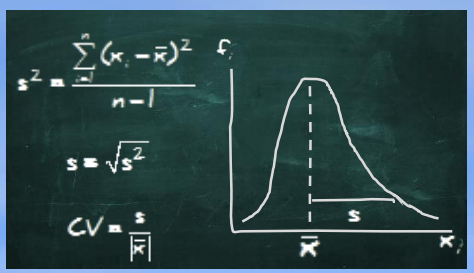
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.6
Análisis de tablas de contingencia

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2015). Análisis de tablas de contingencia. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. 1ª edición. Edición digital: http://ddd.uab.cat/record/131469

Capítulo redactado en abril de 2015

Índice de contenidos

PARTE III
Capítulo 6

Análisis de tablas de contingencia

El análisis de tablas de contingencia (ATC) es una técnica destinada al estudio de la relación entre dos o más variables cualitativas o categóricas, es decir, medidas a nivel nominal y ordinal. En el capítulo anterior vimos que existen diversas técnicas estadísticas que tratan también el análisis de la relación entre variables cualitativas como el análisis log-lineal o el análisis de correspondencias. El análisis de las tablas de contingencia se puede considerar como una técnica de base destinada a la lectura y estudio de las relaciones entre unas pocas variables, entre dos y tres habitualmente, que se utiliza en un ejercicio de análisis descriptivo de sus relaciones así como tratamiento previo para cualquier análisis más complejo de relaciones multidimensionales.

El ATC es una de las técnicas de análisis estadístico más habitual en los trabajos sociológicos y, en general, de tratamiento de los datos de encuesta, donde es habitual que la mayoría de las variables sean cualitativas. Son muchos los ejemplos que se podrían presentar, destacaremos las encuestas del Centro de Investigaciones Sociológicas[1] y la Enquesta de Condiciones de Vida i Hàbits de la Població de Catalunya del Institut d'Estudis Regionals i Metropolitans de Barcelona[2]. Por supuesto los institutos de estadística de los diferentes territorios y otras instituciones y organismos nacionales e internacionales de producción de información estadística son referencias igualmente de la publicación y difusión de este tipo de información[3].

Analizar la relación entre dos o más variables a partir de una taula de contingencia nos conducirá a adquirir la habilidad de lectura de este tipo de información y a interpretar los datos que aparecen en la tabla a partir de los cálculos de porcentajes que se pueden obtener en cada casilla de una tabla. Así podremos determinar la existencia y la naturaleza de la relación de asociación entre las variables consideradas. En segundo

[1] En la página web del CIS (http://www.cis.es) se puede acceder a numerosa información y a las bases de datos. Se puede consultar http://pagines.uab.cat/plopez/content/bases-de-datos-para-la-investigaci%C3%B3n-y-la-docencia#cis.

[2] Desde el año 1985 numerosos análisis y publicaciones han visto la luza en las 6 ediciones de la encuesta. Se pueden consultar las últimas publicaciones de datos que se presentan sistemáticamente con formato de tablas de contingencia en la página http://www.iermb.uab.es/htm/descargaBinaria.asp?idPub=197 y también en http://www.iermb.uab.es/htm/descargaBinaria.asp?idPub=225.

[3] En http://pagines.uab.cat/plopez/content/bases-de-datos-para-la-investigaci%C3%B3n-y-la-docencia se pueden consultar numerosos enlaces a fuentes de información.

lugar, esta lectura e interpretación inicial de la posible asociación entre las variables requerirá una fundamentación estadística que se establecerá mediante la prueba de chi-cuadrado cuyo resultado determinará la significación estadística de la relación. En tercer lugar, el análisis se completa con el cálculo de otros estadísticos destinados a establecer la fuerza de la asociación, ya sea a nivel global, entre variables, o nivel local, en casillas concretas para combinaciones de categorías o valores concretos de estas variables.

Este tipo de análisis lo aplicaremos primero al estudio de las relaciones entre dos variables y lo extenderemos al caso de la introducción de una tercera variable que cumplirá en particular el papel de variable de control de una relación bivariable. De esta forma completaremos lo que podemos denominar como el análisis clásico de las tablas de contingencia. Completaremos esta exposición con dos aspectos de interés que complementan este tipo de análisis: el análisis de diferencias de proporciones y el análisis de razones. El primero dará pie a establecer la significación de las diferencias observadas entre porcentajes, en particular en una tabla de contingencia, y el segundo nos introducirá en un cálculo alternativo de la información de la tabla que nos conduce a fundamentar el razonamiento de base de los denominados modelos log-lineales que veremos en el capítulo siguiente.

El análisis log-lineal nos facilitará el análisis riguroso de la relaciones entre un conjunto reducido de variables, en general no más de cinco, que permitirá superar algunas de las limitaciones que tiene el análisis clásico de tablas de contingencia. Con esa técnica veremos diversos modelos de análisis de enorme interés para la investigación pero caracterizados por el manejo de un número reducido de variables. Una técnica alternativa también de enorme interés y utilidad en el análisis de las variables cualitativas es el análisis de correspondencias, que veremos en el capítulo de análisis factorial, y donde se podrán tener en consideración hasta decenas de variables simultáneamente en un mismo ejercicio de análisis.

Todas estas técnicas permiten además plasmar modelos de análisis destinados tanto al análisis de relaciones de dependencia como de interdependencia. El análisis clásico de tablas de contingencia, si bien nos permitirá razonar en términos de variable dependiente e independiente, se trata de un análisis estadístico esencialmente de carácter simétrico. A este mismo tipo de modelos de interdependencia pertenecen el análisis log-lineal general y el análisis de correspondencias. Cuando nos planteamos un modelo de dependencia que explicita los factores explicativos de otra variable resultado, entonces consideraremos el denominado análisis log-lineal logit o bien el análisis de regresión logística.

Al final del capítulo se incluyen los apartados destinados a precisar cómo obtener tablas de contingencia y representaciones gráficas de las mismas mediante los softwares SPSS y R.

1. Definición y nomenclatura de una tabla de contingencia

Una tabla de contingencia es una tabla de frecuencias que resulta de la distribución conjunta al relacionar o cruzar dos o más variables cualitativas. Iniciaremos nuestro recorrido en el tratamiento de tablas bidimensionales, desarrollando todos los conceptos matemáticos necesarios y adquiriendo la habilidad de la lectura e interpretación de los resultados de este tipo de técnica, a continuación extenderemos ese aprendizaje al análisis multidimensional mediante la introducción de terceras o cuartas variables.

Mediante el estudio de las características de la distribución conjunta de dos variables damos un primer paso en la complejidad de las técnicas de análisis de datos que va más allá del análisis univariable, enriqueciendo así los resultados y las conclusiones del estudio. En la estadística descriptiva univariable disponemos tan solo de la información de la distribución de frecuencias de cada variable por separado; con el ATC describiremos también las variables de interés pero añadiendo la riqueza informativa de la relación con otras variables lo que nos permite estudiar las condiciones que influyen en la distribución de una variable en relación a esas otras consideradas como relevantes, la tarea más frecuente que el investigador/a y que se deriva de los planteamientos de sus modelos de análisis.

Así, por ejemplo, podríamos plantearnos analizar si la satisfacción en el trabajo productivo se relaciona, depende, de la categoría profesional que se ocupa, de la variedad del trabajo realizado, del salario percibido, del tipo de empresa donde se desenvuelve, de la antigüedad, o del sexo, o de la edad, etc. Nos podemos preguntar si el conocimiento o el uso de una lengua están asociados con el origen geográfico, con la antigüedad, con el nivel de estudios, con la categoría socioeconómica, con la edad, con el territorio de residencia, etc. Cada uno de estos vínculos entre la variable de interés o variable dependiente (la satisfacción) con cada una de las variables independientes (la categoría, la variedad, etc.) genera una tabla de contingencia distinta bivariable que nos proporciona resultados parciales que nos facilitan obtener interpretaciones y conclusiones para nuestro estudio del fenómeno.

Una tabla de contingencia con dos variables (de dos dimensiones) es una tabla de doble entrada que relaciona dos variables cualitativas (medidas a nivel nominal u ordinal o que son tratadas en esa escala de medición) dando lugar a la distribución conjunta de frecuencias dispuestas en filas y en columnas según las categorías o valores de cada una de las variables, con tantas celdas como combinaciones de categorías o valores de ambas variables haya.

En la Tabla III.6.1 se presenta un ejemplo de tabla de contingencia[4] donde se relaciona la posesión de coche, sí se tiene o no se tiene, con la clase social, construida a partir de la categoría profesional y agrupada en tres valores: clase alta, clase media y clase baja.

[4] Los datos se han extraído de la *Enquesta Metropolitana sobre Condicions de Vida i Hàbits de la Població de la Regió Metropolitna de Barcelona 1990*, elaborada por el Institut d'Estudis Metropolitans de Barcelona. El formato de presentación sería apropiado para un informe o artículo de investigación.

Tabla III.6.1. La posesión de coche según la clase social*

Posesión de coche	**Clase social**			Total
	Alta	Media	Baja	
Sí	91,0% (650)	78,7% (1234)	58,0% (1430)	69,8% (3314)
No	9,0% (64)	21,3% (333)	42,0% (1036)	30,2% (1433)
Total	100,0% (714)	100,0% (1567)	100,0% (2466)	100,0% (4747)

Fuente: *Enquesta Metropolitana de Barcelona*, 1990.
* Los datos no incluyen ni los NS/NC ni los que han declarado "No tengo categoría". χ^2 =375,583 (0,000), V de Cramer = 0,281.

El cruce de ambas variables genera una tabla cruzada de dimensión 2×3, es decir, con un total de 6 casillas en cada una de las cuales aparece la frecuencia absoluta, entre paréntesis, debajo de la frecuencia relativa, el porcentaje. La tabla se completa con el total por filas y por columnas. De la observación de esa información se constata, en primer lugar, que las personas que poseían coche en área metropolitana de Barcelona en el año 1990 representaba el 70% de la población[5], y en segundo lugar, que esa distribución global no se corresponde con una situación igualitaria entre las diferentes clases sociales, se concluye que las clases altas poseen en mayor proporción coche que las clases medias y sobre todo que las bajas, es decir, que a medida que aumenta la clase social aumenta la posesión de coche.

Veamos con mayor detenimiento la información de una tabla de contingencia y la notación que emplearemos. A partir de una tabla de contingencia de dos dimensiones, constituida de I filas, indexadas por i, con $i = 1...I$, y de J columnas, indexadas por j, con $j = 1...J$, que cruza dos variables cualitativas Y y X.

N(I,J)

Y \ X	1	2	…	j	…	J	Total
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1J}	n_{1+}
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2J}	n_{2+}
⋮	⋮	⋮		⋮		⋮	⋮
i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{iJ}	n_{i+}
⋮	⋮	⋮		⋮		⋮	⋮
I	n_{I1}	n_{I2}	…	n_{Ij}	…	n_{IJ}	n_{I+}
Total	n_{+1}	n_{+2}	…	n_{+j}	…	n_{+J}	n_{++}

[5] Los datos de la encuesta en 2011 para la región metropolitana arroja prácticamente el mismo dato, que el 68,3% tiene automóvil.

De la distribución conjunta de dos variables resultan dos tipos de frecuencias: absolutas y relativas. Consideramos en primer lugar la información de la tabla de frecuencias absolutas $N(I,J)$ con la notación que se observa en la tabla adjunta[6].

El ejemplo que hemos presentado corresponde a una tabla de 2 filas, indexadas por *i*, con $i=1...2$, y de 3 columnas, indexadas por *j*, con $j=1...3$, que cruza dos variables cualitativas *Posesión de coche* (*Coche*) y *Clase social* (*Clase*), siendo la tabla de frecuencias absolutas $N(2,3)$:

Tabla III.6.2. La posesión de coche según la clase social
N(2,3): **frecuencias absolutas observadas**

Clase / Coche	*Alta* 1	*Media* 2	*Baja* 3	Total
Sí 1	650	1234	1430	3314
No 2	64	333	1036	1433
Total	714	1567	2466	4747

Distribución marginal de la variable en filas

Distribución marginal de la variable en columnas

Total de casos

Las frecuencias absolutas son:

- La frecuencia absoluta de cada casilla o celda que surge de la distribución conjunta por combinación de dos valores: número de casos que comparten dos características a la vez (n_{ij}).

- La frecuencia absoluta total de cada valor o categoría de la variable. El conjunto de estos valores dan la distribución marginal absoluta: número de casos que tienen una característica, de fila (n_{i+}) o de columna (n_{+j}).

- El total de casos analizados es n o n_{++}, de una muestra de la población, o del total de unidades de la población.

[6] Para facilitar la identificación de la información de la tabla se emplea el color verde para las frecuencias absolutas de cada casilla, el azul para referirse a las filas, el granate para las columnas y el negro para el total.

La tabla de frecuencias relativas $P(I,J)$ se presenta con la siguiente notación:

P(I,J)

Y \ X	1	2	…	j	…	J	Total
1	p^F_{11} p^C_{11} p_{11}	p^F_{12} p^C_{12} p_{12}	…	p^F_{1j} p^C_{1j} p_{1j}	…	p^F_{1J} p^C_{1J} p_{1J}	p_{1+}
2	p^F_{21} p^C_{21} p_{21}	p^F_{22} p^C_{22} p_{22}	…	p^F_{2j} p^C_{2j} p_{2j}	…	p^F_{2J} p^C_{2J} p_{2J}	p_{2+}
⋮	⋮	⋮		⋮		⋮	⋮
i	p^F_{i1} p^C_{i1} p_{i1}	p^F_{i2} p^C_{i2} p_{i2}	…	p^F_{ij} p^C_{ij} p_{ij}	…	p^F_{iJ} p^C_{iJ} p_{iJ}	p_{i+}
⋮	⋮	⋮		⋮		⋮	⋮
I	p^F_{I1} p^C_{I1} p_{I1}	p^F_{I2} p^C_{I2} p_{I2}	…	p^F_{Ij} p^C_{Ij} p_{Ij}	…	p^F_{IJ} p^C_{IJ} p_{IJ}	p_{I+}
Total	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+J}	p_{++}

Las frecuencias relativas, expresadas bien como proporciones (el tanto por uno) o bien como porcentajes (el tanto por ciento), son:

- La frecuencia relativa o porcentaje de cada casilla. Se pueden calcular 3 porcentajes en cada una:

  - Porcentaje total (p_{ij}): número casos de cada casilla dividido por el total de casos n (y multiplicado por 100 en el caso del porcentaje).

    Tabla III.6.3. La posesión de coche según la clase social
    T(2,3): Porcentajes sobre el total

Clase / Coche	*Alta* 1	*Media* 2	*Baja* 3	Total
Sí 1	13,7 %	26,0 %	30,1 %	**69,8 %**
No 2	1,3 %	7,0 %	21,8 %	**30,2 %**
Total	**15,0 %**	**33,0 %**	**51,9 %**	**100 %**

  - Porcentaje por fila (p^F_{ij}): número casos de cada casilla sobre el total de casos de la fila. El conjunto de estos valores se denomina distribución condicional de filas.

Tabla III.6.4. La posesión de coche según la clase social
F(2,3): Porcentajes por fila

Clase / Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	19,6 %	37,2 %	43,2 %	**100 %**
No 2	4,5 %	23,2 %	72,3 %	**100 %**
Total	**15,0 %**	**33,0 %**	**51,9 %**	**100 %**

- Porcentaje por columna (p_{ij}^{C}): número casos de cada casilla sobre el total de casos de la columna. El conjunto de estos valores se denomina distribución condicional de columnas.

Tabla III.6.5. La posesión de coche según la clase social
C(2,3): Porcentajes por columna

Clase / Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	91,0 %	78,7 %	58,0 %	**69,8 %**
No 2	9,0 %	21,3 %	42,0 %	**30,2 %**
Total	**100 %**	**100 %**	**100 %**	**100 %**

El contenido estadístico de los tres tipos de porcentajes es el mismo, pero cada uno acentúa una distribución distinta y ofrece comparaciones distintas, según el sentido de la predicción. La utilización de los porcentajes nos permitirá eliminar la influencia del tamaño de la muestra y de los marginales, con lo que podremos realizar comparaciones entre valores, entre las distribuciones condicionales, y esta comparación nos indicará la existencia de relación o no entre las variables, así como la naturaleza de la relación.

- La frecuencia relativa total de cada valor o categoría de la variable (el conjunto de estos valores dan la distribución marginal relativa): porcentaje de casos que tienen una característica, de fila o de columna, sobre el total de casos n (p_{i+} o p_{+j}).

Los diferentes cálculos que se pueden realizar para obtener las sumas de frecuencias absolutas y los porcentajes de una taula de contingencia se expresan a través de las fórmulas siguientes:

a) Fórmulas para la suma de frecuencias absolutas:

n_{i+} es la suma de una fila

$\sum_{j=1}^{J} n_{1j} = n_{1+}$ para la fila $i=1$

$\sum_{j=1}^{J} n_{ij} = n_{i+}$ para cualquier fila i

$\sum_{i=1}^{I} n_{i+} = n$ es la suma total

n_{+j} es la suma de una columna

$\sum_{i=1}^{I} n_{i1} = n_{+1}$ para la columna $j=1$

$\sum_{j=1}^{J} n_{ij} = n_{+j}$ para cualquier columna j

$\sum_{j=1}^{J} n_{+j} = n$ es la suma total

b) Fórmulas para el cálculo de proporciones (o porcentajes multiplicadas por 100):

$p_{ij}^{F} = \frac{n_{ij}}{n_{i+}}$ proporción por fila

$p_{ij}^{C} = \frac{n_{ij}}{n_{+j}}$ proporción por columna

$p_{i+} = \frac{n_{i+}}{n}$ proporción marginal de fila

$p_{+j} = \frac{n_{+j}}{n}$ proporción marginal de columna

$p_{ij} = \frac{n_{ij}}{n}$ proporción total

2. Independencia y asociación. Lectura de una tabla

Las tablas de contingencia se pueden considerar como un instrumento de carácter cuantitativo fundamental por el analista social que le proporciona una técnica de lectura rápida de las relaciones entre fenómenos. Con ella se trata de determinar si existe relación (asociación) entre las variables, y cómo es esta relación, o si por el contrario no se da y podemos hablar de independencia entre las variables.

En un análisis de una tabla de contingencia con dos variables (distribución bivariable o de dos dimensiones) se puede examinar la distribución de una de las variables (la considerada como dependiente) según los valores –o dentro de cada una de las categorías– de la otra variable (la considerada como independiente). La lectura de una misma tabla dependerá del punto de vista de la predicción, de la direccionalidad de la relación que el investigador/a sustantivamente necesita establecer.

En algunos casos esta distinción es más evidente: si relacionamos la categoría profesional con el género, es el hecho de ser varón o mujer lo que determina alcanzar mayores o menores posiciones ocupacionales, y no al revés, que según la ocupación se es varón o mujer. En otros casos ambas lecturas podrían ser de interés: cuando analizamos la relación entre la posesión de automóvil y la clase social queda claro que la posición social es la que condiciona el consumo de un bien como el automóvil. Es el caso que presentamos en la Tabla III.6.1. Pero un vendedor de coches en una sociedad capitalista con consumo de masas tiene un interés particular en constatar que las clases bajas son más numerosas, y aunque posean en menor medida coche (el hecho relevante sociológicamente hablando), son más poseyendo y ése aspecto puede ser

relevante en una economía de mercado, es decir, la "estructura social" de los poseedores de coche está "dominada" masivamente por las clases bajas y son muchos potenciales compradores. En otros casos la misma variable puede desempeñar ambos papeles de variable dependiente e independiente: el nivel de ingresos, como dependiente, lo podemos explicar en función de la ocupación, del nivel de estudios, de la antigüedad, etc., pero también el nivel de ingresos es un factor determinante del comportamiento de otros fenómenos como el consumo, la ideología, el nivel educativo de los hijos/as, etc.

Por tanto, al plantearnos el análisis de la relación entre dos variables cualitativas en una tabla de contingencia lo primero que debemos establecer es, sustantivamente, la direccionalidad, qué variable es la dependiente, qué quiero explicar. Eso significa formular una hipótesis para ser contrastada en un análisis de tablas de contingencia. Según la hipótesis, y la direccionalidad afirmada, la lectura de la tabla y la comparación de los porcentajes que comporta serán diferentes: o bien compararemos porcentajes o distribuciones condicionales por fila o bien por columna.

Al considerar los porcentajes para leer la relación entre las variables existe una regla general para determinar la elección de qué porcentaje utilizar: si se considera a una de las variables como factor explicativo (variable independiente) de la distribución de la otra variable (variable dependiente) entonces los porcentajes se calcularán en el sentido de la variable o factor causal. Así, la suma de los porcentajes en cada categoría de la variable independiente referidos al total marginal de esta categoría, tiene que dar el 100%.

Esta misma regla se puede formular de la siguiente manera: si colocamos la variable dependiente en filas y la variable independiente en columnas, para hacer las comparaciones se calcularán los porcentajes por columna[7].

De hecho, esta regla de causa-efecto no pone de manifiesto la existencia de una causalidad real, sino la decisión de considerar simplemente que una variable afecta a la distribución de la otra. En la relación del ejemplo que estamos considerando, Coche es la variable dependiente y Clase la independiente, es en función de la clase social que se posee o no coche, por tanto, la Clase "determina o causa" la posesión de coche, o lo que es lo mismo, la posesión de coche depende de la clase social de la persona; es una hipótesis que se puede formular y evidenciar a través de una tabla de contingencia que cuanto más alta es la clase social más probabilidades habrá de poseer coche. Para mostrar esta relación consideramos pues la tabla de contingencia donde se sitúan en filas las categorías de la variable Coche, la dependiente, y en columnas las de la variable Clase, la independiente, y consideramos en consecuencia los porcentajes por columna.

El análisis y la interpretación de una tabla de contingencia se basa en las comparaciones que se efectúan contrastando las diferencias entre los porcentajes de cada categoría de la variable independiente entre sí, para cada una de las categorías de la variable dependiente, y también en relación al total marginal, que viene a representar el promedio de los porcentajes.

[7] De lo que se deriva que elegiríamos los porcentajes por fila si la variable independiente se coloca en las filas.

Entre dos variables existe asociación cuando la distribución de la variable dependiente difiere entre las diversas categorías de la variable independiente, es decir, cuando las distribuciones condicionales por columna difieren entre sí. Si no existiera asociación entre las variables la distribución porcentual marginal de la variable dependiente se reproduciría en cada una de las categorías de la variable independiente (propiedad de homogeneidad de las proporciones condicionales). Desde el momento en que las distribuciones condicionales se alejen de la "media" se produce una desviación respecto de la independencia.

Estas diferencias las podemos calcular y así se obtienen los valores de la tabla adjunta. La distribución porcentual marginal de la variable dependiente, Coche, da que el 69,8% de las personas tiene coche y que un 30,2% no tiene. Como hemos dicho, la ausencia de asociación implicaría que estos porcentajes globales se reproducirían entre las categorías altas, medias y bajas, es decir, independientemente de la categoría social a la que se pertenece siempre se poseería coche en la misma proporción.

Tabla III.6.6. La posesión de coche según la clase social

C(2,3): Porcentajes por columna

Clase Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	91,0	78,7	58,0	**69,8**
No 2	9,0	21,3	42,0	**30,2**
Total	**100**	**100**	**100**	**100**

Diferencias respecto al marginal

Clase Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	+21,2	+8,9	-11,8	**0**
No 2	-21,2	-8,9	+11,8	**0**
Total	**0**	**0**	**0**	**0**

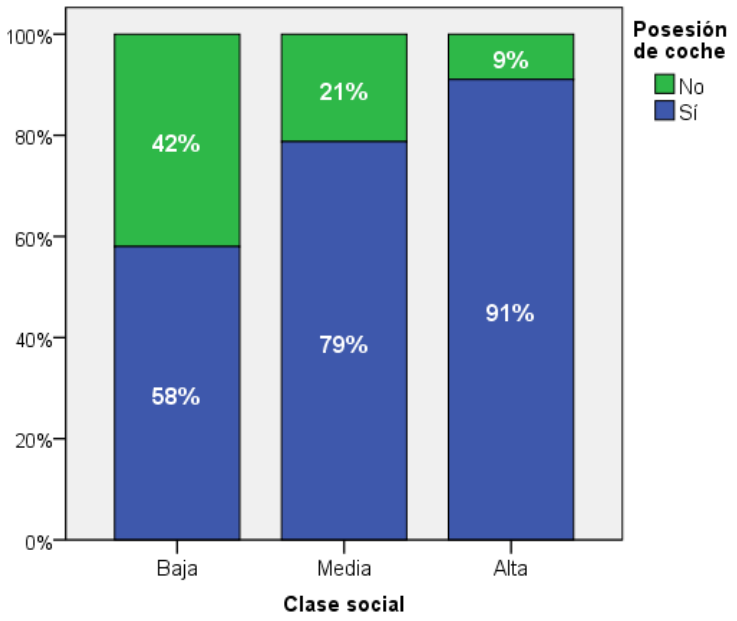
Pero como podemos ver nada más lejos de la realidad. Las distribuciones condicionales se alejan del comportamiento general: resulta que las categorías altas tienen un 21,2% más coche que el conjunto de la población, de forma similar se comportan las categorías medias, pero con una diferencia menor, un 8,9% por encima; finalmente son las categorías bajas las que poseen menos coche que el total, un 11,8% por debajo[8]. Es decir, a medida que aumenta la clase social disminuye la proporción de personas que tienen coche. Similares conclusiones se afirman considerando la no posesión de coche, de hecho es una información complementaria que arroja el mismo resultado y, por ello, redundante. Por lo tanto, hemos podido poner de manifiesto la existencia de una asociación, de diferencias entre las categorías socioeconómicas, la dirección y cuya naturaleza acabamos de describir.

La información de una tabla de contingencia se puede presentar también gráficamente mediante un gráfico de barras, en particular, mediante barras apiladas con los porcentajes que suman el 100%[9].

[8] Obsérvese que las diferencias positivas se compensan con las negativas para dar el valor 0, como sucede siempre que tenemos un cuestión de distribución de un conjunto, en este caso, el reparto de los poseedores de coches se realiza de forma desigual, las clases que más tienen lo poseen "a costa" de los que menos tienen, en un juego de suma cero.

[9] La variable independiente se coloca en el eje de categorías y la dependiente en la leyenda.

Gráfico III.6.1 Gráfico de barras apilado de Coche según Clase



Cuando las barras se igualan estamos ante una situación de independencia y si se observa una disposición escalonada como es el caso nos encontramos ante una (posible) relación de asociación entre las variables.

La caracterización de la distribución conjunta de las dos variables, de la relación entre éstas, se realiza conociendo cuatro aspectos:

1. La existencia o no de asociación. Ya conocemos cómo hacer la lectura de la información de una tabla de contingencia y cómo deducir de esta lectura la existencia de asociación. Nos queda determinar si esta deducción es, desde un punto de vista estadístico, significativa, por lo que realizaremos la denominada prueba de chi-cuadrado.
2. La fuerza o el grado de esta asociación. Una vez comprobado que es estadísticamente significativa la asociación es necesario determinar si és fuerte o débil. Conoceremos la intensidad de la relación a través de diferentes medidas del grado de asociación.
3. La dirección de la asociación. Siempre que haya variables ordinales se trata sencillamente de determinar si la asociación es positiva o negativa, si la asociación es en la misma dirección o en dirección contraria.
4. La naturaleza de la asociación. Finalmente se caracteriza la asociación detallando la forma general en que se distribuyen los datos en la tabla, la forma en que se distribuye la variable dependiente para las diferentes categorías de la variable independiente. Esta distribución puede ser irregular, de progresión uniforme, lineal, etc.

3. La prueba de chi-cuadrado (χ^2)

La asociación captada a través de la lectura de los porcentajes es una forma "intuitiva" de constatación, absolutamente necesaria y relevante, pero que requiere concretarse

(objetivarse) en medidas que nos indiquen, cuando trabajamos con muestras estadísticas, si es significativa o no desde un punto de vista estadístico. De hecho si nuestros datos correspondieran a toda la población, el error estadístico no existiría, y cualquier diferencia, por grande o pequeña que sea, será una diferencia pues la afirmamos de todos los individuos de la población. En el caso de disponer de datos de encuestas muestrales, como en nuestro caso, debemos validar nuestra hipótesis de asociación entre las variables Coche y Clase mediante un cálculo que determine si las diferencias observadas entre las clases sociales son suficientemente relevantes, o por el contrario si pueden deberse simplemente a la aleatoriedad de haber elegido nuestra muestra y no otra, y que en consecuencia las diferencias no constituyen un hecho relevante cuando queremos extrapolarlas, inferirlas, al conjunto de la población.

Para constrastar la hipótesis de asociación entre las variables realizaremos la denominada prueba de independencia de chi-cuadrado de Pearson[10]. Como en toda prueba estadística de contraste de hipótesis podemos distinguir 4 pasos en su realización:

1. Formulación de las hipótesis nula y alternativa, en este caso:
   H_0: Las variables son independientes.
   H_A: Las variables no son independientes, existe asociación.
2. Se calcula el valor del estadístico, aquí el chi-cuadrado (χ^2).
3. Se determina la probabilidad asociada al estadístico.
4. Se toma la decisión aceptando o rechazando la hipótesis nula.

Veamos con detenimiento cómo se obtiene el estadístico de chi-cuadrado y cómo se realiza a continuación el contraste. La forma de conocer la existencia o no de asociación entre dos variables consiste en contrastar las frecuencias observadas o reales con las frecuencias esperadas o teóricas suponiendo que no hubiera asociación, es decir, suponiendo que fueran independientes. Al comparar estas frecuencias, si no existen diferencias entre ellas concluiremos la ausencia de asociación o de relación de interdependencia, y diremos que las variables son independientes entre sí. Pero cuando se producen diferencias entre estos valores, por exceso o por defecto, debemos determinar si esas diferencias son suficientemente importantes, si la asociación que indican es significativa, o bien si las diferencias observadas en la muestra son atribuibles simplemente al azar. En esos términos se plantea la prueba estadística que someterá el contraste de la hipótesis de asociación.

La obtención de las frecuencias teóricas esperadas bajo la hipótesis de independencia supone encontrar una distribución, para cada categoría de la variable independiente, que no difiera de la distribución global de la variable dependiente, es decir, como hemos comentado para hacer la lectura de una tabla de contingencia, que la frecuencia relativa para cada casilla sea la misma que la del total, que se expresa en la ecuación:

$$\frac{n_{ij}^e}{n_{+j}} = \frac{n_{i+}}{n} \quad \text{para todo } i,j. \qquad \textit{Ecuación 1}$$

[10] Karl Pearson (1857–1936), quien estableció la Estadística como disciplina.

En el caso de nuestro ejemplo, cuando *i=1* y *j=1*, la frecuencia esperada es:

$$\frac{n^e_{11}}{n_{+1}} = \frac{n_{1+}}{n} \Rightarrow \frac{n^e_{11}}{714} = \frac{3314}{4747} = 69{,}8\% \Rightarrow n^e_{11} = \frac{714 \times 3314}{4747} = 498{,}5$$

La frecuencia esperada 498,5 es el número de personas de clase alta que se esperaría que hubiera si no existieran diferencias entre las clases sociales (si hubiera independencia) y todas ellas poseyeran coche en la misma proporción y en una magnitud igual al 69,8%. Dicho de otra forma, 498,5 es el 69,8% de las personas de clase alta.

Si este cálculo los realizamos en cada una de las casillas de la tabla obtenemos la distribución de frecuencias esperadas:

$$n^e_{11} = \frac{714 \times 3314}{4747} = 498{,}5 \qquad n^e_{21} = \frac{714 \times 1433}{4747} = 215{,}5 \qquad n^e_{23} = \frac{2466 \times 3314}{4747} = 1721{,}6$$

$$n^e_{12} = \frac{1567 \times 3314}{4747} = 1094{,}0 \qquad n^e_{22} = \frac{1567 \times 1433}{4747} = 473{,}0 \qquad n^e_{23} = \frac{2466 \times 1433}{4747} = 744{,}4$$

De estos cálculos se deriva una fórmula general, la frecuencia esperada es el producto de los totales absolutos marginales dividido por el total:

$$n^e_{ij} = \frac{n_{i+} \cdot n_{+j}}{n} \qquad \textit{Ecuación 2}$$

es decir: $$n^e_{ij} = \frac{total\ \ fila\ i \times total\ columna\ \ j}{total\ de\ casos}$$

Bajo la hipótesis de independencia, la *Ecuación 2* es la llamada condición de independencia que se deriva de la propiedad de la homogeneidad de proporciones (al igualar el porcentaje de una casilla con el marginal) de la *Ecuación 1*.

Por tanto, hemos observado que 650 personas de clase alta tienen coche y que "deberían" ser 498,5 para que hubiera igualdad social en la posesión de coche. La diferencia entre estos valores, entre la frecuencia observada y la frecuencia esperada, 151,5 en este caso, se denomina residuo. Más adelante profundizaremos sobre este concepto, fundamental en el razonamiento estadístico. En la Tabla III.6.7 se recogen las tablas con los tres tipos de datos que acabamos de comentar.

Tabla III.6.7. La posesión de coche según la clase social

N(I,J): Frecuencias observadas

Clase Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	650	1234	1430	3314
No 2	64	333	1036	1433
Total	714	1567	2466	4747

E(I,J): Frecuencias esperadas

Clase Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	498,5	1094,0	1721,6	3314
No 2	215,5	473,0	744,4	1433
Total	714	1567	2466	4747

R(I,J): Residuos

Clase Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	151,5	140,0	-291,6	0
No 2	-151,5	-140,0	291,6	0
Total	0	0	0	0

Observada – Esperada = Residuo

La evaluación estadística de las diferencias se realiza por el criterio de distancias, donde se define un índice como la distancia media cuadrática o de chi-cuadrado (χ^2)[11]:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij} - n_{ij}^e)^2}{n_{ij}^e} \qquad \textit{Ecuación 3}$$

para todo $i=1...I$, $j=1...J$, con $I, J \geq 2$ y con $\nu=(I-1)\times(J-1)$ grados de libertad.

En palabras, la *Ecuación 3* es la suma de todas las casillas donde se calcula en cada una de ellas la diferencia entre la frecuencia observada y esperada (el residuo) que se eleva al cuadrado y se divide por el número de casos esperados en cada casilla. Al elevar al cuadrado los residuos se consigue que no se compensen los positivos con los negativos y sumen cero, y al dividir por las frecuencias esperadas se consigue relativizar las las casillas con mayores y menores contribuciones. El resultado es un valor numérico que llamaremos chi-cuadrado observado (χ_o^2) que expresa la distancia media entre los valores y que será objeto de valoración como veremos seguidamente.

En el ejemplo y utilizando la primera fórmula del χ^2:

$$\chi_o^2 = \frac{(650\text{-}498{,}5)^2}{498{,}5} + \frac{(1234\text{-}1094)^2}{1094} + \frac{(1430\text{-}1721{,}6)^2}{1721{,}6} + \frac{(64\text{-}215{,}5)^2}{215{,}5} + \frac{(333\text{-}473)^2}{473} + \frac{(1036\text{-}744{,}4)^2}{744{,}4} =$$

$$= 46{,}07 + 17{,}92 + 49{,}38 + 106{,}54 + 41{,}46 + 114{,}21 = 375{,}58$$

Calculado el estadístico del chi-cuadrado es necesario valorar si expresa diferencias importantes entre lo observado y lo esperado o bien si ese valor de diferencias no es importante y cabe concluir la independencia entre las variables. El valor obtenido para el estadístico se contrasta entonces con el valor crítico de la distribución teórica de χ^2, el cual se fija según el nivel de significación α (probabilidad a partir de la cual aceptaremos la hipótesis nula). Habitualmente se considera un nivel de significación del 0,05 (el 5% de probabilidades, es decir, un nivel de confianza del 95%) que se corresponde con un valor concreto de la distribución teórica de chi-cuadrado.

Este valor se puede obtener consultando una tabla teórica de los valores de chi-cuadrado como la que se adjunta en el anexo del capítulo. Para localizar nuestro valor crítico además de fijar una significación se deben considerar los grados de libertad, que es un indicador de la dimensión de la tabla y de las sumas que por tanto se realizan al calcular el χ^2, circunstancia que afecta a la distribución teórica.

La noción de grados de libertad es un concepto técnico estadístico, da cuenta de la dimensión de la tabla y es igual al producto $\nu=(I-1)\times(J-1)$: es el resultado de tener en

[11] A efectos de cálculo se pueden utilizar las siguientes expresiones equivalentes:

$$\chi^2 = \left(\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{ij}^e}\right) - n \qquad \chi^2 = n\cdot\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{n_{ij}^2}{n_{i+}\cdot n_{+j}} - 1$$

cuenta el número de casillas de la tabla ($I \times J$) y el número de restricciones. De forma intuitiva la idea es la siguiente. Fijado un tamaño de muestra y las distribuciones marginales de la tabla la cuestión es: "¿qué libertad tengo para distribuir los casos en la tabla?":

Clase / Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	✓	✓	×	3314
No 2	×	×	×	1433
Total	714	1567	2466	4747

Dadas 714 personas de clase alta "puedo decidir" que, por ejemplo, 650 vayan a la primera casilla: primer grado de libertad. Entonces en la casilla (2,1) no puede haber más que 64 personas pues el total son 714. Lo mismo puedo hacer con las clases medias, decido poner 1234 en la casilla (1,2): segundo grado de libertad, pero con ello determino inmediatamente los efectivos de la casilla (2,2), los 333 que completan el total de 1567. Finalmente las casillas de la clase baja están determinadas por las decisiones anteriores, no tengo más libertad de decisión y serán respectivamente 1430 y 1036. En total pues 2 grados de libertad.

La tabla de distribución teórica de χ^2 o "distribución muestral teórica" es una distribución de probabilidades que indica la probabilidad de obtener un valor del estadístico en el supuesto de que la hipótesis nula fuera cierta con $\nu=(I-1)\times(J-1)$ grados de libertad, es decir, los grados de libertad (ν) se obtiene del producto $(I-1)\times(J-1)$, el número de filas menos 1 por el número de columnas menos 1.

