CAPÍTULO 8

ANÁLISIS DE DATOS CATEGÓRICOS

En este capítulo se discutiran técnicas estadísticas para anilizar datos categoricos, los cuales representan atributos o categorías. Primero se dicuten la relación entre las variables que definen las filas y las columna de las tablas y luego se estudian medidas que dan una medida del grado de asociación entre las dos variables categóricas.
Finalmente se estudia la prueba de bondad de ajuste que permite ver si un conjunto de datos sigue una distribución conocida agrupando previamente los datos en categorias.

8.1 Pruebas de Independencia y Homegeneidad

Consideremos datos de dos variables cualitativas A y B como por ejemplo, nivel económico y partido político al cual pertenece.una persona. También podrían ser dos variables cuantitativas que han sido categorizadas, como por ejemplo, Nivel de Educación y Nivel de salario. Como ya se había visto, en la sección 3.7.1 de este texto, los datos se organizan en una tabla de doble entrada, llamada ***Tabla de contingencia,*** cuya forma general es la siguiente:

		VAR A					
		A_1	A_2	A_3	…	A_c	Total
VAR B	B_1	O_{11}	O_{12}	O_{13}		O_{1C}	R_1
	B_2	O_{21}	O_{22}	O_{23}		O_{2C}	R_2
	B_3	O_{31}	O_{32}	O_{33}		O_{3C}	R_3
	…					…	
	B_r	O_{R1}	O_{R2}	O_{R3}	…	O_{RC}	R_r
	Total	C_1	C_2	C_3	…	C_c	N

Aquí O_{ij} es el número de sujetos que tienen las características A_i y B_j a la vez.
R_i (i = 1,…,r) es la suma de la i-ésima fila de la tabla. Es decir, es el total de sujetos que poseen la característica B_i.
C_j {j = 1,…,c) es la suma de la j-ésima columna de la tabla. Es decir, es el total de sujetos que poseen la característica A_j.
n representa el total de observaciones tomadas.
La tabla anterior es llamada una **tabla de contigencia r x c**, porque tiene *r* filas y *c* columnas.

Las tablas más elementales son aquellas con dos variables, donde cada una de ellas asume sólo dos valores distintos, ésta es llamada una tabla 2 x 2. Consideremos la siguiente tabla:

	A1	A2	Total

B1	8	6	14
B2	12	9	21
Total	20	15	35

La primera pregunta que uno se hace es si existirá o no relación entre las variables A y B, es decir si A y B son o no independientes. A y B serán independientes si cada entrada de la tabla es igual al producto de los totales marginales dividido entre el número de datos. Esto es si cumple,

$$O_{ij} = \frac{R_i C_j}{n}$$

para cada celda (i, j). Claramente, esto se cumple para la tabla anterior. Por ejemplo, $8 = (14)(20)/35$. En consecuencia, no hay relación entre las variables A y B.
Otra pregunta que se puede tratar de responder es sí las proporciones de los valores de la variable B en cada columna son iguales. Por ejemplo si A: El estudiante graduando consigue trabajo, B: Sexo del graduando. Uno puede estar interesado en comparar la proporción de mujeres graduandas que consiguen trabajo con la proporción de mujeres graduandas que no consiguen trabajo.
Consideremos ahora la tabla:

	A1	A2	Total
B1	10	6	16
B2	5	16	21
Total	15	22	37

Notar que los valores de la segunda fila están en sentido contrario a los de la primera fila. O sea hay un efecto en la variable A al cambiar los valores de B, en consecuencia aquí si hay relación entre las variables. Es bien obvio, también que la fórmula de independencia no se cumple para ninguna de las entradas. Por otro lado las proporciones de los valores de la variable B no son los mismos en cada columna. Por ejemplo para B1 las proporciones son 10/15 versus 6/22.

Cuando consideramos que los valores de nuestra tabla han sido extraídos de una población, entonces nos interesaría probar las siguientes dos hipótesis:

i) La **prueba de Independencia**, que se efectúa para probar si hay asociación entre la variables categóricas A y B, y

ii) La **prueba de Homogeneidad**, que es una generalización de la prueba de igualdad de dos proporciones, que se discutió en la sección 7.8. En este caso se trata de probar si para cada nivel de la variable B, la proporción con respecto a cada nivel de la variable A es la misma. Si A tiene 3 niveles y B tiene 2 niveles *entonces Ho : p*

Por ejemplo, nos gustaría saber si hay o no relación entre el nivel económico de una persona y su afiliación política. También podríamos estar interesados en determinar si hay relación entre el nivel de educación y el nivel de salario. En ambos casos se usaría una prueba de independencia.

Por otro lado, también podríamos estar interesados en probar si para cada nivel económico hay igual proporción de personas en cada partido politico, o si para cada nivel de educación hay igual proporción de personas en cada nivel de salario. En estos casos se usaría una prueba de homogeneidad.

