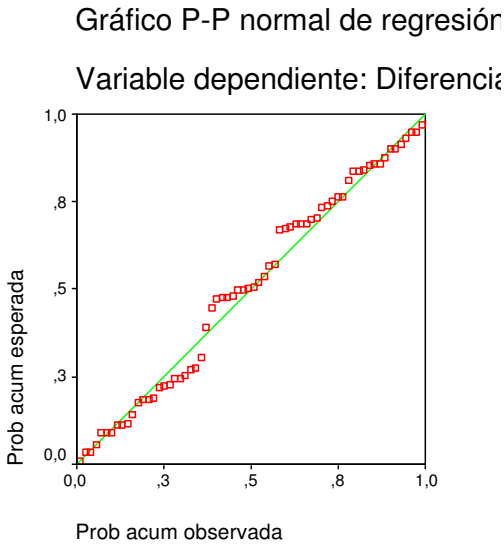


Capítulo 7: Independencia de variables categóricas

Las tablas de contingencia se utilizan para examinar la relación entre dos variables categóricas, o bien explorar la distribución que posee una variable categórica entre diferentes muestras.

Hay diferentes cuestiones que surgen al examinar una tabla de contingencia, y en este tema vamos a tratar la cuestión de la independencia.

La independencia de dos variables consiste en que la distribución de una de las variables es similar sea cual sea el nivel que examinemos de la otra. Esto se traduce en una tabla de contingencia en que las frecuencias de las filas (y las columnas) son aproximadamente proporcionales. Posiblemente sea más cómodo reconocerlo usando en la tabla de contingencias los porcentajes por filas (o columnas) y observando si estos son similares.

La prueba de independencia ji-cuadrado (chi-cuadrado) contrasta la hipótesis de que las variables son independientes, frente a la hipótesis alternativa de que una variable se distribuye de modo diferente para diversos niveles de la otra.

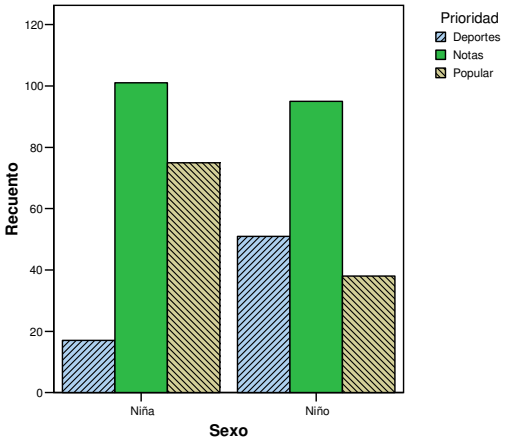
Observe la siguiente tabla, en la que en un estudio con escolares de 10 a 12 años se les preguntó a qué daban más prioridad de entre tres posibilidades: Tener buenas notas, destacar en los deportes o ser popular entre los compañeros.

Tabla de contingencia Sexo * Prioridad

Recuento

		Prioridad			
		Deportes	**Notas**	**Popular**	Total
Sexo	**Niña**	17	101	75	**193**
	Niño	51	95	38	**184**
Total		**68**	**196**	**113**	**377**

Gráfico de barras



Con un poco de atención se observa con facilidad (ya que los grupos de chicos y chicas son de tamaños similares) que ambos sexos valoran de manera aproximadamente similar las notas. Donde más diferencia se observa entre los sexos es en la preferencia que muestran muchos chicos por los deportes y muchas chicas por la popularidad.

Si prestamos atención a la distribución de las prioridades en porcentajes para cada sexo, tal vez la diferencia sea más evidente:

Tabla de contingencia Sexo * Prioridad

% de Sexo

		Prioridad		
		Deportes	Notas	Popular
Sexo	Niña	9%	52%	39%
	Niño	28%	52%	21%

La prueba de ji-cuadrado de Pearson contrasta si las diferencias observadas entre los dos grupos son atribuibles al azar. En este caso se obtiene una significación cercana al 0%, con lo que para al nivel de significación habitual del 5%, se rechaza la hipótesis de independencia de las prioridades de los estudiantes y el sexo (las preferencias no se distribuyen del mismo modo entre chicos y chicas).

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	29,100[a]	2	,000

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 33,19.

7.1 Limitaciones de la prueba de independencia

El contraste de independencia tiene muy pocas limitaciones, aunque es conveniente hacer algunas observaciones:

- Para contrastar la independencia se suele usar el estadístico ji-cuadrado de Pearson. Su cálculo se basa en calcular la diferencia entre las observaciones

observadas para cada par de modalidades de las variables (casillas), y las que serían de esperar en caso de que se satisficiese la condición de independencia. Para que se pueda considerar correcta la significación calculada por el estadístico ji-cuadrado de Pearson, se debe cumplir que las frecuencias esperadas no sean muy pequeñas (inferiores a 5) más que en unas pocas casillas. Si es en muchas las casillas donde esto ocurre (más del 20% por ejemplo) se debe usar una prueba que no incluya aproximaciones, como la prueba exacta de Fisher. Esta la ofrece cualquier programa como opción cuando se hace este tipo de contrastes.

