# Índice general

# Capítulo 5

# Análisis de datos categóricos

El análisis de datos categóricos se ocupa del estudio de variables que no son medibles (color, nacionalidad, enfermedades, sexo, afiliación política, etc.), denominadas también **atributos** o **caracteres cualitativos**. Podemos distinguir entre datos en escala *nominal* (sexo, estado civil, distintas ramas de actividad económica, profesión, ideología política, ...) y datos en escala *ordinal* (nivel de estudios, estratificación de familias por su capacidad de consumo, nivel de autoestima, ..), cuando podemos establecer un determinado orden o *rango* entre las observaciones.

En estos casos no tiene sentido el empleo de promedios, tales como la media aritmética. Cuando las observaciones se nos ofrecen en una escala nominal, sólo la *moda* puede utilizarse como medida resumen; y si éstas responden a una escala ordinal, podría determinarse, además del valor modal, también la mediana.

Una cuestión más interesante es el estudio de la existencia o no de asociación entre dos atributos, y de medidas similares a las de correlación para los casos en que variables no numéricas están relacionadas entre sí.

Para atributos en escala nominal estableceremos los llamados *coeficientes de contingencia*.

Cuando los caracteres estudiados pueden ordenarse de acuerdo con una cierta escala, es posible definir unos coeficientes de correlación que midan el grado de asociación entre ellos de manera parecida a como se mide la asociación entre variables cuantitativas. Estos coeficientes están basados en los rangos u órdenes de las observaciones.

## 5.1. Tablas de contingencia

Una variable cualitativa bidimensional está dada por dos atributos que se observan simultáneamente sobre los individuos de una población. De forma análoga al caso de dos variables numéricas, la distribución de frecuencias conjunta una variable cualitativa bidimensional $(A, B)$ está definida por los pares de datos observados sobre los individuos de la población junto con sus frecuencias absolutas.

Los datos pueden organizarse en **serie** o en una **tabla de doble entrada**. La tabla de doble entrada para caracteres cualitativos recibe el nombre de **tabla de contingencia**.

1. Los datos bidimensionales en serie se presentan en una tabla unidimensional con dos columnas, una

para cada uno de los atributos. Los datos en una misma fila se entiende que han sido observados sobre el mismo individuo.

La siguiente tabla representa los pares de valores $(A_i, B_i)$ de $(A, B)$ observados sobre un total de $n$ individuos.

| $A$ | $B$ |
|---|---|
| $A_1$ | $B_1$ |
| $A_2$ | $B_2$ |
| $\vdots$ | $\vdots$ |
| $A_i$ | $B_i$ |
| $\vdots$ | $\vdots$ |
| $A_n$ | $B_n$ |

2. Si organizamos los datos en una tabla de doble entrada, entonces mostraremos, por ejemplo, las modalidades del atributo $A$ (valores distintos de $A$) por filas en la primera columna de la tabla $(A_1, A_2, \ldots, A_k)$, las modalidades del atributo $B$ (valores distintos de $B$) por columnas en la primera fila de la tabla $(B_1, B_2, \ldots, B_p)$, y las cantidades $n_{ij}$ en el interior de la tabla indican el número de individuos de la población que presentan simultáneamente la modalidad i-ésima de $A$ y la modalidad $j$-ésima de $B$, esto es, la frecuencia absoluta del par de valores $(A_i, B_j)$.

   Así la representación típica de una tabla de contingencia $k \times p$ ($k$ filas y $p$ columnas) es:

| A\B | $B_1$ | $B_2$ | $\cdots$ | $B_j$ | $\cdots$ | $B_p$ | |
|---|---|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1j}$ | $\cdots$ | $n_{1p}$ | |
| $A_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2j}$ | $\cdots$ | $n_{2p}$ | |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | |
| $A_i$ | $n_{i1}$ | $n_{i2}$ | $\cdots$ | $n_{ij}$ | $\cdots$ | $n_{ip}$ | |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | |
| $A_k$ | $n_{k1}$ | $n_{k2}$ | $\cdots$ | $n_{kj}$ | $\cdots$ | $n_{kp}$ | |
| | | | | | | | $n$ |

   En este caso, el número total de individuos de la población, $n$, es la suma de todas las frecuencias $n_{ij}$ del interior de la tabla.

## 5.2. Distribuciones marginales y condicionadas

Las distribuciones marginales están dadas por la distribución unidimensional de cada uno de los atributos independientemente de cuáles sean los valores del otro atributo. Así,

1. La distribución marginal del atributo por filas $A$, está definida por las modalidades de dicho atributo, $A_i$, con frecuencias marginales

$$n_{i.} = \sum_{j=1}^{p} n_{ij}, \ \forall\ i = 1, \ldots, k$$

   es decir, con frecuencias marginales dadas por los totales de frecuencias por filas de la tabla.

2. La distribución marginal del atributo por columnas $B$, está definida por las modalidades de dicho atributo, $B_j$, con frecuencias marginales

$$n_{\cdot j} = \sum_{i=1}^{k} n_{ij},\ \forall\ j = 1, \ldots, p$$

es decir, con frecuencias marginales dadas por los totales de frecuencias por columnas de la tabla.

Es habitual determinar las distribuciones marginales sobre la tabla de doble entrada añadiendo una columna a la derecha con los totales por filas, y una fila en la parte inferior con los totales por columnas, como se indica a continuación:

| A\ B | $B_1$ | $B_2$ | $\cdots$ | $B_j$ | $\cdots$ | $B_p$ | Totales |
|---|---|---|---|---|---|---|---|
| $A_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1j}$ | $\cdots$ | $n_{1p}$ | $n_{1\cdot}$ |
| $A_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2j}$ | $\cdots$ | $n_{2p}$ | $n_{2\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $A_i$ | $n_{i1}$ | $n_{i2}$ | $\cdots$ | $n_{ij}$ | $\cdots$ | $n_{ip}$ | $n_{i\cdot}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $A_k$ | $n_{k1}$ | $n_{k2}$ | $\cdots$ | $n_{kj}$ | $\cdots$ | $n_{kp}$ | $n_{k\cdot}$ |
| Totales | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $\cdots$ | $n_{\cdot j}$ | $\cdots$ | $n_{\cdot p}$ | $n$ |

Es inmediato que

$$\sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{p} n_{\cdot j} = n$$

Las distribuciones condicionadas son las distribuciones unidimensionales de uno de los atributos dado que el otro atributo presenta determinadas modalidades. Las más sencillas son las distribuciones de cada atributo condicionadas a cada una de las modalidades del otro atributo. Así,

1. Las distribuciones del atributo por filas $A$ condicionadas a que el atributo por columnas $B$ presenta el valor $B_j$, que notaremos $A/B = B_j$, están definidas por las modalidades del atributo $A$ con frecuencias condicionadas $n_{ij}$, $i = 1, \ldots, k$.

   Hay $p$ distribuciones de este tipo, y cada una de ellas no está definida sobre el total de individuos, sino sobre la subpoblación de $n_{\cdot j}$ individuos para los que $B = B_j$, $j = 1, \ldots, p$.

   Obsérvese que las frecuencias de la distribución condicionada $A/B = B_j$ están dadas por las frecuencias de la $j$-ésima columna de la tabla de contingencia.