Sin embargo; ambos tipos de hipótesis se pueden probar de la misma manera y el procedimiento se resume en el recuadro que sigue:

Las hipótesis de independencia son:
Ho: No hay asociación entre las variables A y B (es decir hay independencia)
Ha: Si hay relación entre las variables A y B

Las hipótesis de Homogeneidad son:

Ho: Las proporciones de cada valor de la variable B son iguales en cada columna
Ha: Al menos una de las proporciones para cada valor de la variable B no son iguales en cada columna.
Ambas hipótesis se prueban usando una prueba de Ji-Cuadrado:

$$\chi^2 = \sum_{i=1}^{c}\sum_{i=1}^{r}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

donde O_{ij} es la frecuencia observada de la celda que está en la fila i , columna j, y $E_{ij} = \frac{R_i C_j}{n}$, es la frecuencia esperada de la celda (i, j). La frecuencia esperada es aquella que debe ocurrir para que la hipótesis nula sea aceptada.
La prueba estadística se distribuye como una Ji-Cuadrado con $(r\text{-}1)(c\text{-}1)$ grados de libertad.
La hipótesis Nula se rechaza si $\chi^2_{cal} > \chi^2_{1-\alpha}$, donde α es el nivel de significancia o equivalentemente si el "P-value" es menor que 0.5.

Si la tabla de contingencia presenta pocas observaciones en algunas celdas (digamos menos de 5), entonces la prueba no es confiable. Existen pruebas exactas para tablas de contingencia, pero no se han considerado en este texto.

Para analizar tablas de contingencia en **MINITAB** se usa la opción ***Tables*** del menú **STAT**, ésta a su vez tiene un submenú que contiene las opciones ***Cross Tabulation*** y ***Chi Square.*** La opción ***Cross Tabulacion*** se usa en dos situaciones. La primera de ellas es cuando los datos están dados en dos columnas, o sea como si hubiesen sido las contestaciones a dos preguntas de un cuestionario. En el siguiente ejemplo se mostrará este primer uso.

Ejemplo 8.1. Usando los datos del ejemplo 3.16, supongamos que deseamos establecer si hay relación entre las variables tipo de escuela superior y el resultado (aprueba o no aprueba), de la primera clase de matemáticas que toma el estudiante en la universidad, basados en los resultados de 20 estudiantes.

Solución:
Para la prueba de Independencia las hipótesis son:
H_o: No hay relación entre el tipo de escuela y el resultado obtenido en la primera clase de Matemáticas.
H_a: Si hay relación entre ambas variables.
Para la prueba de homogeneidad las hipótesis son:
H_o: La proporción de aprobados en la primera clase de matemáticas es igual tanto para estudiantes que provienen de escuela pública como de escuela privada.
H_a: La proporción de aprobados en la primera clase de matemáticas no es la misma para ambos tipos de escuela.
La ventana de diálogo se completerá como aparece en la siguiente figura:

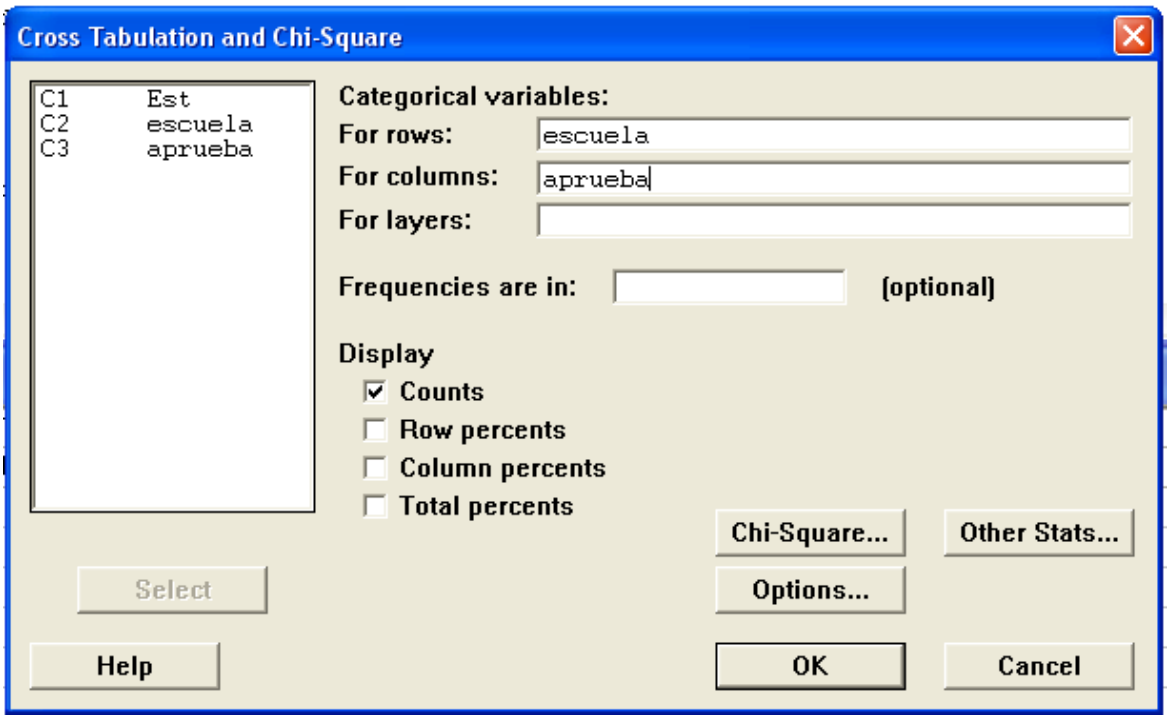

Figura 8.1. Ventana de diálogo de la opción **Cross Tabulation** del submenú **Tables** del menú **Stat**