- Si las muestras son muy grandes, la prueba de independencia dará resultados significativos incluso donde, posiblemente, consideremos que las diferencias no sean en realidad clínicamente interesantes. Es conveniente una inspección visual para confirmar si las diferencias observadas por filas (o columnas, como prefiramos), nos parecen de interés.
- Si las variables poseen muchos niveles posiblemente la prueba no resulte de mucho interés, ya que es lógico esperar que se encuentren diferencias. Eso ocurre si por ejemplo una de las variables es numérica y no hemos agrupado los posibles valores en una cantidad adecuada de intervalos.
- Si una de las variables es numérica u ordinal, posiblemente queramos hacer algo más que contrastar la simple independencia. Después de todo, esto no es tan informativo como saber que en cierto grupo los valores son significativamente mayores que en otro. Lo aconsejable es usar pruebas de tipo t-student, anova o contrastes no paramétricos comos los que se tratan en otros temas.
- El contraste de ji-cuadrado sirve para contrastar la independencia. No hay que considerarla como una medida de la asociación entre variables. Si buscamos estudiar la asociación de variables tenemos otros métodos a nuestra disposición como *regresión logística* que trataremos más adelante.

Ejemplo: Se tienen datos demográficos de más de 130.000 individuos. De ellos se conoce la edad y el nivel de estudios. Se desea contrastar si el nivel de estudios de la población es similar para los individuos de diferentes edades. La sospecha es que en los individuos más jóvenes, el nivel de estudios es superior. Seguramente una prueba ANOVA o un modelo de regresión serían más convenientes, incluyendo posiblemente algunas variables explicatorias, pero vamos a intentar adaptarnos a una prueba ji-cuadrado sobre una tabla de contingencia.

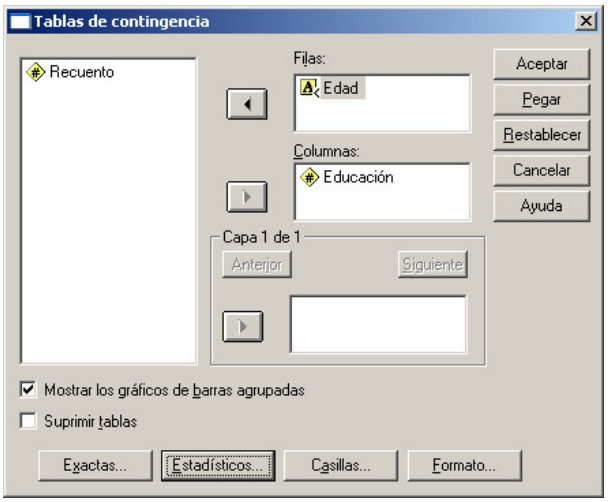
Para adaptar la cuestión a una prueba de ji-cuadrado, recodificaremos la variable nivel de estudios en cuatro categorías que nos simplifiquen la variedad de estudios que podría haber disponibles.

En cuanto a las edades, las simplificamos en 5 grupos. Los grupos no tienen porqué comprender rangos de edades iguales, simplemente deben ser reveladores para nosotros. Por ejemplo, se excluyen de la muestra los individuos de menos de 25 años pues se considera que muchos pueden no haber terminado su formación. Después se forman grupos con un rango de 10 años, pues consideramos que la distribución del nivel de formación los individuos en esos márgenes tan estrechos es razonablemente homogéneo para cualquier nivel de edad del mismo.

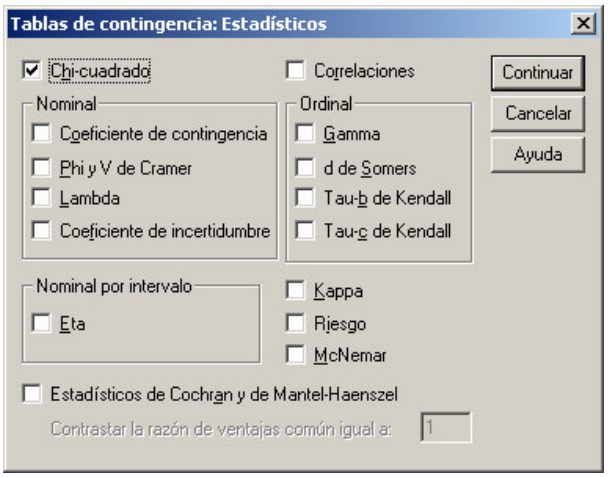
En lugar de contrastar la hipótesis original de que en las submuestras de individuos más jóvenes ocurre con mayor frecuencia un mayor nivel educativo, la debilitamos prescindiendo del orden implícito que hay en la variable edad y en el nivel educativo. Simplemente contrastamos si ambas variables son independientes, es decir, si el nivel educativo se distribuye de la misma manera en cualquier grupo de edad.

