

SEMINARIO DE DOCTORADO

**METODOLOGÍA DE
INVESTIGACIÓN SOCIAL**

**Agustín Salvia
Santiago Poy**

MÓDULO 2A

**ANÁLISIS DE TABLAS DE CONTINGENCIA Y
COEFICIENTES DE ASOCIACIÓN**

Temas del curso

- Formulación de hipótesis bivariadas
- Elaboración y análisis de tablas de contingencia.
- Cálculo de porcentajes
- Concepto de asociación
- Prueba de hipótesis de independencia estadística y su interpretación
- Medidas de asociación global y local

TÉCNICAS DE ANÁLISIS ESTADÍSTICO

ESTADÍSTICA DESCRIPTIVA
(Medidas de tendencia central y dispersión)

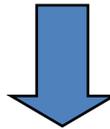
TABLA DE CONTINGENCIA Y ANÁLISIS PORCENTUALES
(Diferencias porcentuales, perfiles y segmentos)

ANÁLISIS DE ASOCIACIÓN
(Correlación y coeficientes de asociación)

TECNICAS DE ANALISIS MULTIVARIADO
(Modelos Lazarsfeld, regresiones, factoriales, clusters)

DE LAS TABLAS DE CONTINGENCIA AL ANÁLISIS DE ASOCIACIÓN MULTIVARIADO

**¿CÓMO ANALIZAR Y EVALUAR HIPÓTESIS
CAUSALES O DE COVARIACIÓN ENTRE
VARIABLES CUANDO LAS MISMAS ESTÁN
MEDIDAS EN ESCALA ORDINAL O NOMINAL?**



**ANÁLISIS DE TABLAS DE CONTINGENCIA
TEST DE SIGNIFICANCIA NO PARAMÉTRICOS
MEDIDAS DE ASOCIACIÓN**

Tablas de contingencia

- Hasta ahora, nos centramos en el análisis detallado de una sola variable (estadística descriptiva univariada).
- Sin embargo, es habitual que nuestro interés sea analizar la relación entre dos o más variables. Precisamente, las **hipótesis** son, generalmente, proposiciones que ponen en relación dos o más variables. Por ejemplo: la proposición “**cuanto mayor sea el grado de satisfacción con su puesto de trabajo, las personas experimentarán mayores niveles de bienestar subjetivo**”, es una hipótesis.
- Los elementos a considerar son:
 - Las **variables** (y su nivel de medición): en este caso, el “grado de satisfacción con el puesto de trabajo” (ordinal) y el “nivel de bienestar subjetivo” (ordinal).
 - La **función de las variables**: estamos asumiendo que el “grado de satisfacción” *provoca* (variable independiente) un mayor “bienestar subjetivo” (variable dependiente).
 - La **unidad de análisis**, en este caso, las personas.
 - El **sentido o dirección** de la relación entre las variables: en este caso, un sentido positivo (mayores valores en la variable independiente → mayores valores en la dependiente).
 - La **consecuencia observacional**: es decir, qué deberíamos encontrar si la hipótesis se comprueba. Aquí, deberíamos observar que las personas más satisfechas con su trabajo tienen mayores niveles de bienestar subjetivo.

Tablas de contingencia

- Una **tabla de contingencia** es una tabla de frecuencias de la distribución conjunta de dos variables *cualitativas* (nominales u ordinales).
- Una tabla de contingencia determina un **espacio de propiedades** de dimensión $n \cdot m$ (filas x columnas), es decir, dependiendo del número de categorías de las variables.
- En general, en una tabla de contingencia las variables cumplen alguna de estas **funciones**:

VARIABLE DEPENDIENTE (Y)

Aquella cuyo comportamiento hay que explicar

VARIABLE INDEPENDIENTE (X)

Es el factor que influye en la variable a explicar

VARIABLE INTERVINIENTE (T)

Se emplea para mantener constante un factor

Tablas de contingencia

UNA TABLA DE CONTINGENCIA ES UNA DISTRIBUCIÓN EN FILAS Y COLUMNAS EN LA QUE LOS INDIVIDUOS DE UNA POBLACIÓN SE CLASIFICAN EN FUNCIÓN DE PARES DE OBSERVACIONES.

La tabla de contingencia es un método de representar simultáneamente dos características diferentes observadas en una misma población. Las dos variables son X e Y . Las categorías de x se escribirán $X.1, X.2.. X.n$, y las de y , se escribirán $Y.1, Y.2... Y.n$. *Los individuos $X.1 Y.1$ son los que reúnen ambos atributos.* Estos valores en una tabla de doble entrada:

X Y	X.1	X.2	...	X.n
Y.1	X.1 Y.1			
Y.2				
...				
Y.n				

Construcción de tablas de contingencia

- **Ejemplo:** Trabajamos con las variables: (1) lectura de diarios impresos y (2) grupo de edad.
- Hipótesis: *“A mayor edad, mayor será la propensión a leer diarios impresos”*.

Cuadro 1. Lectura de diarios impresos según grupo de edad.
Población de 18 años y más, residente en el área Gran Buenos Aires, año 2016. En porcentajes.

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No				
Sí				
Total	(n)	(n)	(n)	(N)

Fuente: Elaboración propia.

- De manera convencional, la variable independiente (X) se coloca en las *columnas* y la variable dependiente (Y) se coloca en *filas*.