Los resultados aparecerán en la ventana **session** como siguen:

```
Tabulated Statistics

 Rows: escuela     Columns: aprueba
```

```
            si       no      All

 priv        7        3       10
           6.00     4.00    10.00

 públ        5        5       10
           6.00     4.00    10.00

 All        12        8       20
           12.00     8.00    20.00
Chi-Square = 0.833, DF = 1, P-Value = 0.361
2 cells with expected counts less than 5.0
  Cell Contents --
                  Count
                  Exp Freq
```

Interpretación: *Como el "P-value" es mayor que .05 se puede concluir que la hipótesis nula de Independencia entre las variables es aceptada. O sea no hay asociación entre el tipo de escuela de donde proviene el estudiante y el resultado que obtiene en la primera clase de matemáticas.*
Por otro lado, la hipótesis nula de homogeneidad también es aceptada y se concluye de que, la proporción de estudiantes que aprueban el curso de matemáticas es la misma para estudiantes de escuela pública y escuela privada.

La segunda situación donde ***Cross Tabulation*** es usada para hacer el análisis de Ji-cuadrado, es cuando los datos ya están resumidos en tablas con filas y columnas, ésta es la manera usual como aparecen en los textos. En este caso, para que **MINITAB** pueda hacer el análisis se deben entrar los datos en 3 columnas. En una columna deben ir las frecuencias observadas en cada celda de la tabla y en las otras dos columnas deben ir los valores de las variables en filas y columnas que permitan identificar a qué celda le corresponde la frecuencia absoluta entrada.

Ejemplo 8.2. Usar los datos del ejemplo 3.17, para tratar de establecer si hay relación entre el Sexo del entrevistado y su opinión.

Solución: Las hipótesis correpondientes son:
Ho: No hay asociación entre el sexo del entrevistado y su opinión, y
Ha: Si hay relación entre las variables.

En este caso los datos son entrados en tres columnas: *Conteo* (frecuencia en cada celda), *Sexo y Opinión.* La ventana de diálogo se completará como se muestra en la figura 8.2
Los resultados serán los siguientes:

```
MTB > Table 'sexo' 'opinion';
SUBC>   Frequencies 'conteo';
SUBC>   ChiSquare 2.

Tabulated Statistics

 Rows: sexo     Columns: opinión
```

```
              si        no      abst       All

 male         10        20        30        60
           10.00     20.40     29.60     60.00

 female       15        31        44        90
           15.00     30.60     44.40     90.00

 All          25        51        74       150
           25.00     51.00     74.00    150.00

Chi-Square = 0.022, DF = 2, P-Value = 0.989

  Cell Contents --
                  Count
                  Exp Freq
```

Interpretación: *Como el "P-value" es mayor que .05, la conclusión en este caso es que la hipótesis nula es aceptada o sea no hay relación entre el sexo y la opinión del entrevistado.*

