; C) = (X Y;C′) = 0; en consecuencia, la relación se debe a los vínculos que el grado de urbanización sostiene con el número de cigüeñas (X Z) y con el número de nacimientos (Y Z): en el campo hay más cigüeñas que en las ciudades y también más nacimientos. En cuanto a la ubicación temporal de las variables es claro que la urbanización es anterior a X y Y.

Otro ejemplo muy usado en la literatura es el de la asociación entre la importancia de los daños (Y) provocados por un incendio y el número de extintores (X). Al incluir la variable control de la gravedad del incendio (Z), desaparecen las parciales (X Y; C) y (X Y; C′). La relación se debe a las marginales (X Z) y (Y Z). Ahora bien, la gravedad del incendio es una variable anterior a los daños que provoca y a la cantidad de extintores utilizados.

Los dos casos anteriores son claros ejemplos de relaciones ilusorias, espurias o equívocas, que se pueden entender a través de la elaboración M.A. (marginal anterior).

El último tipo, el M.I. (marginal intermedia) resulta de la misma forma que la ecuación que origina M.A., pero la tercera variable se ubica temporalmente entre X y Y.

GRÁFICA 4

Relaciones entre las variables en el tipo M.I.

Considérese como ejemplo la relación entre el estado civil de las mujeres (X) y su tasa de ausentismo (Y). Si se introduce como variable control la importancia de las responsabilidades familiares (Z), entonces se tiene que:

$$(X\ Y;\ Z) = (X\ Y;\ Z') = 0$$

y la relación se debe a los vínculos que tiene la responsabilidad familiar con el ausentismo (Y Z) y con el estado civil (X Z). Además, es claro que en la secuencia temporal la variable Z es intermedia, es posterior al estado civil y anterior al ausentismo.

El lector podría analizar qué ocurre si se introduce como tercera variable la menor resistencia física, medida por la edad. ¿A qué caso corresponderá esta situación? ¿Será del tipo M.I. o M.A.?

3.4. Breves comentarios sobre la aplicación de la tipología

En la investigación es bastante normal el empleo de una tercera variable para analizar la relación entre X y Y. También es frecuente que en los desarrollos sólo se tomen en cuenta la forma de la ecuación, sin considerar el tiempo. Es decir, sólo se considera, por así decirlo, los marginales por columna de la tabla 11 en lugar de juzgar el grado de aproximación de la situación examinada con respecto a los casos puros que presenta la tipología.

Pareciera que la incorporación del tiempo complica bastante las posibilidades analíticas. De hecho el planteamiento de Lazarsfeld incorpora el tiempo en su versión casi física. Decimos casi porque de alguna manera, aunque no explícita, reconoce complejidades que derivan de esta variable:

> Ciertas variables pueden ser utilizadas de varias maneras y, en función del problema examinado, tener posiciones diferentes en la secuencia temporal. Un ejemplo típico es la edad. Las personas de 60 años se distinguen de los jóvenes de 20 no solamente por las posibilidades físicas, sino también porque nacieron y fueron educados en otra época. En la proposición *las personas de edad tienen un nivel de instrucción inferior*, lo importante es evidentemente la época en que estas personas han alcanzado la edad de la escolarización: en este caso la edad es anterior a la educación. Si por el contrario se relaciona la edad de fallecimiento y el clima, la edad evidentemente debe considerarse como posterior.[12]

Se puede usar una misma variable como indicador de conceptos que se refieren a procesos cuyas temporalidades son diferentes. En un caso la edad se utilizaría como indicador del tipo de instrucción y socialización que una persona recibió en la escuela; en el otro sería un índice del proceso de envejecimiento. Por otra parte, se pone como referencia a un proceso social que se encuentra vinculado con otros, por ejemplo la evolución de la distribución del poder entre las fracciones hegemónicas de clase, el estilo de desarrollo socioeconómico que impulsan, las normas y valores dominantes en cada periodo histórico, etc. La edad se ubica en relación con el desarrollo físico de los individuos: el efecto del paso de los años, que también puede incorporar (según sea el enfoque) una serie de procesos sociales, por ejemplo los relacionados con la esperanza de vida al nacer, los que determinan las condiciones de salud pública, etcétera.

Lo que importa destacar es que el tiempo es una variable bastante compleja y que cuando se trata de dar cuenta del cambio social no se

[12] *Ibidem*, p. 33

reduce a su concepción física, sino que aparece de múltiples formas en virtud del movimiento de los diferentes procesos sociales. Éste debe ser, tal vez, uno de los problemas más importantes y difíciles de resolver en las ciencias sociales.[13] Es obvio que su tratamiento sistemático está fuera de nuestro alcance; sólo lo incorporamos con el propósito de relativizar las bondades del modelo que presentamos.

Sabemos que Lazarsfeld sostiene que el progreso de la investigación empírica consiste, entre otras cosas, en determinar la secuencia temporal de las variables. Pero, ¿qué pasa si X y Z se ubican en un mismo punto del tiempo? ¿Qué ocurre si no se puede saber cuál fue antes? ¿Es necesaria tal determinación para el avance del conocimiento? ¿Es posible producir conocimiento válido y de interés científico y político sin considerar el tiempo en la forma en que lo hace Lazarsfeld? Con el propósito de esbozar un inicio de respuesta a este tipo de preguntas retomemos el ejemplo de la relación entre hegemonía económica y carácter de la lucha sindical en Argentina. Habíamos concluido que en los sectores económicos centrales para el modelo de acumulación los conflictos obedecían básicamente a razones de carácter organizativo, y en los no centrales a consideraciones reivindicativas o salariales. Supongamos que se decide incorporar como variable control el nivel de concentración económica sectorial (Z). Se argumenta que la relación entre la hegemonía económica (X) y el carácter de la lucha (Y) no es genuina sino que se origina en los vínculos de la concentración económica sectorial con la hegemonía económica (X Z) y con el tipo de lucha (Y Z). Este argumento sostiene que si se introduce en el análisis un tercer factor, las relaciones parciales se harán nulas y que la asociación (X Y) se debe a las marginales. Es decir, a partir del esquema lógico que ha codificado Lazarsfeld se levanta una explicación alternativa que se podría contrastar.

Como la explicación alternativa a (X Y) envía a una elaboración por marginales, sería interesante saber si se trata de un caso M.A. o M.I., puesto que si se puede establecer que la concentración económica sectorial es anterior a la hegemonía, la relación sería ilusoria. Sin embargo parece bastante factible sostener que en el caso argentino sería imposible saber si Z se ubica antes de X o si cronológicamente se encuentra después de X. Es probable que sean simultáneas, en la medida en que el ferrocarril inglés nace económicamente en pocas manos y que la confección de vestidos surge económicamente desconcentrada. Entonces, según el análisis de covarianzas no podríamos distinguir entre los tipos M.I. y M.A. ¿Qué quiere decir esto? Simplemente que no estamos en condiciones de saber si es o no ilusoria la relación entre la hegemonía

[13]Un buen tratamiento de este problema se puede ver en Fernand Braudel, *La historia y las ciencias sociales*, Alianza Editorial, Madrid, 1968, cap. 3.

económica y el tipo de conflicto, cuestión que se dirimiría ubicando Z antes o después de X.

De alguna manera, aunque no expuesto con mucha claridad en el trabajo de Lazarsfeld, el tiempo se usa para determinar la estructura causal entre las variables. Lo que sí es claro es que sus planteamientos sientan las bases para que posteriormente se exploren una serie de caminos analíticos que se englobaron con el título de análisis de causalidad.[14]

De este modo llegamos, a nuestro juicio, al corazón de las dificultades que presenta la incorporación del tiempo en el análisis de covarianzas. De manera implícita o explícita se trata de determinar la causalidad a partir del análisis de los datos; de este modo se llega a uno de los problemas filosóficos más candentes: ¿se puede encontrar el sentido causal a través del análisis de datos o es imprescindible hacer consideraciones de índole teórica? En otros términos, ¿sólo se puede determinar la causalidad a partir de la teoría?

No es posible dar una salida empírica al ejemplo que hemos manejado en esta sección ya que no podemos ubicar Z antes o después de X. Sin embargo, es posible esgrimir una serie de argumentos teóricos y de antecedentes históricos que ayuden a resolver el impasse entre las dos maneras de explicar la relación entre hegemonía económica y el carácter de la lucha sindical.

La dificultad para incorporar el tiempo en el análisis nos orilló a realizar una serie de consideraciones que surgen tanto de la práctica de la investigación como de la interpretación de los desarrollos posteriores a que dio lugar el planteo de Lazarsfeld.[15] Encontramos que los problemas derivados del tiempo pueden provenir, en última instancia, de un intento por encontrar, a partir de los datos, el sentido causal con el que se relacionan las variables y que por tanto surge de una manera velada un serio problema filosófico: ¿cómo determinar la causalidad?

En esta sección también vimos cómo las diferentes formas que puede asumir la ecuación de covarianzas da pie a interpretaciones teóricas diferentes para explicar la ligazón entre los procesos sociales. Tal fue el caso de la inclusión del grado de concentración económica en el análisis de la relación entre hegemonía económica y lucha sindical. También sostuvimos que la simultaneidad entre X y Z, impedía establecer empíricamente si la relación es ilusoria y planteamos que un camino para

[14]Para examinar la relación entre el análisis de Lazarsfeld y el de causalidad, se puede ver Raymond Boudon, *L'Analyse Mathematique de la causalité*, Plon, París, 1968.

[15]Dentro del análisis causal se sostiene que para inferir causalidad entre dos variables es necesario que: *i*) presenten variación concomitante, *ii*) haya asimetría temporal y *iii*) se eliminen otros factores causales. Un buen tratamiento del tema desde el punto de vista técnico se encuentra en Herbert Asher, *Causal Modeling*, Sage Publications, California, 1976.

buscar una solución a este problema se encuentra en la reflexión teórica y en el análisis histórico.

3.5. El concepto de asociación y el análisis de Lazarsfeld

En el momento en que decidimos exponer la tipología de Lazarsfeld siguiendo lo más fielmente posible sus propios desarrollos se produjo un paréntesis en nuestra exposición. Ahora retomaremos nuestro argumento central. La ecuación de covarianzas surge del campo de la estadística y por ello tiene incorporada la definición de asociación por oposición a la independencia estadística. Sabemos que ésta es una de las limitaciones del análisis de asociación que se puede superar a través de la incorporación de proposiciones empíricas.

Hemos visto en las secciones anteriores que los coeficientes Φ y Q se pueden interpretar como una medida del grado de correspondencia entre las proposiciones empíricas particulares y la distribución de las observaciones.

Si incluimos estas consideraciones en la ecuación (16) tendríamos que en el caso de que las proposiciones empíricas fuesen del tipo:

$$P_1: \quad \begin{array}{l} B \longrightarrow A \\ B' \longrightarrow A' \end{array} \qquad \text{o} \qquad \begin{array}{l} B \longrightarrow A' \\ B' \longrightarrow A \end{array}$$

se debería usar la siguiente expresión de la ecuación de covarianzas:

$$\Phi_{xy} = \Phi_{xy \cdot c} \oplus \Phi_{xy \cdot c'} \oplus \Phi_{xz}\, \Phi_{yz} \tag{19}$$

Si por el contrario las proposiciones empíricas fuesen del tipo:

$$P_2: \quad \begin{array}{l} B \longrightarrow A \\ B' \longrightarrow (A \text{ o } A') \end{array} \qquad \text{o} \qquad \begin{array}{l} B \longrightarrow (A \text{ o } A') \\ B' \longrightarrow A' \end{array}$$

o bien

$$P_2': \quad \begin{array}{l} B \longrightarrow A' \\ B' \longrightarrow (A \text{ o } A') \end{array} \qquad \text{o} \qquad \begin{array}{l} B \longrightarrow (A \text{ o } A') \\ B' \longrightarrow A \end{array}$$

el análisis se debería basar en la siguiente forma de la ecuación:

$$Q_{xy} = Q_{xy \cdot c} \oplus Q_{xy \cdot c'} \oplus Q_{xz}\, Q_{yz} \tag{20}$$

Si en el ejemplo con que se pretende explicar la variación del estatus en función de la movilidad educativa, comparando a hijos con padres, incluimos como variable control el estatus socioeconómico del padre y decidimos realizar el análisis a través de coeficientes de asociación, entonces deberíamos usar la ecuación (19), pero si se tratara de la hegemonía económica y los conflictos laborales considerando el papel de la concentración económica sectorial entonces habría que usar la ecuación (20).

En la próxima sección ilustraremos el uso de estas ecuaciones en una investigación concreta.

3.6. Reforma agraria y migraciones (el caso de Chile)

En un trabajo de investigación realizado en Chile[16] se estudió la migración campo-ciudad en su lugar de origen. Por medio de encuestas realizadas en una comuna se detectó la migración potencial[17] en una zona geográfica específica, de manera que a cada campesino se le adjudicó el carácter de migrante (M) o de no migrante (M′).

El enfoque teórico que se usó para identificar los determinantes de la migración sostiene que es necesario considerar las características estructurales de la producción campesina. Este planteamiento se erigía como alternativa a la explicación de la teoría de la modernización que ponía el acento en los factores que influyen en el nivel de modernización del campesino: los campesinos modernos tienden más a migrar que los no modernos.

Una vez que se identifican tres momentos en que se producen cambios en la organización agraria[18] según los diferentes periodos de dominación política, se individualizan sus correspondientes tipos de explotación agraria:

> Los cambios operados en la estructura de dominación tienen una repercusión inmediata y fundamental en la organización de la estructura agraria chilena; comienzan con el inicio de un proceso profundo de movilización campesina y otro proceso importante de reforma agraria durante el gobierno de la Democracia Cristiana. Estos procesos se agudizan y se radicalizan a partir del triunfo electoral de la Unidad Popular en el año 1970. Lo interesante de estos cambios políticos para nuestro objeto de estudio es que cada uno de ellos ha dejado cristalizados diferentes tipos de organización productiva en la estructura agraria. Así tenemos como forma predominante de ex-

[16]Omar Argüello, "Estructura agraria, participación y migraciones internas", en *Migración y desarrollo*, 3. CLACSO, Buenos Aires, 1974.

[17]*Ibidem*, p. 45.

[18]*Ibidem*, p. 6.

> plotación agrícola hasta el año 1964 los "fundos" que es el tipo de empresa privada con grandes extensiones que contrata fuerza de trabajo permanente y temporal, y que en el caso de los permanentes, paga buena parte del salario en especie, fundamentalmente con la concesión de una pequeña extensión de tierra para el usufructo del trabajador. En el periodo 1964-1970 aparece una nueva forma de organización productiva, que toma el nombre de "asentamiento" y que es un tipo de empresa cooperativa con características de autogestión, la que contrata alguna fuerza de trabajo como trabajadores temporales, pagando sus salarios en dinero efectivo. A partir del periodo que comienza en 1970 aparece otra nueva forma de organización productiva que toma el nombre de "Centro de Reforma Agraria" (CERA) que intenta ser un tipo de empresa en transición a formas socialistas de producción, con apropiación social de parte de sus excedentes.[19]

Una vez que el autor ha establecido la heterogeneidad de la estructura agraria chilena (en la época del estudio) continúa analizando una serie de aspectos relativos a participación, relaciones sociales de producción, etc. Para los propósitos de esta ilustración queremos destacar que:

> Llevando en cuenta argumentaciones teórico-metodológicas pueden explicarse adecuadamente las variaciones encontradas en la descripción que acabamos de hacer del fenómeno migratorio, variaciones que se relacionan sistemáticamente con el tipo de organización social de la producción que ha adoptado la empresa agrícola en la cual se inserta el trabajador agrario. Se desprende de lo anterior que aquellas empresas agrícolas organizadas a partir de cambios en la estructura de dominación, según la concepción ideológica de los nuevos grupos dominantes y que se traduce en nuevas relaciones de producción y de nuevas formas de participación en otras relaciones sociales, son las empresas agrícolas que encierran dentro de sí a menores proporciones de trabajadores con proyectos de migración.[20]

Se plantea, entonces, una relación entre el tipo de organización social de la producción que ha adoptado la empresa agrícola y las migraciones. Se nos dice, específicamente, que son las empresas del área reformada (AR) las que tienden a retener la fuerza de trabajo campesina.

Sabemos que en el momento en que se lleva a cabo el estudio coexisten tres formas productivas; las explotaciones tradicionales denominadas fundos y las que surgieron de reformas agrarias, los asentamientos y los Ceras. Al unir estos dos últimos tipos de empresas en la categoría empresas del área reformada concluimos que el factor "organización social de la producción" se objetiva a través de la variable "tipo de empresa agraria" (X), dicotomizada en fundos (empresas tradicionales) y explotaciones reformadas (AR). Tenemos:

[19] *Ibidem*, p. 7.
[20] *Ibidem*, p. 54.

$$\text{Tipo de empresa: } X \begin{cases} B : \text{Fundos} \\ B' : \text{Asentamientos y Ceras (AR)} \end{cases}$$

No nos detendremos en el proceso de objetivación del concepto "migración potencial";[21] sólo señalaremos que se presenta dicotomizado en migrantes (B) y no migrantes (B′):

$$\text{Migración: } Y \begin{cases} A : \text{migrantes} \\ A' : \text{no migrantes} \end{cases}$$

A continuación debemos determinar la estructura lógica de la proposición teórica. Nos dice que en las empresas tipo AR tenderá a desaparecer la migración, pero ¿qué ocurre en las explotaciones tradicionales? Aun cuando el autor no lo sostiene explícitamente parece que se plantea que la estructura fundo promueve la salida de población. Esta afirmación se puede derivar de las fuertes corrientes migratorias de origen campesino que han afectado a la sociedad chilena.

La relación entre la organización social de la producción y la migración da origen a la siguiente proposición empírica:

$$P_1: \quad B \frown A$$
$$B' \frown A'$$

El modo de organizar la producción de las explotaciones tradicionales o las relaciones sociales que las caracterizan, se muestran incapaces para retener a su fuerza de trabajo y a su población. Por el contrario, las empresas del área reformada presentan una serie de características que conducen a mantener la población agraria en su lugar de origen.

El grado de concordancia entre la proposición empírica y la distribución de los datos se mide a través de Φ aplicado a:

CUADRO 19

Distribución de los trabajadores agrarios según proyecto migratorio, por tipo de empresa agrícola

		Tipo de empresa agrícola		
	$Y \backslash X$	*Fundos*	*AR*	*Total*
Proyecto migratorio	Migrantes	34	8	42
	No migrantes	154	378	532
	Total	188	386	574

$$\Phi_{xy} = 0.29$$

[21] El lector interesado puede consultar: *Ibidem*, sección IV: "Reforma agraria, relaciones sociales y migraciones Internas".

Si bien hay una relación entre el tipo de empresa y la migración potencial, el grado de adecuación de los datos (a juzgar por el valor de Φ) a los planteamientos desplegados es débil.

Una vez establecida la relación se continúa el argumento:

> Podría sostenerse que las argumentaciones teóricas y la información empírica presentadas para la explicación sociológica del fenómeno migratorio son válidas al nivel del contexto productivo con sus diversas formas de organizarse socialmente, pero que esas afirmaciones no han sido probadas al nivel individual. En otras palabras, cabe preguntarse si no estaremos en presencia de una falacia ecológica o de nivel equivocado, que consistiría en atribuir a un nivel de análisis explicaciones que son válidas a un nivel diferente. En este caso significaría preguntarse si dentro de los contextos productivos de mayor participación, y que presentan las más bajas proporciones de trabajadores con proyectos de migración, son también los individuos con mayor participación los que menos piensan en migrar, o si ocurre el caso inverso.[22]

Se trata ahora de analizar el vínculo entre la organización social de la producción y las migraciones a la luz de la participación de los campesinos. Este último concepto se objetiva a través de la participación en las decisiones productivas (Z) que se dicotomiza en participa y no participa:

$$\text{Participación en decisiones productivas: } Z \begin{cases} C & \text{participa} \\ C' & \text{no participa} \end{cases}$$

La incorporación de la tercera variable tiene que ver con la construcción de una proposición teórica alternativa: la migración obedecería a factores de carácter individual y no estructural. En caso de que la explicación a través de la participación se muestre más adecuada, entonces para disminuir la migración campo-ciudad no sería necesario realizar reformas agrarias profundas que alteran las relaciones sociales de producción; bastaría con modificar los niveles de participación del campesinado.

Examinemos lo que acontece al incorporar la participación en las decisiones productivas:

[22] *Ibidem*, p. 54.