2. Las distribuciones del atributo por filas $B$ condicionadas a que el atributo por columnas $A$ presenta el valor $A_i$, que notaremos $B/A = A_i$, están definidas por las modalidades del atributo $B$ con frecuencias condicionadas $n_{ij}$, $j = 1, \ldots, p$.

   Hay $k$ distribuciones de este tipo, y cada una de ellas no está definida sobre el total de individuos, sino sobre la subpoblación de $n_{i\cdot}$ individuos para los que $A = A_i$, $i = 1, \ldots, k$.

   Obsérvese que las frecuencias de la distribución condicionada $B/A = A_i$ están dadas por las frecuencias de la $i$-ésima fila de la tabla de contingencia.

## 5.3. Independencia. Test Chi-cuadrado. Tablas $2\times2$

### 5.3.1. Independencia

Diremos que los atributos $A$ y $B$ son independientes si la proporción de individuos que presentan conjuntamente los valores $(A_i, B_j)$ de $(A, B)$ entre los que presentan el valor $A_i$ de $A$ es la misma para cualquier valor de $j$; o equivalentemente, la proporción de individuos que presentan conjuntamente los valores $(A_i, B_j)$ de $(A, B)$ entre los que presentan el valor $B_j$ de $B$ es la misma para cualquier valor de $i$.

Entonces, dos atributos $A$ y $B$ son estadísticamente independientes si y sólo si

$$n_{ij} = \frac{n_{i.}n_{.j}}{n} \ , \quad \forall\ i = 1, 2, \ldots, k\,;\ j = 1, 2, \ldots, p$$

### 5.3.2. Test Chi-cuadrado. Tablas $2\times2$

Existe un contraste formal para la **hipótesis nula de independencia** de los atributos $A$ y $B$ a un determinado nivel de significación $\alpha$, a partir de la información muestral recogida en la tabla de contingencia. **La hipótesis alternativa es la existencia de asociación** entre los atributos $A$ y $B$.

$H_0 : A$ y $B$ son independientes
$H_1 : A$ y $B$ no son independientes

Este test es conocido como **test Chi-cuadrado** y se basa en la distribución bajo la hipótesis nula del llamado **coeficiente de contingencia** $\chi^2$ (coeficiente Chi-cuadrado).

Si designamos $n'_{ij}$ a la frecuencia teórica que correspondería al par de modalidades $(A_i, B_j)$ en el caso de que ambos atributos fueran independientes, conocida como **frecuencia esperada (bajo independencia) del par** $(A_i, B_j)$, esto es,

$$n'_{ij} = \frac{n_{i.}n_{.j}}{n} \ , \quad \forall\ i = 1, 2, \ldots, k\,;\ j = 1, 2, \ldots, p\,;$$

se define el **coeficiente de contingencia** $\chi^2$ como

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{(n'_{ij} - n_{ij})^2}{n'_{ij}}$$

Algunos autores lo denominan **cuadrado de contingencia**, y puede expresarse de forma más sencilla para el cálculo como sigue:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}} - n$$

Puede demostrarse que, bajo la hipótesis nula de independencia de los atributos, el estadístico $\chi^2$ se distribuye según una $\chi^2_{(k-1)(p-1)}$.

Así, para realizar el contraste se halla el valor de una Chi cuadrado con $(k-1)(p-1)$ grados de libertad que deja a la derecha una probabilidad $\alpha$, que denotaremos $\chi^2_{(k-1)(p-1),\alpha}$. Si el valor del estadístico $\chi^2_{exp}$

para los datos observados es mayor que $\chi^2_{(k-1)(p-1),\alpha}$ se rechaza la hipótesis nula de independencia de los atributos $A$ y $B$ al nivel de significación $\alpha$.

O equivalentemente, cómo hace Statgraphics, podemos determinar la **probabilidad que deja a la derecha el valor del estadístico $\chi^2_{exp}$ en una distribución** $\chi^2_{(k-1)(p-1)}$, conocida como ***p*-valor del contraste**. Claramente $\chi^2_{exp} > \chi^2_{(k-1)(p-1),\alpha}$ si y sólo si $p$-valor$< \alpha$.

Por tanto, **si *p*-valor$< \alpha$ se rechaza la hipótesis nula de independencia** de los atributos $A$ y $B$ al nivel de significación $\alpha$, es decir, se acepta la hipótesis alternativa de existencia de asociación entre los atributos $A$ y $B$ al nivel de significación $\alpha$.

Antes de aplicar el test Chi-cuadrado debemos comprobar que se verifican las siguientes condiciones:

1. Ninguna frecuencia esperada es menor que 1
2. Al menos el 80 % de las frecuencias esperadas son mayores que 5

Si estas condiciones no se cumplen, no se puede aplicar el test. En tales casos debemos agrupar las modalidades o aumentar el tamaño muestral con el objetivo de que se cumplan las condiciones de validez del test.

Para tablas $2\times2$, resultan más adecuadas las siguientes condiciones:

1. Las frecuencias marginales son mayores que $\dfrac{n}{10}$
2. Todas las frecuencias esperadas son mayores que 5

Si no se cumplen estas condiciones debe aplicarse otro test conocido como *test exacto de Fisher*. Si el p-valor a 2 colas correspondiente a este test es menor que el nivel de significación considerado se rechaza la hipótesis nula de independencia.

Además, en las tablas $2\times2$ hay que hacer siempre una corrección por continuidad (corrección de Yates) del estadístico de la Chi-Cuadrado, tomando en su lugar el estadístico corregido de Yates, cuya expresión es

$$\chi^2_Y = \sum_{i=1}^{k}\sum_{j=1}^{p}\frac{(|n'_{ij} - n_{ij}| - 0.5)^2}{n'_{ij}}$$

## 5.4. Medidas de asociación: Coeficiente Chi-cuadrado. Otros coeficientes de contingencia

Como concepto contrario al de independencia tenemos el de **asociación**. Se dice que dos atributos $A$ y $B$ están asociados cuando aparecen juntos en mayor número de casos que el que cabría esperar si fuesen independientes.

Según que esa tendencia a coincidir o no coincidir esté más o menos marcada, tendremos distintos grados de asociación. Para medirlos se han ideado diversos coeficientes de asociación.

En la práctica, una vez rechazada la independencia entre los atributos mediante el test Chi-cuadrado, utilizaremos dichos coeficientes de asociación para medir la intensidad de la relación entre los atributos.

Parecería razonable que, puesto que el coeficiente de contingencia $\chi^2$ se emplea en el test Chi-cuadrado para determinar si dos atributos están relacionados, dicho coeficiente proporcionara una medida de asociación entre los atributos. Sin embargo, no es así.

El problema radica en que dicho coeficiente depende del tamaño muestral $n$. En efecto, si todas las frecuencias absolutas bidimensionales de la tabla de contingencia se multiplican por un mismo número $k$, entonces el nuevo valor de $\chi^2_{exp}$ resulta ser el anterior valor de $\chi^2_{exp}$ multiplicado por $k$. Por tanto, la magnitud de $\chi^2_{exp}$ no es una indicación del grado de asociación de los atributos. Dicho de otra forma, el valor $\chi^2_{exp}$ indica únicamente la evidencia de asociación (si es distinto de 0), no su grado.

Obviamente, si los atributos son independientes, entonces

$$n'_{ij} = n_{ij}$$

es decir, las frecuencias esperadas coinciden con las observadas, y $\chi^2 = 0$.

