CAPÍTULO 4-1
EL ANÁLISIS DE LAS TABLAS DE CONTINGENCIA

4-1.1. INTRODUCCIÓN

4-1.1.1. ¿Qué es una tabla de contingencia?

Una tabla de contingencia es una manera de presentar datos de enumeración (de conteo) de individuos previamente clasificados en categorías. Es de constatar, por lo tanto, que una tabla de contingencia es ya el resultado de un tratamiento de datos puesto que los individuos (observaciones) tuvieron que ser el objeto de una previa clasificación y de un previo conteo.

Se determina el formato de la tabla dependiendo del modelo de clasificación que se emplea. Es necesario que el modelo de clasificación se constituya de categorías mutuamente exclusivas y que sea exhaustivo, es decir que en el lenguaje de la teoría de conjuntos, las categorías deben constituir una partición del universo de modo que cada individuo pertenezca a una y a una sola categoría.

Se definen las categorías por medio de una o varias variables de clasificación (variables categóricas) que corresponden a cuantos atributos (dimensiones) tienen los individuos. Se describe a los individuos observados con el fin de clasificarlos en función de los valores de sus atributos. Se efectúa un

conteo de todos los individuos que tengan la misma descripción (los mismos valores de atributos) para luego inscribir su número en la celda correspondiente de la tabla de contingencia resultado de esta clasificación. La tabla de contingencia tiene tantas dimensiones como haya variables de clasificación y tantas celdas como haya combinaciones de categorías.

Examinemos un pequeño ejemplo de construcción de una tabla de contingencia a partir de datos brutos. Teniendo la tabla de observaciones siguiente, donde las observaciones ya están ordenadas por sexo, pues por color de ojos:

	Nombre	Sexo	Color de ojos
1	Dolores	M	Azules
6	Juan	H	Azules
4	Marco	H	Negros
2	Maria	M	Azules
3	Pedro	H	Negros
5	Guadalupe	M	Negros

Se ve en la tabla arriba la estructura matricial de los datos brutos. A partir de estos datos, es posible deducir una tabla de contingencia del color de los ojos en función del sexo:

	Sexo		
Color de ojos	M	H	Total
Azules	2	1	3
Negros	1	2	3
Total	3	3	6

Esa tabla de contingencia tiene también una estructura matricial, pero ya no se trata de datos brutos: la tabla de contingencia es el resultado de un tratamiento. Se nota que una tabla de contingencia de una sola variable es nada más una tabla de frecuencias.

Tabla 1: Población activa empleada
en la región metropolitana de Montreal, 1991
Zona de residencia según el sexo y la profesión

Zona de residencia	Profesiones					
	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Mujeres						
Montreal[178]	24,025	58,204	76,450	24,385	28,825	*211,889*
Resto CUM[179]	22,575	42,207	70,003	14,065	17,435	*166,285*
Anillo norte	16,785	31,699	63,491	11,975	18,630	*142,580*
Anillo sur	18,365	35,674	65,290	10,485	19,380	*149,194*
Fuera RMR[180]	3,265	7,535	11,089	3,190	3,565	*28,644*
Total Mujeres	*85,015*	*175,319*	*286,323*	*64,100*	*87,835*	*698,592*
Hombres						
Montreal	32,336	55,045	43,546	65,340	46,850	*243,117*
Resto de CUM	39,146	39,920	37,819	46,173	28,749	*191,807*
Anillo Norte	33,287	27,560	31,170	62,852	29,329	*184,198*
Anillo Sur	36,006	32,464	30,600	58,778	29,721	*187,569*
Fuera de RMR	8,270	8,590	8,270	22,305	9,099	*56,534*
Total Hombres	*149,045*	*163,579*	*151,405*	*255,448*	*143,748*	*863,225*
Total hombres y mujeres						
Montreal	56,361	113,249	119,996	89,725	75,675	*455,006*
Resto de CUM	61,721	82,127	107,822	60,238	46,184	*358,092*
Anillo Norte	50,072	59,259	94,661	74,827	47,959	*326,778*
Anillo Sur	54,371	68,138	95,890	69,263	49,101	*336,763*
Fuera de RMR	11,535	16,125	19,359	25,495	12,664	*85,178*
Total H + M	*234,060*	*338,898*	*437,728*	*319,548*	*231,583*	*1,561,817*

Fuente: Statistique Canada, Censo de 1991.

[178] Se refiere aquí al municipio de Montreal tal como era definido hasta la fusión en 2002 de todos los municipios que antes formaban la CUM (ver la próxima nota).

[179] CUM = Comunidad Urbana de Montreal: todos los municipios de la Isla de Montreal. La CUM volvió con la fusión del 2002 en el Municipio del Gran Montreal.

[180] RMR = Región Metropolitana de Censo.

Esta tabla de contingencia posee tres dimensiones: la zona de residencia, la profesión y el sexo. La zona de residencia tiene 5 categorías, la profesión, 5 y hay 2 sexos. La tabla contiene, por lo tanto, 5 x 5 x 2 = 50 celdas a las cuales se suman las líneas y columnas de los totales y subtotales.

4-1.1.2. El análisis de las tablas de contingencias entre los métodos de análisis multivariado

De manera general, el análisis mutivariado designa el conjunto de los métodos de análisis estadístico que tratan simultáneamente con más de una variable. En particular, se recurre al análisis multivariado para:

- medir el grado de asociación entre dos o más variables;
- estimar los parámetros de una relación entre dos o más variables;
- evaluar hasta qué punto las diferencias entre dos o varios grupos de observaciones son significativas;
- intentar predecir a cuál grupo pertenece un individuo a partir de sus demás características;
- buscar reconocer una estructura en un conjunto de datos.

Varias técnicas de análisis multivariado permiten distinguir entre las variables dependientes y las variables independientes. Las variables dependientes son las variables cuyo valor se quiere predecir; las otras variables son las independientes.[181] Es posible clasificar los métodos de análisis multi-

[181] Se tomaron los términos "variable dependiente" y "variable independiente" del área de las ciencias experimentales, cuando el investigador fija de manera "independiente" el valor de ciertas variables (como, por ejemplo, la dosificación de un tratamiento) para, luego, observar su efecto en la variable "dependiente". Se da a veces a las variables independientes el nombre de variables "explicativas". Sin embargo, hay que tener sumo cuidado con esta expresión por la connotación de causalidad que transmite. En un modelo con una sola ecuación, la variable dependiente se llama también "endógena", es decir que se

variado en función del número de variables dependientes e independientes y según si las unas o las otras son variables discretas o continuas.[182]

La tabla que sigue presenta una clasificación de los métodos que examinamos en el marco de este curso.

<table>
<tr><th colspan="2">Variable dependiente</th><th>Variables independientes</th><th colspan="2">Método</th></tr>
<tr><td colspan="2" rowspan="2">Ninguna</td><td>2 variables categóricas</td><td rowspan="2">Análisis de tabla de contingencia</td><td>... con 2 dimensiones</td></tr>
<tr><td>Más de 2 variables categóricas</td><td>... con más de 2 dimensiones</td></tr>
<tr><td colspan="2" rowspan="2">Continua</td><td>Discretas (categóricas)</td><td colspan="2">Análisis de varianza o Regresión múltiple</td></tr>
<tr><td>Continuas y/o discretas</td><td colspan="2">Regresión múltiple</td></tr>
<tr><td rowspan="2">Categórica</td><td>2 categorías</td><td rowspan="2">Continuas y/o discretas</td><td rowspan="2">Logit o probit</td><td>... binomial</td></tr>
<tr><td>Más de 2 categorías</td><td>... multinomial</td></tr>
</table>

Este capítulo trata del análisis de las tablas de contingencia. Con este método, es posible examinar las relaciones entre

---

determina al interior del modelo, mientras que las variables independientes son "exógenas", es decir que se determinan al exterior del modelo. Las variables independientes se llaman también "estímulos", y entonces las dependientes son "respuestas". En inglés, es posible encontrar las parejas predictor/criterion, stimulus/response, task/perfomance, input/output.

[182] Se infiere esto de la escala de medición asociada a cada variable, es decir, las variables categóricas son discretas mientras que se considera frecuentemente las variables racionales y de intervalo como variables continuas. En cuanto a las variables ordinales, existen pocos métodos que se adaptan específicamente a ellas; en la practica, se consideran continuas. Sin embargo, en tales condiciones, la interpretación de los resultados debe tomar en cuenta la naturaleza ordinal de las variables.

varias variables categóricas. En el análisis de las tablas de contingencia, ninguna variable toma el papel de variable dependiente.

4-1.1.3. Reglas de presentación de una tabla de contingencia

El principio general de presentación de una tabla de contingencia no difiere de cualquier otra tabla: todo debe encaminarse para que el lector sepa perfectamente de lo que trata.

Las principales reglas de presentación que conviene por lo general respetar, son las siguientes:

1. La tabla se encabeza con un titulo que identifica la población o, si fuera el caso, la muestra a la cual se refiere la tabla (en nuestro ejemplo, la población activa empleada en la RMR de Montreal en 1991); note que la identificación de la población contiene, cuando el caso lo requiere una referencia a la zona geográfica y al periodo de tiempo; indica cuales son las unidades de medición empleadas (miles de personas, millones de dólares o...; se puede omitir este elemento en nuestro ejemplo puesto que se trata de número de personas); identifica las dimensiones de la tabla (variables categóricas de clasificación; aquí, la zona de residencia, el sexo y la profesión).
2. Algunos subtítulos indican a cuál variable corresponden las diferentes dimensiones de la tabla (aquí, las líneas corresponden a las zonas de residencia, las columnas a las profesiones y la tabla se divide en partes en función de la tercera dimensión, el sexo).
3. El encabezado de cada columna, línea o parte de la tabla indica a cuál categoría de la variable corresponde esta columna, línea o parte de la tabla.
4. La tabla contiene líneas y columnas de totales así como del gran total (1,561,817); las líneas y las columnas

se identifican con claridad y resaltan (en nuestro ejemplo, con caracteres en negrillas).
5. Finalmente, se indica la fuente de los datos (aquí en términos generales, cuando lo ideal es procurar una referencia bibliográfica completa).

