UNIVERSIDAD DEL SALVADOR
BUENOS AIRES, ARGENTINA

FACULTAD DE CIENCIAS SOCIALES

DEPARTAMENTO DE COMPUTACIÓN
PARA CIENCIAS SOCIALES

Director Lic. Ricardo Murtagh

APUNTES Nº 2

La relación entre variables: construcción y análisis de tablas de contingencia

Horacio Chitarroni Maceyra

marzo 1996

INDICE

1- INTRODUCCIÓN

Frecuentemente, tanto cuando pretende poner a prueba hipótesis (que implican relaciones esperadas entre ciertas variables), como cuando simplemente procura describir el universo bajo análisis, el cientista social se ve enfrentado a la tarea de construir **tablas de contingencia**. Ellas no son otra cosa que los cuadros de doble entrada: un *"espacio de propiedades"* determinado por dos variables (*Barton*, 1973).

Las técnicas de análisis cuantitativo de datos - potenciadas además por el uso de las computadoras y de programas adecuados - proveen un sinnúmero de instrumentos, algunos de ellos de extrema sofisticación. Sin embargo, el análisis de cuadros estadísticos, la correcta porcentualización de los mismos y su interpretación, constituyen una suerte de herramental básico e imprescindible, que aproxima a la comprensión de la lógica implícita en las relaciones entre variables. Sin el dominio de esta "herramienta clásica", todas las otras técnicas avanzadas devienen ciegas. Y con todas sus limitaciones en cuanto a la incorporación simultánea de muchas variables al análisis y su tratamiento conjunto - no ha de olvidarse que *Emile Durkheim* produjo *"El suicidio"* (aun hoy un clásico de la investigación empírica capaz de realimentar la teoría) casi sin disponer de otro elemento que las tasas y los porcentajes, sabiamente utilizados. Es que nada sustituye a la reflexión acerca de las relaciones teóricas escondidas tras los datos, la cual - no pocas veces - resulta más estimulada por el análisis "artesanal" de los mismos que por el procesamiento mecánico a través de los instrumentos cuantitativos más potentes (sin que esto implique, en modo alguno, desdeñar le utilidad de estos últimos). Aquí nos proponemos sintetizar algunas reglas básicas que deben guiar la construcción y el análisis de las tablas bivariadas, reuniendo elementos dispersos en la bibliografía metodológica corriente (en la que no siempre se presta la necesaria atención y espacio a esta temática).

2- DE LAS MATRICES DE DATOS A LAS TABLAS DE CONTINGENCIA

2.1- La matriz de datos

La *matriz de datos* contiene el resultado - totalmente desagregado - de la tarea de recolección efectuada en una investigación empírica. En ella se encuentran presentes los tres elementos que constituyen la *"estructura tripartita del dato"* (*Galtung*, 1978): las *unidades de análisis*, las *variables* y los respectivos *valores* que cada una de las primeras asumen al ser clasificadas o medidas a través de las segundas.

VAR. Un. Anál.	X	Y	...	Z
1	$r_{.x1}$	$r_{.y1}$	...	$r_{.z1}$
2	$r_{.x2}$	$r_{.y2}$	...	$r_{.z2}$
.	...	...	...	...
.	...	...	...	...
n	$r_{.xn}$	$r_{.yn}$	...	$r_{.zn}$

(Donde "$r_{.x1}$" es el valor que asume la unidad de análisis. "1" en la variable "x")

Cuando se han recolectado datos acerca de una cierta cantidad de variables sobre un número elevado de unidades de análisis (como es la regla en los diseños de tipo cuantitativo), resulta imposible analizar la matriz en sí misma: ella nos dice muy poco. La estadística provee instrumentos que permiten reducir los datos, resumir la información, tales como los porcentajes y tasas, las medidas de tendencia central y las de dispersión (entre otros): se trata de herramientas propias del análisis univa-

riado y hacen posible apreciar la distribución de las unidades de análisis en cada una de las variables por separado. Esta tarea corresponde a una primera etapa del análisis.

Pero casi siempre interesa indagar posibles relaciones (de determinación o de interinfluencia) entre las variables: ya porque contamos con presunciones emergentes de la teoría (hipótesis) acerca de las posibles relaciones, ya porque - careciendo de ellas - procuramos descubrirlas, procediendo inductivamente desde los datos. Allí es cuando debemos "cruzar" las variables en los cuadros de doble entrada, para apreciar la distribución de las unidades de análisis en las celdas determinadas por la combinación de ciertas categorías de una de las variables implicadas, con ciertas categorías de la otra.

2.2- La construcción de las tablas

En términos puramente operativos, el pasaje de la matriz a la tabla de contingencia que cruza "X" con "Y", supone el recorrido, fila por fila, de la primera, identificando la combinación de categorías que asume cada unidad de análisis en las variables seleccionadas, con la finalidad de agruparlas en la celda correspondiente del cuadro: así, todas las unidades de análisis que tengan valor $X._1$ en una variable e $Y._1$ en la otra, caen en la primera celda del cuadro, arriba y a la izquierda:

X Y	$X._1$	$X._2$	...	$X._n$
$Y._1$	$X._1Y._1$			
$Y._2$				
...				
$Y._n$				

El número natural resultante de la cantidad de unidades de análisis que cumplen tal condición es la **frecuencia absoluta** de la celda.

Esta operación supone, asimismo, dos tipos de decisiones: una de ellas, más teórica, consiste en la selección de las variables a cruzar. En este sentido - ya se ha dicho - la guía fundamental es la teoría y las hipótesis que de ella emergen. En ausencia de estas, el sentido común o aun la intuición, sugieren los cruces más aptos (aun en el proceder inductivo, es poco económico cruzar "todo con todo").

