Construcción y análisis de datos. Introducción al uso de técnicas cuantitativas en la investigación social.
Posadas, © 2009 Denis Baranger. ISBN 978-987-05- 6307-5. Se permite la reproducción citando la fuente.

CAPÍTULO 4

TÉCNICAS ELEMENTALES DE ANÁLISIS

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

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

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

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

Forma lógica de la tabla de 2 x 2

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

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

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

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

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

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

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

n = (X) + (-X)
= (Y) + (-Y)
= (-X Y) + (X Y)+ (-X-Y) + (X-Y)

Aplicación del modelo a un ejemplo
Este modelo puede aplicarse para cualquier población y todo tipo de variables:

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

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

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

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

580.522 = 317.098 + 263.424
= 291.007 + 289.515
= 194.397 + 96.610 + 122.701 + 166.814

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

error en la tabulación.

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

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

¿ Para qué sirven los porcentajes?

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

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

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

Fuente: Tabla 4.2

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

$$\frac{292.007}{580522} \times 100 \quad \text{o, en general} \quad Y/n \times 100$$

¿Cuándo redondear los porcentajes?

En realidad, el resultado de la operación aritmética anterior arroja la cifra de 50,12850503, la que nosotros hemos redondeado a un decimal anotando 50,1.[3] ¿ Por qué este redondeo? El interés de los porcentajes es indicar con la mayor claridad las dimensiones relativas de dos o más números, tranformando a uno de esos números, la *base*, en la cifra 100. Es indudable que:

291.007/580.522 = 50/100 = 50%

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

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

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

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

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

¿ Cómo se lee una tabla de contingencia?

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

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

① 194.397 personas vivían en hogares sin NBI en áreas urbanas;

② 96.610 lo hacían en hogares con NBI en áreas urbanas;

③ 122.701 pertenecían a hogares sin NBI en áreas rurales;

④ 166.814 habitaban en áreas rurales en hogares con NBI.

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

Modos alternativos de análisis

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

---

[4] Dada la imprecisión de nuestros instrumentos de medición.

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

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

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

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

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

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

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

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

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

Fuente: Tabla 4.2

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

③ El título finaliza con la expresión '**(%)**'; ello nos indica que las cifras consignadas en las celdas son porcentajes, y no frecuencias absolutas.[7]

④ En el encabezamiento de la última columna aparece la expresión '**Total (100%)**'. Esto quisiera expresar: a) que en dicha columna las cifras **no** son porcentajes sino frecuencias absolutas; y b) que las cifras absolutas de la columna fueron tomadas como base para calcular los porcentajes de las celdas.[8]

¿Cómo se lee una cifra porcentual?

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

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

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

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

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

Igualmente correcto sería escribir:

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

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

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

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

---

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

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

— El total de la columna, ‘317.098’; y
— El “total total”, el ‘n’: ‘580.522’.

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

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

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

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

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

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

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

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

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

---

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

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

¿ Cómo se lee un conjunto de porcentajes?
Podemos ahora leer el conjunto de las cifras de la Tabla 4.2.1, que traducimos en la siguiente serie de enunciados:

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

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

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

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

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

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

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

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

---

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

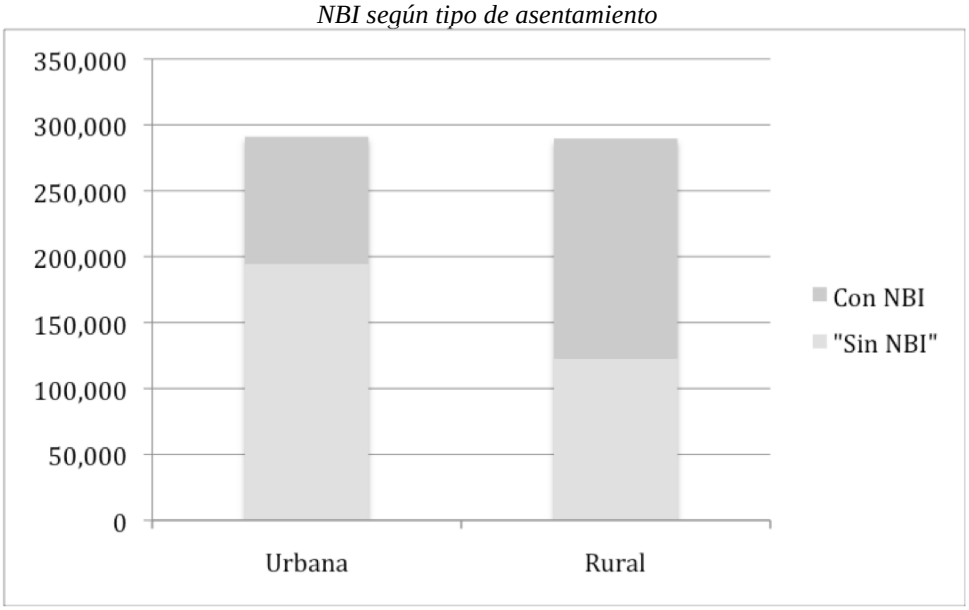
[12] En el caso particular de la tabla que nos ocupa, ambos grupos tienen aproximadamente el mismo peso, por lo que podría calcularse: (33,2 + 57,6) / 2 = 45,4%.