Es posible que una tabla de contingencia contenga también los elementos siguientes:

- proporciones o porcentajes;
- subtotales;
- llamadas y notas correspondientes.

Si la tabla contiene proporciones o porcentajes, es importante poder evidenciar con toda claridad si se trata de proporciones (fracciones contenidas entre cero y uno) o de porcentajes (contenidos entre cero y cien). Además, es necesario indicar claramente la razón de los porcentajes o proporciones (¿porcentaje de qué?). Una manera de llevar a cabo esta tarea es escribiendo "100%" donde conviene hacerlo. Finalmente, es importante no sobrecargar una tabla hasta el punto de dificultar su lectura; puede ser preferible, pues, presentar dos tablas, una para los números y la segunda para los porcentajes.

Lo anterior es también válido para los subtotales. Por ejemplo, en la tabla exhibida más arriba, podríamos pensar que fuese útil presentar el subtotal de la CUM (suma de las dos primeras líneas de cada parte). No obstante, se deben formular los subtotales para que el lector reconozca con claridad lo que se sumó. Además, es importante no sobrecargar la tabla y para este efecto, cabe a menudo presentar los mismos datos en dos tablas: una primera tabla detallada (a veces en anexo) y una segunda, más agregada, que de hecho presenta subtotales.

Para aclarar puntos de los títulos, subtítulos o encabezados sin necesidad de alargarlos indebidamente, se usan notas, las mismas que se pueden emplear para dar definiciones de

algunos términos o para enunciar fórmulas que permitieron calcular las cifras de la tabla.

Finalmente, existen varias maneras de estructurar un tabla con más de dos dimensiones. En el ejemplo anterior, la tercera dimensión, el sexo, corresponde a las diferentes partes de la tabla. Es posible también usar la técnica de la subdivisión de las líneas o de las columnas, técnica que se ilustra con una figura más abajo.

Figura de una tabla
de contingencia con subdivisión de las columnas

	Profesiones, sexo								
Zona de residencia	Directores, gerentes, administradores y similares			Profesionales docentes y cuellos blancos especializados			...	TOTAL todas las profesiones		
	M	H	T	M	H	T		M	H	T
Montreal							...			
					⋮					
Total										

En cuanto se usa, estos métodos de representación, se recomienda acercar hacia la parte interna las variables cuya interacción se desea examinar. La estructura que representamos justo arriba convendría perfectamente para el estudio de la interacción entre sexo y zona de residencia considerando que, en estas condiciones, la profesión toma el papel de una variable de control (se examinan las variables de control más abajo, en el apartado 6); el formato anterior se adapta mejor al examen de la relación entre profesión y zona de residencia mientras que el sexo toma, entonces, el papel de variable de control.

4-1.2 FRECUENCIAS RELATIVAS Y PROBABILIDADES EN UNA TABLA DE CONTINGENCIA

Aunque se puedan generalizar los métodos que a continuación se presentan para las tablas de más de dos dimensiones, porque es más simple, nos limitaremos ahora, a analizar las tablas de dos dimensiones. Con este propósito, retomaremos la tabla anterior pero omitiendo la dimensión del sexo.[183] Para esto, solo basta sumar los hombres y las mujeres, como se efectúa en la tabla siguiente cuyos números son los mismos que aparecen en la tercera parte de la tabla 1.

Tabla 2: Población activa empleada
en la región metropolitana de Montreal, 1991
Zona de residencia según la profesión

	Profesiones						
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trab. especial. en servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones	Repartición $p_{i\bullet}$
Montreal	56,361	113,249	119,996	89,725	75,675	*455,006*	0.29
Resto CUM	61,721	82,127	107,822	60,238	46,184	*358,092*	0.23
Anillo Norte	50,072	59,259	94,661	74,827	47,959	*326,778*	0.21
Anillo Sur	54,371	68,138	95,890	69,263	49,101	*336,763*	0.22
Fuera RMR	11,535	16,125	19,359	25,495	12,664	*85,178*	0.05
Total	*234,060*	*338,898*	*437,728*	*319,548*	*231,583*	*1,561,817*	1.00
Repartic. $p_{\bullet j}$	0.15	0.22	0.28	0.20	0.15	1.00	

[183] Es importante darse cuenta que este procedimiento destruye parte de la información. No se aconseja, por lo tanto, de ninguna manera, esta práctica que se lleva a cabo en este momento sólo por razones pedagógicas.

Usaremos la simbología que sigue:

x_{ij}	número de observaciones de la columna *j* en la línea *i*
$x_{\bullet j} = \sum_i x_{ij}$	número total de observaciones de la columna *j*
$x_{i\bullet} = \sum_j x_{ij}$	número total de observaciones de la línea *i*
$x_{\bullet\bullet} = \sum_i \sum_i x_{ij}$	número total de observaciones en la tabla

Con esta simbología se aplica la convención de que sumar sobre cualquiera de las dos dimensiones puede representarse reemplazando el índice correspondiente con un punto grueso. Por ejemplo, en la tabla de población activa por zona de residencia y por profesión, tenemos:

x_{23} = 107,822, el número de empleados de oficina que viven en la CUM fuera de Montreal;

$x_{\bullet 3}$ = 437,728, el número de empleados de oficina empleados en la RMR;

$x_{2\bullet}$ = 358,092, el número de personas empleadas en la RMR que viven en la CUM fuera de Montreal;

$x_{\bullet\bullet}$ = 1,561,817, el número total de personas empleadas en la RMR.

El análisis de una tabla de contingencia se refiere mucho más a la estructura de los datos que a las magnitudes de los números. Por esta razón se formulan, por lo general, los análisis en términos de las frecuencias relativas que se calculan simplemente con dividir los números por el total pertinente.

Se interpretan las frecuencias relativas como probabilidades. De esta manera, $p_{34} = \frac{74{,}827}{1{,}561{,}817} = 0.048$ corresponde a la probabilidad que un individuo, sorteado entre las 1,561,817 personas empleadas en la RMR, forme parte de la profesión Obrero y viva en el Anillo Norte. Por consiguiente, en el denominador, encontramos el número de individuos donde se efectúa el sorteo (1,561,817) y en el numerador encontramos el número de individuos que reúne la o las características que se pretende examinar (74,827).

Diferentes cálculos de frecuencias relativas corresponden a los diferentes conceptos de probabilidad. Así,

$p_{ij} = {}^{x_{ij}}\!/_{x_{\bullet\bullet}}$ es la probabilidad conjunta de pertenecer al mismo tiempo a *i* y a *j*.

Ejemplo: $p_{34} = \frac{74{,}827}{1{,}561{,}817} = 0.048$ es la probabilidad de formar parte de la profesión Obrero y de vivir en el Anillo Norte.

$p_{i\bullet} = {}^{x_{i\bullet}}\!/_{x_{\bullet\bullet}} = \sum_j {}^{x_{ij}}\!/_{x_{\bullet\bullet}} = \sum_j p_{ij}$ es la probabilidad marginal de pertenecer a *i*.

Ejemplo: $p_{3\bullet} = \frac{326{,}778}{1{,}561{,}817} = 0.209$ es la probabilidad de vivir en el Anillo Norte independientemente de la categoría profesional.

$p_{\bullet j} = {}^{x_{\bullet j}}\!/_{x_{\bullet\bullet}} = \sum_i {}^{x_{ij}}\!/_{x_{\bullet\bullet}} = \sum_i p_{ij}$ es la probabilidad marginal de pertenecer a *j*.

Ejemplo: $p_{\bullet 4} = \frac{319{,}548}{1{,}561{,}817} = 0.205$ es la probabilidad de pertenecer a la profesión Obrero independiente de la zona de residencia.

$p_{j/i\bullet} = {x_{ij}}\big/{x_{i\bullet}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{i\bullet}/x_{\bullet\bullet}} = {p_{ij}}\big/{p_{i\bullet}}$ es la probabilidad condicional de pertenecer a *j* dado que se pertenece a *i*.

Ejemplo: $p_{4/3\bullet} = \frac{74{,}827}{326{,}778} = 0.229$ es la probabilidad de pertenecer a la profesión Obrero dado que la zona de residencia es el Anillo Norte.

$p_{i/\bullet j} = {x_{ij}}\big/{x_{\bullet j}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}} = {p_{ij}}\big/{p_{\bullet j}}$ es la probabilidad condicional de pertenecer a *i* dado que pertenece a *j*.

Ejemplo: $p_{3/\bullet 4} = \frac{74{,}827}{319{,}548} = 0.234$ es la probabilidad de vivir en el Anillo Norte dado que se pertenece a la profesión Obrero.