La otra, de orden más empírico, consiste en determinar el número de categorías que hemos de asignar a cada variable en el cuadro. Hay variables que son naturalmente dicotomías (como el sexo), pero otras (la edad, los ingresos) pueden asumir una multiplicidad de valores, forzando a construir intervalos de clase. Algunas (como el nivel de instrucción formal) poseen categorías preestablecidas (sin instrucción; primaria incompleta; primaria completa; secundaria incompleta; etc.) que pueden eventualmente fusionarse para reducirlas a un número menor: hasta primaria incompleta; hasta secundaria incompleta; secundaria completa y más. Ello, cuidando de preservar los criterios de exclusión y exhaustividad.

- Un número reducido de categorías (en realidad, cualquier reducción o fusión de ellas que se haga) supone perder información: se iguala lo desigual a los efectos del análisis . Asimismo, el corte de cualquier variable implica cierta arbitrariedad y debiera estar justificado teóricamente (*Cicourell*, 1982). Sin embargo, la ciencia no puede prescindir de la abstracción, de la simplificación de la realidad que supone cualquier clasificación.

- Pero un número muy elevado de categorías (un cuadro de muchas celdas) torna difícil y aun imposible el análisis: no permite detectar tendencias en la distribución.

- Aunque las dicotomías (establecer sólo dos valores para cada variable) son fáciles de manejar estadísticamente, suelen ocultar información reveladora: como se verá más adelante, no es posible apreciar las distribuciones curvilíneales. Una solución sugerida por algunos autores (*Galtung*, 1978; *Mora y Araujo*, 1965) consiste en apelar a las tricotomías (tres categorías) cuando ello sea posible. Y resulta aceptable - como regla general - no excederse de cinco o seis. Salvo cuando existan buenas razones teóricas para hacerlo y a condición de que el tamaño de la muestra (el "N" del cuadro) sea grande: con pocas unidades de análisis en un cuadro de muchas celdas, éstas tienden a vaciarse. Y poco se puede afirmar , desde el punto de vista estadístico, con frecuencias muy reducidas.

3- LA POSICIÓN EN LA TABLA: EL ORDEN CAUSAL DE LAS VARIABLES

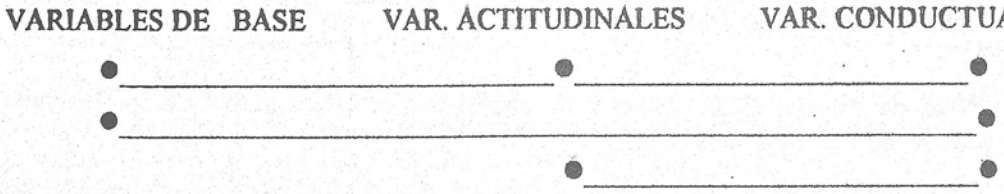
Seleccionadas las variables a cruzar y establecidas sus categorías (el tamaño del cuadro), debe decidirse la posición que ellas asumirán en la tabla. En principio, esta decisión se relaciona con el status lógico que les asignamos: existe una regla convencional - pero que encierra un fundamento lógico - según la cual corresponde ubicar las categorías de la variable más independiente (presunta causa: "X") en los cabezales de las columnas, y las categorías de la dependiente (presunto efecto: "Y") en las filas o renglones del cuadro:

X Y	X_1	X_2	X_3	
Y_1				MY_1
Y_2				MY_2
Y_3				MY_3
	STX_1	STX_2	STX_3	N

De este modo, en la última columna hallaremos las sumatorias de frecuencias que corresponden a cada categoría de Y (prescindiendo de la clasificación por X): se trata de las **frecuencias marginales**, que constituyen la distribución univariada de la muestra o universo según la variable "Y". Mientras que en la última fila tendremos las sumas de frecuencias que corresponden a cada categoría de "X" (prescindiendo ahora de la clasificación por "Y"): tales las llamadas **frecuencias subtotales**, que implican la distribución univariada de todas las unidades de análisis según la variable "X". El "N" total del cuadro resulta tanto de la suma (vertical) de las frecuencias marginales como de la suma (horizontal) de las subtotales.

Al interior del cuadro, en cada celda, tendremos las **frecuencias condicionales**: estas sí, suponen una clasificación bivariada de las unidades de análisis, según su ubicación simultánea en X e Y.

Ahora bien, la posición de las variables en el cuadro no ofrece dificultades cuando contamos con hipótesis que establecen relaciones de determinación entre las mismas (relaciones de "asimetría"). Pero no resulta tan clara cuando sólo estamos en condiciones de suponer la existencia de alguna interdependencia o variación conjunta, o cuando nada decimos del modo en que puedan relacionarse. En tales casos - y como una norma muy general - podemos servirnos de ciertas reglas lógicas sugeridas por *Galtung* (1978): las *variables de base o status* son - en general - determinantes de las variables *actitudinales o de personalidad*, y estas lo son de las *variables conductuales* (que implican conductas: acciones u omisiones).



Así, se podría hipotetizar que la posición socioeconómica (variable de base) tiende a determinar cierto tipo de actitud política que, a su vez, se manifiesta en el voto por un cierto partido (que es una conducta). La hipótesis podría relacionar directamente la clase social con el voto, sin interesarse por la actitud. O bien esta última con el proceder electoral, etc. Pero no tendría sentido suponer que el voto por un partido de extrema derecha influye sobre la actitud o la personalidad: será, en todo caso, su resultante. Sin embargo, esto no es más que un modelo de análisis causal muy esquemático y orientativo: sí puede conjeturarse que cierto tipo de actitud o personalidad (por ejemplo la competitividad) habilita para alcanzar status laborales y socioeconómicos más elevados (aquí, una variable actitudinal determina a una de base).