No obstante, es posible definir a partir del coeficiente de contingencia $\chi^2$ una serie de coeficientes de contingencia que sí constituyen medidas de asociación y que presentamos a continuación.

Podemos eliminar el efecto del tamaño muestral sobre el coeficiente de contingencia sin más que considerar

$$\varphi^2 = \frac{\chi^2}{n} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{p}\frac{n_{ij}^2}{n'_{ij}} - 1$$

Esta medida de asociación es conocida como **cuadrado medio de contingencia.** Es un número comprendido entre 0 (asociación nula o independencia de los atributos) y 1 (asociación máxima o total), pudiendo interpretarse como un coeficiente de correlación lineal.

El coeficiente $\varphi^2$ también presenta una serie de inconvenientes. En general, para tablas $k \times p$ se utiliza el **coeficiente de contingencia de Pearson**, definido como

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}},$$

Este coeficiente varía entre 0 (asociación nula o independencia entre los atributos) y $\sqrt{\frac{q-1}{q}}$ $(< 1)$ (asociación máxima entre los caracteres) con $q = min\{k, p\}$. El coeficiente se aproxima a 1 conforme crecen simultáneamente el número de modalidades de los dos atributos, pero sólo alcanzaría el valor 1 en el caso ideal de infinitas modalidades.

En cualquier caso, el coeficiente $C$ nos revelará un menor grado de asociación entre los atributos cuanto más próximo esté a 0 y un mayor grado de asociación entre los atributos cuanto más se aproxime a $\sqrt{\frac{q-1}{q}}$.

A su vez, Tschuprow propuso un coeficiente que depende nuevamente de $\chi^2$, del número de filas y columnas, y del total de individuos, $n$. El **coeficiente de Tschuprow** está definido por

$$T = \sqrt{\frac{\varphi^2}{\sqrt{(k-1)(p-1)}}} = \sqrt{\frac{\chi^2}{n\sqrt{(k-1)(p-1)}}}$$

El coeficiente varía entre 0 y 1 con la interpretación habitual, y alcanza el valor máximo sólo cuando la tabla es cuadrada $(k = p)$.

Los coeficientes $C$ y $T$ están relacionados por las expresiones

$$C = \sqrt{\frac{\varphi^2}{\varphi^2 + 1}} = \sqrt{\frac{T^2\sqrt{(k-1)(p-1)}}{1 + T^2\sqrt{(k-1)(p-1)}}}$$

y

$$T = \sqrt{\frac{\varphi^2}{\sqrt{(k-1)(p-1)}}} = \sqrt{\frac{C^2}{(1-C^2)\sqrt{(k-1)(p-1)}}}$$

Otro coeficiente, que también depende de $\chi^2$, es el **coeficiente V de Cramer**, cuya expresión es

$$V = \sqrt{\frac{\varphi^2}{m}} = \sqrt{\frac{\chi^2}{m\,n}}$$

donde $m = min\{k-1, p-1\}$. Se trata de un coeficiente que toma el valor 1 cuando hay asociación perfecta entre los atributos, cualquiera que sea el tamaño de la tabla de contingencia. Cuando la tabla es cuadrada $V = T$, y en caso contrario $V > T$.

Existen también una serie de medidas de asociación utilizadas en el caso de atributos en escala ordinal. De ellas comentaremos únicamente que además de evaluar el grado de asociación entre los atributos, indican la dirección de dicha asociación según que la medida sea positiva o negativa. Suele haber tres casos extremos: asociación perfecta positiva, asociación perfecta negativa e independencia (ausencia de asociación).

## 5.5. Análisis de datos categóricos con STATGRAPHICS

Para resumir la distribución de frecuencias una variable bidimensional cualitativa Statgraphics proporciona dos subopciones dentro de la opción **Datos Cualitativos** del menú **Descripción**:

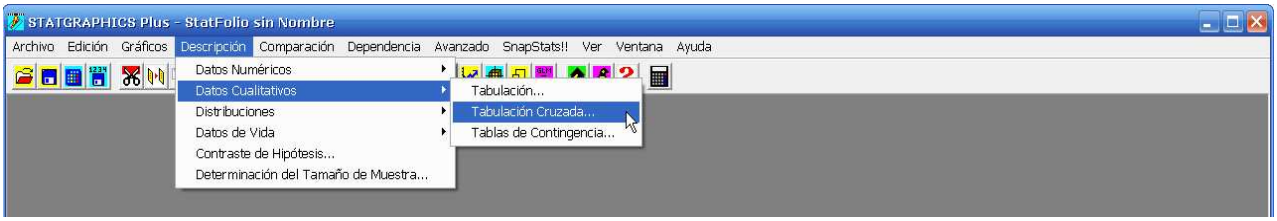
- La subopción **Tabulación Cruzada...**, si los datos están organizados en serie y queremos que Statgraphics los tabule generando la tabla de frecuencias bidimensional.
- La subopción **Tablas de Contingencia...**, si los datos están organizados en una tabla y nuestros datos son las propias frecuencias de la tabla de contingencia.

### 5.5.1. Tabulación Cruzada...

**Ejemplo 1**

*El archivo* **Emplea2.sf3** *contiene los datos observados sobre 36 empleados de una empresa de la variable cualitativa bidimensional definida por los atributos "Sexo" y "Categoría laboral" (CatLab). El atributo Sexo presenta dos categorías: H (hombre) y M (mujer); y el atributo CatLab presenta 3 categorías: A (Administrativos), C (Comerciales) y T (Técnicos)*

Haciendo click en la subopción **Tabulación Cruzada** de la opción **Datos Cualitativos** del menú **Descripción** accedemos al cuadro de diálogo **Tabulación Cruzada - Entrada de Datos**:





