Tablas de contingencia

Tablas de contingencia

Se sabe que la información proporcionada por una tabla bidimensional puede expresarse en términos diversos: frecuencias absolutas conjuntas, relativas conjuntas, condicionadas de una variable a valores de la otra. Además puede derivarse el comportamiento unidimensional de las variables implicadas mediante las distribuciones marginales.

La tabla bidimensional recibe el nombre de tabla de contingencia cuando las características en estudio no son cuantitativas.

Una tabla de doble entrada para las variables X e Y con p filas y k columnas:

	X1	**X2**	**...**	**Xk**
Y1	n_{11}	n_{12}	...	n_{1k}
Y2	n_{21}	n_{22}	...	n_{2k}
...	...	...	...	...
Yp	n_{p1}	n_{p2}	...	n_{pk}

donde **n**ij expresa la frecuencia absoluta observada en las modalidades Xi e Yj refleja la distribución conjunta de X e Y.

La misma tabla puede expresarse en frecuencias relativas o proporciones sin más que dividir cada casilla n_{ij} por el total N. $N = \sum_{j=1}^{k} \sum_{i=1}^{p} n_{ij}$

1.Distribuciones condicionadas de Y a los valores de X

Son distribuciones unidimensionales para la variable Y en distintas condiciones (valores de X). Se obtienen de la tabla anterior dividiendo cada casilla por el total de columna.

2.Distribuciones condicionadas de X a los valores de Y

Son distribuciones unidimensionales para la variable X en distintas condiciones (valores de Y). Se obtienen de la tabla anterior dividiendo cada casilla por el total de fila.

3.Distribuciones marginales:

Marginal de Y

Distribución unidimensional formada por los valores Yi (i=1, ..., p) cuya frecuencia asociada se obtiene sumando las casillas correspondientes a la fila i-ésima.

Marginal de X

Distribución unidimensional formada por los valores Xj (j=1, ..., k) cuya frecuencia asociada se obtiene sumando las casillas correspondientes a la columna j-ésima.

4.Ejemplo 1 (archivo en carpeta ARCHIVOS TEMA2)

Doce individuos se clasificaron según el sexo (hombre, mujer) y su deseo de ver o no una final de campeonato de fútbol que será televisada:

Dos formas de presentar los datos:
a) Tabulados:

Tabla de contingencia desea ver partido * SEXO

Recuento

		SEXO		
		hembra	varon	Total
desea ver partido	si	1	6	7
	no	4	1	5
Total		5	7	12

b) Sin tabular:

Sexo	Futbol
hombre	si
mujer	no
hombre	si
hombre	no
hombre	si
mujer	no
mujer	no
mujer	si
hombre	si
hombre	si
hombre	si
mujer	no

Obtenga:
a) Tabla de contingencia
b) Exprese la tabla anterior con frecuencias relativas (en porcentajes)
c) Determine las condicionadas de Futbol a Sexo
d) Marginales
e) Test de independencia de sexo y futbol

Introducción de los datos

En la ventana del editor de datos se definen dos columnas de nombres sexo y futbol, ambas de tipo cadena (medida nominal).
Para **sexo**, seleccione **tipo** cadena. Introduzca los valores h y m en vez de hombre mujer. Luego , introduzca las etiquetas de las modalidades hombre y mujer, respectivamente.
De modo similar introduzca las etiquetas: SI y NO de la variable **fútbol** para los valores 1 y 2, respectivamente. Luego etiquete los datos.

Y guarde el data frame en un archivo de nombre ejemplo1.

datos=edit(data.frame())
write.table(datos, file="ejemplo1")

Si los archivos están ya creados. Ábralos en un data frame de nombre datos.

>datos=read.table('ejemplo1.dat', header=T)

> datos

```
   sexo futbol
1     h      1
2     m      2
3     h      1
4     h      2
5     h      1
6     m      2
7     m      2
8     m      1
9     h      1
10    h      1
11    h      1
12    m      2
```

datos$sexo=factor(datos$sexo, labels=c("hombre", "mujer")) #Declara factor con etiquetas

datos$futbol= factor(datos$futbol, labels=c("si", "no")) #Declara factor con etiquetas

> datos

```
     sexo futbol
1  hombre     si
2   mujer     no
3  hombre     si
4  hombre     no
5  hombre     si
6   mujer     no
7   mujer     no
8   mujer     si
9  hombre     si
10 hombre     si
11 hombre     si
12  mujer     no
```

Tabla de contingencia:

> ftable(datos$sexo,datos$futbol)

```
        si no
hombre   6  1
mujer    1  4
```

o bien, usando el data frame:

> ftable(datos)

```
       futbol si no
sexo
hombre         6  1
mujer          1  4
```

Marginales:

>td= ftable(datos)
> addmargins(td)

```
     [,1] [,2] [,3]
[1,]    6    1    7
[2,]    1    4    5
[3,]    7    5   12
```