Pero suele ocurrir frecuentemente que relacionemos variables de base entre sí. En este caso, será lógico suponer que las que implican *status adscriptos* sean independientes o determinantes de las que suponen *status adquiridos*: la posición socioeconómica de la familia de origen podrá ser determinante del nivel educativo alcanzado por un individuo y no al revés. Y cuando se trata de dos variables de status adquirido, importará tener en cuenta la prelación temporal: el nivel educativo de las mujeres (en general, aunque no siempre, se completa antes de la maternidad) tenderá a influir sobre el número de hijos. De igual modo, el nivel educativo también determinará la posición ocupacional en la mayor parte de los casos.

En términos generales - y aun cuando no hay reglas fijas - la combinación de este conjunto de criterios con la reflexión teórica y el sentido común suele dar buenos resultados.

Finalmente, cuando se cruzan variables tales como el sexo y la edad (con fines meramente descriptivos, puesto que no es razonable suponer determinación alguna entre ellas), la ubicación dependerá de nuestras preferencias y de los propósitos del análisis.

Conviene destacar que estas reglas básicas referidas a la ubicación de las variables no siempre son observadas rigurosamente: es frecuente encontrar cuadros construidos sin tomarlas en cuenta. Sin embargo, conviene atenerse a ellas cuando elaboramos nuestras propias tablas.

4- ELEMENTOS DE LOS CUADROS

Idealmente, los cuadros deben contar con algunos elementos básicos que facilitan en mucho su análisis e interpretación. Aunque vale la advertencia que cierra el punto anterior (no siempre se incluye la información necesaria en la forma correcta), es importante tomarlos en cuenta.

4.1- Titulación de los cuadros

El título de un cuadro debería permitir identificar fácilmente las unidades de análisis a las que se refiere y las variables implicadas. Estas últimas, mencionando en primer lugar la variable dependiente , en segundo término la independiente y por último la segunda variable independiente o bien la que opere en posición de "control", si las hubiera (nos referiremos más adelante a los cuadros que incluyen más de dos variables).

EJEMPLOS:

"Jefes de familia del Gran Buenos Aires (U.A.), por condición de ocupación (V.D.), según edad (V.I.)"

"Nivel de ingresos (V.D.) según nivel educativo (V.I.), por sexo (V.C.) (Habitantes de la Capital Federal, de 10 y más años de edad)"

"Voto emitido (V.D.) , por nivel socioeconómico (V.I.) y edad (V.C.) (Votantes masculinos de la Provincia de Córdoba)"

"Migrantes limítrofes (U.A.), según nivel educativo (V.D.) y sexo (V.I.)"

"Salario percibido (V.D.), según rama de actividad (V.I.)"

Puede suceder - como en el último de los ejemplos ofrecidos - que no aparezca mencionada en el título la unidad de análisis. En caso de ser así, tal información debiera figurar al pie del cuadro, junto a la fuente de donde provienen los datos (por ejemplo: "Asalariados del Gran Buenos Aires. Datos muestrales").

4.2- Otros elementos de los cuadros

Otros elementos informativos deben ser proporcionados al presentar un cuadro:

- la fuente de donde provienen los datos ("Encuesta Permanente de Hogares - INDEC")

- la fecha y el lugar del relevamiento ("Partido de La Matanza - abril de 1993")

- la indicación acerca de si los mismos se refieren a una muestra ("Datos muestrales")

- la unidad de medida en que están expresados (Porcentajes, cifras absolutas, pesos constantes de 1970, u$s.)

5- LECTURA E INTERPRETACIÓN DE LOS CUADROS

5.1- Las reglas de la porcentualización

La ya aludida convención que aconseja situar la variable independiente en las columnas y la dependiente en las filas se relaciona con la lógica que guía la porcentualización de los datos. Vale decir, la transformación de **frecuencias absolutas** en **frecuencias relativas**. Dicha lógica indica que - a los fines de observar el comportamiento de la variable dependiente cuando ella es sometida a los distintos valores de la variable independiente, los porcentajes deben obtenerse siempre tomando como base los totales de las columnas: esto es, sobre las que hemos designado como **frecuencias subtotales** (que suponen la frecuencia de cada categoría de "X" en la distribución univariada). Ello será así para cada categoría de "X" (la variable independiente), pero también al porcentualizar los **marginales**, es decir la última columna del cuadro que está dada por la frecuencia de cada categoría de "Y" (la variable dependiente). En este último caso, la base será el total de esta última columna, que no es otra cosa que el "N" del cuadro. Sólo para el caso de las frecuencias marginales se empleará como base este total: si se procediera así con las frecuencias interiores (condicionales) no estaríamos agregando ninguna información nueva a la que nos proporcionan las cifras absolutas. Por el contrario, la porcentualización con base en los subtotales logra - precisamente - igualar las bases de comparación entre los grupos o subuniversos en que ha quedado dividido nuestro universo (o muestra) determinados por las categorías de la variable independiente.