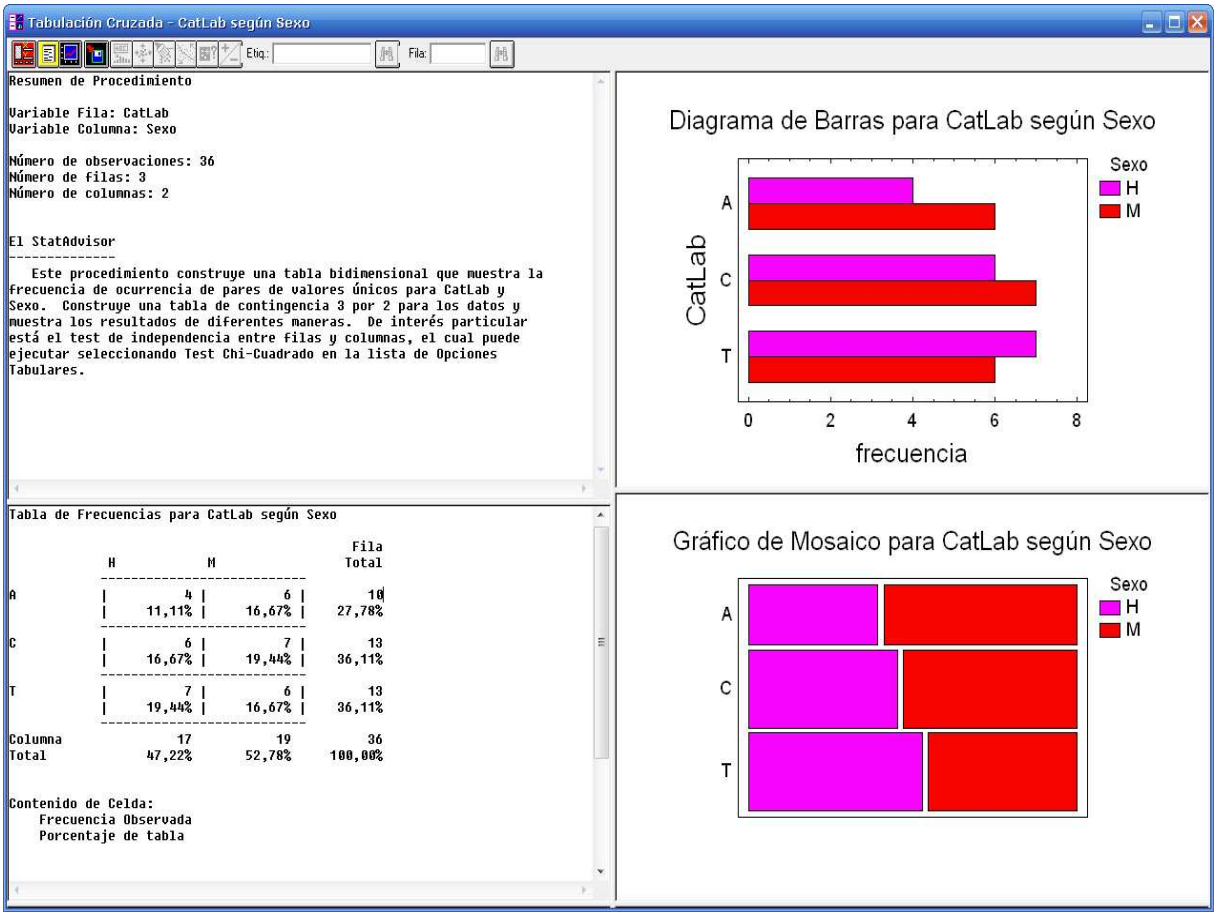
- En el campo **Variable Fila** introduciremos el nombre del atributo cuyas modalidades van a aparecer por filas (en la primera columna). Por ejemplo *CatLab*.

- En el campo **Variable Columna** introduciremos el nombre del atributo cuyas modalidades van a aparecer por columnas (en la primera fila). En nuestro caso, *Sexo*.

- El campo **(Selección:)** es opcional y podemos introducir un operador de selección que acote el conjunto de valores de los atributos, lo que permite trabajar en subpoblaciones de la población total.

- La opción **Ordenar** permite ordenar las modalidades de los atributos alfabéticamente. Está activada por defecto.

Al hacer click en el botón *Aceptar*, se muestra la ventana del análisis de tabulación cruzada:




Podemos observar los siguientes elementos:

- El **resumen del procedimiento**, que indica los atributos fila y columna, el número de observaciones y el tamaño de la tabla de contingencia que se va a construir.
- La **tabla de frecuencias** del atributo por filas según el atributo por columnas, con los totales por filas y columnas que definen las distribuciones marginales de los atributos.

  Por defecto **en cada celda el primer número que aparece es la frecuencia absoluta del par de modalidades correspondientes y el segundo número es el porcentaje de tabla que supone** respecto al número total de datos. Por ejemplo, los datos de la celda intersección de la fila C con la columna M nos indican que del total de 36 empleados, 7 son comerciales y mujeres, esto es, que un 19,44 % del total de empleados son comerciales y mujeres.

  Los **totales por filas** definen las frecuencias absolutas marginales del atributo por filas y el porcentaje del total que representan. En nuestro ejemplo, definen la distribución marginal de la categoría laboral de los 36 empleados, y nos indican que 10 son administrativos, 13 son comerciales y otros 13 son técnicos; o equivalentemente, el 27.78 % son administrativos, el 36.11 % son comerciales y el 36.11 % restante son técnicos.

  Y los **totales por columnas** definen las frecuencias absolutas marginales del atributo por columnas

y el porcentaje del total que representan. En nuestro ejemplo, definen la distribución marginal del sexo de los 36 empleados, y nos indican que 17 son hombres y 19 son mujeres; o equivalentemente, el 47,22 % son hombres y el 52.78 % restantes son mujeres.

Si, estando situados sobre la tabla de frecuencias, hacemos click con el botón derecho del ratón y elegimos la opción **Opciones de Ventana...** del menú emergente que aparece, se abre el cuadro de diálogo **Opciones Tabla de Frecuencias**, en cuyo campo **Incluir** podemos elegir construir la tabla con porcentajes de tabla (opción por defecto), con porcentajes de fila y columna, con frecuencias esperadas, con desviaciones y/o con valores Chi-cuadrado.


Si activamos todas las opciones y hacemos click en el botón *Aceptar*, se obtiene la siguiente tabla de frecuencias:

Tabla de Frecuencias para CatLab según Sexo

| | H | M | Fila Total |
|---|---|---|---|
| A | 4 | 6 | 10 |
| | 11,11% | 16,67% | 27,78% |
| | 40,00% | 60,00% | |
| | 23,53% | 31,58% | |
| | 4,72 | 5,28 | |
| | -0,72 | 0,72 | |
| | 0,11 | 0,10 | |
| C | 6 | 7 | 13 |
| | 16,67% | 19,44% | 36,11% |
| | 46,15% | 53,85% | |
| | 35,29% | 36,84% | |
| | 6,14 | 6,86 | |
| | -0,14 | 0,14 | |
| | 0,00 | 0,00 | |
| T | 7 | 6 | 13 |
| | 19,44% | 16,67% | 36,11% |
| | 53,85% | 46,15% | |
| | 41,18% | 31,58% | |
| | 6,14 | 6,86 | |
| | 0,86 | -0,86 | |
| | 0,12 | 0,11 | |
| Columna | 17 | 19 | 36 |
| Total | 47,22% | 52,78% | 100,00% |

Contenido de Celda:
Frecuencia Observada
Porcentaje de tabla
Porcentaje de fila
Porcentaje de columna
Frecuencia Esperada
Frecuencia esperada - Observada
Contribución a chi-cuadrado

- Los **porcentajes de fila** son los porcentajes que representan las frecuencias de tabla respecto del total de fila. Son, por tanto, los porcentajes que representan las modalidades de la distribución del atributo por columnas condicionada a que el atributo por filas presenta la modalidad correspondiente a la fila considerada.

  En nuestro caso, por ejemplo, si consideramos la distribución del sexo condicionada a ser técnico, los porcentajes de fila que aparecen en la fila T nos indican que de los 13 técnicos que hay, el 53.85 % son hombres (7 de 13) y el 46,15 % restante son mujeres (6 de 13)

- Los **porcentajes de columna** son los porcentajes que representan las frecuencias de tabla respecto del total de columna. Son, por tanto, los porcentajes que representan las modalidades de la distribución del atributo por filas condicionada a que el atributo por columnas presenta la modalidad correspondiente a la columna considerada.

  En nuestro caso, por ejemplo, si consideramos la distribución de la categoría laboral condicionada a ser hombre, los porcentajes de columna que aparecen en la columna H nos indican que de los 17 hombres que hay, el 23.53 % son administrativos (4 de 17), el 35,29 % son comerciales (6 de 17), y el 41,18 % restante son técnicos (7 de 17)

- Las **frecuencias esperadas** son las frecuencias que cabría esperar en cada celda si los atributos fueran independientes.