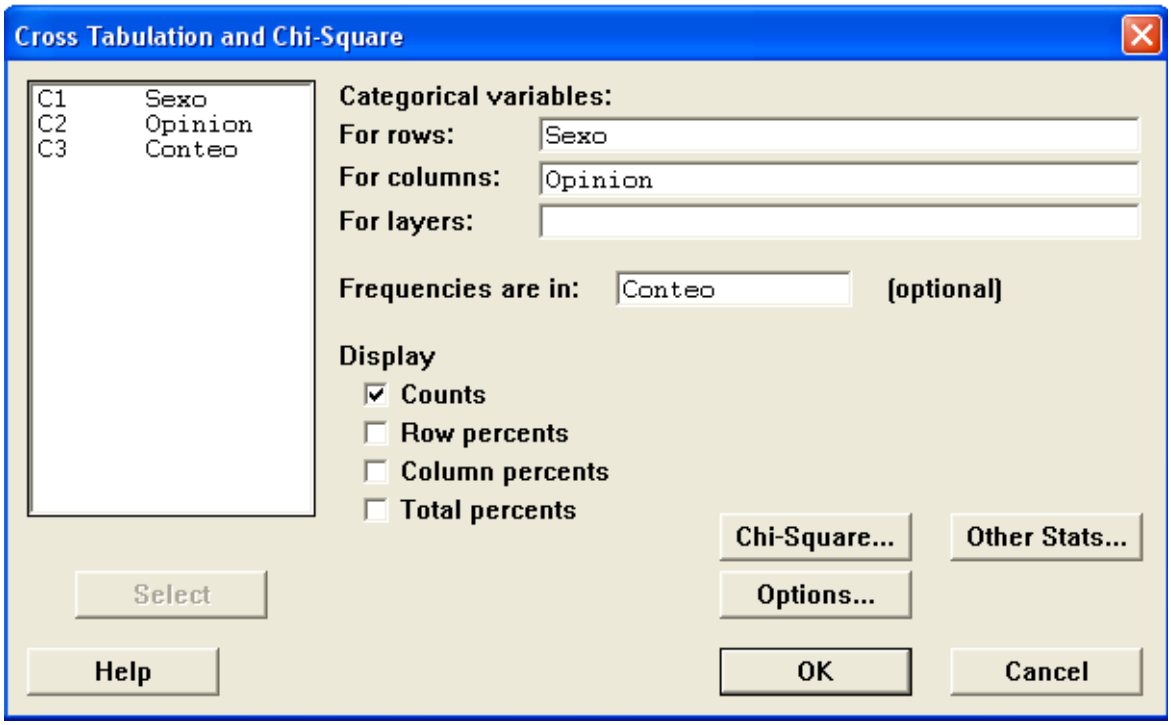

Figura 8.2. Ventana de diálogo de **cross tabulation** para analizar el ejemplo 8.2.

Notar que la opción **Chi-square analysis** aparece seleccionada. Como se ha elegido la opción **above and expected count**, la tabla de salida mostrará las frecuencias absolutas y las frecuencias esperadas de cada celda, en la ventanita de **frecuencies are in:** se asigna la columna conteo.

Existe una última posibilidad de hacer el análisis de la tabla de contingencia usando la opción ***Chi-Square Test.*** En este caso se supone que las columnas de la tabla son entradas columna por columna en el worksheet de **MINITAB**.

Ejemplo 8.3. Para los datos del ejemplo 3.17, donde la tabla es:

	SI	NO	Abst
Hombres	10	20	30
Mujeres	15	31	44

Primero se entran los datos en 3 columnas: SI, NO y ABST y luego se completa la ventana de diálogo de ***Chi-Square Test*** como sigue:

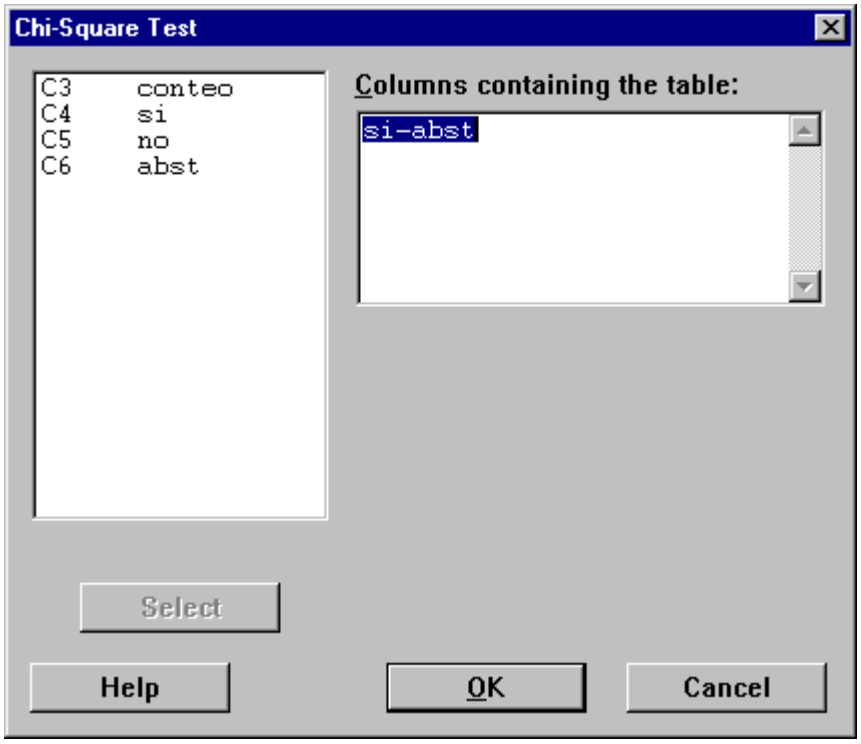

Figura 8.3. Ventana de diálogo para la opción **Chi-Square Test** del menú **Tables**

Los resultados aparecerán de la siguiente manera:

```
MTB > ChiSquare 'si'-'abst'.

Chi-Square Test
Expected counts are printed below observed counts

             si       no     abst    Total
     1       10       20       30       60
```

```
          10.00    20.40    29.60

     2       15       31       44       90
          15.00    30.60    44.40

Total        25       51       74      150

Chi-Sq =  0.000 +  0.008 +  0.005 +
          0.000 +  0.005 +  0.004 =  0.022
    DF = 2, P-Value = 0.989
```

Se puede notar que la presentación de la tabla no es tan buena como en los dos casos anteriores, pero si se presentan los cálculos intermedios de la prueba de Ji-Cuadrado.

8.2 Medidas de Asociación

Asumiendo que se rechaza la hipótesis Nula Ho: No hay relación entre las variables de la tabla, entonces el próximo paso es determinar el grado de asociación de las dos variables categóricas, para ello se usan las llamadas medidas de asociación. Existen un gran número de estas medidas, nosotros sólo consideraremos dos de ellas:

a) El Coeficiente de Contingencia:

Se define por

$C = \sqrt{\frac{\chi^2}{n+\chi^2}}$, donde χ^2 es el valor calculado de la prueba de Ji-Cuadrado y n es el número de datos.