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

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

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

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



Fuente: Tabla 4.2.1

Primera regla de Zeisel

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

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

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

¿Cuál es la variable independiente?

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

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

---

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

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

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

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

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

Fuente: Tabla 4.2

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

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

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

Segunda regla de Zeisel

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

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

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

---

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

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

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

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

Fuente: elaboración propia.

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

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

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

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

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

Conformidad con el examen final	*Carrera*		*Total*
	Turismo	*Trabajo Social*	
Sí	30	70	100
No	58	42	100
			(n = 80)

Fuente: Tabla 4.3

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

Supongamos que de haber trabajado con el universo, se hubiera obtenido las mismas proporciones de conformistas en ambas carreras que las registradas en la Tabla 4.3. Los resultados serían los presentados en la Tabla 4.3.2.[21]

---

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

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

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

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

Fuente: elaboración propia.

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

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

¿Y el modo simétrico?

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

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

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

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

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

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

Fuente: Tabla 4.2.

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

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

2. EL ANÁLISIS DE LA RELACIÓN ENTRE VARIABLES

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

2.1 La diferencia porcentual: una medida de la asociación

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

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

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

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

participación política.[24]

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

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

Fuente: elaboración propia.

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

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

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

Fuente: elaboración propia.

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

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

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

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

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

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

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

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

Fuente: elaboración propia.

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

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

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

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

Fuente: elaboración propia.

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

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

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

¿Es esta relación estadísticamente significativa?

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

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

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

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

	B	A	
A			19
B			102
	65	56	121

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

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

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

	B	A		**a. Máxima**
A	0	19	19	**relación**
B	65	37	102	**posible**
	65	56	121	

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

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

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

¿Qué entendemos por 'ausencia de relación'?

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

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

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

$$10/65 = 9/56 = 19/121.$$[29]

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

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

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

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

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

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

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

---

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

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

[30] Simple transformación de la ecuación, multiplicando ambos términos por la expresión '(a + c)'.

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

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

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

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

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

¿Cómo se calcula x^2?

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

$$x^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

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

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

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

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

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

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

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

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

$$x^2 = 4{,}450$$

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

Tabla 4.6: Tabla resumida de valores críticos para x2

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

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

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

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

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

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

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

---

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

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

[33] En las tablas de los manuales se encontrará ‘df’, del inglés ‘*degree of freedom*’.

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

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

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

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

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

si x^2 > VC, se concluye que la relación es estadísticamente significativa.

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

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

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

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

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

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

¿Cuándo se puede usar x^2?

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

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

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

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

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

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

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

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

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

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

2.3. Algunos coeficientes de asociación

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

¿Qué es un coeficiente de asociación?

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

Coeficientes derivados de x^2

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

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

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

Con los datos de la Tabla 4.4 obtendríamos:

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

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

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

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

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

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

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

'Min (h - 1; c - 1)' significa aquí que se tomará o bien 'h - 1', o bien 'c -1', el que resulte menor de estos dos números. Así, para una tabla de 6 x 4, se tendría Min (h - 1; c - 1) = Min (6-1; 4-1) = 3, y por lo tanto:

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

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

Coeficientes para la tabla de 2 x 2

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

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

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

[35] La letra griega 'ϕ' se pronuncia 'Fi'.