Es obvio que, al sumar todas las probabilidades o frecuencias relativas posibles, el resultado es 1:

$$\sum_i \sum_j p_{ij} = \sum_i p_{i\bullet} = \sum_j p_{\bullet j} = 1$$

$$\sum_j p_{j/i\bullet} = {\sum_j x_{ij}}\big/{x_{i\bullet}} = 1$$

$$\sum_i p_{i/\bullet j} = {\sum_i x_{ij}}\big/{x_{\bullet j}} = 1$$

4-1.3 TEST DE HIPÓTESIS DE INDEPENDENCIA EN UNA TABLA DE CONTINGENCIA

4-1.3.1 Presentación intuitiva

Para cada profesión, la tabla 3 presenta la distribución de los individuos entre las zonas de residencia. Sin que sea una gran sorpresa, constatamos que los individuos de profesiones diferentes se reparten de manera diferente en el espacio, entre las zonas de residencia.

Tabla 3: Población activa empleada
en la región metropolitana de Montreal, 1991
Repartición entre las zonas de residencia según la profesión

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	0.241	0.334	0.274	0.281	0.327	*0.291*
Resto CUM	0.264	0.242	0.246	0.189	0.199	*0.229*
Anillo norte	0.214	0.175	0.216	0.234	0.207	*0.209*
Anillo sur	0.232	0.201	0.219	0.217	0.212	*0.216*
Fuera RMR	0.049	0.048	0.044	0.080	0.055	*0.055*
Total	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*

Sin embargo, ¿son significativas estas diferencias? Para examinar este problema, se comparan las distribuciones observadas con una distribución que sería, de manera hipotética, la misma para todas las profesiones; esta distribución hipotética es simplemente la distribución del total (última columna de la tabla).

Pero, ¿cómo decidir si estas diferencias son o no son "significativas"? Para esto se procede a un test de hipótesis (para más detalles sobre los tests de hipótesis, vea el capítulo 2-3). La hipótesis que pretendemos probar es que las distribuciones son idénticas y las diferencias observadas no son más que accidentes, productos del azar.

Hay tres etapas en este test:

1. Medir la desviación estándar entre lo que se observó y la hipótesis;
2. Determinar cuál es la probabilidad de que una diferencia tan grande sea el producto del azar (cuanto más grande es esta diferencia, menos probable es que sea producto del azar).
3. Tomar una decisión.

Primera etapa: medir la diferencia

Para medir la diferencia entre las observaciones y la hipótesis, es necesario, primero, tener una representación exacta de la hipótesis. Por lo tanto, se calculan las frecuencias que, teóricamente, se obtendrían si las distribuciones fuesen idénticas (tabla 4).

Tabla 4: Población activa empleada
en la Región Metropolitana de Montreal
Frecuencias teóricas
en la hipótesis de distribuciones idénticas

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	68,189.0	98,731.6	127,523.8	93,094.3	67,467.4	*455,006.0*
Resto CUM	53,665.1	77,702.2	100,361.9	73,265.7	53,097.1	*358,092.0*
Anillo norte	48,972.2	70,907.4	91,585.6	66,858.8	48,454.0	*326,778.0*
Anillo sur	50,468.6	73,074.1	94,384.0	68,901.8	49,934.5	*336,763.0*
Fuera RMR	12,765.1	18,482.7	23,872.7	17,427.4	12,630.0	*85,178.0*
Total	*234,060.0*	*338,898.0*	*437,728.0*	*319,548.0*	*231,583.0*	*1,561,817.0*

Se calculan estas frecuencias teóricas con sólo multiplicar el total de cada columna con la distribución de la totalidad (última columna de la tabla de las reparticiones): $x^*_{ij} = x_{\bullet j}\, p_{i\bullet}$ donde el asterisco sirve para distinguir las frecuencias teóricas de las frecuencias observadas. Por ejemplo:[184]

$$x^*_{54} = x_{\bullet 4} \times p_{5\bullet} = 319548 \times 0.0545378 = 17427.4$$

Podemos observar que los totales de las líneas y de las columnas de la tabla 4 y los mismos totales de la tabla 2 de los valores observados son iguales. Esto no es casual y se deduce de la fórmula de cálculo

[184] En la fórmula que sigue y con el fin de obtener frecuencias teóricas exactas, se debe tomar el valor de la probabilidad con 7 decimales puesto que el multiplicador es del orden de cientos de miles. Se busca tal precisión en este contexto para lograr claridad en el desarrollo que sigue, sin embargo esta precisión no es necesaria en la práctica.

$$\sum_i x_{ij}^* = \sum_i x_{\bullet j}\, p_{i\bullet} = \sum_i x_{\bullet j}\left(\frac{x_{i\bullet}}{x_{\bullet\bullet}}\right) = x_{\bullet j}\,\frac{\sum_i x_{i\bullet}}{x_{\bullet\bullet}}$$

$$\sum_i x_{ij}^* = x_{\bullet j}\,\frac{x_{\bullet\bullet}}{x_{\bullet\bullet}} = x_{\bullet j}$$

$$\sum_j x_{ij}^* = \sum_j x_{\bullet j}\, p_{i\bullet} = p_{i\bullet}\sum_j x_{\bullet j} = p_{i\bullet}\, x_{\bullet\bullet}$$

$$\sum_j x_{ij}^* = \frac{x_{i\bullet}}{x_{\bullet\bullet}}\, x_{\bullet\bullet} = x_{i\bullet}$$

Después de calcular las frecuencias teóricas, hay que medir la diferencia entre el conjunto de las frecuencias teóricas y el conjunto de las frecuencias observadas. Para esto, se aplica la fórmula

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij} - x_{ij}^*\right)^2}{x_{ij}^*}$$

Esta estadística es conocida como el Khi-dos (Ji o Chi cuadrado) de Pearson y con el símbolo X^2 tal y como aparece en la fórmula.

Tabla 5: Población activa empleada
en la Región Metropolitana de Montreal
Cálculo del Chi-cuadrado

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	2,051.7	2,134.6	444.4	121.9	998.5	*5,751.1*
Resto CUM	1,209.3	252.0	554.5	2,316.5	900.1	*5,232.4*
Anillo norte	24.7	1,913.6	103.3	949.6	5.1	*2,996.2*
Anillo sur	301.7	333.4	24.0	1.9	13.9	*675.0*
Fuera RMR	118.5	300.8	853.4	3,734.7	0.1	*5,007.5*
Total	*3,706.0*	*4,934.4*	*1,979.6*	*7,124.6*	*1,917.6*	*19,662.2*

Los valores de la tabla 5 son las contribuciones de las celdas individuales al Chi-cuadrado

Así, en el caso de la quinta celda de la cuarta columna

$$\frac{(25{,}495 - 17{,}427.4)^2}{17{,}427.4} = 3{,}734.7$$

El Chi-cuadrado es simplemente la suma de todos los elementos de esta tabla: 19,662.2.

Segunda etapa: determinar la probabilidad

¿Por qué se emplea esta fórmula y no otra? Encontramos la repuesta a esta pregunta en la teoría de la inducción estadística. Se emplea esta fórmula porque, gracias a la estadística matemática, se conoce la distribución de probabilidad del Chi-cuadrado que se calculó de esta manera. En efecto, el Chi-cuadrado posee una distribución asintótica muy conocida: es la ley del χ^2 (decimos “Chi-cuadrado” puesto que el símbolo χ es la letra griega “Chi”). Es posible aplicar este re-

sultado siempre y cuando se use un cierto modelo de muestreo; esto es, un modelo de muestreo es un modelo que describe el proceso aleatorio por el cual, suponemos, se generan las diferencias entre las frecuencias observadas y las frecuencias teóricas (vea capítulo 2-2). No estudiaremos este modelo en este momento pero sí notaremos que este modelo de muestreo es lo suficientemente general como para poder aplicar el test de hipótesis del Chi-cuadrado de Pearson a una gran variedad de situaciones (vea más abajo, 4-1.4).

Es importante, ahora, notar que al momento de usar la ley del χ^2, es necesario tomar en cuenta lo que conocemos como el número de grados de libertad del cual dependen las probabilidades que la ley del χ^2 nos da. Para el test de hipótesis del Chi-cuadrado de Pearson, el número de grados de libertad es igual a

$$(C - 1)\,(L - 1),$$

donde C es el número de columnas y L, el número de líneas en la tabla.

En nuestro ejemplo (tablas 2 a 5), C corresponde al número de profesiones y L, al número de zonas; por consiguiente, en número de grados de libertad es igual a

$$(5 - 1)\,(5 - 1) = 16$$

Hagamos un pequeño paréntesis sobre el número de grados de libertad. Se representa la Ley del Chi-cuadrado con una curva cuya forma varía con el número de valores que el azar es libre de perturbar, por así decirlo. En una tabla de contingencia, los totales de líneas y columnas son fijos, así que en cada una de las C columnas, una vez que este tremendo azar haya "libremente" perturbado $(L - 1)$ valores, el último valor de la columna es determinado por la diferencia entre el total y los otros $(L - 1)$ valores; de igual manera, en cada una de las L líneas, una vez que se haya "libremente" perturbado $(C - 1)$ valores, el último valor de la columna es determinado por la diferencia entre el total y los otros $(C - 1)$ valores. En consecuencia, en la tabla completa, una vez que

introducidas $(C - 1)$ $(L - 1)$ “perturbaciones”, se determinan los demás valores por la necesidad de respetar los totales marginales.

Por medio de una tabla del Chi cuadrado o de la función Ley.Khidos (CHIDIST en inglés o Prueba Chi) del tabulador Excel X^2, es posible ahora determinar la probabilidad de que la diferencia medida entre las frecuencias observadas y las frecuencias teóricas sea tan grande; en particular, el valor de Prueba Chi[185] (19662;16) es inferior a 2.4×10^{-300}.