Ya estamos pues es disposición de localizar este valor teórico crítico de chi-cuadrado de nuestro ejemplo que depende de la significación que fijamos en $\alpha=0{,}05$ y de los grados de libertad $\nu=(I-1)\times(J-1)=(2-1)\times(3-1)=2$. Es decir:

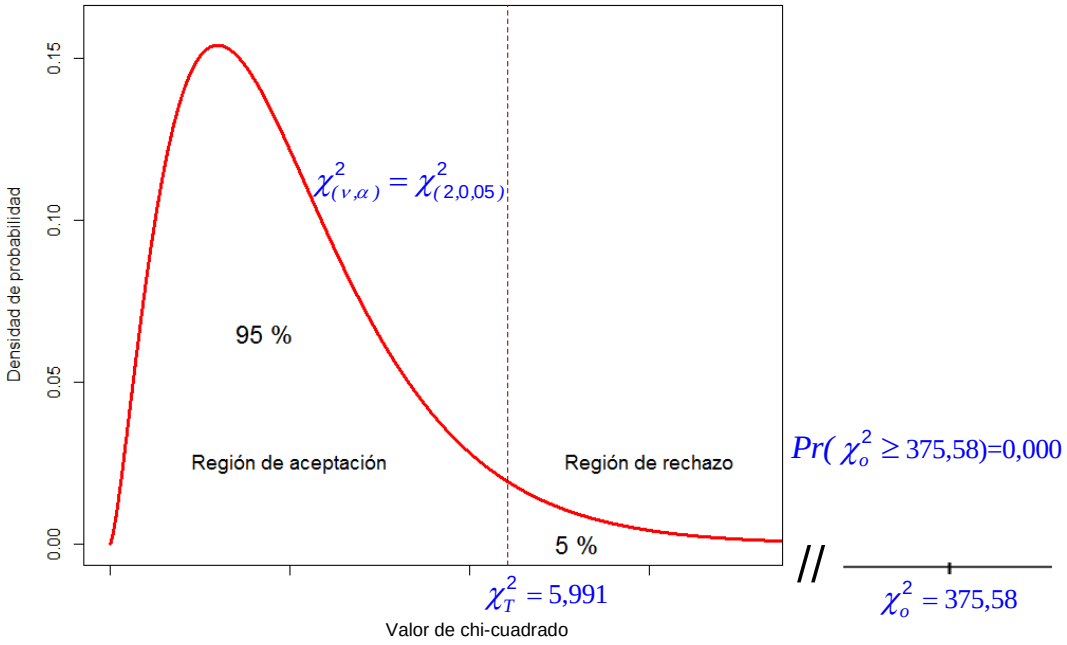
$$\chi^2_T = \chi^2_{(\nu,\alpha)} = \chi^2_{(2,0{,}05)} = 5{,}991$$

El valor 5,991 lo encontraremos en la tabla del anexo en el cruce de la fila de 2 grados de libertad y la columna de probabilidad de 0,05. Este valor se interpreta como el valor máximo para aceptar la hipótesis nula[12]. Como nuestro valor observado es de $\chi^2_o = 375{,}58$, un valor muy por encima, concluimos que rechazamos la hipótesis nula de independencia y aceptamos la alternativa de asociación.

El valor crítico separa las zonas de aceptación y rechazo, la distribución de chi-cuadrado se puede representar gráficamente de la forma siguiente:

[12] La tabla de distribución teórica de χ^2 o "distribución muestral teórica" es una distribución de probabilidades que indica la probabilidad de obtener un valor del estadístico en el supuesto de que la hipótesis nula fuera cierta con $\nu=(I-1)\times(J-1)$ grados de libertad.

Gráfico III.6.2 Prueba de independencia de chi-cuadrado



Cuando realizamos la prueba estadística a través del ordenador la respuesta del software estadístico se expresa en términos de probabilidad, nos indica cuál es la probabilidad de obtener un valor mayor o igual que el chi-cuadrado observado: $Pr(\chi^2_o \geq 375{,}58)=0{,}000$, es decir, cuál es la probabilidad de que se dé la hipótesis nula. Como esta probabilidad es muy baja (de hecho 0 con tres decimales) e inferior a 0,05, que es el límite para aceptar la hipótesis nula, concluimos como antes que no existe independencia entre las variables.

Formalmente las dos maneras de realizar el contraste, la manual mediante la tabla teórica o la automática a través del software estadístico, se expresan de la forma siguiente en la toma de la decisión.

a) Constraste con los valores de la tabla:

Si $\chi^2_o \leq \chi^2_T$ aceptamos la hipótesis nula, las variables son independientes.

Si $\chi^2_o > \chi^2_T$ rechazamos la hipótesis nula, las variables no son independientes (se da una relación de asociación).

Los valores teóricos se presentan en la tabla de distribución teórica del χ^2 indicando la proporción o la probabilidad α de obtener valores superiores o iguales a un valor crítico $\chi^2_T = \chi^2_{(\nu,\alpha)}$, es decir, $Pr(\chi^2 > \chi^2_T) = \alpha$.

En nuestro caso, el valor crítico lo fijamos para $\alpha=0{,}05$:

$$Pr(\chi^2 > \chi^2_{(\nu,\alpha)}) = Pr(\chi^2 > \chi^2_{2,0,05}) = Pr(\chi^2 > 5{,}991) = 0{,}05$$

Mientras que el observado tiene una probabilidad asociada de:

$$Pr(\chi^2 > \chi^2_o) = Pr(\chi^2 > 375{,}58) = 0{,}000$$

b) Constraste en términos de probabilidad:

Si $Pr(\chi_o^2) \geq 0{,}05$ aceptamos la hipótesis nula, las variables son independientes.

Si $Pr(\chi_o^2) < 0{,}05$ rechazamos la hipótesis nula, las variables no son independientes (se da una relación de asociación).

La prueba estadística de χ^2 se aplica al contraste de hipótesis en diferentes situaciones. Aquí nos interesa el caso de la prueba de decisión estadística donde se establece la independencia entre dos variables cualitativas de una tabla de contingencia, pero en general se puede utilizar en situaciones donde se trata de comparar dos distribuciones entre sí. Como en toda prueba estadística es preciso analizar y verificar que se dan las condiciones de aplicación. En el caso del test de χ^2 se suponen seis condiciones:

1. Las observaciones deben ser independientes entre sí.
2. Los sucesos sean mutuamente excluyentes.
3. La distribución de chi-cuadrado observada con variables no continuas se supone que pueden aproximarse a la distribución teórica continua de chi-cuadrado.
4. El nivel de medida mínimo es nominal.
5. El tamaño de la muestra n, el número total de efectivos, debe ser relativamente grande. El criterio que se utiliza habitualmente es que la frecuencia esperada mínima para casilla debe ser de 5 en el 80% de las casillas, considerando además que la frecuencia mínima esperada en cada casilla sea 1 (Cochran, 1952).
6. El test establece únicamente la existencia o no de independencia, nada dice de la intensidad de la relación, pues el tamaño de la muestra y el número de casillas influyen de manera determinante sobre los valores del estadístico χ^2.

Insistiremos en este último punto. Por un lado destacaremos cuáles son los valores inferiores y superiores del χ^2. El valor del estadístico es siempre mayor o igual que cero, siendo el valor cero el que indica que hay independencia perfecta entre les variables pues las frecuencias observadas y esperadas coinciden exactamente. Pero el límite superior del χ^2 varía en cada caso, pues depende del tamaño de la muestra n y del número de casillas de la tabla (de hecho, del número menor entre el número de filas y el número de columnas menos 1), y se expresa en el producto $n \cdot (k-1)$, con $k = mín\{I,J\}$.

En nuestro ejemplo estos valores son los siguientes:

Valor de k: $k = mín\{I,J\} = min\{2,3\} = 2$

Valor máximo de χ^2: $n \cdot (k-1) = 4747 \times (2-1) = 4747$

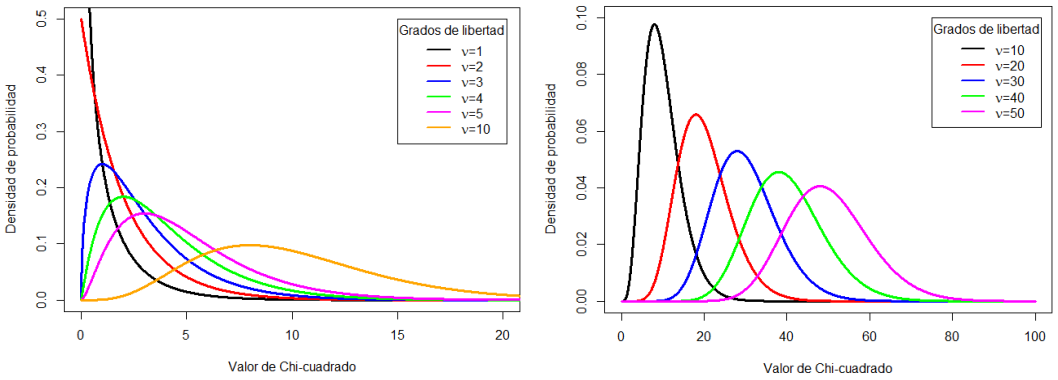
Por otro lado de esta propiedad se deriva que con muestras con un elevado número de casos es fácil establecer la significatividad de la relación entre variables, por débil que ésta sea, de ahí que es importante conocer la fuerza de la relación. Este efecto de tamaño se pone de manifiesto en el siguiente comportamiento: si el número de casillas y los porcentajes de una tabla no varían, cuando se duplica la muestra, el valor del estadístico χ^2 se duplica, o se triplica, etc., es siempre k veces, como muestra la fórmula adjunta:

$$\chi'^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(k\cdot n_{ij} - k\cdot n_{ij}^e)^2}{k\cdot n_{ij}^e} = k\cdot\chi^2 \qquad \textit{Ecuación 4}$$

En consecuencia, se deriva el corolario siguiente: con muestras pequeñas la existencia de relación entre las variables es indicativa de una relación importante; por el contrario, con muestras grandes, la no existencia de relación es indicativa de una real ausencia de relación.

Finalmente comentaremos que la distribución teórica de χ^2 de hecho es una familia de curvas diferentes según los grados de libertad, y a medida que éstos aumentan se aproxima cada vez más a la distribución normal[13]:

Gráfico III.6.3 Forma de la distribución de χ^2 según los grados de libertad



En la Tabla III.6.8 aparecen los resultados de la prueba chi-cuadrado que proporciona el software estadístico SPSS[14]. El largo recorrido que hemos deambulado a través de la prueba estadística se resume en la práctica del análisis en un ejercicio tan automático, sencillo y rápido como constatar que el valor de significación es 0,000 e inferior a 0,05 lo que nos permite afirmar que hay asociación entre las variables Coche y Clase.

[13] Técnicamente se dice que la distribución teórica de χ^2 de hecho es una familia de distribuciones para los diferentes grados de libertad con un rango de valores que varía desde 0 hasta el infinito, asintóticamente hacia a la derecha, dibujando una curva (polígono de frecuencias que se aproxima a la curva matemática) y un área que nos indica la probabilidad de que la suma de cuadrados de n puntuaciones z escogidas aleatoriamente de la distribución normal sea superior a un valor dado y unos grados de libertad dados. La distribución de χ^2 con ν grados de libertad es el resultado de la suma de ν variables normales estandarizadas:

$$\chi_\nu^2 = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$$

El valor medio de una distribución de χ^2 con ν grados de libertad es ν, y su error típico es $\sqrt{2\nu}$. Cuando ν es grande, de 30, el χ^2 se aproxima a la normal.

[14] Junto a la prueba chi-cuadrado aparecen otros estadísticos calculados que comentaremos más adelante.

Tabla III.6.8. Resultados de la prueba de chi-cuadrado con SPSS

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (2 caras)
Chi-cuadrado de Pearson	375,583[a]	2	0,000
Razón de verosimilitud	407,253	2	0,000
Asociación lineal por lineal	367,690	1	0,000
Prueba de McNemar-Bowker	.	.	.[b]
N de casos válidos	4747		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 215,54.
b. Sólo se ha calculado para una tabla PxP, donde P debe ser mayor que 1.

Cuando no se cumple alguna de las condiciones de aplicación de la prueba de chi-cuadrado que hemos visto en relación a las frecuencias esperadas mínimas (mínima de 1 y 80% de las casillas con 5 o más) es necesario agrupar categorías de las variables y reducir el número de casillas con pocos efectivos esperados, siguiendo siempre criterios de coherencia conceptual y de aplicabilidad práctica que hagan pertinentes las uniones. También se puede emplear la prueba exacta de Fisher que es un estadístico alternativo al chi-cuadrado para valorar la existencia de asociación en tablas de 2×2, y en general para tablas con menos de 50 casos.

► **Ejercicio 1.** [15]

A partir de la tabla de contingencia que relaciona la posesión de ordenador y el nivel de ingresos:

	Ingresos	
Ordenador	Altos	Bajos
Sí	**27**	**9**
No	**9**	**15**

Analizar la relación de asociación calculando los porcentajes y los residuos absolutos. Calcular del valor del estadístico chi-cuadrado ¿es significativo?

► **Ejercicio 2.**

A partir de la tabla de contingencia que relaciona el uso del transporte público para ir a trabajar y la clase social:

	Clase social		
Transporte	Alta	Media	Baja
Sí	**10**	**50**	**120**
No	**40**	**50**	**30**

Analizar la relación de asociación calculando los porcentajes y los residuos absolutos. Calcular del valor del estadístico chi-cuadrado ¿es significativo?

[15] Como se comentará más adelante todos los ejercicios propuestos se pueden reproducir con la ayuda del software estadístico.

► **Ejercicio 3.**
Construye tres tablas de contingencia en donde se relacione el abstencionismo electoral y la edad:

	Edad	
Abstención	Joven	Mayor
Sí		
No		

inventando las frecuencias de las casillas para satisfacer cada una de las condiciones siguientes: ausencia de relación, relación moderada y alta asociación. Calcula en cada caso el estadístico de chi-cuadrado.

4. Medidas del grado de asociación

Con el χ^2 hemos establecido una medida de la existencia o no de relación entre dos variables. En caso de que haya relación hay que plantearse si ésta es fuerte o débil, es decir, la intensidad de la relación, pues una asociación aunque exista, sea significativa estadísticamente, puede que reporte poca relevancia dada su baja intensidad. Si la muestra es numerosa, a partir de 2.000 casos, este es un resultado habitual, obtener relaciones significativas, es decir, diferencias entre los porcentajes de las tablas que nos permiten afirmar que son diferencias significativas pero tan pequeñas que no muestran una especial relevancia.

Como comentamos en al apartado anterior el propio estadístico χ^2 podría considerarse una medida de la intensidad de la relación pero tiene el problema de no permitir la comparación entre tablas de contingencia de diferente dimensión, tablas que tengan un número diferente de filas o columnas, y que tengan un tamaño de muestra diferente. Veremos como el estadístico V de Cramer resuelve esta limitación y cómo para medir el grado de asociación existen numerosas medidas de carácter global que resumen en un solo valor estadístico la intensidad de la relación entre las variables. También veremos medidas de carácter local que permiten analizar con mayor detalle el comportamiento de pares de combinaciones de valores de las variables.

4.1. Medidas de asociación global

Una medida de asociación global es un índice numérico que resume la fuerza o la intensidad de la relación entre dos variables en una tabla de contingencia. En función del significado de esta medida además puede mostrar la dirección de la asociación. Hay un gran número de medidas con características diferenciadas en relación a dos aspectos principales:

- La simetría. Si se considera la división entre variable dependiente/independiente entonces el índice variará en función de que una u otra de las variables consideradas sea dependiente, será una medida asimétrica o direccional. Si no se considera la división se estudia el comportamiento conjunto de ambas variables y la medida es simétrica.

- La interpretación. La información de la tabla de contingencia se resume en un solo valor numérico. La mayor parte de estas medidas se interpretan como valores comprendidos entre 0 y 1, o entre -1 y 1 si se da asimetría, según el grado de asociación: el valor 0 es indicativo de nula asociación mientras que el 1 lo es de la asociación perfecta. El significado de los valores intermedios depende de cómo se ha definido operacionalmente la medida.

En general estas medidas tienen dos propiedades que las caracterizan como medidas del grado de intensidad de la relación. Por un lado es la simplicidad de la información que proporcionan al resumir la tabla de contingencia en un solo valor numérico, lo que por un lado es útil en aras de la parsimonia, pero al mismo tiempo no permiten ver el detalle de la relación entre las categorías de las variables. Por otra parte este valor resumen debería tener un significado claro, y no siempre es así, además de que cada una lo hace "a su manera" por lo general son medidas "pesimistas" del grado de asociación ya que rara vez llegan a los valores más altos, si bien ello depende también de la misma naturaleza social de lo medido.

En la Tabla III.6.9 se recogen diversas medidas de asociación global según una clasificación basada en el nivel de medición de las variables y en la Tabla III.6.10 se presentan todas las que calcula el procedimiento de tablas de contingencia del SPSS.

Tabla III.6.9. Clasificación de las medidas de asociación global

1. Variables nominales	2. Variables ordinales
1.1 Basadas en el chi-cuadrado	– *Coeficiente Tau-a de Kendall*
– *Coeficiente V de Cramer*	– *Coeficiente Tau-b de Kendall*
– *Coeficiente de contingencia C de Pearson*	– *Coeficiente Tau-c de Kendall-Stuart*
– *Coeficiente de contingencia C de Pearson con modificación de Sakoda*	– *Coeficiente Gamma de Goodman-Kruskal*
– *Coeficiente de contingencia cuadrático medio de Pearson*	– *Coeficiente d de Sommer*
– *Coeficiente de contingencia de Tschruprow*	– *Coeficiente e de Wilson*
	– *Coeficiente Ro de Spearman*
1.2. Para tablas de 2 × 2	**3. Variable nominal y ordinal**
– *Coeficiente Q de Yule*	– *Correlación biserial-punto*
– *Coeficiente Phi*	– *Correlación biserial*
	– *Eta*
1.3. Medidas de reducción de la predicción del error	**4. Variables de intervalo**
– *Coeficiente Lambda de Goodman-Kruskal*	– *Coeficiente de correlación producto-momento de Pearson*
– *Coeficiente Tau de Goodman-Kruskal*	
– *Coeficiente de incertidumbre*	

Daremos cuenta de algunas medidas de asociación basadas en el chi-cuadrado y que intentan corregir la limitación del χ^2 acotando su valor entre 0 y 1. Para un desarrollo y discusión más precisos de las medidas de asociación para variables cualitativas se puede consultar Reynolds (1977), Hildebrand et al. (1977), Liebetrau (1983), García Ferrando (1994), Sánchez Carrión (1995) o Aguilera (2001).

4.2. Medidas de asociación global basadas en el chi-cuadrado

Las medidas basadas en el chi-cuadrado buscan superar los límites del estadístico como medida del grado de asociación y tienen la ventaja de poder aplicarse a cualquier tabla pues el nivel de medición exigido es el nominal. Todas estas medidas de asociación simétrica dan lugar a un valor cuya significatividad se determina mediante una prueba de significación de χ^2 con $\nu=(I-1)(J-1)$ grados de libertad.

4.2.1. *Coeficiente V de Cramer*

Trabajaremos habitualmente con esta medida, sencilla de interpretar y válida para cualquier tabla de contingencia. La V de Cramer es una medida simétrica que se construye relacionando el valor del estadístico chi-cuadrado con respecto al máximo que éste alcanza $n\cdot(k-1)$[16]. Su fórmula es la siguiente:

$$V=\sqrt{\frac{\chi^2}{n\cdot(k-1)}} \qquad \textit{Ecuación 5}$$

Donde $k=mín\{I,J\}$, es decir, el valor más pequeño entre el número de filas y el número de columnas. La V de Cramer alcanza un valor máximo de 1 en caso de máxima asociación o asociación perfecta y un valor mínimo de 0 en una situación de independencia perfecta. La experiencia muestra que con la V es poco frecuente encontrar valores de alta intensidad próximos a 1, de hecho pocas veces se alcanza un valor de 0,6. En términos empíricos por tanto y tomando el valor de V en sí mismo podemos considerar al 0,6 prácticamente como un valor máximo habitual, por lo que un valor de 0,3, antes que considerarlo como bajo por su proximidad a 0 conviene interpretarlo más bien como un valor empírico intermedio.

Y no puede ser de otra forma. Analicemos brevemente lo que supondría socialmente una medida de asociación V con valor máximo de 1. En ese caso los porcentajes de la tabla serían los que darían las máximas diferencias entre las categorías de la variable independiente con en el ejemplo ficticio siguiente, donde consideramos la posesión de coche (0 coches, 1 coche o 2 y más coches) y la clase social (baja, media y alta) con una distribución extrema como esta:

Clase / Coche	*Alta 1*	*Media 2*	*Baja 3*
2+	**100%**		
1		**100%**	
0			**100%**

una situación donde las clases altas tendrían todas 2 o más coches, las clases medias 1 y las clases bajas no tendrían. Es decir, un determinismo absoluto de la variable dependiente por la independiente, por lo que la posesión de coche sería un clasificador social total: si no tengo coche seguro que seré de clase baja, si tengo uno de clase media y más de un coche significa, "tautológicamente", ser de clase alta. Una realidad poco

[16] Coincide con la T de Tschuprow cuando $I=J$.

real en nuestra sociedad, pero es la que arrojaría un valor máximo de la V de Cramer igual a 1.

En el caso real de la sociedad metropolitana de Barcelona con los datos de la Tabla III.6.1 el valor de la V de Cramer que se obtiene es:

$$V = \sqrt{\frac{\chi^2}{n \cdot (k-1)}} = \sqrt{\frac{375{,}53}{4747 \times (2\text{-}1)}} = 0{,}281$$

Un valor que podemos valorar como intermedio en términos de lo habitual empíricamente. Pero en sí misma una medida absoluta no es lo más interesante. La lógica del análisis social y del análisis estadístico es la comparativa, por lo que lo más interesante es preguntarse, por ejemplo, si en otras sociedades metropolitanas esa medida, que puede considerarse como un indicador de "desigualdad social", es mayor o menor, si ese resultado ha cambiado en el tiempo, si ese valor es diferente entre los distintos municipios, o entre diferentes grupos sociales. Cuando introduzcamos el análisis de tablas de contingencia multidimensionales, la consideración de una tercera variable generará la multiplicación de tablas de contingencia que deberemos comparar mirando cómo varía una medida de asociación entre las diferentes tablas, lo que permite constatar, por ejemplo, si la desigualdad aumenta o disminuye de una situación a otra.

Para determinar la significación estadística, si V_0 es el valor observado de V:

$$Pr(V \geq V_0) = Pr(V^2 \geq V_0^2) = Pr\left(\frac{\chi^2}{n(k-1)} \geq V_0^2\right) = Pr\left(\chi^2 \geq V_0^2 \cdot n(k-1)\right)$$

con $\nu = (I-1)(J-1)$ grados de libertad.

4.2.2. *Coeficiente Phi de Pearson*

El coeficiente Phi (φ) de Pearson es una medida de asociación que se aplica a tablas de dimensión de 2×2, siendo el valor 0 la independencia perfecta y ±1 la asociación perfecta. Se expresa como:

$$\varphi = \sqrt{\frac{\chi^2}{n}} \qquad \textit{Ecuación 6}$$

Coincide con la V de Cramer cuando ésta se calcula en tablas de 2×2. Es un estadístico que equivale también al coeficiente de correlación de Pearson, que veremos en el tema del análisis de regresión, cuando las variables son dicotómicas y se codifican con 0 y 1. Se puede aplicar a tablas con mayor número de filas o columnas pero en ese caso no tiene un máximo.

El cálculo del estadístico en una tabla de 2×2 es equivalente a la fórmula:

$$\varphi = \sqrt{\frac{\chi^2}{n}} = \frac{(n_{11}n_{22} - n_{21}n_{12})^2}{n_{1+}n_{+1} - n_{2+}n_{+2}}$$

Si se eleva al cuadrado se denomina Coeficiente de contingencia cuadrático medio de Pearson (φ^2).

Tabla III.6.10. Medidas de asociación global de Coche por Clase

Medidas direccionales

			Valor	Error estándar asintótico[a]	Aprox. S[d]	Aprox. Sig.
Nominal por Nominal	Lambda	Simétrico	,000	0,000	.[b]	.[b]
		Coche dependiente	0,000	0,000	.[b]	.[b]
		Clase dependiente	0,000	0,000	.[b]	.[b]
	Tau Goodman y Kruskal	Coche dependiente	0,079	0,007		0,000[c]
		Clase dependiente	0,045	0,004		0,000[c]
	Coeficiente de incertidumbre	Simétrico	0,054	0,005	10,841	0,00[e]
		Coche dependiente	0,70	0,06	10,841	0,00[e]
		Clase dependiente	0,43	0,04	10,841	0,00[e]
Ordinal por ordinal	d de Somers	Simétrico	0,265	0,12	21,534	0,00
		Coche dependiente	0,226	0,10	21,534	0,00
		Clase dependiente	0,320	0,14	21,534	0,00
Nominal por intervalo	Eta	Coche dependiente	0,281			
		Clase dependiente	0,278			

a. No se supone la hipótesis nula.
b. No se puede calcular porque el error estándar asintótico es igual a cero.
c. Se basa en la aproximación de chi-cuadrado
d. Utilización del error estándar asintótico que asume la hipótesis nula.
e. Probabilidad de chi-cuadrado de razón de verosimilitud.

Medidas simétricas

		Valor	Error estándar asintótico[a]	Aprox. S[b]	Aprox. Sig.
Nominal por Nominal	Phi	0,281			0,00
	V de Cramer	0,281			0,00
	Coeficiente de contingencia	0,271			0,00
Ordinal por ordinal	Tau-b de Kendall	0,269	0,12	21,534	0,00
	Tau-c de Kendall	0,270	0,13	21,534	0,00
	Gamma	0,541	0,23	21,534	0,00
	Correlación de Spearman	0,281	0,13	20,190	0,00[c]
Intervalo por intervalo	R de persona	0,278	0,12	19,962	0,00[c]
MEdida de acuerdo	Kappa	0,03	0,05	0,543	0,587
N de casos válidos		4747			

a. No se supone la hipótesis nula.
b. Utilización del error estándar asintótico que asume la hipótesis nula.
c. Se basa en aproximación normal.

4.2.3. *Coeficiente de contingencia C de Pearson*

El coeficiente de contingencia C de Pearson se deriva del coeficiente Phi y se puede aplicar a tablas con filas o columnas mayores que 2. Su fórmula es:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \qquad \textit{Ecuación 7}$$

Los valores del estadístico se mueven entre el mínimo 0 y un valor máximo variable que aumenta cuando lo hace el número de casillas de la tabla, sin alcanzar el valor 1. Para tablas cuadradas con I=J, el valor máximo es $\sqrt{(I-1)/I}$ (para: 2×2 es 0,707, par 3×3 es 0,816, para 4×4 es 0,870, etc.).

Para determinar la significación estadística, si C_0 es el valor observado de C:

$$Pr(C \geq C_0) = Pr(C^2 \geq C_0^2) = Pr\left(\frac{\chi^2}{\chi^2 + n} \geq C_0^2\right) = Pr\left(\chi^2 \geq C_0^2 \cdot (n+1) - 1\right)$$

con $\nu = (I-1)(J-1)$ grados de libertad.

Existe una variación denominada Coeficiente de contingencia C de Pearson con modificación de Sakoda, que permite alcanzar el valor 1 máximo:

$$C^S = \sqrt{\frac{k \cdot \chi^2}{(\chi^2 + n) \cdot (k-1)}} \qquad \textit{Ecuación 8}$$

► Ejercicio 4.
Calcula la V de Cramer de las tablas de contingencia de los ejercicios 1, 2 y 3.
¿Cómo se determina la significatividad del estadístico?
¿Qué sucede con el χ^2 y la V de Cramer si los casos de cada casilla se multiplican por 10? Comprobarlo en una de las tablas de contingencia.

4.3. Medidas de asociación local

Tras analizar en una tabla de contingencia la existencia de asociación y la intensidad de esta asociación, resumida en un solo valor estadístico, el interés se centra en analizar con más detalle la naturaleza de la relación, cómo se relacionan los valores de las variables, entre qué categorías se establecen relaciones específicas y si se deduce un modelo o formas distintas de relación con respecto a una o varias variables asociadas.

La observación de las distribuciones porcentuales ofrece una primera y básica información de la naturaleza de la asociación. Otras vías son posibles para analizar la relación desde un punto de vista local: el análisis de diferencias de proporciones, el análisis de residuos y el análisis de las razones. En todos estos casos se obtendrán medidas que nos informarán de la existencia, del grado y de la dirección de la asociación entre las variables a partir de un análisis localizado en cada casilla o entre pares de casillas.

4.3.1. *El análisis de residuos*

El análisis de los residuos permite ver la relación entre pares de categorías y se parte de la misma idea que el contraste de chi-cuadrado pero localizado en una casilla. Los residuos son la diferencia entre las frecuencias observadas y las frecuencias esperadas bajo la hipótesis de independencia:

$$R_{ij} = n_{ij} - n_{ij}^{e} \qquad \textit{Ecuación 9}$$

Cuando mayor sea este valor más se distancia de la situación de independencia y muestra, en consecuencia, la existencia de asociación entre los pares de categorías. El paso siguiente consiste en ver si estas diferencias son significativas estadísticamente y la intensidad, no para el conjunto de la tabla, sino para cada casilla. Para realizar la prueba de significatividad se estandarizan los residuos dividiendo por la desviación típica. Existen dos formas posibles de operar esta transformación:

1. Residuos estandarizados o tipificados. Permiten eliminar el efecto que tienen sobre el valor del residuo las distribuciones marginales de ambas variables, del número de efectivos de las categorías de las variables. Es decir, estandarizar las mayores frecuencias conjuntas, pues cuando mayores son los efectivos mayor probabilidad de que los residuos sean grandes:

$$RT_{ij} = \frac{n_{ij} - n_{ij}^{e}}{\sqrt{n_{ij}^{e}}} \qquad \textit{Ecuación 10}$$

   En el ejemplo, para la primera casilla, $RT_{11} = \frac{650 - 498,5}{\sqrt{498,5}} = 6,8$

2. Residuos corregidos o ajustados (de Haberman). Normalización para compararlos de manera más adecuada con una distribución de probabilidades, mediante la expresión:

$$RA_{ij} = \frac{RE_{ij}}{V_{ij}} = \frac{RE_{ij}}{\sqrt{(1-p_{i+})(1-p_{+j})}} = \frac{n_{ij} - n_{ij}^{e}}{\sqrt{n_{ij}^{e}}\sqrt{(1-p_{i+})(1-p_{+j})}} \qquad \textit{Ecuación 11}$$

   donde V_{ij} es la varianza estimada. En el ejemplo, para la primera casilla,

$$RA_{11} = \frac{650 - 498,5}{\sqrt{498,5}\cdot\sqrt{(1-0,698)\cdot(1-0,15)}} = 13,4$$

Los valores obtenidos con estas dos transformaciones (véase la Tabla III.6.11) siguen una distribución aproximadamente normal, por tanto, para un nivel de significación del *0,05*, todos aquellos valores superiores a *±1,96* son estadísticamente significativos, constituye un test donde se contrasta si el valor obtenido difiere significativamente de *0* con un nivel de confianza del *95%*.

Tabla III.6.11. La posesión de coche según la clase social
R(I,J): **Residuos absolutos, estandarizados y ajustados**

Clase / Coche	*Alta* *1*	*Media* *2*	*Baja* *3*	Total
Sí *1*	151,5 6,8 13,4	140,0 4,2 9,4	-291,6 -7,0 -18,5	0
No *2*	-151,5 -10,3 -13,4	-140,0 -6,4 -9,4	291,6 10,7 18,5	0
Total	0	0	0	0

De esta forma se puede observar:
a) Qué categorías se relacionan significativamente, cuando la diferencia existe y es significativa estadísticamente.
b) La dirección de la asociación, positiva o negativa según los signos.
c) La naturaleza de la relación, deducir perfiles de categorías de ambas variables y una forma o patrón general de la relación entre éstas a partir del conjunto de sus categorías.

En el ejemplo, los valores obtenidos son todos significativos, se muestra la relevancia de todas las casillas y se plasma la naturaleza de la asociación. En particular, por ejemplo, el número de personas de categorías bajas que poseen coche es menor que el que cabría esperar si poseer coche y ser de categoría baja fueran independientes, la relación es negativa y estadísticamente significativa. En general las categorías altas están relacionadas positivamente con tener coche (al igual que el perfil de las medias) mientras que las categorías bajas se relacionan con no tenerlos.

Cuando veamos los modelos log-lineales en el próximo capítulo el concepto de residuo será un concepto central en los análisis para determinar el ajuste de los modelos.

► **Ejercicio 5.**
Calcular los residuos tipificados y ajustados de las tablas de contingencia de los ejercicios 1, 2 y 3 ¿Cómo se relacionan los residuos tipificados en el estadístico de χ^2? ¿Qué residuos son estadísticamente significativos?

4.3.2. *El análisis de diferencias de proporciones*

La diferencia entre dos proporciones o porcentajes (d) se puede considerar una medida de asociación de carácter local y asimétrica que evalúa la asociación entre pares de categorías de la variable independiente para un valor dado de la variable dependiente a

partir del cálculo de la diferencia de los porcentajes de ambas categorías. La diferencia de proporciones puede considerarse como una medida de la existencia de asociación, de su intensidad y de la dirección de la relación entre dos variables a partir de un análisis local que compara sucesivamente parejas de proporciones[17].

En el ejemplo que seguimos, la diferencia en la posesión de coche entre la clase alta y la clase baja es de: $d = 91\% - 58\% = 33\%$, es decir, hay un 33% más de personas que tienen coche entre la clase alta. Pero esta es una de las diferencias de las múltiples existentes en una tabla. Si consideramos las casillas de la primera fila de la variable dependiente (sí posee coche) las diferencias son:

Clase alta – Clase media: 91,0 – 78,7 = 12,3
Clase media – Clase baja: 78,7 – 58,0 = 20,7
Clase alta – Clase baja: 91,0 – 58,0 = 33,0

Si realizamos los mismos cálculos de diferencias entre los que no tienen coche obtendríamos el mismo resultado absoluto pero con signo negativo.

La diferencia de proporciones oscila entre 0 y 1 (o entre 0 y 100 si se emplean porcentajes). La existencia de asociación implica la obtención de diferencias distintas de cero, positivas o negativas, mientras que el cero indica la ausencia de esta asociación. La máxima asociación se produce cuando la diferencia es máxima, es decir, del 100%, positiva o negativa. El valor absoluto de las diferencias es siempre el mismo, tan sólo varía el signo. Es una medida asimétrica y por tanto el sentido de los cálculos de las diferencias da resultados distintos.

Una vez obtenidas las diferencias se trata de ver si éstas son significativas. La prueba de significación supone calcular un intervalo de variación donde se encuentre la diferencia calculada entre dos proporciones P_1 y P_2. Si este intervalo incluye el valor cero se concluye que la diferencia entre dos proporciones no es significativamente distinta de cero. La prueba supone contrastar las hipótesis nula y alternativa siguientes:

H_0: $d = P_1 - P_2 = 0$
H_A: $d = P_1 - P_2 \neq 0$

En una tabla de contingencia bidimensional este contraste implica fijar una fila i de la variable dependiente y comparar dos columnas j y j' donde hemos obtenido dos porcentajes por columna de la variable independiente[18]. Para contrastar si ambas proporciones difieren se trata de buscar el intervalo de confianza donde se encontrará el valor poblacional ($P_{ij}^C - P_{ij'}^C$) a partir de los datos muestrales estimados, dados un nivel de significación y un error muestral de la diferencia:

$$\hat{p}_{ij}^C - \hat{p}_{ij'}^C \pm z \cdot s_d \qquad \textit{Ecuación 12}$$

[17] Sánchez Carrión (1984b: 295, 1989) basándose en el trabajo de J. A. Davis, y anteriormente de Lazarsfeld y Rosenberg, desarrolla el denominado Sistema de Diferencias de Proporciones donde la base son estos cálculos que comentamos y en donde se aplican ecuaciones lineales y la teoría de grafos para modelizar relaciones múltiples de dependencia entre variables cualitativas en tablas de contingencia. Véase también Latiesa (1991a).

[18] Es una generalización a cualquier par de categorías de una tabla de contingencia del contraste de homogeneidad de dos probabilidades binomiales independientes basados en la aproximación Normal a la distribución binomial (Alvarez, 2001: 62).

Es un contraste que supone la existencia de dos muestras aleatorias e independientes de tamaño n_1 y n_2 de poblaciones normales (o con n suficientemente grande): $N\left(P_{ij}^C,\sigma_{p_j}^2\right)$ y $N\left(P_{ij'}^C,\sigma_{p_{j'}}^2\right)$. La distribución muestral de la diferencia entre ambas proporciones es también normal del tipo $N\left(P_{ij}^C-P_{ij'}^C,\sigma_{p_j}^2+\sigma_{p_{j'}}^2\right)$. Las proporciones y varianzas poblacionales son estimadas a través de $\hat{p}_{ij}^C$, $\hat{p}_{ij'}^C$, $s_{p_j}^2$ y $s_{p_{j'}}^2$ siendo el error típico de la diferencia:

$$s_d=\sqrt{s_{p_j}^2+s_{p_{j'}}^2}=\sqrt{\frac{p_{ij}^C\cdot(1-p_{ij}^C)}{n_{+j}}+\frac{p_{ij'}^C\cdot(1-p_{ij'}^C)}{n_{+j'}}} \qquad \textit{Ecuación 13}$$

Veámoslo con un ejemplo contrastando la diferencia en la posesión de coche (fila i=1) entre la clase alta (columna j=1) y baja (columna j'=3). Las proporciones por columna estimados en la muestra son: $\hat{p}_{11}^C=0{,}910$ y $\hat{p}_{13}^C=0{,}580$, y la diferencia $d=\hat{p}_{11}^C-\hat{p}_{13}^C=0{,}91-0{,}53=0{,}33$. Su error típico da:

$$s_d=\sqrt{\frac{p_{11}^C\cdot(1-p_{11}^C)}{n_{+1}}+\frac{p_{13}^C\cdot(1-p_{13}^C)}{n_{+3}}}=\sqrt{\frac{0{,}91\times(1\text{-}0{,}91)}{714}+\frac{0{,}53\times(1\text{-}0{,}53)}{2466}}=0{,}0146$$

Por tanto, considerando un nivel de confianza del 95% (z=1,96), la diferencia poblacional se situará en un intervalo de valores:

$$0{,}33\pm1{,}96\times0{,}0146=0{,}33\pm0{,}0286$$

es decir, entre los valores (0,3014 , 0,3586), o en porcentaje (30,14% , 35,86%), intervalo que no incluye el valor cero y, por tanto, la diferencia del 33% es significativamente distinta de cero.

El software SPSS proporciona una opción, que comentaremos también después, destinada a las comparaciones de proporciones por columna en una tabla de contingencia facilitándonos la interpretación de este tipo de análisis sin necesidad de reproducir los cálculos anteriores. De hecho los cálculos no se visualizan sino que se presentan los resultados de significación (al nivel 0,05) de todas las comparaciones de proporciones a través del denominado sistema de notación de estilo APA utilizando subíndices de letras. En el análisis de la relación entre Coche y Clase los datos se presentan como en la **Error! Reference source not found.**.

Siguiendo la interpretación que se recoge en la nota a pie de la tabla, se constata que existen diferencias significativas entre todos los porcentajes pues aparecen los subíndices a, b y c en cada columna indicando que existen tres grupos de porcentajes, tantos como los iniciales, que son diferentes entre sí. Si dos de esos porcentajes no hubieran sido diferentes hubieran aparecido con la misma letra de subíndice.