Así como la porcentualización se realiza, entonces, en sentido vertical (en el sentido de la variable independiente), la "lectura" o comparación de los porcentajes (que se denominan "frecuencias relativas") debe efectuarse horizontalmente:

- En primer término, será útil comparar las distribuciones relativas (%) de la variable dependiente al interior del cuadro, con la que corresponde al marginal. Esta primera comparación responde a la pregunta más general: ¿parece haber algún tipo de relación entre las variables que se han cruzado?. En efecto, en la medida en que la distribución de "Y" para las distintas categorías de "X" (frecuencias relativas condicionales) sea muy semejante a la que se registra en el total de la muestra o universo (frecuencias relativas marginales), ello será una primera indicación de ausencia de vinculación entre las variables. Por el contrario, el hecho de que el desagregado por la variable "X" tenga por efecto una diferente distribución de los valores de "Y", sugiere que ambas distribuciones no son independientes.

- En segundo término, pasaremos a comparar las frecuencias relativas condicionales entre sí. Esto da respuesta a una cuestión más sustantiva: ¿de qué modo se relacionan las variables? ¿qué valores de "X" tienden a coincidir con qué valores de "Y"?. Esto, la particular "forma " de la relación, queda gráficamente plasmada en el "espacio de propiedades" (*Barton*, 1973) que constituye el cuadro y no es discernible - paradojalmente - mediante otra clase de instrumentos estadísticos más sofisticados (del tipo de los *coeficientes de contingencia* o las *pruebas de significación*). La comparación horizontal de los porcentajes dentro de cada fila (categoría de "Y") y entre cada columna (categoría de "X") permite observar, pues, cómo varía el peso relativo de los individuos que asumen un cierto valor en "Y" para las distintas categorías de "X". La idea básica que está detrás de la lógica de la porcentualización es afín a la lógica del *diseño experimental*: los grupos que comparamos resultan de la división de la muestra (o universo) en segmentos que se diferencian entre sí en que - en ellos - opera con desigual intensidad la variable independiente (estímulo). Al comparar horizontalmente, observamos los efectos diferenciales sobre la variable dependiente.

A esta altura, conviene emplear un ejemplo: supongamos que contamos con datos provenientes del Censo Nacional 1991 para un partido del Gran Buenos Aires con elevada proporción de migrantes internos y limítrofes. Nos interesa cruzar esta variable (la *condición migratoria*) con el *nivel educativo formal*: detrás de esta elección, subyace el supuesto de que los nativos poseen un nivel educativo superior al de los migrantes.

Partido de ... : Población de 10 años y más, según nivel de instrucción formal, por condición migratoria (en miles de habitantes)

Condic. migratoria Nivel educativo	Nativos	Migrantes internos	Migrantes limítrofes	Migrantes no limítrofes	Total
H/ prim. incompleta	72.6	104.1	15.8	5.0	197.5
H/ secund. incompleta	210.1	191.8	18.7	29.4	450.0
Sec. completa y más	99.3	29.3	1.5	1.4	131.5
Totales	382.0	325.2	36.0	35.8	779.0

Fuente: datos ficticios

Por cierto que, en este caso, la condición migratoria será la variable más independiente. No por ser - seguramente - *causa* del nivel educativo alcanzado: este, sin duda, depende de un conjunto de factores. Pero con bastante probabilidad la condición migratoria se contará entre tales factores y estará asociada a un buen número de ellos (por ejemplo, el nivel socioeconómico). Se puede conjeturar razonablemente que ejercerá influencia sobre la instrucción formal: la situamos, pues, en las columnas del cuadro. El siguiente paso consiste, entonces, en obtener frecuencias relativas (%), tomando como base los subtotales:

Condic. migratoria Nivel educativo	Nativos	Migrantes internos	Migrantes limítrofes	Migrantes no limítrofes	Total
H/ prim. incompleta	19%	32%	44%	14%	25%
H/ secund. incompleta	55%	59%	52%	82%	58%
Sec. completa y más	26%	9%	4%	4%	17%
Totales	100% (382.0)	100% (325.2)	100% (36.0)	100% (35.8)	100% (779.0)

Las cifras entre paréntesis, bajo los subtotales, indican las bases sobre las que se obtuvieron los porcentajes y permiten reconstruir las frecuencias absolutas de las celdas: 19*382/100 = 72.6 para la primera celda de la primera columna; es conveniente incluir este valor de las frecuencias absolutas en todas las tablas, a fin de que el lector pueda, si fuera de su interés, reconstruir los valores absolutos, o saber de cuántos casos se está tratando. Estamos ahora en condiciones de emprender el análisis del cuadro a través de las frecuencias relativas:

- Mirando el marginal, podemos ver (y esto es un análisis *univariado*) que más de la mitad - un 58% - de los habitantes del partido considerado han completado el ciclo primario y aun cursado (ignoramos qué proporción de ellos) algunos años del secundario. La cuarta parte (un 25%) no alcanzó a completar la educación básica, mientras que un 17% al menos completó el secundario (tampoco sabemos cuántos de ellos cursaron, parcial o enteramente, educación superior).

- Una rápida observación permite apreciar que la distribución de los porcentajes en el marginal (vale decir, para la población total) difiere de la que aparece en las distintas columnas interiores: no es igual para los nativos ni para las distintas clases de migrantes. Esto basta para afianzar la idea de que hay alguna asociación entre las variables.