CUADRO 20

Distribución de los trabajadores agrarios según proyecto migratorio, por participación en las decisiones productivas y tipo de empresa agrícola

	Participan			*No participan*		
Proyecto migratorio	*Fundos*	*AR*	*Total*	*Fundos*	*AR*	*Total*
Migrantes	0	5	5	34	3	37
No migrantes	2	319	321	152	59	211
Total	2	324	326	186	62	248
		(a)			(b)	

Antes de proceder al análisis hay que aclarar que en las conclusiones que se obtengan se deberá tomar en cuenta que en el cuadro 20 (a) hay una categoría casi inexistente, pues sólo tenemos dos personas que participan en las decisiones productivas en los fundos. Esta limitación en las observaciones, que prácticamente hace constante la variable tipo de empresa en la subpoblación que muestra participación en las decisiones productivas, cuestiona la significación estadística de los hallazgos. Esto no pasó inadvertido al autor, quien en el párrafo siguiente al último que transcribimos nos dice:

> La pregunta metodológica es correcta, sólo que en el caso de nuestro estudio la asociación entre el tipo de organización productiva y la participación de los trabajadores es tan fuerte que dificulta técnicamente no sólo que se encuentre esa falacia de nivel equivocado sino que, incluso, dificulta que se la someta a prueba. De hecho ocurre que en la categoría de participación se concentren casi todos los casos de trabajadores del sector reformado con ausencia casi total de trabajadores de fundos, mientras en las categorías de no participación ocurre la situación inversa.[23]

Hecha la advertencia realicemos un somero análisis de los cuadros condicionales 20 (a) y (b). En la subpoblación conformada por los que no participan en las decisiones productivas se mantiene el mismo tipo de relación que en la tabla original, aunque un poco más debilitado: el coeficiente Φ alcanza el valor 0.16 ($\Phi_{xy;c'} = 0.16$). Para el cuadro 20(a), tenemos que $\Phi_{xy \cdot c} = -0.01$, es decir, habría independencia estadística entre el tipo de empresa y la migración potencial; sin embargo, este resultado no es confiable por el escaso número de observaciones en los fundos en que sí hay participación en las decisiones productivas.

En el cuadro 21 mostramos la información que permite calcular las asociaciones marginales:

[23] *Ibidem*, p. 54.

CUADRO 21

Distribución de los trabajadores agrícolas

a) Según participación en las decisiones productivas, por tipo de empresa agrícola

Z \ X	*Fundos*	*AR*	*Total*
Participan	2	324	326
No participan	186	62	248
Total	188	386	574

b) Según proyecto migratorio, por participación en las decisiones productivas

Y \ Z	*Participan*	*No participan*	*Total*
Migrantes	5	37	42
No migrantes	321	211	532
Total	326	248	574

Los coeficientes de asociación son: $\Phi_{xz} = -0.79$ y $\Phi_{yz} = -0.26$

En consecuencia, la ecuación de covarianzas de Lazarsfeld asume la siguiente forma particular:

$$0.29 = -0.01 \oplus 0.16 \oplus (-0.79)\,(-0.26)$$

La estructura de esta ecuación no corresponde a ninguno de los casos puros de la tipología de Lazarsfeld. La asociación original se conforma tanto por parciales como por marginales. A pesar de los problemas de la muestra (carencia de significación estadística de $\Phi_{xy \cdot c}$) podemos afirmar que al introducir la participación en las decisiones productivas las relaciones parciales son más débiles que la original. Esto llevaría a privilegiar el análisis de las marginales, donde encontramos una fuerte relación entre las variables explicativa y de control.

Sabemos que si vamos a hacer una elaboración por marginales la ubicación temporal de la variable Z es de fundamental importancia para estar en condiciones de detectar si la relación es ilusoria. La participación en las decisiones productivas es posterior a X, porque una de las consecuencias de la reforma agraria consistió en democratizar el manejo de las unidades productivas agrarias. Esta conclusión concuerda con la del autor:

> Los datos muestran lo que habíamos sostenido anteriormente. Por un lado, dentro de los fundos encontramos sólo dos trabajadores que aparecen con participación en las decisiones productivas de la empresa. Por otro lado, dentro de los Asentamientos y de los Cera aquellos trabajadores que participan en esas decisiones productivas, la abrumadora mayoría, son los que muestran una proporción menor de quienes tienen proyecto de migrar. De esta manera probamos que en nuestro estudio no existe una falacia de nivel equivocado cuando afirmamos que las nuevas relaciones sociales constitutivas

de las nuevas formas de organizar la producción llevan a una menor presencia de proyectos de migración.[24]

La relación entre el tipo de organización social de la producción y las migraciones está mediada por la participación de los trabajadores agrícolas. Esto quiere decir que si se llevan a cabo modificaciones en las unidades productivas, su efecto no se hará sentir necesariamente en las migraciones. Para retener a la población rural se deben propiciar mayores niveles de participación.

3.7. Algunas limitaciones del método para estudiar la relación entre más de dos variables dicotómicas

Una de las restricciones que se deja sentir con fuerza cuando se intenta trabajar con varias variables simultáneamente es el número de casos de que se dispone. Las casillas crecen geométricamente con la introducción conjunta de variables adicionales, tal como se ve en la siguiente tabla:

TABLA 12

Relación entre el número de variables dicotómicas cruzadas simultáneamente y el número de casillas

Variables	*Casillas*
2	4
3	8
4	16
5	32
6	64
.	.
.	.
.	.

Como el total de casillas crece geométricamente con razón dos, el cruce simultáneo de variables demanda una cantidad creciente de información. Si combinamos este hecho con las limitaciones presupuestarias para realizar la investigación social, entonces resulta evidente que el volumen de información constituye una severa restricción a los análisis que pretenden considerar distribuciones conjuntas definidas por varias variables.

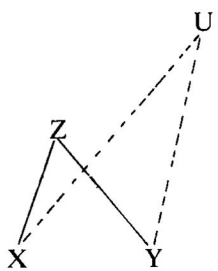
Pero, ¿por qué considerar que el número de casos es una limitación seria? ¿No bastaría con examinar la relación entre X y Y a la luz de

[24] *Ibidem*, p. 55.

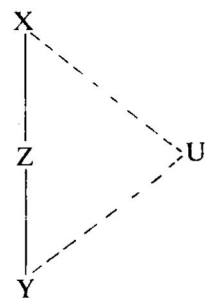
una variable control cada vez? ¿Para qué cruzar simultáneamente varias variables? Si para evitar la restricción que nos impone el número de casos sólo consideramos una variable control cada vez, puede ocurrir que las relaciones parciales no se anulen debido a la acción de una cuarta variable no incluida en el sistema de relaciones:

GRÁFICA 5

Relaciones entre cuatro variables



(a)



(b)

La gráfica 5(a) nos muestra que si la relación entre X y Y es del tipo M.A., es decir, corresponde a una asociación ilusoria y si controlamos por Z, entonces las parciales no se anularán debido a que hay una cuarta variable (U) no controlada que mantiene el vínculo entre X y Y. En la gráfica 5(b) se muestra una situación equivalente pero esta vez referida al tipo marginal intermedia. En los casos representados las parciales se anularían sólo al controlar simultáneamente por ambas variables.

El método que consiste en analizar la relación entre dos variables incluyendo una tercera cada vez, no permitiría detectar asociaciones ilusorias que se explican por dos o más variables test, ni dar cuenta de vínculos genuinos en que participan dos o más variables control. De aquí deriva la importancia de usar simultáneamente varias variables control.

Por último, quisiéramos recordar que hemos señalado que una de las limitaciones más serias para aplicar la tipología de Lazarsfeld radica en las complejidades que derivan de la ubicación de la secuencia temporal de las variables. Pero aun haciendo caso omiso de esta restricción, la utilidad de la tipología se ve menoscabada en tanto las situaciones examinadas escapen a los casos puros que la constituyen.

Ejercicios [25]

1. A continuación se presentan dos tablas de asociación, provenientes de la investigación de Lipset ya aludida en el ejercicio 2 del capítulo 2.

Los profesores comunistas deben enseñar	*Estudiantes que empiezan estudios*		*Estudiantes que terminan estudios*	
	Liberales	*Conservadores*	*Liberales*	*Conservadores*
Sí	74	25	118	16
No	74	38	72	32
Total	148	63	190	48

Entre las variables control que introduce hemos seleccionado el tiempo de permanencia del estudiante en la universidad. Se le pide: *i*) que establezca algunas proposiciones teóricas acerca del porqué se introduce esta variable test, y *ii*) que con la información entregada lleve a cabo un análisis de variables múltiples.

2. Con el propósito de indagar la contribución que haría la diferenciación sexual a la formación de la población económicamente activa, hemos construido, a partir del censo de 1970, la siguiente tabla, referida al área metropolitana de la ciudad de México:

	Sexo	
Población	*Hombres*	*Mujeres*
Económicamente activa	1 548 575	1 456 238
No económicamente activa	2 092 079	2 320 687
Total	3 640 654	3 776 925

Un breve análisis nos conduce a comprobar la existencia de un cierto grado de relación entre las variables. Con el propósito de investigar este extraño resultado, hemos decidido incluir como variable control la procedencia de la población que lo ha originado:

	Migrantes		*No migrantes*	
Población	*Hombres*	*Mujeres*	*Hombres*	*Mujeres*
Económicamente activa	831 755	847 579	716 820	608 659
No económicamente activa	370 773	555 175	1 721 306	1 765 512
Total	1 202 528	1 402 754	2 438 126	2 374 171

[25] Véase la bibliografía al final de los ejercicios.

Con el objetivo de lograr una adecuada comprensión de estos hechos empíricos, se le pide que: *i*) intente una construcción conceptual que le permita interpretar los resultados numéricos. (Sugerencia: a través de la noción de desarrollo social desequilibrado, característico de las formaciones sociales capitalistas, introduzca las peculiaridades del desarrollo capitalista mexicano y el posible aporte de las migraciones en la formación del mercado de trabajo.) *ii*) Destaque las diferencias entre la idea de asociación y la de *nivel* y *iii*) ¿tiene o no sentido, en este caso concreto, aplicar la ecuación de covarianzas de Lazarsfeld?

3. En un trabajo de Américo Saldívar, titulado "Diferencias ideológicas entre obreros y empleados", encontramos la siguiente tabla:

	Empleados		*Obreros*	
	Estrato social		*Estrato social*	
Clase social que debería tener el poder	*Alto*	*Bajo*	*Alto*	*Bajo*
Alta	404	141	178	134
Trabajadora	180	101	81	121
Total	584	242	259	255

Se le pide:

a) Realizar una discusión teórica sobre el posible sentido que podrían tener estas tablas. Para ello le sugerimos que intente movilizar las categorías de "conciencia real" y "conciencia atribuida".

b) Abordar el análisis de la información en función de la teorización realizada, y

c) Presentar ordenadamente las conclusiones a las que haya llegado. Éstas deberían mostrar la articulación entre las categorías conceptuales y la información numérica provista por los cuadros estadísticos. Vale decir, no se trata de hacer una simple descripción de resultados.

4. Silvia Sigal, en su artículo "Crisis y conciencia obrera: la industria azucarera tucumana", se dedica a examinar las consecuencias de la amenaza de desocupación sobre las orientaciones obreras.

La información que proporciona proviene de un conjunto de entrevistas con trabajadores pocos meses después de la intervención de los ingenios en 1966.

Como indicador de la amenaza de desocupación usa la situación del ingenio, ordinalmente tricotomizada en "intervenidos", "amenazados" y "seguros".

Las respuestas a las entrevistas se clasificaron en aquellas que: *i*) apuntaban hacia la solución de problemas inmediatos; *ii*) buscaban resolver la crisis con el mantenimiento de la actividad en la industria azucarera y *iii*) planteaban que los problemas de la provincia debían

resolverse con la diversificación productiva, es decir, con la creación de fuentes opcionales de trabajo.

El cuadro 1 del estudio permite concluir que la propensión hacia la movilidad laboral fuera del azúcar aumenta al pasar de ingenios seguros a intervenidos. Del cuadro 2 se desprende que frente al "problema del azúcar" aparece una clara tendencia hacia la diversificación de la actividad productiva.

De allí la autora concluye que frente a la amenaza de desocupación aumentan las respuestas de cambio individual (movilidad ocupacional) y de cambio global (diversificación económica). Pero entre estos dos tipos de respuesta no hay una transcripción directa puesto que no hay asociación entre aspiraciones de movilidad social y respuestas frente al problema del azúcar.[26]

La explicación de este vínculo se encuentra en la acción ideológica de la actividad sindical, dirigida siempre a la defensa de la industria como una manera de defender el trabajo y al mismo tiempo abogar por la diversificación como una solución de largo plazo.

A continuación presentamos el cuadro 3 de la investigación de referencia; en él se muestra la distribución de las respuestas a la pregunta ¿le gustaría trabajar en algo que no fuera el azúcar? Para los propósitos de este ejercicio hemos fundido en una sola las categorías de ingenios amenazados e intervenidos y omitimos los casos en que no hubo respuesta.

Participación sindical

	Alta			*Baja*		
	Seguros	*No seguros*	*Total*	*Seguros*	*No seguros*	*Total*
Sí	27	59	86	35	43	78
No	31	27	58	27	28	55
Total	58	86	144	62	71	133

Sobre la base de esta información examine la argumentación presentada.

5. En su estudio de las determinantes sociales de la violencia en Colombia, Paul Oquist establece un vínculo entre la polarización y la continuidad de los partidos políticos colombianos y la presencia de violencia. La objetivación del concepto polarización-continuidad se realizó a través de un índice (denominado "de identificación partidaria") que permitió clasificar los municipios en aquellos cuyo porcentaje prome-

[26]Este juicio se basa en información empírica.

dio de votos en las elecciones de 1930 y 1946: *i*) favoreció en un 80% o más a un partido; *ii*) presentó una votación en favor de uno de los partidos entre 60 y 79% y *iii*) una categoría que incluye todos los casos restantes. A la primera clase la denominó "hegemonía" en virtud de que corresponde a una situación en que un solo partido domina toda la escena política. A la segunda, "control" porque corresponde a un caso en que hay cierto equilibrio, aun cuando de todas maneras un partido es depositario de la mayoría, y a la tercera le llamó "competencia" ya que corresponde a una situación de equilibrio.

Una vez que establece el vínculo entre identificación partidaria y violencia, introduce como control el partido político que ejerce la dominación y genera los siguientes cuadros:[27]

Identificación partidaria

	Conservador			*Liberal*		
	Hegemonía	*Control*	*Total*	*Hegemonía*	*Control*	*Total*
Ausencia de violencia	96	81	177	124	76	200
Presencia de violencia	30	23	53	69	58	127
Total	126	104	230	193	134	327

Sobre la base de esta información, ¿qué puede concluir respecto a la violencia en Colombia?

6. De un estudio sobre opiniones de niños mexicanos realizado por Rafael Segovia extrajimos el siguiente párrafo referido a la intervención del Estado en la economía:

> Todo esto depende en gran medida de antecedentes familiares, pero también de la escuela, la cual discrimina desde un punto de vista social porque previamente los niños están discriminados desde un punto de vista económico. De ahí que las tendencias observadas estén aún más marcadas (las actitudes se desarrollan antes) entre los escolares de los establecimientos privados, y sean más reticentes frente a la intervención del Estado en la economía (p. 137).

i) Exprese a través del modelo de covarianzas de Lazarsfeld una relación entre tres variables (acorde con el párrafo); *ii*) indique el papel de cada variable en el modelo; *iii*) identifique a cuál de los cuatro tipos propuestos por Lazarsfeld corresponde el análisis y porqué; *iv*) suponga que las tres variables son dicotómicas y proponga para cada una

[27]Para los propósitos de este ejercicio se omitió la categoría competencia en la identificación partidaria.

las dos categorías de manera que pueda llegar a alguna(s) proposición(es) empírica(s) derivada(s) del párrafo citado.

REFERENCIAS BIBLIOGRÁFICAS DE LOS EJERCICIOS

Lipset, S. M. *op. cit.*

Oquist, Paul. *Violencia, conflicto y política en Colombia*, Instituto de Estudios Colombianos, Bogotá, 1978, cap. II.

Saldívar, Américo. "Diferencias ideológicas entre obreros y empleados", en H. Muñoz, Orlandina de Oliveira y Claudio Stern: *Migración y desigualdad social en la ciudad de México*, El Colegio de México e Instituto de Investigaciones Sociales, UNAM, 1981.

Segovia, Rafael. *La politización del niño mexicano*, El Colegio de México, segunda edición, 1977.

Sigal, Silvia. "Crisis y conciencia obrera: la industria azucarera tucumana", en *Revista Latinoamericana de Sociología 70/1*. Centro de Investigaciones Sociales, Instituto Torcuato di Tella, Buenos Aires, 1970.

4. Asociación en tablas de r × c

En el capítulo anterior presentamos una extensión del análisis de asociación al considerar más de dos variables dicotómicas. En éste nos dedicaremos a exponer las complejidades que surgen al estudiar la distribución conjunta de variables pluricotómicas. Se trata, entonces, de considerar tablas con *r* renglones o líneas y *c* columnas.

TABLA 13

Distribución de frecuencias conjuntas de dos variables pluricotómicas

$Y \setminus X$	B_1	B_2	B_3	$\cdots$	B_c	*Total*
A_1	$N(A_1 \wedge B_1)$	$N(A_1 \wedge B_2)$	$N(A_1 \wedge B_3)$	$\cdots$	$N(A_1 \wedge B_c)$	$N(A_1)$
A_2	$N(A_2 \wedge B_1)$	$N(A_2 \wedge B_2)$	$N(A_2 \wedge B_3)$	$\cdots$	$N(A_2 \wedge B_c)$	$N(A_2)$
A_3	$N(A_3 \wedge B_1)$	$N(A_3 \wedge B_2)$	$N(A_3 \wedge B_3)$	$\cdots$	$N(A_3 \wedge B_c)$	$N(A_3)$
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.
A_r	$N(A_r \wedge B_1)$	$N(A_r \wedge B_2)$	$N(A_r \wedge B_3)$	$\cdots$	$N(A_r \wedge B_c)$	$N(A_r)$
Total	$N(B_1)$	$N(B_2)$	$N(B_3)$	$\cdots$	$N(B_c)$	N

Las variables X y Y tienen *c* y *r* categorías respectivamente. La frecuencia conjunta que corresponde a la casilla genérica formada por A_i y B_j, se simboliza por $N(A_i \wedge B_j)$ y representa el número de casos que presentan simultáneamente ambos atributos. También tenemos dos distribuciones marginales, que nos muestran cómo se distribuyen el total de las observaciones con respecto a la variable X y a la variable Y. La frecuencia del renglón genérico A_i se simboliza por $N(A_i)$ y la que corresponde a la columna B_j por $N(B_j)$; ambas frecuencias denotan el número de casos en cada categoría.

Presentaremos dos grandes líneas de desarrollo. La primera versará sobre aquellas medidas que surgen de la definición de asociación por

oposición a la independencia estadística. Examinaremos las potencialidades y limitaciones del análisis porcentual y la óptica que compara las frecuencias observadas con las esperadas. Veremos el uso de ji cuadrado (χ^2) como indicador de la fuerza de la relación entre variables y una serie de coeficientes que se definen como función de χ^2.

El desarrollo de la segunda línea que constituye esta sección mostrará una serie de coeficientes que derivan de los esfuerzos por definir el concepto de asociación más allá de su simple consideración por oposición a la independencia estadística. Estos coeficientes pretenden medir el grado de correspondencia entre una proposición empírica, teóricamente justificada, y la distribución de las observaciones.

4.1. Asociación e independencia estadística

En primer lugar veremos las bondades del análisis de porcentajes para medir la asociación.

4.1.1. Asociación y análisis porcentual

Al estudiar la asociación en tablas de dos por dos veíamos que si los porcentajes tomados sobre el total marginal por columna eran iguales, entonces las variables dicotómicas X y Y eran estadísticamente independientes y que si la diferencia entre ellos era grande entonces concluíamos que había relación y que ésta era más estrecha en tanto más grande fuera la discrepancia. También planteamos que bastaba con examinar la diferencia entre los porcentajes de la primera línea porque la de la segunda debía ser de igual tamaño aunque de signo opuesto. Estas características (propias y particulares de tablas de dos por dos) permiten sostener que hay un vínculo estrecho entre las diferencias porcentuales y el grado de relación definido por oposición a la independencia estadística. ¿Se seguirá manteniendo este vínculo para tablas pluricotómicas?

Con el propósito de esbozar una respuesta a la pregunta examinemos previamente algunos casos concretos.