Construcción de tablas de contingencia

NUMERO DEL CUADRO

TITULO

Cuadro 1. Lectura de diarios impresos según grupo de edad.
Población de 18 años y más residente en el área Gran Buenos Aires, año 2016. En porcentajes.

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No				
Sí				
Total	(n)	(n)	(n)	(N)

Fuente: Elaboración propia.

FUENTE

Componentes de un cuadro

1. Número del Cuadro

2. Título

¿Cómo dar título a un Cuadro? Primero nombramos a la **variable dependiente**, luego a la **variable independiente**. En el Cuadro de ejemplo: “Consumo de diario impreso según Grupo de edad”. También es común agregar el **universo** al que remite el Cuadro. Así: “Población de 18 años y más residente en el Área Gran Buenos Aires”.

3. Nota superior: aparece debajo del título, y permite entender mejor el Cuadro. Brinda información adicional sobre la población, o bien sobre las unidades (pesos, dólares, porcentajes) en que se encuentra la información.

4. Notas al pie: brindan la información adicional de cualquier tipo de información que sea necesario agregar.

5. Fuente

Tipo de análisis que permite una tabla de contingencia

- Veamos la siguiente tabla:

Cuadro 1. Lectura de diarios impresos según grupo de edad.
Población de 18 años y más residente en el área Gran Buenos Aires, año 2016. Frecuencias absolutas.

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No	152	163	71	386
Sí	29	80	56	165
Total	181	243	127	551

Frecuencias
condicionales

Frecuencias
marginales

Frecuencias
marginales

Fuente: Elaboración propia.

- Para posibilitar la interpretación, debemos calcular **porcentajes**. De acuerdo con cómo se calculen los porcentajes vamos a obtener información diferente sobre la relación entre las variables X e Y.
- Hay tres formas de calcular los porcentajes: por **columna**, por **fila** y a nivel de **total**.

Tipo de análisis que permite una tabla de contingencia

- Para posibilitar la interpretación, debemos calcular *porcentajes*. De acuerdo con cómo se calculen los porcentajes vamos a obtener información diferente sobre la relación entre las variables X e Y.
- Hay tres formas de calcular los porcentajes: por *fila*, a nivel de *total* y por *columna*.
- Cada una de estas formas nos dice algo diferente:
 - *Fila*: análisis de perfiles o características poblacionales.
 - *Total*: análisis descriptivo de grupos o segmentos de población.
 - *Columna* : análisis de asociación entre variables.

Cálculo de porcentajes: por columna

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No	152	163	71	386
Sí	29	80	56	165
Total	181	243	127	551

% de 18 a 30 que lee diarios:
 $29/181*100=16\%$

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No	84,0	67,1	55,9	70,1
Sí	16,0	32,9	44,1	29,9
Total	100,0	100,0	100,0	100,0
	(181)	(243)	(127)	(551)

Fuente: Elaboración propia.

- Por regla general, el cálculo de los porcentajes suele hacerse en la dirección de la variable independiente. Es decir, con “100%” en cada una de las columnas.
- Es la lectura que hacemos cuando buscamos comprobar la hipótesis.

Cálculo de porcentajes: por fila

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No	152	163	71	386
Sí	29	80	56	165
Total	181	243	127	551

Fuente: Elaboración propia.

% de los que leen diarios que tienen 31 a 45 años:
 $80/165 * 100 = 48,5\%$

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No	39,4	42,2	18,4	100,0
Sí	17,6	48,5	33,9	100,0
Total	32,8	44,1	23,0	100,0

Fuente: Elaboración propia.

- La lectura por filas nos dice otra cosa: cómo se “compone” la población en términos de la variable independiente.

Cálculo de porcentajes: total

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No	152	163	71	386
Sí	29	80	56	165
Total	181	243	127	551

Fuente: Elaboración propia.

Lectura de diarios impresos	Grupo de edad			Total
	18 a 30	31 a 45	46 y más	
No	27,6	29,6	12,9	70,1
Sí	5,3	14,5	10,2	29,9
Total	32,8	44,1	23,0	100,0

% del total que son jóvenes y que leen diarios
 $29/551 * 100 = 5,3\%$

- La lectura sobre el total nos dice cómo está compuesta nuestra población en términos del cruce de las dos variables.
- En general, esto es útil cuando las dos variables nos ayudan a construir una **tipología**. Por ejemplo: “jóvenes desinformados”, “jóvenes informados”, etc.

Asociación entre variables: concepto

- “Cuando establecemos una clasificación cruzada de dos variables, nuestro interés se centra (...) en el conocimiento de la forma en que se distribuye la variable dependiente para las diferentes categorías de la variable independiente o causal. La forma en que se relacionan dos variables se denomina **asociación entre dos variables**” (García Ferrando, p. 217)
- Se puede caracterizar la relación entre dos variables mediante el estudio de las siguientes características:
 1. **Existencia:** ¿hay asociación o hay **independencia** estadística?
 2. **Fuerza:** ¿la asociación es fuerte o es débil?
 3. **Dirección:** ¿la asociación es positiva o negativa?
 4. **Naturaleza:** ¿la asociación es lineal, curvilínea o rinconal?

Análisis de asociación: 1) existencia

- Hablamos de **existencia** de asociación cuando la distribución de una variable difiere de algún modo entre las diversas categorías de la segunda variable. Su determinación requiere de la estimación de los porcentajes en la forma apropiada (por columna).
- La asociación se manifiesta en que las distribuciones porcentuales condicionales difieren en mayor o menor grado entre sí. Si no ocurre, hay **independencia estadística**.