Por ejemplo, la frecuencia esperada bajo independencia de la celda intersección de la fila C con la columna M está dada por

$$\frac{13 \times 19}{36} = 6.86$$

- Las **desviaciones** son las diferencias entre las frecuencias de tabla observadas y las esperadas. Las desviaciones positivas corresponden a individuos que se presentan en más casos de los que cabría esperar bajo independencia. Por el contrario, las desviaciones negativas corresponden a individuos que se presentan en menos casos de los que cabría esperar bajo independencia.

  Así, la desviación de la celda intersección de la fila T con la columna M está dada por $6 - 6.86 = -0.86$

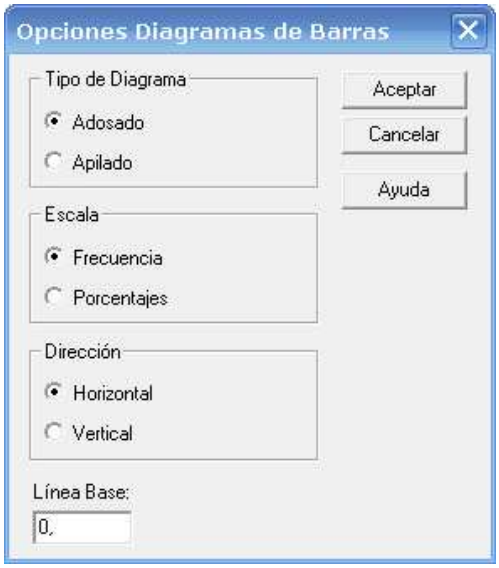
- Los **valores Chi-cuadrado** muestran la contribución de cada celda al estadístico $\chi^2$ que se utiliza para contrastar la independencia entre los atributos.

  Por ejemplo, la aportación al estadístico $\chi^2$ de la celda intersección de la fila A con la columna H está dada por

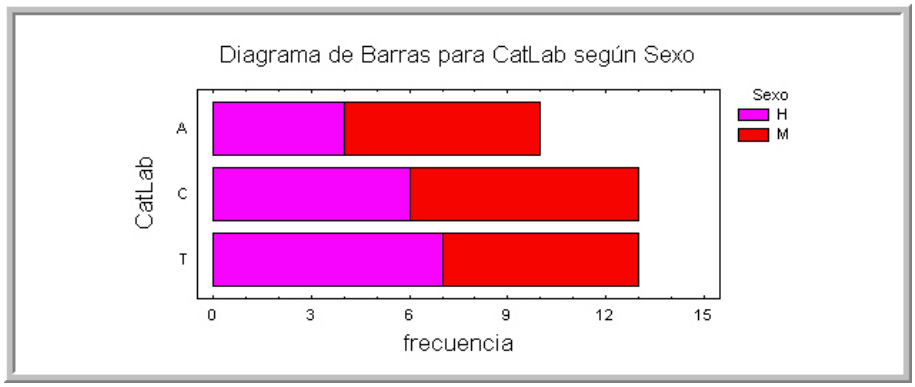
$$\frac{(-0.72)^2}{4.72} = 0.11$$

- EL **diagrama de barras** adosadas para el atributo por filas según el atributo por columnas, que muestra un gráfico de barras múltiples bidimensional sobre un mismo eje. Las longitudes de las barras son proporcionales a las frecuencias absolutas de la tabla de frecuencias; y las barras se agrupan, adosadas o apiladas, según el atributo por columnas para cada modalidad del atributo por filas.

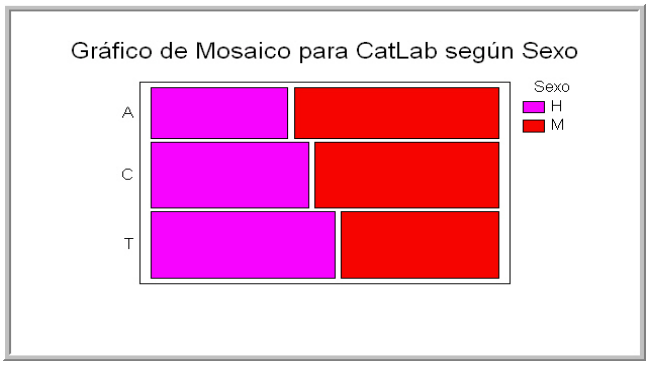
  Si, estando situados sobre este gráfico, hacemos click con el botón derecho del ratón y elegimos la opción **Opciones de Ventana...** del menú emergente que aparece, se abre el cuadro de diálogo **Opciones Diagrama de Barras**, que en el campo **Tipo de Diagrama** nos permite apilar las barras en lugar de adosarlas, en el campo **Escala** podemos elegir entre porcentajes o frecuencias, y en el campo **Dirección** podemos determinar la dirección horizontal o vertical para el gráfico.



El diagrama de barras apiladas puede identificarse con el diagrama de barras de la distribución marginal del atributo por filas sin más que considerar cada grupo de barras apiladas como una sola barra.




- El **gráfico de mosaico** para el atributo por filas según el atributo por columnas, formado por un mosaico de rectángulos cuyas áreas son proporcionales a las frecuencias absolutas de las celdas de la tabla de frecuencias, siendo la altura de los rectángulos para cada modalidad del atributo por filas proporcional a los totales por filas. De esta forma, la anchura de los rectángulos para cada modalidad del atributo por filas es proporcional a los porcentajes de fila.




Los rectángulos para una misma modalidad del atributo por filas representan un diagrama de barras apiladas de la distribución del atributo por columnas condicionada a dicha modalidad del atributo por filas.

Si, estando situados sobre este gráfico, hacemos click con el botón derecho del ratón y elegimos la opción **Opciones de ventana...** del menú emergente que aparece, se abre el cuadro de diálogo **Opciones Gráfico de Mosaico**, en cuyo campo **Dirección** podemos elegir la dirección vertical u horizontal para el gráfico.


Como en todo análisis de Statgraphics, el icono **Opciones tabulares** (segundo icono por la izquierda de la barra de herramientas de la ventana del análisis de tabulación cruzada), abre el cuadro de diálogo **Opciones Tabulares** que permite manejar todas las opciones del análisis:


- La opción tabular **Contraste de Chi-cuadrado** realiza el contraste cuya hipótesis nula es la independencia de los dos atributos. Se presenta en pantalla el valor del estadístico, los grados de libertad y el p-valor. Si el p-valor es menor que el nivel de significación $\alpha$ se rechaza la hipótesis nula de independencia a dicho nivel de significación.

```
Contraste de Chi-cuadrado
-----------------------------------------
    Chi-cuadrado        GL          P-Valor
-----------------------------------------
            0,44         2           0,8009
-----------------------------------------
Precaución: La frecuencia de alguna celda es inferior a 5.


El StatAdvisor
--------------
   El test chi-cuadrado realiza un contraste de hipótesis para
determinar si se rechaza o no la idea de que la fila y la columna
seleccionadas son independientes.  Dado que el p-valor es superior o
igual a 0.10, no podemos rechazar la hipótesis de que las filas y
columnas son independientes.  En consecuencia, el valor observado de
CatLab para un caso particular puede no tener relación con su valor en
Sexo.
```

En nuestro caso, el valor del estadístico $\chi^2$ es $\chi^2_{exp} = 0.44$ y dicho estadístico se distribuye según una Chi-cuadrado con $(3-1)\times(2-1) = 2$ grados de libertad ($\chi^2_2$). Y el p-valor está dado por $P(\chi^2_2 > 0.44) = 0.8009$. Entonces, al nivel de significación habitual $\alpha = 0.05$ no hay evidencia para rechazar la independencia de los atributos, lo que nos indica que la categoría laboral de un empleado

no tiene relación con su sexo, y, por tanto,la empresa no discrimina a sus empleados por razones de sexo.