CUADRO 22

Distribución de los votos en las elecciones municipales de 1963 en Chile, por condición del elector

	Condición del elector				
Partidos políticos	*Viejos (1)*	*% (2)*	*Nuevos (3)*	*% (4)*	*(4) − (2)*
Conservador	59 297	14.4	41 497	13.1	−1.3
Liberal	43 400	10.7	30 981	9.9	−0.8
D. Cristiano	107 716	26.2	94 301	29.9	+3.7
Radical	87 316	21.2	62 278	19.7	−1.5
Socialista	49 059	11.9	37 642	11.9	—
Comunista	64 309	15.6	48 795	15.5	−0.1
Total	411 097	100.0	315 494	100.0	0.0

El cuadro 22 aporta evidencia empírica para juzgar la idea teórica[1] de que la apertura en la participación electoral genera inestabilidad. Se trata de examinar si los nuevos electores (aquellos que se inscribieron en los registros electorales con posterioridad a la elección parlamentaria de 1961) votan de manera similar o no que los viejos electores. Si los datos mostrasen que la condición de elector y la preferencia electoral son independientes, entonces la apertura del sistema electoral no acarrearía consecuencias de ningún tipo. Si las variables estuviesen relacionadas sería de esperar cambios en la distribución del poder entre partidos debidos a la extensión del sufragio.

Ahora bien, un rápido vistazo a las diferencias porcentuales nos permite *describir* las preferencias partidarias por condición del elector. Vemos que, en general, las diferencias entre los porcentajes son pequeñas; por tanto, la estructura del voto tiende a ser muy similar. Podríamos agregar que hay una leve tendencia a que los nuevos electores prefieran a la Democracia Cristiana. Los partidos de izquierda (Socialista y Comunista) mantienen su base electoral independientemente de la condición del elector, y los de derecha (Conservador y Liberal) así como el Partido Radical, tienden a penetrar menos en la juventud. Todos estos resultados son interesantes e ilustrativos, pero ¿cuál de las diferencias porcentuales nos permite medir el grado de relación? Tenemos seis diferencias, de las cuales una no es independiente de las restantes porque la suma total debe ser igual a cero; entonces, ¿cuál de las cinco o qué combinación de ellas permite construir un índice que mida la fuerza de la relación?

Antes de ver si es posible dar una respuesta a esta inquietud veamos un ejemplo más.

[1] A. Flisfisch y J. Sulbrandt, *Movilización electoral: cambio y permanencia*, FLACSO, Santiago, 1976, p. 31.

CUADRO 23*

Distribución de la PEA, por grupos de ingreso y sector económico de la población masculina de 15 a 64 años de edad, México, 1970

	Sectores económicos						
*Grupos de ingreso mensual***	*Manufactura*	*Construcción*	*Serv. distrib.*	*Serv. produc.*	*Serv. sociales*	*Serv. personales*	*Total*
Menos de 1 152	293	45	140	18	80	133	709
	(26.5)	(29.6)	(26.4)	(11.3)	(21.7)	(43.3)	(27.0)
1 153-1 920	375	67	186	27	119	89	863
	(33.9)	(44.1)	(35.0)	(17.0)	(32.2)	(29.0)	(32.9)
1 921-3 840	291	17	130	52	88	34	612
	(26.3)	(11.2)	(24.5)	(32.7)	(23.8)	(11.1)	(23.3)
3 841 y más	148	23	75	62	82	51	441
	(13.4)	(15.1)	(14.1)	(39.0)	(22.3)	(16.6)	(16.8)
Total	1 107	152	531	159	369	307	2 625
	(42.2)	(5.8)	(20.2)	(6.1)	(14.0)	(11.7)	(100.0)

*Tomado de: Muñoz, De Oliveira y Stern, ***Migración y desigualdad ...***, *op. cit.*
**El salario mínimo mensual era de $847.50 (U.S. $68).

En este cuadro se muestra la distribución de la población económicamente activa por sector económico y tramos de ingreso mensual. Entre paréntesis y al lado de cada frecuencia absoluta agregamos los porcentajes calculados respecto al total de cada columna.

La comparación entre los porcentajes de los distintos sectores económicos nos proporciona valiosa información descriptiva. Sin proponernos realizar un examen exhaustivo del cuadro podemos destacar:

i) que más de una cuarta parte de la PEA que labora en manufactura, construcción y servicios distributivos percibe un ingreso mensual menor que 1 152 pesos mensuales y que en el mismo tramo se encuentra más de 40% de las personas que se desempeñan en servicios personales;

ii) que la situación es la opuesta para el sector servicios productivos, donde más de 70% de los trabajadores se ubican en los dos tramos superiores y

iii) que la distribución tiende a ser uniforme en el sector de servicios sociales.

El análisis de los porcentajes de las distribuciones marginales nos indica que una gran proporción de la PEA se dedica a la manufactura y que junto con los servicios distributivos asciende a más de 60%. La distribución marginal de los grupos de ingreso muestra un sesgo hacia los tramos bajos: casi 60% de la PEA obtiene un ingreso mensual menor que 1 920 pesos.

Como se puede ver, el análisis de porcentajes es un instrumento muy valioso para destacar las características descriptivas de un cuadro de

distribución de frecuencias. Pero, ¿qué podemos decir respecto a la *fuerza* de la relación?

Si quisiéramos formarnos una idea del grado de asociación entre las dos variables del cuadro 23 siguiendo el camino que orienta el análisis en tablas de dos por dos, es decir, por medio de un análisis de diferencias de porcentajes, tendríamos que tomar en cuenta 45 discrepancias; 15 por cada línea, multiplicados por el número de líneas independientes (3).

En tablas de dos por dos hay una relación nítida entre la diferencia porcentual y el grado de proximidad o lejanía respecto a la independencia estadística. De los dos ejemplos que hemos presentado en esta sección se deriva que en las tablas de r × c se rompe el vínculo entre las discrepancias porcentuales y la asociación, lo que inhibe su uso en este sentido. Pero ello no impide que el análisis de los porcentajes sea particularmente útil para remarcar las características más relevantes de una clasificación cruzada.

A continuación examinaremos el otro camino que mide la asociación por oposición a la independencia estadística.

4.1.2. Asociación y diferencias entre frecuencias observadas y esperadas. Índices derivados de ji-cuadrado

En una tabla de dos por dos basta con calcular la discrepancia entre las frecuencias observada y esperada en una casilla (usualmente D_{11}) para obtener una medida de la cercanía o lejanía respecto de la independencia estadística (o fuerza de la relación).

Para calcular el grado de asociación en una tabla de r × c habrá que encontrar una medida resumen de todas las discrepancias entre las frecuencias observadas y sus correspondientes frecuencias esperadas. En efecto, para cada casilla tendremos un valor D_{ij} que simboliza la discrepancia en la casilla genérica formada por la i-ésima línea y la j-ésima columna:

$$D_{ij} = N(A_i \Lambda B_j) - E(A_i \Lambda B_j) \qquad (i = 1, 2, \ldots, r) \text{ y } (j = 1, 2, \ldots, c) \tag{21}$$

donde:

$$E(A_i \Lambda B_j) = \frac{N(A_i)\, N(B_j)}{N} \qquad (i = 1, 2, \ldots, r) \text{ y } (j = 1, 2, \ldots, c) \tag{22}$$

en que $N(A_i \Lambda B_j)$ y $E(A_i \Lambda B_j)$ denotan las frecuencias observada y esperada respectivamente en la casilla (i, j) y $N(A_i)$, $N(B_j)$ sus frecuencias marginales.

La aplicación de la ecuación (22) a los datos del cuadro 23 genera r × c = 4 × 6 = 24 frecuencias esperadas y la sustitución de estos

valores en la ecuación (21) produce igual número de discrepancias D_{ij}. Los cálculos se presentan en el cuadro 24. En cada casilla hay dos números, el primero es D_{ij} y el segundo, que hemos encerrado entre paréntesis, es la frecuencia esperada.

CUADRO 24

Discrepancias (D_{ij}) y frecuencias esperadas $E(A_i \Lambda B_j)$ de la PEA por grupos de ingreso y sector económico de la población masculina de 15 a 64 años de edad, México, 1970

	Sectores económicos						
Grupos de ingr. mensual (pesos)	*Manufactura*	*Construcción*	*Serv. distrib.*	*Serv. produc.*	*Serv. sociales*	*Serv. personales*	*Total*
Menos de 1 152	−6	4	−3	−25	−20	50	0
	(299)	(41)	(143)	(43)	(100)	(83)	(709)
1 153-1 920	11	17	11	−25	−2	−12	0
	(364)	(50)	(175)	(52)	(121)	(101)	(863)
1 921-3 840	33	−18	6	15	2	−38	0
	(258)	(35)	(124)	(37)	(86)	(72)	(612)
3 841 y más	−38	−3	−14	35	20	0	0
	(186)	(26)	(89)	(27)	(62)	(51)	(441)
Total	0	0	0	0	0	0	0
	(1 107)	(152)	(531)	(159)	(369)	(307)	(2 625)

En este cuadro destacan dos resultados cuya validez es general para cualquier tabla de r × c: *i*) las frecuencias esperadas marginales son iguales a sus correspondientes frecuencias observadas. Es decir, $E(A_i) = N(A_i)$ para $i = 1, 2, \ldots, r$ y $E(B_j) = N(B_j)$ para $j = 1, 2, \ldots, c$, y *ii*) las sumas de las discrepancias por líneas o por columnas son iguales a cero.[2] En símbolos:

[2]Sabemos que

$$E(A_i \Lambda B_j) = \frac{N(A_i)N(B_j)}{N}, \text{ luego:}$$

$$E(B_j) = \sum_{i=1}^{r} E(A_i \Lambda B_j) = N(B_j) \sum_{i=1}^{r} N(A_i)/N = N(B_j).$$

De igual modo:

$$E(A_i) = \sum_{j=1}^{c} E(A_i \Lambda B_j) = N(A_i).$$

Por tanto, son idénticas las distribuciones marginales observada y esperada. Por otra parte,

$$\sum_i D_{ij} = \sum_j D_{ij} = 0$$

como consecuencia de la coincidencia entre frecuencias marginales observadas y esperadas.

$$\sum_i D_{ij} = \sum_j D_{ij} = 0.$$

Sabemos que las D_{ij} tienen relación directa con el grado de asociación. Cuando todas son cero las variables son estadísticamente independientes y se alejan de cero en la medida en que aumenta la fuerza de la relación. Podemos usar este hecho con el propósito de proponer una medida resumen que nos indique el grado de asociación.

No sería conveniente proponer la suma de las discrepancias como medida resumen, puesto que, independientemente del grado de asociación, será siempre igual a cero:

$$\sum_j D_{ij} = \sum_i D_{ij} = \sum_i \sum_j D_{ij} = 0$$

Este resultado es producto de que los D_{ij} asumen valores positivos y negativos (ver cuadro 24) cuyas sumas globales son de la misma magnitud en valor absoluto; por ello, para que el vínculo entre D_{ij} y la asociación se manifieste en toda la tabla será necesario recurrir a un procedimiento que haga caso omiso del signo.

Una posibilidad consiste en tomar el cuadrado de la discrepancia por casilla. De este modo tendríamos que para la celda genérica (i, j):

$$(23) \qquad D_{ij}^2 = [N(A_i \cap B_j) - E(A_i \cap B_j)]^2 \qquad (i = 1, 2, \ldots, r) \text{ y } (j = 1, 2, \ldots, c)$$

De esta manera dos casillas que muestran la misma discrepancia pero de signo contrario harán contribuciones de igual magnitud a la formación de la medida de asociación. Por ejemplo, en el cuadro 24 tenemos que:

$$D_{15} = -20 \qquad \text{y} \qquad D_{45} = 20$$

Luego:

$$D_{15}^2 = 400 \qquad \text{y} \qquad D_{45}^2 = 400$$

A partir de las discrepancias por casilla se puede proponer como medida de asociación la suma de sus cuadrados:

$$\sum_i \sum_j D_{ij}^2$$

Pero con este procedimiento asignamos la misma importancia a una discrepancia de igual valor, sin tomar en cuenta el orden de magnitud de las cifras que la conforman. Es lógico suponer que una diferencia de 20 que surge de 100 − 80, debiera tener más peso que una que se origina en 100 020 − 100 000. En el primer caso se compara 20 con 80, mientras que en el segundo con 100 000, lo que hace necesario ubicar un mismo valor en escalas que le adjudiquen un peso distinto. De acuerdo con esta idea los valores
$D_{14} = -25 = 18 - 43$ y $D_{15} = -20 = 80 - 100$,
aunque parecidos, debieran hacer contribuciones diferentes al grado de asociación.

Para asignar un sentido relativo a la discrepancia de cada casilla habrá que ponerla en relación con el orden de magnitud de las cifras que la conforman. Para llevar a cabo esta operación sólo disponemos de las frecuencias observadas y de las frecuencias esperadas. Las primeras no constituyen una opción viable por cuanto es frecuente que algunos N_{ij} sean iguales a cero. Normalmente las frecuencias esperadas no serán nulas pues para que asuman este valor se requiere que por lo menos uno de los totales marginales sea cero (ya sea en la línea, en la columna o en ambas), es decir, que no exista una categoría en cuando menos una de las variables. Si ésta fuese la situación, eliminaríamos la dificultad de que la frecuencia esperada sea igual a cero, redefiniendo de manera conveniente las categorías de la variable. Se concluye entonces que la mejor solución consiste en expresar las discrepancias en términos relativos, tomando como base de comparación las frecuencias esperadas.

Esta operación se aplica a cada D_{ij}^2, definiéndose:

$$(24) \qquad D'_{ij} = \frac{D_{ij}^2}{E(A_i \Lambda B_j)} = \frac{[N(A_i \Lambda B_j) - E(A_i \Lambda B_j)]^2}{E(A_i \Lambda B_j)}$$

$$(i = 1, 2, \ldots, r) \qquad y \qquad (j = 1, 2, \ldots, c)$$

En el cuadro 25 mostramos el resultado de aplicar las ecuaciones (23) y (24) a la información del cuadro 24.

CUADRO 25

Cuadrados de discrepancias (D_{ij}^2) y discrepancias relativas (D'_{ij}) entre las frecuencias observadas y esperadas de la PEA por grupos de ingreso y sector económico de la población masculina de 15 a 64 años de edad, México, 1970

	Sectores económicos						
Grupos de ingreso mensual (pesos)	*Manufactura*	*Construcción*	*Servicios distrib.*	*Servicios produc.*	*Servicios sociales*	*Servicios personales*	*Total*
Menos de 1 152	36	16	9	625	400	2 500	
	(0.120)	(0.390)	(0.063)	(14.535)	(4.000)	(30.120)	(49.228)
1 153-1 920	121	289	121	625	4	144	
	(0.332)	(5.780)	(0.691)	(12.019)	(0.033)	(1.426)	(20.281)
1 921-3 840	1 089	324	36	225	4	1 444	
	(4.221)	(9.257)	(0.290)	(6.081)	(0.047)	(20.056)	(39.952)
3 841 y más	1 444	9	196	1 225	400	0	
	(7.763)	(0.346)	(2.202)	(45.370)	(6.452)	(0.000)	(62.133)
Total	(12.436)	(15.773)	(3.246)	(78.005)	(10.532)	(51.602)	(171.594)

En cada casilla hay dos cifras; la primera es D_{ij}^2 y la segunda, entre paréntesis, es la discrepancia relativa. Nótese que D'_{14} es bastante mayor que D'_{15} y que a pesar de que $D_{15}^2 = D_{45}^2 = 400$ al expresar el cuadrado de la discrepancia en términos relativos se tiene que $D'_{15} < D'_{45}$. Estos dos casos muestran el impacto que se produce en la medida de asociación por casilla, al considerar el orden de magnitud de las frecuencias que generan la discrepancia.

Una vez que eliminamos las dificultades que provienen del signo de la diferencia y que tomamos en cuenta el peso relativo que ésta tiene en cada casilla, se puede proponer como medida de asociación la suma de los D'_{ij} en toda la tabla. La resultante de esta operación se denomina coeficiente de asociación ji-cuadrado (χ^2).[3] Tenemos entonces que:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} D'_{ij}$$

o bien que:

$$(25) \quad \chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{[N(A_i \Lambda B_j) - E(A_i \Lambda B_j)]^2}{E(A_i \Lambda B_j)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{D_{ij}^2}{E(A_i \Lambda B_j)}$$

[3]En este caso usamos ji-cuadrado como una medida del grado de aproximación o lejanía respecto con la independencia estadística. Su distribución de probabilidades tiene muchos usos dentro de la inferencia estadística. Frecuentemente se aplica para saber si una relación es o no significativa.

En el caso de independencia estadística las frecuencias observadas y esperadas coinciden en todas las casillas, por lo que el numerador de (25) se anula y χ^2 valdrá también cero. Cuando las variables no son independientes se tiene que las D_{ij} son diferentes de cero y en la medida en que el grado de asociación es mayor, más grande es su valor, de modo que a mayor lejanía respecto de la independencia estadística mayor será el valor de χ^2.

Con la información del cuadro 25 mostraremos el cálculo de χ^2 para el cruce de las variables sector económico y grupo de ingreso mensual. El valor que resulta de la suma de las cifras entre paréntesis es:

$$\chi^2 = \overbrace{0.120 + 0.390 + 0.063 + 14.535 + 4.000 + 30.120}^{49.228} +$$

$$+ \overbrace{0.332 + 5.780 + 0.691 + 12.019 + 0.033 + 1.426}^{20.281} +$$

$$+ \overbrace{4.221 + 9.257 + 0.290 + 6.081 + 0.047 + 20.056}^{39.952} +$$

$$+ \overbrace{7.763 + 0.346 + 2.202 + 45.370 + 6.452 + 0.000}^{62.133} = 171.594$$

Como este valor se encuentra alejado de cero podríamos decir que las variables están asociadas. Pero para estar en condiciones de juzgar la fuerza de la relación no basta con saber que nos alejamos del valor nulo, sino que hay que tener alguna idea del grado de cercanía con respecto al valor que denota la total asociación entre las variables. En este punto nos encontramos con dos problemas que afectan a χ^2 y que limitan sus bondades para usarlo como índice de asociación:

i) A mayor asociación entre las variables mayor valor de χ^2 pero lo inverso no es necesariamente cierto. No siempre un valor más alto de χ^2 está representando un mayor grado de relación entre las variables. Para ilustrar esta idea supongamos que decidimos arbitrariamente multiplicar por 10 las frecuencias del cuadro 23, de modo que:

$$N'(A_i \Lambda B_j) = 10\ N(A_i \Lambda B_j)\ ;\ N'(A_i) = 10\ N(A_i);$$

$$N'(B_j) = 10\ N(B_j)\ ;\ N' = 10\ N;$$

$$E'(A_i \Lambda B_j) = \frac{N'(A_i)\ N'(B_j)}{N'} = 10\ \frac{N(A_i)N(B_j)}{N}$$

de modo que:

$$E'\ (A_i \Lambda B_j) = 10\ E(A_i \Lambda B_j)$$

De acuerdo con (25):

$$\chi'^2 = \sum\ \sum \frac{[N'(A_i \Lambda B_j) - E'\ (A_i \Lambda B_j)]^2}{E'(A_i \Lambda B_j)}$$

luego:

$$\chi'^2 = \sum\ \sum\ 10\ \frac{[N(A_i \Lambda B_j) - E(A_i \Lambda B_j)]^2}{E(A_i \Lambda B_j)}$$

y

$$\chi'^2 = 10 \sum\ \sum \frac{[N(A_i \Lambda B_j) - E(A_i \Lambda B_j)]^2}{E(A_i \Lambda B_j)} = 10\ \chi^2$$

Al multiplicar por 10 las frecuencias del cuadro, la fuerza de la relación entre las variables permanece inalterada, pero el valor de χ^2 aumentó también al multiplicarse por 10. Este resultado es totalmente general; si modificamos las frecuencias de una tabla por un factor de proporcionalidad k, el nuevo valor de χ^2 será igual al antiguo, multiplicado por esa constante. Es decir,

$$\chi'^2 = k\ \chi^2$$

ii) El otro problema que afecta a este coeficiente es que su valor límite superior depende del número de observaciones y del tamaño de la tabla,[4] lo que constituye un impedimento para comparar la fuerza de la relación en tablas diferentes.

4.1.2.1. El coeficiente de contingencia cuadrado medio (Φ^2)

Este índice de asociación corrige la sensibilidad de χ^2 respecto de los cambios proporcionales. Al definirse como:

[4]El valor máximo de χ^2 es:

$$\chi^2_{máx} = N\ mín\ [(r - 1),\ (c - 1)]$$

Ver, por ejemplo, M.G. Kendall y Alan Stuart, *The Advanced Theory of Statistics*, vol. 2, Charles Griffin, Londres, 1961, p. 557.

