en un par de ejes cartesianos, sino que toma la figura de un conjunto ordenado de celdas, cada una de las cuales representa una determinada combinación de valores o categorías.

Reducciones de espacios de propiedades cualitativas

Supongamos que estamos realizando una investigación en el campo de la Sociología del Trabajo y que, inquiriendo a un conjunto de obreros acerca de su satisfacción con la tarea que realizan, hemos llegado a diferenciarlos -mediante el uso de algún tipo de índice o de escala- según su satisfacción con la tarea sea Alta (A), Media (M) o Baja (B). Al mismo tiempo, hemos preguntado a los mismos obreros si se consideraban satisfechos con el salario que percibían, obteniendo una clasificación análoga en Alta, Media y Baja satisfacción con la remuneración. Si ahora nos interesa ver cómo se distribuyen estas unidades (los obreros) simultáneamente en las dos clasificaciones, la combinación de ambas variables dará lugar a un espacio de propiedades de nueve (3 x 3) **variantes**. No por sencillo resulta menos esencial captar que al conformar ese conjunto de nueve variantes lo que hicimos fue generar una nueva clasificación: cada obrero encuestado deberá necesariamente encontrar su ubicación en una celda y solamente en una celda. Para decirlo con toda claridad, lo que ha ocurrido es que, combinando dos variables (cuyos valores originales eran, en ambos casos, A, M y B), **hemos creado una nueva variable**,

Figura 5.2: Ejemplo de espacio de propiedades cualitativas

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

T10

con los valores AA, AM, AB, MA, MM, MB, BA, BM y BB[7].

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

Reducción por simplificación de dimensiones: sin duda, el caso más obvio. Es sabido que cualquier variable es susceptible de ser reducida a una dicotomía. Si entonces reducimos

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

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

[9] BARTON hace referencia a un cuarto tipo: la reducción **funcional**: ésta es la que se aplica en el análisis de escalograma de GUTTMAN (cf. *infra*: sección 2 de este capítulo) y en las distintas técnicas de análisis factorial (algunas de las cuales serán abordadas en el capítulo 6).

el número de valores de una o más de las dimensiones que conforman un espacio de propiedades, se reducirá necesariamente el número de celdas de dicho espacio.

Figura 5.2.a: Reducción por simplificación de dimensiones

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

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

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

Figura 5.2.b: Reducción numérica

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

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

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

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

4 ó 3 puntos ⇨ Alta satisfacción con el trabajo
2 puntos ⇨ Mediana satisfacción con el trabajo
1 ó 0 puntos ⇨ Baja satisfacción con el trabajo

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

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

2 (salario) + 0 (tarea)
1 (salario) + 1 (tarea)

UUAA -personas- serían las unidades de muestreo de última etapa.

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

Figura 3.1: Resumen de los principales tipos de muestras

No-probabilísticas	Probabilísticas
Casual Por cuotas	Al azar simple Sistemática Estratificada Por conglomerados Por etapas múltiples

T10

Capítulo IV

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 este 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.

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

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

"n" representa el **total** de unidades de análisis incluidas en la muestra, lo que se suele denominar "la frecuencia de

[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.

orden cero". Por su parte, "-X", "X", "Y" y "-Y" son las **frecuencias marginales** o de **primer orden**; así, por ejemplo, "-X" representa el número total de casos que no presenta el atributo X, independientemente de que posean o no el atributo Y. Por último, "-XY", "XY", "-XY" y "X-Y" representan las **frecuencias condicionales**, o de **segundo orden**; de este modo, "-XY" significa el número absoluto de observaciones que combinan la ausencia del atributo X con la presencia de Y. Es importante notar que:

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

Aplicación del modelo a un ejemplo

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

Tabla 4.2: Misiones, 1980 - Población según tipo de asentamiento y pertenencia a hogares con Necesidades Básicas Insatisfechas

Tipo de asentamiento	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:

Tabla 4.2.a: Misiones, 1980 - Población según tipo de asentamiento

Tipo de asentamiento	Número de habitantes
Urbano	291007
Rural	289515
Total	580522

Fuente: Tabla 4.2.

¿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. Pero, considerando intrínsecamente los datos de la Tabla 4.2.a, la manera de apreciar la importancia de esas 1.492 personas de diferencia es poniéndolas en relación con el total de la población provincial.

Tabla 4.2.a1: Misiones, 1980 - Población según tipo de asentamiento (%)

Tipo de asentamiento	Número de habitantes	%
Urbano	291007	50,1
Rural	289515	49,9
Total	580522	100,1

Fuente: Tabla 4.2.

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:

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

¿Cuándo redondear los porcentajes?

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

291.007/580.522 = 50/100 = 50%

Matemáticamente, estas son expresiones equivalentes -o casi- pero es evidente que en un sentido psicológico "50%" es la manera más concisa, sencilla y ventajosa de denotar la relación que nos interesa. Si se conservan muchos decimales, sólo se logra tornar más engorrosa la lectura de la tabla y se pierde la ventaja de expresar las cifras en porcentajes. Por ende se puede recomendar, siempre que sea ello posible, **como regla general, prescindir totalmente de los decimales**. "50,12850503" parece más preciso que "50 %"; mas 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 4.2.a.1 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

[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".

[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).

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**

[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.

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 con 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 el 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.

¿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 No	Hogares con NBI Sí	Total
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 cuales 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:

[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 hilera: "291.007", en este caso;
- 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"; y

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

[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...

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"; y

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.

¿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;

[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).

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 de-

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

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

[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%.

[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 probable que muchas personas percibirán mejor una relación expresada visualmente, aunque se pierda algo de precisión con respecto a los datos numéricos. En este caso, el agregado de los porcentajes dentro de cada columna permite conservar lo esencial de la información numérica.

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.

[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: Turismo		Carrera: Trabajo Social		Total	
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 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%).

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 POBLACIÓN, LOS PORCENTAJES DEBEN COMPUTARSE EN LA DIRECCIÓN 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

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

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: Turismo	Carrera: Trabajo Social	Total
Sí	30	70	100
No	58	42	100

(n = 80)

Fuente: Tabla 4.2.

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].

[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: Turismo	Carrera: Trabajo Social	Total
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 4.3.3: Misiones, 1980 - Distribución de la población por tipo de asentamiento y pertenencia a hogares con NBI

Tipo de asentamiento	Hogares con NBI No	Hogares con NBI SÍ	Total
Urbano	33,5	16,6	50,1
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 el 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.

[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".

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 participación política[24].

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

Participación política	Conocimiento político: Bajo	Conocimiento político: Alto	Total
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:

[23] Los datos provienen del estudio "Participación política del estudiante de la FHyCS-UNaM" (inédito) realizado por alumnos de Antropología Social en 1984.
[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, este es un ejemplo de relación **reversible e interdependiente** entre las variables (1968: 59 y ss.).

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

Participación política	Conocimiento político: Bajo	Conocimiento político: Alto	Dif. %
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 DIRECCIÓN EN QUE SE REALIZA LA COMPARACIÓN.

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].

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

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

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

Participación política	Conocimiento político Bajo	Alto	Total.
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.

La Tabla 4.5 permite ilustrar este problema[26]. Puesto que

[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.

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.

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.

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 un 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 FHyCS-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 pueden 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:

a. Máxima relación posible

	B	A	
A	0	19	19
B	65	37	102
	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 total de relación

	B	A	
A	10	9	19
B	55	47	102
	65	56	121

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

¿Qué entendemos por "ausencia de relación"?

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

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

[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).

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