HAY ASOCIACIÓN

Cuadro 1. Lectura de diario impreso según grupo de edad. Población de 18 años y más residente en el área Gran Buenos Aires, año 2016. En porcentajes.

Lectura de diarios impresos	Grupo de edad	
	Jóvenes	Adultos
No	84,0%	63,2%
Sí	16,0%	36,8%
Total	100,0%	100,0

Fuente: Elaboración propia.

NO HAY ASOCIACIÓN

Cuadro 1. Lectura de diario impreso según grupo de edad. Población de 18 años y más residente en el área Gran Buenos Aires, año 2016. En porcentajes.

Lectura de diarios impresos	Grupo de edad	
	Jóvenes	Adultos
No	84,0%	84,0%
Sí	16,0%	16,0%
Total	100,0%	100,0%

Fuente: Elaboración propia.

Análisis de asociación: 1) existencia

- ✓ La forma más intuitiva de entender el concepto de asociación en el caso de tablas de contingencia de 2x2 (es decir, con variables de dos categorías) es comparar los porcentajes.
- ✓ Así se llega al más sencillo de todos los *coeficientes de asociación* que miden la existencia de relación entre dos variables.
- ✓ El *épsilon* (ϵ) son las diferencias porcentuales calculadas en la dirección contraria a la que se han realizado los porcentajes.
- ✓ En el ejemplo anterior, obtendríamos épsilon del siguiente modo:

$$\epsilon = \% \text{ de adultos que leen diarios} - \% \text{ jóvenes}$$

$$\epsilon = 36,8\% - 16\% = \mathbf{20,8 \text{ pp.}}$$

- ✓ Notar que la medida es simétrica: no importa cómo lo calculemos, nos interesa el valor absoluto.

La diferencia entre dos porcentajes se expresa en puntos porcentuales

Análisis de asociación: 1) existencia

- La lectura de los porcentajes es una forma intuitiva de evaluar la existencia de asociación. Sin embargo, no nos informa si es *significativa* desde el punto de vista estadístico.
- Al trabajar con datos de una muestra, nos interesa saber si la asociación observada existe en la población estudiada o se debe al azar (es decir, al hecho de haber tomado una muestra y no otra).
- Debemos acompañar la lectura de porcentajes con una *prueba de hipótesis* de independencia estadística mediante el test de *chi-cuadrado* (χ^2).

Chi cuadrado χ^2 : frecuencias observadas y esperadas

- La prueba de χ^2 compara la **distribución observada** de X e Y con la **distribución esperada** en caso de que no hubiera asociación (bajo el supuesto de **independencia** estadística).
- Es fácil imaginarse cómo sería la tabla de contingencia en caso de que no hubiera asociación: la variable dependiente no cambiaría en las distintas columnas de la variable independiente.

Frecuencias observadas

Lectura de diarios impresos	Grupo de edad		Total
	Jóvenes	Adultos	
No	152	234	386
Sí	29	136	165
Total	181	370	551

Frecuencias esperadas (bajo H de independencia)

Lectura de diarios impresos	Grupo de edad		Total
	Jóvenes	Adultos	
No	127	259	386
Sí	54	111	165
Total	181	370	551

$$f_e = \frac{\text{total de fila} * \text{total columna}}{\text{total}}$$

$$\frac{386 * 181}{551} = 127$$

Chi cuadrado χ^2

- Por lo tanto, el test de χ^2 opera como un clásico testeo de hipótesis:
 - Nuestra **hipótesis nula** H_0 (que esperamos rechazar) es *que ambas variables son independientes*. Lógicamente, la **hipótesis alternativa** H_A es el complemento: *ambas variables no son independientes (están asociadas)*.
 - Se calcula el valor del estadístico (χ^2).
 - Se determina la probabilidad del estadístico
 - Se concluye sobre la aceptación o rechazo.

Chi cuadrado χ^2

- Para obtener χ^2 trabajamos sobre las frecuencias observadas y las frecuencias esperadas en caso de no asociación. Siguiendo la fórmula:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Tabla cruzada Lectura de diarios impresos*Grupo de edad

			Grupo de edad		Total
			Jóvenes	Adultos	
Lectura de diarios impresos	No	Recuento	152	234	386
		Recuento esperado	126,8	259,2	386,0
	Sí	Recuento	29	136	165
		Recuento esperado	54,2	110,8	165,0
Total	Recuento	181	370	551	
	Recuento esperado	181,0	370,0	551,0	

$$\chi^2 = \frac{(152 - 126,8)^2}{126,8} + \frac{(234 - 259,2)^2}{259,2} + \frac{(29 - 54,2)^2}{54,2} + \frac{(136 - 110,8)^2}{110,8}$$

$$\chi^2 = 5,0 + 2,45 + 11,7 + 5,73$$

$$\chi^2 = 24,9$$

χ^2 no está estandarizado.
Asumirá un valor más grande cuantos más efectivos tenga la tabla

Prueba de independencia estadística χ^2

Tabla cruzada Lectura de diarios impresos*Grupo de edad

% dentro de Grupo de edad

		Grupo de edad		Total
		Jóvenes	Adultos	
Lectura de diarios impresos	No	84,0%	63,2%	70,1%
	Sí	16,0%	36,8%	29,9%
Total		100,0%	100,0%	100,0%

Interpretamos la significación del coeficiente:

Si $p < 0,05 \rightarrow$ rechazamos hipótesis nula de no asociación

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	24,909 ^a	1	,000		
Corrección de continuidad ^b	23,930	1	,000		
Razón de verosimilitud	26,710	1	,000		
Prueba exacta de Fisher				,000	,000
Asociación lineal por lineal	24,864	1	,000		
N de casos válidos	551				

Análisis de asociación: 2) fuerza

- El análisis de asociación entre variables se interesa también por la *intensidad* o fuerza con la que dichas variables se relacionan.
- Volvamos a épsilon. Si los valores son altos no es lo mismo que si son bajos. Ello nos da una idea de la fuerza con que las variables están asociadas.