(26) $$\Phi^2 = \frac{\chi^2}{N}$$

se logra que su valor no se altere por modificaciones proporcionales en las frecuencias de la tabla. En efecto, si se aplica k a todas las frecuencias, tenemos que:

$$\chi'^2 = k\,\chi^2$$

pero también se cumple que:

$$N' = k\,N.$$

Entonces:

$$\Phi'^2 = \frac{\chi'^2}{N'} = \frac{k\,\chi^2}{k\,N} = \frac{\chi^2}{N} = \Phi^2$$

y el coeficiente de contingencia cuadrado medio no se modifica.

Si las variables son estadísticamente independientes, $\chi^2 = 0$ y por tanto $\Phi^2 = 0$. Pero el problema del límite superior sigue sin resolverse.

Sabemos que para la información del cuadro 23, $\chi^2 = 171.594$ y $N = 2\,625$. Luego, $\Phi^2 = 171.594/2\,625 = 0.065$. Aunque cercano a cero, este valor del coeficiente no nos proporciona una idea de la fuerza de la relación, a menos que procedamos a calcular su valor máximo.

Una vez que se ha resuelto con Φ^2 el problema del efecto sobre χ^2 de transformaciones proporcionales en las frecuencias, queda aún la tarea de encontrar una función de χ^2 cuyos valores extremos sean independientes del número de observaciones y del tamaño de la tabla.

4.1.2.2. Coeficiente de contingencia de Tschuprow

Este coeficiente de asociación responde a la fórmula:

(27) $$T^2 = \frac{\Phi^2}{\sqrt{(r-1)(c-1)}} = \frac{\chi^2}{N\sqrt{(r-1)(c-1)}}$$

Cuando las variables X y Y son estadísticamente independientes, se tiene que:

$$\chi^2 = \Phi^2 = T^2 = 0$$

Si: *i*) la tabla es cuadrada, es decir, r = c, y *ii*) en la línea y la columna que definen cada intersección de la tabla hay sólo una frecuencia no nula, entonces T^2 asume como valor máximo la unidad.

Las dos condiciones establecen que, por ejemplo, para una tabla de r × r (condición de tabla cuadrada) se debe cumplir que:

TABLA 14

Distribución conjunta de las variables X y Y correspondiente a $T^2 = 1$

$Y \setminus X$	B_1	B_2	···	B_r	*Total*
A_1	0	$N(A_1 \Lambda B_2)$	···	0	$N(A_1)$
A_2	$N(A_2 \Lambda B_1)$	0	···	0	$N(A_2)$
·	·	·		·	·
·	·	·		·	·
·	·	·		·	·
A_r	0	0	···	$N(A_r \Lambda B_r)$	$N(A_r)$
Total	$N(B_1)$	$N(B_2)$		$N(B_r)$	N

En todas las líneas y columnas de la tabla hay *sólo una* frecuencia no nula que puede estar localizada en cualquier casilla de la línea o columna correspondiente. Ahora bien, es fácil ver que el valor de χ^2 y por consiguiente el de Φ^2 y T^2 no se alteran si intercambiamos líneas entre sí o columnas entre sí;[5] por tanto cualquier tabla cuadrada que tenga la característica de la 14 se puede transformar en otra en que las frecuencias no nulas se ubiquen sobre la diagonal principal. Esto quiere decir que el coeficiente de Tschuprow alcanzará su valor máximo en el caso en que:

$$N(A_i \Lambda B_j) = N(A_i) = N(B_i)$$

Bajo esta condición se tiene el valor máximo de χ^2 que es $N(r - 1)$[6] y al remplazarlo en (27) se llega a:

$$T^2 = \frac{N(r-1)}{N\sqrt{(r-1)(r-1)}} = \frac{(r-1)}{(r-1)} = 1$$

Como conclusión tenemos que $T^2 = 0$ cuando las variables son independientes y $T^2 = 1$ para cada casilla de la tabla haga sólo una frecuencia no nula, en sus correspondientes líneas y columnas.

[5] χ^2 resulta de la agregación de discrepancias relativas calculadas en cada casilla y los valores de cada D_{ij} son insensibles a intercambios de líneas o de columnas.

[6] Véase la nota 4.

Hay que recalcar que la solución que se obtiene con el coeficiente de Tschuprow para determinar el valor máximo de esta función de χ^2 sólo tiene validez en el caso de tablas cuadradas. A continuación presentaremos un coeficiente que da una solución de carácter general.

4.1.2.3. Coeficiente de asociación de Cramer

Su fórmula es:

$$(28) \qquad C^2 = \frac{\chi^2}{N \text{ mín } (r - 1, c - 1)} = \frac{\Phi^2}{\text{mín } (r - 1, c - 1)}$$

Sabemos que si hay independencia estadística tanto χ^2 como Φ^2 asumen el valor nulo. Cuando en cada línea y en cada columna de la tabla sólo hay una frecuencia conjunta no nula tanto χ^2 como Φ^2 asumen el valor máximo:

$$\chi^2_{\text{máx.}} = N \text{ mín } (r - 1, c - 1)$$

$$\Phi^2_{\text{máx.}} = \text{mín } (r - 1, c - 1)$$

En este caso:

$$C^2_{\text{máx.}} = 1$$

En consecuencia, el coeficiente de Cramer asume el valor cero cuando las variables son estadísticamente independientes y el valor unitario cuando en cada línea y columna de la tabla todas las frecuencias son nulas, excepto una.

Al aplicar la fórmula (28) a los datos del cuadro 25, tenemos que:

$\chi^2 = 171.594$; $N = 2\,625$;

$\text{mín}(r - 1; c - 1) = \text{mín}(4 - 1, 6 - 1) = \text{mín}(3; 5) = 3$ y $\Phi^2 = 0.065$

$$C^2 = \frac{171.594}{2\,625(3)} = \frac{0.065}{3} = 0.022$$

Este valor del coeficiente nos permite concluir que la relación es débil. Se puede afirmar que las variables son estadísticamente independientes porque el 0.022 no sólo se halla muy cerca del valor 0 sino también muy alejado del valor máximo 1.

4.1.2.4. Tres breves comentarios sobre χ^2, Φ^2, T^2 y C^2

Hay que destacar que estos cuatro coeficientes están ligados entre sí por una lógica que obedece a la solución de problemas netamente estadísticos. En efecto, comparten la definición de asociación en términos de grado de aproximación o lejanía respecto de la independencia estadística. El paso de χ^2 a Φ^2 obedece a la necesidad de corregir la sensibilidad que presenta el primero a transformaciones proporcionales de las frecuencias o, si se quiere poner en términos más generales, a eliminar la dependencia que χ^2 tiene respecto del número total de observaciones: el valor máximo que puede alcanzar tiene una relación directa con N. El paso de Φ^2 a T^2 proviene de la búsqueda de un límite superior fijo; de hecho el coeficiente de Tschuprow puede alcanzar el valor máximo 1, siempre que la tabla sea cuadrada, mientras que el valor superior de Φ^2 depende de cada caso particular, ya que

$$\Phi^2_{\text{máx}} = \text{mín}(r - 1, c - 1),$$

lo que inhibe la comparación directa de coeficientes obtenidos en tablas de diferente tamaño. El coeficiente de Cramer propone otra solución al problema de la determinación de un valor máximo; presenta la ventaja de que su recorrido está limitado por los valores 0 y 1, independientemente del número de observaciones y de las dimensiones de la tabla.

Estos coeficientes, por ser función de χ^2, son insensibles a intercambios de líneas entre sí y de columnas entre sí. Esto quiere decir que son indiferentes al orden de las categorías o, en otros términos, que si disponemos del cruce de dos variables ordinales y las desordenamos, el valor de los coeficientes no se altera. Esto significa que χ^2, Φ^2, T^2 y C^2 son coeficientes útiles para medir el grado de relación entre variables nominales[7] y que si se aplican sobre variables ordinales se pierde la jerarquización de las categorías inherente a la medición.

Una vez que se establece que estos cuatro coeficientes sólo recuperan la diferenciación por clases pero no por orden, encontramos un inconveniente al uso de la raíz cuadrada de los coeficientes. Es común que el signo positivo de la raíz se asocie con una relación directa y el negativo con una inversa, pero ¿qué sentido tiene hablar de relación directa o inversa en el cruce de dos variables nominales? Si X y Y son nominales y la disposición de las categorías en la tabla arroja como resultado una

[7]Las ideas básicas respecto a escalas de medición se exponen en Hubert Blalock *Estadística social*, Fondo de Cultura Económica, México, 1978, pp. 26-32. Un tratamiento con mayor profundidad se encuentra en Warren Torgerson, *Theory and methods of scaling*, John Wiley, Nueva York, 1967, cap. 2.

relación directa, podemos transformarla en negativa invirtiendo la disposición de las clases en una de las variables y mantenerla en la otra. Es evidente que no tiene ningún sentido calificar como inversa o directa la asociación entre dos variables nominales.

Antes de pasar a la sección 4.2 donde examinaremos una serie de coeficientes que asignan un sentido concreto a la idea de asociación, expondremos brevemente una función de χ^2 cuya lógica se aparta de las que hemos estudiado.

4.1.2.5. Coeficiente de contingencia de Pearson

Karl Pearson se planteó como problema encontrar un coeficiente aplicable al cruce de dos variables nominales que fuese equivalente al coeficiente de correlación producto momento. La solución a que llegó lo condujo a proponer el coeficiente de contingencia de Pearson:

$$P^2 = \frac{\chi^2}{N + \chi^2} = \frac{\Phi^2}{1 + \Phi^2} \tag{29}$$

Este índice de asociación es equivalente al coeficiente de correlación en el caso límite en que el número de categorías en ambas variables tiende a infinito. Se aproxima al valor uno cuando la tabla es cuadrada y las observaciones se distribuyen de manera tal que en cada renglón y en cada columna hay sólo una frecuencia no nula.

El valor máximo que puede asumir P^2 en una tabla cuadrada específica es:

$$P^2 = \frac{N(r - 1)}{N + N(r - 1)} = \frac{r - 1}{r}$$

En la tabla 15 se muestra el valor máximo que puede asumir P^2 en tablas cuadradas de distinto tamaño.

TABLA 15

Valores máximos que puede asumir el coeficiente de contingencia de Pearson en tablas cuadradas

Tamaño de la tabla r	*Valor máximo de P^2*
2	0.500
3	0.667
4	0.750
5	0.800
6	0.833
7	0.857
8	0.875
9	0.889
10	0.900

El coeficiente P^2 se diferencia de las otras funciones de χ^2 en que se define por aproximación al coeficiente de correlación producto momento de Pearson. Desde el punto de vista netamente estadístico corresponde a una línea argumental diferente a la que origina χ^2, Φ^2, T^2 y C^2.

4.2. Asociación como el grado de correspondencia entre proposiciones empíricas y distribuciones de frecuencias

En la sección 4.1., presentamos los desarrollos vinculados a la definición de asociación como grado de "no independencia". En ésta expondremos la línea estadística cuya preocupación central es la de definirla en función de las hipótesis formuladas por el investigador.

Expondremos en detalle el índice de asociación delta-ro[8] (∇_e) a partir del cual se pueden derivar como casos particulares una serie de coeficientes de uso habitual. Nuestra presentación no será muy profunda, por lo que el lector interesado en los detalles deberá consultar el artículo mencionado en la bibliografía.

A continuación entregamos un ejemplo que nos servirá para exponer ∇_e e ilustrar las distintas formas que puede asumir.

4.2.1. Un ejemplo.[9] Relación entre la situación de clase y el ingreso

Según la CEPAL una de las características del desarrollo socioeconómico de América Latina en la década de los setenta sería:

> El crecimiento económico de América Latina, que en la última década alcanzó una tasa media de 5.5% anual, no ha logrado mejorar las condiciones de vida de los distintos grupos de la población y ha hecho más visible y condenable la existencia de la actual pobreza masiva que alcanza a 40% de la población de la región.[10]

Este párrafo sostiene que el desarrollo de América Latina es concentrador, en el sentido de que los beneficios económicos se reparten discriminadamente entre los miembros de la sociedad. Algunas teorías que explican la desigualdad económica ponen el énfasis en las productividades diferenciales del trabajo y entre el trabajo y el capital; otras destacan factores individuales como el nivel educativo alcanzado; algunas más

[8]Hildebrand *et al., op. cit.*

[9]Se basa en un ejercicio de clase presentado por Jesús Castañeda, Gustavo Lozano, Fernando Macías y Guillermo Macías.

[10]Informe al XVIII periodo de sesiones de la Comisión Económica para América Latina, Bolivia, 1979.

privilegian aspectos organizacionales y de poder, etc. Para los propósitos de este ejemplo supondremos que el análisis del problema ha llevado a sostener que la diferenciación económica se explica, en gran medida, por la inserción en la estructura productiva. Se plantea una relación entre la estructura de clase y el acceso a los logros alcanzados por el desarrollo económico.

Para examinar esta hipótesis es necesario construir una proposición empírica que permita establecer la ligazón con los datos. El concepto inserción en la estructura productiva se objetiva a través de la situación de clase, distinguiéndose cuatro categorías:

Situación de clase: X
- B_1: Trabajadores manuales no calificados
- B_2: Empleadores de 1 a 4 empleados y autónomos manuales
- B_3: Autónomos no manuales, empleados no manuales con personal, no manuales sin personal y manuales calificados
- B_4: Propietarios capitalistas

El acceso a los bienes y servicios se objetiva por medio de la variable ingreso, la que se ha tricotomizado en:

Ingreso: Y
- A_1, bajo: 0 a 4 000 pesos
- A_2, medio: 4 001 a 8 000 pesos
- A_3, alto: de 8 001 o más

Para estudiar la relación entre la situación de clase y la distribución del ingreso, se consideraron 90 observaciones tomadas de una encuesta[11] que se realizó en 1970 en la ciudad de México. Del cruce de las variables se obtuvo:

[11]Humberto Muñoz, Orlandina de Oliveira y Claudio Stern, *Migración y desigualdad social...*, *op. cit.*

CUADRO 26

Distribución bidimensional de la situación de clase y del ingreso en 90 observaciones, ciudad de México, 1970

Ingreso $Y \setminus X$	*Situación de clase* B_1	B_2	B_3	B_4	*Total*
A_1	12	9	40	1	62
A_2	0	1	20	1	22
A_3	0	0	3	3	6
Total	12	10	63	5	90

En adelante usaremos la información de este cuadro para desarrollar e ilustrar las diferentes aplicaciones del coeficiente ∇_e.

4.2.2. El coeficiente delta-ro (∇_e) para el caso de predicciones bidimensionales de observaciones únicas

El coeficiente delta-ro puede entenderse como una medida del grado de adecuación entre una proposición empírica y la distribución de frecuencias observada. En efecto, una vez que hemos operacionalizado en una variable estadística cada concepto envuelto en la teoría y que expresamos en una proposición la relación entre ellos, estaremos en condiciones de medir el grado de concordancia entre la distribución efectiva de los datos y la que deriva de la proposición empírica.[12]

Para aplicar el coeficiente ∇_e será necesario, en primer lugar, establecer la proposición empírica ro (e)[13]. Supongamos que en el caso de nuestro ejemplo es la siguiente:

[12]No se puede examinar directamente la correspondencia entre la proposición teórica y la distribución de los datos por cuanto median los procesos de objetivación y de operacionalización.

[13]El nombre de este índice destaca el hecho de que depende de la proposición empírica.

ϱ: $B_1 \frown A_1$

$B_2 \frown A_1$

$B_3 \frown (A_1 \text{ o } A_2)$

$B_4 \frown A_3$

De acuerdo con la proposición ϱ, si se conoce la situación de clase, se podría pronosticar el nivel de ingreso de cada observación. Esto querría decir que si una persona tiene la situación de clase B_1 o B_2 debería tener un ingreso bajo (A_1); si se ubica en B_3, bajo o medio (A_1 o A_2) y a un propietario capitalista (B_4) le correspondería un ingreso alto (A_3). Los casos que se desvían del pronóstico proporcionan una primera medida del grado de adecuación de la proposición empírica.

Supongamos que se conoce la situación de clase de cada una de las 90 observaciones y que a través de ϱ se pronostica el ingreso correspondiente. Si la predicción es errónea se agrega uno a la suma de errores o error total. Así, según la proposición ϱ los 12 trabajadores manuales no calificados (B_1) (cuadro 26) deberían tener un ingreso bajo, lo que concuerda perfectamente con las observaciones del cuadro; por tanto, no hay errores. A diez propietarios de talleres artesanales (empleadores de 1 a 4 empleados y trabajadores autónomos manuales B_2), les debería corresponder un ingreso bajo; sin embargo hay uno con ingreso medio que cuenta como un error. Las 63 personas que tienen la situación de clase B_3 deberían tener un ingreso bajo (A_1) o medio (A_2) y como hay tres que poseen uno alto, contamos tres errores más, que sumados al anterior hacen, por el momento, cuatro en total. Para los propietarios capitalistas (B_4) esperábamos que el ingreso fuese alto (A_3); en consecuencia hay dos casos desviados, que sumados a los cuatro anteriores arrojan un error total de seis.

Una manera simple de contar el total de errores consiste en ubicar las casillas que según la proposición empírica no deberían tener casos (que denominaremos casillas de error y que están sombreadas en el cuadro 26) y enseguida sumar sus frecuencias.[14]

Sin embargo, el total de errores o su proporción 6/90 no nos da información suficiente para medir el grado de concordancia (asociación) entre ϱ y la distribución de las observaciones. En efecto, supongamos que las 90 observaciones se distribuyen de la siguiente manera:

[14]Todos los desarrollos que presentaremos suponen que ninguna línea o columna completa se considera de error. Si se previera que ello pueda ocurrir debería eliminarse del análisis la correspondiente línea o columna.

CUADRO 27

Distribución bidimensional (hipotética) de la situación de clase y el ingreso en 90 observaciones

Ingreso	*Situación de clase*				
Y \ X	B_1	B_2	B_3	B_4	*Total*
A_1	84	0	0	0	84
A_2	4	0	0	0	4
A_3	2	0	0	0	2
Total	90	0	0	0	90

Aun cuando en este cuadro también hay seis errores no podríamos decir que el grado de asociación es el mismo, porque una de las variables es constante (la situación de clase) y por lógica consecuencia la relación debería ser nula. De aquí que sea necesario introducir un criterio adicional para juzgar sobre el grado de asociación.

El coeficiente delta-ro se define como la disminución relativa en el error como consecuencia de pasar de una situación en que no conocemos a una en que sí conocemos, para cada observación, su categoría en la variable independiente. Hemos visto cómo se calcula el error total cuando se conoce la categoría de X. Entonces nos faltaría obtener una medida del mismo, en el supuesto de que ignoramos si la observación es B_1, B_2, B_3 o B_4.

Examinemos ahora el procedimiento que nos permitirá predecir Y sin conocer X. Entre las 90 personas que componen la muestra hay 28 (22 + 6) que tienen un ingreso medio o alto (A_2 o A_3) que si perteneciesen a la categoría trabajadores manuales no calificados o empleadores de 1 a 4 empleados y autónomos manuales constituirían una parte del error total. Si seleccionásemos al azar 12 personas (el total de la columna B_1) de entre las 90, esperaríamos que (28/90) × 12 = 3.7 cayesen en las casillas de error de la primera columna y si por el mismo procedimiento elegimos 10 deberíamos tener (20/90) × 10 = 3.1 errores de la segunda columna (B_2). Para la situación de clase B_3 los errores posibles se generan en los seis casos con ingreso alto (la casilla de error es $A_3 \Lambda B_3$); por tanto al seleccionar aleatoriamente 63 personas entre las 90 esperaríamos que aparezcan 6/90 × 63 = 4.2 personas

con ingreso A_3. El cálculo se hace de manera análoga para la clase de los propietarios capitalistas (B_4); en efecto, hay 84 personas que tienen ingresos medios o bajos y al tomar una muestra de 5 (total de la columna B_4) esperaríamos $84/90 \times 5$ casos, que perteneciendo a B_4, sean al mismo tiempo A_1 o A_2.