Tabla III.6.12. La posesión de coche según la clase social. Pruebas de significación de comparación de proporciones

			Clase Clase social			
			1 Alta	2 Media	3 Baja	Total
Coche Posesión de coche	1 Sí	Recuento	650[a]	1234[b]	1430[c]	3314
		% por columna	91,0%	78,7%	58,0%	69,8%
	2 No	Recuento	64[a]	333[b]	1036[c]	1433
		% por columna	9,0%	21,3%	42,0%	30,2%
Total		Recuento	714	1567	2466	4747
		% por columna	100,0%	100,0%	100,0%	100,0%

Cada letra del subíndice denota un subconjunto de categorías de **Clase** (Clase social) cuyas proporciones de columna no difieren de forma significativa entre sí en el nivel 0,05.

Las comparaciones se realizan con estos estadísticos:

$$p^{C}_{i(jj')} = \frac{n_{+j} \cdot p^{C}_{ij} + n_{+j'} \cdot p^{C}_{ij'}}{n_{+j} + n_{+j'}} \qquad \textit{Ecuación 14}$$

$$z = \frac{p^{C}_{ij} - p^{C}_{ij'}}{\sqrt{\left(\frac{p^{C}_{i(jj')} \cdot \left(1 - p^{C}_{i(jj')}\right)}{n_{+j}} + \frac{p^{C}_{i(jj')} \cdot \left(1 - p^{C}_{i(jj')}\right)}{n_{+j'}} \right)}} \qquad \textit{Ecuación 15}$$

Además existe la opción de ajustar el nivel de significación por la transformación de Bonferroni cuando se realizan múltiples comparaciones, cuyo efecto es establecer un nivel de significación más exigente para no cometer errores de aceptación de la hipótesis alternativa (diferencias significativas) cuando es falsa:

$$sig_B = mín\left(\frac{sig \times J \times (J-1)}{2}, 1 \right) \qquad \textit{Ecuación 16}$$

► **Ejercicio 6.**
Realiza un análisis de diferencia de proporciones con los datos de las tablas de contingencia de los ejercicios 1 y 2.

4.3.3. *El análisis de razones*

El cálculo de las razones (ratios) corresponde a una finalidad de análisis local e implica la comparación de las categorías entre sí tomadas dos a dos, no en relación al marginal o en la relación a las frecuencias esperadas, sino a partir del cálculo de sus cocientes o ratios (*odds* en inglés)[19]. Se pueden calcular tres tipos de razones:

[19] En los acontecimientos deportivos, sobre todo en el ámbito anglosajón, son habituales las apuestas sobre los resultados de éstos y se expresan con afirmaciones como: *Manchester United vs Manchester City Odds: Man United to win has been priced at 6/5, the draw is valued at 12/5 while Man City to get the win stands at 11/4*.

1. Razones marginales: a partir de la distribución marginal de cada variable. Por ejemplo la razón de tener coche a no tener es: 3314/1433=2,3, es decir, por cada uno que no tiene coche hay 2,3 que tienen. O en sentido inverso, la razón de no tener coche a tener es: 1433/3314=0,4.

2. Razones condicionales: a partir de la distribución condicional, el cociente, para el valor de una variable, entre dos valores de la otra variable:
   - Entre las categorías altas, la razón de tener coche a no tener es: 650/64=10,2.
   - Entre las categorías medias de 1234/333=3,7.
   - Entre las categorías bajas de 1430/1036=1,4.

   La presencia de razones condicionales diferentes muestra la existencia de asociación entre las categorías consideradas, y por tanto entre las variables. Resultados de razones iguales serían una muestra de la independencia. Tan sólo la existencia y por dos categorías.

Tabla III.6.13. La posesión de coche según la clase social
Cálculo de razones

Categoría de referencia *baseline*

Clase / Coche	*Alta* 1	*Media* 2	*Baja* 3	Total
Sí 1	650	1234	1430	3314
No 2	64	333	1036	1433
Total	714	1567	2466	4747
Razón	**10,2**	**3,7**	**1,4**	**2,3**
Razón de razones	**7,4**	**2,7**	**1**	

3. Razón de razones (*odds ratio*) o cociente de productos cruzados (*cross-product ratio*)[20]. El cociente entre razones condicionales simples permite ver la intensidad de la relación al comparar las diferentes categorías:
   - La razón de tener a no tener entre las categorías altas y bajas es: 10,2/1,4=7,4.
   - Entre altas y medias de: 10,2/3,7=2,7.
   - Entre medias y bajas de: 3,7/1,4=2,7.

La doble relación que expresa la razón de razones constituye una medida de asociación. A partir de cuatro casillas de una tabla:

	j	*j'*
i	n_{ij}	$n_{ij'}$
i'	$n_{i'j}$	$n_{i'j'}$

[20] Otras denominaciones utilizadas de razón y razón de razones son ratios: chances y chances relativas, ventajas y razón de ventajas (o ventajas relativas), momios y razones de momios.

la medida es:

$$\alpha = \frac{n_{ij} / n_{i'j}}{n_{ij'} / n_{i'j'}} \qquad \textit{Ecuación 17}$$

La interpretación de este cociente es sencilla al considerar las siguientes propiedades de la razón de razones:

1. El valor 1 indica la existencia de independencia entre las categorías consideradas, las razones condicionales respectivas son las mismas. Cualquier otro valor por encima o por debajo implica la existencia de asociación, con una dirección positiva o negativa, y cuanto más lejos del 1 mayor será la intensidad de la relación. Así:
   - 1: independencia
   - [0,1): dependencia negativa
   - (1,+∞): dependencia positiva

2. La razón de razones, como las razones condicionales, dan valores distintos en función de qué cantidad se coloca en el numerador o en el denominador. No obstante ambas magnitudes son indicativas de la misma asociación, pero no de la dirección. Así se verifica que son magnitudes inversas: $\alpha' = 1/\alpha$, que miden lo mismo pero en diferente escala: [0,1) o bien (1,+∞), es decir, no es simétrica respecto al valor 1. Para solucionar este problema se calcula el logaritmo de la razón de razones que da el mismo valor absoluto independientemente del cociente empleado, tan sólo cambiará el signo.

   $$\alpha^* = \log \alpha = \log \frac{n_{ij} / n_{i'j}}{n_{ij'} / n_{i'j'}} \qquad \textit{Ecuación 18}$$

   con $i=1...I-1$ y $j=1...J-1$.

   Que se puede expresar así:

   $$\alpha^* = \log \alpha = \log(n_{ij}) - \log(n_{i'j}) - \log(n_{ij'}) + \log(n_{i'j'})$$

   En general, $\log \alpha' = \log 1/\alpha = Log\ 1 - Log\ \alpha = -\log \alpha$.

   En este caso el rango de variación es simétrico con los valores siguientes:
   - 0: independencia
   - (-∞,0): dependencia negativa
   - (0,+∞): dependencia positiva

3. Ahora es una medida simétrica, se obtienen resultados idénticos considerando a cualquiera de las variables como dependiente.

4. Es invariable ante multiplicaciones de efectivos, lo que permite comparar tablas de tamaños de muestra diferentes. Esta propiedad se extiende a los marginales, así la razón de razones es invariante ante cambios en la distribución de los marginales de fila y de columna.

5. La razón de razones alcanza su máximo valor en situaciones de débil asociación perfecta.

Una vez obtenidas las medidas se trata de ver si éstas son significativas, calcular el intervalo de posibles valores donde se encontrará el valor poblacional. Para ello es necesario estimar la varianza que se obtiene con la fórmula:

$$\hat{\sigma}^2(\log \hat{\alpha}) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} \qquad \textit{Ecuación 19}$$

con $n_{ij} = \hat{N}_{ij}$, $\alpha = \dfrac{N_{ij}/N_{i'j}}{N_{ij'}/N_{i'j'}}$ y $\hat{\alpha} = \dfrac{n_{ij}/n_{i'j}}{n_{ij'}/n_{i'j'}}$

El intervalo de confianza es: $\left(\log \hat{\alpha} - 1{,}96 \cdot \hat{\sigma}^2(\log \hat{\alpha}), \log \hat{\alpha} - 1{,}96 \cdot \hat{\sigma}^2(\log \hat{\alpha})\right)$ con un nivel de significación del 0,05. Es decir, el valor poblacional se encuentra en un intervalo que si no incluye el valor cero se puede concluir que $\log \hat{\alpha}$ es significativamente distinto de cero.

El cálculo de la razón de razones (*odds ratio*) está en la base del razonamiento matemático que caracteriza el análisis logarítmico lineal (análisis log-lineal) que veremos en el próximo capítulo.

► **Ejercicio 7.**
Realiza un análisis de razones con los datos de las tablas de contingencia de los ejercicios 1 y 2.

4.3.4. *El estudio de la movilidad social*

El estudio de la movilidad social es una de las aplicaciones características de los análisis de tablas de contingencia.

La movilidad social intergeneracional es la que se establece entre la posición del origen social del sujeto, sea ésta la ocupación o el nivel educativo familiar en relación a su posición ocupacional actual o cualquier otra medida que pueda obtenerse de estatus social. De esta manera valoramos el cambio que se ha producido en un período de tiempo que ha transcurrido entre un origen social determinado y un destino social observado.

A partir de esta relación expresada en una tabla de contingencia se puede analizar tanto la movilidad absoluta como la movilidad relativa. Presentaremos a continuación un ejemplo de movilidad absoluta.

El estudio de Fachelli y Planas (2014)[21] sobre los titulados universitarios analiza la relación entre el origen social (expresado por el máximo nivel ocupacional del padre o de la madre) y el destino (representado por el nivel ocupacional de los hijos). Se consideran los titulados en el curso académico 2006-2007 en las universidades públicas catalanas que fueron encuestados en 2011 por la Agencia para la Calidad del Sistema

[21] Fachelli, S.; Planas, J. (2014). Inserción profesional de los universitarios: de la expansión a la crisis. *Revista Española de Sociología*, 21, 69-98. http://ddd.uab.cat/pub/artpub/2014/125654/revespsoc_a2014n21p69iSPA.pdf

Universitario de Cataluña (AQU) y que en este año estaban trabajando a tiempo completo. La tabla de contingencia que relaciona ambas variables es la siguiente.

Tabla III.6.14. La categoría ocupacional de los pares (Origen) e hijos/as (Destino)

Cataluña, 2011		Clase de Destino (Hijos e hijas)					Total	%
		Dirección	Técnico Superior	Cualificado	Cuenta propia	No Cualificado		
Clase de origen (Padre o Madre)	Dirección	426	720	191	20	8	1365	16,3
	Téc. Superior	541	1074	263	33	21	1932	23,0
	Cualificado	614	1293	406	17	50	2380	28,4
	Cuenta propia	464	871	271	33	23	1662	19,8
	No Cualificado	251	559	208	8	26	1052	12,5
	Total	2296	4517	1339	111	128	8391	100,0
	Porcentaje	27,4	53,8	16,0	1,3	1,5	100,0	

χ^2 = 77,386 (0,000); V de Cramer= 0,048

Vemos que la muestra total es de 8.391 titulados, los totales marginales de fila reflejan la clase de Origen representada por la máxima categoría ocupacional del padre o la madre y en los totales marginales de columna se representan los valores absolutos de la clase de Destino, expresada por la categoría ocupacional de los titulados universitarios. Cabe destacar que el 97,2% de los titulados se acumula en tres de las cinco categorías ocupacionales: dirección, técnico superior y cualificado.

Si bien existe asociación entre origen y destino, el hecho que la V de Cramer sea tan baja, nos está hablando de que los graduados universitarios han logrado una movilidad que se acerca a los niveles de independencia, es decir, que el origen no los está condicionando para insertarse en el mercado laboral.

Si comparamos ambas estructuras podemos observar qué categorías han aumentado y cuáles han descendido. En este caso al tratarse de graduados vemos cómo las categorías superiores han aumentado, superando a los padres en el tipo de ocupación, y los padres quedan con porcentajes altos en las categorías bajas. Estas diferencias ya nos indican un importante proceso de movilidad ascendente en un colectivo específico como es el de los graduados, que es muy particular y no representativo de la sociedad en su conjunto.

Cataluña, 2011	Padres (1)	Hijos (2)	Diferencia (2) - (1)
Dirección	16,3	27,4	11,1
Téc. Superior	23,0	53,8	30,8
Cualificado	28,4	16,0	-12,4
Cuenta propia	19,8	1,3	-18,5
No Cualificado	12,5	1,5	-11,0
Total	100,0	100,0	0,0

Analizaremos la movilidad absoluta y calcularemos la movilidad ascendente, la descendente i la inmovilidad. Realizaremos un análisis de *inflow* y *outflow*, y veremos el grado de asociación entre las variables, es decir, determinaremos hasta qué punto el origen de los padres determina el destino de los hijos e hijas (Fachelli y López-Roldán, 2013).

El porcentaje de personas con un mismo destino que procede de distintas posiciones de origen, los porcentajes que totalizan cien en cada columna, se denomina *inflow*.

Tabla III.6.15. La categoría ocupacional de los padres (Origen) según la de los hijos/as (Destino). *Inflow*

Cataluña, 2011		Clase de Destino (Hijos e hijas)				
		Dirección	Téc. Superior	Cualificado	Cuenta propia	No Cualificado
Clase de origen (Padre o Madre)	Dirección	18,6	15,9	14,3	18,0	6,3
	Téc. Superior	23,6	23,8	19,6	29,7	16,4
	Cualificado	26,7	28,6	30,3	15,3	39,1
	Cuenta propia	20,2	19,3	20,2	29,7	18,0
	No Cualificado	10,9	12,4	15,5	7,2	20,3
	Total	100,0	100,0	100,0	100,0	100,0

Fuente: Fachelli y Planas (2014) sobre la base de AQU 2011

Esta distribución permite comparar entre columnas si el origen de los titulados se diferencia mucho según realicen tareas de dirección, técnico superior o cualificado por ejemplo (recordemos que el 97,2% de los titulados se distribuyen entre estas tres categorías).

Por su parte el *outflow* es el porcentaje de personas de un mismo origen que terminan en cada una de las distintas posiciones de destino (el marginal de cada fila totaliza cien)[22].

Tabla III.6.16. La categoría ocupacional de los padres (Origen) según la de los hijos/as (Destino). *Outflow*

Cataluña, 2011		Clase de Destino (Hijos e hijas)					Total
		Dirección	Téc. Superior	Cualificado	Cuenta propia	No Cualificado	
Clase de origen (Padres)	Dirección	31,2	52,7	14,0	1,5	0,6	100,0
	Téc. Superior	28,0	55,6	13,6	1,7	1,1	100,0
	Cualificado	25,8	54,3	17,1	0,7	2,1	100,0
	Cuenta propia	27,9	52,4	16,3	2,0	1,4	100,0
	No Cualificado	23,9	53,1	19,8	0,8	2,5	100,0

Fuente: Fachelli y Planas (2014) sobre la base de AQU 2011

Al observar las distintas tareas realizadas por los titulados según origen social es interesante advertir que la influencia del origen de los padres sobre la ocupación del hijo no es muy importante, pues los titulados están relativamente representados en forma similar provengan del origen que provengan. La excepción se constata en las categorías extremas de los hijos directores, pues provienen de padres directores en un 31% mientras que lo que provienen de padres no cualificados son un 24%.

En los análisis de movilidad social junto al concepto de movilidad absoluta se maneja el concepto de movilidad relativa. Ésta responde a la pregunta siguiente: ¿cuánta

[22] A este tipo de movilidad Carabaña (1999) lo denomina movilidad particular, es decir, cuando nos preguntamos por los destinos de las personas que proceden de cada una de las categorías. La movilidad sería global cuando se toma en cuenta a un país entero o a cualquier unidad geográfica tomando a todos sus individuos conjuntamente, analizando el cambio de los porcentajes totales de origen y destino.

diferencia hay en la probabilidad de ocupar un lugar más que otro entre las personas provenientes de diferentes orígenes de clases?[23]. Es decir, expresa la diferencia en la oportunidad de ocupar un lugar en destino según el origen social de los padres.

El supuesto sobre el que se basan los modelos de movilidad relativa es que ésta se independiza del problema que tiene la movilidad absoluta, este es, depender de dos momentos históricos diferentes y de varias estructuras laborales muy diferentes. Al dejar de lado los marginales de las tablas de movilidad absoluta y concentrarnos en medidas de independencia de las casillas interiores, observamos la dinámica entre origen y destino en forma pura, esto es, independiente de la influencia de la estructura social de los momentos en que los padres y los hijos estudiaron o trabajaron respectivamente. En otras palabras, refleja la apertura de una estructura de clases, es decir, la facilidad con que la gente pasa de unas clases a otras (Carabaña, 1999: 33)[24]

La situación de perfecta movilidad o perfecta fluidez social alude a la ausencia de diferencias en las probabilidades de las personas de llegar a un destino proviniendo de un origen más que de otro. El término fluidez social y movilidad relativa son sinónimos (Breen, 2004: 21)[25] y la medida básica de la fluidez es el cálculo de la razón de razones (*odds ratio*) que bajo movilidad perfecta es igual a 1.

Estos análisis han derivado de la discusión de dos hipótesis clásicas, la de la Convergencia y la de la Constancia. La primera formulada por Lipset y Zetterberg (1959)[26] que planteaba que el proceso de industrialización llevaba a la convergencia en la estructura ocupacional de las sociedades occidentales, por tanto existirá una apertura social con lo que los orígenes sociales de los padres tendrían menos intensidad en el destino de sus hijos. La segunda hipótesis planteada por Featherman, Jones y Hausser (1975)[27] refuta la hipótesis de la convergencia y afirma que la movilidad relativa se mantiene constante. Erikson y Goldthorpe (1993)[28] confirman la hipótesis FJH y establecen los métodos para analizar la movilidad relativa en su obra magna *The Constant Flux*.

No obstante, nosotros puntualizamos que Lipset y Zetterberg aciertan en sus pronósticos cuando el análisis se hace sobre la movilidad absoluta. Efectivamente se observa un nivel de independencia entre origen y destino en el devenir de la sociedad industrial. Por su parte la segunda hipótesis también se ha venido constando cuando se analiza la sociedad occidental con los métodos derivados de la movilidad relativa.

---

[23] Cabe destacar que estas probabilidades son ex post, dado que se trata de un análisis de un hecho, esto es importante porque las ventajas o desventajas asociadas a haber nacido en una clase más que otra o de tener un nivel educativo más que otro es un dato observado. Estos resultados reflejan la desigualdad de oportunidades pero también son el producto de otras cosas y por eso debemos ser cuidadosos en su interpretación (Breen, 2004: 20).

[24] Carabaña, J. (1999). *Dos estudios sobre movilidad intergeneracional*. Madrid: Fundación Argentaria, Visor.

[25] Breen, R. (2004). *Social Mobility in Europe*. New York: Oxford University Press.

[26] Lipset, S. M.; Zetterberg, H. L. (1959). Social mobility in industrial societies. En S. M. Lipset y R. Bendix, *Social mobility in industrial society*. Berkeley: University of California Press.

[27] Featherman, D. L.; Jones, F. L.; Hauser, R. M. (1975) Assumptions of Mobility Research in the US: The Case of Occupational Status. *Social Science Research*, 4, 329-360.

[28] Erikson, R.; Golthorpe, J. H. (1993). *The constant flux*. New York: Oxford University Press.

En ese sentido es fundamental aclarar desde qué perspectiva se discute cuando se aborda la movilidad. Porque una sociedad con mucha movilidad absoluta puede tener nula movilidad relativa. Veremos este punto a continuación.

Como comentamos previamente la movilidad relativa se mide a partir de los *odds ratio* o razón de razones, cálculo que nos permite eliminar el efecto de los marginales, es decir, los cambios de estructura ocupacional en el tiempo entres padres y madres e hijos e hijas. Habiendo cambiado esta estructura la pregunta es ¿en qué medida las oportunidades de llegar a un destino han aumentado o disminuido, existe mayor o menor fluidez social? Para ilustrar esta idea seguiremos un sencillo ejemplo ilustrativo presentado por Vallet (2001: 18-19)[29]. Consideramos 2 clases ocupacionales de origen y dos de destino (profesional y trabajador) y analizamos el cambio en el tiempo entre t1 y t2:

Como se puede observar entre t1 y t2, teniendo el mismo número de casos total (1000), se produce un aumento de los profesionales (y la consecuente disminución de los trabajadores). Si calculamos la movilidad absoluta en t1: (125+75)/1000 obtenemos que es del 20%, mientras que en t2: (200+50)/1000 obtenemos que es del 25%. Podemos concluir que la movilidad absoluta aumenta y que la sociedad ha experimentado una mejora de ascenso social. Ahora la pregunta es ¿en términos de movilidad relativa han aumentado las oportunidades de llegar a la clase profesional viniendo de clase trabajadora y de clase profesional? Realicemos los cálculos:

	Probabilidad de ser profesional en:	
Origen	t1	t2
Con padres profesionales:	125/200 = 62,5%	150/200 = 75%
Con padres trabajadores:	125/800 = 15,6%	200/800 = 25%

¿Podemos decir que entre los dos momentos ha aumentado las oportunidades de llegar a profesional, ya sea viniendo de padres profesionales o de padres trabajadores? En términos relativos, calculando la razón de razones, resulta que las oportunidades son idénticas:

$$\frac{125/75}{125/75} = \frac{150/50}{200/60} = 9 \qquad \text{o bien} \qquad \frac{1,7}{0,2} = 9 = \frac{3}{0,3}$$

En t1 la razón de llegar a profesional es 1,7 viniendo de padre profesional y de 0,2 viniendo de padre trabajador. En t2 estos valores cambian a 3 y 0,3. En ambos casos la relación entre origen social y destino es la misma, una razón de 9, siendo la razón de

[29] Vallet, L. A. (2001). Forty Years of Social Mobility in France: Change in Social Fluidity in the Light of Recent Models. *Revue Française de Sociologie*, 42, 5-64.

razones invariante, 9/9=1. La sociedad cambia su estructura social (movilidad absoluta) pero las oportunidades sociales (movilidad relativa) siguen siendo las mismas.

► **Ejercicio 8.**

Con datos del Panel de Desigualtats 2009 de la Fundació Jaume Bofill, Martínez Celorrio y Marín Saldo (2010)[30] analizan la movilidad social a partir de las clases o categorías ocupacionales de Erikson, Golthorpe y Portocarero (1979)[31] y tratadas por estos autores de forma extensa en Erikson y Golthorpe (1993).

Con la siguiente tabla:

Relación entre la categoría ocupacional de los pares (Origen) e hijos (Destino)
Personas de 25 a 64 años

Cataluña, 2009		Clase de Destino (Hijos)				Total
		I-II Directivos y Profesionales	IV Pequeña Burguesía	III-V Clase media funcional	VI-VII Obreros manuales	
Clase de origen (Padre)	I-II Directivos y Profesionales	149	37	107	41	334
	IV Pequeña Burguesía	93	78	96	99	366
	III-V Clase media funcional	126	57	191	112	486
	VI-VII Obreros manuales	153	148	318	381	1000
	Total	521	320	712	633	2186

Nota: categorías ocupacionales de Erikson, Goldthorpe y Portocarero

Fuente: Martínez Celorrio y Marín Saldo, Panel de Desigualtats, 2009

- Analizar la movilidad social a partir de la lectura y la interpretación de los porcentajes de la tabla.
- Calcular las frecuencias esperadas y comparalas con las observadas y calcular los residuos.
- Obtener los residuos estandarizados y ajustados y analizarlos.
- Calcular los *odds ratio* e interpretarlos.

5. Análisis de tablas de contingencia multidimensionales

El análisis de tablas de contingencia multidimensionales nos ofrece la posibilidad de estudiar la realidad compleja y multidimensional que caracteriza a los fenómenos sociales. Después de haber encontrado una relación significativa entre dos variables la cuestión que se plantea es si la introducción de una tercera variable, llamada de control, modifica la existencia de esta relación, su naturaleza o su intensidad. En todo caso nos

[30] Martínez Celorrio, X.; Marín Saldo, A. (2010). *Educació i mobilitat social a Catalunya*. Barcelona: Fundació Jaume Bofill.

[31] Erikson, R.; Golthorpe, J. H.; Portocarero, L. (1979). Intergenerational Class Mobility in Three Western European Societies: England, France and Sweden. *The British Journal of Sociology*, 30, 4: 415-441.

planteamos cómo se relacionan simultáneamente tres o más variables y en este sentido enriquecemos nuestro análisis al tratar con más información.

La oportunidad de introducir una tercera variable o de realizar un análisis de la relación entre tres o más variables debe responder a un modelo de análisis que dé sentido a la elección de las variables o establezca hipótesis sobre los vínculos que se esperan entre ellas.

Cuando se introduce una tercera variable en el análisis se obtienen y analizan tantas tablas bidimensionales como valores tiene la tercera variable incorporada. Así por ejemplo, la relación entre la posesión de coche y la clase social podríamos analizarla según una tercera variable como el lugar de residencia, diferenciando las personas que viven en la ciudad de Barcelona o en resto del ámbito metropolitano. Analizaríamos la tabla de contingencia entre coche y clase primero para los residentes en Barcelona y la compararíamos con la de los residentes fuera de la capital catalana.

En general, la problemática que introduce el análisis multivariable de tablas de contingencia consiste en estudiar y controlar el efecto de terceras variables sobre una relación bidimensional. Se trata de ver si una tercera o cuarta variable pueden evidenciar una relación espuria, es decir, una relación inexistente, o bien menos intensa de lo que inicialmente podíamos haber previsto, o ver su interacción con otras variables, al tiempo que permite la validación interna de la relación inicial entre dos variables.

La forma de realizar este control ofrece distintas posibilidades que han caracterizado diversas disciplinas. Hay al menos tres métodos posibles de controlar este tipo de relaciones:

1. El método experimental: se realiza un experimento a partir de diferentes grupos seleccionados aleatoriamente para medir la relación entre dos variables en las distintas situaciones de experimentación. Este diseño de investigación es menos habitual en la investigación sociológica. Los otros son más sencillos y se hacen mediante el control estadístico.
2. El método de estandarización o de homogenización de poblaciones: supone recalcular las frecuencias observadas para que haya igual porcentaje de personas en cada categoría de la variable independiente, es decir, para que haya ausencia de relación entre la variable dependiente y la variable independiente.
3. El método de control por una tercera variable (*factor test*). Se trata de examinar la relación original entre dos variables en cada nivel o dentro de cada uno de los grupos que determina la tercera variable o variable de control.

Por otro lado el análisis multidimensional de una tabla de contingencia se puede extender a otras técnicas de análisis según dos tipos de objetivos metodológicos planteados en el modelo de análisis:

a) Para efectuar un análisis de dependencia donde se conceptualiza la existencia de variables dependientes e independientes. A este objetivo se destinan los modelos logarítmicos de dependencia (modelos logit), la regresión logística o el análisis de correspondencias no simétrico.

b) Para efectuar un análisis de interdependencia donde todas las variables consideradas juegan un papel simétrico, de variables independientes, sin direccionalidad preestablecida. A este objetivo se destinan los modelos generales logarítmicos lineales o el análisis factorial de correspondencias.

A continuación veremos cómo realizar la lectura de una tabla de contingencia con la introducción de una tercera variable de control, y detallaremos los modelos de análisis que se derivan de la relación entre tres variables a partir tanto de un objetivo de análisis de dependencia como de interdependencia.

5.1. El análisis con una variable de control

Un primer trabajo de análisis de tipo multidimensional consiste en considerar inicialmente una tabla de contingencia bidimensional e introducir una tercera variable de control para observar el comportamiento de los porcentajes y de las medidas de asociación para cada valor o categoría de esta tercera variable. Se trata pues de reproducir un análisis bivariable dentro de cada una de las subtablas que definen los valores de la tercera variable. Nos encontramos así con una descomposición de la relación original entre dos variables que nos permite una lectura parcial y la obtención de medidas de asociaciones parciales para cada subrupo o submuestra que establece cada categoría de la nueva variable.

En una tabla de contingencia de tres dimensiones tendremos filas, columnas y ahora niveles (o capas). La nueva notación diferenciará en la tabla I filas, indexadas por *i*, con $i=1...I$, J columnas, indexadas por *j*, con $j=1...J$, y K niveles, indexados por *k*, con $k=1...K$ que cruza tres variables cualitativas *Y*, *X* y *Z*.

La notación se complica y la tabla de frecuencias absolutas $N(I,J,K)$ se hace compleja. Para visualizar una tabla de tres dimensiones consideraremos entonces el caso más sencillo donde I=J=K=2:

N(I,J,K)		**Z**					
		k=1			*k=2*		
		X			**X**		
		j=1	*j=2*	Total	*j=1*	*j=2*	Total
Y	*i=1*	n_{111}	n_{121}	n_{1+1}	n_{112}	n_{122}	n_{1+2}
	i=2	n_{211}	n_{221}	n_{2+1}	n_{212}	n_{222}	n_{2+2}
	Total	n_{+11}	n_{+21}	n_{++1}	n_{+12}	n_{+22}	n_{++2}

En este caso tenemos dos subtablas, para *k=1* y *k=2*, donde se analiza por separado la relación entre Y y X. Los totales hacen referencia a la relación bidimensional de Y y X en cada subgrupo.

A partir de una tabla tridimensional podemos pasar a una bidimensional colapsando una de las variables, es decir, sumando las frecuencias de sus valores: si colapsamos Z sumando las frecuencias de la subtabla *k=1* con la subtabla *k=2* generaremos la tabla

de contingencia N(I,J) entre Y y X; si colapsamos X generaremos la tabla de contingencia N(I,K) entre Y y Z, y si colapsamos Y la tabla es N(J,K) entre X y Z.

N(I,J)		X		
		j=1	*j=2*	Total
Y	*i=1*	n_{11+}	n_{12+}	n_{1++}
	i=2	n_{21+}	n_{22+}	n_{2++}
	Total	n_{+1+}	n_{+2+}	n_{+++}

N(I,K)		Z		
		k=1	*k=2*	Total
Y	*i=1*	n_{1+1}	n_{1+2}	n_{1++}
	i=2	n_{2+1}	n_{2+2}	n_{2++}
	Total	n_{++1}	n_{++2}	n_{+++}

N(J,K)		Z		
		k=1	*k=2*	Total
X	*j=1*	n_{+11}	n_{+12}	n_{+1+}
	j=2	n_{+21}	n_{+22}	n_{+2+}
	Total	n_{++1}	n_{++2}	n_{+++}

Consideremos de nuevo el ejemplo de la posesión de coche relacionado con la clase social, pero ahora con solo dos categorías: alta (sumando alta y media) y baja, y considerando como tercera variable el lugar de residencia: si reside en Barcelona ciudad o en el área metropolitana. La tabla que cruza las tres variables se presenta con los datos absolutos (Tabla III.6.17) y con la distribución de porcentajes por columna (Tabla III.6.18).

Para interpretar estos resultados recordemos que el porcentaje global de posesión de coche era del 69,8%. Cuando diferenciamos entre Barcelona y el área metropolitana observamos cómo baja algo entre los primeros (67,8%) y aumenta entre los segundos (71,4%)[32], si bien las diferencias son pequeñas. Si analizamos las distribuciones condicionales según la clase social constatamos que las diferencias entre clases son algo más pronunciadas en Barcelona (79,8% - 51,9% = 27,9%) que en el área metropolitana Barcelona (85,7% - 61,5% = 24,2%). La diferente estructura social de ambos territorios, con una mayor presencia de la clase alta en la ciudad de Barcelona, da como resultado esa acentuación de la posesión entre las clases altas en relación al área metropolitana.

Tabla III.6.17. La posesión de coche según la clase social y el lugar de residencia. Datos absolutos

		Residencia Lugar de residencia					
		1 Barcelona			2 Área metropolitana		
		Clase Clase social		Total	**Clase** Clase social		Total
		1 Alta	2 Media		1 Alta	2 Media	
Coche Posesión de coche	1 Sí	958	469	1427	926	961	1887
	2 No	242	435	677	155	601	756
Total		1200	904	2104	1081	1562	2643

Fuente: *Enquesta Metropolitana de Barcelona*, 1990.

[32] Este comportamiento diferenciado se fue incrementando en el tiempo hasta nuestros días: en Barcelona el 50% dispone de automóvil frente al 74% del resto del área metropolitana de Barcelona y el 82% más amplio de la región metropolitana.

Tabla III.6.18. La posesión de coche según la clase social y el lugar de residencia. Distribución porcentual por columna

		Residencia Lugar de residencia					
		1 Barcelona			2 Área metropolitana		
		Clase Clase social		Total	**Clase** Clase social		Total
		1 Alta	2 Media		1 Alta	2 Media	
Coche	1 Sí	79,8%	51,9%	67,8%	85,7%	61,5%	71,4%
Posesión de coche	2 No	20,2%	48,1%	32,2%	14,3%	38,5%	28,6%
Total		100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Fuente: *Enquesta Metropolitana de Barcelona*, 1990.

Estos comportamientos diferenciados en cada subtabla son estadísticamente significativos como expresan las pruebas de chi-cuadrado de cada una (Tabla III.6.19). Los valores del estadístico se denominan ahora chi-cuadrados parciales.

Tabla III.6.19. La posesión de coche según la clase social y el lugar de residencia. Pruebas de chi-cuadrado

Residencia Lugar de residencia		Valor	gl	Sig.
1 Barcelona	Chi-cuadrado de Pearson	184,601[a]	1	0,000
2 Área metropolitana	Chi-cuadrado de Pearson	182,265[b]	1	0,000

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 290,88.
b. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 309,21.

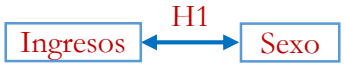
Y tiene una traducción en las medidas de intensidad de la relación que se observa entre Coche y Clase. En relación a la V de Cramer global de 0,281 obtenida de la relación bivariable, cuando se diferencian los dos territorios se concluye un aumento del grado de relación en Barcelona ciudad (V de Cramer parcial de 0,296) y una atenuación en el área metropolitana (V de Cramer parcial de 0,263) como se recoge en la Tabla III.6.20.

Tabla III.6.20. La posesión de coche según la clase social y el lugar de residencia Medida de asociación V de Cramer

Residencia Lugar de residencia		Valor	Sig.
1 Barcelona	V de Cramer	0,296	0,000
	N de casos válidos	2104	
2 Área metropolitana	V de Cramer	0,263	0,000
	N de casos válidos	2643	

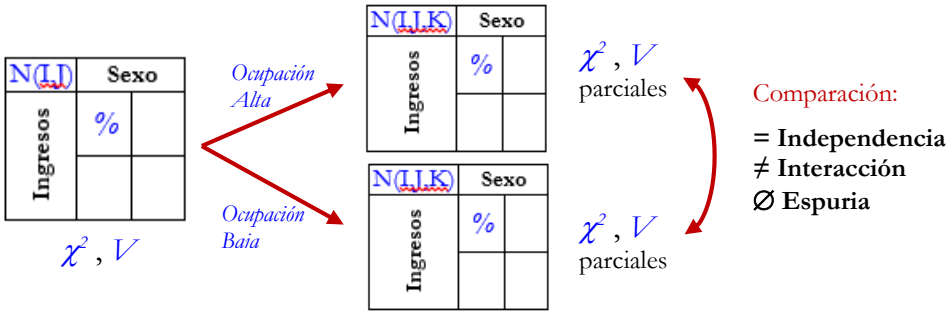
Así pues la introducción de una tercera variable en el análisis de tablas de contingencia multiplica los ejercicios de interpretación a través de tablas parciales que se comparan en su lectura y a partir de los cálculos de las medidas que estiman la asociación parcial. Las cuestiones que se suscita a continuación son: ¿Qué conclusiones se pueden extraer de estas mutuas comparaciones? ¿Cómo se establece la existencia de asociación entre las tres variables? ¿Cuál es la fuerza de esa asociación? Veremos cómo intentar responder a estas cuestiones y las limitaciones que tiene el análisis clásico de las tablas de contingencia que estamos viendo.

Consideremos ahora la relación entre la variable Ingresos, con dos categorías: altos y bajos, y las variables Sexo y Ocupación, con dos categorías también: alta y baja. Para explicar el nivel de ingresos nuestro modelo contempla una hipótesis inicial (Hipótesis 1) que establece que la distribución de los ingresos es diferente entre varones y mujeres, siendo inferiores en el caso de estas últimas[33]:



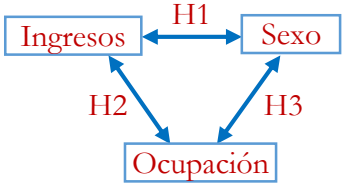
Podríamos validar esta hipótesis obteniendo un chi-cuadrado significativo y un cierto nivel de asociación entre ambas variables que mediría la V de Cramer como indicador de desigualdad social entre varones y mujeres. La pregunta que nos hacemos a continuación es qué sucede cuando consideramos la ocupación. Analizaremos por tanto dos tablas de contingencia: la que relaciona Sexo e Ingresos entre las ocupaciones altas y entre las ocupaciones bajas, tal y como se representa en el Gráfico III.6.4.

Gráfico III.6.4 Esquema del análisis con una variable de control



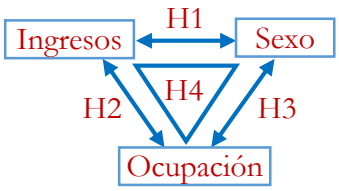
Aquí las hipótesis a testar pueden ser varias pues nos podríamos encontrar en diferentes situaciones:

- Que la relación original no cambiara, en este caso tanto entre las ocupaciones altas como entre las ocupaciones bajas las diferencias de ingresos se mantendrían constantes y el grado de asociación parcial sería similar al global. La hipótesis establecería que en cualquier nivel ocupacional la desigualdad de ingresos entre varones y mujeres se mantiene constante con igual intensidad. La conclusión sería por tanto que la tercera variable no introduce nueva información, Ocupación es independiente de la relación entre Ingresos y Sexo. No existe relación entre las tres variables simultáneamente. Por otro lado, podrían darse (o no) diferencias entre pares de ellas, constatando por ejemplo que las ocupaciones más altas tienen mayores niveles de ingresos (Hipótesis 2) y que las categorías ocupacionales son diferentes entre varones y mujeres (Hipótesis 3):

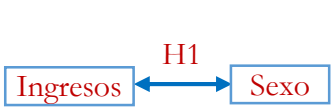


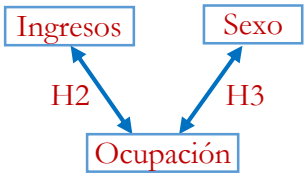
[33] Con el análisis de tablas de contingencia estamos aplicando una técnica que trata simétricamente a las variables y nos permite hablar de interrelación, por ello se representa gráficamente el vínculo con una doble flecha y no unidireccional que expresaría una relación entre variable dependiente e independiente.