Cuadro 1. Lectura de diario impreso según grupo de edad.

Lectura de diarios impresos	Grupo de edad		Total
	Jóvenes	Adultos	
No	84,0%	63,2%	84,0%
Sí	16,0%	36,8%	16,0%
Total	100,0%	100,0	100,0%

$$\varepsilon = 36,8 - 16,0 = \pm 20,8 \text{ pp.}$$

Cuadro 1. Consumo de diario impreso según grupo de edad.

Lectura de diarios impresos	Grupo de edad		Total
	Jóvenes	Adultos	
No	89,0	58,9	68,8
Sí	11,0	41,1	31,2
Total	100,0	100,0	100,0

$$\varepsilon = 41,1 - 11,0 = \pm 30,1 \text{ pp.}$$

- Nuestra hipótesis luce más fuerte o robusta en el cuadro derecho, justamente porque la asociación es más fuerte.

Análisis de asociación: 3) dirección

- La idea de **dirección** de la asociación sólo tiene sentido cuando las variables son, al menos, de nivel de medición ordinal.
- Cuando los valores altos (bajos) de una variable varían con los niveles altos (bajos) de la otra variable, hablamos de una **asociación positiva**. Cuando los valores altos (bajos) se corresponden con los valores bajos (altos) de la otra variable, nos referimos a una **asociación negativa (o inversa)**.
- En el ejemplo a continuación, vemos una relación positiva entre origen socioeconómico y destino socioeconómico.

		Origen socioeconómico (padre)			Total
		Alto	Medio	Bajo	
Destino socioeconómico (hijo)	Alto	63,4	10,0	4,4	25,5
	Medio	23,6	65,6	34,0	43,4
	Bajo	13,1	24,4	61,6	31,1
Total		100,0	100,0	100,0	100,0

**ASOCIACIÓN
POSITIVA**

- En una asociación positiva, los porcentajes altos se ubican en la **diagonal principal**.
- En una asociación negativa, los porcentajes altos se ubican en la **diagonal secundaria**.

Análisis de asociación: 4) naturaleza

- La **naturaleza** de la asociación se refiere a la forma en que se distribuyen los datos en la tabla. Para evaluar la naturaleza tenemos que ver cómo cambian los porcentajes de una categoría a otra de la variable.
- Veamos los siguientes esquemas, donde las celdas sombreadas representan *aquellas en las que esperamos encontrar porcentajes elevados*.

Relación lineal

Relación curvilínea

Relación rinconal

- “A mayor nivel socioeconómico, mayor nivel de bienestar subjetivo” (lineal)
- “Las personas que realizan trabajos de calificación media tienen más chances de ser reemplazados por robots que los de calificación baja o alta” (curvilínea)
- “Las personas de bajo nivel socioeconómico tienen propensión a experimentar alto déficit en la realización de ejercicio físico” (rinconal)

Medidas de asociación

- Obviamente, el coeficiente épsilon es muy limitado. Veremos dos tipos de medidas que permiten evaluar la fuerza de la asociación:
- Una *medida de asociación global* resume la **fuerza** o el grado de asociación de la relación entre las dos variables de una tabla de contingencia. Tenemos aquí los *coeficientes de asociación*.
- Una *medida de asociación local* nos informa sobre las características de la asociación bivariada, su naturaleza y la forma en que se relacionan las distintas categorías de las variables. Mencionaremos el *análisis de residuos* y el *análisis de razones*.

Medidas de asociación global

- Los coeficientes de asociación ofrecen una medida sintética del grado de asociación entre dos variables. Existen muchos coeficientes, que debemos seleccionar según el *nivel de medición* de las variables en la tabla.
- Los coeficientes de asociación se dividen en dos grupos:
 - *Los basados en χ^2* : son coeficientes en cuyo cálculo interviene chi-cuadrado.
 - Los basados en la *reducción proporcional del error*: estos coeficientes miden la fuerza de la asociación comparando cuánto cambia la predicción que podríamos realizar sobre el ordenamiento de casos conociendo sólo la variable dependiente con respecto a la predicción que realizaríamos al introducir la variable independiente.

Medidas de asociación global: coeficientes basados en χ^2

- El coeficiente χ^2 no está “estandarizado”: como pudimos comprobar, cuando el n de la tabla es más grande, el valor de χ^2 cambia sustantivamente. Esto hace que no podamos comparar la asociación entre dos tablas de contingencia de distinto n . Distintos coeficientes basados en χ^2 tratan de subsanar este punto.
- Para tablas de contingencia de 2x2 (dos variables con dos categorías cada una), el coeficiente de asociación más habitual es **Phi**:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

- Cuando la tabla tiene más de 2x2, utilizaremos el coeficiente **V de Cramer**. Este coeficiente casi nunca alcanza un valor de 1 y, de hecho, valores de 0,3 pueden considerarse muy aceptables (López-Roldán y Fachelli, 2016).