Si para cada observación suponemos que no conocemos X el error total será igual a:

$$\frac{28}{90} \times 12 + \frac{28}{90} \times 10 + \frac{6}{90} \times 63 + \frac{84}{90} \times 5 = 3.7 + 3.1 + 4.2 + 4.7 = 15.7$$

El miembro de la izquierda se puede escribir de la siguiente manera:

$$\frac{(22)(12)}{90} + \frac{(6)(12)}{90} + \frac{(22)(10)}{90} + \frac{(6)(10)}{90} + \frac{(6)(63)}{90} + \frac{(62)(5)}{90} + \\ + \frac{(22)(5)}{90} = 15.7$$

Cada sumando de esta última expresión es igual al producto de las frecuencias marginales correspondientes a cada casilla de error dividido entre el total de observaciones. Se trata entonces *de las frecuencias esperadas en las casillas de error en el supuesto de independencia estadística.*

Dijimos que ∇_e mide la disminución relativa que se produce en el error total si pasamos de una situación en que desconocemos a qué categoría de X pertenece una observación a otra en que sí lo sabemos. Hemos establecido una manera de calcular el error esperado (sin conocimiento de X) y el modo de obtener el error observado (sabiendo cuál es el valor de X), por tanto estamos en condiciones de computar la disminución en el error total como: error esperado menos error observado. Sabemos que el tamaño de una discrepancia como ésta tendrá un sentido diferente dependiendo del orden de magnitud de las cifras involucradas lo que nos lleva a introducirlo como base de la comparación. Para este efecto podríamos usar el valor de las frecuencias observadas o el de las esperadas, pero como las primeras pueden ser iguales a cero, se opta por recurrir al valor esperado. De este modo llegamos a definir:

(30)

$$\nabla_e = \frac{\text{Error total esperado} - \text{Error total observado}}{\text{Error total esperado}} = \frac{\text{E.E.} - \text{E.O.}}{\text{E.E.}}$$

o su equivalente

(31)

$$\nabla_e = 1 - \frac{\text{E.O.}}{\text{E.E.}}$$

En nuestro ejemplo tenemos que E.O. = 6 y E.E. = 15.7, luego:

$$\nabla_{\varrho} = 1 - \frac{6}{15.7} = 0.62$$

En el caso de que la distribución de frecuencias derivada de la proposición empírica corresponda exactamente con las frecuencias efectivas, el error total observado será igual a cero y por tanto $\nabla_{\varrho} = 1$.

Si conocer la categoría de X a que pertenece cada observación es irrelevante para el pronóstico correcto de las frecuencias (es decir, entre X y Y no hay el tipo de asociación expresado por ϱ), tendremos que los errores observados y esperados serán iguales y por tanto $\nabla_{\varrho} = 0$.

Este coeficiente también puede asumir valores negativos. Dicha situación se presenta cuando el error total observado es mayor que el esperado, es decir, en aquellos casos en que la distribución es tal que las frecuencias se desvían tanto de la predicción que el error supera al que se habría generado por la asignación aleatoria de las observaciones muestrales a las diferentes casillas.

En consecuencia: *i*) el valor $\nabla_{\varrho} = 1$ debe interpretarse como una adecuación perfecta entre la distribución que proviene de la proposición empírica y la distribución de los datos; *ii*) el valor $\nabla_{\varrho} = 0$ significa que entre las variables no hay la relación planteada por ϱ, es decir X y Y son independientes en el sentido establecido por la proposición empírica, de manera que los errores observado y esperado son iguales; *iii*) valores $\nabla_{\varrho} < 0$ indican un total desacuerdo entre la proposición empírica y la distribución de frecuencias observada.

Con el propósito de generalizar esta medida de asociación más allá del ejemplo que hemos venido manejando consideremos una tabla con *r* renglones y *c* columnas, pero en lugar de las frecuencias absolutas (como en la tabla 13) consideremos las proporciones respecto al total.

TABLA 16

Distribución de frecuencias conjunta (en proporciones) de dos variables pluricotómicas

$Y \setminus X$	B_1	B_2	B_3	$\cdots$	B_c	*Total*
A_1	$P(A_1 \Lambda B_1)$	$P(A_1 \Lambda B_2)$	$P(A_1 \Lambda B_3)$	$\cdots$	$P(A_1 \Lambda B_c)$	$P(A_1)$
A_2	$P(A_2 \Lambda B_1)$	$P(A_2 \Lambda B_2)$	$P(A_2 \Lambda B_3)$	$\cdots$	$P(A_2 \Lambda B_c)$	$P(A_2)$
A_3	$P(A_3 \Lambda B_1)$	$P(A_3 \Lambda B_2)$	$P(A_3 \Lambda B_3)$	$\cdots$	$P(A_3 \Lambda B_c)$	$P(A_3)$
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.
A_r	$P(A_r \Lambda B_1)$	$P(A_r \Lambda B_2)$	$P(A_r \Lambda B_3)$	$\cdots$	$P(A_r \Lambda B_c)$	$P(A_r)$
Total	$P(B_1)$	$P(B_2)$	$P(B_3)$	$\cdots$	$P(B_c)$	1.00

Las tablas 13 y 16 tienen una relación simple, puesto que la proporción en la casilla genérica formada por la i-ésima línea y j-ésima columna, resulta de:

$$P(A_i \Lambda B_j) = \frac{N(A_i \Lambda B_j)}{N} \qquad \begin{matrix} (i = 1, 2, \ldots, r) \\ (j = 1, 2, \ldots, c) \end{matrix} \quad y$$

Al aplicar esta transformación al cuadro 26 tenemos:

CUADRO 28

Distribución bidimensional (en proporciones) de la situación de clase y del ingreso en una muestra de 90 observaciones. Ciudad de México, 1970

Ingreso	*Situación de clase*				
$Y \setminus X$	B_1	B_2	B_3	B_4	*Total*
A_1	0.13	0.10	0.45	0.02*	0.70
A_2	0.00	0.01	0.22	0.01	0.24
A_3	0.00	0.00	0.03	0.03	0.06
Total	0.13	0.11	0.70	0.06	1.00

*El valor se ajustó un centésimo para que la suma sea igual a 1.

Ahora bien, supongamos que a la casilla genérica (i, j) de la tabla le asignamos una medida de error ω_{ij}. Si la proposición empírica nos dice que se trata de una casilla de error entonces $\omega_{ij} = 1$; en caso contrario $\omega_{ij} = 0$. A partir de la definición de ω_{ij} se puede construir la siguiente tabla para nuestro ejemplo:

TABLA 17

Medida de error (ω_{ij}) para la relación entre la situación de clase y el ingreso según la proposición ϱ

Ingreso	*Situación de clase*			
$Y \setminus X$	B_1	B_2	B_3	B_4
A_1	0	0	0	1
A_2	1	1	0	1
A_3	1	1	1	0

Dada una proposición empírica cualquiera siempre será posible construir la tabla de la distribución de las medidas de error, que aplicadas convenientemente sobre las proporciones permiten computar el coeficiente delta-ro.

Interesa ahora determinar una expresión matemática que permita computar ∇_ϱ a partir de las proporciones en lugar de las frecuencias absolutas. Para ello recordemos que:

$$\nabla_e = 1 - \frac{\text{E.O.}}{\text{E.E.}}$$

en que E.O. y E.E., denotan el número de casos en las casillas de error, observados y esperados respectivamente. Al dividir entre N el numerador y el denominador de la fracción (lo que aritméticamente no altera el cociente) se tiene que el valor de ∇_e no se alteraría si se usara:

$$P(A_i \Lambda B_j) = \frac{N(A_i \Lambda B_j)}{N}$$

en lugar de

$$N(A_i \Lambda B_j)$$

y

$$P(A_i)\, P(B_j) = \frac{N(A_i)}{N} \times \frac{N(B_j)}{N}$$

en vez de

$$\frac{N(A_i)\, N(B_j)}{N}$$

La medida de error ω_{ij} se introduce con el fin de disponer de una fórmula que distinga entre las casillas de error y las de acierto. Esto se logra aplicando cada ω_{ij} como factor a la frecuencia de la casilla correspondiente:

$$\nabla_e = 1 - \frac{\sum\sum \omega_{ij} P(A_i \Lambda B_j)}{\sum\sum \omega_{ij} P(A_i) P(B_j)} = 1 - \frac{\text{E.O.}}{\text{E.E.}} \tag{32}$$

Con la información del cuadro 28 y de la tabla 17 se pueden computar las sumas que componen el coeficiente ∇_e:

$$\begin{aligned}\sum\sum \omega_{ij} P(A_i \Lambda B_j) = {} & (0)(0.13) + (0)(0.10) + (0)(0.45) + (1)(0.02) + \\ & + (1)(0.00) + (1)(0.01) + (0)(0.22) + (1)(0.01) + \\ & + (1)(0.00) + (1)(0.00) + (1)(0.03) + (0)(0.03) = \\ & = 0.07\end{aligned}$$

y

$$\begin{aligned}\sum \omega_{ij}P(A_i)P(B_j) = {} & (0)(0.70)(0.13) + (0)(0.70)(0.11) + \\ & + (0)(0.70)(0.70) + (1)(0.70)(0.06) + \\ & + (1)(0.24)(0.13) + (1)(0.24)(0.11) + \\ & + (0)(0.24)(0.70) + (1)(0.24)(0.06) + \\ & + (1)(0.06)(0.13) + (1)(0.06)(0.11) + \\ & + (1)(0.06)(0.70) + (0)(0.06)(0.06) = 0.17\end{aligned}$$

Luego

$$\nabla_e = 1 - \frac{0.07}{0.17} = 0.59^{15}$$

En este cálculo se ha supuesto que la importancia del error es la misma para cualquier caso que se desvía de la proposición empírica. Sin embargo, la fórmula (32) deja abierta la posibilidad de asignar pesos diferentes a los errores. La decisión de seguir uno u otro camino (es decir, suponer que para cada casilla de error $\omega_{ij} = 1$ o bien adjudicar pesos distintos) podría depender de la escala de medición de las variables. Como en el ejemplo el ingreso se usa como variable ordinal, el investigador podría estar interesado en dar más importancia a las observaciones que se ubican más lejos de la casilla pronosticada que a las que están más cerca. Así, el caso de un empresario capitalista que tiene un ingreso bajo tendría más importancia en los errores totales que uno que tiene un ingreso medio. Ese criterio para diferenciar la medida de error por casilla (ω_{ij}) no se podría aplicar cuando las variables son nominales porque no hay un orden entre sus valores.

El coeficiente ∇_e se puede aplicar a variables registradas en cualquier escala de medición. Si las variables son nominales todos los ω_{ij} serán iguales entre sí en las casillas de error; como normalmente se usa el criterio de asignar el valor uno al mayor ω_{ij}, entonces en las casillas de error $\omega_{ij} = 1$ y en las casillas de acierto $\omega_{ij} = 0$. Si por lo menos la variable dependiente es ordinal entonces el investigador deberá asignar pesos diferentes a los errores en cada casilla. Como la única recomendación es dar el valor 1 a la casilla de error de mayor gravedad, queda abierta la posibilidad de adscribir en las casillas restantes cualquier número del conjunto de los mayores que cero y menores que uno, lo que significa que, para una misma tabla e igual proposición empírica, se pueden obtener diferentes valores de ∇_e en función del sistema de pesos que se utilice. De aquí que sea conveniente proceder, en cada caso,

[15]La diferencia con el valor obtenido sobre la base de las frecuencias absolutas se explica por la aproximación que realizamos en la casilla (1,4) del cuadro 28.

a un análisis de la sensibilidad del coeficiente ante diferentes conjuntos de ω_{ij}, tema que queda fuera de los límites de este texto.[16]

Sabemos que cada coeficiente ∇_ϱ mide el grado de concordancia entre la proposición empírica y la distribución de frecuencias observadas. Un valor específico de ∇_ϱ depende de ϱ y de la distribución bidimensional; por tanto, no tiene sentido comparar coeficientes que difieren en sus proposiciones empíricas, aplicadas sobre un mismo o diferentes conjuntos de observaciones.

Las dificultades para contrastar los valores que asume ∇_ϱ se deben a: *i*) el número de casillas de error correspondiente a cada proposición empírica; *ii*) a las frecuencias de estas casillas y *iii*) a los pesos ω_{ij}.

Saber si un ∇_ϱ es igual, mayor o menor que otro implica, como mínimo, disponer de una medida que dé cuenta de los efectos "casillas" y "frecuencias". Para este fin se podría proponer como indicador el error total observado o bien el error total esperado, ya que ambos dependerán de las casillas de error (número de sumandos) y de las frecuencias. Entre estas dos posibilidades parece conveniente inclinarse en favor de los errores esperados, en virtud de que nos proporcionan una idea del tamaño del error que se debe reducir (a través de la proposición empírica) y que se vincula directamente con la *precisión envuelta en* ϱ. *Diremos que una proposición* ϱ_2 *es más precisa que una* ϱ_1 *si el error esperado de la primera es mayor que el de la segunda.*

Si la proposición ϱ_2 origina un conjunto de casillas de error que es un subconjunto de las casillas de error que se derivan de ϱ_1, se dice que ϱ_2 está contenida en ϱ_1; si además ambas se aplican sobre una misma distribución de frecuencias, entonces ϱ_2 es más precisa que ϱ_1. Si se usaran para medir la asociación en diferentes distribuciones no podríamos decir cuál es más precisa (aunque lógicamente ϱ_2 esté incluida en ϱ_1) por cuanto el error esperado en una y otra dependerá de las distribuciones marginales de las variables X y Y.

Supongamos que ϱ_1 es la proposición que se usó para ilustrar el cálculo del coeficiente y que dio como resultado $\nabla_\varrho = 0.59$. Examinemos el grado de asociación entre las variables X y Y para la proposición:

ϱ_2: $B_1 \rightarrow A_1$

$B_2 \rightarrow A_1$

$B_3 \rightarrow A_2$

$B_4 \rightarrow A_3$

[16]Las personas interesadas en este tema pueden ver David Hildebrand, *op. cit.*, pp. 30-35.

ϱ_2 se diferencia de ϱ_1 en que el tercer enunciado determina dos casillas de error la (A_1, B_3) y la (A_3, B_3) en lugar de una, lo que se manifiesta en que el error esperado (E.E.) en ϱ_2 es mayor. En efecto, si simbolizamos por $(E.E.)_1$ el error esperado total que proviene de la proposición ϱ_1 y $(E.E.)_2$ el que deriva de ϱ_2 tendremos que:[17]

$$\begin{aligned}(E.E.)_2 = {} & (1)(0.70)(0.06) + (1)(0.24)(0.13) + (1)(0.24)(0.11) + \\ & + (1)(0.24)(0.06) + (1)(0.06)(0.13) + (1)(0.06)(0.11) + \\ & + (1)(0.06)(0.70) + (1)(0.70)(0.070)\end{aligned}$$

La suma de los términos de la derecha excepto el último conforma el error esperado total según la proposición ϱ_1. Luego,

$$(E.E.)_2 = (E.E.)_1 + (1)(0.70)(0.70) = 0.17 + 0.49 = 0.66$$

Lo que permite concluir que ϱ_2 es más precisa que ϱ_1. El coeficiente $\nabla_{\varrho 2}$ es igual a:

$$\nabla_{\varrho 2} = 1 - \frac{0.52}{0.66} = 1 - 0.79 = 0.21$$

Diremos que el valor de un coeficiente $\nabla\varrho$ domina a otro si es mayor y su precisión es por lo menos igual.

En este ejemplo $\nabla_{\varrho 2}$ no domina a $\nabla_{\varrho 1}$ porque aunque es más preciso, no es mayor.

Sabemos que si dos variables X y Y son estadísticamente independientes entonces:

$$P(A_i \Lambda B_j) = P(A_i)P(B_j) \qquad \begin{matrix}(i = 1, 2, \ldots, r) \\ (j = 1, 2, \ldots, c)\end{matrix} \quad y$$

si sustituimos en (32) tenemos que:

$$\nabla_{\varrho} = 1 - \frac{\sum\sum \omega_{ij} P(A_i)P(B_j)}{\sum\sum \omega_{ij} P(A_i)P(B_j)} = 0$$

Por tanto, cuando dos variables son estadísticamente independientes el coeficiente $\nabla_{\varrho} = 0$. Pero lo inverso no es necesariamente cierto. El índice de asociación delta-ro puede ser cero sin que las variables sean estadísticamente independientes. *Asume el valor nulo si los atributos de las casillas de error son independientes.* Es decir, si

[17]Eliminamos todos los términos cuyos $\omega_{ij} = 0$.

(33) $$P(A_i \Lambda B_j) = P(A_i)P(B_j)$$

para todas las casillas en que $\omega_{ij} = 1$.

Dos variables son estadísticamente independientes si lo son todos sus atributos (y por ello $\nabla_e = 0$), pero no lo son si sólo algunas de sus categorías (que al coincidir con las casillas de error producen $\nabla_e = 0$) cumplen con la condición especificada en (33).

Ahora estamos en condiciones de precisar el significado del coeficiente: ∇_e *mide simultáneamente el grado de adecuación entre* e *y la distribución de los datos y la lejanía respecto a la independencia estadística entre los atributos que definen las casillas de error.*

4.2.3. Un caso particular de ∇_e: el coeficiente $\varkappa$ de Cohen

Si *i*) las dos variables que se cruzan tienen el mismo número de categorías (r = c), *ii*) la proposición empírica es:

$$\varkappa: \quad B_i \frown A_i \qquad (i = 1, 2, \ldots r)$$

y *iii*) a las casillas de error se les asigna $\omega_{ij} = 1$, entonces el coeficiente ∇_e es igual al coeficiente de $\varkappa$ de Cohen.

Las tres condiciones se sintetizan en las dos tablas siguientes:

TABLA 18

Distribución de frecuencias conjunta (en proporciones) de dos variables pluricotómicas con el mismo número de categorías

$Y \backslash X$	B_1	B_2	B_3	$\cdots$	B_r	*Total*
A_1	$P(A_1 \Lambda B_1)$	$P(A_1 \Lambda B_2)$	$P(A_1 \Lambda B_3)$	$\cdots$	$P(A_1 \Lambda B_r)$	$P(A_1)$
A_2	$P(A_2 \Lambda B_1)$	$P(A_2 \Lambda B_2)$	$P(A_2 \Lambda B_3)$	$\cdots$	$P(A_2 \Lambda B_r)$	$P(A_2)$
A_3	$P(A_3 \Lambda B_1)$	$P(A_3 \Lambda B_2)$	$P(A_3 \Lambda B_3)$	$\cdots$	$P(A_3 \Lambda B_r)$	$P(A_3)$
.	.	.	.		.	.
.	.	.	.		.	.
.	.	.	.		.	.
A_r	$P(A_r \Lambda B_1)$	$P(A_r \Lambda B_2)$	$P(A_r \Lambda B_3)$	$\cdots$	$P(A_r \Lambda B_r)$	$P(A_r)$
Total	$P(B_1)$	$P(B_2)$	$P(B_3)$		$P(B_r)$	1.00

TABLA 19

Medida de error (ω_{ij}) para la proposición $\varkappa$

$Y \setminus X$	B_1	B_2	B_3	···	B_r
A_1	0	1	1	···	1
A_2	1	0	1	···	1
A_3	1	1	0	···	1
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
A_r	1	1	1	···	0

Al aplicar los ω_{ij} sobre las correspondientes proporciones de la tabla 18 y desarrollando se llega al coeficiente $\varkappa$ de Cohen:

$$\nabla_{\varkappa} = \frac{\sum P(A_i \Lambda B_i) - \sum P(A_i)P(B_i)}{1 - \sum P(A_i)P(B_i)} \tag{34}$$

Para ilustrar el uso de este coeficiente supongamos que se decide unir las categorías B_1 y B_2 del cuadro 28, de manera que tenemos:

CUADRO 29

Distribución bidimensional (en proporciones) de la situación de clase (tricotomizada) y del ingreso en una muestra de 90 observaciones. Ciudad de México, 1970

Ingreso	*Situación de clase*			
$Y \setminus X$	B_1^*	B_2^*	B_3^*	*Total*
A_1	0.23	0.45	0.02	0.70
A_2	0.01	0.22	0.01	0.24
A_3	0.00	0.03	0.03	0.06
Total	0.24	0.70	0.06	1.00

Nota: B_1^* resulta de la unión de B_1 y B_2 en el cuadro 28. B_2^* es B_3 y B_3^* es B_4.

La proposición empírica es:

$$\varkappa: \quad B_1^* \frown A_1$$
$$B_2^* \frown A_2$$
$$B_3^* \frown A_3.$$

Si aceptamos que la situación de clase está socialmente jerarquizada de menor a mayor, la proposición $\varkappa$ nos estaría diciendo que a mayor situación de clase mayor ingreso. Para evaluar esta relación calculamos:

$$\nabla_{x} = \frac{0.4800 - 0.3396}{0.6604} = \frac{0.1404}{0.6604} = 0.2126$$

obteniendo de este modo una idea del grado de concordancia entre la proposición empírica x y la distribución de las observaciones.

4.2.4. Coeficiente delta-ro para el caso de predicciones bidimensionales sobre pares de observaciones

La versión de ∇_{ϱ} que expusimos en la sección anterior permite medir el grado de asociación entre variables de cualquier nivel de medición. El hecho de que esta sección se dedique a predicciones sobre pares de observaciones lleva implícita la idea de comparación, por lo cual todos los desarrollos que presentamos serán aplicables a variables que en sí sean ordinales o bien que se puedan expresar ordinalmente.

El próximo apartado es de carácter introductorio y en él incluimos una serie de consideraciones que son básicas para los desarrollos posteriores.