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

siguientes:

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

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

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

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

¿Cómo interpretar el valor de un coeficiente?
Indicando '1,00' (o '-1,00') una correlación perfecta, se suele hablar de una 'correlación 'muy fuerte' para valores superiores a 0,80, de una 'fuerte' correlación entre 0,60 y 0,80, 'moderada' de 0,40 a 0,60, y 'débil' de 0,20 a 0,40 . Por debajo de 0.20, probablemente no exista correlación, y se trate simplemente de un producto del azar -al menos que se cuente con una muestra muy grande- (Fitz-Gibbon, 1987: 82). Si aceptamos este modo de interpretar los valores de nuestros coeficientes, nos encontramos con que, en tanto Q indicaría una asociación moderada entre las variables, de acuerdo a ϕ no existiría ni siquiera una asociación débil. El valor de Q es mucho mayor que el de ϕ. De hecho, esto ocurrirá toda vez que calculemos ambos coeficientes; es una **ley**: para cualquier tabla de 2 x 2 se cumple siempre que $Q > \phi$.[37]

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

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

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

Ejemplo de asociación completa

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

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

---

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

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

Ejemplo de asociación absoluta

		Atributo X: B	Atributo X: A	
Atributo Y	A	0	70	70
	B	30	0	30
ϕ= 1,00		30	70	100

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

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

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

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

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

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

¿Qué hacer con el signo de ϕ y Q?

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

---

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

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

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

	B	A
A	n	N
B	N	n

Asociación positiva

	A	B
A	n	N
B	N	n

Asociación negativa

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

Dicotomizar abusivamente es riesgoso

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

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

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

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

Fuente: elaboración propia.

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

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

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

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

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

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

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

Fuente: elaboración propia.

El coeficiente Gamma de Goodman y Kruskal

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

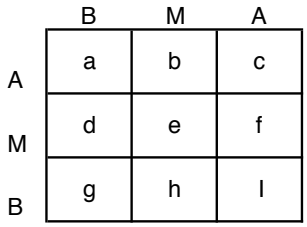
La fórmula para el cálculo de γ es la siguiente:

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

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

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

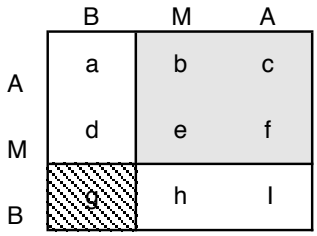
[44] Una exposición sistemática sobre todos los coeficientes para tablas de h x c está más allá de los objetivos de este trabajo de carácter introductorio. El lector interesado podrá consultar a Galtung (1966: II), y Zelditch (1959), así como los trabajos más recientes de Hildebrand, Laing y Rosenthal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

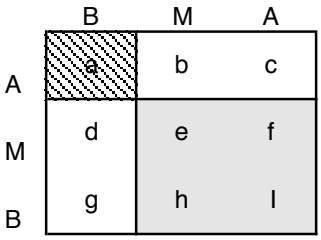


Comparando dos casos cualesquiera ubicados en distintas celdas pueden darse tres situaciones: (1) que sus órdenes sean iguales (cualquier caso en g será más bajo en ambas variables que un caso de e, b, c o f); (2) que estén parcialmente empatados (casos ubicados en una misma hilera –como a y b, o en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

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

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





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

Podemos ahora aplicar γ a un ejemplo.

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

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

Fuente: elaboración propia.

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

Desarrollemos el cálculo de n_i :

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

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

Con análogo procedimiento calculamos los n_d:

1 (49 + 26 + 11 + 5) = 91
13 (26 + 5) = 403
15 (11 + 5) = 240
49 (5) = 245
Σn_d = 979

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

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

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

3. DISTRIBUCIONES MULTIVARIANTES CONJUNTAS

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

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

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

---

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

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

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

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

indicadores agrarios construidos en base a datos censales.

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

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

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

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

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

Fuente: Bartolomé, L., 1982: 33.

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

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

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

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

[49] Y no simplemente para trabajarlos.

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

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

¿Por qué no utilizar la tabla de contingencia ?

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

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

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

Fuente: Tabla 4.9

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

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

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

El coeficiente ρ de Spearman

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

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

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

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

Departamentos	*Nº núcleos en 1973 (I)*	*% explot. 5-50 has. (II)*	*Rango en (I) (III)*	*Rango en (II) (IV)*	*d (V)*	*d2 (VI)*
Oberá	44	17,7	1	2	-1	1
Caingúas	37	18,9	2	1	1	1
Leandro N. Alem	20	11,3	3	3	0	0
Apóstoles	17	4,6	4	9	-5	25
Montecarlo	16	3,0	5	11	-6	36
San Ignacio	15	7,1	6	4	2	4
L.G. San Martín	12	5,9	7,5	7	0,5	0,25
Veint. de Mayo	12	6,7	7,5	5	2,5	5,25
Guaraní	10	4,0	9	10	-1	1
San Javier	8	4,9	10	8	2	4
San Pedro	6	1,5	11	14,5	3,5	12,25
Eldorado	5	6,1	12	6	6	36
Iguazú	3	1,4	13	16	-3	9
Candelaria	*0*	*2,9*	*15,5*	*12*	*3,5*	*12,25*
Concepción	0	1,5	15,5	14,5	1	1
General Belgrano	0	1,7	15,5	13	2,5	6,25
Capital	0	0,8	15,5	17	-1,5	2,25
				Total	0	157,5