Mejor presentación si se usa previamente **table** en vez de **ftable**:

```
> td1=table(datos)
> addmargins(td1)
```

```
       futbol
sexo     si no Sum
  hombre  6  1   7
  mujer   1  4   5
  Sum     7  5  12
```

Expresión en proporciones:

Distribución bidimensional conjunta en frecuencias relativas:

```
> prop.table(td)
```

```
       futbol         si         no
sexo
hombre        0.50000000 0.08333333
mujer         0.08333333 0.33333333
```

Expresión en proporciones: Condicionadas de futbol a valores del sexo

```
> prop.table(td1,1)
```

```
        futbol
sexo             si        no
  hombre 0.8571429 0.1428571
  mujer  0.2000000 0.8000000
```

Expresión en proporciones: Condicionadas de sexo a valores del futbol

```
> prop.table(td1,2)
```

```
        futbol
sexo             si        no
  hombre 0.8571429 0.2000000
  mujer  0.1428571 0.8000000
```

5.Estudio de la asociación

Sean **X** e **Y** dos características, cualitativas o cuantitativas, con i=1, ..., p y j=1, ...q modalidades o categorías, respectivamente, presentadas en una tabla pxq.

Una de las medidas de asociación más usadas en la práctica es:

CHI-CUADRADO

Medida resumen que compara los valores ($\mathbf{n_{ij}}$) observados en la tabla, con los que teóricamente se obtendría ($\mathbf{t_{ij}}$), en el supuesto de que las variables **X** e **Y** fuesen independientes.

$$\chi^2 = \sum_{i}^{p}\sum_{j}^{q}\frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Los valores teóricos $\mathbf{t_{ij}}$ se obtienen mediante:

$t_{ij} = \frac{n_{i*} n_{*j}}{N}$ siendo n_{i*} y n_{*j} las frecuencias marginales

Este estadístico toma valores comprendidos entre **0** y **N·mín{p-1, q-1}**, el valor 0 indica que el numerador de la expresión anterior es nulo, por tanto las frecuencias observadas coinciden con las que habría si las variables fuesen independientes; de donde se admite la independencia de **X** e **Y**. El hecho de que sus valores dependan tanto del número de elementos de la tabla (**N**), como del nº de filas y columnas, hace difícil su interpretación e impracticable la comparación entre tablas.

El estadístico Chi-cuadrado permite contrastar la hipótesis de independencia de **X** e **Y**, basándose en el conocimiento del comportamiento de Chi-cuadrado bajo la hipótesis de independencia: Modelo **Chi-cuadrado** con **(p-1)(q-1)** grados de libertad.

6.Analizar Tablas de contingencia (Continuación con el ejemplo1)

Test chi-cuadrado de independencia de factores:

El estadístico Chi-cuadrado de Pearson seguirá el modelo Chi-cuadrado con (p-1)(q-1) g.l. si N es suficientemente grande. Cuando N es pequeño se intenta mejorar el comportamiento efectuando una corrección, que suele ser complicada para tablas generales pxq, con p y q mayores a 2.

R proporciona la corrección por continuidad para tablas 2x2 y la prueba exacta de Fisher, que aporta mejores resultados.

El sistema avisa sobre la proporción de casillas que presentan valores esperados inferiores a 5. Si la proporción supera al 20% de las celdas, el estadístico Chi-cuadrado no cumple los requisitos necesarios para poder interpretarlo sin problemas. En este ejemplo el 100% de las casillas presentan valores inferiores a 5, en cuyo caso la interpretación de su valor no merece confianza. No obstante, pueden usarse otras pruebas, tales como el estadístico exacto de Fisher. Cuando las frecuencias esperadas son menores que 5, en tablas 2x2, será aconsejable el uso del test exacto de Fisher. Si lo que se desea contrastar es la independencia se tomará el p-valor correspondiente a dos colas (significación bilateral). (vea en el ejemplo: 0,072).

La prueba exacta de Fisher se basa en el modelo de la distribución hipergeométrica, para estimar la probabilidad de obtener las frecuencias observadas en la tabla, u otras frecuencias menos consistentes con la hipótesis de independencia, correspondientes a situaciones aún más extremas que la observada.