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

Donde n es el tamaño de la muestra y k es el valor más pequeño entre filas o columnas en la tabla de contingencia. Cuando la tabla es de 2x2, $V = \phi$, dado que $k=2$

Medidas de asociación global: coeficientes basados en χ^2

Tabla cruzada Lectura de diarios impresos*Grupo de edad

% dentro de Grupo de edad

		Grupo de edad		Total
		Jóvenes	Adultos	
Lectura de diarios impresos	No	84,0%	63,2%	70,1%
	Sí	16,0%	36,8%	29,9%
Total		100,0%	100,0%	100,0%

Aplicando la fórmula tenemos que:

$$V = \sqrt{\frac{24,97}{551 \cdot (2 - 1)}} = 0,213$$

Notar que ϕ arrojaría el mismo valor

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)	Sigra (b)
Chi-cuadrado de Pearson	24,909 ^a	1	,000	
Corrección de continuidad ^b	23,930	1	,000	
Razón de verosimilitud	26,710	1	,000	
Prueba exacta de Fisher				
Asociación lineal por lineal	24,864	1	,000	
N de casos válidos	551			

Medidas simétricas

		Valor	Significación aproximada
Nominal por Nominal	Phi	,213	,000
	V de Cramer	,213	,000
N de casos válidos		551	

Medidas de asociación global

- Como todos los coeficientes basados en χ^2 , los anteriores son especialmente apropiados para evaluar relaciones de asociación **lineales**.
- Esto significa que el coeficiente será muy alto cuando se aprecie un claro agrupamiento en torno a la diagonal principal.
- Con un supuesto menos exigente, que asuma relaciones **rinconales** (y tablas de 2x2) podemos usar **Q de Yule**:

a	b
c	d

$$Q = \frac{a * d - b * c}{a * d + b * c}$$

Medidas de asociación global

Lectura de diarios impresos	Grupo de edad	
	Jóvenes	Adultos
No	152	234
Sí	29	136

$$Q = \frac{a * d - b * c}{a * d + b * c}$$

$$Q = \frac{152 * 136 - 234 * 29}{152 * 136 + 234 * 29} = 0,506$$

- Cuando las tablas son de más de 2x2, utilizaremos el coeficiente **Gamma**, que también es apropiado para medir el grado de asociación entre variables ordinales. Gamma es un coeficiente basado en la reducción proporcional del error y toma valores entre -1 y +1. **Cuando la tabla es de 2x2, Gamma=Q de Yule.**

Medidas simétricas

		Valor	Error estándar asintótico ^a	T aproximada ^b	Significación aproximada
Ordinal por ordinal	Gamma	,506	,085	5,516	,000
N de casos válidos		551			

a. No se presupone la hipótesis nula.

b. Utilización del error estándar asintótico que presupone la hipótesis nula.

Medidas de asociación global

- Para saber qué coeficiente de asociación usar, hay que: 1) evaluar el nivel de medición de las variables de la tabla; 2) elegir el coeficiente a partir de esta tabla, pero teniendo en cuenta en función de la variable que tiene *menor* nivel de medición (si tengo una variable nominal y otra ordinal, debo elegir coeficientes para variables nominales).

1. Variables nominales	2. Variables ordinales
1.1 Basadas en el chi-cuadrado <ul style="list-style-type: none"> – <i>Coeficiente V de Cramer</i> – <i>Coeficiente de contingencia C de Pearson</i> – <i>Coeficiente de contingencia C de Pearson con modificación de Sakoda</i> – <i>Coeficiente de contingencia cuadrático medio de Pearson</i> – <i>Coeficiente de contingencia de Tschruprow</i> 	<ul style="list-style-type: none"> – <i>Coeficiente Tau-a de Kendall</i> – <i>Coeficiente Tau-b de Kendall</i> – <i>Coeficiente Tau-c de Kendall-Stuart</i> – <i>Coeficiente Gamma de Goodman-Kruskal</i> – <i>Coeficiente d de Sommer</i> – <i>Coeficiente e de Wilson</i> – <i>Coeficiente Ro de Spearman</i>
1.2. Para tablas de 2×2 <ul style="list-style-type: none"> – <i>Coeficiente Q de Yule</i> – <i>Coeficiente Phi</i> 	3. Variable nominal y ordinal <ul style="list-style-type: none"> – <i>Correlación biserial-punto</i> – <i>Correlación biserial</i> – <i>Eta</i>
1.3. Medidas de reducción de la predicción del error <ul style="list-style-type: none"> – <i>Coeficiente Lambda de Goodman-Kruskal</i> – <i>Coeficiente Tau de Goodman-Kruskal</i> – <i>Coeficiente de incertidumbre</i> 	4. Variables de intervalo <ul style="list-style-type: none"> – <i>Coeficiente de correlación producto-momento de Pearson</i>

Coefficientes de asociación: síntesis

Tabla cruzada Lectura de diarios impresos*Grupo de edad

% dentro de Grupo de edad

		Grupo de edad		Total
		Jóvenes	Adultos	
Lectura de diarios impresos	No	84,0%	63,2%	70,1%
	Sí	16,0%	36,8%	29,9%
Total		100,0%	100,0%	100,0%

Mientras que los coeficientes que asumen relaciones diagonales nos “penalizan”, los coeficientes rinconales nos muestran que la asociación es alta.