- Que la relación cambiara para cada categoría ocupacional; estaríamos ante una situación de interacción en donde la intensidad de la relación original sería más fuerte en una tabla que en otra. Podríamos hipotetizar por ejemplo que las en niveles ocupacionales superiores, en la medida en que las mujeres no suelen desempeñar puestos de mando y no alcanzan los mayores niveles de ingresos, las diferencias entre varones y mujeres se acentúan. En cambio en los niveles ocupacionales inferiores las diferencias se atenúan pues no se dan tanto las situaciones de discriminación. Pero podríamos formular la hipótesis en sentido contrario, lo que nos daría igualmente una situación de interacción, afirmando que en los niveles ocupacionales superiores las diferencias de ingresos persisten pero no generan tantas diferencias entre varones y mujeres como en los niveles inferiores: en las ocupaciones altas no hay tantas diferencias ocupacionales y, en consecuencia, de ingresos, mientras que en las ocupaciones bajas se producen mayores diferencias pues las mujeres desempeñan mayoritariamente las ocupaciones menos cualificadas y menos remuneradas. Tanto en uno como en otro caso estamos ante una relación de asociación entre las tres variables que denominamos interacción y que implica un cambio de la relación bidimensional inicial (Ingresos y Sexo) a cada nivel de la tercera variable (Ocupación). Tendríamos así una hipótesis adicional (Hipótesis 4) que representamos de la forma siguiente:



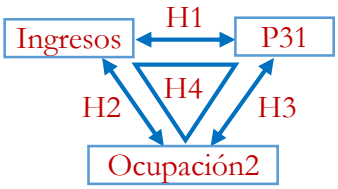
- Finalmente nos podríamos encontrar en la situación en que la relación inicial entre las variables desapareciera, en este caso la asociación parcial estaría próxima o sería cero y la prueba de significación de chi-cuadrado no sería significativa. Estaríamos ante una situación de relación espuria, es decir, donde la aparente relación de asociación que observamos en una relación bidimensional, los varones cobran más que las mujeres, desaparece cuando controlamos por la tercera variable que era la causante del vínculo entre aquéllas. Cuando se tiene en cuenta se pone de manifiesto su relación con cada una desapareciendo la relación inicial:





La hipótesis que formularíamos sería que las diferencias de ingresos de varones y mujeres de hecho existen pero no por un efecto directo sino a través de la mediación de la ocupación estableciendo un "mecanismo causal" como el siguiente: en la medida en que las mujeres ocupan profesiones menos cualificadas y éstas son menos remuneradas finalmente ingresan menos que los varones, pero a igual ocupación (variable de control) no existen diferencias de ingresos entre varones y mujeres.

Tres situaciones distintas, tres posibles modelos de análisis a probar con la realidad empírica. Veámoslo con los datos del Barómetro del CIS de la matriz de datos CIS3041. Relacionaremos la variable Ingresos (la variable original P46 de ingresos personales recodificada en dos valores: nivel de ingresos alto y bajo), la variable P31 del sexo y la variable Ocupación2 donde se han agrupado las diferentes ocupaciones de la variable original OCUMAR11 en nivel ocupacional alto y bajo. Analizaremos primero todas las relaciones bivariables entre les tres variables formulando las tres primeras hipótesis bidimensionales y a continuación analizaremos la relación multidimensional entre las tres para ver si se da una relación de interacción o no.



Las hipótesis bivariables formuladas esquemáticamente serían las siguientes:
Hipótesis 1: El nivel de ingresos de los varones es superior al de las mujeres.
Hipótesis 2: El nivel de ingresos es superior entre las ocupaciones altas.
Hipótesis 3: El nivel ocupacional de los varones es superior al de las mujeres.

La Tabla III.6.21, la Tabla III.6.22 y la Tabla III.6.23 permiten contrastarlas junto al Gráfico III.6.5.

Tabla III.6.21. Tabla de contingencia entre Ingresos y Sexo

			P31 Sexo		
			1 Hombre	2 Mujer	Total
Ingresos Nivel de ingresos	1 Bajo	Frecuencia	463	774	1237
		Frecuencia esperada	595,0	642,0	1237,0
		% por columna	**50,2%**	**77,7%**	**64,5%**
		Residuo corregido	-12,6	12,6	
	2 Alto	Frecuencia	460	222	682
		Frecuencia esperada	328,0	354,0	682,0
		% por columna	**49,8%**	**22,3%**	**35,5%**
		Residuo corregido	12,6	-12,6	
Total		Frecuencia	923	996	1919
		Frecuencia esperada	923,0	996,0	1919,0
		% por columna	**100,0%**	**100,0%**	**100,0%**

Fuente: *Centro de Investigaciones Sociológicas, Estudio 3041 de 2014.*
χ^2 =158,698 (sig. 0,000), V de Cramer = 0,288.

Tabla III.6.22. Tabla de contingencia entre Ingresos y Ocupación

			Ocupación2 Nivel ocupacional		
			1 Bajo	2 Alto	Total
Ingresos Nivel de ingresos	1 Bajo	Frecuencia	947	275	1237
		Frecuencia esperada	827,0	395,0	1237,0
		% por columna	**74,0%**	**45,0%**	**64,5%**
		Residuo corregido	12,4	-12,4	
	2 Alto	Frecuencia	332	336	682
		Frecuencia esperada	452,0	216,0	682,0
		% por columna	**26,0%**	**55,0%**	**35,5%**
		Residuo corregido	-12,4	12,4	
Total		Frecuencia	923	1279	611
		Frecuencia esperada	923,0	1279,0	611,0
		% por columna	**100,0%**	**100,0%**	**100,0%**

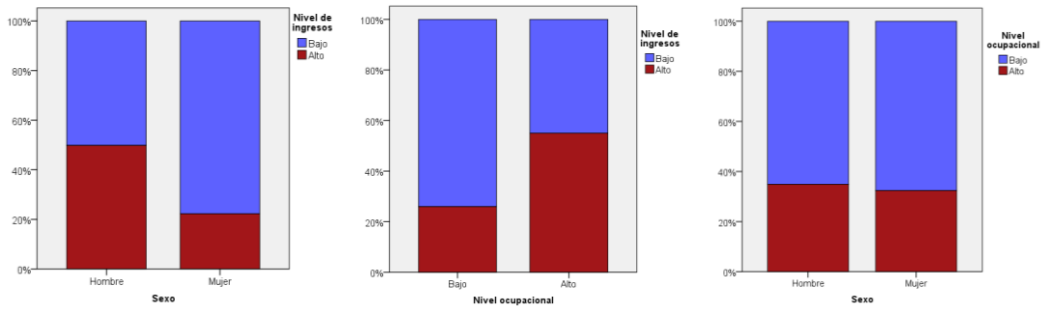
Fuente: *Centro de Investigaciones Sociológicas, Estudio 3041 de 2014*.
χ^2 =152,525 (sig. 0,000), 0% de casillas con frecuencia esperada inferior a 5, frecuencia mínima esperada 215,95, V de Cramer = 0,284.

Tabla III.6.23. Tabla de contingencia entre Ocupación y Sexo

			P31 Sexo		
			1 Hombre	2 Mujer	Total
Ocupación2 Nivel ocupacional	1 Bajo	Frecuencia	770	848	1618
		Frecuencia esperada	785,1	832,9	1618,0
		% por columna	**65,1%**	**67,6%**	**66,4%**
		Residuo corregido	-1,3	1,3	
	2 Alto	Frecuencia	412	406	818
		Frecuencia esperada	396,9	421,1	818,0
		% por columna	**34,9%**	**32,4%**	**33,6%**
		Residuo corregido	1,3	-1,3	
Total		Frecuencia	923	1182	1254
		Frecuencia esperada	923,0	1182,0	1254,0
		% por columna	**100,0%**	**100,0%**	**100,0%**

Fuente: *Centro de Investigaciones Sociológicas, Estudio 3041 de 2014*.
χ^2 =1,678 (sig. 0,195), 0% de casillas con frecuencia esperada inferior a 5, frecuencia mínima esperada 215,95, V de Cramer no significativa.

Gráfico III.6.5 Gráficos de barras bivariables entre Ingresos, Sexo y Ocupación

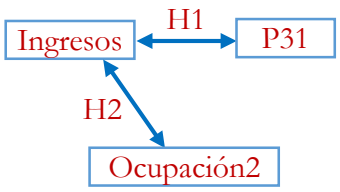


(a) Ingresos según sexo (b) Ingresos según ocupación (c) Ocupación según sexo

La hipótesis 1 se confirma, el 49,8% de los varones tiene nivel de ingresos alto frente al 22,3% de las mujeres, una brecha de 27,5 puntos porcentuales, con un grado de asociación de 0,288 medido a través de la V de Cramer.

La hipótesis 2 se valida igualmente, entre quienes tienen una ocupación alta existe una probabilidad del 55% de alcanzar niveles de ingresos altos, mientras que esa probabilidad se reduce al 26% cuando el nivel ocupacional es bajo. En este caso la intensidad de la asociación es de 0,284.

La hipótesis 3, por el contrario, no se puede verificar. En contra de nuestra afirmación inicial esperada, varones y mujeres muestran niveles ocupacionales similares en torno al 66% en ocupaciones bajas y al 34% en ocupaciones altas. Resultado relativamente sorprendente[34].
Nuestro modelo de relaciones bivariables quedaría así:



Seguidamente analizaremos la relación simultánea entre las tres variables. Ante todo cabe destacar que una relación bivariable con dos variables solamente no tiene que arrojar los mismos resultados cuando su relación se examina en el conjunto de tres variables. La introducción de una tercera variable puede alterar completamente la relación inicial bidimensional. Nuestro interés inicial era preguntarnos hasta qué punto los ingresos diferían entre hombres y mujeres como resultado de ocupar categorías ocupacionales diferentes, entendiendo que a igual ocupación los ingresos de varones y mujeres deberían ser los mismos. Podemos formular nuestra cuarta hipótesis en este sentido y afirmar que la desigualdad de ingresos entre varones y mujeres se acentúa en los niveles ocupacionales más bajos mientras se atenúa en los más altos, poniendo de manifiesto la existencia de una interacción entre las variables.

Si analizamos los datos de la Tabla III.6.24 y Tabla III.6.25 parece que podemos validar nuestra hipótesis pues las diferencias de ingresos entre varones y mujeres son de un 29,2% entre las ocupaciones bajas y de un 21,5% entre las altas. Estas diferencias son estadísticamente significativas e implican distintos grados de intensidad que se cifran en 0,332 y 0,216 respectivamente según se obtiene de la V de Cramer. Por lo tanto, las diferencias de ingresos se mantienen al mirar el tipo de ocupación y se agravan entre las ocupaciones menos cualificadas. La explicación de este comportamiento viene dado en parte por la variabilidad interna de las ocupaciones y la segregación ocupacional

[34] Invitamos al lector/a a que analice la relación entre las variables OCUMAR11 y Ocupación, ésta última creada en el capítulo III.2 y disponible en matriz CIS3041+.sav, con el sexo. Con la variable original desagregada los niveles ocupacionales más bajos diferencian a hombres y mujeres especialmente: ellas están más presentes en las ocupaciones elementales y como trabajadoras de servicios frente a los varones que predominan en las categorías de operadores y trabajadores cualificados. En los niveles ocupacionales altos sin embargo, las diferencias internas apenas se observan. Un análisis más a fondo comparando las ocupaciones en el tiempo mostraría tanto el aumento de los niveles altos y bajos de nuestra estructura ocupacional, hecho que alimentaría la hipótesis de la polarización, y de la incorporación creciente de la mujer al mercado de trabajo siguiendo este patrón pero alcanzado especialmente niveles ocupacionales altos.

entre varones y mujeres, y en parte por el inferior nivel de ingresos que reciben las mujeres para trabajos que tienen un nivel ocupacional similar al de los varones[35].

Tabla III.6.24. Tabla de contingencia entre Ingresos, Sexo y Ocupación

			Ocupación2 Nivel ocupacional					
			1 Bajo			2 Alto		
			P31 Sexo			**P31** Sexo		
			1 Hombre	2 Mujer	Total	1 Hombre	2 Mujer	Total
Ingresos Nivel de ingresos	1 Bajo	Frecuencia	352	595	947	104	171	275
		Frecuencia esperada	445,0	502,0	947,0	136,8	138,2	275,0
		% por columna	**58,6%**	**87,8%**	**74,0%**	**34,2%**	**55,7%**	**45,0%**
		Residuo corregido	-11,9	11,9		-5,3	5,3	
	2 Alto	Frecuencia	249	83	332	200	136	336
		Frecuencia esperada	156,0	176,0	332,0	167,2	168,8	336,0
		% por columna	**41,4%**	**12,2%**	**26,0%**	**65,8%**	**44,3%**	**55,0%**
		Residuo corregido	11,9	-11,9		5,3	-5,3	
Total		Frecuencia	601	678	1279	304	307	611
		Frecuencia esperada	601,0	678,0	1279,0	304,0	307,0	611,0
		% por columna	**100,0%**	**100,0%**	100,0%	**100,0%**	**100,0%**	100,0 %

Fuente: *Centro de Investigaciones Sociológicas, Estudio 3041 de 2014*.

Tabla III.6.25. Tabla de contingencia entre Ingresos, Sexo y Ocupación
Prueba de chi-cuadrado y V de Cramer

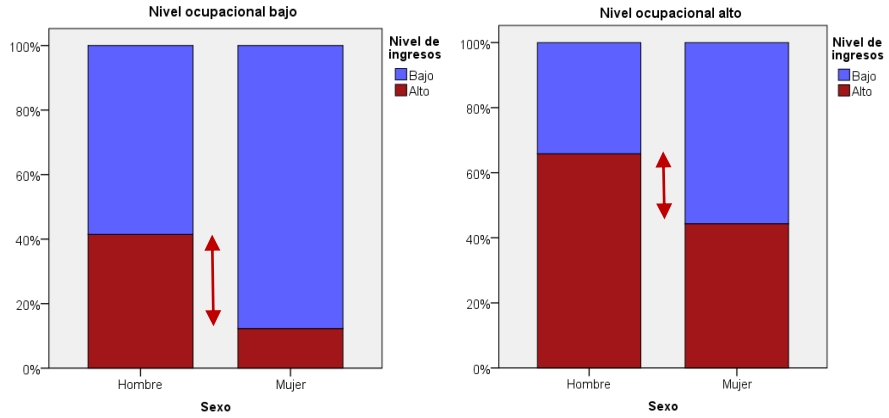
Ocupación2 Nivel ocupacional		Valor	gl	Sig.
1 Bajo	Chi-cuadrado de Pearson	141,230[a]	1	0,000
	V de Cramer	0,332	1	0,000
	N de casos válidos	1279		
2 Alto	Chi-cuadrado de Pearson	28,500[b]	1	0,000
	V de Cramer	0,216	1	0,000
	N de casos válidos	611		

a. 0% de casillas con frecuencia esperada inferior a 5, frecuencia mínima esperada 156,01.
b. 0% de casillas con frecuencia esperada inferior a 5, frecuencia mínima esperada 136,82.

El Gráfico III.6.6 ilustra gráficamente el resultado: además de constatar que las ocupaciones altas ingresan más, ya sean varones o mujeres, la brecha entre ambos sexos es mayor en las ocupaciones inferiores.

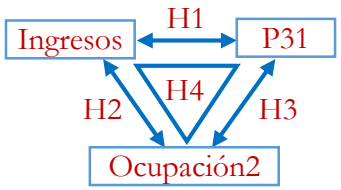
[35] Un análisis más detallado de los niveles ocupacionales con la variable OCUMAR11 permite ver este aspecto.

Gráfico III.6.6 Gráfico de barras de Ingresos, Sexo y Ocupación



Cuando se analizan tres variables existen distintas alternativas de tablas cruzadas en función de la lectura entre variables dependientes e independientes. En particular no hemos comentado la relación entre ocupación y sexo controlada por ingresos. Hemos visto cómo ocupación y sexo no tenían una relación bivariable, si hiciéramos el ejercicio de controlar por ingresos conseguiríamos hacer emerger una relación oculta existente entre ellas: a niveles bajos de ingresos la estructura ocupacional de varones y mujeres es la misma pero a niveles altos la estructura de ocupaciones de las mujeres es mayor que la de los hombres, dicho de otro modo, para llegar a alcanzar niveles de ingresos altos las mujeres tienen que acreditar una cualificación alta, condición necesaria que se le exige mayor medida que a los varones.

Resultados de todos estos análisis se constata pues la cuarta hipótesis de interacción y también de todas las relaciones bivariables. Nuestro modelo final sería el siguiente:



No obstante, estas conclusiones basadas en la comparación de porcentajes y de medidas de asociación ¿hasta qué punto muestran diferencias significativas de las relaciones entre las tres variables?, ¿qué prueba estadística establece las existencia de diferencias significativas entre las dos subtablas para afirmar la existencia de interacción? No existe esta prueba en un análisis clásico de tablas de contingencia como el que acabamos de realizar, y, en consecuencia, no podemos afirmar de forma categórica que las diferencias observadas sean suficientemente significativas. Esta es una importante limitación de la técnica: la imposibilidad de establecer pruebas de significación estadística sobre la asociación entre tres o más variables. Así que nos quedaremos con el interrogante de saber si existe interacción entre las tres variables y si existen diferencias importantes entre los niveles ocupacionales alto y bajo. Veremos en el próximo capítulo cómo el análisis log-lineal nos permitirá resolver la cuestión: podremos confirmar que efectivamente estas diferencias son significativas.

► **Ejercicio 9.**
El 14 de abril del año 1912 el barco del correo real *Titanic* se hundió en su viaje inaugural. Disponemos de los datos sobre si sobrevivió el pasajero en relación a la su clase del pasaje y su sexo que presentamos a continuación en la tabla de contingencia que las relaciona:

Sexo del pasajero	Sobrevive al hundimiento	Clase del pasaje			Total
		1a	2a	3a	
Varón	Sí	61	22	85	168
	No	111	150	419	680
Mujer	Sí	126	40	101	267
	No	6	13	107	126
Total		304	225	712	1241

Analizar si sobrevive el pasajero en función de la clase, por un lado, y de su sexo, por el otro, y contestar a las cuestiones siguientes:
- De la lectura de los porcentajes ¿a qué conclusión se llega?, ¿cuál es el perfil del "no superviviente"?
- ¿Qué fue más importante para sobrevivir, la clase social o el sexo? Calcula las medidas de asociación de cada tabla bivariable para determinarlo.
- Calcula manualmente los valores del chi-cuadrado observado de cada tabla y la V de Cramer.
- ¿Cuántos grados de libertad se consideran en cada tabla? ¿Cómo se determinan? ¿Por qué son diferentes?
- En cada caso calcula el valor máximo del chi-cuadrado. ¿Cuál es el valor mínimo?
- En cada caso ¿cuál es el valor del chi-cuadrado observado y del chi-cuadrado teórico? Considera un nivel de significación del 0,05.

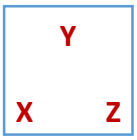
Analizar la mortandad según el sexo para cada clase de pasaje ¿qué conclusión se extrae? Obtener los estadísticos de chi-cuadrado y V de Cramer parciales para realizar las comparaciones.

5.2. Modelos de relaciones de interdependencia

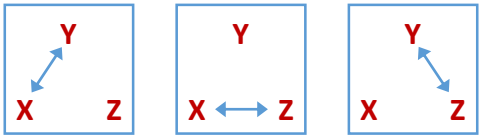
Al analizar la relación entre un conjunto de variables cualitativas sin considerar una direccionalidad de las relaciones que distingue entre variable dependiente e independiente, podemos distinguir diferentes tipos de modelos de interrelación en función de la presencia y ausencia de asociaciones.

Si consideramos el caso de tres variables X, Y, Z se pueden contemplar los siguientes modelos:

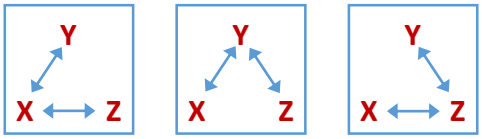
1. Modelo de independencia mutua: las tres variables son independientes entre sí.



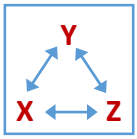
2. Modelos de asociación condicional (o de independencia parcial): se da una asociación entre dos de las tres variables y esta relación es independiente de la tercera variable, es decir, se repite la misma pauta de asociación a cada nivel de la tercera, produciéndose una independencia parcial. Los tres modelos posibles son los siguientes:



3. Modelos de doble asociación (o de independencia condicional): se da una doble asociación parcial entre dos de las tres variables. Cada par de variables asociadas es condicionalmente independiente, las relación muestra un patrón que se repite para cada valor de la tercera variable, aunque no lo sean cuando no se considera esta tercera variable. Son modelos donde se puede evidenciar la existencia de relaciones espurias como hemos visto anteriormente. Los tres modelos posibles son los siguientes:



4. Modelos de asociación homogénea. Se dan tres dobles asociaciones entre las variables, y cada pareja de relaciones vuelve a ser independiente del valor de la tercera variable. El modelo es único:



5. Modelo de interacción (de asociación triple, de orden 3). Cualquier par de variables que se considere están relacionadas y esta relación varía en intensidad o en su naturaleza para cada valor o categoría de la tercera variable. Se trata del modelo más complejo:

Los dos ejemplos de tablas de contingencia multidimensionales que hemos tratado en el capítulo, analizando la posesión de coche y los ingresos, corresponden a este último modelo de interacción, a tenor de los resultados de un análisis clásico donde, recordemos, no disponemos de una respuesta definitiva de significación estadística del modelo. En los ejercicios con el software estadístico veremos algún caso de modelo más parsimonioso donde no se da la interacción. Cuando veamos el análisis log-lineal en el próximo capítulo, y podamos aplicar pruebas rigurosas de significación de las relaciones entre tres y más variables, destacaremos también los efectos particularmente "adversos" de la tendencia a obtener relaciones siempre significativas cuando el tamaño de la muestra es grande, con muestras de 2000 y más casos, que obliga a considerar modelos complejos aunque la intensidad de la relación sea débil.

5.3. El análisis de relaciones de dependencia

Cuando formulamos un modelo de análisis donde se establece la existencia de variables dependientes e independientes y de relaciones de dependencia entre ellas disponemos de un modelo explicativo. Para poder afirmar teórica y empíricamente que entre dos variables de este modelo existe además una relación causal se exigen condiciones particulares: además de que ambas variables estén relacionadas, una variable tiene que ser causa o anteceder el comportamiento de la otra variable o efecto, y esta relación no debe ser aparente (espuria o casual).

Los modelos de relaciones de dependencia entre variables pueden ser diversos y no suponen en un análisis de tablas de contingencia un tratamiento distinto de la información. Recordemos que estamos ante una técnica que trabaja de forma simétrica, si bien algunas medidas son direccionales y nuestros modelos teóricos y la lectura que hacemos de la información las establecemos en términos explicativos de la forma "los cambios en la variable previa X produce cambios en la variable efecto Y":

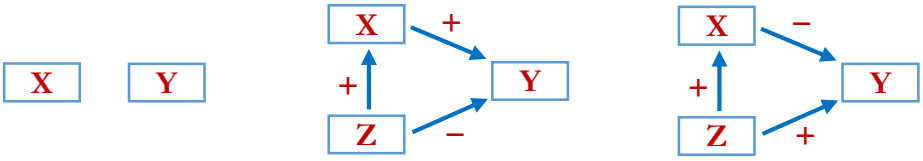
En un contexto multivariable con la introducción de terceras se pueden plantear igualmente modelos y análisis diversos:

1. Para evidenciar relaciones espurias que revelen la inexistencia de una relación inicial entre una pareja de variables. En un análisis de relaciones de dependencia, un primer trabajo de análisis consiste en determinar si la relación aparente e inicial relación entre X e Y por una tercera variable Z que hace que la relación desaparezca, sea puramente estadística, pues aquéllas dependen de ésta.

2. Para plantear modelos de secuencias causales donde X está indirectamente relacionada con Y a través de la variable interviniente Z.

   Un cambio en X produce un cambio en Z que a su vez produce un cambio en Y.

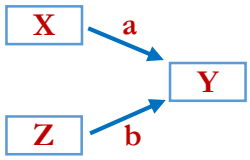
3. Para hacer emerger relaciones ocultas que inicialmente no se constatan. Una relación entre dos variables puede ser baja o inexistente y puede intensificarse o emerger cuando introducimos una tercera. Se trata de situaciones donde las relaciones entre las variables son de signo diferente y de una magnitud tal que el efecto se contrarresta por el otro.



En los gráficos adjuntos se muestra que la aparente inicial no relación entre X e Y en realidad sí existe, pero la suma de efectos negativos con positivos se compensan. Nos encontramos ante una relación oculta, situaciones donde la introducción de una tercera variable de control provoca su aparición.

4. Para plasmar, de forma similar al caso anterior, situaciones donde se produce una supresión de la relación en un cierto grado.

5. Para modelizar causas múltiples donde dos de las variables son variables independientes que simultáneamente actúan en el comportamiento de la tercera. No se trata estrictamente de analizar la relación entre dos variables según una tercera, sino de una sola variable dependiente en función de otras dos.



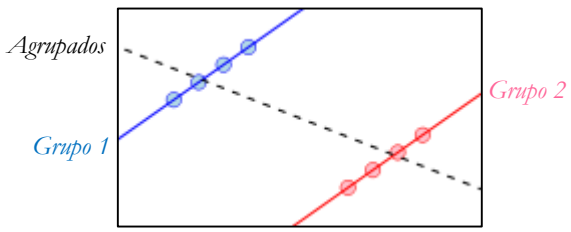
En próximos capítulos veremos cómo otras técnicas como el análisis log-lineal logit o el análisis de regresión logística nos facilitaran el tratamiento simultáneo y más amplio de relaciones de dependencia.

5.4. La paradoja de Simpson

Trataremos un caso particular de relaciones entre variables que se denomina como la Paradoja de Simpson en honor a su creador Eduard H. Simpson (1951), si bien fue descrita anteriormente por George Udny Yule (1903), por lo que se conoce también como el efecto Yule-Simpson.

Como paradoja se plantea la existencia de una contradicción u oposición en relación a un estado de las cosas, siendo verosímil y con apariencia de certeza se puede mostrar que no es cierta. En este caso se aplica en situaciones donde se analiza la relación entre dos variables que conduce a una conclusión frente a la conclusión opuesta que muestra la interacción con una tercera variable. Así, todo cambio, desaparición o emergencia del sentido de una asociación entre dos variables (numéricas o cualitativas) cuando se controla el efecto de una tercera variable (variable de confusión) se puede identificar como la Paradoja de Simpson.

La contraposición de resultados es posible cuando la población de estudio se divide en subpoblaciones definidas por una tercera variable que revierte el sentido de la asociación. Por ejemplo una tendencia positiva de dos grupos, cuando se agrupan, genera una tendencia negativa:



Existen diversos ejemplos que la literatura ha presentado, se pueden consultar Blyth (1972), quien le dio el nombre a la paradoja, Bickel, Hammel y O'Connell (1975), Wainer (1986), Gaviria (1999) o Schneiter y Symanzik (2013). Aquí trataremos el caso de Bickel et al. sobre la discriminación por sexo en la admisión a los estudios universitarios. Tras recoger la información de los resultados de la admisión a los estudios de postgrado de la *Univesity of California Berkeley* en el año 1973 se cuestionaron dichos resultados por arrojar resultados que se consideraron sexistas pues el 44,5% de los varones fueron admitidos frente al 30,4% de las mujeres. La pregunta que se planteó fue ¿hasta qué punto fueron discriminadas las mujeres en el acceso a los estudios de postgrado?

La información original que detonó la denuncia de los resultados fue la relación entre la admisión y el sexo que se recoge en la Tabla III.6.26, donde observamos los resultados comentados de admisión entre varones y mujeres.

Tabla III.6.26. Tabla de contingencia entre Admisión y Sexo

			S Sex of applicants		
			1 Woman	2 Men	Total
A Admission outcome	1 Admitted	Frecuencia	557	1198	1755
		% por columna	**30,4%**	**44,5%**	**38,8%**
	2 Rejected	Frecuencia	1278	1494	2772
		% por columna	**69,6%**	**55,5%**	**61,2%**
Total		Frecuencia	1835	2692	4527
		% por columna	**100,0%**	**100,0%**	**100,0%**

Fuente: Bickel et al. (1975)

Una de las informaciones adicionales disponibles sugirió analizar qué había sucedido en los distintos departamentos de la universidad donde se realizaron las pruebas de admisión. La variable del departamento diferencia seis valores, de A a F. Si miramos en primer lugar el porcentaje de admisión según el departamento (Tabla III.6.27) constatamos que algunos departamentos rechazaron a más candidatos, el C y el D y sobre todo el E y F.

Tabla III.6.27. Tabla de contingencia entre Admisión y Departamento

			D Department						
			1 A	2 B	3 C	4 D	5 E	6 F	Total
A Admission outcome	1 Admitted	Frecuencia	601	370	322	269	147	46	1755
		% por columna	**64,3%**	**63,2%**	**35,1%**	**34,0%**	**25,2%**	**6,4%**	**38,8%**
	2 Rejected	Frecuencia	333	215	596	523	437	668	2772
		% por columna	**35,7%**	**36,8%**	**64,9%**	**66,0%**	**74,8%**	**93,6%**	**61,2%**
Total		Frecuencia	934	585	918	792	584	714	4527
		% por columna	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Fuente: Bickel et al. (1975)

Si vemos a qué departamentos se presentaron varones y mujeres vemos que la mujeres se presentaron en mayor medida que los varones en los departamentos C a F (Tabla III.6.28) que fueron los que más rechazaron.

Tabla III.6.28. Tabla de contingencia entre Departamento y Sexo

			S Sex of applicants		
			1 Woman	2 Men	Total
D Department	1 A	Frecuencia	108	826	934
		% por columna	**5,9%**	**30,7%**	**20,6%**
	2 B	Frecuencia	25	560	585
		% por columna	**1,4%**	**20,8%**	**12,9%**
	3 C	Frecuencia	593	325	918
		% por columna	**32,3%**	**12,1%**	**20,3%**
	4 D	Frecuencia	375	417	792
		% por columna	**20,4%**	**15,5%**	**17,5%**
	5 E	Frecuencia	393	191	584
		% por columna	**21,4%**	**7,1%**	**12,9%**
	6 F	Frecuencia	341	373	714
		% por columna	**18,6%**	**13,9%**	**15,8%**
Total		Recuento	1835	2692	4527
		% por columna	**100,0%**	**100,0%**	**100,0%**

Fuente: Bickel et al. (1975)

Ya podemos intuir qué sucedió. Cuando analizamos la admisión según el sexo controlando por el departamento obtenemos la Tabla III.6.29 que muestra cómo el porcentaje de admitidos en ambos sexos fue muy similar en todos los departamentos, excepto en el departamento A donde las mujeres fueron más admitidas que los varones.

Tabla III.6.29. Tabla de contingencia de Admisión según Sexo y Departamento

				S Sex of applicants		
D Department				1 Woman	2 Men	Total
1 A	**A** Admission outcome	1 Admitted	Frecuencia	89	512	601
			% por columna	**82,4%**	**62,0%**	**64,3%**
		2 Rejected	Frecuencia	19	314	333
			% por columna	**17,6%**	**38,0%**	**35,7%**
	Total		Frecuencia	108	826	934
			% por columna	**100,0%**	**100,0%**	**100,0%**
2 B	**A** Admission outcome	1 Admitted	Frecuencia	17	353	370
			% por columna	**68,0%**	**63,0%**	**63,2%**
		2 Rejected	Frecuencia	8	207	215
			% por columna	**32,0%**	**37,0%**	**36,8%**
	Total		Frecuencia	25	560	585
			% por columna	**100,0%**	**100,0%**	**100,0%**
3 C	**A** Admission outcome	1 Admitted	Frecuencia	202	120	322
			% por columna	**34,1%**	**36,9%**	**35,1%**
		2 Rejected	Frecuencia	391	205	596
			% por columna	**65,9%**	**63,1%**	**64,9%**
	Total		Frecuencia	593	325	918
			% por columna	**100,0%**	**100,0%**	**100,0%**

4 D	**A** Admission outcome	1 Admitted	Frecuencia	131	138	269
			% por columna	**34,9%**	**33,1%**	**34,0%**
		2 Rejected	Frecuencia	244	279	523
			% por columna	**65,1%**	**66,9%**	**66,0%**
	Total		Frecuencia	375	417	792
			% por columna	**100,0%**	**100,0%**	**100,0%**
5 E	**A** Admission outcome	1 Admitted	Frecuencia	94	53	147
			% por columna	**23,9%**	**27,7%**	**25,2%**
		2 Rejected	Frecuencia	299	138	437
			% por columna	**76,1%**	**72,3%**	**74,8%**
	Total		Frecuencia	393	191	584
			% por columna	**100,0%**	**100,0%**	**100,0%**
6 F	**A** Admission outcome	1 Admitted	Frecuencia	24	22	46
			% por columna	**7,0%**	**5,9%**	**6,4%**
		2 Rejected	Frecuencia	317	351	668
			% por columna	**93,0%**	**94,1%**	**93,6%**
	Total		Frecuencia	341	373	714
			% por columna	**100,0%**	**100,0%**	**100,0%**

Fuente: Bickel et al. (1975)

Si realizamos las pruebas de significación estadística de chi-cuadrado comprobamos que las diferencias no son significativas entre hombres y mujeres en ningún departamento excepto en el primero.

Tabla III.6.30. Pruebas de chi-cuadrado de Admisión y Sexo según Departamento

D Department	Valor de Chi-cuadrado	gl	Sig.
1 A	17,363	1	0,000
2 B	0,254	1	0,614
3 C	0,754	1	0,385
4 D	0,298	1	0,585
5 E	1,001	1	0,317
6 F	0,384	1	0,535

Se produce un resultado paradójico, la mujeres aparentemente fueron discriminadas sin embargo la razón que explica la menor admisión es que la mayor parte se presentó en los departamentos que más rechazaron, pero donde fueron consideradas por igual que sus compañeros masculinos, por lo que en el balance final resultan más rechazadas pero no discriminadas. Es un efecto de concentración en determinados valores de una tercera variable que revierte la información y produce dos informaciones aparentemente contradictorias: la mirada bivariable y la mirada multivariable, pero no lo son y en el primer caso extraeríamos una conclusión equivocada[36].

[36] Todos estos resultados se pueden reproducir con el archivo ATC-Berkeley.sps de la página web. En http://www.math.usu.edu/~schneit/CTIS/SP/index.html o en http://vudlab.com/simpsons/ se puede reproducir un *applet* de éste y otros ejemplos que reproducen gráficamente la paradoja.

► **Ejercicio 10.**
Radelet (1981)[37] analiza la relación entre la pena impuesta (pasillo de la muerte o prisión) a un grupo de condenados que se clasifican según la raza de éstos (blanca o negra) y la raza de las víctimas a partir de los datos siguientes:

Tabla de contingencia

			V Victim's race								
			1 White			2 Black			Total		
			R Murder's Race			R Murder's Race			R Murder's Race		
			1 White	2 Black	Total	1 White	2 Black	Total	1 White	2 Black	Total
P Penalty imposed	1 Death row	Recuento	19	11	30	0	6	6	19	17	36
		% dentro de R Murder's Race	12,6%	17,5%	14,0%	0,0%	5,8%	5,4%	11,9%	10,2%	11,0%
	2 Prison	Recuento	132	52	184	9	97	106	141	149	290
		% dentro de R Murder's Race	87,4%	82,5%	86,0%	100,0%	94,2%	94,6%	88,1%	89,8%	89,0%
Total		Recuento	151	63	214	9	103	112	160	166	326
		% dentro de R Murder's Race	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%	100,0%

Comparar la tabla bidimensional que relaciona la pena impuesta con la raza del condenado y contrastarla con los resultados de introducir la raza de la víctima.

6. Análisis de tablas de contingencia con SPSS y R

Veremos a continuación cómo utilizar el software para analizar la relación entre variables cualitativas mediante tablas de contingencia. Presentaremos una primera parte destinada al análisis de dos variables que extenderemos al caso multivariable relacionando tres variables. En cada apartado veremos cómo generar la información de las tablas para dar cuenta de la(s) hipótesis estudiada, realizaremos el ejercicio de lectura de las mismas y obtendremos las representaciones gráficas de barras de ayuda para la presentación de resultados (aspecto descriptivo y de contenido del análisis) y de las medidas de asociación para dar cuenta estadísticamente de la(s) hipótesis (aspecto inferencial del análisis). También veremos un procedimiento instrumental para introducir directamente los datos de frecuencia de una tabla de contingencia sin que sea necesario disponer de la matriz de datos original[38].

6.1. Análisis de tablas de contingencia con SPSS

6.1.1. *Análisis descriptivo con dos variables*

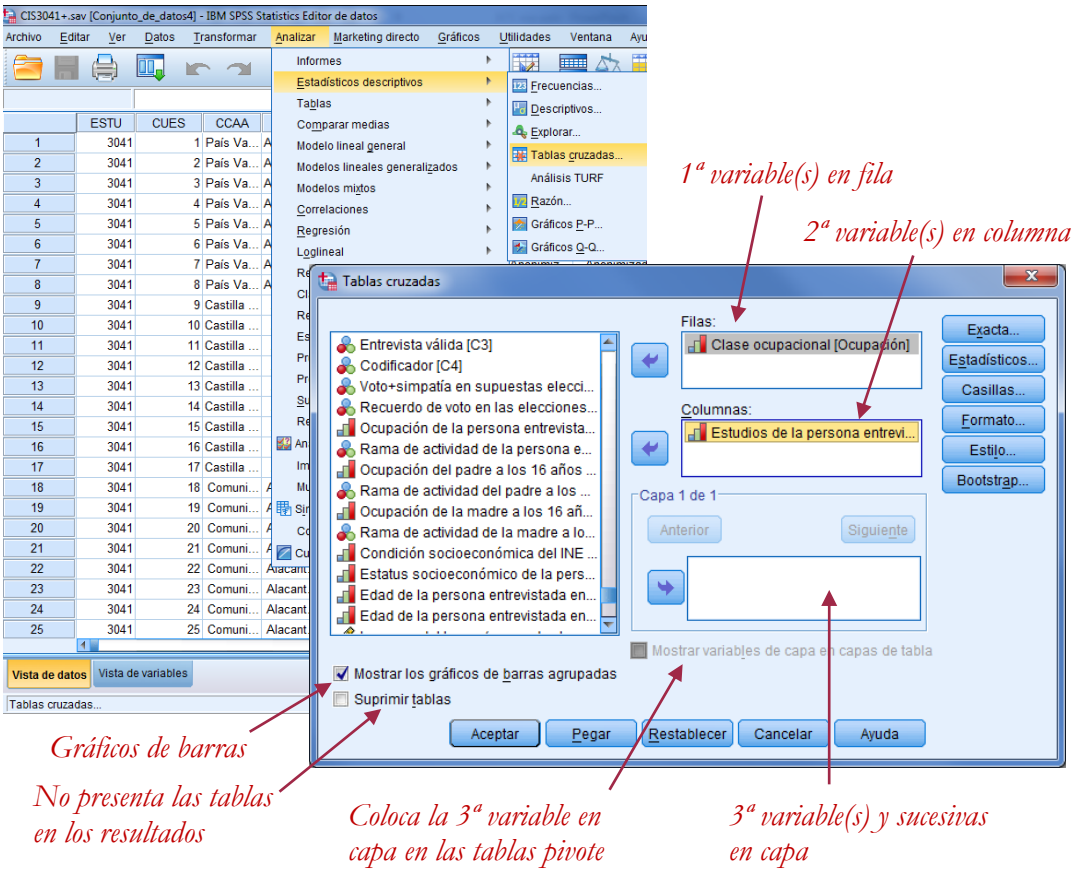
Para analizar la relación entre dos (o más variables) de tipo cualitativo, variables medidas a nivel nominal u ordinal, disponemos del procedimiento de Tablas cruzadas[39] que nos proporciona tablas de distribución conjunta de frecuencias con el cálculo de varias medidas que nos evalúan la existencia de asociación y su intensidad.