El valor de C varía entre 0 y 1. Si C = 0, significa que no hay asociación entre las variables. El coeficiente de contigencia tiene la desventaja de que no alcanza el valor de uno aún cuando las dos variables sean totalmente dependientes. Otra desventaja es que su valor tiende a aumentar a medida que el tamaño de la tabla aumenta.

En general, un valor de C mayor que .30, indica una buena asociación entre las variables. Sin embargo hay que tomar en consideración también el tamaño de la tabla.
A diferencia de otros programas estadísticos como **SPSS** y **SAS, MINITAB** no calcula el coeficiente de contingencia directamente. Se tiene que usar **Calculator** del menú **CALC.**

Ejemplo 8.4. Calcular el coeficiente de contingencia para la siguiente tabla, donde se trata de relacionar las variables: asistir a servicios religiosos y faltar a clases.

```
Rows: va a igl     Columns: falta a

          de vez e frecuent    nunca      All

 de vez e      78      119      140      337
            75.56   103.44   158.01   337.00

 frecuent     106       90      296      492
           110.31   151.01   230.68   492.00

 nunca         68      136       91      295
            66.14    90.55   138.31   295.00

 All          252      345      527     1124
           252.00   345.00   527.00  1124.00
```

La ventana de diálogo de **Calculator** se debe completar de la siguiente manera:

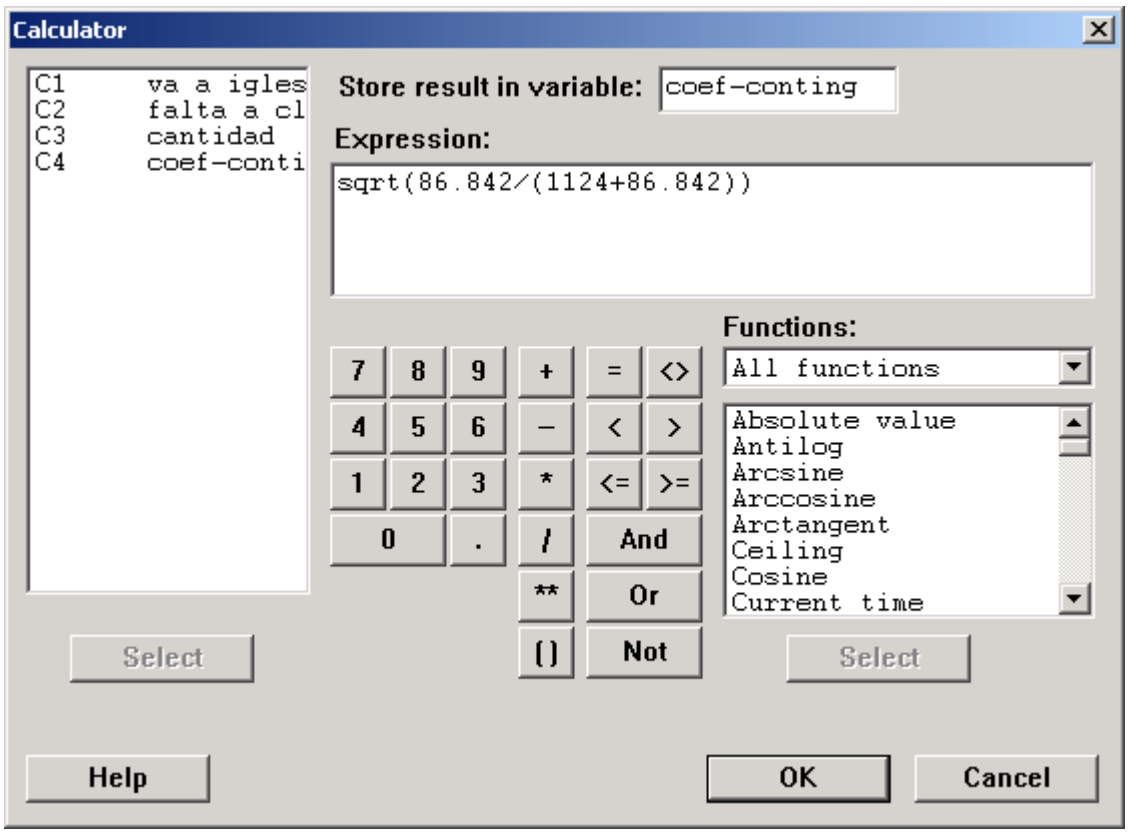

Figura 8.4. Ventana de diálogo de Calculator para hallar el coeficiente de contigencia del ejemplo 8.4

Data Display

```
coef-conting
   0.267807
```

***Interpretación**:*
No existe una buena asociación entre asistir a la iglesia y faltar a clases.

b) El Coeficiente de Cramer:

Se calcula por

$V = \sqrt{\frac{\chi^2}{nt}}$, donde t es el menor de los números r-1 y c-1, aqui r representa el número de filas y c el número de columnas. Si V=0 entonces, no hay asociación entre las variables. El coeficiente de Cramer si alcanza un máximo de 1. Un valor de V mayor .30 indica ya un cierto grado de asociación entre las variables. En el ejemplo anterior el coeficiente de Cramer es .1965, lo que reafirma que no existe buena asociación entre las variables.