Medidas simétricas

		Valor	Significación aproximada
Nominal por Nominal	Phi	,213	,000
	V de Cramer	,213	,000
N de casos válidos		551	

Medidas simétricas

		Valor	Error estándar asintótico ^a	T aproximada ^b	Significación aproximada
Ordinal por ordinal	Gamma	,506	,085	5,516	,000
N de casos válidos		551			

a. No se presupone la hipótesis nula.
 b. Utilización del error estándar asintótico que presupone la hipótesis nula.

Leemos el coeficiente:
 Tenemos un nivel de asociación medio-bajo, con $V=0,130$

La asociación es estadísticamente significativa
(Esto ya lo sabíamos por la prueba de Chi Cuadrado)

Medidas de asociación local: análisis de residuos

- El análisis de residuos permite ver la relación entre pares de categorías en una tabla cruzada.
- El *residuo* es la diferencia entre las frecuencias observadas y las frecuencias esperadas en caso de no asociación. *Cuanto mayor sea el residuo, estaremos observando una mayor asociación, en esa celda en particular.*

Tabla cruzada Clase social de destino*Clase social de origen

Recuento

		Clase social de origen			Total
		Alta	Media	Baja	
Clase social de destino	Alta	392	547	160	1099
	Media	245	1431	664	2340
	Baja	56	627	557	1240
Total		693	2605	1381	4679

Tabla cruzada Clase social de destino*Clase social de origen

Recuento esperado

		Clase social de origen			Total
		Alta	Media	Baja	
Clase social de destino	Alta	162,8	611,9	324,4	1099,0
	Media	346,6	1302,8	690,6	2340,0
	Baja	183,7	690,4	366,0	1240,0
Total		693,0	2605,0	1381,0	4679,0

Tabla cruzada Clase social de destino*Clase social de origen

Residual

		Clase social de origen		
		Alta	Media	Baja
Clase social de destino	Alta	229,2	-64,9	-164,4
	Media	-101,6	128,2	-26,6
	Baja	-127,7	-63,4	191,0

$$R = 392 - 162,8 = 229,2$$

Vemos que hay más “exceso” de asociación entre las variables en algunas celdas

Medidas de asociación local: análisis de residuos

- Trabajar con los *residuos estandarizados corregidos* nos va a permitir sacar conclusiones acerca de la significancia de los residuos.
- El análisis de residuos estandarizados o tipificados corregidos nos permite considerar como estadísticamente significativos todas las celdas en que $R \geq \pm 1,96$.

Tabla cruzada Clase social de destino*Clase social de origen

Residuo corregido

		Clase social de origen		
		Alta	Media	Baja
Clase social de destino	Alta	22,3	-4,5	-12,4
	Media	-8,4	7,5	-1,7
	Baja	-11,9	-4,2	13,9

No es estadísticamente significativo

Medidas de asociación local: análisis de razones

- Existen tres tipos de razones que podemos calcular en una tabla de contingencia. Retomemos nuestro ejemplo:

Recuento

		Grupo de edad		Total
		Jóvenes	Adultos	
Lectura de diarios impresos	No	152	234	386
	Sí	29	136	165
Total		181	370	551

- Razones marginales:** se obtienen a partir de las frecuencias marginales de cada variable. Por ejemplo, la razón de no leer diarios impresos frente a leerlos es: $386/165=2,46$. Es decir, hay 2,46 personas que no consumen diarios impresos por cada persona que sí lo hace.
- Razones condicionales:** se obtienen de la misma forma, pero comparando dentro de cada categoría de la variable independiente. 1) La razón de no leer diarios impresos frente a leerlos, entre jóvenes, es: $152/29=5,24$ y 2) entre adultos es: $234/136= 1,72$.

Las razones condicionales nos dan una idea de asociación, porque si ambas razones fueran iguales no habría diferencia en la lectura de diarios.

Medidas de asociación local: análisis de razones

Recuento

		Grupo de edad		Total
		Jóvenes	Adultos	
Lectura de diarios impresos	No	152	234	386
	Sí	29	136	165
Total		181	370	551

- **Razón de razones (odds ratio) o razón de momios:** sintetiza la información anterior. Compara, entre grupos, las razones condicionales. Así, tendremos que:

$$OR = \frac{5,24}{1,72} = 3,04$$

Entre los jóvenes, las chances de no leer diarios impresos (frente a consumirlos) son 3,04 veces las que se registran entre los adultos.