4.2.4.1. Tabla ordinal condensada

Supongamos que se ha cruzado una variable X, dicotomizada en alto (A) y bajo (A'), con una variable Y tricotomizada en alto (A), medio (M) y bajo (B) y que se ha llegado a la siguiente distribución:

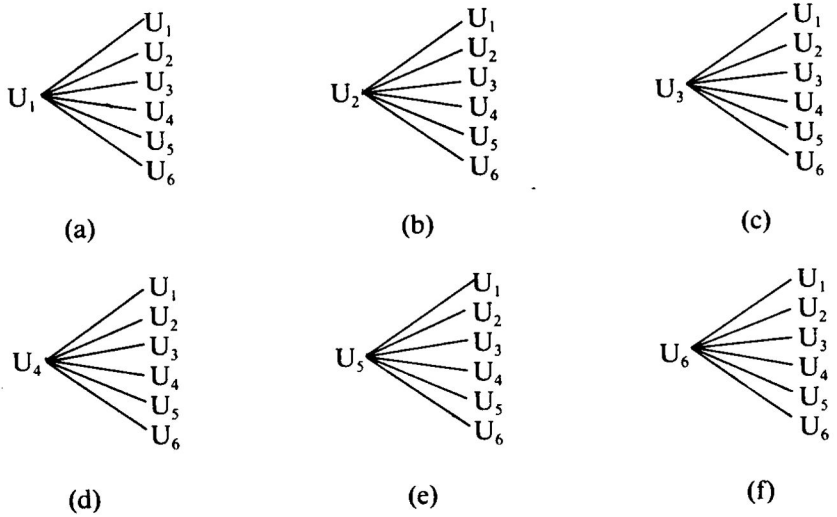
CUADRO 30

Distribución bidimensional (hipotética) de las variables X y Y

$Y \setminus X$	*Alto* *(A)*	*Bajo* *(A')*	*Total*
Alto (A)	2	0	2
Medio (M)	2	0	2
Bajo (B)	0	2	2
Total	4	2	6

Sobre la base de esta información, ¿cómo proceder a realizar todas las comparaciones posibles entre las seis observaciones? ¿Cuáles son los tipos de resultados que obtendremos? ¿Hay regularidades que permitan establecer rutinas fijas de cálculo? Con el propósito de responder a estas interrogantes introduciremos una simbología que nos permitirá examinar el detalle del proceso de comparación.

Si U_1, U_2, U_3, U_4, U_5 y U_6 representan las seis unidades del cuadro 30, tendremos 36 comparaciones posibles, que se pueden observar en los siguientes esquemas:



Cada línea indica no sólo las observaciones envueltas en la comparación sino también su sentido. Por ejemplo, la segunda línea de (a) y la primera de (b) comparan U_1 con U_2; sin embargo, el orden es diferente. En (a) tenemos U_1 comparado con U_2 y en (b) U_2 comparado con U_1.

Para expresar U_1 en términos de sus atributos, usaremos dos coordenadas. La primera representará a X y la segunda a Y. Así, si las dos observaciones de la casilla formada por la primera línea y la primera columna del cuadro 30 son U_1 y U_2, entonces:

$$U_1 = (A, A) \qquad \text{y} \qquad U_2 = (A, A)$$

Suponiendo que U_3 y U_4 son las observaciones de la casilla identificada por la segunda línea y la primera columna y U_5 y U_6 las dos restantes, tendremos que:

$$U_3 = (A, M);\ U_4 = (A, M);\ U_5 = (A', B) \quad \text{y} \quad U_6 = (A', B)$$

Tomemos, a manera de ejemplo, algunas comparaciones entre los U_1. Supóngase que nos interesa:

U_1 c U_1; expresión que debe leerse "U_1 comparado con U_1". Una manera equivalente de plantear lo mismo, pero esta vez en término de los atributos (X, Y) sería:

$$(A, A) \text{ c } (A, A)$$

El resultado es "igual" en X y en Y o bien (I, I). Nótese que la comparación es ordenada porque sólo tiene sentido hacerla entre los respectivos atributos de X y los de Y.

Veamos ahora el resultado de U_1 c U_5:

$$(A, A) \text{ c } (A', B) \longrightarrow (M_a, M_a)$$

La flecha significa "da por resultado" y M_a significa mayor. Por tanto U_1 es mayor que U_5 tanto en X como en Y.

Al realizar la comparación inversa tendremos:

$$(A', B) \text{ c } (A, C) \longrightarrow (M_e, M_e),$$

donde M_e significa "menor".

Este resultado es interesante pues nos dice que si en la comparación de dos observaciones obtenemos el par (M_a, M_a) automáticamente en el conjunto de comparaciones se da también su simétrico (M_e, M_e).

Si nos interesa U_1 c U_3, tenemos:

$$(A, A) \text{ c } (A, M) \longrightarrow (I, M_a)$$

y al realizar U_3 c U_1:

$$(A, M) \text{ c } (A, A) \longrightarrow (I, M_e)$$

En la comparación siempre aparecerán estos casos en pareja, lo mismo ocurrirá con el (M_a, I) y el (M_e, I) que en este ejemplo no se presentó.

En la tabla 20 se detallan las 36 comparaciones posibles entre las seis observaciones.

TABLA 20

Comparación entre las 6 observaciones del cuadro 30

	Observaciones	*Atributos*	*Resultados*
(1)	U_1 c U_1	(A, A) c (A, A)	(I, I)
(2)	U_1 c U_2	(A, A) c (A, A)	(I, I)
(3)	U_2 c U_1	(A, A) c (A, A)	(I, I)
(4)	U_1 c U_3	(A, A) c (A, M)	(I, M_a)
(5)	U_3 c U_1	(A, M) c (A, A)	(I, M_c)
(6)	U_1 c U_4	(A, A) c (A, M)	(I, M_a)
(7)	U_4 c U_1	(A, M) c (A, A)	(I, M_c)
(8)	U_1 c U_5	(A, A) c (A', B)	(M_a, M_a)
(9)	U_5 c U_1	(A', B) c (A, A)	(M_c, M_c)
(10)	U_1 c U_6	(A, A) c (A', B)	(M_a, M_a)
(11)	U_6 c U_1	(A', B) c (A, B)	(M_c, M_c)
(12)	U_2 c U_2	(A, A) c (A, A)	(I, I)
(13)	U_2 c U_3	(A, A) c (A, M)	(I, M_a)
(14)	U_3 c U_2	(A, M) c (A, A)	(I, M_c)
(15)	U_2 c U_4	(A, A) c (A, M)	(I, M_a)
(16)	U_4 c U_2	(A, M) c (A, A)	(I, M_c)
(17)	U_2 c U_5	(A, A) c (A', B)	(M_a, M_a)
(18)	U_5 c U_2	(A', B) c (A, A)	(M_c, M_c)
(19)	U_2 c U_6	(A, A) c (A', B)	(M_a, M_a)
(20)	U_6 c U_2	(A', B) c (A, A)	(M_c, M_c)
(21)	U_3 c U_3	(A, M) c (A, M)	(I, I)
(22)	U_3 c U_4	(A, M) c (A, M)	(I, I)
(23)	U_4 c U_3	(A, M) c (A, M)	(I, I)
(24)	U_3 c U_5	(A, M) c (A', B)	(M_a, M_a)
(25)	U_5 c U_3	(A', B) c (A, M)	(M_c, M_c)
(26)	U_3 c U_6	(A, M) c (A', B)	(M_a, M_a)
(27)	U_6 c U_3	(A', B) c (A, M)	(M_c, M_c)
(28)	U_4 c U_4	(A, M) c (A, M)	(I, I)
(29)	U_4 c U_5	(A, M) c (A', B)	(M_a, M_a)
(30)	U_5 c U_4	(A', B) c (A, M)	(M_c, M_c)
(31)	U_4 c U_6	(A, M) c (A', B)	(M_a, M_a)
(32)	U_6 c U_4	(A', B) c (A, M)	(M_c, M_c)
(33)	U_5 c U_5	(A', B) c (A', B)	(I, I)
(34)	U_5 c U_6	(A', B) c (A', B)	(I, I)
(35)	U_6 c U_5	(A', B) c (A', B)	(I, I)
(36)	U_6 c U_6	(A', B) c (A', B)	(I, I)

Al tomar dos observaciones y comparar sus correspondientes atributos en X y Y sólo habrá tres resultados posibles para cada variable: son iguales (I), mayores (M_a) o menores (M_c). Esto se puede ver con claridad en la tabla 20. Esta regularidad permite resumir la información obtenida en un solo cuadro:

CUADRO 31

Forma ordinal condensada: distribución del total de comparaciones posibles tomando dos observaciones cada vez

Y\X	*Mayor*	*Igual*	*Menor*	*Total*
Mayor	8	4	0	12
Igual	0	12	0	12
Menor	0	4	8	12
Total	8	20	8	36

Según la tabla 20 hay ocho comparaciones que entregaron el resultado "mayor en X y mayor en Y", por simetría sabemos que la casilla (3, 3) también debe tener frecuencia 8. En la casilla "igual en X y mayor en Y" tenemos cuatro casos y sabemos que la misma frecuencia debe aparecer en igual en X y menor en Y. Como la tercera casilla de la primera línea es cero, lo mismo debe ocurrir con su simétrica (3, 1) es decir, en la comparación no hubo ningún caso en que X fuese menor y Y mayor y por tanto el caso exactamente opuesto tampoco existe. Una situación análoga se presenta con la casilla formada por "mayor en X e igual en Y" y su casilla simétrica "menor en X e igual en Y". Por último, tenemos que la frecuencia de casilla igual en X e igual en Y es 12.

Nótese que las marginales también son simétricas debido a que si en algunas comparaciones el atributo X es mayor, habrá la misma cantidad en que es menor. Esto también es válido para los atributos de Y.

Cuando el número total de observaciones es grande, la cantidad de comparaciones que habría que realizar consumiría gran cantidad de tiempo. Note que los cálculos que presentamos se refieren a sólo seis observaciones. Afortunadamente existe un procedimiento de cálculo que permite llegar de manera expedita a la forma ordinal condensada. Esto se logra a través de la siguiente rutina de cálculo:

1) Se determina el número total de comparaciones posibles, que es igual al cuadrado del número de casos: en el ejemplo, como $N = 6$, el total es $6^2 = 36$.

2) El total de empates (igual) en X y en Y se obtiene como la suma de los cuadrados de las frecuencias marginales correspondientes: la frecuencia marginal del igual en X es $4^2 + 2^2 = 20$ y del igual en Y es $2^2 + 2^2 + 2^2 = 12$ (véase el cuadro 30).

3) Para obtener las restantes frecuencias marginales se opera por simetría. Sabemos que quedan $36 - 20 = 16$ comparaciones que clasificar en la marginal de X, y que deben distribuirse por igual entre la categoría mayor y menor de modo que la frecuencia de cada una sea 8. Un razonamiento análogo permite determinar las frecuencias restantes para la variable Y. En efecto, como hay 12 comparaciones que arrojan

el resultado igual en Y, quedan 24 que se reparten 12 y 12 entre mayor

Al completarse el tercer paso se conocen el total y las marginales del cuadro 31. De aquí en adelante veremos cómo se determinan las frecuencias dentro del cuadro.

4) La frecuencia de la casilla igual en X e igual en Y resulta de la suma de cuadrados de las frecuencias conjuntas del cuadro 30.

$$2^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 = 12$$

5) Como sabemos que el cuadro es simétrico, resulta fácil determinar las frecuencias de la columna y de la línea que corresponden al igual. Para la columna, tenemos que distribuir $20 - 12 = 8$ comparaciones equitativamente entre las casillas restantes, lo que entrega frecuencia 4 para las celdas (1, 2) y (3, 2). De manera análoga se obtienen los ceros de la segunda línea.

6) Para calcular el número de comparaciones que dan por resultado mayor en X y mayor en Y hay que operar sobre el cuadro 30 de la siguiente manera: *i*) eliminar las observaciones de la línea y columna correspondientes a la casilla seleccionada como pivote[18] por cuanto no pueden dar como resultado mayor en X y simultáneamente mayor en Y; *ii*) examinar caso por caso las frecuencias restantes y seleccionar las que resultarían en "mayor, mayor" comparadas con el pivote. Así, por ejemplo, las dos observaciones que estén en la casilla (1, 1) del cuadro 30 no pueden ser mayores en X y en Y que las que se encuentran en la primera línea y en la primera columna. Se cumple la condición de que las observaciones tengan valores mayor en X y Y que los casos de la casilla pivote, en las casillas (2, 2) y (3, 2) cuyas respectivas frecuencias son 0 y 2. El número de comparaciones que entregan un resultado mayor, se obtiene por medio de:

$$2 \times \text{la suma de } \begin{bmatrix} 0 \\ 2 \end{bmatrix} = 4$$

Al tomar como pivote la casilla (2, 1) los únicos casos que al compararse contra las dos observaciones seleccionadas tienen simultáneamente mayor X y mayor Y son las dos que se encuentran en la casilla (3, 2). Luego el número de comparaciones "mayor, mayor" es:

$$2 \times \text{la suma de } [2] = 4$$

[18]Se denomina pivote a la casilla que se toma como base de comparación. El pivote se va cambiando de manera que todas las casillas hayan sido consideradas como tal.

Cuando se selecciona como pivote la casilla (3, 1) se tiene que el número de comparaciones que da un resultado "mayor, mayor" es cero porque *i*) la frecuencia en la casilla pivote es cero y *ii*) no hay ninguna observación que tenga mayor X y mayor Y que los casos de esa casilla pivote. Estas mismas razones son válidas para las casillas (1, 2) y (2, 2). Si se toma como pivote la casilla (3, 2) el número de comparaciones que entregan un valor mayor en X y mayor en Y es cero, debido a la segunda de las razones que acabamos de señalar.

Los 8 casos se registran en la casilla "mayor, mayor" del cuadro 31 y por simetría debe corresponder el mismo número a la casilla "menor, menor".

7) Las frecuencias que quedan por determinarse, en la esquina superior derecha e inferior izquierda del cuadro condensado, se obtienen por diferencia.

Al aplicar esta secuencia de cálculo al cuadro 26, se obtuvo:

CUADRO 32

Forma ordinal condensada para la situación de clase y el ingreso en una muestra de 90 observaciones. Ciudad de México, 1970

$Y \backslash X$	*Mayor*	*Igual*	*Menor*	*Total*
Mayor	805	996	67	1 868
Igual	1 059	2 246	1 059	4 364
Menor	67	996	805	1 868
Total	1 931	4 238	1 931	8 100

Esta distribución se puede expresar en proporciones, aplicando la secuencia de cálculo señalada a las frecuencias relativas del cuadro 28. Otra manera de obtener el mismo resultado consiste en dividir las frecuencias entre el número total de comparaciones (casos) en el cuadro condensado:

CUADRO 33

Forma ordinal condensada (en proporciones) para la situación de clase y el ingreso en una muestra de 90 observaciones. Ciudad de México, 1970

$Y \backslash X$	*Mayor*	*Igual*	*Menor*	*Total*
Mayor	0.0994	0.1229	0.0083	0.2306
Igual	0.1307	0.2774	0.1307	0.5388
Menor	0.0083	0.1229	0.0994	0.2306
Total	0.2384	0.5232	0.2384	1.000

Estableceremos una simbología que nos permita identificar de manera general los componentes de la tabla ordinal condensada en proporciones y llegar a una fórmula para el cálculo de ∇_e. Denotaremos por $P(E_x)$ la proporción de empates en X; por tanto, $P(E'_x) = 1 - P(E_x)$ es lo que falta para completar las tres categorías del total por columna y sabemos que se distribuye por igual entre las dos casillas restantes, de manera que tendremos en cada una $\frac{1 - P(E_x)}{2} = \frac{P(E'_x)}{2}$. De manera análoga se definen las proporciones de la variable Y. $P(E_y)$ para la casilla intermedia del total por renglón y $P(E'_y) = 1 - P(E_y)$ para las otras dos; correspondiendo a cada casilla $\frac{P(E'_y)}{2}$.

Por otra parte, la comparación que resulta en las combinaciones mayor en X y mayor en Y, o bien menor en X y menor en Y, reflejan una correspondencia entre ambas variables, por lo que la denotaremos como proporción de concordancias, P(C).[19] Este total se reparte por mitades entre las casillas (1, 1) y (3, 3).

En las casillas extremas de la diagonal secundaria están las proporciones en que no concuerdan los atributos de las variables. Representan los casos en que hay discordancia y las simbolizaremos por P(D). Sabemos que se distribuyen ½P(D) en la casilla (1, 3) y el otro ½P(D) en la (3, 1).

La proporción de casos en que la comparación arroja el resultado igual en X e igual en Y nos indica el peso, $P(E_{xy})$, de los empates. La característica de las dos celdas que resta identificar en la segunda línea tienen la particularidad de que hay empate en Y, pero no en X, $P(E_{x'y})$ y que son del mismo tamaño por lo que en cada una habrá una proporción de $\frac{1}{2}P(E_{x'y})$. De manera similar denotamos las proporciones de las casillas (1, 2) y (3, 2) de la segunda columna: $\frac{1}{2}P(E_{xy'})$.

Tenemos, entonces que:

$$P(C) = 0.1988 \qquad P(E_{x'y}) = 0.2614$$
$$P(D) = 0.0166 \qquad P(E_x) = 0.5232$$
$$P(E_{xy}) = 0.2774 \qquad P(E'_x) = 0.4768$$
$$P(E_{xy'}) = 0.2458 \qquad P(E_y) = 0.5388$$
$$P(E'_y) = 0.4612$$

Ahora estamos en condiciones de presentar la forma general de la tabla ordinal condensada.

[19] El igual se considerará aparte.

TABLA 21

Tabla ordinal condensada (forma general)

$Y \setminus X$	*Mayor*	*Igual*	*Menor*	*Total*
Mayor	$\frac{1}{2}P(C)$	$\frac{1}{2}P(E_{xy'})$	$\frac{1}{2}P(D)$	$\frac{1}{2}P(E'_y)$
Igual	$\frac{1}{2}P(E_{x'y})$	$P(E_{xy})$	$\frac{1}{2}P(E_{x'y})$	$P(E_y)$
Menor	$\frac{1}{2}P(D)$	$\frac{1}{2}P(E_{xy'})$	$\frac{1}{2}P(C)$	$\frac{1}{2}P(E'_y)$
Total	$\frac{1}{2}P(E'_y)$	$P(E_x)$	$\frac{1}{2}P(E'_x)$	1.00

Para aplicar ∇_ϱ sobre la tabla ordinal condensada bastará con expresar cualquier predicción en la distribución de la medida de error (ω_{ij}), que aplicada sobre la tabla 21 producirá fórmulas diferentes.

La proposición empírica ϱ determina los ω_{ij} y el coeficiente ∇_ϱ asume una forma particular, que es la que permite evaluar el grado en que ϱ concuerda con la distribución de las observaciones.

Todas las medidas de asociación que presentaremos en las secciones siguientes son casos particulares de ∇_ϱ.

4.2.4.2. Coeficientes de asociación de Sommers y Kim

El índice de asociación d de Sommers resulta ser equivalente al ∇_ϱ bajo dos condiciones. La primera es que la proposición empírica sea:[20]

d_{xy}: Mayor ⟶ Mayor
Igual ⟶ (Mayor, igual o menor)
Menor ⟶ Menor

y la segunda que se aplique el siguiente sistema de pesos a las casillas de error.

TABLA 22

Medidas de error (ω_{ij}), para obtener el coeficiente de Sommers a partir de ∇_ϱ

$Y \setminus X$	*Mayor*	*Igual*	*Menor*
Mayor	0	0	1
Igual	½	0	½
Menor	1	0	0

[20] La simbología utilizada para representar las proposiciones empíricas puede prestarse a confusiones cuando éstas se plantean sobre la tabla ordinal condensada (ya que las categorías tienen el mismo nombre) si no recordamos que el término a la izquierda de ⟶ se refiere a una categoría de la variable independiente (x) y el de la derecha a la correspondiente categoría de la variable dependiente (y).

Según la fórmula (ecuación 32) bastará con aplicar los ω_{ij} de la tabla 22 sobre las proporciones correspondientes de la tabla ordinal condensada, para obtener ∇_{dyx}, el que después de algunas simplificaciones algebraicas resulta ser:

$$(35) \quad \nabla_{dxy} = \frac{P(C) - P(D)}{P(E'_x)} = \frac{P(C) - P(D)}{P(C) + P(D) + P(E_{x'y})} = d_{xy}$$

La proposición empírica nos dice que: *i*) a mayor X (la situación de clase) mayor Y (el ingreso). Esto quiere decir que cada vez que comparamos dos observaciones debe ocurrir que la que tiene un mayor X debe tener también un mayor Y; *ii*) el enunciado menor —⌒→ menor es la consecuencia lógica del primero y *iii*) cuando hay empate en X no se hace ninguna afirmación; si se da igualdad para un X particular, Y puede tomar cualquier valor.