[37] Radelet, M. L. (1981). Racial characteristics and imposition of the death penalty. *American Sociological Review*, 46, 6, 918-927. Radelet, M. L., Pierce, G. L. (1991). Choosing Those Who Will Die: Race and the Death Penalty in Florida. *Florida Law Review*, 43, 1, 1-34.
[38] Todos los ejercicios propuestos hasta ahora se pueden reproducir con el software estadístico como veremos en este apartado, permitiendo verificar los resultados del análisis y ejercitar el uso del mismo software.
[39] Tablas cruzadas es la traducción directa del inglés de *crosstabs*. que se utiliza desde la versión 22 en lengua española para identificar este procedimiento que siempre se llamó Tablas de contingencia.

Analizaremos la relación entre la ocupación y los estudios para estudiar la asociación entre el nivel profesional alcanzado (variable dependiente) según el nivel de educación formal alcanzado (variable independiente). Nuestra hipótesis plantea en términos generales que a mayor nivel de estudios cabe esperar un mayor nivel ocupacional. Consideraremos los datos de la CIS3041+.sav donde generó la variable agrupada de ocupación con el nombre Ocupación y la variable original de la base de datos del CIS ESTUDIOS.

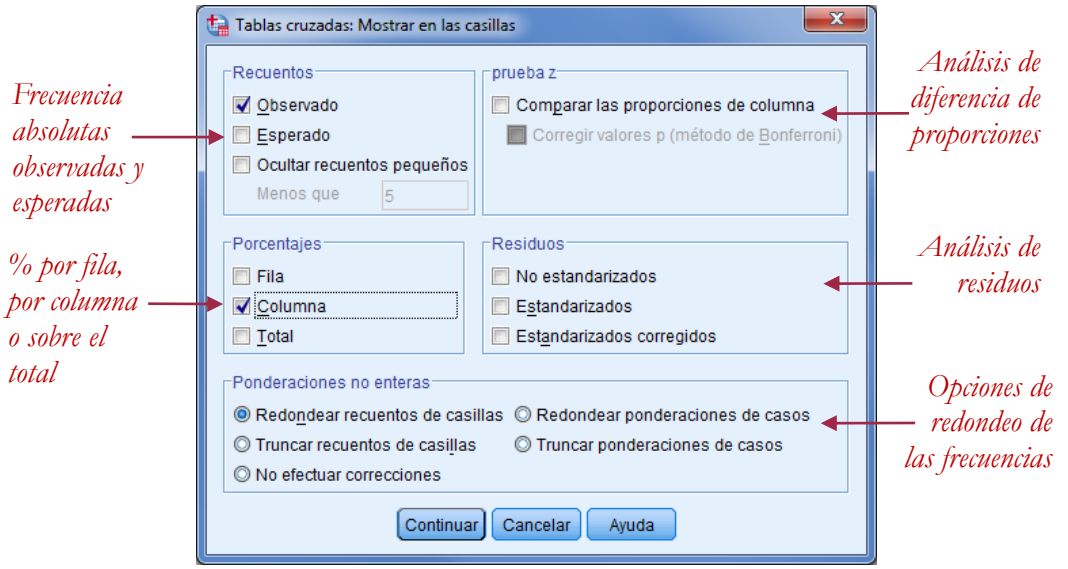
El análisis de tablas de contingencia se realiza a través del comando **CROSSTABS**, el cual se puede ejecutar a través del menú mediante la siguiente selección: Analizar / Estadísticos descriptivos / Tablas cruzadas.

El cuadro de diálogo inicial que aparece a continuación nos muestra por un lado el recuadro con el listado de variables de la matriz de datos y unos recuadros donde se trasladarán las variables según se sitúen en filas o en columnas, o bien en capas si se consideran tablas de más de dos variables o dimensiones. Situaremos la variable dependiente en las filas (Ocupación), y la variable independiente en las columnas (ESTUDIOS)[40]. Podemos marcar igualmente sobre Mostrar los Gráficos de barras Agrupados para obtener una representación gráfica.



[40] Qué variable se coloca en fila o en columna es pura convención, el análisis y el resultado es simétrico y se genera la misma información.

A continuación disponemos de diversas opciones que permiten especificar la información que queremos obtener del análisis. En un primer ejercicio de análisis descriptivo solicitaremos simplemente las frecuencias absolutas observadas o "recuento observado" (opción por defecto) y los porcentajes por columna (pues la variable independiente está en columnas) a través del botón Casillas:



En el apartado de Porcentajes de este cuadro de diálogo también podemos seleccionar los tanto por ciento por fila y sobre el total. En un análisis descriptivo no requerimos más información que estas frecuencias y una representación gráfica. Los resultados que se obtiene son los siguientes.

Ocupación Clase ocupacional*ESTUDIOS Estudios de la persona entrevistada tabulación cruzada

			ESTUDIOS Estudios de la persona entrevistada						
			1 Sin estudios	2 Primaria	3 Secundaria 1ª etapa	4 Secundaria 2ª etapa	5 F.P.	6 Superiores	Total
Ocupación Clase ocupacional	1 Clase alta	Recuento	1	14	23	54	33	268	393
		% dentro de ESTUDIOS Estudios de la persona entrevistada	0,7%	2,8%	3,8%	16,2%	8,0%	59,8%	16,1%
	2 Clase media	Recuento	2	44	69	101	107	102	425
		% dentro de ESTUDIOS Estudios de la persona entrevistada	1,4%	8,9%	11,5%	30,2%	25,8%	22,8%	17,5%
	3 Trababajadores cualificados	Recuento	77	265	311	128	203	65	1049
		% dentro de ESTUDIOS Estudios de la persona entrevistada	55,4%	53,3%	51,7%	38,3%	48,9%	14,5%	43,1%
	4 Trabajadores no cualificados	Recuento	59	174	199	51	72	13	568
		% dentro de ESTUDIOS Estudios de la persona entrevistada	42,4%	35,0%	33,1%	15,3%	17,3%	2,9%	23,3%
Total		Recuento	139	497	602	334	415	448	2435
		% dentro de ESTUDIOS Estudios de la persona entrevistada	100,0%	100,0%	100,0%	100,0%	100%	100,0%	100,0%

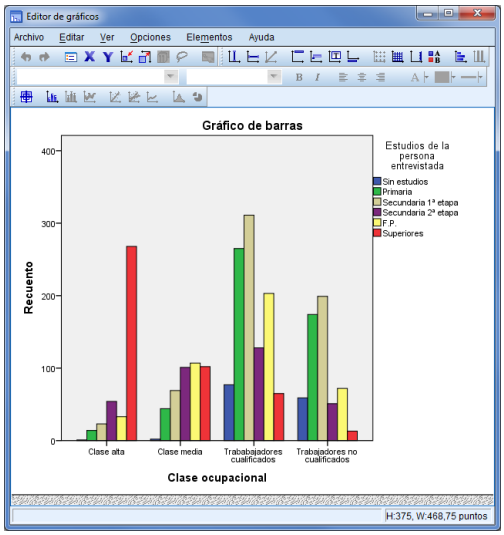
Si nos fijamos en los porcentajes de la tabla observamos cómo a medida que el nivel de estudios es más alto va aumentando el porcentaje de personas con mayor nivel ocupacional. Así por ejemplo, entre los que tienes estudios superiores casi el 60%

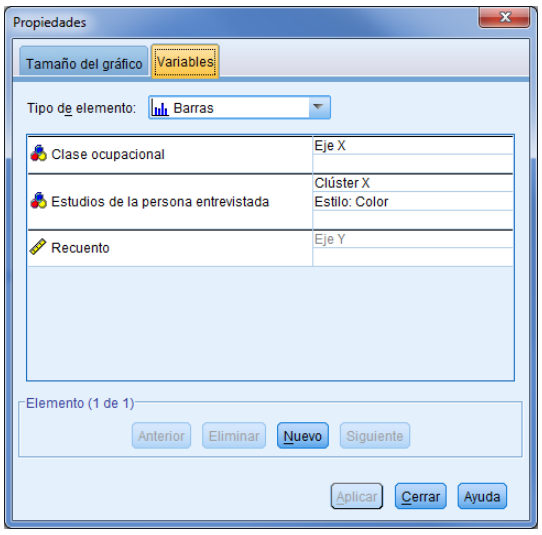
puede alcanzar la clase ocupacional alta, y a medida que baja el nivel de estudios se va reduciendo hasta la situación donde, si no se tienen estudios, no es solo poco probable sino imposible alcanzar dicha categoría profesional. Semejante lectura pero en sentido inverso podemos extraer al considerar en nivel inferior de los trabajadores no cualificados, donde observamos en particular que si no se tienen estudios es el nivel ocupacional más probable que se puede alcanzar.

En la tabla observamos también el particular comportamiento del nivel de FP que se corresponde más con la posición 4 en el orden de niveles educativos que con la posición o valor 5 que tiene: las personas con la segunda etapa de secundaria obtienen mejores ocupaciones que las personas con formación profesional.

En un gráfico podemos observar visualmente esta tendencia de asociación positiva entre estudios y ocupación. Se presentan dos gráficos, el primero es de barras agrupadas, sin apilar, con el recuento y con la disposición de las variables que proporciona originalmente el resultado de la ejecución del procedimiento. El segundo es de barras apiladas al 100% y se ha intercambiado la posición de las variables: la variable dependiente Ocupación va en la leyenda y la independiente ESTUDIOS en el eje de categorías. Para obtener este resultado es necesario editar el gráfico (doble clic) y operar los siguientes cambios:

- En la ventana de propiedades pulsamos sobre la pestaña de Variables y efectuamos los cambios siguientes: cambiamos de la variable ESTUDIOS la especificación Clúster X por Eje X, y pulsamos sobre el botón Aplicar.
- A continuación de la variable Ocupación cambiamos la especificación Clúster X por Pila y de nuevo pulsamos sobre el botón Aplicar.
- Finalmente desde la barra de botones pulsamos sobre el botón  que cambia la escala al 100%.[41]



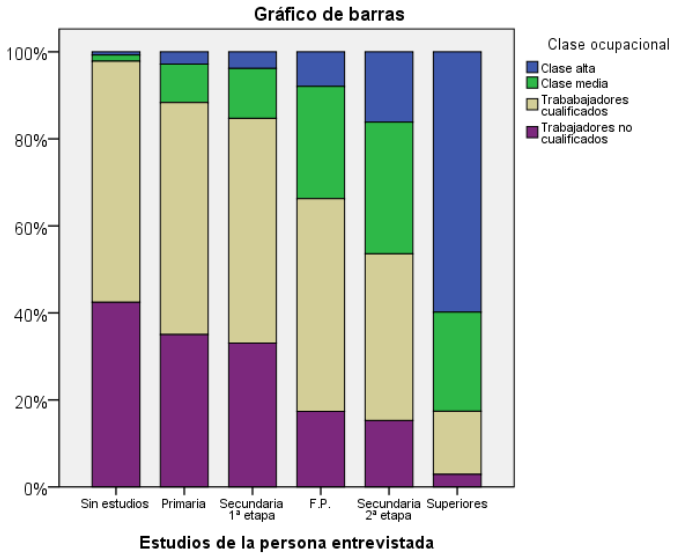


- Podemos modificar otros aspectos de la presentación. Por ejemplo podemos clicar sobre las barras del gráfico y elegir la pestaña Opciones de las barras de la ventana de propiedades para hacer que la anchura sea 70% y así espaciarlas entre ellas.

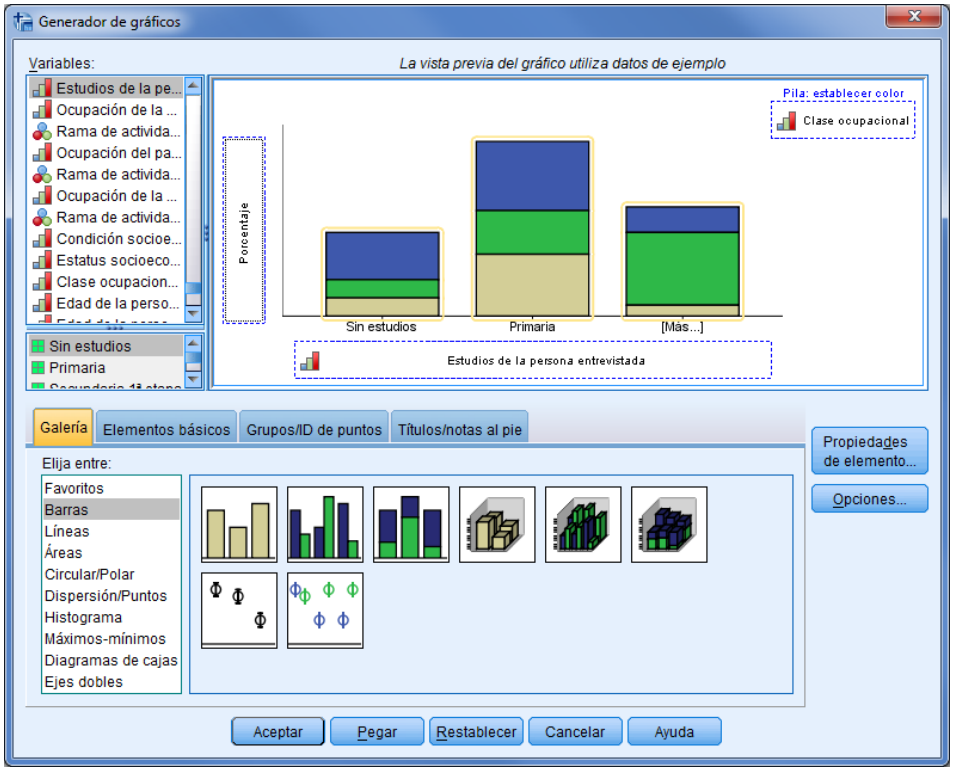
[41] Adicionalmente, si no la cambia automáticamente, hay que cambiar la etiqueta Recuento por Porcentaje o bien suprimirla como en este caso.

También cambiar el número de decimales a 0 desde la pestaña Formato de numeración o el tamaño de letra. Por último hemo reordenado las categoría de la variable ESTUDIOS desde la pestaña Categorías para colocar FP antes de Secundaria 2ª etapa.

El resultado de estos cambios es el gráfico siguiente:

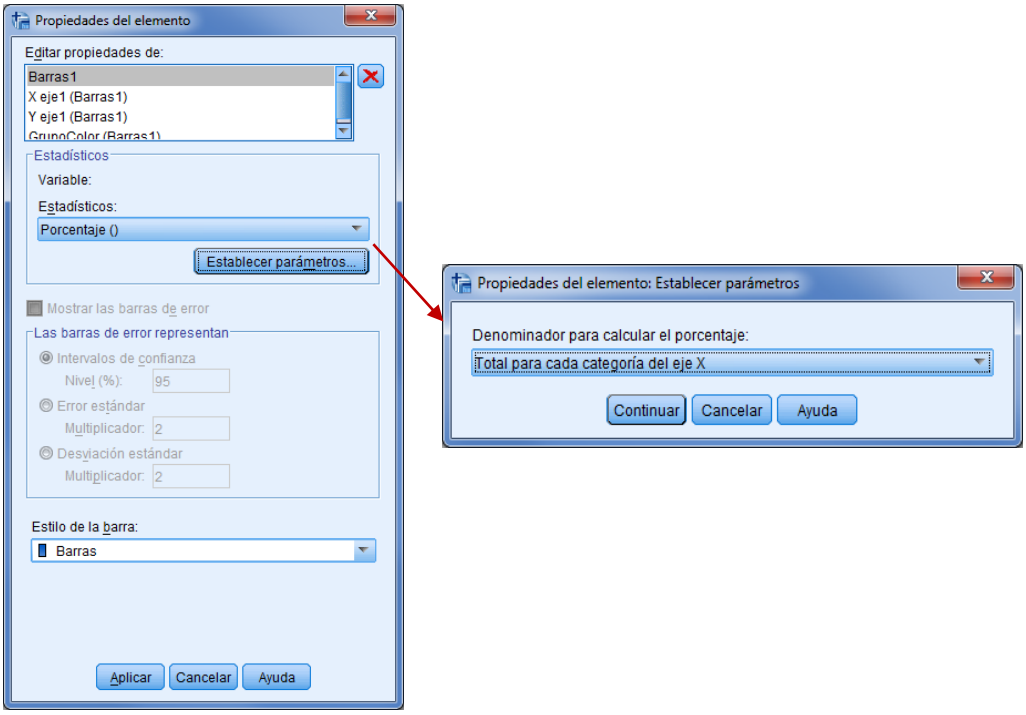


Otra forma de obtener el mismo gráfico es utilizar el menú Gráficos / Generador de Gráficos. Después de tener en cuenta la advertencia (si el nivel de medida de la variable no es el correcto no se podrá utilizar en determinados tipos de gráficos) nos aparecerá un cuadro de diálogo como el siguiente, y escogeremos un gráfico de barras, de tipo Barras apilado:

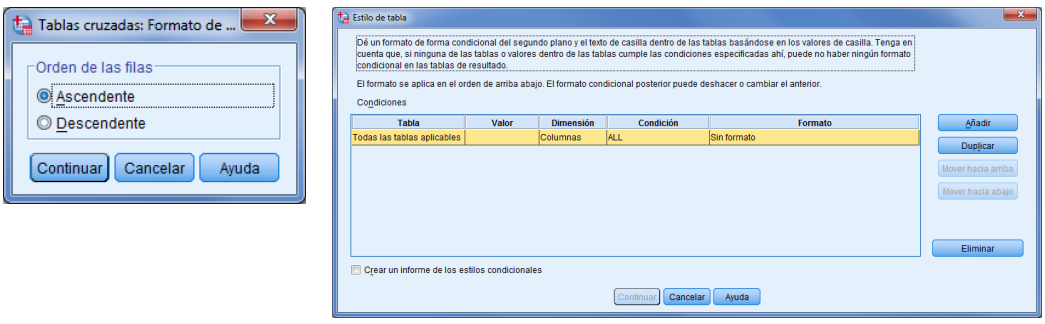


Se coloca en el eje de categorías (Eje X) la variable independiente (ESTUDIOS) y las pilas vendrán definidas por la variable dependiente (Ocupación).

Cuando lo hacemos el Eje Y recoge el recuento de casos. Para obtener los porcentajes cambiaremos una opción a través de la ventana de propiedades. En el apartado Estadístico seleccionaremos Porcentaje (), y además seleccionaremos tras pulsar sobre Establecer parámetros con Total de cada categoría del eje X, lo que supone reproducir los porcentajes por columna de la tabla de contingencia. Tras hacer clic sobre Continuar hay que hacerlo igualmente sobre Aplicar, y finalmente sobre Aceptar.



Los botones de Opciones y Estilo de tablas permite configurar aspectos adicionales de presentación de las tablas. En primer casos simplemente si las categorías se presentan de forma ascendente o descendente, y en el segundo se especifican las condiciones para cambiar automáticamente las propiedades de las tablas, por ejemplo, que aparezca en rojo no hay significación y en verde si sí la hay.



► **Ejercicio 11.**
Con la matriz de datos CIS3041+.sav analizar la relación entre parejas de variables cualitativas. Por ejemplo se puede analizar el comportamiento electoral según diversas variables independientes como la edad, el sexo, la ocupación, el lugar de residencia, etc. En relación a cada tabla:
- Observa los totales marginales.
- Calcula los porcentajes marginales.
- Calcula los porcentajes condicionales (fila y columna).
- Comenta la distribución condicional que hay que interpretar en cada tabla según la definición de la variable dependiente y la independiente para determinar la existencia de asociación entre las variables.
- Crea un gráfico de barras apiladas al 100%.

6.1.2. *Introducción de datos de una tabla de contingencia*

Si queremos trabajar con una tabla de contingencia que aparezca publicada en algún medio y reproducir los porcentajes y las gráficas, es posible hacerlo a partir de la introducción de las frecuencias de la tabla y la ponderación los datos lo que nos permite trabajar como si dispusiéramos de la información. El procedimiento de ponderación lo vimos en el capítulo III.2 y lo aplicamos entonces al caso de una tabla de frecuencias de una sola variable. Ahora si disponemos de una tabla de doble entrada como la siguiente será necesario tener en cuenta la combinación de valores de las variables. Hemos consultado en la web del INE la Encuesta de Población Activa (EPA) correspondiente al cuarto trimestre de 2014 para ver la tabla de distribución de la población desocupada según grupos de edad y sexo, con estos resultados:

Tabla III.6.31. Desocupados según Edad y Sexo en 2014
Datos absolutos (en miles)

	Hombres	Mujeres
De 16 a 19 años	85,9	82,2
De 20 a 24 años	341,4	304,2
De 25 a 29 años	357,2	324,6
De 30 a 34 años	350,4	348,3
De 35 a 39 años	359,2	366,5
De 40 a 44 años	343,1	376,9
De 45 a 49 años	329,9	320,6
De 50 a 54 años	291,7	261,5
De 55 a 59 años	263,4	177,3
De 60 a 64 años	96,1	69,6
De 65 y más años	5,3	2,3

Fuente: EPA 4º trimestre 2014

La introducimos en el SPSS en una nueva base de datos (Archivo / Nuevo / Datos) de la siguiente forma:

	Edad	Sexo	Frecuencia
1	1	1	85,9
2	2	1	341,4
3	3	1	357,2
4	4	1	350,4
5	5	1	359,2
6	6	1	343,1
7	7	1	329,9
8	8	1	291,7
9	9	1	263,4
10	10	1	96,1
11	11	1	5,3
12	1	2	82,2
13	2	2	304,2
14	3	2	324,6
15	4	2	348,3
16	5	2	366,5
17	6	2	376,9
18	7	2	320,6
19	8	2	261,5
20	9	2	177,3
21	10	2	69,6
22	11	2	2,3

Hemos creado tantas filas, 22, como casillas tiene la distribución conjunta, es decir, 11 categorías de edad × 2 categorías de sexo = 22 categorías de la distribución conjunta.

Se trata de una matriz de datos particular. En efecto, disponemos de una matriz de 22 filas y 3 columnas que habitualmente se entendería que hacen referencia a 22 casos o individuos y 3 variables. De hecho así es. Pero en este caso utilizaremos las 22 filas de la matriz de forma instrumental para identificar las 22 casillas de la tabla de contingencia que relaciona la edad con el sexo. Para la identificación de las casillas utilizamos las dos primeras columnas de la matriz, las que contienen los valores de las dos variables que relacionamos, mientras que la tercera columna contiene la frecuencia de cada casilla de la tabla de contingencia. Las frecuencias tienen decimales, dado que la unidad son miles de desocupados. Así tenemos que la primera fila identificada en la primera casilla de la tabla, la que tiene la combinación (1,1), es decir, se refiere a 16-19 y Hombres, con un total de 85,9 individuos (en miles, es decir, 85.900). Esta lógica se extiende a las 21 filas restantes hasta completar todas las casillas de la tabla.

Una vez introducidos los datos asignamos etiquetas a los valores numéricos con los que hemos codificado cada variable, asignamos el formato y extraemos la tabla de contingencia. Lo que se presenta a continuación (Gráfico III.6.7) es el programa de instrucciones del SPSS para reproducir el análisis de las dos variables del ejemplo mediante el lenguaje de comandos[42]. El programa incluye el comando **CROSSTABS** con la especificación de la tabla de contingencia bivariable, con las opciones de contenido de las casillas que habíamos detallado al comentar el menú del procedimiento. Pero además se han incluido una serie de instrucciones previas destinadas a identificar los datos que son objeto de tratamiento en el procedimiento.

Se trata de un conjunto de instrucciones de la programación del SPSS destinadas a la generación e identificación de datos que nosotros utilizaremos de forma instrumental, a modo de "truco", para poder tratar los datos de una tabla de contingencia sin necesidad de disponer de la matriz de datos original. El comando **DATA LIST**, en

[42] Se puede encontrar en la página web con el nombre ATC-Desempleo.sps.

formato libre (opción FREE), especifica el nombre de las variables de la matriz de datos que se incluye entre las palabras clave **BEGIN DATA** / **END DATA**. A continuación se detallan las etiquetas de las variables (instrucción **VARIABLE LABELS**) y los valores (instrucción **VALUE LABELS**) y su formato (instrucción **FORMATS**).

Gráfico III.6.7 Programa de instrucciones de SPSS para identificar la tabla de contingencia sobre desempleo según edad y sexo

```
* Identificación de los datos de la tabla.
DATA LIST FREE / Edad Sexo Frecuencia.
BEGIN DATA.
1 1 85,9
2 1 341,4
3 1 357,2
4 1 350,4
5 1 359,2
6 1 343,1
7 1 329,9
8 1 291,7
9 1 263,4
10 1 96,1
11 1 5,3
1 2 82,2
2 2 304,2
3 2 324,6
4 2 348,3
5 2 366,5
6 2 376,9
7 2 320,6
8 2 261,5
9 2 177,3
10 2 69,6
11 2 2,3
END DATA.

VARIABLE LABELS  Edad 'Edad de la persona parada'
                 Sexo 'Sexo de la persona parada'.
VALUE LABELS Edad  1 'De 16 a 19 años'    2 'De 20 a 24 años'    3 'De 25 a 29 años'
                   4 'De 30 a 34 años'    5 'De 35 a 39 años'    6 'De 40 a 44 años'
                   7 'De 45 a 49 años'    8 'De 50 a 54 años'    9 'De 55 a 59 años'
                  10 'De 60 a 64 años'   11 'De 65 y más años'
  /Sexo 1 'Hombres' 2 'Mujeres'.
FORMATS Edad (F2.0) Sexo (F1.0).
WEIGHT BY Frecuencia.
FREQUENCIES ALL.
CROSSTABS Edad  BY Sexo
  /CELLS= COUNT COLUMN
  /BARCHART.
```

Una vez identificada la tabla de esta forma lo que tenemos, de hecho, son 22 individuos que se caracterizan por el perfil de cada casilla de la tabla. El paso siguiente consistirá en indicarle al SPSS que no contabilice cada uno de estos 22 individuos como uno solo, sino que cuente tantos individuos como especifique la tercera variable de la matriz, es decir, la variable Frecuencia que detalla la frecuencia de cada casilla. Esto se hace mediante el comando **WEIGHT** (Ponderar casos en el menú), el cual transforma el peso original que cada individuo tiene, de una unidad, en el peso que se especifica en una variable, en este caso Frecuencia. Por tanto, el primer individuo que identifica a la primera casilla (1,1) de la tabla pasa de valer 1 a valer 85,9, tendremos por tanto 85,9 individuos -de hecho en miles- con las mismas características en las dos variables. El

segundo individuo que identifica a la segunda casilla (1,2) de la tabla pasa de valer 1 a valer 341,4, y así sucesivamente. De esta forma pasamos, una vez hecha la ponderación, de tener 22 individuos a tener los 5.457 que es el total de la tabla que utilizamos de ejemplo.

De esta forma podemos introducir cualquier tabla de contingencia que podamos ver publicada, tan sólo hay que saber la frecuencia absoluta de cada casilla de la tabla para poder tratarla mediante este procedimiento. Estas mismas instrucciones se podrían ejecutar mediante los menús. Hay primero que introducir los datos en la ventana del editor de datos del SPSS, a continuación se identifican las variables, sus valores y el formato, y por último, se ejecuta la ponderación a través de: Datos / Ponderar casos. En el cuadro de diálogo sólo hay que marcar primero Ponderar casos mediante y a continuación traspasar la variable de ponderación, en este caso la variable Frecuencia. Una vez ejecutado el comando tendremos los individuos ponderados y en la barra de estado de la ventana del editor de datos aparecerá la indicación Ponderación activada. Solo queda pedir la tabla de contingencia.

► **Ejercicio 12.**
Introduce los datos de la tabla de contingencia siguiente y analiza la relación entre las variables:

		P31 Sexo		
		Hombre	Mujer	Total
P1 Valoración de la situación económica general de España	Buena	19	13	32
	Regular	203	196	399
	Mala	510	480	990
	Muy mala	474	576	1050
Total		1206	1265	2471

6.1.3. *Análisis inferencial con dos variables*

Completaremos el análisis de una tabla de contingencia con el estudio inferencial de la relación que nos permite establecer la significación estadística de la relación entre las variables y poder inferir el resultado al conjunto de la población.

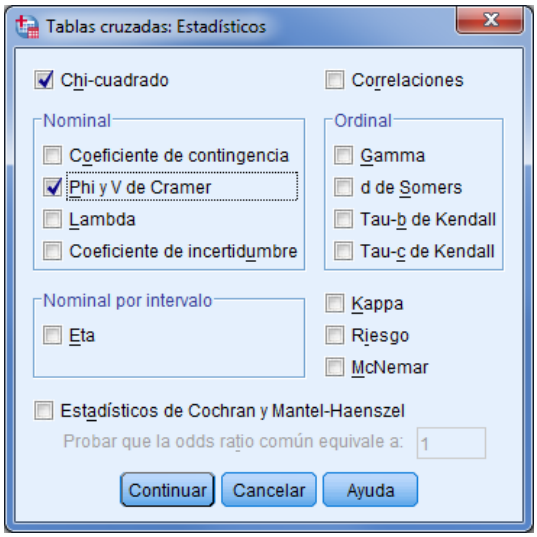
Después de analizar descriptivamente la tabla con los porcentajes y el gráfico, y llegar a la conclusión de que a medida que aumentan los estudios aumenta la ocupación, nos formulamos ahora dos preguntas para completar nuestro estudio: estas diferencias observadas son suficientemente significativas desde un punto de vista estadístico? Si lo son, ¿cuál es el grado de relación entre estas variables?

Para responder rigurosamente y con objetividad desde un punto de vista estadístico debemos realizar la prueba estadística del test de independencia de chi-cuadrado, donde contrastaremos dos hipótesis:

- la hipótesis nula: consiste en asumir que las dos variables son independientes y que no existe ninguna relación de asociación entre ellas,
- la hipótesis alternativa: consiste en aceptar que sí existe algún tipo de relación de asociación o dependencia pues la hipótesis alternativa no es falsable.

El objetivo es saber si, con un cierto nivel de confianza, tenemos evidencias suficientes como para rechazar la hipótesis nula y concluir que las diferencias porcentuales son significativas. Una vez constatada la significación de la relación tiene sentido contestar a la segunda pregunta, se trata de calcular una medida de la intensidad de la relación. Consideraremos la V de Cramer, la cual incorpora también una prueba de significación, de hecho, la misma que la del chi-cuadrado ya que se trata de una medida basada en aquel estadístico.

Vamos a ver cómo realizar este test con el SPSS. En el procedimiento de Tablas cruzadas pulsamos sobre el botón de Estadísticos y marcamos las opciones Chi-cuadrado y Phi y V de Cramer:



Los resultados son los siguientes:

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (2 caras)
Chi-cuadrado de Pearson	1075,801[a]	15	,000
Razón de verosimilitud	1010,809	15	,000
Asociación lineal por lineal	687,987	1	,000
N de casos válidos	2435		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 22,43.

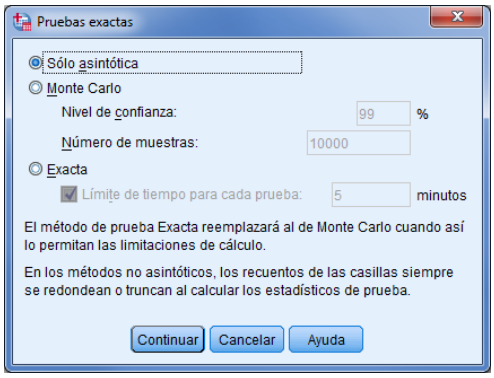
Medidas simétricas

		Valor	Aprox. Sig.
Nominal por Nominal	Phi	,665	,000
	V de Cramer	,384	,000
N de casos válidos		2435	

Como la significación es de 0,000 <0,05, podemos concluir que hay relación entre las variables, que las diferencias porcentuales son significativas con un nivel de confianza del 95% (con un 5% de riesgo). Esta afirmación se mantiene siempre y cuando las condiciones para interpretar el test se den: la frecuencia mínima esperada en cada casilla sea 1 como mínimo y el porcentaje de casillas con una frecuencia esperada inferior a 5

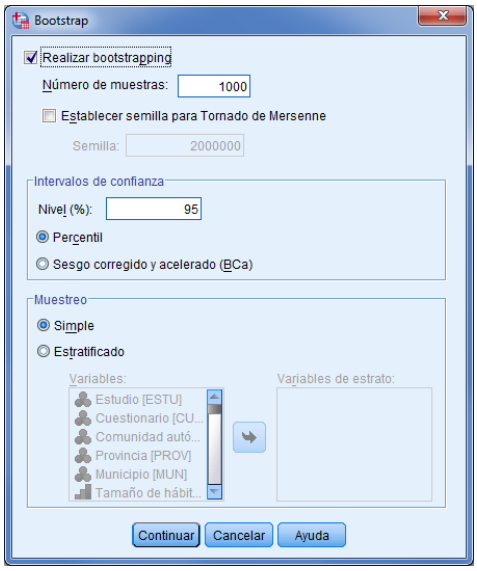
sea inferior al 20% como se puede comprobar en la nota a pie de la tabla de la prueba de chi-cuadrado. Finalmente, el grado de esta relación observada es de 0,384, un valor intermedio, importante y significativo (0,000), pero que no alcanza un valor elevado como 0,6. Es decir, el grado en que se determina el nivel ocupacional por el nivel educativo existe pero es limitado.

Además de la V de Cramer se pueden elegir otros estadísticos que tienen una función similar de evaluación de la intensidad de la relación y varían según la escala de medición de las variables. También en el cuadro de diálogo inicial disponemos de una opción adicional: las pruebas exactas[43]. Si clicamos sobre Exacta se visualiza un cuadro diálogo como el siguiente:



donde podemos elegir dos métodos: el exacto y el de Monte Carlo, destinados a obtener resultados exactos cuando los datos no cumplen alguno de los supuestos: son muestras pequeñas, con pocos efectivos en las casillas o presentan dispersiones importantes. Por defecto se realizan las pruebas asintóticas.

Con la opción Bootstrap se puede emplear un método para obtener estimaciones robustas de errores típicos e intervalos de confianza para estimaciones de los estadísticos a partir de reiteraciones en numerosas muestras (por defecto 1000 y con muestreo aleatorio simple).



[43] Las pruebas exactas aparecen si se tiene instalado un módulo específico del SPSS.

► **Ejercicio 13.**
Con la matriz de datos CIS3041+.sav y siguiendo el ejemplo del Ejercicio 11 completar el análisis con los aspectos siguientes:
- Interpreta si hay relación entre las variables según el estadístico Chi cuadrado.
- Interpreta la fuerza de la relación (si es que la hay) observando el estadístico V de Cramer.

6.1.4. *Análisis de tablas de contingencia multidimensionales*

Para analizar la relación entre variables cualitativas en tablas multidimensionales consideraremos algunos ejemplos de asociaciones trivariables. En primer lugar presentamos el ejemplo de la relación entre las variables abandono de los estudios universitarios (ABA), la actividad laboral (ACT) y el horario (HOR), a partir de un total de 474 casos[44]. La tabla de contingencia con las frecuencias absolutas es la siguiente:

Tabla III.6.32. El abandono de los estudios universitarios según la actividad laboral y el horario

Recuento

		HOR Horario								
		1 Mañana			2 Tarde			3 Noche		
		ACT Actividad laboral			**ACT** Actividad laboral			**ACT** Actividad laboral		
		1 No	2 Sí	Total	1 No	2 Sí	Total	1 No	2 Sí	Total
ABA Abandono de los estudios universitarios	1 No	100	17	117	70	50	120	75	55	130
	2 Sí	10	7	17	25	20	45	10	35	45
Total		110	24	134	95	70	165	85	90	175

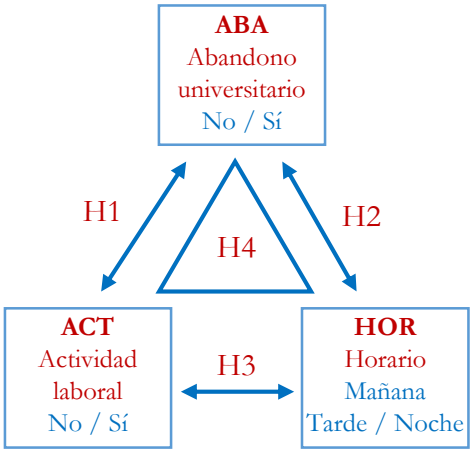
Fuente: Latiesa (1991)

Con estas variables podríamos plantear un modelo de análisis con diferentes hipótesis. Dos de estas hipótesis nos conducen inicialmente a explicar el abandono como variable dependiente en función de la actividad laboral (Hipótesis 1: trabajar penaliza con mayor abandono) y en función del horario de clases (Hipótesis 2: el estudiantado de la tarde y de la noche abandona más). Adicionalmente podemos plantear una tercera relación entre las variables independientes (Hipótesis 3: el estudiantado que tiene una actividad laboral tiende a matricularse sobre todo en los grupos de tarde y de noche). Ahora bien, cabe preguntarse hasta qué punto la razón del abandono tiene que ver realmente con el horario de clases, de hecho podemos pensar que se está dando un mecanismo secuencial donde los trabajadores tienden a matricularse por la tarde-noche y en consecuencia estos grupos tienen una mayor tasa de abandono, por tanto, que la verdadera razón es la actividad laboral y no el grupo de clase. En este sentido una posible y aparente relación entre abandono y horario deberá desaparecer al controlar por la actividad laboral, poniendo de manifiesto una relación espúrea. En este sentido nuestra Hipótesis 4 afirmaría que no existe una relación de interacción entre las variables y que el modelo que cabe esperar es aquel donde el abandono viene explicado

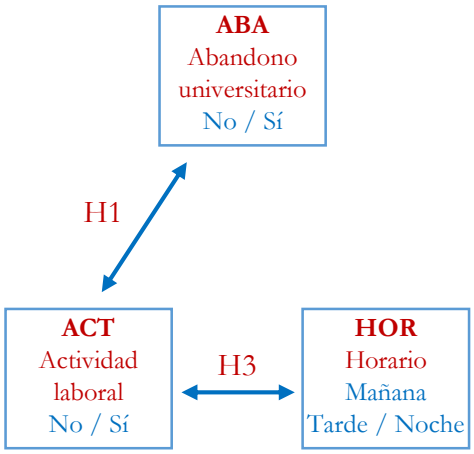
[44] Información extraída de un estudio sobre los alumnos de la Facultad de Ciencias Políticas y Sociología de la Universidad Complutense de Madrid. El ejemplo está publicado en la revista *Papers* por Latiesa (1991).

solamente por la actividad laboral (Hipótesis 1) y el horario de clase del estudiante es diferente según la actividad laboral (Hipótesis 3). Si por el contrario mantuviéramos la hipótesis de la interacción entre las tres variables deberíamos observar que el abandono no solamente varía en función de la actividad laboral sino que es mayor entre los grupos de tarde y de noche que en los de mañana; podríamos pensar por ejemplo que el estudiantado de la mañana tiene un perfil de estudiante que trabaja y el estudiantado de la noche un perfil de trabajador que estudia y por tanto entre esos últimos la penalización de trabajar es más acentuada.