Es importante señalar que Statgraphics nos avisa de que al menos una celda tiene una frecuencia esperada inferior a 5. Si observamos las frecuencias esperadas de la tabla de frecuencias es fácil ver que todas son mayores que 1 y que 5 de las 6 frecuencias son mayores que 5, es decir más del 80 % de las frecuencias esperadas son mayores que 5. Por tanto, se cumplen las condiciones de validez del contraste de la Chi-cuadrado.

- La opción tabular **Resumen Estadístico** calcula diferentes medidas de asociación y correlación por rangos que permiten determinar el grado de asociación entre dos atributos.

  En nuestro caso, no tiene sentido utilizar esta opción dado que los atributos son independientes. No obstante, si la utilizamos obtenemos la siguiente salida

Resumen Estadístico

| Estadístico | Simétrico | Con Filas Dependientes | Con Columnas Dependientes |
|---|---|---|---|
| Lambda | 0,0500 | 0,0435 | 0,0588 |
| Coef. Incertidumbre | 0,0069 | 0,0057 | 0,0089 |
| D de Somer | -0,1037 | -0,1207 | -0,0909 |
| Estadístico Eta | | 0,1108 | 0,1111 |

| | Valor | P-Valor | Gl |
|---|---|---|---|
| Coef. Contingencia | 0,1104 | | |
| V de Cramer | 0,1111 | | |
| Gamma condicional | -0,1814 | | |
| R de Pearson | -0,1108 | 0,2600 | 34 |
| Tau b de Kendall | -0,1048 | 0,5113 | |
| Tau c de Kendall | -0,1204 | | |

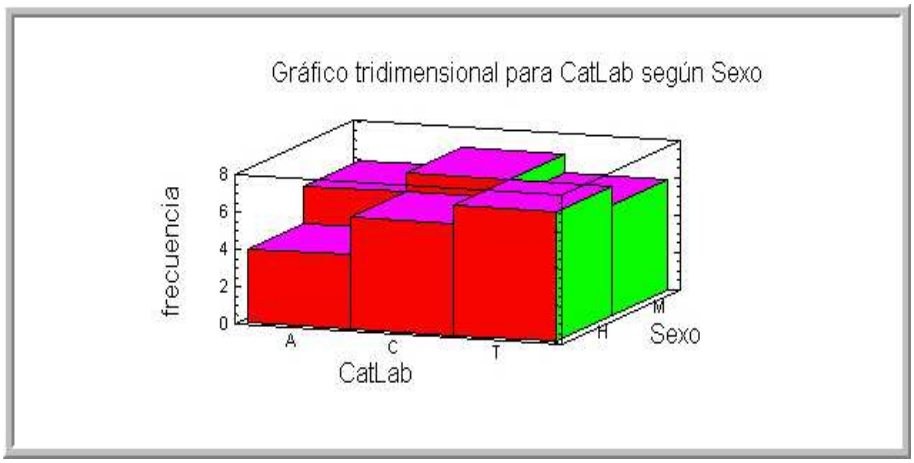
El StatAdvisor

Aquí las estadísticas muestran la medida del grado de asociación entre filas y columnas. De interés particular están el coeficiente de contingencia y lambda, que miden el grado de asociación en una escala de 0 a 1. Lambda mide la utilidad del factor de la fila (o columna) en la predicción de otro factor. Por ejemplo, el valor de lambda con columnas dependientes es igual a 0,0588235. Esto significa que hay un 5,88235% de reducción en el error cuando CatLab se utiliza para predecir Sexo. Para esas estadísticas con P-valores, los P-valores inferiores a 0.05 indican una asociación significativa entre filas y columnas con un nivel de confianza del 95%.

Statgraphics nos muestra, entre otros, el **coeficiente de contingencia** de Pearson $C = 0.1104$ y el **coeficiente V de Cramer** $V = 0.1111$.

Por otra parte, el icono **Opciones Gráficas** (tercer icono por la izquierda de la barra de herramientas de la ventana del análisis de tabulación cruzada), permite manejar todas las opciones gráficas del análisis de tabulación cruzada a través de la ventana **Opciones Gráficas**. Por defecto están seleccionadas las opciones **Diagrama de barras** y **Gráfico de mosaico**.


La opción **Gráfico Tridimensional** permite obtener un diagrama de barras tridimensional para la variable cualitativa bidimensional, en el que la altura de sus barras es proporcional a la frecuencia absoluta de cada celda de la tabla de frecuencias.




Si, estando situados sobre este gráfico, hacemos click con el botón derecho del ratón y elegimos la opción **Opciones de ventana...** del menú emergente que aparece, se abre el cuadro de diálogo **Opciones Gráfico Tridimensional**, en cuyo campo **Gráfico** podemos elegir la representación basada en frecuencias absolutas de celdas o en porcentajes.


### 5.5.2. Tablas de Contingencia...

**Ejemplo 2**

*La siguiente tabla clasifica a un grupo de personas atendiendo a la frecuencia con que leen la prensa y si escuchan o no las tertulias de radio:*

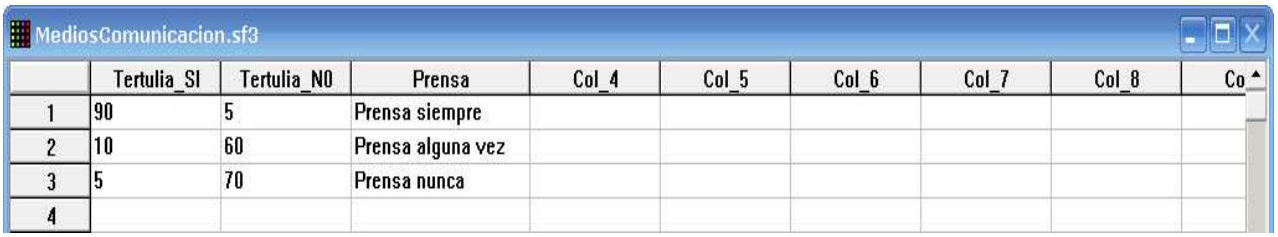
| | | *Tertulias de radio* | |
|---|---|---|---|
| | | *Si* | *No* |
| *Prensa* | *Siempre* | 90 | 5 |
| | *Alguna vez* | 10 | 60 |
| | *Nunca* | 5 | 70 |

*Estudia la asociación, si la hay, entre leer prensa y escuchar las tertulias de radio.*

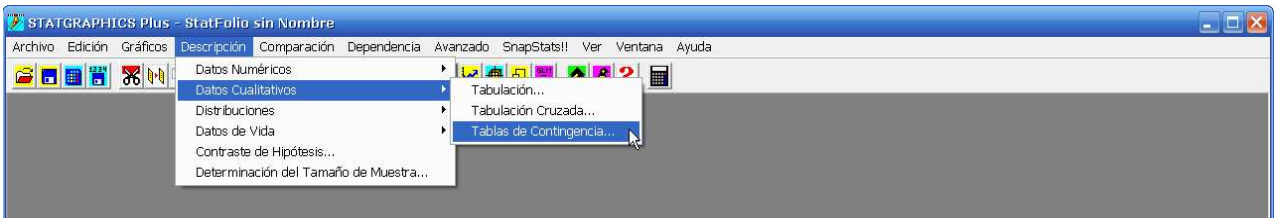
La primera tarea es introducir los datos de la tabla de contingencia en 2 columnas de la hoja de cálculo de Statgraphics, con nombres, por ejemplo, "Tertulia_SI" y "Tertulia_N0".