Tercera etapa: tomar una decisión

Una probabilidad de 2.4×10^{-300} es una probabilidad tan pequeña que es en extremo improbable que las desviaciones de las frecuencias observadas con relación a las frecuencias teóricas se deben únicamente al azar. De hecho, es tan improbable que, al menos que surjan circunstancias excepcionales, la buena decisión que se debe tomar es rechazar esta hipótesis y, por lo contrario, concluir que existe ciertamente una relación entre la profesión y la zona de residencia.

4-1.3.2 ¿¡Datos idénticos, nueva pregunta... respuesta idéntica?!

Acabamos de examinar si era significativo que los individuos con profesiones diferentes se repartieran entre zonas de residencia diferentes. Ahora pretendemos saber si la composición profesional de la población empleada es, de manera significativa, diferente de una zona de residencia a otra. Los datos pertinentes se encuentran en la tabla 6, más abajo, la cual procura, para cada zona de residencia, la distribución de los individuos entre las profesiones.

[185] Prueba Chi es la función correspondiente a Xi cuadrado en la versión en español de Excel.

Para ser más precisos, queremos probar la hipótesis de que no existen diferencias significativas entre las zonas con relación a la composición profesional de las personas empleadas que ahí viven. En la tabla 6 se compara, por lo tanto, las diferentes distribuciones con aquellas que, teóricamente, encontraríamos si las distribuciones fueran idénticas, lo que corresponde a la distribución de la totalidad (último renglón).

Tabla 6: Población activa empleada
en la Región Metropolitana de Montreal
Composición profesional de las zonas de residencia, 1991

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	0.124	0.249	0.264	0.197	0.166	*1.000*
Resto CUM	0.172	0.229	0.301	0.168	0.129	*1.000*
Anillo norte	0.153	0.181	0.290	0.229	0.147	*1.000*
Anillo sur	0.161	0.202	0.285	0.206	0.146	*1.000*
Fuera RMR	0.135	0.189	0.227	0.299	0.149	*1.000*
Total	*0.150*	*0.217*	*0.280*	*0.205*	*0.148*	*1.000*

Es posible constatar que existen, de hecho diferencias entre las zonas de residencia en cuanto a la composición profesional. Para saber si estas diferencias son significativas, procedemos de la misma manera que anteriormente, es decir, empezamos por calcular las frecuencias teóricas. Sin embargo, ¡qué sorpresa!, las frecuencias teóricas que arroja el cálculo son las mismas que las frecuencias teóricas calculadas cuando examinábamos el problema de la distribución entre las zonas de las diferentes profesiones (el lector puede verifi-

carlo por sí mismo). Es inútil seguir, pues llegaremos forzosamente a la misma conclusión.

Claro está que esto no es el resultado del azar. En el primer caso, tenemos

$$x^*_{ij} = x_{\bullet j} \times p_{i\bullet}$$

Por ejemplo:

$$x^*_{54} = x_{\bullet 4} \times p_{5\bullet} = 319{,}548 \times 0.0545378 = 17{,}427.4$$

En el caso presente,

$$x^*_{ij} = x_{i\bullet} \times p_{\bullet j}$$

Por ejemplo:

$$x^*_{54} = x_{5\bullet} \times p_{\bullet 4} = 85{,}178 \times 0.2046002 = 17{,}427.4$$

Las dos fórmulas permiten llegar al mismo resultado numérico por ser completamente equivalentes:

$$x^*_{ij} = x_{\bullet j}\ p_{i\bullet} = x_{\bullet j} \frac{x_{i\bullet}}{x_{\bullet\bullet}} = x_{i\bullet} \frac{x_{\bullet j}}{x_{\bullet\bullet}} = x_{i\bullet}\ p_{\bullet j}$$

De por sí, en la práctica cambiamos por lo general de la tabla de las frecuencias observadas a la tabla de las frecuencias teóricas por medio de la fórmula

$$x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}}$$

La tabla de las frecuencias teóricas es, por lo tanto, biproporcional, es decir, las columnas son proporcionales entre sí, así como las líneas.

4-1.3.3 Generalización: la independencia estadística en una tabla de contingencia

El análisis de una tabla de contingencia se basa en el postulado de que el número de individuos observados en las celdas de la tabla depende de una estructura subyacente. El análisis tiene como objetivo descubrir esta estructura. Evocaremos

más tarde y de manera breve el modelo log-lineal que sirve para representar esta estructura. En este instante, sólo nos interesa un aspecto particular de esta estructura: la independencia estadística.

¿Qué es la independencia estadística? En la teoría de las probabilidades, un evento aleatorio *A* es independiente de otro evento *B* cuando la probabilidad que el evento *A* suceda siga la misma de que el evento *B* suceda o no. Por ejemplo, en la tabla de la población activa por zona de residencia y por profesión, hay independencia entre las variables zona de residencia y profesión si, por un individuo sorteado, la probabilidad de vivir en una zona dada es la misma no importando la profesión de este individuo. De manera simétrica, hay independencia si la probabilidad de pertenecer a un grupo de profesión dada es la misma no importando la zona de residencia del individuo.

Por ejemplo, digamos que el evento *A* es "el individuo vive en el Anillo Norte" y el evento *B* es "el individuo es empleado de oficina": si hubiera independencia, la probabilidad de que un individuo sorteado viva en el Anillo Norte (probabilidad del evento *A*) sería la misma no importando que este individuo fuera empleado de oficina (evento *B*) o no.

Examinemos de más cerca cómo se manifiesta la independencia en una tabla de contingencia. Con este objetivo, es importante empezar por interpretar las frecuencias relativas de la tabla como si fueran probabilidades observadas, las mismas que se confrontarán en contra de las probabilidades teóricas del modelo o de la hipótesis. Así, para un individuo sorteado entre 1,561,817 trabajadores censados de la RMR, la probabilidad de que sea un obrero y que viva en el Anillo Sur se define con

$$p_{44} = \frac{x_{44}}{x_{\bullet\bullet}} = \frac{69{,}263}{1{,}561{,}817} = 0.044$$

De la misma manera se calculan las probabilidades marginales observadas. Así, para un individuo sorteado entre

1,561,817 trabajadores censados de la RMR, la probabilidad de que sea un obrero se define con

$$p_{\bullet 4} = \sum_i p_{i4} = \sum_i \left(\frac{x_{i4}}{x_{\bullet\bullet}} \right) = \frac{x_{\bullet 4}}{x_{\bullet\bullet}} = \frac{319{,}548}{1{,}561{,}817} = 0.205$$

Además, la probabilidad que un individuo sorteado entre 1,561,817 trabajadores censados de la RMR, viva en el Anillo Sur se define con

$$p_{4\bullet} = \sum_j p_{4j} = \sum_j \left(\frac{x_{4j}}{x_{\bullet\bullet}} \right) = \frac{x_{4\bullet}}{x_{\bullet\bullet}} = \frac{336{,}763}{1{,}561{,}817} = 0.216$$

Y con todo esto, ¿dónde está la independencia? Si la probabilidad de ser un obrero es independiente de la zona de residencia entonces la fracción de obreros en cada zona debería ser igual a $p_{\bullet 4}$, o sea a 20.5%. Como sabemos que la fracción de trabajadores que viven en la Anillo Sur es igual a $p_{4\bullet}$, o sea a 21.6%, entonces, entre los 1,561,817 trabajadores censados de la RMR, los que son obreros y viven en la Anillo Sur deberían representar 20.5% de 21.6% del total, o sea

$$p_{\bullet 4} \times p_{4\bullet} = 0.205 \times 0.216 = 0.044$$

Esto, lo recordamos, en caso que los dos eventos (ser obrero y vivir en la Couronne Sud) sean independientes. En este particular, pasa que el resultado es muy cercano al valor de p_{44}, lo que permite creer que en efecto los dos eventos podrían ser independientes.

Sin embargo, este análisis es incompleto porque cada una de las dos variables, zona de residencia y profesión, contiene más de dos categorías. Queriendo, por lo tanto, generalizar, si dos variables son independientes, es de esperarse que la probabilidad observada de pertenecer al mismo tiempo a la categoría *i* de la primera variable y a la categoría *j* de la segunda sea igual al producto de las probabilidades marginales:

$p_{ij} = p_{i\bullet} \times p_{\bullet j}$ para todos los pares *i*,*j*

Podemos llegar a la misma conclusión tomando otro camino aunque partiendo de la misma definición, o sea: "Un evento *A* es independiente de otro evento *B* cuando la probabilidad de que el evento *A* suceda siga la misma de que el evento *B* suceda o no." En el lenguaje de la teoría de las probabilidades, este enunciado equivale a decir que la probabilidad condicional de *A* es igual a su probabilidad marginal, es decir, en una tabla de contingencia de 2 dimensiones:

$$p_{i/\bullet j} = p_{i\bullet}$$

Puesto que $p_{i/\bullet j} = p_{ij} \Big/ p_{\bullet j}$, esto implica $p_{i\bullet} = p_{ij} \Big/ p_{\bullet j}$, o sea $p_{ij} = p_{i\bullet}\ p_{\bullet j}$

De una manera equivalente, las 2 variables categóricas son independientes si:

$$p_{j/i\bullet} = p_{\bullet j}$$

Puesto que $p_{j/i\bullet} = p_{ij} \Big/ p_{i\bullet}$, esto implica $p_{\bullet j} = p_{ij} \Big/ p_{i\bullet}$, o sea $p_{ij} = p_{i\bullet}\ p_{\bullet j}$.