MINITAB no calcula el coeficiente de contingencia directamente. Se tiene que usar **Calculator** del menú **CALC.**

Ejemplo 8.5. Calcular el coeficiente de Cramer para la siguiente tabla, donde se trata de relacionar las variables: sobrevivir a un ataque cardiaco y tener mascota ("pet").

Tabulated Statistics

```
Rows: status     Columns: pet?

            no       si      All

muere       11        3       14
          5.93     8.07    14.00

vive        28       50       78
         33.07    44.93    78.00

All         39       53       92
         39.00    53.00    92.00

Chi-Square = 8.851, DF = 1, P-Value = 0.003
```

En este caso r=2 y c=2, luego t es el menor de r-1=1 y c-1=1, asi t=1

La ventana de diálogo de **Calculator** se debe completar de la siguiente manera:

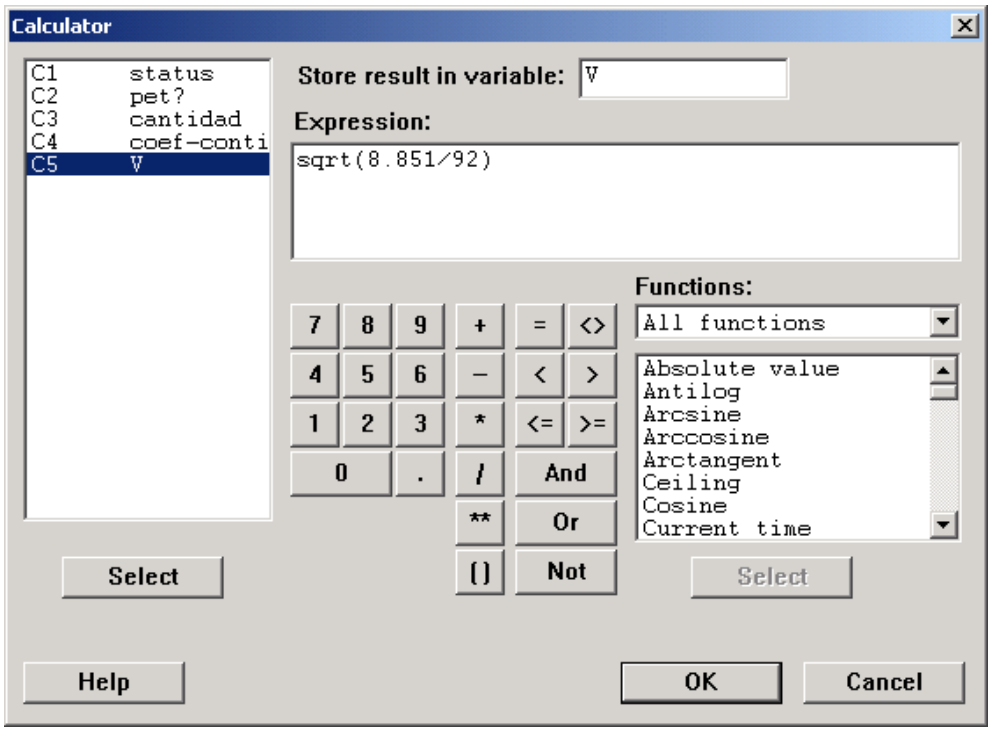

Figura 8.5. Ventana de diálogo de Calculator para hallar el coeficiente de Cramer del ejemplo 8.5.

Usando la secuencia **Manip** ‣ **Data Display**, se obtiene:

Data Display

```
V
0.310172
```

Por otro lado, el coeficiente de contingencia C resultó ser .3121

Intrepretación: *Se concluye que existe buena asociación entre tener mascota y sobrevivir a un ataque cardíaco.*

8.3. Prueba de Bondad de Ajuste

Otra aplicación de la prueba de Ji-Cuadrado, es la prueba de Bondad de Ajuste. Aquí se trata de probar si los datos de una muestra tomada siguen una cierta distribución predeterminada. Los n datos tomados deben estar divididos en categorias.

Categoría	1	2	3	…	K	
Frecuencia observada	Obs_1	Obs_2	Obs_3		Obs_k	N

Se asume que las probabilidades p_i , de caer en la categoría i deben ser conocidos.

La hipótesis nula es H_o: $p_1 = p_{10}$, $p_2 = p_{20} = \ldots = p_k = p_{k0}$, es decir los datos siguen la distribución deseada, y la hipótesis alterna es H_a: al menos una de las p_i es distinta de la probabilidad dada p_{i0}.
La prueba estadística es:

$$\sum_{i=1}^{k} \frac{(Obs_i - np_{io})^2}{np_{io}}$$

donde p_{io} representa la proporción deseada en la i-ésima categoría, Obs_i la frecuencia observada en la categoría i y n es el tamaño de la muestra. La prueba estadística se distribuye como una Ji-Cuadrado con k-1 grados de libertad donde, k es el número de categorias. Si el valor de la prueba estadística es mayor que $\chi^2_{1-\alpha}$ se rechaza la hipótesis nula.