Fuente: Tabla 4.9

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

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

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

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

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

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

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

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

Aplicando esta fórmula a nuestro ejemplo, se obtiene:

$$\rho = 1 - \frac{6x157,50}{17(17^2 - 1)} = 1 - 0,193 = 0,807$$

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

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

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

El coeficiente r de Pearson

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

$$y = a + bx$$

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

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

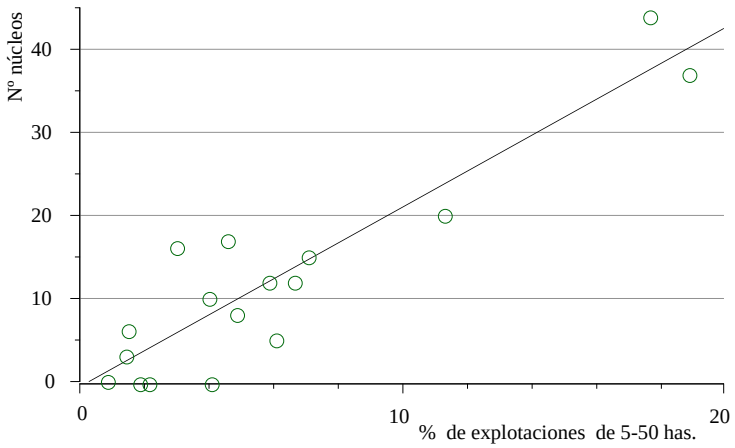
Adjudicando en esta ecuación cualquier valor a x, se obtiene el valor previsto para Y.

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

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

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

Figura 4.2: Diagrama de dispersión de los departamentos y recta de regresión para las variables Nº de nuúcleos del MAM y % de explotaciones de 5-50 has.



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

$$r = 0{,}927$$

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

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

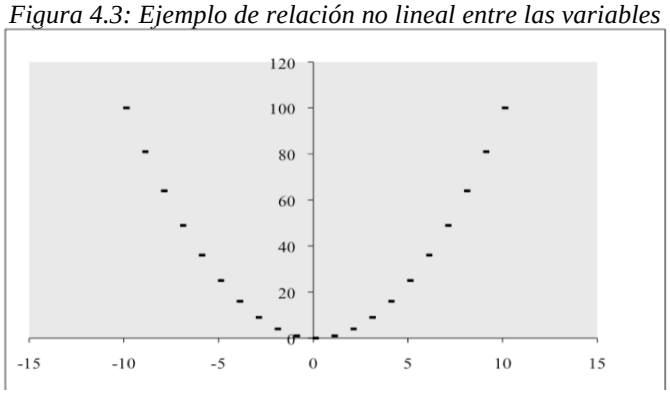
$$r^2 = 0{,}859$$

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

Examinar siempre primero el diagrama de dispersión

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

Figura 4.3: Ejemplo de relación no lineal entre las variables



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

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

¿Es legítimo inferir relaciones de un nivel a otro?

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

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

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

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

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

La matriz de correlaciones

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

---

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

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

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

Tabla 4.9.3: Matriz de correlaciones entre cuatro indicadores

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

Fuente: Tabla 4.9

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

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

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

Para las tablas de contingencia, el análisis multivariable encuentra sus fundamentos en el clásico trabajo de Lazarfeld (1968). En lo que hace a variables intervalares, existe una variedad de técnicas que se adaptan a diferentes situaciones. Así, por ejemplo, para analizar la relación entre una variable intervalar dependiente, y varias variables independientes de nivel nominal y/o ordinal, se puede recurrir al análisis de variancia (Iversen y Norpoth, 1976); si en cambio tanto las variables independientes como la dependiente son intervalares, el análisis de regresión múltiple puede ser una alternativa adecuada. Recomendamos la lectura de los buenos manuales[58] para familiarizarse con estas técnicas un poco más complejas. Por último, y siempre dentro de las técnicas estadísticas clásicas, también se puede recurrir al análisis factorial (Kim y Mueller; 1978a y 1978b) para una multiplicidad de propósitos. En el capítulo 6 nos ocuparemos de un tipo particular de técnicas factoriales más recientes que resultan especialmente útiles para la investigación sociológicamente orientada.

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

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

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

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

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

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