Es ésta la definición exacta de la independencia estadística entre dos variables categóricas. Resalta a la vista que esta definición es perfectamente simétrica con relación a las dos variables. Además, es posible constatar que las frecuencias teóricas de las cuales se trato más arriba son las probabilidades que se esperaría ver en la hipótesis de la independencia estadística. En efecto:

$$x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}} = \left(\frac{x_{i\bullet}}{x_{\bullet\bullet}}\right)\left(\frac{x_{\bullet j}}{x_{\bullet\bullet}}\right) x_{\bullet\bullet} = p_{i\bullet}\ p_{\bullet j}\ x_{\bullet\bullet}$$

La tabla de frecuencias teóricas (tabla 4) es, por lo tanto, una representación exacta de la hipótesis de independencia.

Pero, ¿por qué nos interesa tanto la hipótesis de la independencia? Porque, si dos variables son independientes, se puede considerar que ninguna de las dos tiene una influencia sobre la otra (observe que, en esta formulación, no se identifica a ninguna de las dos variables para que tomen el papel de variable independiente o "explicativa").

Puede pasar que los datos se conformen perfectamente a la hipótesis de independencia. Sin embargo, es mucho más frecuente que los datos difieran de lo que predice el modelo de independencia. En caso de que no difieran "demasiado", se podrá juzgar el modelo aceptable siempre y cuando admitamos que no es más que una aproximación y que entre la realidad y el modelo, interviene un elemento aleatorio que hemos llamado "perturbación debido al azar". Las hipótesis que emitimos en cuanto a este elemento aleatorio permiten delimitar la incertidumbre en cuanto al "verdadero" modelo.

El test de hipótesis que acabamos de describir es, por lo tanto, un procedimiento de inducción estadística que apoya la decisión de rechazar o no rechazar el modelo de la independencia estadística.

4-1.3.4 Otro test: el test de la relación de verosimilitud

Se usa también la estadística de la relación de verosimilitud (con más exactitud, menos dos veces el logaritmo de la razón de las funciones de verisimilitud). Para una tabla rectangular, esta estadística se define con[186]

[186] La fórmula que enseñamos a continuación es de hecho la correcta. Difiere de la definición informal que da Upton (1981, p. 36, definición del Y^2).

$$G^2 = -2\sum_i \sum_j x_{ij} \ln\left(\frac{x^*_{ij}}{x_{ij}}\right)$$

$$G^2 = 2\sum_i \sum_j x_{ij} \ln\left(\frac{x_{ij}}{x^*_{ij}}\right)$$

Como en el caso del Chi-cuadrado de Pearson, G^2 posee una distribución asintótica χ^2 con, bajo la hipótesis de independencia, $(L - 1)\ (C - 1)$ grados de libertad.[187]

Se da un ejemplo del cálculo de esta variable-test tomando los datos de la tabla de la población activa en función de la profesión y la zona de residencia.

Tabla 7: Población activa empleada
en la Región Metropolitana de Montreal
Cálculo de la estadística de la relación de verosimilitud G^2

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	–10,737.1	15,536.0	–7,301.1	–3,307.6	8,687.8	*2,878.0*
Resto CUM	8,632.4	4,548.4	7,730.8	–11,793.9	–6,442.2	*2,675.5*
Anillo norte	1,112.0	–10,634.5	3,126.5	8,425.2	–492.4	*1,536.8*
Anillo sur	4,049.5	–4,765.5	1,517.9	362.2	–826.5	*337.6*
Fuera RMR	–1,168.8	–2,200.5	–4,057.2	9,699.2	34.0	*2,306.7*
Total	*1,888.0*	*2,484.0*	*1,016.8*	*3,385.1*	*960.7*	*9,734.6*

[187] Si bien X^2 y G^2 tienen la misma distribución asintótica, esto no implica que tengan el mismo valor.

Los valores de la tabla 7 son las contribuciones de las celdas individuales al G^2. Así, para la quinta celda de la cuarta columna,

$$25{,}495 \times \ln\left(\frac{25{,}495}{17{,}427.4}\right) = 9{,}699.2$$

El G^2 es sencillamente igual al doble de la suma de todos los elementos de esta tabla: o sea, 19,469.2. La probabilidad crítica correspondiente es inferior a 2.4×10^{-300}.

4-1.4 UN ESPECIAL VISTAZO SOBRE EL CHI-CUADRADO DE PEARSON

4-1.4.1 Las infinitas aplicaciones del test del Chi-cuadrado de Pearson a las tablas de contingencia

Test sobre una sola celda de la tabla

Es posible interpretar cada uno de los términos de la doble sumatoria que integra el Chi-cuadrado como la "contribución" de la celda correspondiente al Chi-cuadrado. Esto nos permite detectar las celdas más "desviadas" con relación a la hipótesis.

Es también posible probar de manera formal la hipótesis de que una celda especifica de la tabla es significativamente "desviada". Sólo se necesita construir una tabla donde se agregue todas las demás líneas y columnas. Por ejemplo, en caso de querer efectuar el test para la celda [*h*, *k*], se construye una tabla agregada 2×2 como lo muestra el modelo siguiente:

x_{hk}	$\sum_{j} x_{hj}$
$\sum_{i} x_{ik}$	$\sum_{i \neq h} \sum_{j \neq k} x_{ij}$

Luego, aplicamos a esta tabla el test del Chi-cuadrado de Pearson con 1 grado de libertad:

$(L - 1)\ (C - 1) = (2 - 1)\ (2 - 1) = 1$

Consideremos, por ejemplo, la fracción de los empleados que viven al exterior de la RMR y que pertenecen a las profesiones Trabajadores especializados en los servicios, personal de explotación de los transportes, etc. En la tabla 5, se puede observar que esta celda de la tabla no contribuye más que por 0.1 con el valor total del Chi-cuadrado. Podemos probar la hipótesis de que esta desviación no es significativa con relación a la hipótesis de independencia. A partir de la tabla de contingencia, se construye la tabla agregada que sigue:

Tabla 8: Tabla agregada Población activa empleada,
Región Metropolitana de Montreal, 1991
Zona residencial, según el sexo y la profesión

	Todas las professones menos →	Trabajadores especializados en los servicios, personal de explotación de los transportes, etc.	Total
RMR	1,257,720	218,919	*85,178*
Fuera de RMR	72,514	12,664	*1,476,639*
Total	*231,583*	*1,330,234*	*1,561,817*

Fuente: Statistique Canada, Censo de 1991.

El valor del chi cuadrado que se calculó a partir de esta tabla es de 0.11, lo que queda bastante alejado del 19,662.2 obtenido con la tabla detallada. La probabilidad crítica aso-

ciada a 0.11 es de 74%, lo que, claramente, no nos permite rechazar la hipótesis de independencia en la tabla agregada.

Test de homogeneidad entre dos o más grupos o muestras

A menudo sucede que tengamos que analizar tablas que comparan dos grupos de individuos distribuidos en varias categorías. En el caso de dos grupos, la tabla de contingencia tiene el aspecto siguiente:

	Grupo *A*	Grupo *B*	Total *A*+*B*
Categorías			
Total			

Un test de homogeneidad entre dos o más grupos busca determinar si, desde el punto de vista de su repartición entre las categorías de una variable de clasificación dada, ambos son o no significativamente diferentes. Podríamos pretender, por ejemplo, comparar la repartición de los hombres y de las mujeres entre las profesiones. No será muy difícil reconocer que el problema por saber si la repartición de las mujeres entre las profesiones es significativamente diferente de aquella de los hombres no es más que el problema de independencia entre la variable Profesión y la variable Sexo.

Test de homogeneidad entre una subpoblación y el resto de la población

El test de homogeneidad sirve entre otras cosas para comparar un grupo particular con el resto de la población. En particular, se emplea para comparar una muestra con la población

de donde se obtiene para saber si es representativa de algunas características conocidas de la población.

Supongamos, por ejemplo, que efectuemos un sondeo por medio de entrevistas a los residentes de Montreal, los cuales se escogen al azar en el cruce de las calles Sainte-Catherine y Jeanne-Mance. Si consideramos que la variable lingüística es importante para alcanzar el objetivo de tal estudio, tendremos que verificar, al final, si la proporción de francófonos y de anglófonos entrevistados es representativa de la proporción lingüística de Montreal. Con este fin, se construye una tabla basándose en el modelo siguiente:

	Grupo *A*	Resto de la población	Total
Categorías		Calcular por sustracción	
Total			

En nuestro ejemplo, las categorías pertinentes son obviamente Francófono, Anglófono y Otros.[188] Los datos del grupo *A* son aquellos de la muestra o de otro grupo específico del estudio; es posible obtener los datos de la columna Total en

[188] No queremos mencionar ahora la dificultad que existe para definir la pertenencia lingüística de modo operacional y aún más grande dificultad para encontrar en los datos del Censo de *Statistique Canada* la información pertinente (la formulación de las preguntas del censo con relación a la pertenencia lingüística es el objeto de abiertas y fuertes críticas).

cualquier fuente oficial como un censo. Se efectúa el cálculo de las cifras del Resto de la población por sustracción.[189]

Test de la hipótesis de una distribución particular

Generalizando, el test del Chi-cuadrado puede servir para evaluar cualquier hipótesis sobre una distribución de un conjunto de individuos entre categorías.[190] Para esto el conjunto de individuos estudiado es considerado como si fuera una muestra obtenida de una población infinita, la cual debe estar distribuida según la hipótesis que se pretende evaluar.