Planteamos así dos posibles alternativas de modelo de análisis con hipótesis distintas que se trata de verificar seguidamente. La representación gráfica en ambos modelos sería la siguiente:



Modelo 1: Interacción

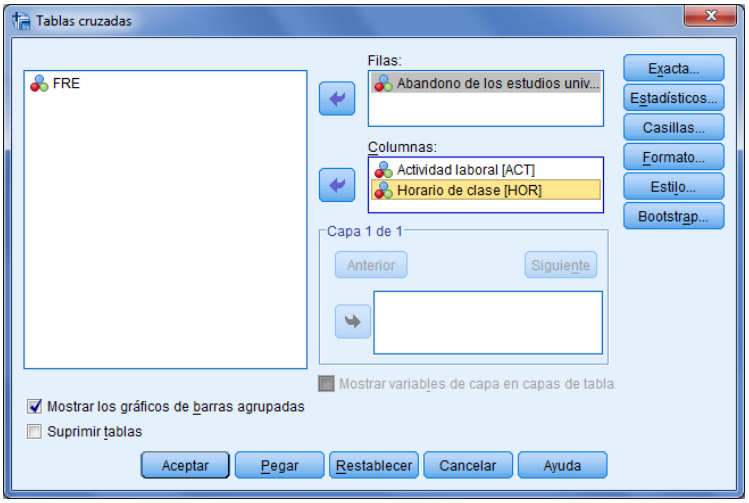


Modelo 2: Doble asociación

Veamos cómo analizar estos modelos con el software estadístico.

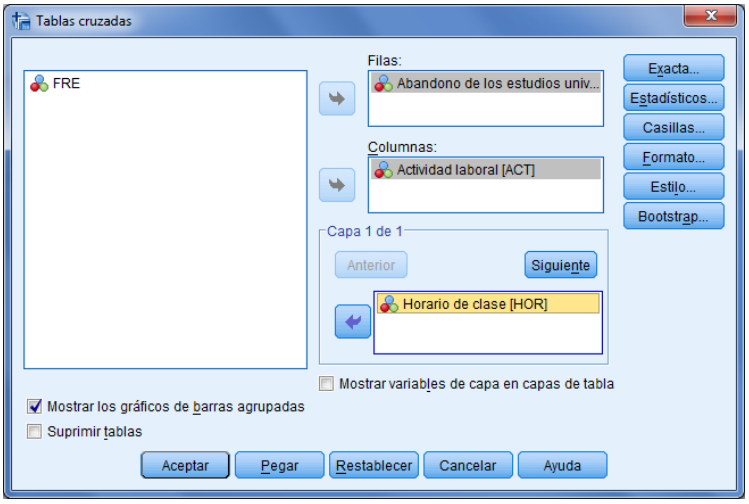
A través del procedimiento Tablas cruzadas del menú cuando consideramos sólo relaciones bivariables, las tablas de contingencia que se obtienen relacionan todas las combinaciones de las variables incluidas en los recuadros de Filas y Columnas. Si por ejemplo colocamos la variable A en el recuadro de Filas y las variables B y C en el cuadro Columnas, obtendremos dos tablas de contingencia, las que relacionan A×B y A×C. Y si por ejemplo colocamos las variables A y Z en el cuadro Filas y las variables B y C en el cuadro Columnas, obtendremos cuatro tablas de contingencia, las que relacionan A×B, A×C, Z×B y Z×C.

En el caso del ejemplo sobre el abandono de los estudios podemos considerar la petición de dos tablas de contingencia que relacionen ABA×ACT y ABA×HOR. En este caso, y por nuestra convención, estamos considerando la variable ABA como dependiente, y la colocamos en filas, y las variables ACT y HOR como independientes, y las colocamos en columnas. El cuadro de diálogo correspondiente a este análisis sería el siguiente:



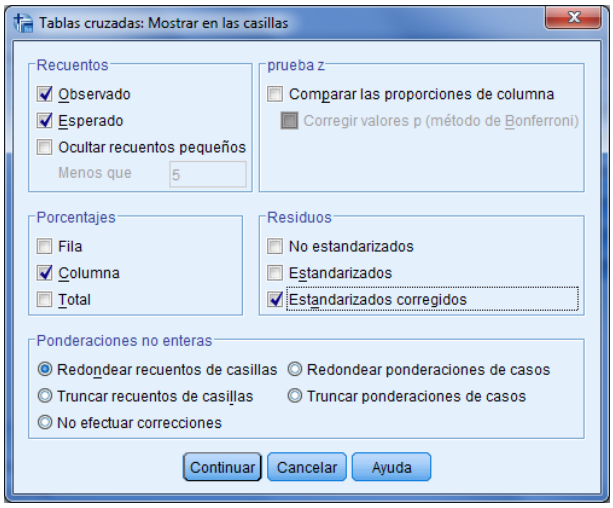
Si quisiéramos obtener además la tabla de contingencia entre ACT y HOR entonces deberíamos ejecutar una segunda vez el procedimiento dado que no hay forma de especificar simultáneamente este cruce con los anteriores. Sin embargo, con las instrucciones introducidas con el lenguaje de comandos sí sería posible.

Si consideramos ahora relaciones entre tres o más variables entonces debemos utilizar el recuadro de Capa para trasladar las variables que definen la tercera y sucesivas dimensiones. Así, por ejemplo, si colocamos las variable A y Z en el cuadro Filas, la variable B en el cuadro Columnas, y la variable C en el cuadro Capa 1 de 1, obtendremos dos tablas de contingencia, las que relacionan A×B×C y Z×B×C. Si además deseáramos reproducir las mismas tablas con una cuarta variable D deberíamos colocar de nuevo las variable A y Z en el recuadro Filas, la variable B en el cuadro Columnas, la variable C en el cuadro Capa 1 de 1, la que se convertirá más tarde en la Capa 1 de 2, y la variable D en el cuadro Capa 2 de 2 , obtendremos así dos tablas de contingencia, las que relacionan A×B×C×D y Z×B×C×D. Recordemos que la obtención de tablas de contingencia de tres (o más dimensiones) significa reproducir tantas tablas y estadísticos bivariables como valores (o combinaciones de valores) tiene la tercera variable (o combinaciones entre valores de las variables de tercera y sucesivas dimensiones). A continuación se ilustra el caso de la especificación de la tabla tridimensional que cruza ABA×ACT×HOR:



En estos análisis solicitamos los diagramas de barras de la variable especificada en las columnas (con la asignación de un color a cada valor) que se representa para cada valor de la variable especificada en filas. Si hay una tercera variable se reproduce este gráfico para cada valor de la tercera variable, o para combinaciones de valores en el caso de más de tres dimensiones. Los gráficos presentan las frecuencias absolutas, tal y como vimos en el caso bivariable en un apartado anterior. Nos interesa por tanto disponer de los valores porcentuales, por lo que es preciso cambiar la escala en porcentajes y considerarlos como barras apiladas (ver apartado 6.1.1).

Para completar las especificaciones del procedimiento de análisis detallaremos las siguientes opciones de los apartados de Casillas y Estadísticos. Consideremos en primer lugar la información que queremos mostrar en las casillas de la tabla de contingencia a partir de su cuadro de diálogo:

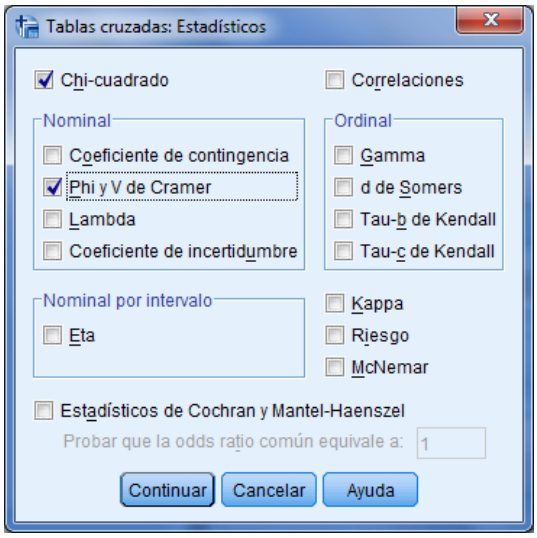


Por defecto, si no hacemos ninguna especificación, el SPSS nos sacará las frecuencias absolutas observadas. En este caso hemos pedido también las frecuencias esperadas bajo la hipótesis de independencia entre las variables, los porcentajes por columna ya que consideraremos, por convención, a la variable colocada en filas como la variable dependiente y la variable de las columnas como la variable independiente, y también el valor de las frecuencias esperadas y los residuos tipificados corregidos.

Por su parte también pediremos los estadísticos que nos determinan de un lado la existencia de asociación y de la otra la intensidad de ésta. En el primer caso marcaremos la opción Chi-cuadrado que nos proporciona la prueba de chi-cuadrado de Pearson junto con el chi-cuadrado de la razón de verosimilitud, que es una reformulación del de Pearson con resultados similares[45]. Estas pruebas se aplican a tablas con cualquier número de filas o columnas, pero también se calculan dos estadísticos más destinados a establecer la existencia de asociación en el caso particular de que disponemos de una mesa de 2×2: la prueba exacta Fisher, que se utiliza cuando una casilla tiene una frecuencia esperada menor que 5, y el chi-cuadrado corregido de Yates para el resto de tablas 2×2. Por último, si ambas variables de la tabla fueran cuantitativas entonces se interpreta la prueba de asociación lineal por lineal.

[45] Sobre este estadístico, llamado también L^2 o G^2, volveremos en el capítulo siguiente pues con él estableceremos la significatividad de los modelos log-lineales.

Si disponemos de variables ordinales o de intervalo tanto en las filas como en las columnas podemos hacer uso de la opción de Correlaciones. En este caso se calculan la rho de Spearman que es una medida de asociación para órdenes de rangos, y el coeficiente de correlación de Pearson que mide la asociación lineal para variables cuantitativas.



Para determinar el grado de asociación entre las variables disponemos de diferentes medidas en función del nivel de medida de éstas, las cuales se caracterizan además por ser o bien direccionales o simétricas. En el cuadro de diálogo que aparece a continuación se pueden ver las diferentes medidas en función de si son nominales, ordinales, nominal/intervalo, más las medidas de Kappa de Cohen, la prueba no paramétrica de McNemar y los estadísticos de Cochran y Mantel-Haenszel. Como en casos anteriores nos limitaremos a considerar la V de Cramer.

La sintaxis del comando **CROSSTABS** que procesa las tablas de contingencia se presenta en el Gráfico III.6.8, con dos posibilidades de ejecución: general mode es el procedimiento habitual en el que sólo es necesario especificar las variables y las opciones que se quieran utilizar en cada análisis; integer mode es el procedimiento que requiere la especificación tanto de las variables como los valores mínimo y máximo de las mismas, lo que nos permitirá considerar la distribución de los valores perdidos (*missing values*) en la tabla de contingencia, sin formar parte de los cálculos, simplemente se verán a título informativo. Esta opción es posible mediante la especificación: /MISSING=REPORT.

A continuación se presenta el programa de instrucciones del SPSS para reproducir el análisis de las tres variables del ejemplo mediante el lenguaje de comandos[46]. El programa de instrucciones del SPSS incluye el comando **CROSSTABS** con la especificación de tres tablas de contingencia bivariables y una trivariable, con las opciones de estadísticos y de contenido de las casillas que habíamos detallado al comentar el menú del procedimiento. Pero además se han incluido una serie de instrucciones previas destinadas a identificar los datos que son objeto de tratamiento en el procedimiento. Se trata de las instrucciones destinadas a la generación e

[46] Se puede encontrar en la página web con el nombre ATC-Abandono.sps.

identificación de datos que nosotros utilizaremos de forma instrumental como vimos en un ejemplo anterior. En este caso se involucran tres variables siguiendo la misma lógica.

Gráfico III.6.8 Esquema de la sintaxis del comando CROSSTABS

CROSSTABS

CROSSTABS is available in the Statistics Base option.

General mode:

```
CROSSTABS [TABLES=]varlist BY varlist [BY...] [/varlist...]

 [/MISSING={TABLE**}]
           {INCLUDE}

 [/WRITE[={NONE**}]]
          {CELLS }

 [/HIDESMALLCOUNTS [COUNT = {5      }] ]
                            {integer}

 [/SHOWDIM = integer]

 [/CELLS = [PROP] [BPROP]
```

Integer mode :

```
CROSSTABS VARIABLES=varlist(min,max) [varlist...]

 /TABLES=varlist BY varlist [BY...] [/varlist...]

 [/MISSING={TABLE**}]
           {INCLUDE}
           {REPORT }

 [/WRITE[={NONE**}]]
          {CELLS }
          {ALL   }
```

Both modes:

```
 [/FORMAT= {AVALUE**}  {TABLES**}]
           {DVALUE  }  {NOTABLES}

 [/COUNT = [{ASIS}] [{ROUND   }]
            {CASE}   {TRUNCATE}
            {CELL}

 [/CELLS=[{COUNT**}]  [ROW   ]  [EXPECTED]  [SRESID ]]
         {NONE   }   [COLUMN]  [RESID   ]  [ASRESID]
                     [TOTAL ]              [ALL    ]

 [/STATISTICS=[CHISQ]  [LAMBDA]  [BTAU ] [GAMMA  ]  [ETA    ]]
              [PHI  ]  [UC    ]  [CTAU ] [D      ]  [CORR   ]
              [CC   ]  [RISK  ]  [KAPPA] [MCNEMAR]  [CMH(1*)]
              [ALL  ]  [NONE  ]

 [/METHOD={MC [CIN({99.0 })] [SAMPLES({10000})]]††
                  {value}            {value}
          {EXACT [TIMER({5    })]             }
                        {value}

 [/BARCHART]
```

**Default if the subcommand is omitted.

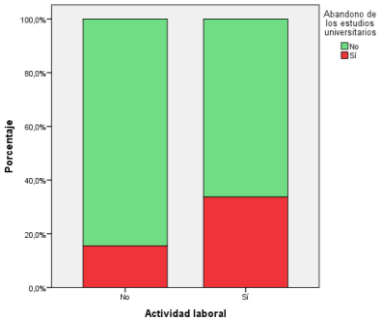
†† The METHOD subcommand is available only if the Exact Tests option is installed (available only on Windows operating systems).

```
* Generación e identificación de los datos de la tabla de contingencia.
DATA LIST FREE/ ABA ACT HOR FRE.
BEGIN DATA
1 1 1 100
1 1 2 70
1 1 3 75
1 2 1 17
1 2 2 50
1 2 3 55
2 1 1 10
2 1 2 25
2 1 3 10
2 2 1 7
2 2 2 20
2 2 3 35
END DATA.
VARIABLE LABELS ABA 'Abandono de los estudios universitarios'
                ACT 'Actividad laboral'
                HOR 'Horario de clase'.
VALUE LABELS ABA ACT 1 'No' 2 'Sí'/
             HOR 1 'Mañana' 2 'Tarde' 3 'Noche'.
FORMATS ABA ACT HOR (F1.0).
WEIGHT BY FRE.
CROSSTABS ABA BY ACT BY HOR.
* Tablas de contingencia bivariables i trivariables.
CROSSTABS ABA BY ACT /ABA BY HOR /HOR BY ACT
 /ABA BY ACT BY HOR   /ABA BY HOR  BY ACT
 /CELLS=COUNT COLUMN EXPECTED ASRESID
 /STATISTICS=CHISQ PHI
 /BARCHART.
```

A continuación se presentan los diversos resultados de tablas y gráficos que se obtienen con el procedimiento de tablas de contingencia y los datos del ejemplo.

Tabla de contingencia ABA x ACT

			ACT Actividad laboral		
			1 No	2 Sí	Total
ABA Abandono de los estudios universitarios	1 No	Recuento	245	122	367
		Frecuencia esperada	224,5	142,5	367,0
		% dentro de ACT Actividad laboral	**84,5%**	**66,3%**	**77,4%**
		Residuos corregidos	4,6	-4,6	
	2 Sí	Recuento	45	62	107
		Frecuencia esperada	65,5	41,5	107,0
		% dentro de ACT Actividad laboral	**15,5%**	**33,7%**	**22,6%**
		Residuos corregidos	-4,6	4,6	
Total		Recuento	290	184	474
		Frecuencia esperada	290,0	184,0	474,0
		% dentro de ACT Actividad laboral	100,0%	100,0%	100%



En primer lugar constatamos que la tasa de abandono general de la Facultad es el 22,6% pero que afecta en mayor medida a los que trabajan (33,7%) que a los que no trabajan (15,5%), un diferencia del 18,2%. Este comportamiento diferenciado motiva que las frecuencias observadas y esperadas difieran generando un residuo estadísticamente significativo (±4,6 supera el valor ±1,96). Este resultado de diferencia de porcentajes y de residuos nos indica lo que la prueba estadística de chi-cuadrado evidencia: la existencia de una relación estadísticamente significativa que nos permite concluir que las variables están asociadas, con una intensidad del 0,212 según se obtiene con la V de Cramer. Se verifica pues la Hipótesis 1 en un análisis bivariable.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	21,284[b]	1	,000		
Corrección de continuidad [a]	20,257	1	,000		
Razón de verosimilitud	20,838	1	,000		
Estadístico exacto de Fisher				,000	,000
Asociación lineal por lineal	21,239	1	,000		
Prueba de McNemar				,000[c]	
N de casos válidos	474				

a. Calculado sólo para una tabla de 2x2.

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 41,54.

c. Distribución binomial utilizada

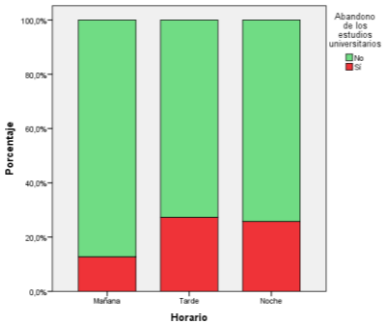
Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,212	,000
	V de Cramer	,212	,000
	Coeficiente de contingencia	,207	,000
N de casos válidos		474	

Analizamos a continuación la Hipótesis 2 en un análisis bidimensional. Planteábamos en ella que el abandono sería mayor entre el estudiantado de la tarde o de la noche. Los datos muestran que la tasa de abandono efectivamente aumenta en los grupos de clase de la tarde y la noche, frente a un porcentaje del 12,7% de la mañana, el de tarde tiene un valor del 27,3% y similar al de noche con un 25,7%. Estas diferencias en grupos de clase son claramente significativas para el horario de mañana, 12,7% difiere del comportamiento global de 22,6%. Pero en los grupos de tarde y noche, si bien sus tasas de abandono son superiores al promedio, las diferencias son reducidas en relación a la mañana y resultan localmente no significativas estadísticamente (los residuos corregidos son inferiores a 1,96). Por tanto, las casillas de tarde y noche no contribuyen a generar asociación y es el grupo de la mañana el que genera la fuente de asociación entre las variables.

Tabla de contingencia ABA x HOR

			HOR Horario			
			1 Mañana	2 Tarde	3 Noche	Total
ABA Abandono de los estudios universitarios	1 No	Recuento	117	120	130	367
		Frecuencia esperada	103,8	127,8	135,5	367,0
		% dentro de HOR Horario	**87,3%**	**72,7%**	**74,3%**	**77,4%**
		Residuos corregidos	3,2	-1,8	-1,3	
	2 Sí	Recuento	17	45	45	107
		Frecuencia esperada	30,2	37,2	39,5	107,0
		% dentro de HOR Horario	**12,7%**	**27,3%**	**25,7%**	**22,6%**
		Residuos corregidos	-3,2	1,8	1,3	
Total		Recuento	134	165	175	474
		Frecuencia esperada	134,0	165,0	175,0	474,0
		% dentro de HOR Horario	100,0%	100%	100%	100%



Si miramos el comportamiento global de la asociación a través del chi-cuadrado verificamos que se cumple la hipótesis alternativa, estableciendo un nivel de asociación de 0,149:

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)	Sig. exacta (bilateral)
Chi-cuadrado de Pearson	10,567[a]	2	,005	
Razón de verosimilitud	11,480	2	,003	
Asociación lineal por lineal	6,568	1	,010	
Prueba de McNemar				,[b]
N de casos válidos	474			

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 30,25.

b. Sólo se efectuará el cálculo para tablas de PxP, donde P debe ser mayor que 1.

Medidas simétricas

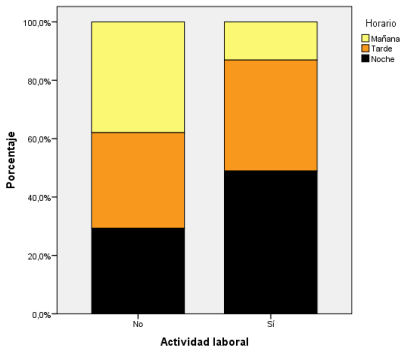
		Valor	Sig. aproximada
Nominal por nominal	Phi	,149	,005
	V de Cramer	,149	,005
	Coeficiente de contingencia	,148	,005
N de casos válidos		474	

Se comprueba también la Hipótesis 2 en un análisis bivariable.

Analizamos la tercer y última relación entre parejas de variables. En la Hipótesis 3 planteábamos que los que trabajan tienden a matricularse sobre todo por la tarde y noche.

Tabla de contingencia HOR x ACT

			ACT Actividad laboral		
			1 No	2 Sí	Total
HOR Horario	1 Mañana	Recuento	110	24	134
		Frecuencia esperada	82,0	52,0	134,0
		% dentro de ACT Actividad laboral	37,9%	13,0%	28,3%
		Residuos corregidos	5,9	-5,9	
	2 Tarde	Recuento	95	70	165
		Frecuencia esperada	100,9	64,1	165,0
		% dentro de ACT Actividad laboral	32,8%	38,0%	34,8%
		Residuos corregidos	-1,2	1,2	
	3 Noche	Recuento	85	90	175
		Frecuencia esperada	107,1	67,9	175,0
		% dentro de ACT Actividad laboral	29,3%	48,9%	36,9%
		Residuos corregidos	-4,3	4,3	
Total		Recuento	290	184	474
		Frecuencia esperada	290,0	184,0	474,0
		% dentro de ACT Actividad laboral	100,0%	100,0%	100%



Se observa ante todo que los grupos más numerosos son el de tarde y sobre todo el de noche. Cuando los separamos entre trabajadores y no trabajadores vemos como el porcentaje de estudiantes que trabajan se reduce al 13% en el grupo de la mañana y sube al 38% y al 49% en los de la tarde y la noche, respectivamente. Por tanto se evidencia un comportamiento diferenciado con residuos significativos en la mañana y en la noche (el de tarde no porque tiene un comportamiento cercano al promedio). La relación es significativa según el test de independencia de chi-cuadrado y la V de Cramer arroja valor de asociación de 0,280.

Se comprueba igualmente la Hipótesis 3 en un análisis bivariable.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)	Sig. exacta (bilateral)
Chi-cuadrado de Pearson	37,285[a]	2	,000	
Razón de verosimilitud	39,834	2	,000	
Asociación lineal por lineal	34,505	1	,000	
Prueba de McNemar				,[b]
N de casos válidos	474			

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 52,02.

b. Sólo se efectuará el cálculo para tablas de PxP, donde P debe ser mayor que 1.

Medidas simétricas

		Valor	Sig. aproximada
Nominal por nominal	Phi	,280	,000
	V de Cramer	,280	,000
	Coeficiente de contingencia	,270	,000
N de casos válidos		474	

Cuando introducimos la tercera variable podemos elegir distintas alternativas de lectura, en particular, analizar el abandono según la actividad, comparando mañana, tarde y noche, o bien analizar el abandono según el horario controlando por la actividad laboral. Comentamos en nuestro modelo de análisis inicial que el interés estaba en contrastar la hipótesis de hasta qué punto la relación entre abandono y horario era de carácter espúrea (Hipótesis 4).

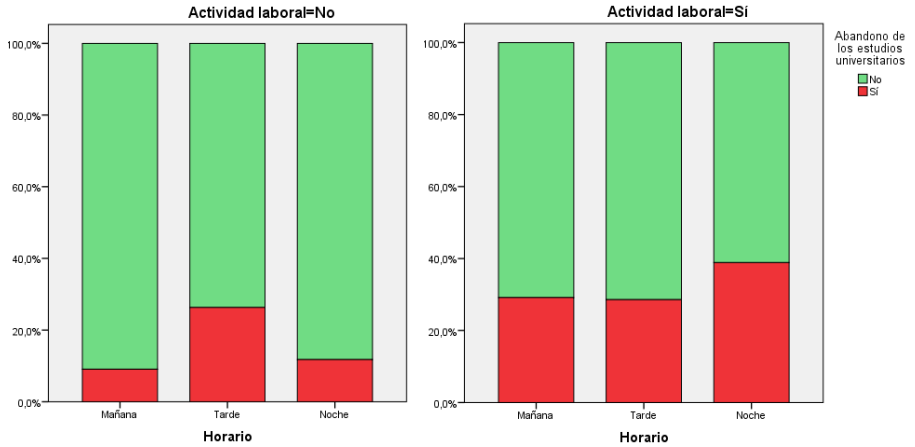
Por ello analizaremos esta relación controlando por actividad para intentar evidenciar que el abandono se debe al hecho de trabajar y que si analizamos solo al estudiantado que trabaja, entre ellos, deben tener tasas de abandono similares, y lo mismo entre los que no trabajan. Veámoslo. La tabla de contingencia y los gráficos de barras que se obtienen se presentan seguidamente.

Tabla de contingencia ABA x HOR x ACT

			ACT Actividad laboral							
			1 No				2 Sí			
			HOR Horario				HOR Horario			
			1 Mañana	2 Tarde	3 Noche	Total	1 Mañana	2 Tarde	3 Noche	Total
ABA Abandono de los estudios universitarios	1 No	Recuento	100	70	75	245	17	50	55	122
		Frecuencia esperada	96,0	69,1	63,1	224,5	21,0	50,9	66,9	142,5
		% dentro de ACT Actividad laboral	**90,9%**	**73,7%**	**88,2%**	**84,5%**	**70,8%**	**71,4%**	**61,1%**	**66,3%**
		Residuos corregidos	2,7	,3	4,1	4,6	-2,7	-,3	-4,1	-4,6
	2 Sí	Recuento	10	25	10	45	7	20	35	62
		Frecuencia esperada	14,0	25,9	21,9	65,5	3,0	19,1	23,1	41,5
		% dentro de ACT Actividad laboral	**9,1%**	**26,3%**	**11,8%**	**15,5%**	**29,2%**	**28,6%**	**38,9%**	**33,7%**
		Residuos corregidos	-2,7	-,3	-4,1	-4,6	2,7	,3	4,1	4,6
Total		Recuento	110	95	85	290	24	70	90	184
		Frecuencia esperada	110,0	95,0	85,0	290,0	24,0	70,0	90,0	184,0
		% dentro de ACT Actividad laboral	100%	100%	100%	100%	100%	100%	100%	100%

La tasa de abandono entre los que trabajan se observan algo superior entre el estudiantado del grupo de noche en relación a la mañana y la tarde. El residuo local es significativo pero globalmente la prueba de chi-cuadrado que relaciona abandono con horario, entre los que sí trabajan, no resulta significativa estadísticamente. Es decir, a pesar de observar ciertas diferencias en la muestra para el grupo de noche debemos considerar que las diferencias se deben al azar y no son extrapolables al conjunto del alumnado. Por tanto, desaparece la relación entre abandono y horario. ¿Y entre los que

no trabajan, su comportamiento también se puede considerar similar y podemos concluir que desaparece la relación? Pues no. Si nos fijamos en el gráfico o en la tabla vemos como el grupo de tarde tiene un comportamiento claramente diferenciado, su tasa de abandono es del 26,3% cuando los que no trabajan de la mañana y de la noche tienen porcentajes del 9,1% y del 11,8%. En este caso el chi-cuadrado confirma que las diferencias son significativas y existe una asociación que la V de Cramer parcial cifra en 0,210.



Este resultado nos permite llegar a la conclusión de que existen dos comportamientos diferenciados, el de los que no tienen actividad laboral (existe asociación) y el de los que tienen actividad laboral (desaparece la relación). En consecuencia, al observar dos patrones de comportamiento, la relación original entre abandono y horario varía a cada nivel de la tercera variable, concluimos la existencia de una interacción verificándose el modelo de interacción.

Pruebas de chi-cuadrado

ACT Actividad laboral		Valor	gl	Sig. asintótica (bilateral)
1 No	Chi-cuadrado de Pearson	12,83[b]	2	,002
	Razón de verosimilitudes	12,216	2	,002
	Asociación lineal por lineal	,593	1	,441
	N de casos válidos	290		
2 Sí	Chi-cuadrado de Pearson	2,129[c]	2	,345
	Razón de verosimilitudes	2,134	2	,344
	Asociación lineal por lineal	1,636	1	,201
	N de casos válidos	184		

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 13,19.

c. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 8,09.

Medidas simétricas

ACT Actividad laboral			Valor	Sig. aproximada
1 No	Nominal por nominal	Phi	,210	,002
		V de Cramer	,210	,002
	N de casos válidos		290	
2 Sí	Nominal por nominal	Phi	,108	,345
		V de Cramer	,108	,345
	N de casos válidos		184	

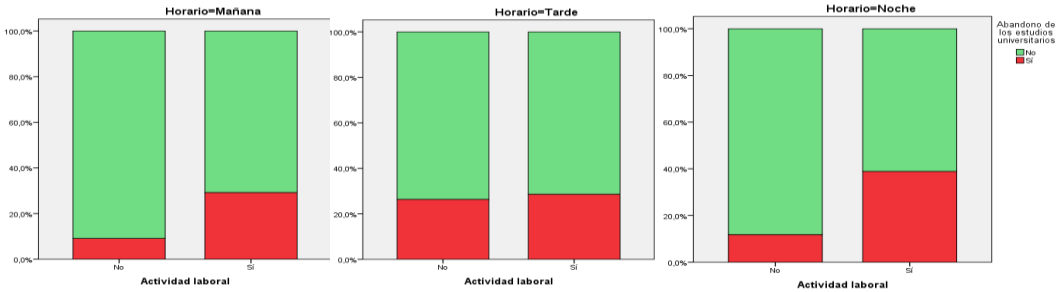
a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Para completar el ejercicio analizaremos la tabla de contingencia entre abandono y actividad, controlando por horario, donde tenemos esta información:

Tabla de contingencia ABA x ACT x HOR

			HOR Horario								
			1 Mañana			2 Tarde			3 Noche		
			ACT Actividad laboral			ACT Actividad laboral			ACT Actividad laboral		
			1 No	2 Sí	Total	1 No	2 Sí	Total	1 No	2 Sí	Total
ABA Abandono de los estudios universitarios	1 No	Recuento	100	17	117	70	50	120	75	55	130
		Frecuencia esperada	92,9	15,9	103,8	80,3	46,4	127,8	71,8	59,7	135,5
		% dentro de HOR Horario	**90,9%**	**70,8%**	**87,3%**	**73,7%**	**71,4%**	**72,7%**	**88,2%**	**61,1%**	**74,3%**
		Residuos corregidos	2,4	,5	3,2	-3,5	1,2	-1,8	1,1	-1,5	-1,3
	2 Sí	Recuento	10	7	17	25	20	45	10	35	45
		Frecuencia esperada	17,1	8,1	30,2	14,7	23,6	37,2	13,2	30,3	39,5
		% dentro de HOR Horario	**9,1%**	**29,2%**	**12,7%**	**26,3%**	**28,6%**	**27,3%**	**11,8%**	**38,9%**	**25,7%**
		Residuos corregidos	-2,4	-,5	-3,2	3,5	-1,2	1,8	-1,1	1,5	1,3
Total		Recuento	110	24	134	95	70	165	85	90	175
		Frecuencia esperada	110,0	24,0	134,0	95,0	70,0	165,0	85,0	90,0	175,0
		% dentro de HOR Horario	100%	100%	100%	100%	100%	100%	100%	100%	100%



Podemos comprobar que los porcentajes son los mismos que en la tabla trivariable anterior pero se disponen en subtablas diferentes. En este caso la lectura de la información nos dice que entre los de la mañana y los de la noche existen diferencias de abandono según se trabaje o no se trabaje, así lo muestra el test de chi-cuadrado. Pero en el de la tarde las diferencias desaparecen. No tenemos más información para dilucidar qué está pasando en el grupo de tarde, pero sigue una pauta diferente de la esperada que se verifica en el de la mañana y la noche. De esta forma, el comportamiento diferente de la tarde está provocando la interacción y que no podamos validar el modelo de independencia condicional.

Pruebas de chi-cuadrado

HOR Horario de clase		Valor	gl	Sig. asintótica (2 caras)	Significación exacta (2 caras)	Significación exacta (1 cara)
1 Mañana	Chi-cuadrado de Pearson	7,168[c]	1	,007		
	Corrección de continuidad[b]	5,471	1	,019		
	Razón de verosimilitud	5,949	1	,015		
	Prueba exacta de Fisher				,014	,014
	Asociación lineal por lineal	7,115	1	,008		
	N de casos válidos	134				
2 Tarde	Chi-cuadrado de Pearson	,103[d]	1	,748		
	Corrección de continuidad[b]	,021	1	,885		
	Razón de verosimilitud	,103	1	,748		
	Prueba exacta de Fisher				,860	,441
	Asociación lineal por lineal	,103	1	,749		
	N de casos válidos	165				
3 Noche	Chi-cuadrado de Pearson	16,837[e]	1	,000		
	Corrección de continuidad[b]	15,447	1	,000		
	Razón de verosimilitud	17,656	1	,000		
	Prueba exacta de Fisher				,000	,000
	Asociación lineal por lineal	16,740	1	,000		
	N de casos válidos	175				

b. Sólo se ha calculado para una tabla 2x2

c. 1 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 3,04.

d. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 19,09.

e. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 21,86.

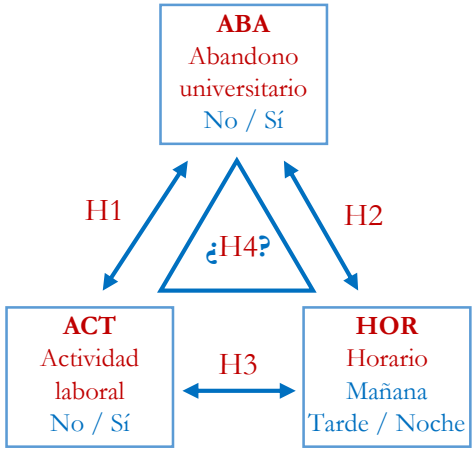
Medidas simétricas

HOR Horario			Valor	Sig. aproximada
1 Mañana	Nominal por nominal	Phi	,231	,007
		V de Cramer	,231	,007
	N de casos válidos		134	
2 Tarde	Nominal por nominal	Phi	,025	,748
		V de Cramer	,025	,748
	N de casos válidos		165	
3 Noche	Nominal por nominal	Phi	,310	,000
		V de Cramer	,310	,000
	N de casos válidos		175	

a. Asumiendo la hipótesis alternativa.

b. Empleando el error típico asintótico basado en la hipótesis nula.

Dicho lo cual, y como anunciamos al inicio, el hecho de que observemos dos patrones distintos ¿hasta qué punto es un resultado concluyente estadísticamente en un análisis de tablas de contingencia como el realizado? Los resultados estadísticos parciales de las subtablas no son objeto de contraste entre sí en un análisis clásico de tablas de contingencia por lo que no podemos establecer con certeza un posible modelo de interacción como este:



Nos quedamos con el interrogante que resolveremos en el próximo capítulo con un análisis log-lineal.

► Ejercicio 14.

Proponer un modelo de relación entre las variables ACT (actitud: grado de acuerdo con la afirmación "Las mujeres deben quedarse en su casa"), EST (el nivel de estudios) y SEX (el sexo de la persona entrevistada) y contrastar las hipótesis con los datos siguientes de forma similar al ejercicio realizado con el ejemplo del abandono universitario. El archivo de sintaxis ATC-Actitud.sps de la página web contiene la sintaxis que genera los datos y obtiene las tablas de contingencia.