- Ahora podemos penetrar al interior del cuadro y comparar las frecuencias relativas de las celdas (*condicionales*) entre sí: esto sí ya supone análisis *bivariado*. La proporción de quienes no han completado la primaria (que incluye, también, a quienes no han recibido instrucción alguna) aumenta entre los migrantes internos y alcanza su máximo peso (44%) entre los provenientes de países limítrofes. Es mínima, en cambio, entre los migrantes no limítrofes (14%) y menor que el marginal en los nativos. El porcentaje de quienes han completado la educación elemental (y eventualmente cursado parcialmente la secundaria) no difiere demasiado entre nativos, migrantes internos y limítrofes, pero es muy elevada (82%) entre los no limítrofes. La categoría superior (secundaria completa y más) crece entre los nativos: 26% de ellos están en esta situación , mientras que la proporción cae a 9% entre los migrantes internos y a 4% para el resto.

- Efectivamente, podríamos sostener que los nativos están relativamente más educados que los migrantes internos y limítrofes: un 81% de aquellos han completado - al menos - el ciclo primario. Sólo el 68% de los migrantes internos y el 56% de los limítrofes obtuvo ese logro.

- Es posible ahora centrar la atención en las celdas en que aparecen frecuencias elevadas. De algún modo, nuestra hipótesis esperaba frecuencias relativas mayores (superiores a las del marginal) en algunas casillas del cuadro: las que combinan bajos niveles de instrucción con la condición de migrante. Se ve confirmada por el 32% y el 44%, respectivamente, de migrantes internos y limítrofes que no han completado la primaria. Asimismo, por el 26% de nativos con *secundaria completa y más*. En cambio, constituye un dato no previsto el 82% de no limítrofes con *hasta secundaria incompleta*. No es ya tan fácil decidir si ellos están menos educados que los nativos: es menor la proporción de los que no han completado el primer ciclo, pero también la de quienes han terminado el intermedio.

- Como se aprecia, estas conclusiones pueden obtenerse a partir de los porcentajes. Con frecuencias absolutas podríamos, hasta cierto punto, comparar a nativos y migrantes internos (cuyo número total no difiere en mucho). pero sería difícil decidir, a simple vista, si 104.100 migrantes internos con *hasta primaria incompleta* representan más o menos que 15.800 limítrofes que cumplen igual condición.

5.2- Variables ordinales: el "sentido" de la relación

Cuando las variables cruzadas no son meros criterios clasificatorios, sino que suponen *órdenes*, cobra importancia la *posición* de las celdas con más altas frecuencias. Supongamos que hemos elegido cruzar el *nivel de instrucción* con los *ingresos*. En tal caso, presumiblemente esperamos hallar que "*a mayor nivel de instrucción, mayores ingresos*". Entonces, construiremos la tabla haciendo converger la categoría más alta de cada variable en el vértice superior izquierdo:

NIVEL DE INSTRUCCION / INGRESOS	ALTO (Sec. completa y más)	MEDIO (Hasta sec. incompleta)	BAJO (Hasta prim. incompleta)
ALTOS (2000 y más)			
MEDIOS (700 hasta 1999)			
BAJOS (699 y menos)			

En el supuesto de que las frecuencias tendieran a concentrarse en las celdas sombreadas (la llamada *diagonal positiva*) diríamos que existe *asociación positiva* entre las variables, según lo suponía nuestra hipótesis. La intensidad de dicha asociación será mayor cuanto más concentración se verifique en dicha diagonal, y menor cuanto más dispersión exista (frecuencias altas en las celdas ajenas a ella). Inversamente, cuando las frecuencias se concentran en la *diagonal negativa* (desde arriba y a la derecha hacia abajo y a la izquierda) habrá *asociación de signo negativo*: a más alto valor en una variable, más bajo valor en la otra.

5.3- Forma de la relación

En el ejemplo precedente, la relación era *positiva en su sentido* y *lineal en su forma.* Pero, en ocasiones, las relaciones observadas no son diagonales o lineales sino *curvilineales*: supongamos que vinculamos el nivel socioeconómico con el tamaño familiar (medido por el número de hijos):

N. SOCIO ECON. CANT. DE HIJOS	ALTO	MEDIO	BAJO
5 Y MAS			
3 A 4			
HASTA 2			

Podría suceder que las celdas con frecuencias relativas más altas se dispusieran del modo indicado : habría allí una relación de tipo *curvilineal* (no observable en la diagonal del cuadro) y que no podría señalarse como "negativa" o "positiva": esta forma de relación puede presentarse, por supuesto, también con variables nominales (atributos o criterios clasificatorios): a diferencia del *sentido*, la apreciación de la *forma* no demanda ordinalidad. Es útil señalar que, en caso de haberse dicotomizado ambas variables (por ejemplo en "alto" y "bajo"), podrían neutralizarse las diferencias y aparecería un cuadro que no revelaría mayor vinculación entre el nivel socioeconómico y el tamaño familiar: se perdería de vista la relación curvilineal existente. Por ello, es que vale la recomendación formulada en el punto 2 acerca de la conveniencia de trabajar, al menos, con tres valores.

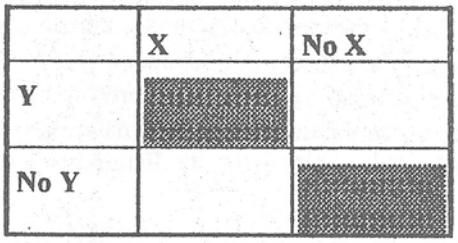
En el caso de las dicotomías, de todas maneras, también puede ocurrir que la concentración de frecuencias no se sitúe en la diagonal, sino en un rincón del cuadro. *Cortes y Rubalcaya* (1987) citan el ejemplo de la relación entre *pobreza* y *marginalidad*: la marginalidad (definida como generalizada ausencia de participación) casi siempre se relaciona con la pobreza, pero no a la inversa (los pobres pueden ser o no ser marginales).