MINITAB no tiene un comando que lleve a cabo la prueba de bondad de ajuste, pero ésta se puede efectuar escribiendo algunas lineas de comandos.

Ejemplo 8.6. Los siguientes datos representan los nacimientos por mes en PR durante 1993. Probar si hay igual probabilidad de nacimiento en cualquier mes del año. Usar un nivel de significación del 5%.

```
5435    4830    5229    4932    5052    5072    5198    5712
6126    5972    5748    5936
```

Solución:
La hipótesis nula es H_o: Hay igual probabilidad de nacer en cualquier mes del año (es decir, $p_1 = p_2 = \ldots = p_{12} = 1/12 = .083$). La hipótesis alterna es que no hay igual probabilidad de nacer en cualquier mes del año.
La ventana **Session** es como sigue:

```
MTB > let c3=sum(Obs)*p
MTB > let c4=(Obs-c3)**2/c3
MTB > let k5=sum(c4)
Esta es la prueba de Ji-Cuadrado para Bondad de ajuste
MTB > print k5

Data Display
K5    402.384
```

La ventana **Data** contendrá lo siguiente:

	C1	C2	C3	C4	C5	C6	C7
↓	Obs	p	np	(Obs-np)^2/np			
1	5435	0.08333	5436.62	0.0005			
2	4830	0.08333	5436.62	67.6860			
3	5229	0.08333	5436.62	7.9285			
4	4932	0.08333	5436.62	46.8374			
5	5052	0.08333	5436.62	27.2098			
6	5072	0.08333	5436.62	24.4536			
7	5198	0.08333	5436.62	10.4730			
8	5712	0.08333	5436.62	13.9492			
9	6126	0.08333	5436.62	87.4166			
10	5972	0.08333	5436.62	52.7233			
11	5748	0.08333	5436.62	17.8346			
12	5936	0.08333	5436.62	45.8713			
13							
14							

El valor de $\chi^2_{.95}$ con 11 grados de libertad es 19.6751, se encuentra usando la opción ***Probability distribution*** del menú ***Calc***

Interpretación: *Comparando el valor de la prueba estadística con una Ji-Cuadrado con 11 grados de libertad y nivel de significación del 5 por ciento que es 19.6751 se concluye que se rechaza la hipótesis nula, es decir no hay igual probabilidad de nacimiento para los meses.*

Ejemplo 8.7. Según el último censo se sabe que la distribución porcentual del estado marital de las personas adultas en los Estados Unidos es como sigue:

Soltero	Casado	Viudo	Divorciado
30	40	12	18

De acuerdo al censo de 1990, en Puerto Rico se tiene la siguiente distribución de personas adultas por estado marital:

Soltero	Casado	Viudo	Divorciado
811,291	1'279,628	198,553	189,346

Se desea establecer si la distribución del estado marital en Puerto Rico, es igual a la de los Estados Unidos. Usar un nivel de significación del 5%.

Solución:
La hipotesis nula Ho: Los datos tomados en Puerto Rico siguen la misma distribución de la de Estados Unidos, mientras que la hipótesis alterna Ha: Los datos no siguen la misma distribución.

Las ventanas **Session** y **Data** aparecerán como sigue:

```
MTB > Let 'np' = sum(obs)*p
MTB > Let '(Obs-np)^2/np' = (obs-np)**2/np
MTB > Let k5 = sum('(Obs-np)^2/np')
```

Esta es la prueba de Ji-Cuadrado

```
MTB > print k5
```

Data Display

```
K5    270598
```

Worksheet 1 ***

	C1	C2	C3	C4	C5	C6	C7
→	Obs	p	np	(Obs-np)^2/np			
1	811291	0.30	743645	6153			
2	1279628	0.40	991527	83711			
3	198553	0.12	297458	32886			
4	189346	0.18	446187	147847			
5							

Interpretación: *Claramente la prueba estadística es mayor que una Ji-Cuadrado con 3 grados de libertad al nivel de significación del 5 por ciento. Luego se rechaza la hipótesis nula y se concluye que la distribución del estado marital en Puerto Rico es distinta a la de Estados Unidos.*

Existen muchas otras pruebas de bondad de ajuste, especialmente pruebas noparamétricas.

EJERCICIOS

1. La siguiente tabla muestra los resultados de un estudio para mostrar la relación entre asistir a la iglesia los domingos y la ausencia a clases para jovenes entre 13 y 18 años:

	Falta a Clases		
Va a la Iglesia	Nunca	De vez en Cuando	Frecuentemente
Nunca	91	68	136
De vez en Cuando	140	78	119
Frecuentemente	296	106	90

   a) Usando la siguiente salida de **MINITAB**, probar la hipotesis de independencia entre faltar a clases e ir a la iglesia los domingos. En la salida deden aparecer los nombres de las filas y columnas

   b) ¿Cuál es la frecuencia esperada de los que nunca van a la Iglesia y faltan frecuentemente a clase?. Explicar cómo se calcula e interpretarlo.