Recuento

		SEX Sexo							
		1 Varón				2 Mujer			
		EST Nivel de estudios				EST Nivel de estudios			
		1 Primarios	2 Secundarios	3 Superiores	Total	1 Primarios	2 Secundarios	3 Superiores	Total
ACT Actitud: permanencia de la mujer en el hogar	1 De acuerdo	72	110	44	226	86	173	28	287
	2 En desacuerdo	47	196	179	422	38	283	187	508
Total		119	306	223	648	124	456	215	795

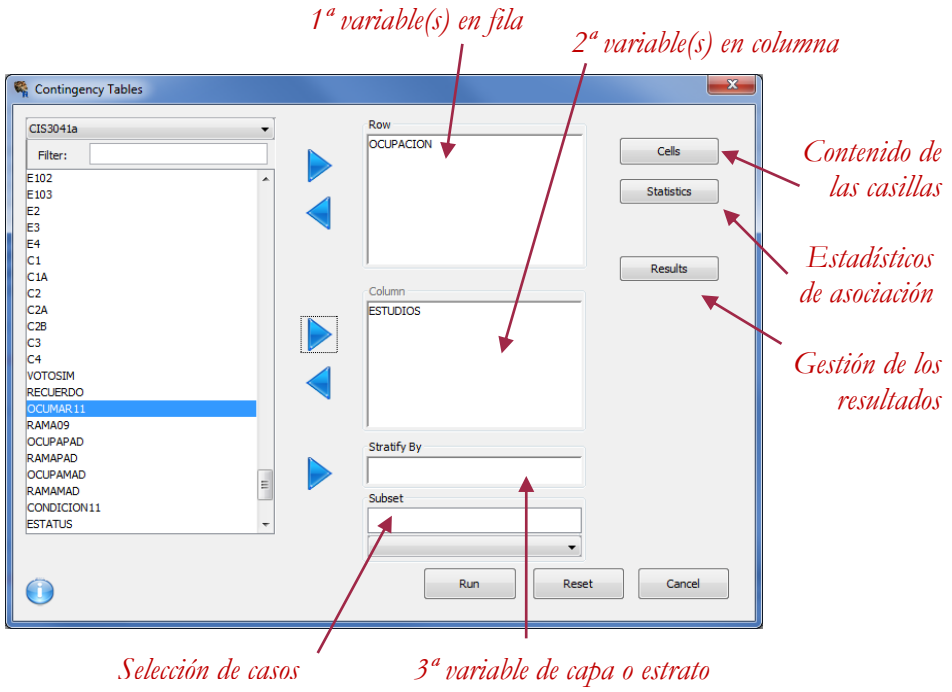
6.2. Análisis de tablas de contingencia con R

6.2.1. *Análisis descriptivo con dos variables*

Para analizar la relación entre dos (o más variables) de tipo cualitativo, variables medidas a nivel nominal u ordinal, disponemos de diversas alternativas en R. Nos centraremos en el procedimiento Contingency Tables de Deducer que nos proporciona tablas de distribución conjunta de frecuencias con el cálculo de varias medidas que nos evalúan la existencia de asociación y su intensidad.

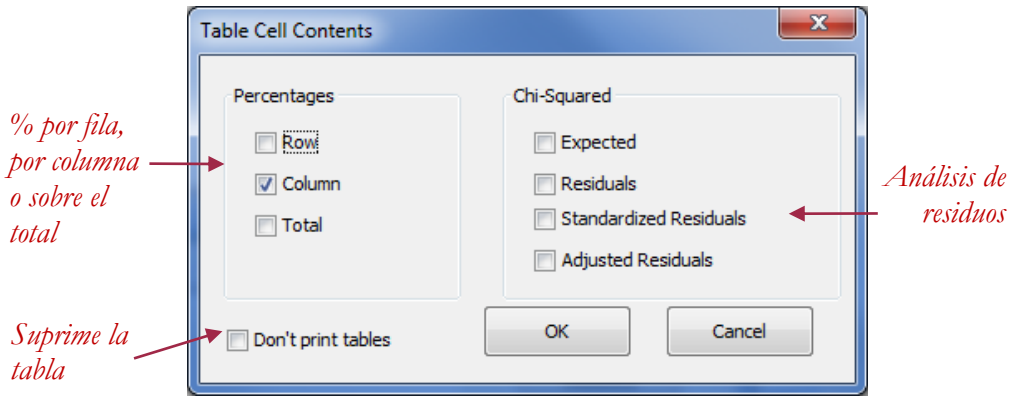
Analizaremos la relación entre la ocupación y los estudios para estudiar la asociación entre el nivel profesional alcanzado (variable dependiente) según el nivel de educación formal alcanzado (variable independiente). Nuestra hipótesis plantea en términos generales que a mayor nivel de estudios cabe esperar un mayor nivel ocupacional. Consideraremos los datos de la matriz CIS3041a.sav donde se generó la variable agrupada de ocupación con el nombre OCUPACION y la variable original de la base de datos del CIS ESTUDIOS.

El análisis de tablas de contingencia se realiza a través del menú Analysis / Contingency Tables. El cuadro de diálogo inicial que aparece a continuación nos muestra por un lado el recuadro con el listado de variables de la matriz de datos y unos recuadros donde se trasladarán las variables según se sitúen en filas o en columnas, o bien en capa o estrato si se considera una tabla de tres dimensiones. Situaremos la variable dependiente en las filas (OCUPACION), y la variable independiente en las columnas (ESTUDIOS)[47].



[47] Qué variable se coloca en fila o en columna es pura convención, el análisis y el resultado es simétrico y se genera la misma información.

A continuación disponemos de diversas opciones que permiten especificar la información que queremos obtener del análisis. En un primer ejercicio de análisis descriptivo solicitaremos simplemente las frecuencias absolutas observadas o "recuento observado" (opción por defecto) y los porcentajes por columna (pues la variable independiente está en columnas) a través del botón Cells:



Las frecuencias absolutas de la tabla aparecen por defecto y en el apartado de Percentages de este cuadro de diálogo además de los porcentajes por columna también podemos seleccionar los tanto por ciento por fila y sobre el total. En un análisis descriptivo no requerimos más información que estas frecuencias y una representación gráfica que solicitaremos a continuación. El resultado que se obtiene es el siguiente.

Contingency Tables

OCUPACION by ESTUDIOS across levels of

		ESTUDIOS						
OCUPACION		Sin	Primaria	Secundaria1	Secundaria2	FP	Superiores	Row Total
Alta	Count	1	14	23	54	33	268	393
	Column %	0.719%	2.82%	3.82%	16.17%	7.95%	59.82%	
Media	Count	2	44	69	101	107	102	425
	Column %	1.44%	8.85%	11.46%	30.24%	25.78%	22.77%	
Cualificada	Count	77	265	311	128	203	65	1049
	Column %	55.40%	53.32%	51.66%	38.32%	48.92%	14.51%	
No cualificada	Count	59	174	199	51	72	13	568
	Column %	42.45%	35.01%	33.06%	15.27%	17.35%	2.90%	
Column Total		139	497	602	334	415	448	**2435**
Column %		5.71%	20.41%	24.72%	13.72%	17.04%	18.40%	

Si nos fijamos en los porcentajes de la tabla observamos cómo a medida que el nivel de estudios es más alto va aumentando el porcentaje de personas con mayor nivel ocupacional. Así por ejemplo, entre los que tienen estudios superiores casi el 60% puede alcanzar la clase ocupacional alta, y a medida que baja el nivel de estudios se va reduciendo hasta la situación donde, si no se tienen estudios, no solo es poco probable, sino imposible alcanzar dicha categoría profesional. Semejante lectura pero en sentido inverso podemos extraer al considerar el nivel inferior de los trabajadores no cualificados, donde observamos en particular que si no se tienen estudios es el nivel ocupacional más probable que se puede alcanzar.

En la tabla observamos también el particular comportamiento del nivel de FP que se corresponde más con la posición 4 en el orden de niveles educativos que con la posición 5 que tiene: las personas con la segunda etapa de secundaria obtienen mejores ocupaciones que las personas con formación profesional.

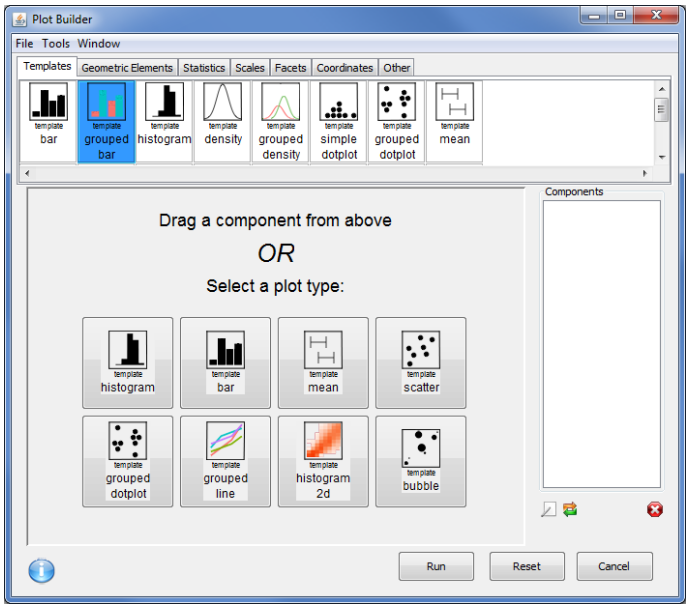
Por defecto Deducer extrae los porcentajes por fila y por columna. Nosotros hemos solicitado solamente éstos últimos para simplificar la información de la tabla. Pero de esta forma no tenemos la información del marginal de fila. Si queremos disponer de él deberemos pedir también los porcentajes por fila. La tabla es la siguiente:

Contingency Tables

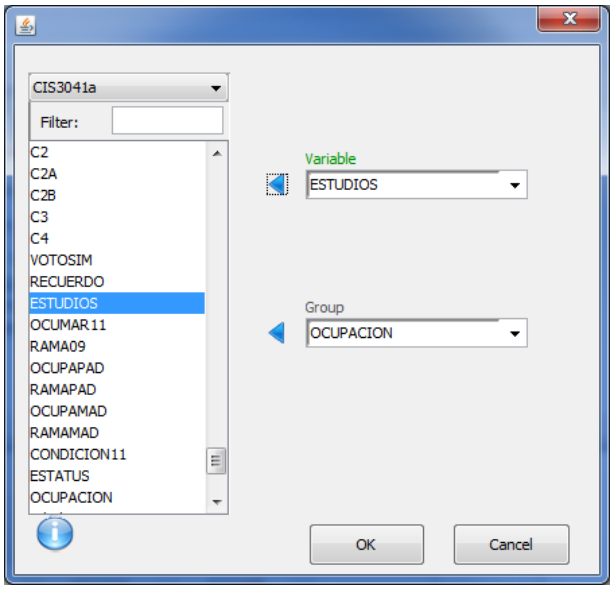
OCUPACION by ESTUDIOS across levels of

		ESTUDIOS						
OCUPACION		Sin	Primaria	Secundaria1	Secundaria2	FP	Superiores	Row Total
Alta	Count	1	14	23	54	33	268	393
	Row %	0.254%	3.56%	5.85%	13.74%	8.40%	68.19%	16.14%
	Column %	0.719%	2.82%	3.82%	16.17%	7.95%	59.82%	
Media	Count	2	44	69	101	107	102	425
	Row %	0.471%	10.35%	16.24%	23.76%	25.18%	24.00%	17.45%
	Column %	1.44%	8.85%	11.46%	30.24%	25.78%	22.77%	
Cualificada	Count	77	265	311	128	203	65	1049
	Row %	7.34%	25.26%	29.65%	12.20%	19.35%	6.20%	43.08%
	Column %	55.40%	53.32%	51.66%	38.32%	48.92%	14.51%	
No cualificada	Count	59	174	199	51	72	13	568
	Row %	10.39%	30.63%	35.04%	8.98%	12.68%	2.29%	23.33%
	Column %	42.45%	35.01%	33.06%	15.27%	17.35%	2.90%	
Column Total		139	497	602	334	415	448	**2435**
Column %		5.71%	20.41%	24.72%	13.72%	17.04%	18.40%	

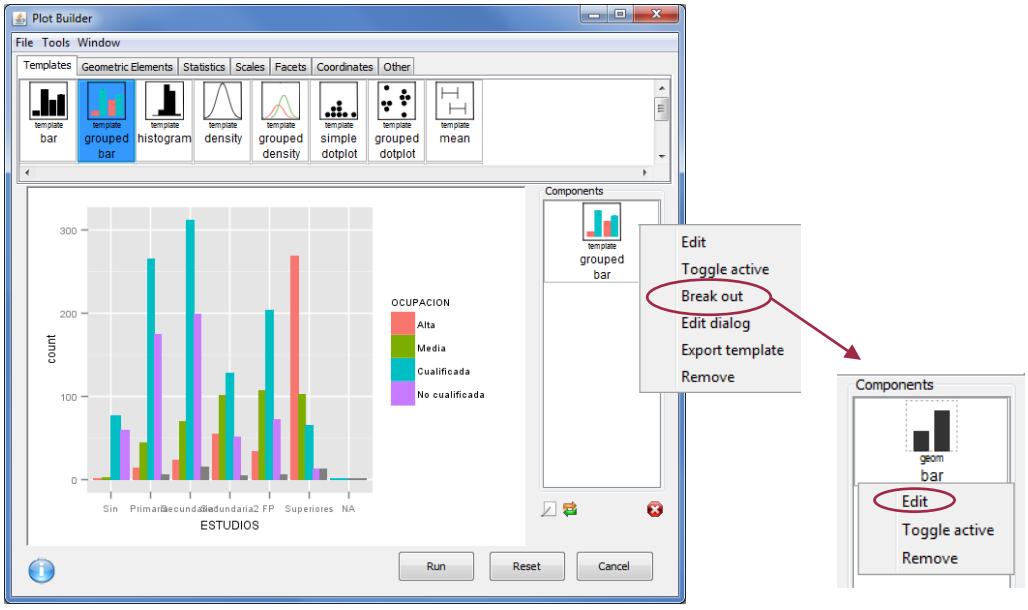
En un gráfico podemos observar visualmente la tendencia de asociación positiva entre estudios y ocupación. Para obtener una representación gráfica iremos al menú Plot / Plot Builder y seguiremos los siguientes pasos para obtener un gráfico de barras apiladas al 100%. Desde el cuadro de diálogo de Plot Builder elegimos grouped bar y hacemos doble-clic:



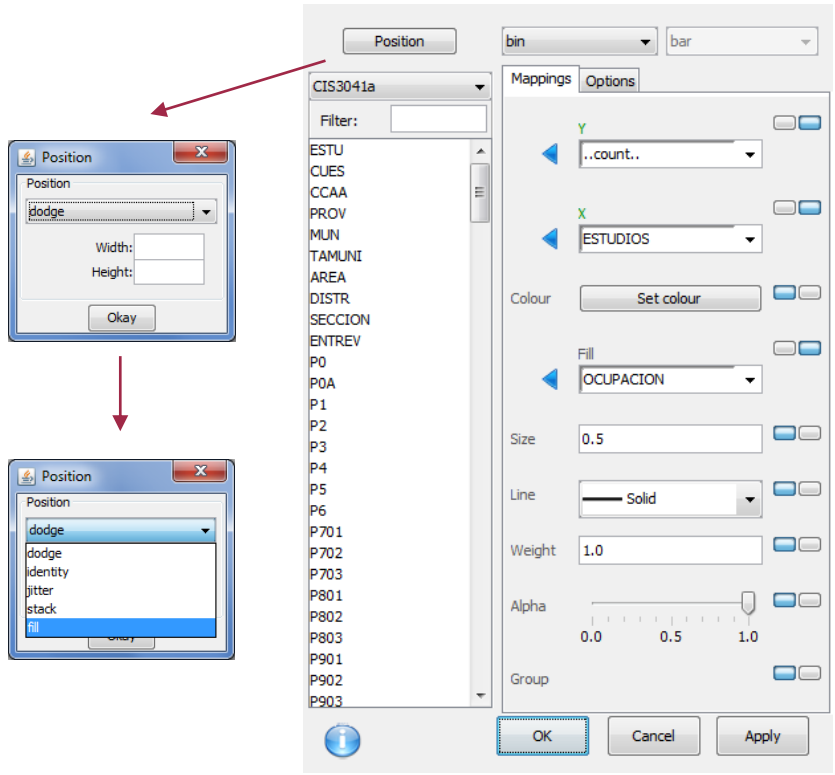
Nos aparece el siguiente cuadro de diálogo donde colocaremos la variable independiente ESTUDIOS en Variable (aparecerá en el eje de categorías X) y la variable dependiente OCUPACION en Group (será la leyenda del gráfico):



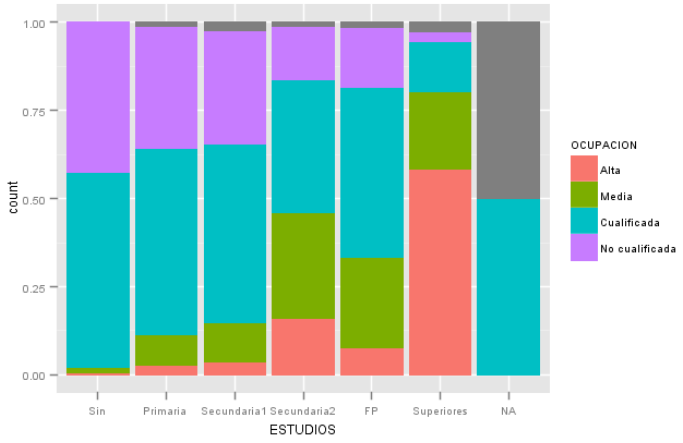
Ejecutando el procedimiento con OK se obtiene un gráfico de barras agrupadas con las frecuencias absolutas (count):



Presionamos el botón derecho sobre el icono grouped bar, en el recuadro de Components y seleccionamos la opción Break out. Luego observaremos que el icono de grouped bar ha cambiado a bar. Nuevamente presionamos presionamos el botón derecho sobre el nuevo icono y seleccionamos la opción Edit, se abrirá junto a la ventana un nuevo cuadro de diálogo:



Presionaremos sobre el botón Position y elegimos del menú desplegable la opción fill. Clicamos OKay, sobre OK y luego sobre Run, obtendremos finalmente el siguiente gráfico:



Observamos que se incluyen los valores perdidos (NA) y que la escala ha cambiado a proporciones entre 0 y 1. Modificaremos ambos aspectos con la ayuda de la sintaxis de R[48]. Para obtener un gráfico sin los valores perdidos recurriremos a la línea de comandos para ejecutar la misma instrucción seleccionando los casos válidos. Podemos recuperar la instrucción que ejecutó Deducer y que aparece en la consola:

[48] Plot Builder de Deducer corresponde al comando de elaboración de gráficos ggplot de R, se puede ampliar la información en: http://ggplot2.org/.

```
> dev.new()
> ggplot() +
+ geom_bar(aes(y = ..count..,x = ESTUDIOS,fill =
OCUPACION),data=CIS3041a,position = position_fill())
```

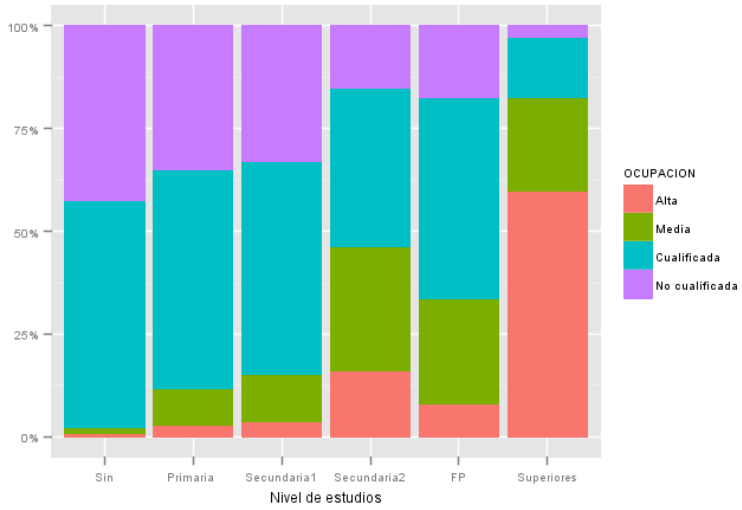
y la modificamos añadiendo la especificación que selecciona los casos (subset) que no (!) son perdidos (is.na) en ambas variables:

```
> dev.new()
> ggplot() + geom_bar(aes(y = ..count..,x = ESTUDIOS,fill =
OCUPACION),data=subset(CIS3041a,!is.na(ESTUDIOS) &
!is.na(OCUPACION)), position = position_fill())
```

Si adicionalmente cambiamos la escala del eje y donde aparece la etiqueta count, cambiamos su etiqueta por un texto alternativo como % o bien un texto en blanco, y etiquetamos el eje de categorías x:

```
> dev.new()
> ggplot() + geom_bar(aes(y = ..count..,x = ESTUDIOS,fill =
OCUPACION),data=subset(CIS3041a,!is.na(ESTUDIOS) &
!is.na(OCUPACION)), position = position_fill())+
scale_y_continuous(labels = percent_format()) + ylab ("") + xlab
("Nivel de estudios")
```

El gráfico finalmente cambia a este formato:



► Ejercicio 15.

Con la matriz de datos CIS3041a.rda analizar la relación entre parejas de variables cualitativas. Por ejemplo se puede analizar el comportamiento electoral según diversas variables independientes como la edad, el sexo, la ocupación, el lugar de residencia, etc. En relación a cada tabla:

- Observa los totales marginales y calcula los porcentajes marginales.
- Calcula los porcentajes condicionales (fila y columna).
- Comenta la distribución condicional que hay que interpretar en cada tabla según la definición de la variable dependiente y la independiente para determinar la existencia de asociación entre las variables.
- Crea un gráfico de barras apiladas al 100%.

6.2.2. *Introducción de datos de una tabla de contingencia*

Si queremos trabajar con una tabla de contingencia que aparezca publicada en algún medio y reproducir los porcentajes y las gráficas, es posible hacerlo a partir de la introducción directa de las frecuencias de la tabla y tratarla como una clase particular de objeto de R como es una table. Consideremos el caso de la Tabla III.6.1 que relaciona la posesión de coche y la clase social, las frecuencias absolutas son:

Clase Coche	*Alta 1*	*Media 2*	*Baja 3*	Total
Sí 1	650	1234	1430	3314
No 2	64	333	1036	1433
Total	714	1567	2466	4747

La introducción de los datos y el análisis de la tabla lo haremos con el lenguaje de comandos[49]. En primer lugar creamos la tabla de contingencia y le asignamos el nombre de objeto tabla con una primera instrucción que crea una matrix de 2 filas y 3 columnas y la convierte en una table de R. A continuación se etiquetan las filas y las columnas obteniendo este resultado:

```
> tabla=as.table(matrix(c(650,64,1234,333,1430,1036),
      nrow=2,ncol=3))
> colnames(tabla)=c("Alta","Media","Baja")
> rownames(tabla)=c("Sí","No")
> tabla
   Alta Media Baja
Sí  650  1234 1430
No   64   333 1036
```

Las instrucciones que siguen se destinan a obtener las proporciones de la tabla con el comando **prop.table** que en el último caso se multiplica por 100 para convertirlos en porcentajes con un decimal:

```
> prop.table(tabla) # proporción total
         Alta      Media       Baja
Sí 0.13692859 0.25995365 0.30124289
No 0.01348220 0.07014957 0.21824310
> prop.table(tabla,1) # proporción fila
         Alta      Media       Baja
Sí 0.19613760 0.37235969 0.43150272
No 0.04466155 0.23237962 0.72295883
> prop.table(tabla,2) # proporción columna
         Alta      Media       Baja
Sí 0.91036415 0.78749202 0.57988646
No 0.08963585 0.21250798 0.42011354
> round(prop.table(tabla,2)*100,1) # % columna con 1 decimal
   Alta Media Baja
Sí 91.0  78.7 58.0
No  9.0  21.3 42.0
```

[49] Las instrucciones que comentaremos se recogen en el script Coche.R que se enceuntraen la página web. Podemos editar y ejecutar las instrucciones desde la consola de R a través de Deducer o desde R-Studio.

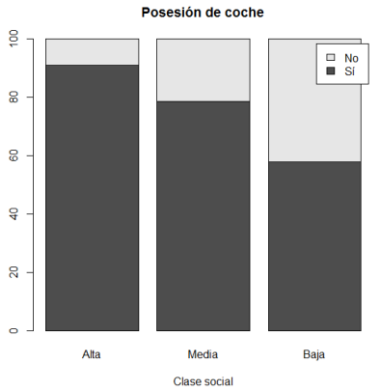
Las siguientes instrucciones extraen y añaden los marginales a la tabla:

```
> margin.table(tabla) # marginal total
[1] 4747
> margin.table(tabla,1) # marginal de fila
  Sí   No
3314 1433
> margin.table(tabla,2) # marginal de columna
 Alta Media  Baja
  714  1567  2466
> addmargins(tabla, margin=1) # Se añade marginal de fila
    Alta Media Baja
Sí   650  1234 1430
No    64   333 1036
Sum  714  1567 2466
> addmargins(tabla, margin=2) # Se añade marginal de columna
   Alta Media Baja  Sum
Sí  650  1234 1430 3314
No   64   333 1036 1433
> addmargins(tabla) # Se añaden marginal de fila y columna
    Alta Media Baja  Sum
Sí   650  1234 1430 3314
No    64   333 1036 1433
Sum  714  1567 2466 4747
> addmargins(round(prop.table(tabla)*100,1)) # % total con
  marginal de fila y columna
    Alta Media Baja  Sum
Sí  13.7  26.0 30.1 69.8
No   1.3   7.0 21.8 30.1
Sum 15.0  33.0 51.9 99.9
```

Por último se obtiene un gráfico de barras y se realiza el test de chi-cuadrado con el comando **`chisq.test`** o bien pidiendo un **`summary`** de la tabla:

```
> # Gráfico de barras
> barplot(round(prop.table(tabla,2)*100,1),legend=TRUE,
          xlab="Clase social",main="Posesión de coche")
> chisq.test(tabla)

        Pearson's Chi-squared test

data:  tabla
X-squared = 375.5831, df = 2, p-value < 2.2e-16
> summary(tabla)
Number of cases in table: 4747
Number of factors: 2
Test for independence of all factors:
        Chisq = 375.6, df = 2, p-value = 2.774e-82
```

El gráfico que se obtiene es el siguiente:



► **Ejercicio 16.**
Introduce los datos de la tabla de contingencia siguiente y analiza la relación entre las variables:

		P31 Sexo		
		Hombre	Mujer	Total
P1 Valoración de la situación económica general de España	Buena	19	13	32
	Regular	203	196	399
	Mala	510	480	990
	Muy mala	474	576	1050
Total		1206	1265	2471

6.2.3. *Análisis inferencial con dos variables*

Completaremos el análisis de una tabla de contingencia con el estudio inferencial de la relación que nos permite establecer la significación estadística de la relación entre las variables y poder inferir el resultado al conjunto de la población.

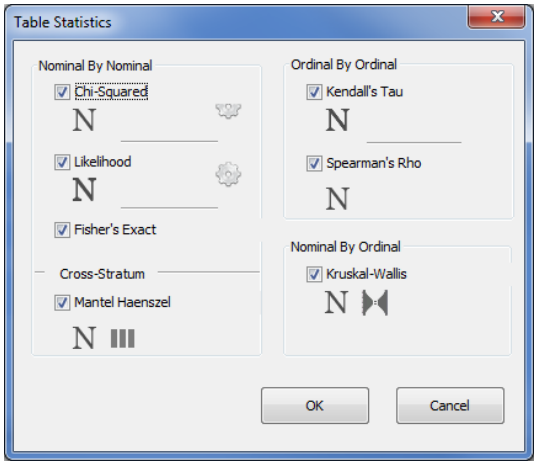
Después de analizar descriptivamente la tabla con los porcentajes y el gráfico, y llegar a la conclusión que a medida que aumentan los estudios aumenta la ocupación, nos formulamos ahora dos preguntas para completar nuestro estudio: estas diferencias observadas son suficientemente significativas desde un punto de vista estadístico? Si lo son, ¿cuál es grado de relación entre estas variables?

Para responder rigurosamente y con objetividad desde un punto de vista estadístico debemos realizar la prueba estadística del test de independencia de chi-cuadrado, donde contrastaremos dos hipótesis:
- la hipótesis nula: consiste en asumir que las dos variables son independientes y que no existe ninguna relación de asociación entre ellas,
- la hipótesis alternativa: consiste en aceptar que sí existe algún tipo de relación de asociación de dependencia pues la hipótesis alternativa no es falsable.

El objetivo es saber si, con un cierto nivel de confianza, tenemos evidencias suficientes como para rechazar la hipótesis nula y concluir que las diferencias porcentuales son significativas. Una vez constatada la significación de la relación tiene sentido contestar a la segunda pregunta, se trata de calcular una medida de la intensidad de la relación. Consideraremos la V de Cramer, la cual incorpora también una prueba de significación, de hecho, la misma que la del chi-cuadrado ya que se trata de una medida basada en aquel estadístico.

Vamos a ver cómo realizar este test con Deducer. En el procedimiento de Contingency Tables primero pulsamos sobre el botón Cells para pedirle todos los cálculos de frecuencias esperadas y residuos en el recuadro de Chi-Squared. A continuación en Statistics podemos marcar todas las opciones de medidas de asociación disponibles para ver los resultados. Junto con el chi-cuadrado disponemos de la razón de verosimilitud o *Likelihood ratio*, que es una reformulación del de Pearson con resultados

similares[50]. Estas pruebas se aplican a tablas con cualquier número de filas o columnas, pero también se calculan dos estadísticos más destinados a establecer la existencia de asociación en el caso particular de que disponemos de una tabla de 2×2: la prueba exacta de Fisher (Fisher's Exact), que se utiliza cuando una casilla tiene una frecuencia esperada menor que 5, y la prueba de Mantel-Haenszel. Con variables ordinales se calcula la Tau de Kendall o la Rho de Spearman, combinando una nominal con una ordinal utilizamos el estadístico de Kruskal-Wallis. En nuestro caso las dos variables son ordinales y los diferentes estadísticos son interpretables, excepto los de tablas de 2×2.



Los resultados son los siguientes:

Contingency Tables

OCUPACION by ESTUDIOS across levels of

		ESTUDIOS						
OCUPACION		Sin	Primaria	Secundaria1	Secundaria2	FP	Superiores	Row Total
Alta	Count	1	14	23	54	33	268	393
	Row %	0.254%	3.56%	5.85%	13.74%	8.40%	68.19%	16.14%
	Column %	0.719%	2.82%	3.82%	16.17%	7.95%	59.82%	
	Expected	22.43	80.21	97.16	53.91	66.98	72.31	
	Adj Resid	-5.09	-9.05	-9.47	0.015	-4.98	27.82	
Media	Count	2	44	69	101	107	102	425
	Row %	0.471%	10.35%	16.24%	23.76%	25.18%	24.00%	17.45%
	Column %	1.44%	8.85%	11.46%	30.24%	25.78%	22.77%	
	Expected	24.26	86.75	105.07	58.30	72.43	78.19	
	Adj Resid	-5.12	-5.66	-4.46	6.63	4.91	3.28	
Cualificada	Count	77	265	311	128	203	65	1049
	Row %	7.34%	25.26%	29.65%	12.20%	19.35%	6.20%	43.08%
	Column %	55.40%	53.32%	51.66%	38.32%	48.92%	14.51%	
	Expected	59.88	214.11	259.34	143.89	178.78	193.00	
	Adj Resid	3.02	5.17	4.90	-1.89	2.64	-13.52	
No cualificada	Count	59	174	199	51	72	13	568
	Row %	10.39%	30.63%	35.04%	8.98%	12.68%	2.29%	23.33%
	Column %	42.45%	35.01%	33.06%	15.27%	17.35%	2.90%	
	Expected	32.42	115.93	140.43	77.91	96.80	104.50	
	Adj Resid	5.49	6.90	6.51	-3.75	-3.16	-11.32	
Column Total		139	497	602	334	415	448	**2435**
Column %		5.71%	20.41%	24.72%	13.72%	17.04%	18.40%	

[50] Sobre este estadístico, llamado también L^2 o G^2, volveremos en el capítulo siguiente pues con él estableceremos la significatividad de los modelos log-lineales.

Contingency Table Tests

Tests for OCUPACION by ESTUDIOS across levels of

	statistic	*df*	*asymptotic p-value*	*exact p-value*	*ES measure*	*ES est.*
Chi Squared	1075.80	15	7.51e-220			
Likelihood	1010.81	15.00	0.00			
Fishers Exact						
spearman's Correlation	3.64e+09		2.6e-162		rho	-0.511
kendall's Correlation	-25.86		1.78e-147		tau	-0.43
Kruskal-W (nominal rows)	710.62	3.00	1.05e-153			
Kruskal-W (nominal cols)	786.76	5.00	8.46e-168			

En Deducer no se calculan las medidas de asociación basadas en el chi-cuadrado que comentamos anteriormente. Para obtenerlas podemos instalar y cargar el paquete vcd[51] y ejecutar la función **`assocstats`** a partir de la tabla que relaciona las dos variables:

```
> install.packages("vcd")
> library(vcd)
> assocstats(table(CIS3041a$OCUPACION, CIS3041a$ESTUDIOS))
                    X^2 df P(> X^2)
Likelihood Ratio 1010.8 15        0
Pearson          1075.8 15        0

Phi-Coefficient   : 0.665
Contingency Coeff.: 0.554
Cramer's V        : 0.384
```

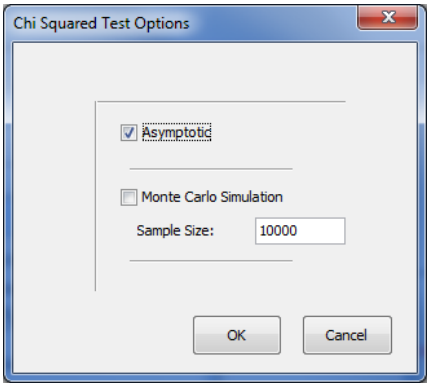
A tenor de los resultados constatamos cómo la significación de la prueba de chi-cuadrado es un valor muy pequeño[52], de 0,000<0,05, con lo que podemos concluir que hay relación entre las variables, que las diferencias porcentuales son significativas con un nivel de confianza del 95% (con un 5% de riesgo). Esta afirmación se mantiene siempre y cuando las condiciones para interpretar el test se den: la frecuencia mínima esperada en cada casilla sea 1 como mínimo y el porcentaje de casillas con una frecuencia esperada inferior a 5 sea inferior al 20% como se puede comprobar en la tabla de contingencia anterior. Finalmente, el grado de esta relación observada es de 0,384 según la V de Cramer, un valor intermedio, importante y significativo (0,000), pero que no alcanza un valor elevado como 0,6. Es decir, el grado en que se determina el nivel ocupacional por el nivel educativo es relevante pero limitado.

Además de la V de Cramer se pueden elegir otros estadísticos que tienen una función similar de evaluación de la intensidad de la relación y varían según la escala de medición de las variables y que deben ser interpretados desde la lógica de su construcción que es diferente en cada caso. Así, el coeficiente de contingencia por ejemplo arroja un valor más alto de 0,554 o la tau de Kendall de -0,43.

Dentro de las opciones del procedimiento también es posible realizar una simulación de Montecarlo a partir de 10000 muestras aleatorias de nuestros datos que generan una

[51] vcd corresponde al acrónimo de *Visualizing Categorical Data* ya que se trata de un paquete inspirado en el libro del mismo nombre de Michael Friendly (2000, 2013). Se puede ver en ver http://www.datavis.ca/books/vcd/.
[52] El valor 7,51e-220 es 7,51 por 10^{-220} , es decir, un número muy bajo que podemos considerar 0 (es 0,0000… hasta 219 ceros a la derecha de la coma y 751) y representar por 0,000.

estimación más robusta de la significación del estadístico chi-cuadrado. En el cuadro diálogo de Statistics si clicamos sobre al lado de Chi-Squared accedemos al segundo cuadro de diálogo donde se activa:



► **Ejercicio 17.**
Con la matriz de datos CIS3041+.sav y siguiendo el ejemplo del ejercicio del Ejercicio 15 completar el análisis con los aspectos siguientes:
- Interpreta si hay relación entre las variables según el estadístico Chi cuadrado.
- Interpreta la fuerza de la relación (si es que la hay) observando el estadístico V de Cramer.

6.2.4. *Análisis de tablas de contingencia multidimensionales*

Para analizar la relación entre variables cualitativas en tablas multidimensionales consideraremos algunos ejemplos de asociaciones trivariables. En primer lugar presentamos el ejemplo de la relación entre las variables abandono de los estudios universitarios (ABA), la actividad laboral (ACT) y el horario (HOR), a partir de un total de 474 casos[53]. La tabla de contingencia con las frecuencias absolutas es la siguiente:

Tabla III.6.33. El abandono de los estudios universitarios según la actividad laboral y el horario

Contingency Tables

ABA by ACT across levels of HOR

Stratum: HOR = Mañana

		ACT		
ABA		No	Sí	Row Total
No	Count	100	17	117
Sí	Count	10	7	17
Column Total		110	24	**134**

Stratum: HOR = Tarde

		ACT		
ABA		No	Sí	Row Total
No	Count	70	50	120
Sí	Count	25	20	45
Column Total		95	70	**165**

Stratum: HOR = Noche

		ACT		
ABA		No	Sí	Row Total
No	Count	75	55	130
Sí	Count	10	35	45
Column Total		85	90	**175**

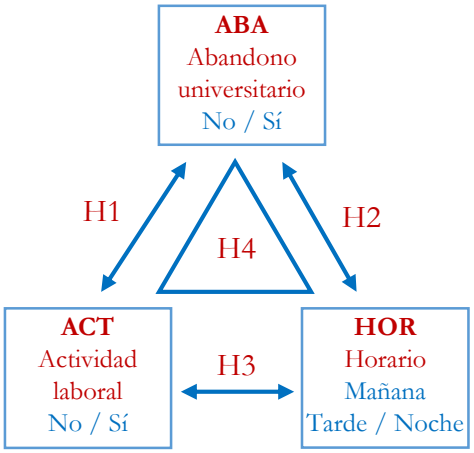
Fuente: Latiesa (1991)

Con estas variables podríamos plantear un modelo de análisis con diferentes hipótesis. Dos de estas hipótesis nos conducen inicialmente a explicar el abandono como variable dependiente en función de la actividad laboral (Hipótesis 1: trabajar penaliza con

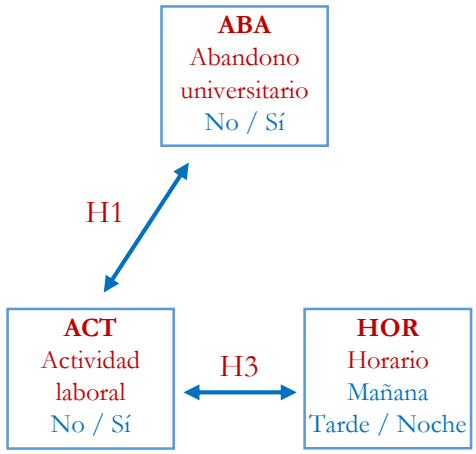
[53] Información extraída de un estudio sobre los alumnos de la Facultad de Ciencias Políticas y Sociología de la Universidad Complutense de Madrid. El ejemplo está publicado en la revista *Papers* por Latiesa (1991).

mayor abandono) y en función del horario de clases (Hipótesis 2: el estudiantado de la tarde y de la noche abandona más). Adicionalmente podemos plantear una tercera relación entre las variables independientes (Hipótesis 3: el estudiantado que tiene una actividad laboral tiende a matricularse sobre todo en los grupos de tarde y de noche). Ahora bien, cabe preguntarse hasta qué punto la razón del abandono tiene que ver realmente con el horario de clases, de hecho podemos pensar que se está dando un mecanismo secuencial donde los trabajadores tienden a matricularse por la tarde-noche y en consecuencia estos grupos tienen una mayor tasa de abandono, por tanto, que la verdadera razón es la actividad laboral y no el grupo de clase. En este sentido una posible y aparente relación entre abandono y horario deberá desaparecer al controlar por la actividad laboral, poniendo de manifiesto una relación espúrea. En este sentido nuestra Hipótesis 4 afirmaría que no existe una relación de interacción entre las variables y que el modelo que cabe esperar es aquel donde el abandono viene explicado solamente por la actividad laboral (Hipótesis 1) y el horario de clase del estudiante es diferente según la actividad laboral (Hipótesis 3). Si por el contrario mantuviéramos la hipótesis de la interacción entre las tres variables deberíamos observar que el abandono no solamente varía en función de la actividad laboral sino que es mayor entre los grupos de tarde y de noche que en los de mañana; podríamos pensar por ejemplo que el estudiantado de la mañana tiene un perfil de estudiante que trabaja y el estudiantado de la noche un perfil de trabajador que estudia y por tanto entre esos últimos la penalización de trabajar es más acentuada.