POBREZA / MARGINALIDAD	SI	NO
SI		
NO		

En este caso tendería a vaciarse una celda del cuadro (no hay *no pobres* que sean *marginales*) y a cargarse intensamente otra. Se trata de una relación de tipo *rinconal.*

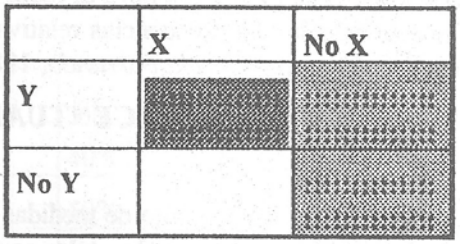
Diagonalidad: X .....Y

No X.....No Y



Rinconalidad: X ..... Y

No X ..... Y o No Y



5.4- La porcentualización descriptiva

Las reglas enunciadas acerca de la porcentualización en el sentido de la *variable independiente* pueden aparecer, eventualmente, transgredidas en algunos cuadros. A veces, sucederá que nos encontremos con porcentajes obtenidos horizontalmente: ello no altera - en lo fundamental - la lógica precedente a condición de que se haya rotado la posición de las variables, situando la independiente en las filas. Pero en ocasiones, los porcentajes aparecen calculados tomando como base los marginales (en el sentido de la *variable dependiente*). En este caso, la lectura porcentual tiene otro significado.

Intención de voto de los sectores sociales en Capital Federal - Elecciones legislativas (%)

Partido votado / Clase social	P. Justic.	U.C.R.	U.CeDé	F. Grande
Alta y M. Alta	5%	18%	50%	15%
M. Med. y M. Baja	24%	68%	50%	70%
Baja	31%	14%	-	15%
	100%	100%	100%	100%

Fuente: Diario "Clarín" - 21/8/93

No puede suponerse que el voto influye sobre la clase, sino al revés. Pero en el cuadro que precede se han calculado los porcentajes para cada agrupación política: en este caso se trata de una **porcentualización descriptiva** de la *composición de clase* del voto de cada partido. Revela, por ejemplo, el fuerte aporte de sectores medios al voto de la U.C.R. y el Frente Grande y la composición más popular del voto justicialista. Si lo que se quiere obtener es esta suerte de "perfil clasista" de las

distintas fuerzas políticas, es lícito proceder de este modo. Pero a condición de que se cumpla con una condición: la muestra deberá guardar estricta correspondencia con el universo en cuanto a la proporción de cada clase social. Si, por acaso, se hubieran sobrerrepresentado en ella los sectores altos (u otro cualquiera), aparecerían aportando una proporción mayor que la real al voto de cada fuerza política, distorsionando las conclusiones (*Zeizel*, 1974). Este inconveniente desaparece si se está trabajando con el universo total.

5.5- Frecuencias mínimas para calcular porcentajes

Es preciso tener en cuenta que no deben obtenerse porcentajes sobre bases muy pequeñas: no tiene sentido alguno decir que 1 es el 50% de 2 casos. El desplazamiento de una unidad de análisis de una a otra celda puede resultar por entero casual y no debiera significar una variación superior al 5% en términos de porcentajes (*Galtung*, 1978). Esto nos lleva al requisito de una frecuencia mínima de 20 casos en la base sobre la que se calculan las frecuencias relativas. Algunos autores son aún más conservadores y exigen no menos de 50 casos (*García Ferrando*, 1985).

6- EL USO DE LA DIFERENCIA PORCENTUAL COMO MEDIDA DE ASOCIACIÓN

Si bien la estadística provee un amplio conjunto de medidas capaces de evaluar la intensidad con que se relacionan dos variables cruzadas en un cuadro, tales como los *coeficientes de contingencia* (*Blalock*,1986; *García Ferrando*, 1985), es frecuente emplear - como medida básica de asociación - la **diferencia porcentual** (usualmente indicada como "D %", o bien con la letra griega épsilon: ε). Cuando la variable independiente es dicotómica, consiste en restar, para cada fila del cuadro, el porcentaje correspondiente a la segunda columna del que hallamos en la primera.

	X1	X2	D%
Y1	60%	25%	35
Y2	40%	75%	-35
	100%	100%	

Si el cuadro cruza dos dicotomías, la D% asume un solo valor (con signo positivo en una fila y negativo en la otra). Se suele tomar el correspondiente a la primera fila, que arroja valor positivo si la mayor carga está en la diagonal positiva. Resulta así un único valor de D% (representativo de la relación) estandarizado entre 100 (asociación perfecta) y 0 (independencia de las variables):

Asociación perfecta:

	X1	X2	D%
Y1	100%	-	100
Y2	-	100%	-100
	100%	100%	

Independencia:

	X1	X2	D%
Y1	50%	50%	0
Y2	50%	50%	0
	100%	100%	

Si la variable dependiente "Y" posee más de dos valores (siendo "X" dicotómica), ya no tendremos un solo valor de D%: podremos obtener uno diferente para cada fila. Ya no existiría, en este caso, una sola medida de la fuerza de la asociación (aunque de todas formas, la sumatoria algebraica de las D% seguiría siendo igual a cero).