2. El consumo de alcohol y nicotina (cigarrillos) durante el periodo de gestación puede afectar al bebé. Se hizo un estudio en 452 madres y se las clasificó de acuerdo a su consumo de alcohol (medido en onzas por dia), y al de nicotina (medida en miligramos por dia). Los datos están en el archivo **alcohoynico**, que está disponible en la pagina de internet del texto.

   a) Usando la salida de **MINITAB**, probar la hipótesis de independencia entre el consumo de alcohol y nicotina. En la salida deben aparecer los nombres de las filas y columnas
   b) Escribir la hipótesis de homogeneidad.
   c) ¿Cuál es la frecuencia esperada de las madres que consumen 1 onza o más por día y no fuman. Cómo se calcula dicho valor e Interpretar el significado de dicho valor.

3. En una ciudad se hace un estudio para relacionar los hábitos de fumar de los estudiantes de escuela superior con las de sus padres. Los resultados que se obtienen aparecen en la siguiente tabla:

	Estudiante Fuma	Estudiante no Fuma
Ambos padres fuman	400	1380
Sólo uno de los padres fuma	416	1823
Ninguno de los padres fuma	188	1168

   a) Calcular la proporción de estudiantes que fuman para cada uno de los grupos de padres. ¿Qué puede concluir de estos resultados?

b) Calcular las frecuencias esperadas de cada celda de la tabla si no hubiera relación entre los hábitos de fumar de los estudiantes con las de sus padres.
c) Probar la hipótesis de que no hay relación entre los hábitos de fumar de los estudiantes con los de sus padres.

4. La siguiente tabla reporta información acerca del sexo, status económico de la mayoría de los pasajeros del TITANIC, un crucero británico de lujo que se hundió en 1912.

	Hombres		Mujeres	
Status	Murió	Sobrevivió	Murió	Sobrevivió
Alto	111	61	6	126
Medio	150	22	13	40
Bajo	419	85	107	101
Total	680	168	126	317

a) ¿Hay suficiente evidencia para concluir que la proporción de hombres que murieron fue mayor que el de las mujeres?
b) Para cada uno de los sexos, probar si hay relación entre el status económico del pasajero y si sobrevivió o nó al hundimiento.

5. Las encuestas sobre asuntos sensitivos pueden dar diferentes resultados dependiendo de como se hace la pregunta. Se hace una encuesta a 2400 personas para estimar el uso de cocaína. Se dividieron al azar a los encuestados en 3 grupos de 800 cada uno, y se les preguntó si alguna vez habían usado cocaína El primer grupo fue entrevistado por teléfono, y 21% dijeron que habían usado cocaina. El Segundo grupo fue entrevistado personalmente, y 25% dijeron que habían usado cocaína. En el tercer grupo, donde se permitió una respuesta escrita anónima, el 28% contestaron positivamente a la pregunta.
Probar si hay efecto del método de hacer la pregunta en la estimación de la proporción de usuarios de cocaína.

6. En una ciudad se hace una encuesta a 103 personas entre los 25 y 30 años acerca de su estado marital. Los resultados están resumidos en la siguiente tabla:

Estado Marital	Hombre	Mujer
Nunca Casado	20	9
Casado	19	39
Viudo, Divorciado, Separado	9	7

a) ¿Piensa Ud. que la distribución del estatus marital es la misma para ambos sexos?.
b) Si las distribuciones son diferentes, con quiénes se están casando las mujeres?

7. En un estudio acerca de hábitos de fumar de los estudiantes de una universidad realizado en 1990, se reportó que 40 % de los fumadores proceden de la facultad de Administración de Empresas, 30 % de la facultad de Artes y Ciencias, 25% de Ingeniería, y un 5% de Agricultura. Un estudiante de la clase de Estadística quiere comprobar si esos porcentajes se mantienen aún en 1998 para ello toma una muestra de estudiantes fumadores de las distintas facultades de la universidad y obtiene los siguientes resultados:

Empresas	Artes y Ciencias	Ingenieria	Agricultura
45	40	22	8

Usar un nivel de significación del 1%.

8. La siguiente tabla reporta la distribución de la población de un país de acuerdo a su nivel educacional y el número de alcaldes elegidos en cada una de las categorías en las últimas elecciones:

Nivel Educacional	País	Alcaldes electos
Elemental	30%	6
Secundaria	45%	15
Universitaria Incompleta	12%	27
Universitaria Completa	13%	30

¿Habrá suficiente evidencia para concluir que la distribución del nivel educacional de los alcaldes electos sigue la misma distribución del país?. Usar un nivel de significación del 5%.

9. Un Sociólogo piensa que hay más probabilidad de que un crimen ocurra durante los fines de semana. En particular él piensa que la probabilidad de que un crimen ocurra el sábado es igual a la probabilidad de que un crimen ocurra el domingo, y éstas a su vez son el doble de probabilidad de que un crimen ocurra un dia de semana. Para probar su afirmación usa los siguientes datos de crímenes ocurridos en un mes cualquiera del año.

Lunes	Mártes	Miércoles	Jueves	Viernes	Sábado	Domingo
18	23	19	16	21	42	37

Usar un nivel de significación del 1%.