Por ejemplo, según el censo de la población de 1984 en Costa Rica, este país contaba entonces con 630,995 hombres y 649,619 mujeres (tabla 9). Confrontemos estas cifras con la hipótesis de una distribución 50-50 entre los sexos.

Tabla 9: Población masculina y femenina, Costa Rica, 1984

	Datos del censo	Frecuencias teóricas
Hombres	630 995	640 307
Mujeres	649 619	640 307
Total	*1 280 614*	*1 280 614*

Fuente: http://populi.eest.ucr.ac.cr/observa/estima/cuadro1.htm

Calculemos el valor del Chi-cuadrado como

$$X^2 = \frac{(630{,}995 - 640{,}307)^2}{640{,}307} + \frac{(649{,}619 - 640{,}307)^2}{640{,}307}$$

$$X^2 = 270.85$$

[189] En caso de que el grupo *A* no represente más que una mínima fracción de toda la población, es posible, en la practica, calcular el Khi-dos entre el grupo *A* y la totalidad aunque no sea teóricamente exacto.
[190] Blalock (1972, p. 312), ejercicio núm. 3.

La probabilidad crítica con 1 grado de libertad es igual a 7.4×10^{-61}, lo que lleva a rechazar la hipótesis de que la distribución de la población entre hombres y mujeres no es significativamente diferente de la distribución 50-50.

En apariencia, este procedimiento difiere de aquel empleado hasta el momento. Pero tal no es el caso, puesto que este test se fundamenta en la comparación implícita que se efectúa entre la población estudiada y una población hipotética de tamaño infinito, la cual respeta la distribución hipotética que se pretende probar. De manera explicita, presentamos a continuación la tabla de contingencia subyacente a este test.

Tabla 10: Población masculina y femenina, Costa Rica, 1984

	Población Costa Rica, 1984	Resto	Población hipotética infinita
Hombres	630 995	$0.5 \times Y - 630\,995$	$0.5 \times Y$
Mujeres	649 619	$0.5 \times Y - 649\,619$	$0.5 \times Y$
Total	1 280 614	$Y - 1\,280\,614$	Y

Fuente: http://populi.eest.ucr.ac.cr/observa/estima/cuadro1.htm

Cálculo de las frecuencias teóricas

	Población Costa Rica 1984	Resto
Hombres	$\frac{(630995 - 640307)^2}{640307}$	$\frac{\{(0{,}5\,Y - 630995) - [0{,}5\,(Y - 1280614)]\}^2}{0{,}5\,(Y - 1280614)} = \frac{(640307 - 630995)^2}{0{,}5\,(Y - 1280614)}$
Mujeres	$\frac{(649619 - 640307)^2}{640307}$	$\frac{\{(0{,}5\,Y - 649307) - [0{,}5\,(Y - 1280614)]\}^2}{0{,}5\,(Y - 1280614)} = \frac{(640307 - 649307)^2}{0{,}5\,(Y - 1280614)}$

Si Y es infinitamente grande, la contribución de la columna Resto en el valor del Chi-cuadrado es despreciable (infinitamente pequeño) puesto que el divisor 0.5 (Y – 1,280,614) es infinitamente grande de tal modo que el cálculo equivale a lo hecho anteriormente. Además, el número de grados de libertad es efectivamente igual a $(C - 1)(L - 1)$.[191]

4-1.4.2 Condiciones de validez del test del Chi-cuadrado de Pearson

El test del Chi-cuadrado de Pearson se basa en una aproximación: la distribución del $\chi 2$ es la distribución asintótica de la estadística del Chi-cuadrado de Pearson. Para que sea válido el test, tiene que ser bastante buena la aproximación. Por lo general, se considera que la aproximación es bastante buena y que el test es válido, cuando el número total de observaciones respeta la condición

$$x_{\bullet\bullet} > 10 \times L \times C$$

donde C es el número de columnas y L, el número de líneas de la tabla (Legendre y Legendre, 1998, p. 218).

En la práctica la mayoría de los autores afirman que el test del Chi-cuadrado de Pearson podría no ser válido si existen una o más celdas que contienen menos de 5 observaciones (Freund y Williams, 1973, p. 379).

Según Legendre y Legendre (1998, p. 218), el test podría no ser válido si $x_{\bullet\bullet} < 5 \times L \times C$. Esa condición es muy cercana a la anterior: cuando cumple esa condición, hay necesariamente por lo menos una celda con frecuencia teórica abajo de 5, tal como se demuestra a continuación.

[191] Puede haber situaciones en que el cálculo de las frecuencias teóricas se someta a más de una restricción; en estas condiciones, el número de grados de libertad se calcula de otra manera. Vea Blalock (1972, ejercicio 3, p. 312).

Tenemos :

$\underset{i}{MIN}\left[x_{i\bullet}\right] \leq \frac{x_{\bullet\bullet}}{L}$ et $\underset{j}{MIN}\left[x_{\bullet j}\right] \leq \frac{x_{\bullet\bullet}}{C}$, de manera que

$\underset{i,j}{MIN}\left[x_{i\bullet} x_{\bullet j}\right] \leq \frac{(x_{\bullet\bullet})^2}{L \times C}$, o sea $\underset{i,j}{MIN}\left[\frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}\right] \leq \frac{x_{\bullet\bullet}}{L \times C}$

Sigue que, si $x_{\bullet\bullet} < 5 \times L \times C$, o sea si $\frac{x_{\bullet\bullet}}{L \times C} < 5$, entonces la más pequeña frecuencia teórica $x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}}$ estará abajo de 5.

En cambio, Cochran (1954) y Siegel (1956), a los cuales se refieren Legendre y Legendre (1998, p. 218) emiten las condiciones siguiente, menos restrictivas, que invalidarían el test del Chi-cuadrado:

- Existe una o más celdas *ij* cuya frecuencia teórica x^*_{ij} es inferior a 1.

NOTA: Puesto que $x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}}$, esta condición equivale a decir que existe al menos una línea *i* y una columna *j* tales que $x_{i\bullet}\ x_{\bullet j} < x_{\bullet\bullet}$

O bien:

- Existe 20% de las celdas *ij* cuya frecuencia teórica x^*_{ij} es inferior a 5.

Ahora bien, según otros autores, hasta esta última condición parece ser un tanto severa. Legendre y Legendre (1998, p. 218) citan a Fienberg (1950), para quien el test es válido con

un umbral de significación de 5% siempre y cuando todas las frecuencias teóricas sean superiores a 1.

Concretamente, recordaremos que al momento de aplicar el test del Chi-cuadrado de Pearson a una tabla de contingencia, es importante desconfiar de los resultados cuando algunas de las frecuencias teóricas son demasiadas pequeñas.

En caso de tener buenas razones para desconfiar de la validez del test del Chi-cuadrado, ¿qué podemos hacer? Una primera posibilidad es agrupar unas categorías para así fusionar las filas o columnas que sólo contienen un pequeño número de observaciones. Se obtendrán, de esta manera, frecuencias teóricas más altas en las celdas fusionadas. Pero, ¡no basta agrupar categorías de cualquier manera! Agrupar categorías equivale a cambiar el modo de operacionalización de la hipótesis (vea el inicio del capítulo 2-2). Hay que justificarlo con relación al modelo conceptual subyacente a la investigación.

Por otra parte, a menudo es preferible descartar del análisis las categorías que dificultan la interpretación (por ejemplo, las respuestas “No sé” en los datos de encuestas). Se pensará descartar hasta categorías que sí tienen un contenido analítico, pero tienen un pequeño número de observaciones, mientras que no se pueden agrupar con otras para constituir nuevas categorías que sean pertinentes en relación con el modelo conceptual.

4-1.4.3 Algunas propiedades numéricas del test del Chi-cuadrado de Pearson

El Chi-cuadrado de Pearson posee las propiedades que, a continuación enumeramos:

1. Chi-cuadrado es no negativo.
2. Chi-cuadrado es nulo cuando $x_{ij} = x_{ij}^{*}$ para todas las celdas i,j de la tabla.

3. Chi-cuadrado aumenta con el número de observaciones $x_{\bullet\bullet}$.

4. $X^2 \leq x_{\bullet\bullet} Min(L-1, C-1)$

donde C es el número de columnas y L, el número de líneas de la tabla y donde la expresión $Min(L-1, C-1)$ representa el más pequeño valor entre $(L-1)$ y $(C-1)$.