Dado que **0,072 < α = 0,1** se rechaza la hipótesis de no asociación o independencia entre las variables al nivel alfa del 10%..

```
> chisq.test(table(datos))

        Pearson's Chi-squared test with Yates' continuity correction

data:  table(datos)
X-squared = 2.831, df = 1, p-value = 0.09246
```

```
Warning message:
In chisq.test(table(datos)) : Chi-squared approximation may be incorrect
```

Como es una tabla 2x2 con pocas observaciones, realizaremos también el test exacto de Fisher

```
> fisher.test(table(datos))  #Realiza el test de independencia exacto de Fisher

        Fisher's Exact Test for Count Data

data:  table(datos)
p-value = 0.07197
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.747344 1351.222783
sample estimates:
odds ratio
  15.99491
```

Valores esperados bajo independencia, observados y residuos estandarizados (observado menos esperado entre la raiz cuadrada del valor esperado):

```
> a$expected

        futbol
sexo           si       no
  hombre 4.083333 2.916667
  mujer  2.916667 2.083333

> a$observed

        futbol
sexo     si no
  hombre  6  1
  mujer   1  4

> a$residuals

        futbol
sexo            si         no
  hombre  0.948504 -1.122285
  mujer  -1.122285  1.327906
```

7.Ejemplo 2

La tabla siguiente clasifica a un grupo de personas según su opinión sobre un documental televisivo y el nivel de estudios:

Tabla de contingencia Nivel de estudios y opinión sobre documental

Recuento

		opinión sobre documental			
		malo	regular	bueno	**Total**
Nivel de estudios	bajo	1	10	30	41
	medio	40	80	60	180
	alto	25	12		37
Total		66	102	90	258

Cree un archivo con los datos anteriores, definiendo las variables estudios (nivel de estudios) y opinión (opinión sobre documental).

A) Exprese las frecuencias en proporciones a) respecto al total (frecuencias relativas conjuntas) y b) respecto a la variable independiente nivel de estudios (condicionadas de opinión a estudios). Interprete sus valores. ¿Cómo han de efectuarse las comparaciones entre las proporciones para establecer la posible existencia de asociación?

B) Contraste la hipótesis de independencia del nivel de estudios y opinión sobre el documental.

C) Caso de resultar dependientes las variables, determine algunas medidas del grado de asociación.

El archivo de datos creado en el editor de R presentará un aspecto similar a:

Estudios	**Opinion**	**numper**
bajo	malo	1
medio	malo	40
alto	malo	25
bajo	regular	10
medio	regular	80
alto	regular	12
bajo	bueno	30
medio	bueno	60
alto	bueno	0

Vea el archivo ejemplo2.dat en carpeta ARCHIVOS TEMA2
Lea el archivo:

```
> d=read.table('ejemplo2.dat',header=T)
> d
```

```
  estudios opinión numper
1        1       1      1
2        2       1     40
3        3       1     25
4        1       2     10
5        2       2     80
6        3       2     12
7        1       3     30
8        2       3     60
```

Colocaremos etiquetas a los códigos de las modalidades de los factores:

```
> d$estudios=factor(d$estudios, labels=c("bajo", "medio", "alto"))

> d$opinión=factor(d$opinión, labels=c("malo","regular",'bueno'))

> d
```

```
  estudios opinión numper
1     bajo    malo      1
2    medio    malo     40
3     alto    malo     25
4     bajo regular     10
5    medio regular     80
6     alto regular     12
7     bajo   bueno     30
8    medio   bueno     60
```

Antes de comenzar el análisis de la tabla de contingencia es preciso tener en cuenta que los datos están tabulados con las frecuencias (**numper**).

```
> xtabs(numper ~ ., d)


        opinión
estudios malo regular bueno
   bajo     1      10    30
   medio   40      80    60
   alto    25      12     0

> tabla=xtabs(numper ~ ., d)
> tabla
        opinión
estudios malo regular bueno
   bajo     1      10    30
   medio   40      80    60
   alto    25      12     0
```

A) Las tablas que se muestran a continuación representan las proporciones o frecuencias relativas conjuntas y las proporciones condicionadas de opinión sobre estudios.

TABLA 1: Distribución bidimensional de Estudios y Opinión. Frecuencias relativas

```
> #conjunta
> prop.table(tabla)

      opinión
estudios        malo     regular       bueno
   bajo  0.003875969 0.038759690 0.116279070
   medio 0.155038760 0.310077519 0.232558140
   alto  0.096899225 0.046511628 0.000000000
```

La tabla expresa el **comportamiento conjunto** de los individuos atendiendo a dos dimensiones: **estudios** y **opinión**.

```
> round(prop.table(tabla),3)
    opinión
estudios  malo regular bueno
   bajo  0.004   0.039 0.116
   medio 0.155   0.310 0.233
   alto  0.097   0.047 0.000
```

La tabla presenta en cada casilla la frecuencia conjunta (en proporciones) respecto al total. Cada valor representa la proporción de veces que aparece cada valor bidimensional en la población total. Por ejemplo, podemos afirmar que el 23,3% (12/258 x100) del total de individuos del análisis tienen estudios medios y califican el documental como bueno.

```
> tabla2=round(prop.table(tabla),3)
> addmargins(tabla2)

        opinión
estudios  malo regular bueno   Sum
   bajo  0.004   0.039 0.116 0.159
   medio 0.155   0.310 0.233 0.698
   alto  0.097   0.047 0.000 0.144
   Sum   0.256   0.396 0.349 1.001
>
```

La fila y la columna **Sum** representan las **frecuencias marginales**. Por ejemplo, la fila de frecuencias relativas Sum indica cómo se distribuye la variable opinión para los 258 individuos, sin tener en cuenta su nivel de estudios.