	X1	X2	D%
Y1	40%	20%	20
Y2	50%	30%	20
Y3	10%	50%	-40
	100%	100%	

Pero el problema mayor surge cuando es "X" (la variable independiente) la que es politómica ("n" valores):

	X1	X2	X3	D%
Y1	60%	40%	20%	40 (60-20)
Y2	30%	40%	75%	-45 (30-75)
Y3	10%	20%	5%	
	100%	100%	100%	

Aquí, existe más de una alternativa para obtener D%. Si (para cada fila) los porcentajes crecen - o decrecen - en forma monótona (como en las dos primeras filas del cuadro precedente) convendrá calcular la diferencia porcentual entre las columnas extremas. Esas serían siempre las más significativas. Sin embargo, puede suceder que - como en la última fila - resulten más importantes las diferencias interiores: -10 entre la primera y la segunda y 15 entre la segunda y la tercera. Convendrá valernos de estas, puesto que si calculásemos la existente entre las puntas no evaluaríamos adecuadamente la relación entre las variables. De todas maneras, cuando tenemos cuadros que cruzan más que dicotomías, las diferencias porcentuales ya no resultan tan claras y útiles como modo de caracterizar la relación entre las variables, y es pertinente recurrir a otros coeficientes más adecuados (que no trataremos aquí).

7- MAS DE DOS VARIABLES

Es frecuente que se encuentren cuadros que incluyen más de dos variables (y muchas veces hay buenas razones para emplear más de dos criterios clasificatorios, en forma simultánea, al tabular los datos). Para valernos de un ejemplo, al considerar la relación presentada en el **punto 2**, entre la *condición migratoria* y el *nivel educativo*, acaso nos preguntemos si ella no puede estar influenciada por la distinta composición por *edad* de los migrantes y nativos: pudiera ocurrir que una proporción muy alta de jóvenes en alguno de los grupos aumente el peso de aquellos que aún no han concluido el ciclo secundario. Indagar este aspecto - como muchos otros - requeriría examinar la relación entre condición migratoria e instrucción para distintos segmentos de edad (en forma separada), con la finalidad de observar si ella permanece estable o varía de uno a otro segmento. Esta es, apenas, **una** razón para incorporar al análisis más de dos criterios. A estos efectos, deberá situarse la primera *variable de corte o de control* ("Z") en las columnas, por fuera de "X":

	Z1		Z2	
	X1	X2	X1	X2
Y1				
Y2				
	100%	100%	100%	100%

Entonces se comenzará por considerar cada una de las dos mitades de la tabla por separado, examinando la relación entre "X" e "Y". Luego se estará en condiciones de comparar la relación existente entre "X" e "Y" para los distintos valores de "Z".

A veces, aparece considerada aun una cuarta variable ("T"), y se la sitúa en las filas, por fuera de "Y":

		Z1		Z2	
		X1	X2	X1	X2
T1	Y1				
	Y2				
T2		100%	100%	100%	100%
	Y1				
	Y2				
		100%	100%	100%	100%

En cada caso, el 100% indica cuál es la base sobre la que habrían de calcularse porcentajes. De todas formas, la consideración simultánea de más de tres variables torna complejo el análisis, y exige un gran número de casos para evitar que desciendan mucho las frecuencias al multiplicarse las celdas. (o que disminuyan mucho las frecuencias en cada celda)

El decidir cuál debe ser la variable de *corte* (la más externa del cuadro) es cuestión teórica. *Mora y Araujo* (1965) aconseja situar en tal posición la más independiente de todas, examinando la relación entre las otras dos al interior de cada una de sus categorías (a modo de regla general).

8- RECOMENDACIONES GENERALES

No siempre - ya ha sido dicho - la presentación de los cuadros respeta estas "reglas de buena práctica". De hecho, la información incluida en las tablas es, a veces, compleja y hasta confusa.

Para clarificar de qué cosas nos habla un cuadro y qué tipo de información provee respecto de ellas, pueden ser de utilidad algunas recomendaciones generales:

- **La unidad de análisis**: resulta fundamental saber, antes que nada, cuál es la unidad de análisis considerada. De quiénes - o de qué - se predican las variables. Puede tratarse de individuos, de conjuntos de ellos, de instituciones, de distritos geográficos o aún de otros elementos, tales como productos culturales (libros, periódicos o trabajos de investigación, por ejemplo). Debiera estar mencionada en el título del cuadro o en alguna información al pie del mismo. No siempre ocurre así y, a veces , resulta difícil reconocerlas.

- **Las variables**: es importante establecer cuántas variables están consideradas en el cuadro, cuál es su nivel de medición y cuáles sus categorías o valores. También cuál es el *status lógico* que les ha adjudicado el autor al situarlas en la tabla (cuál es la *independiente* y cuál la *dependiente*). *Variables*, *categorías* de ellas y *unidades de análisis* suelen confundirse a veces entre sí. Cuando decidimos claramente qué es cada cosa, se nos torna evidente la estructura del cuadro. Por ejemplo, en la siguiente tabla, tanto podría entenderse que las *unidades de análisis* son cada uno de los *habitantes* de la Argentina, clasificados en función de su *lugar de residencia* (una variable *nominal*), como también podrían tomarse los *distritos* como *unidades de análisis*, siendo su *población total* una variable *intervalar* (lo que en el primer caso era *categoría de variable*, pasa a ser *unidad de análisis* cuando es otra la perspectiva que dirige nuestro interés: en la primera alternativa, nos estaríamos interesando por las personas y su distribución geográfica; en la segunda, por los distritos y el tamaño de su población).