En la columna "Tertulia_SI" introduciremos como valores las frecuencias de la columna "Si" de la tabla de contingencia. Análogamente, en la columna "Tertulia_N0" introduciremos como valores las frecuencias de la columna "No" de la tabla de contingencia.

Aunque no es necesario, vamos a añadir otra columna con nombre "Prensa" para almacenar las modalidades de la frecuencia con que se lee la prensa y así etiquetar más adecuadamente las tablas de frecuencias y los gráficos del análisis.




A continuación haremos click en la subopción **Tablas de Contingencia...** de la opción **Datos Cualitativos** del menú **Descripción** para acceder al cuadro de diálogo **Tablas de Contingencia** de entrada de datos:




- En el campo **Columnas:** especificamos las columnas en las que hemos almacenado las columnas de frecuencias del atributo por columnas. En nuestro caso, *Tertulia_SI* y *Tertulia_N0*.
- El campo **Etiquetas** es opcional y en él podemos especificar las modalidades del atributo por filas. Como nosotros las hemos almacenado en la columna *Prensa*, introduciremos dicha columna.
- El campo **(Selección:)** es opcional y podemos introducir un operador de selección que acote el conjunto de valores de los atributos, lo que permite trabajar en subpoblaciones de la población total.
- La opción **Ordenar** permite ordenar las modalidades de los atributos alfabéticamente. Está activada por defecto.

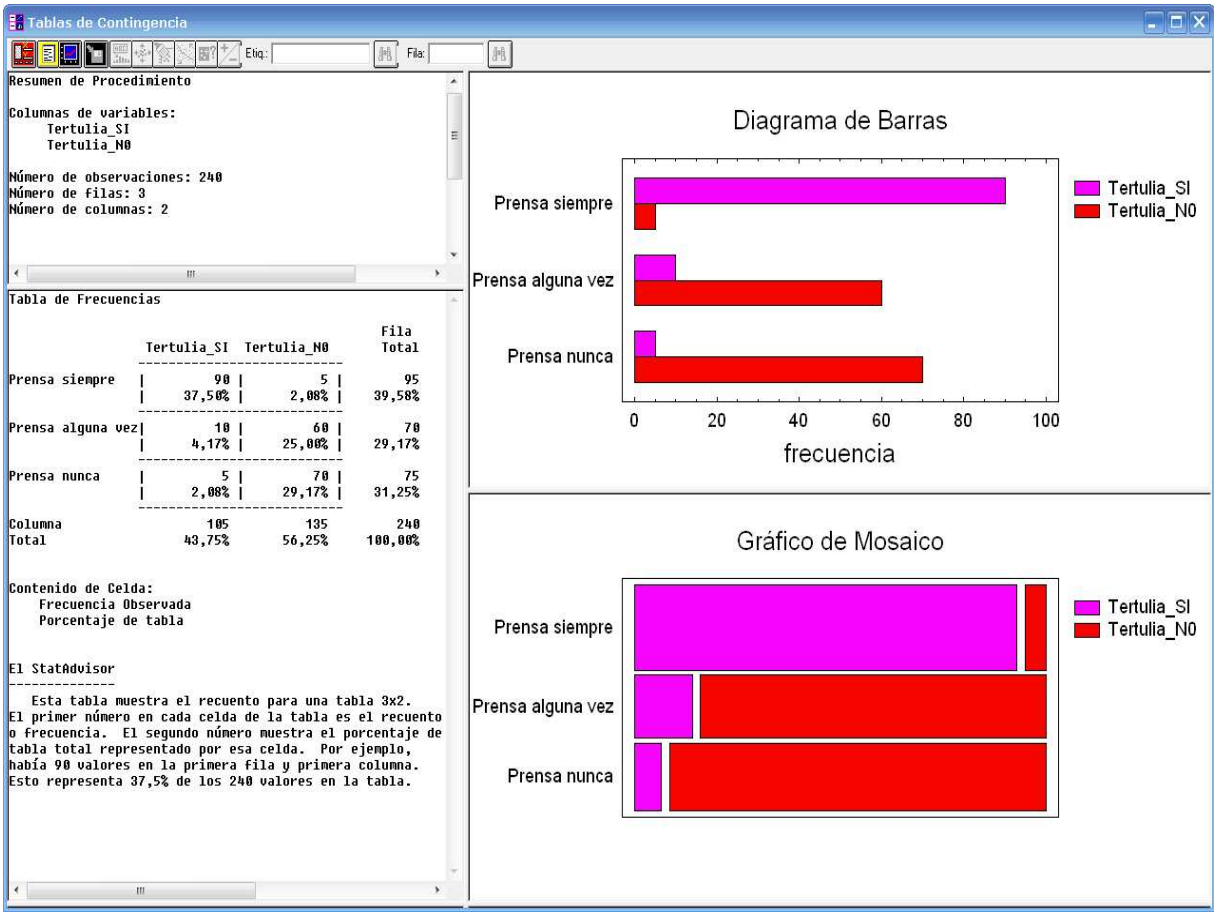
Al hacer click en el botón *Aceptar*, se muestra la ventana del análisis de tablas de contingencia, que es completamente análoga a la del análisis de tabulación cruzada del Ejemplo 1. Como en aquel caso muestra por defecto:

- El resumen del procedimiento
- La tabla de frecuencias con recuentos y porcentajes de tabla.

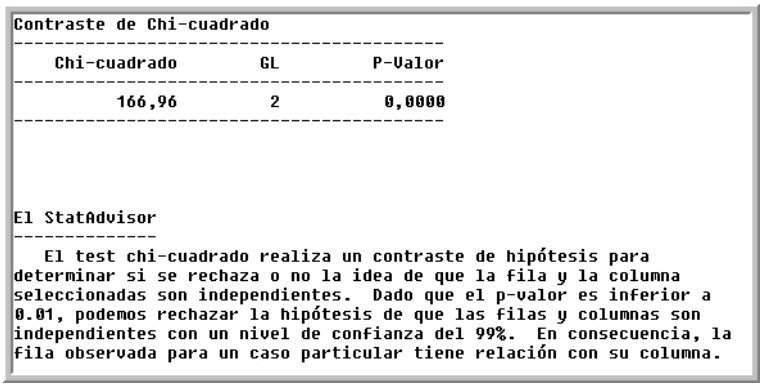
  Si hacemos click con el botón derecho del ratón sobre la tabla y elegimos la opción **Opciones de Ventana...** del menú emergente que aparece, se abre el cuadro de diálogo **Opciones Tabla de Frecuencias**, y podemos seleccionar el contenido de la tabla: porcentajes de tabla, porcentajes de fila y columna, frecuencias esperadas, desviaciones y/o valores Chi-cuadrado.
- El diagrama de barras multiple del atributo por filas según el atributo por columnas.