Es importante observar que esta es una gran debilitación de la hipótesis original. Al adaptar la hipótesis a una prueba de independencia hemos "ignorado" información relativa al orden en nuestras variables.

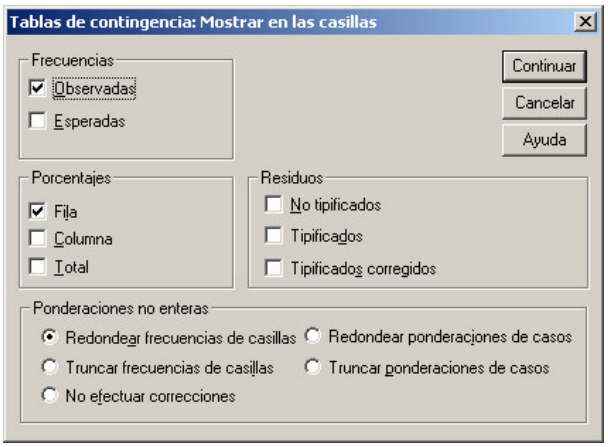
En SPSS elegimos la opción del menú "Analizar - Estadísticos descriptivos - Tablas de contingencia…" y situamos una de las variables en filas y otra en columnas para crear la tabla de contingencia.



Pulsamos el botón "Estadísticos…" para marcar que queremos realizar la prueba ji-cuadrado de Pearson.



Pulsando en el botón "Casillas…" podemos elegir ver no sólo las frecuencias observadas sino los porcentajes por filas.

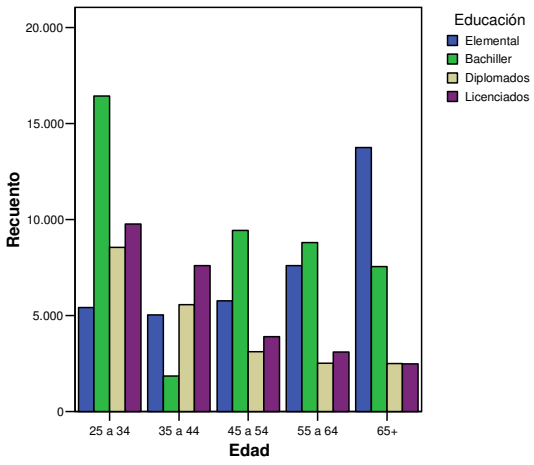


Observando detenidamente los porcentajes por filas no es difícil constatar que la distribución del nivel educativo no es la misma en los diferentes grupos de edad.

Tabla de contingencia Edad * Educación

			Educación				
			Elemental	Bachiller	Diplomados	Licenciados	Total
Edad	25 a 34	Recuento	5416	16431	8555	9771	40173
		% de Edad	13%	41%	21%	24%	100,0%
	35 a 44	Recuento	5030	1855	5576	7596	20057
		% de Edad	25%	9%	25%	38%	100,0%
	45 a 54	Recuento	5777	9435	3124	3904	22240
		% de Edad	26%	42%	14%	18%	100,0%
	55 a 64	Recuento	7606	8795	2524	3109	22034
		% de Edad	35%	40%	11%	14%	100,0%
	65+	Recuento	13746	7558	2503	2483	26290
		% de Edad	52%	29%	10%	9%	100,0%
Total		Recuento	37575	44074	22282	26863	130794
		% de Edad	29%	34%	17%	21%	100,0%

Gráfico de barras



La prueba de significación de ji-cuadrado, nos indica que la significación es cercana a cero, como era de esperar[6].

[6] Incluso aunque la diferencia hubiese sido mínima entre los grupos de edad, la prueba hubiese dado inevitablemente significativa. Con muestras tan grandes todo sale siempre significativo. Así que no es

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	22373,566[a]	12	,000
Razón de verosimilitud	23365,060	12	,000
N de casos válidos	130794		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 3416,90.

Hay una nota al pie de la última tabla que nos indica que todas las casillas de la tabla de contingencia cumplen la condición de validez. Si no se hubiese cumplido deberíamos haber pulsado el botón "Exactas…" para que el cálculo de significación se realice exactamente.

cierto como mucha gente pueda sospechar que para contrastar hipótesis sea bueno tener muestras grandes. Lo ideal es que no sean ni grandes ni pequeñas. Deben tener un tamaño "razonable" para que la diferencia que pretendemos contrastar sea observable y a la vez no achacable al azar.