En cuanto a las medidas de error hay que notar que se le adjudica el peso mayor a las casillas que implican más lejanía respecto de lo postulado por los enunciados. En la tabla 22 tenemos que $\omega_{31} = \omega_{13} = 1$. Y sólo la mitad de importancia a las casillas de error más cercanas a las de acierto: $\omega_{21} = \omega_{23} = ½$. También hay que destacar que los pesos se distribuyen simétricamente, lo que no es más que una necesidad lógica derivada de la estructura de la tabla ordinal condensada.

Al aplicar la versión d de Sommers de ∇_e al cuadro 33 obtenemos como resultado:

$$\nabla_{d_{xy}} = \frac{0.1988 - 0.0166}{0.4768} = \frac{0.1822}{0.4768} = 0.3821$$

que mide el grado de concordancia entre la proposición d_{xy} y la distribución de los datos.

El coeficiente asimétrico de Kim se obtiene a partir de:

$d_{x'y}$: Mayor —⌒→ (Mayor o igual)
Igual —⌒→ Igual
Menor —⌒→ (Igual o menor)

y del siguiente sistema de ponderaciones:

TABLA 23

Medida de error (ω_{ij}) para obtener el coeficiente asimétrico de Kim a partir de ∇_e

Y \ X	*Mayor*	*Igual*	*Menor*
Mayor	0	½	1
Igual	0	0	0
Menor	1	½	0

Al aplicar estos pesos sobre la tabla ordinal condensada en proporciones llegamos a:

$$(36) \qquad \nabla_{d_{x \cdot y}} = \frac{P(C) - P(D)}{P(E'_y)} = \frac{P(C) - P(D)}{P(C) + P(D) + P(E_{xy'})} = d_{x \cdot y}$$

En el caso de nuestro ejemplo tenemos:

$$\nabla_{d_{x \cdot y}} = \frac{0.1988 - 0.0166}{0.4612} = \frac{0.1822}{0.4612} = 0.3951$$

El coeficiente simétrico de Kim es el ∇_ϱ correspondiente a la proposición empírica:

d: Mayor —⌒→ Mayor
Igual —⌒→ Igual
Menor —⌒→ Menor

y a las ponderaciones:

TABLA 24

Medidas de error (ω_{ij}) para obtener el coeficiente simétrico de Kim a partir de ∇_ϱ

$Y \backslash X$	*Mayor*	*Igual*	*Menor*
Mayor	0	¼	1
Igual	¼	0	¼
Menor	1	¼	0

Con estos pesos obtenemos:

$$\nabla_d = \frac{P(C) - P(D)}{\tfrac{1}{2}\{P(E'_y) + P(E'_x)\}} = \frac{P(C) - P(D)}{\tfrac{1}{2}\{2P(C) + 2P(D) + P(E_{x'y}) + P(E_{xy'})\}}$$

En nuestro ejemplo:

$$\nabla_d = \frac{0.1988 - 0.0166}{\tfrac{1}{2}\{0.4612 + 0.4768\}} = \frac{0.1822}{0.4690} = 0.3885$$

La línea argumental que lleva al coeficiente de asociación ∇_ϱ permite aclarar el sentido en que se mide la asociación: como éste difiere en los tres coeficientes no debe sorprender que den resultados distintos.

Tienen que ser distintos porque miden aspectos diferentes. Sus valores provienen de las proposiciones empíricas y de los pesos adjudicados a las casillas de error. *Miden el grado de concordancia de la distribución de los casos con sus correspondientes proposiciones empíricas y como son lógicamente diferentes, también lo serán los valores de los coeficientes aplicados a la misma información.*

4.2.4.3. El coeficiente de asociación gamma (γ)

Para obtenerlo como un caso particular de ∇_e es necesario redefinir la tabla ordinal condensada eliminando la línea y la columna de los empates. Sabemos que:

$$P(C) + P(D) + P(E) = 1$$

por tanto la proporción de no empates será:

$$P(E') = 1 - P(E) = P(C) + P(D)$$

Como necesitamos que las proporciones en la nueva tabla ordinal condensada sumen la unidad, entonces las dividimos entre P(E').

TABLA 25

Forma ordinal condensada al eliminar los empates

$Y \setminus X$	*Mayor*	*Menor*	*Total*
Mayor	P(C) 2P(E')	P(D) 2P(E')	0.5
Menor	P(D) 2P(E')	P(C) 2P(E')	0.5
Total	0.5	0.5	1.0

Cuando la proposición empírica es:

γ: Mayor ⟶ Mayor
Menor ⟶ Menor

tenemos que:

$$\nabla_\gamma = 1 - \frac{\frac{1}{2}\frac{P(D)}{P(E')} + \frac{1}{2}\frac{P(D)}{P(E')}}{\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}}$$

$$\nabla_\gamma = \frac{P(C) - P(D)}{P(E')}$$

(37)
$$\nabla_\gamma = \frac{P(C) - P(D)}{P(C) + P(D)} = \gamma$$

Adviértase que no fue necesario establecer los pesos (ω_{ij}) de las casillas de error por cuanto se trata de dos celdas simétricas cuyos errores deberían tener la misma importancia.

Para nuestro ejemplo el valor de γ

$$\nabla_\gamma = \frac{0.1988 - 0.0166}{0.1988 + 0.0166} = \frac{0.1822}{0.2154} = 0.8459 = \gamma$$

Como se puede observar gamma entregó un valor bastante superior al de los otros coeficientes que hemos presentado. Hildebrand *et al.*,[21] señalan que una situación como ésta es un caso particular de un hecho de validez universal: gamma siempre entregará un valor mayor que cualquiera de las medidas "d". Señalan que al no considerar los empates, gamma mide distorsionadamente la reducción en el error. Agregan que:

> Por esta razón, aunque apareció tempranamente en la literatura, y por consecuencia, se ha usado más ampliamente que cualquier otra de las medidas d, gamma aparece inadecuado para evaluar "A más X, más Y". Un ejemplo particularmente notable es la siguiente tabla de 2 × 2 y su forma ordinal condensada con y sin empates:

		X	
		x	x′
Y	y	0.1	0.0
	y′	0.8	0.1

Forma ordinal condensada con empates

Y \ X	*Mayor*	*Igual*	*Menor*	*Total*
Mayor	0.01	0.08	0.00	0.09
Igual	0.08	0.66	0.08	0.82
Menor	0.00	0.08	0.01	0.09
Total	0.09	0.82	0.09	1.00

[21] *Op. cit.*, p. 46.

Forma ordinal condensada al eliminar los empates

$Y \backslash X$	*Mayor*	*Menor*	*Total*
Mayor	0.5	0	0.5
Menor	0	0.5	0.5
Total	0.5	0.5	1.0

Claramente Y no aumenta fuertemente con X. Cuando x ocurre, 8/9 de las observaciones son aún y'. Pero, para esta tabla gamma −1.0.

En contraste, todas las medidas d son igual a 0.111.[22]

4.2.4.4. Consideraciones finales sobre ∇_ϱ

En las secciones anteriores entregamos el material que, en nuestra opinión, constituye los desarrollos básicos referidos al coeficiente de asociación delta-ro. Nuestra exposición no ha pretendido ser exhaustiva sino que sólo ha tenido como propósito delinear sus rasgos fundamentales.

No presentamos una serie de otros coeficientes que también son casos particulares de ∇_ϱ. Tampoco incluimos los índices de asociación que no son derivables de la forma delta-ro, tales como τ_b^2 y τ_c.

El trabajo que originalmente presentó este coeficiente[23] se explaya sobre su aplicación al análisis de variables múltiples, así como al estudio de asociación en distribuciones de frecuencias que combinan variables con distintos niveles de medición. El lector interesado en profundizar su conocimiento del índice de asociación ∇_ϱ tendrá que revisar detenidamente el trabajo que ha servido de base para la exposición del material de esta sección.

Ejercicios[24]

1. La siguiente tabla, tomada de Muñoz, De Oliveira y Stern, muestra el cruce entre escolaridad y migración en 1970 en la ciudad de México:

[22] *Ibidem.*
[23] David Hildebrand, *et al., op. cit.*
[24] Véase la bibliografía al final de los ejercicios.

Categoría migratoria y escolaridad, ciudad de México, 1970

	Migrantes según periodo de exposición					
Escolaridad terminada	*Menos de 10 años*	*10-19 años*	*20 años y más*	*Nativos por adopción*	*Nativos por nacimiento*	*Total*
Ninguna	16 305	9 478	12 474	8 873	11 552	58 602
Primaria incompleta	83 854	64 770	68 605	53 239	127 071	397 539
Primaria completa	48 915	28 436	26 507	37 267	150 176	291 301
Secundaria y preparatoria	37 268	25 276	17 151	47 916	150 176	277 787
Universidad incompleta	18 634	14 218	4 678	12 422	63 536	113 488
Universidad completa	27 951	15 798	26 507	17 746	75 088	163 090
Total	232 927	157 976	155 922	177 463	577 599	1 301 887

Establezca una proposición empírica teóricamente justificada y proceda a evaluar su grado de correspondencia con la distribución bidimensional de frecuencias.

2. José Nun destaca en los siguientes términos la relevancia de estudiar la población flotante: "las prácticas del empleador contribuirán cada vez más a reforzar objetivamente las estrategias de bloqueo y/o ruptura de la unidad obrera que sirven a la reproducción del sistema capitalista en su conjunto" (p. 62).

Respecto al estudio de este problema en las sociedades periféricas nos dice que las modalidades de la superpoblación relativa adquieren una funcionalidad específica en el proceso de acumulación dado el distorsionado desarrollo capitalista. En consecuencia no se puede seguir el análisis en abstracto sino que es necesario proceder a investigar cómo opera el mecanismo de expulsión-absorción de fuerza de trabajo.

Con este propósito tomó, en 1970, una muestra de obreros despedidos entre el último cuatrimestre de 1966 y el primero de 1968, por las plantas argentinas de Chrysler y Fiat.

De los cuadros incluidos en el estudio seleccionamos el siguiente:

Ocupaciones actuales por cuenta propia,
por calificación en el empleo de origen

	No calificados	*Calificados*	*Total*
Reparación de automotores	9	8	17
Otras industrias	4	2	6
Comercios varios	6	3	9
Transportes	3	2	5
Total	22	15	37

Entre los comentarios que realiza el autor respecto a este cuadro destacamos:

> Por lo demás, si el acceso a estas actividades puede reputarse comparativamente de "fácil" desde el punto de vista del conjunto de los poseedores de capital, no ocurre lo mismo cuando se adopta como criterio de referencia a la masa de asalariados. Primeramente, con la sola excepción de dos comisionistas, las ocupaciones que estoy analizando hicieron necesaria una inversión de capital que, en la mitad de los casos, excedió $m.n. 500 000. Si se tiene en cuenta que el ingreso medio anual neto del sector asalariado activo fue, en 1967, de alrededor de $m.n. 350 000, resulta que ese 50% tuvo que disponer de recursos equivalentes a unos 18 meses de salario medio... A este respecto, importa recordar que 4 de cada 5 trabajadores autónomos debieron contar con un local (22 casos) o con un medio de transporte (5 casos). Más aún, estas restricciones en el acceso tendieron a verse incrementadas por exigencias específicas de calificación, como lo evidencia el subgrupo mayoritario de dueños de talleres de reparación de automotores (p. 76.)

Señale cuál sería la variable independiente y cuál la dependiente. Cómo se justifica o puede justificar la relación planteada. Examine la posibilidad de establecer una proposición empírica a partir del argumento del autor.

3. En un estudio sobre los empresarios industriales brasileños Fernando H. Cardoso hace una revisión de las ideas que conducían a esperar que la burguesía industrial latinoamericana fuese capaz de crear una economía industrial urbana y simultáneamente ensanchar el margen de autonomía en cuanto a las decisiones políticas y económicas. Dice que:

> esta última expectativa presupone la mediación de dos procesos: 1) que la alianza desarrollista tendría suficiente fuerza para imponer su hegemonía a los demás grupos sociales y llegaría a controlar el Estado y 2) que el sector industrial urbano, al cual se vincularía el sector financiero a medida que se desligara del esquema importador-exportador, estaría controlado por una burguesía nacional (p. 3.).

También sostiene que el proceso de industrialización incitaría a las clases empresariales industriales no sólo a politizarse sino a prever su misión histórica de aglutinar a los demás grupos urbanos y concebir una política de defensa de sus intereses, es decir, crear las condiciones para la expansión del mercado interno. Sin embargo, en el caso brasileño se produce una apertura creciente de los mercados al capital extranjero durante el gobierno de Kubitschek y en lo que se refiere a política agraria se puede percibir un rechazo latente o manifiesto a la reforma agraria.

> Ante esta evidencia política se han esgrimido varias razones originadas más bien en inferencias lógicas basadas en el sentido común que en análisis sistemáticos que se pueden clasificar en dos grandes grupos: *i*) argumentos que dicen relación con las consecuencias políticas de los orígenes sociales

de la burguesía industrial brasileña (agricultores capitalistas, artesanos y comerciantes prósperos). *ii*) Argumentos que ponen el énfasis en obstáculos estructurales, coyunturales o culturales a la cristalización en la burguesía nacional de un punto de vista capaz de darle conciencia de sus verdaderos intereses.

Basándose en un trabajo empírico de Luciano Martins, [25] Cardoso establece la hipótesis que el proceso de sustitución de importaciones brasileño conduce a una inestabilidad de las capas empresariales y que su actuación política es una forma de ajustarse a esta situación. Además plantea que

> objetivamente la diferenciación de las capas productivas nacionales llevaría a la formación de grupos económicos que incluyen sectores industriales, sectores de servicios, sectores financieros y sectores agrarios como un recurso normal de ajuste para el desarrollo capitalista de un país que sufre un rápido proceso de sustitución de importaciones (p. 11).

En el análisis de referencia interesa discutir el sentido de la presencia política y los límites de la participación de la burguesía industrial en la política de alianza desarrollista. Para aquilatar mejor esta problemática se examina la forma en que se presenta el desarrollo industrial en dos dimensiones:

i) Las características de la estructura de producción de los grupos económicos que operan en Brasil y *ii*) las formas de control interno o externo que predomina en ellos.

Respecto al primero de estos dos aspectos se sostiene la existencia de grupos económicos multimillonarios (con un capital y reservas de más de 4 000 millones de cruceiros de 1962) y el de los millonarios que comprende conjuntos de empresas con capitales y reservas que fluctúan entre 1 000 y 4 000 millones de cruceiros de 1962. En el primero de estos grupos se encontraban 55 empresas que controlaban la mayor parte de la producción y circulación de bienes. Se plantea que al parecer el desarrollo de la economía brasileña estaría condicionado por la formación de grupos económicos que actúan en condiciones de competencia imperfecta. Por tanto, es fundamental analizar la estructura productiva de estas empresas y determinar los mecanismos políticos y sociales por medio de los cuales son controladas financieramente. El análisis pone de relieve que hay necesidad de aclarar algunos puntos para comprender el funcionamiento de la economía brasileña y posteriormente de los estratos sociales que la controlan.

Para dar sustento a su planteamiento Fernando H. Cardoso agrega la siguiente tabla de distribución de frecuencias:

[25] Luciano Martins, *Formação do Empresariado Industrial no Brasil*, edición mimeografiada del Instituto de Ciencias Sociales de la Universidad Federal de Rio de Janeiro, 1966.

Distribución de los grupos multimillonarios por sector de actividad y por tipo de control financiero

	Tipo de control			
Sectores	*Nacionales*	*Extranjeros*	*Mixtos*	*Total*
No industrial (exportación, importación, servicios industriales e investigación)	8	6	1	15
Consumo no durable	8	5	—	13
Consumo durable	1	7	—	8
Mecánica pesada	1	4	—	5
Industria básica	6	7	1	14
Total	24	29	2	55

Se le pide que sobre la base de la argumentación teórica identifique la hipótesis que conduce al autor a presentar esta tabla. Enseguida tradúzcala a una proposición empírica y realice el análisis estadístico de la información.

4. En la sección 2.3.3 ilustramos el uso del coeficiente fi por medio del cruce entre educación y estatus socioeconómico en el caso de Estados Unidos, tal como nos lo presenta Raymond Boudon.

En este ejercicio presentamos la tabla original que habíamos redimensionalizado a una de dos por dos. Tomando en cuenta el argumento de Boudon ya expuesto en la sección aludida y la información:

Estatus socioeconómico del hijo en relación al padre como función de su nivel educacional con relación al padre

Estatus socioeconómico del hijo con relación al padre	*Nivel educacional del hijo con relación al padre*			
	Mayor	*Igual*	*Menor*	*Total*
Mayor	134	23	7	164
Igual	96	33	16	145
Menor	61	24	22	107
Total	291	80	45	416

Se le pide que: *i*) establezca la proposición empírica compatible con el argumento presentado en el texto y *ii*) examine el grado de correspondencia entre su proposición y la distribución empírica de frecuencias. (Sugerencia: discuta si se trata de predicciones para observaciones únicas o para pares de ellas.)

5. Peter H. Smith estructura su investigación sobre la élite política mexicana, en el siglo XX, sobre la base de tres preguntas de carácter general, pero esenciales: ¿quién gobierna?, ¿quiénes tienen acceso al poder? y ¿cuáles son las condicionantes sociales de ese acceso? (p. 5).

Estos propósitos generales se especifican en tres campos temáticos de indudable interés en la historia contemporánea de México: *i*) "El efecto a largo plazo que la Revolución mexicana haya o no podido tener sobre la composición política de la élite (p. 5). *ii*) Dado que en México impera un régimen autoritario interesa "determinar quiénes quedan dentro y quiénes quedan fuera de los límites del poder" (p. 5). *iii*) El análisis de la relación "entre las alteraciones que ocurren en el seno de la sociedad en general y las alteraciones en el seno de la composición de los grupos dominantes" (p. 6).

Además de realizar una periodización por sexenio, define tres cohortes políticas diferenciadas en función "del periodo a lo largo del cual se ocupó un cargo público de importancia nacional" (p. 25).

> La primera cohorte comprende a todos aquellos que ocuparon un cargo público entre 1900 y 1911 (N = 610); la segunda incluye a quienes lo hicieron entre 1917 y 1940 (N = 2 289), y la tercera a aquellos que ocuparon un cargo público de nivel nacional entre 1946 y 1971 (N = 2 008) [...] Para minimizar traslapes entre cohortes suprimí conscientemente dos periodos históricos; a) el periodo de violencia y desorden que va de 1911 hasta 1917, en el que una multitud de facciones luchaban por afirmar su supremacia, provocando con ello una gran inestabilidad en el poder y b) 1940-1946, cuando Manuel Ávila Camacho presidió una etapa de transición en la que ante el ascenso de los futuros líderes de la cohorte de 1946-1971 muchos remanentes del grupo 1917-1949 todavía se aferraban a algún puesto público (pp. 25 y 26).

Al analizar las condiciones del poder político (cap. 3) examina tres hipótesis en relación al efecto que tiene el origen social de la élite: *i*) la revolución amplió sustancialmente las tasas de movilidad ascendente (p. 93); *ii*) la Revolución mexicana, por lo menos en lo que respecta a su liderazgo, fue un movimiento esencialmente burgués (p. 93); *iii*) tomando en cuenta los datos referidos a las élites políticas en Gran Bretaña, Francia, Alemania y los Estados Unidos revelan el debilitamiento del papel político de la clase alta, la creciente importancia de la clase media y el ascenso, dentro de ciertos límites, de la clase baja, aunados a la lentitud y los constantes cambios en los orígenes de las élites llevaría a sostener que los cambios se pueden explicar más bien por alteraciones en las estructuras sociales y económicas que de la revolución de 1910-1920 (p. 94).

El autor examina estas tres hipótesis a partir de la siguiente distribución bidimensional de frecuencias:

Élites políticas: clases sociales por cohortes[26]

	Cohortes			
Clases sociales	*Prerrevolucionaria*	*Revolucionaria*	*Posrevolucionaria*	*Total*
Alta	12	15	11	38
Media	19	69	135	223
Baja	4	17	46	67
Total	35	101	192	328

i) Discuta la forma como se objetivó el concepto de clase social; *ii*) establezca las proposiciones empíricas correspondientes a las tres hipótesis y *iii*) examine el grado de correspondencia entre las proposiciones empíricas y la distribución de frecuencias. (Indicación: decida previamente, en cada caso, si la hipótesis requiere de la construcción de un cuadro ordinal consensado.)