  Si hacemos click con el botón derecho del ratón sobre el gráfico y elegimos la opción **Opciones de Ventana...** del menú emergente que aparece, se abre el cuadro de diálogo **Opciones Diagrama de Barras**, que nos permite apilar las barras en lugar de adosarlas, elegir entre porcentajes o frecuencias, y determinar la dirección horizontal o vertical para el gráfico.
- El gráfico de mosaico del atributo por filas según el atributo por columnas.

Si hacemos click con el botón derecho del ratón sobre el gráfico y elegimos la opción **Opciones de ventana...** del menú emergente que aparece, se abre el cuadro de diálogo **Opciones Gráfico de Mosaico**, y podemos elegir la dirección vertical u horizontal para el gráfico.




Para realizar el contraste de la hipótesis nula de independencia de los atributos haremos click en el icono **Opciones tabulares** de la barra de herramientas del análisis y seleccionaremos la opción tabular **Test Chi-Cuadrado** del cuadro de diálogo **Opciones Tabulares**, que proporciona la siguiente salida:




Ahora, el estadístico $\chi^2$ toma el valor $\chi^2_{exp} = 166{,}96$; y como el p-valor es del orden de $10^{-5}$, entonces

es menor que el nivel de significación habitual $\alpha = 0.05$ y hay evidencia estadística para rechazar la independencia entre la frecuencia con que se lee la prensa y el que se escuche o no las tertulias de radio.

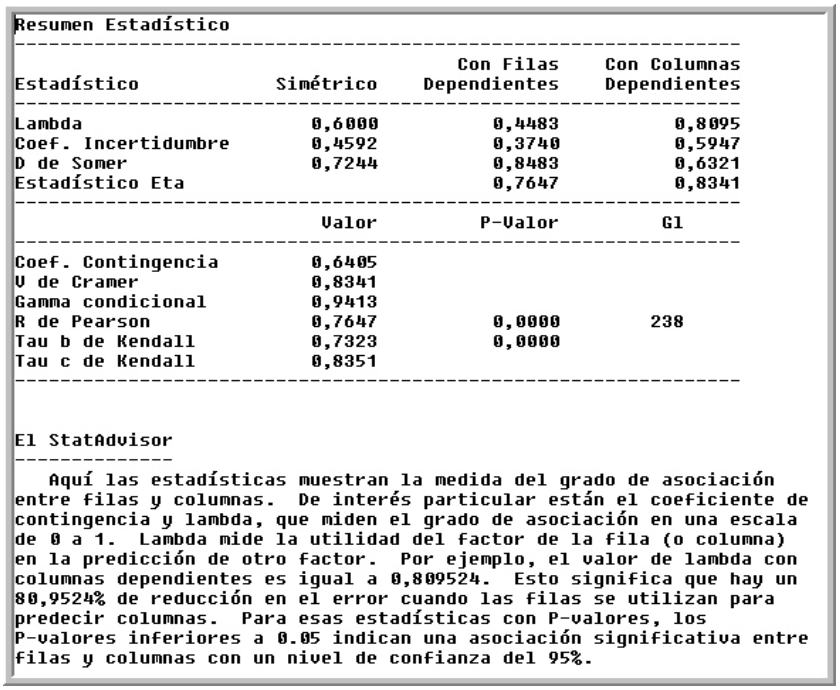
Statgraphics no nos advierte de que ninguna celda sea inferior a 5, por lo que el contraste Chi-cuadrado es válido. No obstante podemos comprobar que las frecuencias esperadas bajo independencia son todas mayores que 1 y más del 80 % de ellas son superiores a 5. Haciendo click con el botón derecho del ratón sobre la ventana **Tabla de Frecuencias** del análisis elegiremos la opción **Opciones de Ventana...** del menú emergente que se despliega. Así accedemos al cuadro de diálogo **Opciones Tabla de Frecuencias** y elegimos la opción **Frecuencias Esperadas**. La ventana *Tabla de Frecuencias* del análisis muestra ahora las frecuencias esperadas:

Tabla de Frecuencias

| | Tertulia_SI | Tertulia_NO | Fila Total |
|---|---|---|---|
| Prensa siempre | 90 | 5 | 95 |
| | 37,50% | 2,08% | 39,58% |
| | 41,56 | 53,44 | |
| Prensa alguna vez | 10 | 60 | 70 |
| | 4,17% | 25,00% | 29,17% |
| | 30,63 | 39,38 | |
| Prensa nunca | 5 | 70 | 75 |
| | 2,08% | 29,17% | 31,25% |
| | 32,81 | 42,19 | |
| Columna Total | 105 43,75% | 135 56,25% | 240 100,00% |

Contenido de Celda:
Frecuencia Observada
Porcentaje de tabla
Frecuencia Esperada

Dado que hay relación entre la frecuencia con que se lee la prensa y el que se escuche o no las tertulias de radio, el siguiente paso será cuantificar el grado de asociación mediante algún coeficiente de asociación. Para ello, haremos click sobre el icono **Opciones tabulares** de la barra de herramientas del análisis y seleccionaremos la opción tabular **Resumen Estadístico** del cuadro de diálogo **Opciones Tabulares**, que proporciona la siguiente salida:



Resumen Estadístico

| Estadístico | Simétrico | Con Filas Dependientes | Con Columnas Dependientes |
|---|---|---|---|
| Lambda | 0,6000 | 0,4483 | 0,8095 |
| Coef. Incertidumbre | 0,4592 | 0,3740 | 0,5947 |
| D de Somer | 0,7244 | 0,8483 | 0,6321 |
| Estadístico Eta | | 0,7647 | 0,8341 |

| | Valor | P-Valor | Gl |
|---|---|---|---|
| Coef. Contingencia | 0,6405 | | |
| V de Cramer | 0,8341 | | |
| Gamma condicional | 0,9413 | | |
| R de Pearson | 0,7647 | 0,0000 | 238 |
| Tau b de Kendall | 0,7323 | 0,0000 | |
| Tau c de Kendall | 0,8351 | | |

El StatAdvisor

Aquí las estadísticas muestran la medida del grado de asociación entre filas y columnas. De interés particular están el coeficiente de contingencia y lambda, que miden el grado de asociación en una escala de 0 a 1. Lambda mide la utilidad del factor de la fila (o columna) en la predicción de otro factor. Por ejemplo, el valor de lambda con columnas dependientes es igual a 0,809524. Esto significa que hay un 80,9524% de reducción en el error cuando las filas se utilizan para predecir columnas. Para esas estadísticas con P-valores, los P-valores inferiores a 0.05 indican una asociación significativa entre filas y columnas con un nivel de confianza del 95%.

El **coeficiente de contingencia** de Pearson es $C = 0.6405$, valor que se aproxima bastante al valor máximo de dicho coeficiente en el caso de asociación total $\sqrt{\frac{2-1}{1}} = \sqrt{0.5} = 0.7071$

Y el **coeficiente V de Cramer** es $V = 0.8341$, que se aproxima también bastante al valor 1 que corresponde a una asociación total.

Luego, podemos afirmar que hay una asociación bastante alta entre la frecuencia con que se lee la prensa y el que se escuche o no las tertulias de radio.