Las dos primeras propiedades son relativamente evidentes si observamos la fórmula de cálculo

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*}$$

Se ilustra la tercera propiedad con el ejemplo que sigue:

Tabla 11: Sensibilidad del Chi-cuadrado al número de observaciones. Ilustración numérica

Tablas de contingencia

			$p_{i\bullet}$
15	10	*25*	0.5
10	15	*25*	0.5
25	*25*	*50*	
0.5	0.5	← $p_{\bullet j}$	

			$p_{i\bullet}$
30	20	*50*	0.5
20	30	*50*	0.5
50	*50*	*100*	
0.5	0.5	← $p_{\bullet j}$	

			$p_{i\bullet}$
60	40	*100*	0,5
40	60	*100*	0,5
100	*100*	*200*	
0.5	0.5	← $p_{\bullet j}$	

Frecuencias teóricas

12.5	12.5	*25*
12.5	12.5	*25*
25	*25*	*50*

25	25	*50*
25	25	*50*
50	*50*	*100*

50	50	*100*
50	50	*100*
100	*100*	*200*

Cálculo del Chi-cuadrado

0.5	0.5
0.5	0.5

Chi-cuadrado = 2
núm. de lineas = 2
núm. de col. = 2
grad. de libertad = 1
Prob. crítica = 0.157

1	1
1	1

Chi-cuadrado = 4
núm. de lineas = 2
núm. de col. = 2
grad. de libertad = 1
Prob. crítica = 0.046

2	2
2	2

Chi-cuadrado = 8
núm. de lineas = 2
núm. de col. = 2
grad. de libertad = 1
Prob. crítica = 0.005

Las tres tablas de contingencia de arriba poseen estructuras idénticas. La única cosa que las distingue es el número de observaciones, que son 25, 50 y 100 respectivamente. El test del Chi-cuadrado nos conduce a rechazar la hipótesis de independencia en el tercer caso y, aunque de manera no tan categórica, también en el segundo; por lo contrario, en el primer caso, se tomaría, por lo general, la decisión de no rechazar la hipótesis, al menos dentro de los criterios que se acostumbran usar en ciencias sociales.

Generalizando, cuando para una estructura dada el número de observaciones aumenta en proporción en toda las celdas, el valor del Chi-cuadrado aumenta en la misma proporción. De manera formal, cuando el número de observaciones es multiplicado por α, tenemos:

$$\sum_i \sum_j \frac{\left(\alpha x_{ij}^* - \alpha x_{ij}\right)^2}{\alpha x_{ij}^*} = \sum_i \sum_j \frac{\alpha^2 \left(x_{ij}^* - x_{ij}\right)^2}{\alpha x_{ij}^*}$$

$$\sum_i \sum_j \frac{\left(\alpha x_{ij}^* - \alpha x_{ij}\right)^2}{\alpha x_{ij}^*} = \alpha \left[\sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*} \right]$$

¿Por qué nos interesa esta propiedad? Porque si el número total de observaciones $x_{\bullet\bullet}$ es grande, esto nos puede incitar a rechazar la hipótesis de independencia (y por lo tanto, a considerar que las diferencias entre las distribuciones son estadísticamente significativas) cuando, por lo contrario, no son, de manera segura, científicamente significativas. Al revés, puede suceder que las diferencias reales parezcan estadísticamente no significativas si el número de observaciones es pequeño.

Supongamos, por ejemplo, que en lugar de usar los datos del Censo sobre la profesión y la zona de residencia, tomáramos una muestra de 1 sobre 1000. Supongamos, también, que por una suerte increíble, la muestra sea un buen reflejo de la población de tal manera que las frecuencias observadas fueran iguales a una milésima de las frecuencias observadas del Censo tomando en cuenta el error de redondeo que se podría cometer puesto que es imposible tener fracciones de personas en la muestra. En estas condiciones, obtendríamos la tabla siguiente:

Tabla 12: Muestra ficticia de la población activa empleada en la Región Metropolitana de Montreal
Zona de residencia según la profesión, 1991

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	56	113	120	90	76	455
Resto CUM	62	82	108	60	46	358
Anillo norte	50	59	95	75	48	327
Anillo sur	54	68	96	69	49	337
Fuera RMR	12	16	19	25	13	85
Total	234	339	438	320	232	1,562

Con los datos de esta muestra ficticia pero eminentemente representativa, el valor del Chi-cuadrado de Pearson no es más que 19.79 y la probabilidad correspondiente es de 0.23, por lo tanto ¡no es posible rechazar la hipótesis de independencia!

Regresaremos a este punto cuando tratemos las mediciones de la intensidad de la relación entre dos variables categóricas. De por sí, la cuarta propiedad del Chi-cuadrado de Pearson,

$$X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$$

interviene en la definición de algunas de estas mediciones.

Demostración de que $X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$

La demostración de esta última propiedad requiere una fórmula de cálculo del Chi-cuadrado que difiere de la fórmula que dimos anteriormente. Esta fórmula se deriva de la primera:

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left[\left(x_{ij}^*\right)^2 - 2\, x_{ij}^*\, x_{ij} + x_{ij}^2\right]}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j x_{ij}^* - 2\sum_i \sum_j x_{ij} + \sum_i \sum_j \frac{x_{ij}^2}{x_{ij}^*}$$

$$X^2 = x_{\bullet\bullet} - 2x_{\bullet\bullet} + \sum_i \sum_j \frac{x_{ij}^2}{\left(\frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}\right)}$$

$$X^2 = x_{\bullet\bullet}\left[\sum_i \sum_j \frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} - 1\right]$$

Para demostrar la cuarta propiedad, sólo basta constatar que,

por una parte $\frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} \leq \frac{x_{ij}}{x_{i\bullet}}$

de tal manera que

$$\sum_{i=1}^{L} \sum_{j=1}^{C} \frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} \leq \sum_{i=1}^{L} \sum_{j=1}^{C} \frac{x_{ij}}{x_{i\bullet}} = \sum_{i=1}^{L} \frac{x_{i\bullet}}{x_{i\bullet}} = L\,,$$

y por otra, $\frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} \leq \frac{x_{ij}}{x_{\bullet j}}$

de tal manera que

$$\sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} \le \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}}{x_{\bullet j}} = \sum_{i=1}^{C}\frac{x_{\bullet j}}{x_{\bullet j}} = C$$

Se deduce que

$$X^2 = x_{\bullet\bullet}\left[\sum_i\sum_j \frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} - 1\right] \le x_{\bullet\bullet}\left[L-1\right]$$

y

$$X^2 = x_{\bullet\bullet}\left[\sum_i\sum_j \frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} - 1\right] \le x_{\bullet\bullet}\left[C-1\right]$$

Es lo que queríamos demostrar.

4-1.4.4 Post scriptum: una nueva mirada sobre el cociente de localización

En el capítulo 1-2 presentamos el cociente de localización como un instrumento que sirve para analizar una tabla del empleo por rama y por ciudad o región. Además, mencionamos que este tipo de tabla es una tabla de contingencia (de dos dimensiones). Es posible, por lo tanto, usar el mismo cálculo para cualquier tabla de contingencia que tenga dos dimensiones (aunque el término "localización" sea un tanto incongruente en algunas ocasiones).

Aun más interesante es poder reexaminar el cociente de localización bajo la luz de los tests de hipótesis aplicados a las tablas de contingencias. Concretamente, existe una relación muy simple entre los cocientes de localización y los números que se esperan de la hipótesis de independencia. Se definen estos últimos con

$$x_{ij}^* = \frac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}}$$

En cuanto a los cocientes de localización, se calculan con la fórmula

$$QL_{ij} = \frac{\left(x_{ij} \Big/ x_{\bullet j} \right)}{\left(x_{i\bullet} \Big/ x_{\bullet\bullet} \right)},$$

lo que es equivalente, con $QL_{ij} = \dfrac{\left(x_{ij} \Big/ x_{i\bullet} \right)}{\left(x_{\bullet j} \Big/ x_{\bullet\bullet} \right)}$

sea, $QL_{ij} = \dfrac{x_{ij}}{\left(\dfrac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}} \right)} = \dfrac{x_{ij}}{x^*_{ij}}$

El cociente de localización es, por consiguiente, la razón de la frecuencia observada entre la frecuencia teórica bajo la hipótesis de independencia; como lo vimos, esta hipótesis se traduce con una tabla biproporcional. Esta relación permite también expresar el Chi-cuadrado de Pearson en función de los cocientes de localización. En efecto, puesto que tenemos

$$x^*_{ij}\ QL_{ij} = x_{ij}$$

tenemos también

$$X^2 = \sum_i \sum_j x^*_{ij} \left(QL_{ij} - 1 \right)^2$$

Demostración de $X^2 = \sum_i \sum_j x^*_{ij} \left(QL_{ij} - 1 \right)^2$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij} - x^*_{ij} \right)^2}{x^*_{ij}}$$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* \; QL_{ij} - x_{ij}^*\right)^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left[x_{ij}^* \left(QL_{ij} - 1\right)\right]^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^*\right)^2 \left(QL_{ij} - 1\right)^2}{x_{ij}^*}$$

El Chi-cuadrado de Pearson es una suma ponderada de los cuadrados de las desviaciones de los cocientes de localización con relación al valor de referencia 1; en particular, el peso de cada celda es la frecuencia teórica bajo la hipótesis de independencia.

Tratándose del estudio de una tabla de empleo por ramo y por ciudad o región, parece claro que casi siempre el test del Chi-cuadrado desembocará en rechazar de manera categórica la hipótesis de independencia. Así, el verdadero interés de examinar esta relación es poder interpretar cada uno de los términos de la doble sumatoria como la contribución de la celda correspondiente al Chi-cuadrado. En términos relativos, la razón

$$\frac{x_{ij}^* \left(QL_{ij} - 1\right)^2}{X^2}$$

es la parte de la desviación total (con relación a la biproporcionalidad) que se puede atribuir a la celda *i*,*j*.

4-1.5 MEDICIONES DE LA INTENSIDAD DE LA RELACIÓN ENTRE DOS VARIABLES CATEGÓRICAS

Cuando rechazamos la hipótesis de independencia, esto significa que decidimos que existe una relación estadísticamente significativa entre las dos variables. No obstante, esta relación estadísticamente significativa no es necesariamente pertinente o importante desde un punto de vista científico o práctico. Se destaca sobremanera la necesidad de esta distinción, en particular, cuando examinemos la tercera propiedad que analizamos en 4-3 (el Chi-cuadrado aumenta con el número de observaciones). De aquí, lo útil que representa medir la intensidad de la relación entre las dos variables categóricas.