Planteamos así dos posibles alternativas de modelo de análisis con hipótesis distintas que se trata de verificar seguidamente. La representación gráfica en ambos modelos sería la siguiente:



Modelo 1: Interacción



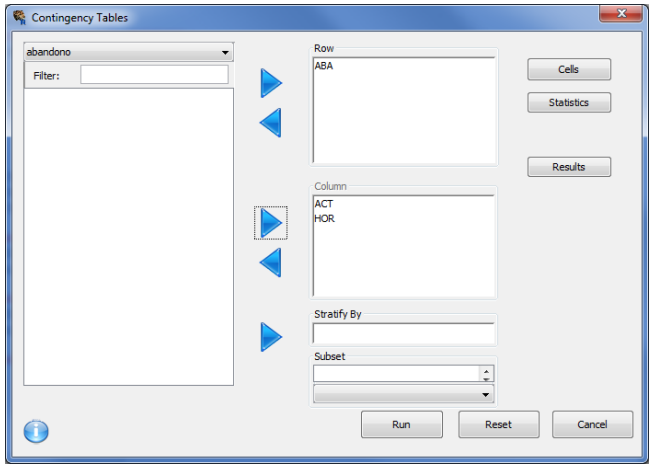
Modelo 2: Doble asociación

Veamos cómo analizar estos modelos con el software estadístico.

A través del procedimiento Contingency Tables del menú Analysis cuando consideramos sólo relaciones bivariables, las tablas de contingencia que se obtienen relacionan todas las combinaciones de las variables incluidas en los recuadros de Row y Column. Si por ejemplo colocamos la variable A en el recuadro de Row y las variables B y C en el cuadro Column, obtendremos dos tablas de contingencia, las que relacionan A×B y A×C. Y si por ejemplo colocamos las variables A y Z en el cuadro Row y las

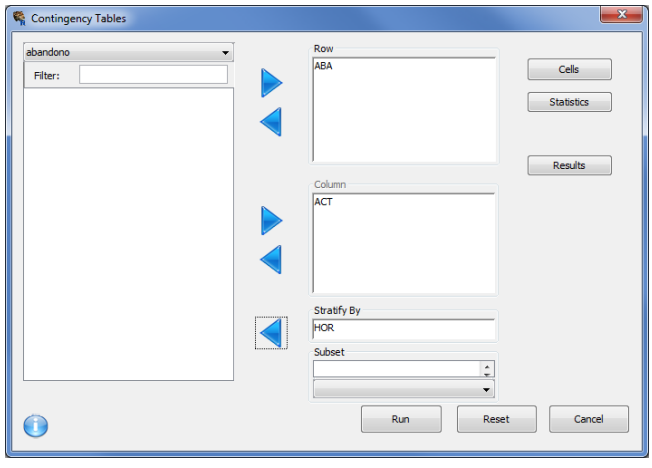
variables B y C en el cuadro Column, obtendremos cuatro tablas de contingencia, las que relacionan A×B, A×C, Z×B y Z×C.

En el caso del ejemplo sobre el abandono de los estudios podemos considerar la petición de dos tablas de contingencia que relacionen ABA×ACT y ABA×HOR. En este caso, y por nuestra convención, estamos considerando la variable ABA como dependiente, y la colocamos en filas, y las variables ACT y HOR como independientes, y las colocamos en columnas. El cuadro de diálogo correspondiente a este análisis sería el siguiente:

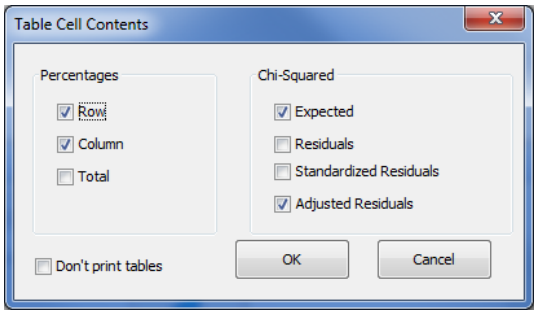


Si quisiéramos obtener además la tabla de contingencia entre ACT y HOR entonces deberíamos ejecutar una segunda vez el procedimiento dado que no hay forma de especificar simultáneamente este cruce con los anteriores.

Si consideramos ahora relaciones entre tres variables (y no más de tres) entonces debemos utilizar el recuadro de Stratify By para trasladar la variable que define la tercera dimensión. Así, por ejemplo, si colocamos las variables A y Z en el cuadro Row, la variable B en el cuadro Column, y la variable C en el cuadro Stratify By, obtendremos dos tablas de contingencia, las que relacionan A×B×C y Z×B×C. Recordemos que la obtención de tablas de contingencia de tres dimensiones significa reproducir tantas tablas y estadísticos bivariables como valores tiene la tercera variable. A continuación se ilustra el caso de la especificación de la tabla tridimensional que cruza ABA×ACT×HOR:



Adicionalmente vemos que podríamos hacer intervenir una cuarta variable si la empleamos para seleccionar un grupo de casos. Para completar las especificaciones del procedimiento de análisis detallaremos las siguientes opciones de los apartados de Cells y Statistics. Consideremos en primer lugar la información que queremos mostrar en las casillas de la tabla de contingencia a partir de su cuadro de diálogo:



Por defecto, si no hacemos ninguna especificación, Deducer nos sacará las frecuencias absolutas observadas y los porcentajes por fila y por columna, si bien interpretaremos los de columna ya que consideraremos, por convención, a la variable colocada en filas como la variable dependiente y la variable de las columnas como la variable independiente. En este caso hemos pedido también las frecuencias esperadas bajo la hipótesis de independencia entre las variables, el valor de la frecuencia esperada y los residuos tipificados corregidos.

Por su parte también pediremos los estadísticos que nos determinan de un lado la existencia de asociación y, por otro, la intensidad de ésta. En el primer caso marcaremos la opción Chi-Squared que nos proporciona la prueba de chi-cuadrado de Pearson. En el segundo caso consideraremos la V de Cramer que ejecutaremos con el lenguaje de comandos a partir del paquete vcd. En estos diferentes análisis solicitaremos igualmente los gráficos de barras para completar la descripción visual de la relación entre las variables[54].

La información de esta tabla de contingencia es la que aparece en el artículo que reseñamos. Mediante el lenguaje de comandos hemos reproducido los datos de los 474 alumnos y las tres variables y los hemos guardado en el archivo Abandono.rda. Las instrucciones has sido estas[55]:

```
# Generación del Data frame a partir de las frecuencias
abandono=data.frame(
  ABA=factor(rep(1:2, c(367,107)), labels=c("No","Sí")),
  ACT=factor(c(rep(1:2, c(245,122)), rep(1:2, c(45,62))),
             labels=c("No","Sí")),
  HOR=factor(c(rep(1:3, c(100,70,75)), rep(1:3, c(17,50,55)),
             rep(1:3, c(10,25,10)), rep(1:3, c(7,20,35))),
             labels=c("Mañana","Tarde","Noche")) )
save(abandono,file='D:/Datos/Abandono.rda')
```

[54] Todos los análisis se han recogido mediante instrucciones de R que se presentan en el *script* Abandono.R que se encuentra en la página web.

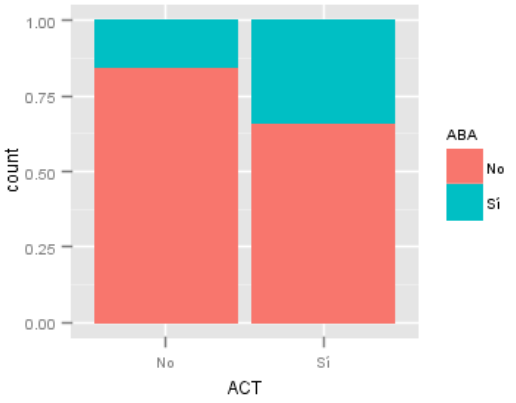
[55] Para algunos de los valores se ha mantenido el acento y se ha empleado la letra ñ. Ejecutando sobre RStudio no hemos encontrado problemas, en Deducer se pueden encontrar en algún momento. Como alternativa se pueden suprimir los acentos y no utilizar la ñ. Si se trabaja en RStudio para que funcione en particular el comando contingency.tables de Deducer es necesario tener instalado y cargar el paquete Deducer desde la consola de RStudio, junto a los paquetes rJava, JGR y ggplot2.

A continuación se presentan los diversos resultados de tablas y gráficos que se obtienen con los procedimientos de tratamiento de las tablas de contingencia y los datos del ejemplo.

Contingency Tables

ABA by ACT across levels of

		ACT		
ABA		No	Sí	Row Total
No	Count	245	122	367
	Row %	66.76%	33.24%	77.43%
	Column %	84.48%	66.30%	
	Expected	224.54	142.46	
	Adj Resid	4.61	-4.61	
Sí	Count	45	62	107
	Row %	42.06%	57.94%	22.57%
	Column %	15.52%	33.70%	
	Expected	65.46	41.54	
	Adj Resid	-4.61	4.61	
Column Total		290	184	**474**
Column %		61.18%	38.82%	



En primer lugar constatamos que la tasa de abandono general de la Facultad es el 22,6% pero que afecta en mayor medida a los que sí trabajan (33,7%) que a los que no trabajan (15,5%), un diferencia del 18,2%. Este comportamiento diferenciado motiva que las frecuencias observadas y esperadas difieran generando un residuo estadísticamente significativo (±4,6 supera el valor ±1,96). Este resultado de diferencia de porcentajes y de residuos nos indica lo que la prueba estadística de chi-cuadrado evidencia[56]: la existencia de una relación estadísticamente significativa que nos permite concluir que las variables están asociadas, con una intensidad del 0,212 según se obtiene con la V de Cramer. Se verifica pues la Hipótesis 1 en un análisis bivariable.

Contingency Table Tests

Tests for ABA by ACT across levels of

	statistic	*df*	*asymptotic p-value*
Chi Squared	20.26	1	6.77e-06

```
> assocstats(table(abandono$ABA, abandono$ACT))
                    X^2 df   P(> X^2)
Likelihood Ratio 20.838  1 4.9976e-06
Pearson          21.284  1 3.9599e-06

Phi-Coefficient   : 0.212
Contingency Coeff.: 0.207
Cramer's V        : 0.212
```

Analizamos a continuación la Hipótesis 2 en un análisis bidimensional. Planteábamos en ella que el abandono sería mayor entre el estudiantado de la tarde o de la noche.

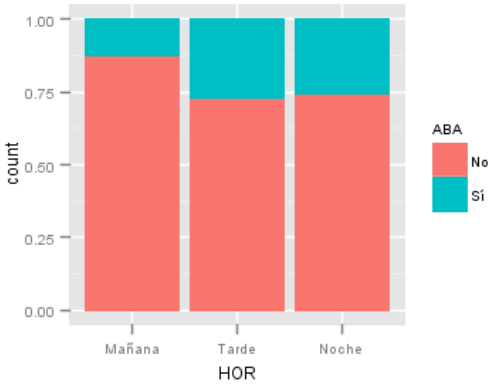
[56] Se presentan dos cálculos de chi-cuadrado, el que ofrece Deducer es el que tiene una corrección de continuidad que se aplica sólo en tablas de 2×2, como es el caso. El paquete vcd lo calcula sin la corrección. En ambos casos la conclusión es la misma.

Los datos muestran que la tasa de abandono efectivamente aumenta en los grupos de clase de la tarde y la noche, frente a un porcentaje del 12,7% de la mañana, el de tarde tiene un valor del 27,3% y similar al de noche con un 25,7%. Estas diferencias en grupos de clase son claramente significativas para el horario de mañana, 12,7% difiere del comportamiento global de 22,6%. Pero en los grupos de tarde y noche, si bien sus tasas de abandono son superiores al promedio, las diferencias son reducidas en relación a la mañana y resultan localmente no significativas estadísticamente (los residuos corregidos son inferiores a 1,96). Por tanto, las casillas de tarde y noche contribuyen a generar asociación y es el grupo de la mañana el que genera la fuente de asociación entre las variables.

Contingency Tables

ABA by HOR across levels of

		HOR			
ABA		Mañana	Tarde	Noche	Row Total
No	Count	117	120	130	367
	Row %	31.88%	32.70%	35.42%	77.43%
	Column %	87.31%	72.73%	74.29%	
	Expected	103.75	127.75	135.50	
	Adj Resid	3.23	-1.79	-1.25	
Sí	Count	17	45	45	107
	Row %	15.89%	42.06%	42.06%	22.57%
	Column %	12.69%	27.27%	25.71%	
	Expected	30.25	37.25	39.50	
	Adj Resid	-3.23	1.79	1.25	
Column Total		134	165	175	**474**
Column %		28.27%	34.81%	36.92%	



Si miramos el comportamiento global de la asociación a través del chi-cuadrado verificamos que se cumple la hipótesis alternativa, estableciendo un nivel de asociación de 0,149:

Contingency Table Tests

Tests for ABA by HOR across levels of

	statistic	*df*	*asymptotic p-value*
Chi Squared	10.57	2	0.00508

```
> assocstats(table(abandono$ABA, abandono$HOR))
                    X^2 df  P(> X^2)
Likelihood Ratio 11.480  2 0.0032154
Pearson          10.567  2 0.0050753

Phi-Coefficient   : 0.149
Contingency Coeff.: 0.148
Cramer's V        : 0.149
```

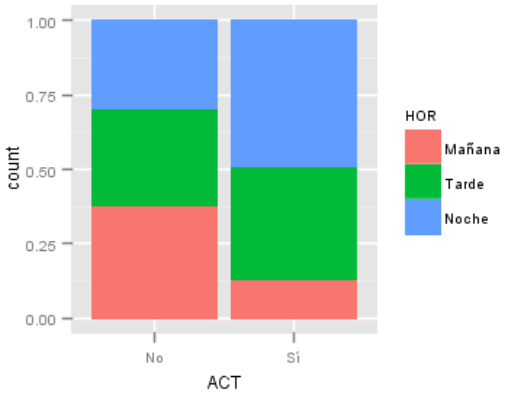
Se comprueba también la Hipótesis 2 en un análisis bivariable.

Analizamos la tercera y última relación entre parejas de variables. En la Hipótesis 3 planteábamos que los que trabajan tienden a matricularse sobre todo por la tarde y noche.

Contingency Tables

ACT by HOR across levels of

ACT		Mañana	Tarde	Noche	Row Total
No	Count	110	95	85	290
	Row %	37.93%	32.76%	29.31%	61.18%
	Column %	82.09%	57.58%	48.57%	
	Expected	81.98	100.95	107.07	
	Adj Resid	5.86	-1.18	-4.31	
Sí	Count	24	70	90	184
	Row %	13.04%	38.04%	48.91%	38.82%
	Column %	17.91%	42.42%	51.43%	
	Expected	52.02	64.05	67.93	
	Adj Resid	-5.86	1.18	4.31	
Column Total		134	165	175	**474**
Column %		28.27%	34.81%	36.92%	



Se observa ante todo que los grupos más numerosos son el de tarde y sobre todo el de noche. Cuando los separamos entre trabajadores y no trabajadores vemos que el porcentaje de estudiantes que trabajan se reduce al 13% en el grupo de la mañana y sube al 38% y al 49% en los de la tarde y la noche, respectivamente. Por tanto se evidencia un comportamiento diferenciado con residuos significativos en la mañana y en la noche (el de tarde no porque tiene un comportamiento cercano al promedio). La relación es significativa según el test de independencia de chi-cuadrado y la V de Cramer arroja valor de asociación de 0,280.

Contingency Table Tests

Tests for ACT by HOR across levels of

	statistic	df	asymptotic p-value
Chi Squared	37.28	2	8.01e-09

```
> assocstats(table(abandono$ACT, abandono$HOR))
                    X^2 df   P(> X^2)
Likelihood Ratio 39.834  2 2.2398e-09
Pearson          37.285  2 8.0117e-09

Phi-Coefficient   : 0.28
Contingency Coeff.: 0.27
Cramer's V        : 0.28
```

Se comprueba igualmente la Hipótesis 3 en un análisis bivariable.

Cuando introducimos la tercera variable podemos elegir distintas alternativas de lectura, en particular, analizar el abandono según la actividad, comparando mañana, tarde y noche, o bien analizar el abandono según el horario controlando por la actividad laboral. Comentamos en nuestro modelo de análisis inicial que el interés estaba en contrastar la hipótesis de hasta qué punto la relación entre abandono y horario era de carácter espúrea (Hipótesis 4). Por ello analizaremos esta relación controlando por actividad para intentar evidenciar que el abandono se debe al hecho de trabajar y que si analizamos solo al estudiantado que trabaja, entre ellos, deben tener tasas de abandono similares, y lo mismo entre los que no trabajan. Veámoslo. La tabla de contingencia y los gráficos de barras que se obtienen son los siguientes:

Contingency Tables

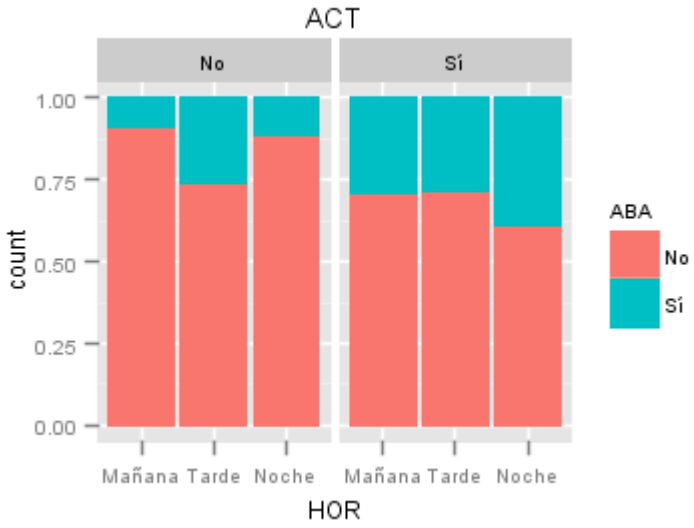
ABA by HOR across levels of ACT

Stratum: ACT = No

ABA		Mañana	Tarde	Noche	Row Total
No	Count	100	70	75	245
	Row %	40.82%	28.57%	30.61%	84.48%
	Column %	90.91%	73.68%	88.24%	
	Expected	92.93	80.26	71.81	
	Adj Resid	2.36	-3.55	1.14	
Sí	Count	10	25	10	45
	Row %	22.22%	55.56%	22.22%	15.52%
	Column %	9.09%	26.32%	11.76%	
	Expected	17.07	14.74	13.19	
	Adj Resid	-2.36	3.55	-1.14	
Column Total		110	95	85	**290**
Column %		37.93%	32.76%	29.31%	

Stratum: ACT = Sí

ABA		Mañana	Tarde	Noche	Row Total
No	Count	17	50	55	122
	Row %	13.93%	40.98%	45.08%	66.30%
	Column %	70.83%	71.43%	61.11%	
	Expected	15.91	46.41	59.67	
	Adj Resid	0.503	1.15	-1.46	
Sí	Count	7	20	35	62
	Row %	11.29%	32.26%	56.45%	33.70%
	Column %	29.17%	28.57%	38.89%	
	Expected	8.09	23.59	30.33	
	Adj Resid	-0.503	-1.15	1.46	
Column Total		24	70	90	**184**
Column %		13.04%	38.04%	48.91%	



La tasa de abandono entre los que trabajan se observan algo superior entre el estudiantado del grupo de noche en relación a la mañana y la tarde. El residuo local es significativo pero globalmente la prueba de chi-cuadrado que relaciona abandono con horario, entre los que sí trabajan, no resulta significativa estadísticamente. Es decir, a pesar de observar ciertas diferencias en la muestra para el grupo de noche debemos considerar que las diferencias se deben al azar y no son extrapolables al conjunto del alumnado. Por tanto, desaparece la relación entre abandono y horario. ¿Y entre los que no trabajan, su comportamiento también se puede considerar similar y podemos concluir que desaparece la relación? Pues no. Si nos fijamos en el gráfico o en la tabla vemos que el grupo de la tarde tiene un comportamiento claramente diferenciado, su tasa de abandono es del 26,3% cuando los que no trabajan de la mañana y de la noche tienen porcentajes del 9,1% y del 11,8%. En este caso el chi-cuadrado confirma que las diferencias son significativas y existe una asociación que la V de Cramer parcial cifra en 0,210[57].

[57] El paquete vcd solamente analiza tablas bidimensionales por lo que en las instrucciones que se adjuntan se ha ejecutado el comando **`assocstats`** seleccionando la submuestra que corresponde a cada categoría de la tercera variable con el comando **`subset`**.

Este resultado nos permite llegar a la conclusión de que existen dos comportamientos diferenciados, el de los que no tienen actividad laboral (existe asociación) y el de los que tienen actividad laboral (desaparece la relación). En consecuencia, al observar dos patrones de comportamiento, la relación original entre abandono y horario varía a cada nivel de la tercera variable, concluimos la existencia de una interacción verificándose el modelo de interacción.

Contingency Table Tests

Tests for ABA by HOR across levels of ACT

Stratum: ACT = No

	statistic	*df*	*asymptotic p-value*
Chi Squared	12.83	2	0.00164

Stratum: ACT = Sí

	statistic	*df*	*asymptotic p-value*
Chi Squared	2.13	2	0.345

```
> summary(subset(abandono,subset=ACT=="No"))
 ABA      ACT          HOR
 No:245   No:290   Mañana:110
 Sí: 45   Sí:  0   Tarde : 95
                   Noche : 85
> ACT1=subset(abandono,subset=ACT=="No")
> assocstats(table(ACT1$ABA, ACT1$HOR))
                    X^2 df  P(> X^2)
Likelihood Ratio 12.216  2 0.0022255
Pearson          12.829  2 0.0016380

Phi-Coefficient   : 0.21
Contingency Coeff.: 0.206
Cramer's V        : 0.21
> summary(subset(abandono,subset=ACT=="Sí"))
 ABA      ACT          HOR
 No:122   No:  0   Mañana:24
 Sí: 62   Sí:184   Tarde :70
                   Noche :90
> ACT2=subset(abandono,subset=ACT=="Sí")
> assocstats(table(ACT2$ABA, ACT2$HOR))
                    X^2 df P(> X^2)
Likelihood Ratio 2.1335  2  0.34412
Pearson          2.1295  2  0.34482

Phi-Coefficient   : 0.108
Contingency Coeff.: 0.107
Cramer's V        : 0.108
```

Para completar el ejercicio analizaremos la tabla de contingencia entre abandono y actividad, controlando por horario, donde tenemos esta información:

Contingency Tables

ABA by ACT across levels of HOR

Stratum: HOR = Mañana

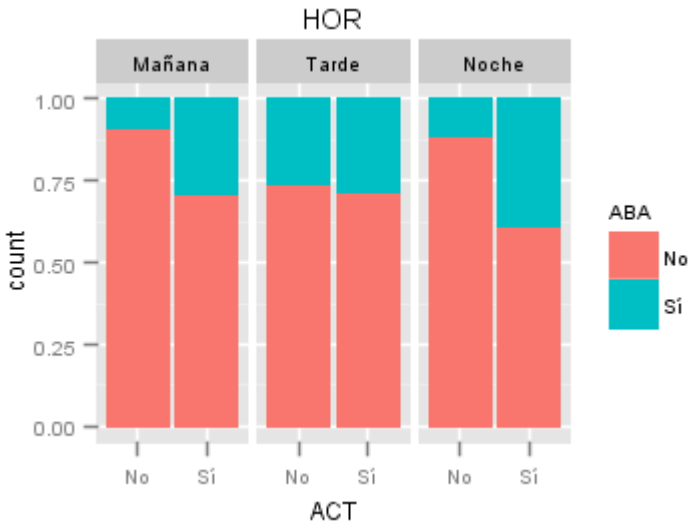
ABA		ACT: No	ACT: Sí	Row Total
No	Count	100	17	117
	Row %	85.47%	14.53%	87.31%
	Column %	90.91%	70.83%	
	Expected	96.04	20.96	
	Adj Resid	2.68	-2.68	
Sí	Count	10	7	17
	Row %	58.82%	41.18%	12.69%
	Column %	9.09%	29.17%	
	Expected	13.96	3.04	
	Adj Resid	-2.68	2.68	
Column Total		110	24	134
Column %		82.09%	17.91%	

Stratum: HOR = Tarde

ABA		ACT: No	ACT: Sí	Row Total
No	Count	70	50	120
	Row %	58.33%	41.67%	72.73%
	Column %	73.68%	71.43%	
	Expected	69.09	50.91	
	Adj Resid	0.322	-0.322	
Sí	Count	25	20	45
	Row %	55.56%	44.44%	27.27%
	Column %	26.32%	28.57%	
	Expected	25.91	19.09	
	Adj Resid	-0.322	0.322	
Column Total		95	70	165
Column %		57.58%	42.42%	

Stratum: HOR = Noche

ABA		ACT: No	ACT: Sí	Row Total
No	Count	75	55	130
	Row %	57.69%	42.31%	74.29%
	Column %	88.24%	61.11%	
	Expected	63.14	66.86	
	Adj Resid	4.10	-4.10	
Sí	Count	10	35	45
	Row %	22.22%	77.78%	25.71%
	Column %	11.76%	38.89%	
	Expected	21.86	23.14	
	Adj Resid	-4.10	4.10	
Column Total		85	90	175
Column %		48.57%	51.43%	

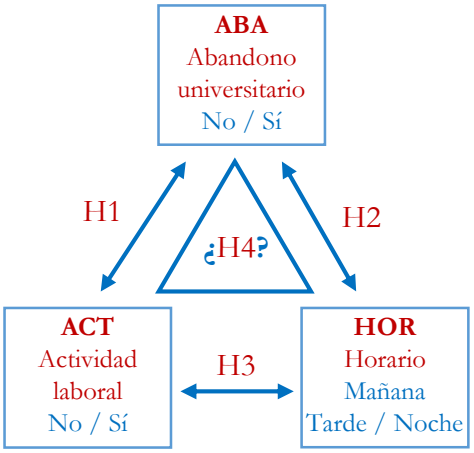


Podemos comprobar que los porcentajes son los mismos que en la tabla trivariable anterior pero se disponen en subtablas diferentes. En este caso la lectura de la información nos dice que entre los de la mañana y los de la noche existen diferencias de abandono según se trabaje o no se trabaje, así lo muestra el test de chi-cuadrado. Pero en el de la tarde las diferencias desaparecen. No tenemos más información para dilucidar qué está pasando en el grupo de tarde, pero sigue una pauta diferente de la esperada que se verifica en el de la mañana y la noche. De esta forma, el comportamiento diferenciado de la tarde está provocando la interacción y que no podamos validar el modelo de independencia condicional.

```
> summary(subset(abandono,subset=HOR=="Mañana"))
 ABA       ACT          HOR
 No:117   No:110   Mañana:134
 Sí: 17   Sí: 24   Tarde :  0
                   Noche :  0
> HOR1=subset(abandono,subset=HOR=="Mañana")
> assocstats(table(HOR1$ABA, HOR1$ACT))
                    X^2 df  P(> X^2)
Likelihood Ratio 5.9486  1 0.0147291
Pearson          7.1683  1 0.0074202

Phi-Coefficient   : 0.231
Contingency Coeff.: 0.225
Cramer's V        : 0.231
```

```
> summary(subset(abandono,subset=HOR=="Tarde"))
 ABA      ACT         HOR
 No:120   No:95   Mañana:  0
 Sí: 45   Sí:70   Tarde :165
                  Noche :  0
> HOR2=subset(abandono,subset=HOR=="Tarde")
> assocstats(table(HOR2$ABA, HOR2$ACT))
                    X^2 df P(> X^2)
Likelihood Ratio 0.10314  1  0.74810
Pearson          0.10338  1  0.74781

Phi-Coefficient   : 0.025
Contingency Coeff.: 0.025
Cramer's V        : 0.025
> summary(subset(abandono,subset=HOR=="Noche"))
 ABA      ACT         HOR
 No:130   No:85   Mañana:  0
 Sí: 45   Sí:90   Tarde :  0
                  Noche :175
> HOR3=subset(abandono,subset=HOR=="Noche")
> assocstats(table(HOR3$ABA, HOR3$ACT))
                   X^2 df   P(> X^2)
Likelihood Ratio 17.656  1 2.6468e-05
Pearson          16.837  1 4.0738e-05

Phi-Coefficient   : 0.31
Contingency Coeff.: 0.296
Cramer's V        : 0.31
```

Dicho lo cual, y como anunciamos al inicio, el hecho de que observemos dos patrones diferentes ¿hasta qué punto es un resultado concluyente estadísticamente en un análisis de tablas de contingencia como el realizado? Los resultados estadísticos parciales de las subtablas no son objeto de contraste entre sí en un análisis clásico de tablas de contingencia por lo que no podemos establecer con certeza un posible modelo de interacción como este:



Nos quedamos con el interrogante que resolveremos en el próximo capítulo con un análisis log-lineal.

► **Ejercicio 18.**
Proponer un modelo de relación entre las variables ACT (actitud: grado de acuerdo con la afirmación "Las mujeres deben quedarse en su casa"), EST (el nivel de estudios) y SEX (el sexo de la persona entrevistada) y contrastar las hipótesis con los datos siguientes de forma similar al ejercicio realizado con el ejemplo del abandono universitario. El archivo de sintaxis ATC-Actitud.R de la página web contiene la sintaxis que genera los datos y obtiene las tablas de contingencia.

Contingency Tables

ACT by EST across levels of SEX

Stratum: SEX = Varón

ACT		EST: Primarios	Secundarios	Superiores	Row Total
Acuerdo	Count	72	110	44	226
Desacuerdo	Count	47	196	179	422
Column Total		119	306	223	**648**

Stratum: SEX = Mujer

ACT		EST: Primarios	Secundarios	Superiores	Row Total
Acuerdo	Count	86	173	28	287
Desacuerdo	Count	38	283	187	508
Column Total		124	456	215	**795**

7. Bibliografía

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, New Jersey: John Wiley & Sons.

Andersen, P. B. (1990). *The Statistical Analysis of Categorical Data*. Berlin: Springer-Verlag.

Aguilera, A. M. (2001). *Tablas de contingencia bidimensionales*. Madrid: La Muralla.

Aguilera, A. M. (2006). *Modelización de tablas de contingencia multidimensionales*. Madrid: La Muralla.

Alvira Martín, F. (1989). Introducción al anàlisis de los datos. En *El análisis de la realidad social. Métodos y técnicas de investigación*, editado por M. García Ferrando, J. Ibáñez i F. Alvira. 2a edición. Madrid: Alianza. Alianza Universidad Textos, 105, 325-358.

Ato, M.; López, J. J. (1996). *Análisis estadístico para datos categóricos*. Madrid: Síntesis.

Bardina, X.; Farré, M.; López-Roldán, P. (2005). *Estadística: un curs introductori per a estudiants de ciències socials i humanes. Volum 2: Descriptiva i exploratòria bivariant. Introducció a la inferència*. Bellaterra (Cerdanyola del Vallès): Servei de Publicacions de la Universitat Autónoma de Barcelona.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187 (4175), 398-404. http://www.unc.edu/~nielsen/soci708/cdocs/Berkeley_admissions_bias.pdf

Bishop, Y. M.; Fienberg, S. E.; Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge: Cambridge University Press.

Blalock, H. M. Jr. (1981). *Estadística Social*. México: Fondo de Cultura Económica.

Blyth, C.R. (1972), On Simpson's Paradox and the Sure-Thing Principle. *Journal of the American Statistical Association*, 67 (338), 364-366.

Bryman, A.; Cramer, D. (1990). *Quantitative Data Analysis for Social Scientist*. London: Routledge.

Boudon, R.; Lazarsfeld, P. (1985). *Metodología de las ciencias sociales. II. Análisis empírico de la causalidad*. Barcelona: Laia.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, 23 (3), 315-345.

Davis, J. (1980). Contingency tables analysis: proportions and flow graphs. *Quality & Quantity*, 14, número especial, 117-153.

Everitt, B. S. (1992). *The Analysis of Contingency Tables*. London: Chapman and Hall.

Fachelli, S.; López-Roldán, P. (2013). *Análisis de datos estadísticos. Análisis de movilidad social*. Bellaterra: Universitat Autònoma de Barcelona. http://ddd.uab.cat/record/88747

Friendly. M. (2000). *Visualizing Categorical Data*. SAS Insitute, Carey, NC. http://www.math.yorku.ca/SCS/vcd/

Friendly. M. (2013). Working with categorical data with R and the vcd and vcdExtra packages. http://cran.us.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*. Cambridge: MIT Press.

García Ferrando, M. (1987). *Socioestadística. Introducción a la estadística en sociología*. 2a edición amp. Madrid: Alianza. Alianza Universidad Textos, 96.

Goodman, L. A.; Kruskal, W. H. (1954). Measures of Association for cross-clessifications. *Journal of American Statistical Association*, 49, 732-764.

Goodman, L. A. (1963). On methods for comparing contingency tables. *The Journal of the Royal Statistical Society, series A*, 126.

Goodman, L. A. (1972). A General Model for the Analysis of Surveys. *American Journal of Sociology*, 77, 6, 1035-1086.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika, 60*, 179–192.

Goodman, L. A. (2011). Measures, Models, and Graphical Displays in the Analysis of Cross-Classified Data. *Journal of the American Statistical Association*, 86 (416), 1085–1111.

Haberman, S. J. (1979). *Analysis of Qualitative Data*. New York: Academic Press.

Hellevik, O. (1988). *Introduction to Causal Analysis. Exploring Survey Data by Crosstabulation*. Oslo: Norwegian University Press.

Hildebrand, D. K.; Laing, J. D.; Rosenthal, H. (1977). *Analysis of ordinal data*. Beverly Hills: Sage Publications.

Kateri, M. (2014). *Contingency Table Analysis: Methods and Implementation Using R*. New York: Springer. http://cta.isw.rwth-aachen.de/

Latiesa, M. (1991a). El análisis multivariable de tablas de contingencia: sistema de ecuaciones y grafos. *Papers. Revista de Sociologia*, 37, 77-96. http://ddd.uab.cat/pub/papers/02102862n37/02102862n37p77.pdf

Latiesa, M. (1991b). Introducción a los modelos logarítmicos lineales. *Papers. Revista de Sociologia,* 37, 97-112. http://ddd.uab.cat/pub/papers/02102862n37/02102862n37p97.pdf

Liebetrau, A. M. (1983). *Measures of Association*. Beverly Hills: Sage Publications.

López-Roldán, P.; Lozares Colina, C. (1999). *Anàlisi bivariable de dades estadístiques*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona. Colección Materials, 79.

Powers, D.; Xie, Y. (2008). *Statistical Methods for Categorical Data Analysis*. Bingley (UK): Emeral. 2a. edición.

Reynolds, H. T. (1977). *The Analysis of Cross-classifications*. New York: Free-Press.

Reynolds, H. T. (1984). *Analysis of Nominal Data*. Sage Publications, Beverly Hills.

Rudas, T. (1998). *Odds ratios in the analysis of contingency tables*. Thousand Oaks: Sage.

Ruiz-Maya, L. et al. (1991). *Metodología estadística para el análisis de datos cualitativos*. Madrid: Centro de Investigaciones Sociológicas.

Ruiz-Maya, L. et al. (1995). *Análisis estadístico de encuestas: datos cualitativos*. Madrid: Editorial AC.

Sánchez Ramos, M. A. (2005). Uso metodológico de las tablas de contingencia en la ciencia política. *Espacios Públicos*, 8 (16), agosto, 60-84.

Sánchez Carrión, J. J. (1984). Análisis de Tablas de Contingencia: Sistema de las Diferencias de Proporciones (Exégesis del trabajo de James A. Davis). En *Introducción a les tècniques de anàlisi multivariable aplicadas a las ciencias sociales*, editado per J. J. Sánchez Carrión. Madrid: Centro de Investigaciones Sociológicas, 295-321.

Sánchez Carrión, J. J. (1989a). Técnicas de análisis de datos nominales. *Revista Española de Investigaciones Sociológicas*, 45, enero-marzo, 133-155. http://www.reis.cis.es/REIS/PDF/REIS_045_08.pdf

Sánchez Carrión, J. J. (1989b). *Análisis de tablas de contingencia. El uso de los percentatges en ciencias sociales*. Madrid: Centro de Investigaciones Sociológicas-Siglo XXI.

Sánchez Carrión, J. J. (1999*). Manual de análisis de datos*. Madrid: Alianza.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society, Series B*, 13, 238–241.

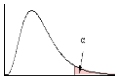
Schneiter, K.; Symanzik, J. (2013). An Applet for the Investigation of Simpson's Paradox. *Journal of Statistics Education*, 21 (1). http://www.amstat.org/publications/jse/v21n1/schneiter.pdf

Wainer, H. (1986). Minority Contributions to the SAT Score Turnaround: An Example of Simpson's Paradox. *Journal of Educational Statistics*, 11 (4), 239-244.

Upton, G. (1978). *The Analysis of Cross-Tabulated Data*. New York: John Wiley.

Yule, G.U. (1903), Notes on the Theory of Association of Attributes in Statistics. *Biometrika*, 2 (2), 121-134.

Anexo. Tabla de distribución teórica de Chi-cuadrado (χ^2)



	Probabilidades α (*áreas a la derecha del valor crítico*)									
Grados de libertad	**0,995**	**0,99**	**0,975**	**0,95**	**0,90**	**0,10**	**0,05**	**0,025**	**0,01**	**0,005**
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	40,256	43,773	46,979	50,892	53,672
40	20,707	22,164	24,433	26,509	29,051	51,805	55,758	59,342	63,691	66,766
50	27,991	29,707	32,357	34,764	37,689	63,167	67,505	71,420	76,154	79,490
60	35,534	37,485	40,482	43,188	46,459	74,397	79,082	83,298	88,379	91,952
70	43,275	45,442	48,758	51,739	55,329	85,527	90,531	95,023	100,425	104,215
80	51,172	53,540	57,153	60,391	64,278	96,578	101,879	106,629	112,329	116,321
90	59,196	61,754	65,647	69,126	73,291	107,565	113,145	118,136	124,116	128,299
100	67,328	70,065	74,222	77,929	82,358	118,498	124,342	129,561	135,807	140,169