6. Del libro *Hogares y trabajadores...*, de García, Muñoz y De Oliveira, presentamos la distribución de los jefes hombres (90% del total de los jefes) de las unidades domésticas estudiadas, según el sector social al que pertenecen y su nivel de ingreso:

	Sector social			
Nivel de ingreso	*Trabajadores por cuenta propia*	*Trabajadores asalariados no manuales*	*Trabajadores asalariados manuales*	*Total*
Menos del salario mínimo	68	18	157	243
De 1 a 3.9 veces el salario mínimo	108	286	635	1 029
De 4 a 9.9 veces el salario mínimo	23	167	31	221
10 o más veces el salario mínimo	8	55	0	63
Total	207	526	823	1 556

¿Puede sostenerse que el sector social y el nivel de ingreso están asociados? Calcule la fuerza de la relación.

REFERENCIAS BIBLIOGRÁFICAS DE LOS EJERCICIOS

Boudon, Raymond. *Education, Opportunity & Social Inequality: Changing prospect in western society*, John Wyley, Nueva York, 1974.
Cardoso, Fernando Henrique. *Empresarios industriales y desarrollo nacional en Brasil*,

[26]Este cuadro es un resumen del que presenta el autor en la página 92 de su obra. La variable clase agrupa un conjunto detallado de ocupaciones cuyo listado se encuentra en el cuadro 3-3 del trabajo de referencia. Además, hay que consignar que el mismo autor dedica parte de la página 91 a establecer las limitaciones de la información que presenta.

Centro de Estudios Socio-económicos (CESO), Universidad de Chile. Santiago, 1968.
García, Brígida, Humberto Muñoz y Orlandina de Oliveira. *Hogares y trabajadores en la ciudad de México*, El Colegio de México y UNAM, México, 1982.
Muñoz, Humberto, Orlandina de Oliveira y Claudio Stern. *Migración y desigualdad social en la ciudad de México*, El Colegio de México e Instituto de Investigaciones Sociales, UNAM, México, 1980.
Nun, José. "Despidos en la industria automotriz argentina: estudio de un caso de superpoblación flotante", en *Revista Mexicana de Sociología, 1/78*, año XL, vol. XL, núm. 1, México, enero-marzo de 1978.
Smith, Peter. *Los laberintos del poder: el reclutamiento de las élites políticas en México, 1900-1971*, El Colegio de México, México,1981.

Apéndice

A continuación incluimos dos de los ejercicios propuestos en este libro que han sido desarrollados por Ángel Blas y Fernando Zúñiga, estudiantes de la quinta promoción (1984-1986) de la maestría de la FLACSO. Hemos decidido tomar ejercicios referidos a la misma investigación y seguir el orden del texto, por lo que primero presentamos el cruce de dos variables dicotómicas. Luego incorporamos una tercera variable, también dicotómica (variable control) y, por último, con la información del primer ejercicio analizamos el cruce de las dos variables tratadas en el primer ejercicio, pero esta vez una de ellas con cuatro categorías.

Ejercicio 2, capítulo 2

La información que se presenta se obtuvo del trabajo de S.M. Lipset: "La formación de opiniones en una situación de crisis". Es un estudio de opinión de los estudiantes de la Universidad de Berkeley, California, en plena época del macarthismo. Uno de los aspectos que interesaba en el estudio era la situación docente de los profesores comunistas.

Entre los objetivos del estudio se quería establecer la relación entre la posición ideológica de los estudiantes y su opinión respecto a la política de empleo que debía seguir la universidad con los profesores comunistas.

Si bien ambas variables constan de cuatro categorías, para propósitos de la primera parte de este ejercicio las hemos dicotomizado.

La información que permite establecer la relación ideología-opinión se encuentra en el cuadro 1.

CUADRO 1

Número de estudiantes según opinión respecto a la política de empleo de profesores comunistas por ideología (1950)

	Ideología		
Política de empleo a profesores comunistas	*Liberal*	*Conservador*	*Total*
Favor	130	76	206 (52.8%)
Contra	82	102	184 (47.2%)
Total	212 (54.4%)	178 (45.6%)	390

Del análisis de las distribuciones marginales se desprende que la mayoría de los estudiantes son liberales (54.4%) y que 52.8% de los entrevistados está a favor de la política de empleo de los profesores comunistas. De esta información pareciera desprenderse que la ideología liberal predispone al estudiante a tener una postura más abierta, en tanto que la conservadora sería uno de los factores explicativos de la posición que está en contra del empleo de los profesores comunistas.

Para establecer la relación insinuada, no basta con el análisis de las marginales. La concomitancia porcentual (54% de estudiantes liberales y 52.2% de estudiantes a favor de la contratación de profesores comunistas) no es suficiente para dar evidencia en favor o en contra de la hipótesis que relaciona ambas variables. En efecto, estos porcentajes ocultan aquellos casos de estudiantes liberales que están contra la política y los de estudiantes conservadores que están a favor. Del cuadro 1 se ve que éstos son 82 y 76 casos respectivamente.

Además, para medir el grado de relación entre las variables hay que definir la hipótesis con precisión. Supongamos que se sostiene que la posición ideológica determina unívocamente la postura del alumno respecto a la política de contratación, es decir que, los estudiantes liberales deben estar a favor y los conservadores en contra.

Si simbolizamos por B y B′ las categorías liberal y conservadora respectivamente y por A y A′ las posiciones a favor y en contra, entonces se puede escribir la hipótesis de la siguiente forma:

$$H1: \quad B \longrightarrow A$$

y

$$B' \longrightarrow A'$$

El cálculo de porcentajes nos proporciona una primera aproximación al análisis de la relación, esta información aparece en el cuadro 2. Seguimos la norma de calcular los porcentajes sobre los totales de las categorías de las variables independientes.

CUADRO 2

Distribución porcentual de los estudiantes de Berkeley, según su ideología y por posición respecto a la contratación de profesores comunistas

	Ideología		
Política de empleo a profesores comunistas	*Liberal*	*Conservador*	*Total*
Favor	61.3	42.7	52.8
Contra	38.7	57.3	47.2
Total	100.0	100.0	100.0

La distribución porcentual muestra evidencia en favor de la hipótesis. Los porcentajes más altos se encuentran en la diagonal principal, lo que quiere decir que hay una tendencia a que los estudiantes liberales estén a favor de la contratación de los profesores comunistas y los conservadores en contra.

Para saber la fuerza de la relación hay que decidir qué coeficiente utilizar. La naturaleza de la hipótesis exige que se calcule el coeficiente fi:

$$\Phi = \frac{(130)(102) - (82)(76)}{\sqrt{(212)(178)(206)(184)}} = 0.186$$

La conclusión es relativamente clara. Si bien parece haber una relación, como la postulada, entre ambas variables (decimos parece, porque decidir si la relación es significativa o no, requiere de conocimientos de inferencia estadística) ésta ha resultado débil.

También es posible analizar la hipótesis que sostiene que los estudiantes ideológicamente conservadores se opondrán a la contratación de profesores comunistas y que los liberales tendrán una posición ambigua. Algunos a favor y otros en contra. El argumento conceptual para cambiar la hipótesis podría estar basado en la consideración del impacto que tendría una serie de factores no considerados sobre las posiciones de los estudiantes. La presión social que experimentan los estudiantes en la universidad, las condicionantes de la interacción familiar, el impacto de los medios de comunicación de masas, etc., fortifican las posturas de los conservadores y debilitan las de los liberales.

En símbolos, la nueva hipótesis sería:

$$\text{H2:} \quad B \frown\!\!\rightarrow (A \text{ o } A')$$
$$B' \frown\!\!\rightarrow A'$$

La estructura lógica de esta hipótesis requiere del coeficiente Q de Yule:

$$Q = \frac{(130)(102) - (82)(76)}{(130)(102) + (82)(76)} = 0.36$$

Se concluye entonces, que hay una relación débil aún bajo esta segunda hipótesis, que sabemos, es menos exigente.

Ejercicio 1 del capítulo 3

Uno de los factores que pueden explicar las débiles relaciones obtenidas al cruzar la ideología y la posición respecto a la política de contratación

de profesores, es el efecto que tiene la permanencia en la universidad. El desarrollo académico e intelectual que logran los estudiantes en la universidad mediaría en la forma como se traduce la ideología en posturas políticas. En efecto, se podría argumentar que uno de los efectos importantes de la formación universitaria es el de ganar en capacidad de pensamiento libre y analítico de manera que los estudiantes tenderían a filtrar las consecuencias de las posiciones ideológicas para adaptarlas a circunstancias diferentes. Si esta hipótesis es cierta deberíamos observar una menor relación en la subpoblación de los estudiantes que terminan estudios que los que empiezan.

Con el propósito de investigar estas hipótesis someteremos la información a un análisis de covarianzas de Lazarsfeld.

La información de base se despliega en el cuadro 1 (los totales no coinciden con los del cuadro 1 del ejercicio anterior debido a que hemos eliminado la categoría residual no sabe o no contesta).

CUADRO 1

Opinión de los estudiantes de acuerdo con su permanencia en la universidad; por posición respecto a la política de empleo de profesores comunistas

	Estudiantes que empiezan estudios		
Los profesores comunistas deben enseñar	*Liberales*	*Conservadores*	*Total*
Sí	74	25	99
No	74	38	112
Total	148	63	211
	Estudiantes que terminan estudios		
Los profesores comunistas deben enseñar	*Liberales*	*Conservadores*	*Total*
Sí	118	16	134
No	72	32	104
Total	190	48	238

En este ejercicio tenemos tres variables dicotómicas, la variable independiente "ideología" (X), la dependiente "posición respecto a la política de contratación" (Y) y la variable control o variable test (Z, con categorías C y C') "antigüedad en la universidad". Si suponemos que la posición ideológica se adquiere en la vida cotidiana, entonces podemos suponer que Z es anterior a X. Para identificar a cuál de los cuatro tipos ideales se aproxima la estructura de las relaciones entre estas variables tenemos que decidir si el modelo es parcial o marginal.

Hasta este punto hemos operado con criterios estrictamente estadísticos y con las líneas que se desprenden de los planteamientos de Lazarsfeld. Debemos agregar ahora las consideraciones teóricas.

Los argumentos que justifican la introducción de la tercera variable implican que los datos del cuadro 1 deberían mostrar una relación entre X y Y en la subpoblación de los estudiantes que empiezan los estudios y que ésta debería debilitarse en la subpoblación de los que están por terminarlos. Por otra parte, parece consistente con los razonamientos expuestos, que la hipótesis sería del tipo:

$$B \longrightarrow A$$

y

$$B' \longrightarrow A'.$$

Esto nos llevaría a calcular los valores de los coeficientes fi en las dos subpoblaciones que están contenidas en el cuadro 1.

Estudiantes que empiezan estudios:

$$\Phi_{xy.c} = \frac{(74)(38) - (74)(25)}{\sqrt{(148)(63)(99)(112)}} = 0.09$$

Estudiantes que terminan estudios:

$$\Phi_{xy.c} = \frac{(118)(32) - (72)(16)}{\sqrt{(190)(48)(134)(104)}} = 0.23$$

Estos resultados rechazan totalmente la hipótesis del efecto de los estudios universitarios sobre la capacidad de pensamiento libre que en lugar de jugar el papel de filtro parecieran más bien reforzar el papel de la ideología.

El valor prácticamente nulo, que toma el coeficiente fi, en la subpoblación de los estudiantes que empiezan estudios hace pensar que el medio social (a las edades de estos alumnos) no genera una ideología más o menos definida.

En este punto se abren dos posibilidades para seguir el análisis: extraer las consecuencias que tiene el planteamiento general sobre las relaciones marginales o incorporar el resultado que hemos obtenido del análisis de las parciales. Seguiremos este último camino.

Deberíamos esperar que la relación entre la permanencia en la universidad y la posición respecto a la contratación de los profesores comunistas debería ser del tipo:

$$C \longrightarrow (A \text{ o } A')$$

$$C' \longrightarrow (A \text{ o } A')$$

Esta hipótesis sostiene que no debería esperarse ninguna relación entre la permanencia en la universidad y la posición respecto a la política de contrataciones (o sea que las variables son independientes).

CUADRO 2

Distribución de los estudiantes según su permanencia en la universidad por posición respecto a la política de emplear profesores comunistas

	Permanencia		
Los profesores comunistas deben enseñar	*Empiezan*	*Terminan*	*Total*
Sí	99	134	233
No	112	104	216
Total	211	238	449

La hipótesis de independencia admite el uso del coeficiente fi o Q. El cálculo de fi arrojó el valor −0.09, lo que implica que la estructura de las relaciones es del tipo parcial. Por la forma de la ecuación de covarianzas de Lazarsfeld basta que una de las marginales sea cero para que el modelo sea parcial. Ya sabemos que la variable control es anterior a X, en consecuencia, el tipo que mejor ajusta es el parcial anterior.

Tenemos entonces que la relación entre la ideología y la posición respecto a la política de emplear profesores comunistas se debe básicamente, a que los alumnos que están por terminar sus estudios mantienen posiciones relativamente coherentes con su ideología.

Para terminar este segundo ejercicio analizaremos la relación entre la permanencia y la ideología. En este caso no adelantaremos hipótesis por lo que nos limitaremos a interpretar lo que dicen los datos:

CUADRO 3

Permanencia de los estudiantes en la universidad

Los profesores comunistas deben enseñar	*Empiezan*	*Terminan*	*Total*
Sí	148	190	338
No	63	48	111
Total	211	238	449

El cálculo del coeficiente Φ arroja un resultado de −0.11. Si bien la relación es baja nos indica que los estudiantes que empiezan tienden a ser conservadores y los que terminan liberales. ¿Será éste un impacto de la universidad sobre la transformación de la postura ideológica? En realidad esta pregunta no se puede responder debido a que se trata de la comparación de dos grupos y no del seguimiento de uno solo, además de que no se ha considerado el control de otras variables.

El coeficiente Q de Yule da un valor de −0.25, lo que se interpreta de la mismma manera que el resultado de fi.

Continuación del ejercicio 2, capítulo 2

Este ejercicio es una extensión del primero incluido en este apéndice y sólo una de las variables es dicotómica. La variable independiente (ideología) tiene cuatro categorías ordenadas, que van desde los liberales extremistas hasta los conservadores moderados. La variable explicada es dicotómica y diferencia a los alumnos que están a favor de la política de contratación de profesores comunistas, de los que se oponen.

La información se despliega en el cuadro 1.

CUADRO 1

Actitud de los estudiantes de la Universidad de Berkeley sobre la política de empleo de los profesores comunistas según ideología

Actitud	*Liberal extrema*	*Liberal moderada*	*Conservadora moderada*	*Conservadora extrema*	*Total*
A favor	96	34	51	25	206
En contra	48	34	56	46	184
Total	144	68	107	71	390

El coeficiente ji cuadrado, arroja los siguientes resultados:

$$\chi^2 = 21.272$$

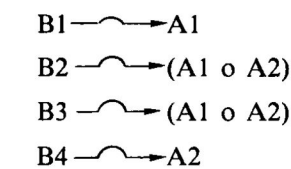
El coeficiente de Cramer es:

$$C = 0.2274$$

Este valor no se diferencia sustancialmente del que obtuvimos para Φ en el problema 1. Muestra que las frecuencias observadas se alejan levemente de las esperadas bajo el supuesto de independencia estadística. Por lo que la conclusión ya obtenida no se altera. Parece haber una relación débil entre la ideología y la actitud respecto a la contratación de profesores comunistas.

En el caso en que definamos asociación, no por lejanía, respecto a la independencia estadística, sino conceptualmente su medición adecuada implica el uso del coeficiente delta-ro.

Supongamos que nuestro planteamiento teórico especifica que sólo es posible esperar una relación nítida entre la ideología y la actitud sólo en los extremos ideológicos. Esto originaría la siguiente proposición empírica:



Donde:

B1 representa la ideología liberal extrema
B2 representa la ideología liberal moderada
B3 representa la ideología conservadora moderada
B4 representa la ideología conservadora extrema

y

A1 representa la actitud a favor de la contratación de profesores comunistas.

A2 representa la actitud en contra de la contratación de profesores comunistas.

Las casillas de error en el cuadro 1 son la extrema inferior izquierda (2, 1) y la extrema superior derecha (1, 4).

En consecuencia los errores observados son: $73 = 48 + 25$. Y los errores esperados:

$$105.4 = 67.9 + 37.5 = 144 \times 184/390 + 71 \times 206/390.$$

Por lo tanto:

$$\nabla_e = 1 - 73/105.4 = 0.307$$

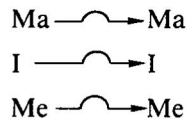
Se concluye entonces que hay una relación entre ambas variables, aunque débil.

Examinemos ahora una posibilidad conceptualmente distinta para definir asociación. Supongamos que se plantea que al comparar a dos estudiantes se puede predecir su actitud, de manera que si empatan en la variable independiente (ambos son liberales extremos, liberales moderados, conservadores moderados o conservadores extremos), entonces también empatarán en la variable dependiente (ambos estarán a favor o en contra de la política de contratación).

Pero si el primero tiene una posición relativamente más liberal que el segundo (lo que quiere decir que si el primero es liberal extremo el segundo es ya sea liberal moderado o conservador moderado o conservador extremo, o bien si, por ejemplo, el primero es conservador moderado, el segundo tiene que ser conservador extremo), entonces la actitud del primero será a favor de la contratación de profesores comunistas, en tanto que la del segundo en contra.

En caso de que el primero tenga una posición ideológica más conservadora que la del segundo, la relación sería inversa a la examinada en el párrafo anterior.

Las tres situaciones descritas conducen a la proposición empírica:



que debe aplicarse sobre la tabla ordinal condensada.

En efecto, a lo largo del texto se ha visto que la construcción de esta tabla se basa en la comparación entre los valores de X para un par de observaciones y los correspondientes valores de Y, lo que coincide con el desarrollo lógico de la hipótesis.

Aplicando el procedimiento de cálculo para generar la tabla ordinal condensada a partir de la original, se obtiene el cuadro 2 (construido sobre la base de los datos del cuadro 1).

CUADRO 2

Tabla ordinal condensada correspondiente a la actitud de los estudiantes de Berkeley, respecto a la política de empleo de los profesores comunistas, según posición ideológica

	Posición ideológica			
Actitud	*Mayor*	*Igual*	*Menor*	*Total*
Mayor	18 870	9 770	9 264	37 904
Igual	26 991	22 310	26 991	76 292
Menor	9 264	9 770	18 870	37 904
Total	55 125	41 850	55 125	152 100

Las casillas de error son todas aquellas que se encuentran fuera de la diagonal principal y la frecuencia de los errores observados es:

$$72\,510 = 26\,991 + 9\,264 + 9\,770 + 9\,770 + 9\,264 + 26\,991.$$

Un cálculo de frecuencias esperadas, aplicado sobre las mismas casillas arroja un resultado 82 774.8.

El valor del coeficiente de asociación delta-ro es igual a:

$$\nabla_{\varrho} = 1 - 72\,510/82\,774.8 = 0.124$$

La relación (si es que existe), nuevamente es débil.

Métodos estadísticos aplicados a la investigación en ciencias sociales. Análisis de asociación
se terminó de imprimir en julio de 1987
en los talleres de Programas Educativos, S.A. de C.V.,
Chabacano 65-A, 06850 México, D.F.
Se imprimieron 2000 ejemplares más
sobrantes para reposición.
Diseñó la portada Mónica Díez-Martínez.
La edición estuvo al cuidado de José Gabriel Guerra

Centro de Estudios
Sociológicos

Esta obra tiene particular importancia, en especial en el campo de la docencia, donde se observa una marcada carencia de textos adecuados a la realidad de la investigación social en América Latina, aunque ello no excluye su uso como libro de consulta para investigadores que requieran del análisis de asociación.

Los autores nos ofrecen un libro claro y sencillo, sin dejar con ello de lado el rigor en la exposición. Se trata de una búsqueda —si así quiere llamársele— en la que se pretende establecer una articulación entre el componente teórico y el proceso de investigación, que desea establecer la relación entre lo que explica, de manera tentativa, un problema u objeto de estudio (las hipótesis) y su constatación empírica correspondiente (el mundo de los datos).

El libro tiene una ventaja adicional, desde la perspectiva de las ciencias sociales: ilustra las maneras de correlacionar variables con base en ejemplos que se desprenden de la investigación social. Ello hace que el lector no enfrente una serie de fórmulas abstractas, sino que permite que esas fórmulas tengan un contenido que se vincula a problemáticas con las cuales está familiarizado el investigador de estas disciplinas.

Ilustración de la portada: Obra del escultor *Sebastián*.
Fotografía de *Enrique Bostelmann*.

El Colegio de México