4-1.5.1 Mediciones derivadas del Chi-cuadrado de Pearson

Como pudimos ver, el Chi-cuadrado de Pearson posee algunas propiedades numéricas no deseables como medición de la intensidad de la relación entre dos variables categóricas:

X^2 aumenta con el número de observaciones $x_{\bullet\bullet}$

$$X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$$

donde *C* es el número de columnas y *L*, el número de líneas de la tabla.

Nos interesa una medición que refleje la estructura más que el número de observaciones y que, en condiciones ideales, variaría entre 0 y 1 en lugar de entre 0 y $x_{\bullet\bullet}\, Min(L-1, C-1)$.

A continuación, enlistamos algunas mediciones que se derivan del Chi-cuadrado de Pearson.

1. $\varphi^2 = \frac{X^2}{x_{\bullet\bullet}}$

Este "Fi-cuadrado" (el símbolo φ es la letra griega "Fi") varía entre 0 y 1 para las tablas 2 x 2, pero, en general, su valor máximo es, de manera considerable, mucho más elevado.

2. El T^2 de Tschuprow:

$$T^2 = \frac{\varphi^2}{\sqrt{(L-1)(C-1)}} = \frac{X^2}{x_{\bullet\bullet}\sqrt{(L-1)(C-1)}}$$

Su valor máximo es igual a 1 si $L = C$; de otra manera, es estrictamente inferior a 1.

3. El V^2 de Cramer

$$V^2 = \frac{\varphi^2}{Min(L-1, C-1)} = \frac{X^2}{x_{\bullet\bullet} Min(L-1, C-1)}$$

El V^2 de Cramer es equivalente al T^2 de Tschuprow cuando $L = C$, pero al opuesto de este último, puede tomar el valor máximo de 1 cuando $L \neq C$.

4-1.5.2 Otras mediciones (tau y lambda)

Principio general

El tau de Goodman y Kruskal (del nombre de la letra griega τ, que tiene el valor fonético de "T") y el lambda (del nombre de la letra griega λ que tiene el valor fonético de "L") no son simétricos puesto que sus mediciones se basan en una distinción entre la variable dependiente y la variable independiente. En caso que la variable dependiente corresponde a las

categorías *j* (columnas), el tau y el lambda miden la intensidad de la relación por medio de la reducción relativa promedio de los errores de asignación que se efectúan al momento de predecir a cuál categoría *j* pertenece un individuo cuando se sabe a cuál categoría *i* pertenece. Su expresión general es por lo tanto:

$$1 - \frac{\begin{array}{l}\text{Número promedio de errores de}\\ \text{asignación cuando se conoce } i\end{array}}{\begin{array}{l}\text{Número promedio de errores de}\\ \text{asignación cuando no se conoce } i\end{array}}$$

Las dos mediciones difieren en cuanto a la regla que se respeta para predecir a cual categoría *j* pertenece el individuo.

El tau de Goodman y Kruskal

Regla de asignación. Se distribuyen los individuos entre las categorías *j* de manera proporcional a los $p_{\bullet j}$ cuando no se conoce *i*, y de manera proporcional a los p_{ij} cuando se conoce *i*.

Fórmula.

$$\tau_J = 1 - \frac{\sum_i \sum_j p_{ij}\left(1 - \frac{p_{ij}}{p_{i\bullet}}\right)}{\sum_j p_{\bullet j}(1 - p_{\bullet j})} = 1 - \frac{\sum_i \sum_j p_{ij}\left(1 - p_{j/i\bullet}\right)}{\sum_j p_{\bullet j}(1 - p_{\bullet j})}$$

Valores límites. El tau es igual a cero cuando las dos variables son perfectamente independientes, o sea cuando $p_{ij} = p_{i\bullet}\ p_{\bullet j}$. Es igual a 1 cuando, en cada línea *i* de la tabla, existe solamente una sola celda no nula, lo que permite predecir *j* con certeza cuando se conoce *i*.

La lambda

Regla de asignación. En caso de no saber a cuál categoría *i* pertenecen los individuos, se asignan todos en la categoría que contiene el más grande número de observaciones, o sea la categoría con la probabilidad marginal $p_{\bullet j}$ más grande; en caso de conocer *i*, se asignan los individuos en la categoría *j* que contiene el más grande número de observaciones en el interior de la categoría *i*, o sea la categoría con la probabilidad condicional $p_{j/i\bullet}$ más grande.

Fórmula.

$$\lambda_J = 1 - \frac{\sum_i \left(1 - \frac{p_{i,Max}}{p_{i\bullet}}\right) p_{i\bullet}}{(1 - p_{\bullet Max})} = 1 - \frac{\sum_i (1 - p_{Max/i\bullet})\, p_{i\bullet}}{(1 - p_{\bullet k})}$$

,

donde

$$p_{\bullet Max} = \underset{j}{Max}\, p_{\bullet j}\,,\ p_{i,Max} = \underset{j}{Max}\, p_{ij}$$

y

$$p_{Max/i\bullet} = \underset{j}{Max}\, p_{j/i\bullet}$$

Valores límites. El lambda es igual a 1 cuando, en cada línea *i* de la tabla, existe una única celda no nula y que $p_{i,Max} = p_{i\bullet}$, lo que permite predecir *j* con certeza cuando se conoce *i*. Es igual a cero cuando $p_{i,Max} = p_{i\bullet}\ p_{\bullet Max}$ para todo *i*, aunque las variables no sean independientes, es decir, aunque, para las columnas menos aquella donde se encuentra

$p_{i,Max} = \underset{j}{Max}\, p_{ij}$, tengamos $p_{ij} \neq p_{i\bullet}\ p_{\bullet j}$. Es por esta última propiedad que se prefiere tau.

4-1.6 LAS VARIABLES DE CONTROL EN LAS TABLAS CON MÁS DE DOS DIMENSIONES

La tabla 1 tenía tres dimensiones: el sexo, la zona de residencia y la profesión. Hasta el momento, nuestro análisis consideró las dos últimas dimensiones ignorando la posibilidad de que existan diferencias entre los hombres y las mujeres. Sin embargo, es muy probable que la tabla de contingencia zona de residencia-profesión sea muy diferente para las mujeres y los hombres. El rigor científico nos forza a tomar en cuenta esta posibilidad.

Generalizando, cuando se examina la relación entre dos variables categóricas, es importante preguntarse si otras variables no pudiesen influenciar la intensidad o la forma de esta relación. Y si tal fuera el caso, es necesario tomar en cuenta estas nuevas variables, conocidas como variables de control. Esta expresión proviene del lenguaje de las ciencias experimentales, cuando las condiciones del laboratorio permiten "controlar" el nivel de las variables que podrían influenciar la relación en el estudio. Por ejemplo, en caso de probar un medicamento en ratas y de pensar que la alimentación podría influenciar el rendimiento del medicamento, se efectuarán pruebas sobre diferentes grupos con diferentes regímenes alimenticios "controlados".

Un modo simple de tomar en cuenta estas variables de control es examinar la relación en el estudio (con tests de hipótesis y medición de la intensidad de la relación) de manera separada para cada grupo homogéneo de individuos. En nuestro ejemplo, esto significaría examinar dos tablas de contingencias, una para las mujeres y otra para los hombres. Sin embargo, este procedimiento tiene límites. En particular,

cuando existen variables de control con varias categorías cada una, el número de tablas de contingencia para analizar aumenta rápidamente. Por ejemplo, si pretendemos considerar el sexo y la edad, con cinco grupos de edad, es necesario analizar diez tablas de contingencia. Además, cuando el número de observaciones es limitado, es posible que el número de observaciones sea demasiado pequeño como para confiar en la validez de los tests (por ejemplo, con 1000 observaciones, con 5 profesiones, 5 zonas de residencia, 5 grupos de edad y 2 sexos, tendremos frecuencias teóricas de solamente 4 en promedio, y es muy probable que algunas celdas tuvieran frecuencias teóricas abajo de 1).

Visto desde otra perspectiva, el problema es el de la multiplicidad de las interacciones posibles, la cual aumenta rápidamente con el número de variables (dimensiones de la tabla). Así, en una tabla con dos dimensiones, no existe más que una interacción posible y, por consiguiente, existe solamente una hipótesis de independencia para probar. En una tabla con tres dimensiones, existen cuatro interacciones posibles, es decir tres entre pares de variables y una entre las tres variables al mismo tiempo. En el caso de una tabla con cuatro dimensiones, hay 17 interacciones posibles (cuatro por cada una de las tercias posibles entre las cuatro variables, más una cuádruple interacción que implica todas las variables)...

El modelo log-lineal constituye un marco que permite examinar las diferentes interacciones posibles. El modelo "saturado", el cual incluye todas las interacciones posibles, reproduce perfectamente los datos observados. Una versión generalizada del test de hipótesis de independencia permite seleccionar, entre las numerosas interacciones posibles, aquellas que debemos guardar para representar la estructura subyacente.

Para profundizar en este tema...

Es posible consultar Upton (1981) y encontrar una presentación informal y pragmática del modelo log-lineal, así como un ejemplo de su uso en el contexto de las ciencias regionales. Button *et al*. (1995) ofrecen un ejemplo más reciente de uso del modelo log-lineal.