Tablas de contingencia en SPSS

SPSS Menú: **Análisis** > **Tablas cruzadas...**

Diálogo: **Tablas cruzadas**

- Filas: info_diario
- Columnas: edad_grupo2

Mostrar en las casillas (Sub-diálogo):

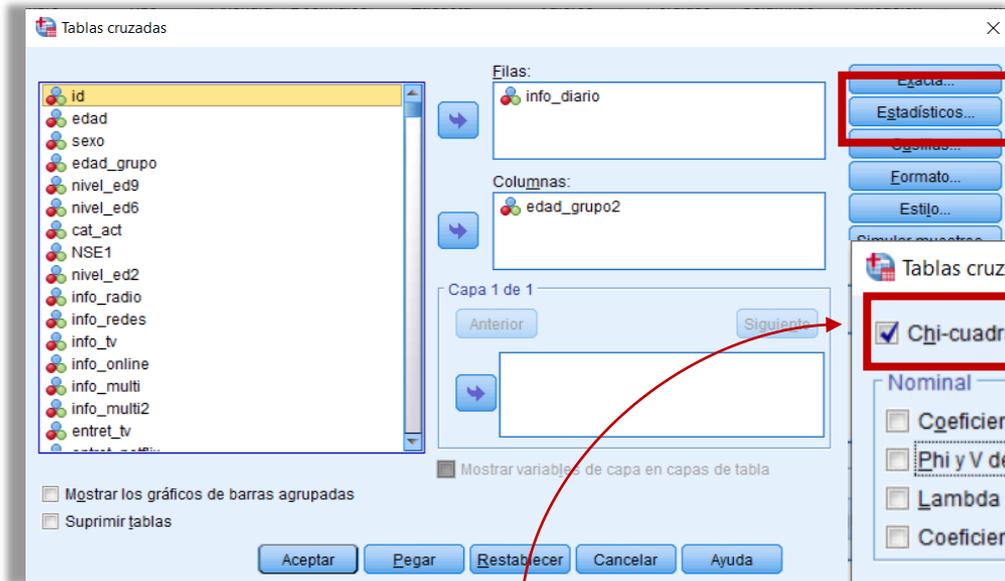
- Recuentos: Observado, Esperado, Ocultar recuentos pequeños (Menor que: 5)
- Prueba z: Comparar las proporciones de columna (Corregir valores p (método de Bonferroni))
- Residuos: No estandarizados, Estandarizados, Estandarizados corregidos
- Porcentajes: Ella, Columna, Total
- Ponderaciones no enteras: Redondear recuentos de casillas, Redondear ponderaciones de casos, Truncar recuentos de casillas, Truncar ponderaciones de casos, No efectuar correcciones

Elegir las variables de filas y columnas

Botón > Casillas

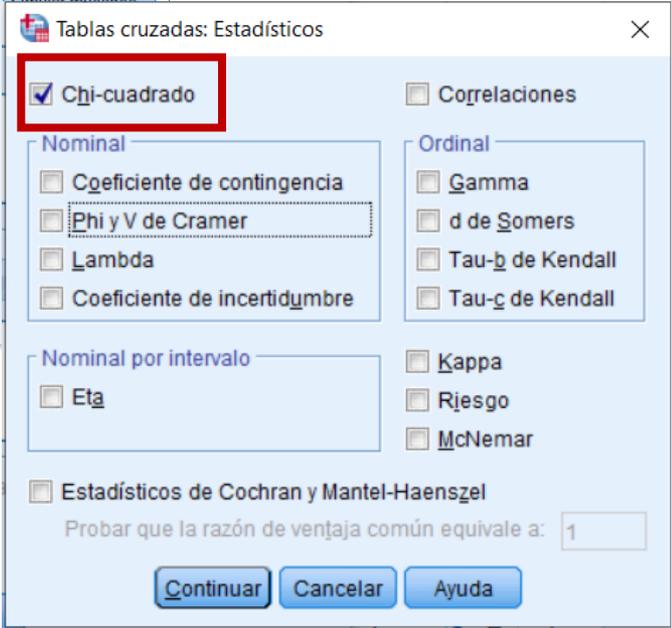
Seleccionamos los porcentajes que queremos