Argentina: población total por distrito geográfico

Distrito geográfico	Población total
Buenos Aires	10.865.408
Capital Federal	2.922.829
Catamarca	207.717
Córdoba	2.407.754

Fuente: Censo Nacional 1980

- **Las frecuencias**: Si el cuadro ya se presenta en porcentajes (como es frecuente), será esencial determinar en qué sentido estos fueron calculados (cuáles son las bases). Pero también habrá que reparar en si las frecuencias absolutas son suficientes como para que tenga sentido usar porcentajes y sacar conclusiones respecto de ellos. Asimismo es importante el tamaño total de la muestra, porque cuenta para saber en qué medida pueden generalizarse las observaciones emergentes del cuadro (sobre este último aspecto versan las llamadas *pruebas de significación estadística*, que no trataremos aquí).

9- OTRAS FORMAS DE CUADROS

A veces, los cuadros se presentan de forma tal que pueden desconcertar al observador poco habituado. Por ejemplo, cuando una de las variables es *dicotómica*, es usual (y correcto) presentar tan solo los porcentajes correspondientes a una de las categorías, entendiendo que a la restante le corresponde la diferencia a 100. Tal sería el caso del siguiente cuadro:

Porcentaje de personas que aprueban la integración regional (MERCOSUR), por país de origen

País de origen	%	
Argentina	42%	(58%)
Brasil	58%	(42%)
Paraguay	39%	(61%)
Uruguay	35%	(65%)

(Datos ficticios)

La *columna oculta*, correspondiente a quienes no aprueban la iniciativa integracionista, aparece entre paréntesis.

Es posible hallar presentaciones más complejas de los datos que, en realidad, resultan de practicar algunas operaciones estadísticas sobre el cuadro original. Por ejemplo:

Tasa de actividad por grupos de edad y sexo (Provincia de La Rioja - población de 14 y más años)

Edades	Varones	Mujeres
15-19	28,3	21,7
20-24	80,5	42,6
25-29	94,7	46,7
30-39	98,5	50,3
40-49	95,3	44,6
50-59	71,7	25,1
60-69	31,3	13,3
70 y más	12,2	0,0

(Fuente: EPH - Octubre'82)

¿Cuál es aquí la *unidad de análisis*?. ¿Cuáles las *variables*?. En principio tenemos una *variable nominal* (el *sexo*) y otra *intervalar* (*la tasa de actividad*). Pero, ¿de qué *unidad de análisis* se predicaría esta última?: no puede una *tasa* obtenerse de individuos, sino de grupos de ellos. ¿Acaso se tratará de los *grupos de edad*?

Podría pensarse en estos *grupos etarios* como *unidades colectivas*[1]. Pero sin embargo, el *sexo* es un atributo de los individuos y no de los grupos (debiera, entonces, hablarse de grupos definidos por dos atributos: *edad* y *sexo*). En verdad, en el origen del cuadro anterior están los *individuos* (unidades de análisis), clasificados por *edad*, *sexo* y *condición de actividad*:

VARONES

Condición de actividad

Edad	Activos	No activos	Total
15-19	28,3%	71,7%	100%
20-24	80,5%	19,5%	100%
25-29	94,7%	5,3%	100%
.....	.....	.....	.....
70 y más	12,2%	87,8%	100%

MUJERES

Condición de actividad

Edad	Activos	No activo	Total
15-19	21,7%	78,3%	100%
20-24	42,6%	57,4%	100%
25-29	46,7%	53,3%	100%
.....	.....	.....	.....
70 y más	0,0%	100%	100%

El primer cuadro resulta de la porcentualización - en sentido horizontal - realizada en el segundo: el porcentaje de activos para cada segmento de edad es la **tasa de actividad**: que es la columna que se traslada al primer cuadro.

Este tipo de cuestiones solo pueden aclararse mediante el análisis cuidadoso y la reflexión. De todos modos - y como lo ilustra también el ejemplo del punto 8 - un cuadro no es otra cosa que un modo particular de organizar y presentar la información, que sirve a los propósitos del investigador: sus elementos constituyentes (y el papel que ellos desempeñan en cada ocasión) dependen de cierto número de abstracciones que se practican sobre la realidad en función de tales propósitos.

---

[1]Aunque no cumplirían el requisito exigido por *Francis Korn*, consistente en poseer - al menos - una *propiedad global* diferente del tamaño (en número de miembros).

BIBLIOGRAFÍA UTILIZADA:

BARTON, Allen - "El concepto de espacio de propiedades en la investigación social", en Francis Korn y otros: "Conceptos y variables en la investigación social" - Ediciones Nueva Visión - Bs.As., 1973

BENSON, Oliver - "El laboratorio de la ciencia política" - Editorial Amorrortu - Bs.As., 1974

BLALOCK, Hubert - "Estadística social" - F.C.E. - México, 1986

CICOURELL, Aaron - "El método y la medida en sociología" - Editora Nacional - Madrid, 1982

CORTES, Fernando y RUBALCAYA, Rosa - "Métodos estadísticos aplicados a la investigación en ciencias sociales" - El Colegio de México - México, 1987

GALTUNG, Johan - "Teoría y método de la investigación social" - EUDEBA - Bs.As., 1978

GARCÍA FERRANDO, Manuel - "Socioestadística" - Alianza Editorial - Madrid, 1985

KORN, Francis - "El significado del término *variable* en sociología", en Korn y otros: "Conceptos y variables en la investigación social" - Ediciones Nueva Visión - Bs.As., 1973

MORA Y ARAUJO, Manuel - "Recomendaciones para la lectura y análisis de cuadros" - Facultad de Filosofía y Letras U.B.A. - Bs.As., 1965

SIERRA BRAVO, Restituto - "Técnicas de investigación social - Teoría y ejercicios" - Paraninfo - Madrid, 1988

ZEIZEL, Hans - "Dígalo con números" - F.C.E. - México, 1974