Prueba de independencia estadística χ^2

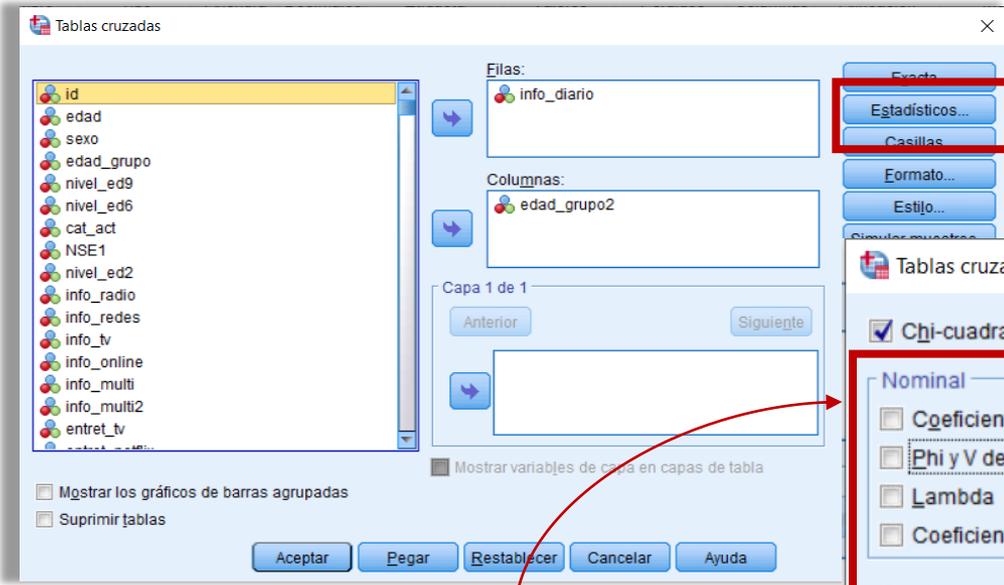


Botón > Estadísticos

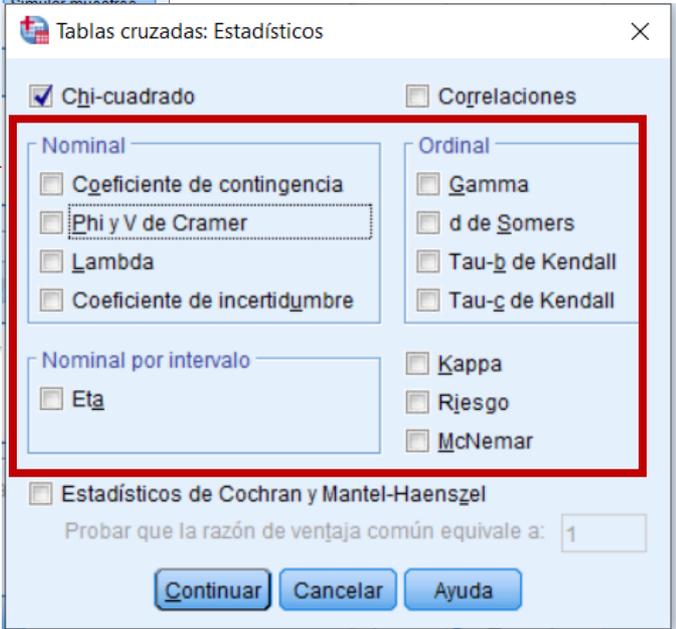
Seleccionamos la prueba



Coeficientes de asociación

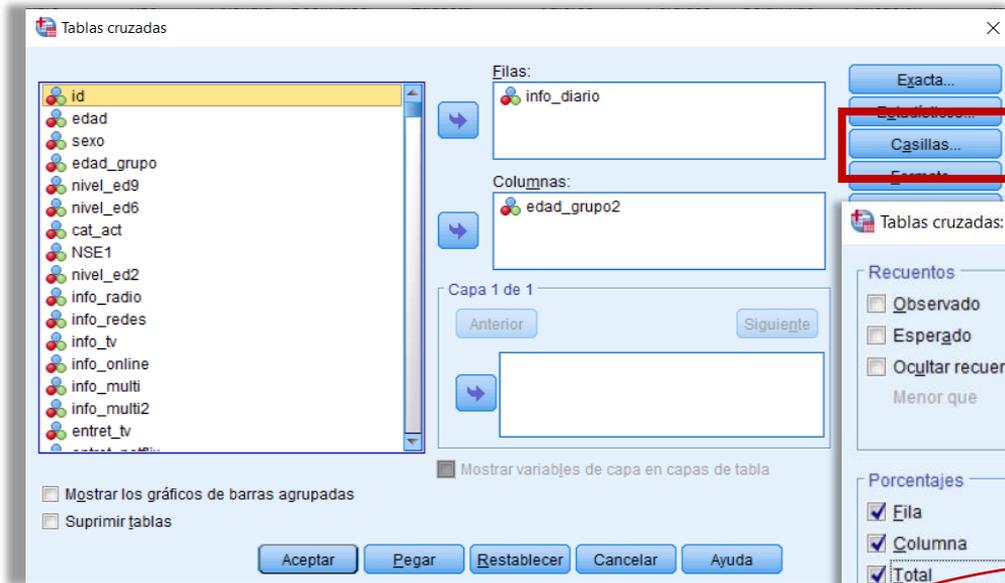


Botón > Estadísticos

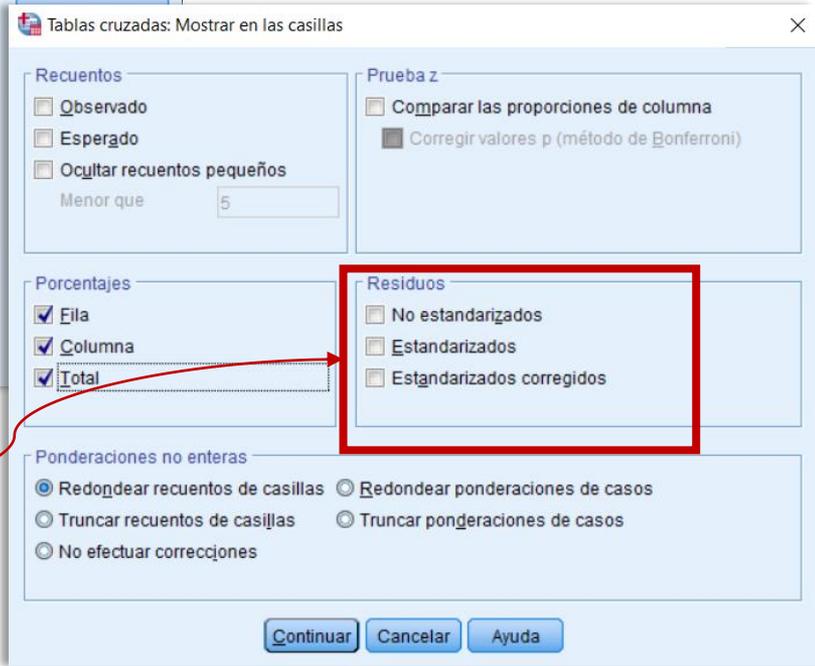


Seleccionamos el coeficiente

Análisis de residuos tipificados



Botón > Casillas



Seleccionamos los residuos