Condicionadas

La tabla de distribuciones condicionadas de opinión/estudios se obtiene determinando en vez de proporción respecto al total, respecto a **la suma de cada fila**:

```
> prop.table(tabla,1)

        opinión
estudios       malo    regular      bueno
   bajo  0.02439024 0.24390244 0.73170732
   medio 0.22222222 0.44444444 0.33333333
   alto  0.67567568 0.32432432 0.00000000
```

A diferencia de la tabla 1, ésta presenta no sólo una distribución, sino 3. Mientras la primera tabla es bidimensional (distribución conjunta de opinión y estudios), aquí sólo tenemos conocimiento sobre la distribución de una dimensión: **opinión**. Nada sabemos acerca de cómo se distribuye el nivel de estudios.

La tabla nos indica cómo se **distribuye la opinión** en el grupo de individuos con nivel de estudios bajo, con nivel medio y con nivel alto.
Si las variables opinión y estudios fueran independientes, los individuos opinarían de modo similar, que es tanto como afirmar que cualquiera que sea su nivel de estudios, la distribución de la opinión es la misma: distribuciones condicionadas de opinión a estudios son iguales.
En la medida en que las distribuciones condicionas se diferencien nos estaremos alejando del concepto de independencia y habrá que admitir que las variables están asociadas.

Observe que en la tabla condicionada anterior, los cálculos se han efectuado en sentido horizontal (dividiendo cada casilla entre el total de fila); por tanto, la lectura deberá efectuarse en sentido vertical: comparando las casillas por columnas. Por ejemplo:0,024 con 0,226 con 0,676. Las grandes diferencias que existen entre estas proporciones no pueden deberse al azar. Cabe pensar que las variables están asociadas. El nivel de estudios afecta a la opinión.

Tabla de condicionadas de estudios a opinión:

```
> prop.table(tabla,2)
        opinión
estudios       malo    regular      bueno
   bajo  0.01515152 0.09803922 0.33333333
   medio 0.60606061 0.78431373 0.66666667
   alto  0.37878788 0.11764706 0.00000000
```

B) Para responder con cierto rigor científico a la cuestión de existencia o no de asociación, efectuaremos un contraste de hipótesis mediante la prueba Chi-cuadrado. Admitamos un nivel de significación, alfa, igual a 0,05.

```
> a=chisq.test(tabla)
> a

        Pearson's Chi-squared test

data:  tabla
X-squared = 69.0831, df = 4, p-value = 3.544e-14
```

> a$expected

```
        opinión
estudios      malo  regular    bueno
   bajo  10.488372 16.20930 14.30233
   medio 46.046512 71.16279 62.79070
   alto   9.465116 14.62791 12.90698
```

> a$observed

```
        opinión
estudios malo regular bueno
   bajo     1      10    30
   medio   40      80    60
   alto    25      12     0
```

> a$residuals

```
        opinión
estudios       malo    regular      bueno
   bajo  -2.9297979 -1.5422708  4.1508017
   medio -0.8910591  1.0475835 -0.3521804
   alto   5.0494611 -0.6870983 -3.5926281
```

El valor del estadístico **chi-cuadrado** se obtiene de la expresión:

$$\chi^2 = \sum_i^p \sum_j^q \frac{(n_{ij} - t_{ij})^2}{t_{ij}}$$

Las frecuencias teóricas t_{ij} se obtienen mediante:

$t_{ij} = \frac{n_{i*} n_{*j}}{N}$ siendo n_{i*} y n_{*j} las frecuencias marginales

Por ejemplo:

$$t_{11} = \frac{n_{1*} n_{*1}}{N} = \frac{66 \cdot 41}{258} = 10{,}5$$

$$t_{12} = \frac{n_{1*} n_{*2}}{N} = \frac{66 \cdot 180}{258} = 46{,}0$$

....

$$t_{33} = \frac{n_{3*} n_{*3}}{N} = \frac{90 \cdot 37}{258} = 12{,}9$$

De donde:

$$\chi^2 = \frac{(1-10{,}5)^2}{10{,}5} + \frac{(40-46{,}0)^2}{46{,}0} + \ldots + \frac{(0-12{,}9)^2}{12{,}9} = 69{,}083$$

Bajo la hipótesis nula:

H_0: Las variables estudios y opinión son independientes

el estadístico Chi-cuadrado sigue un modelo de probabilidad Chi-cuadrado con (p-1)(q-1) grados de libertad. Siendo p=número de filas y q=número de columnas.

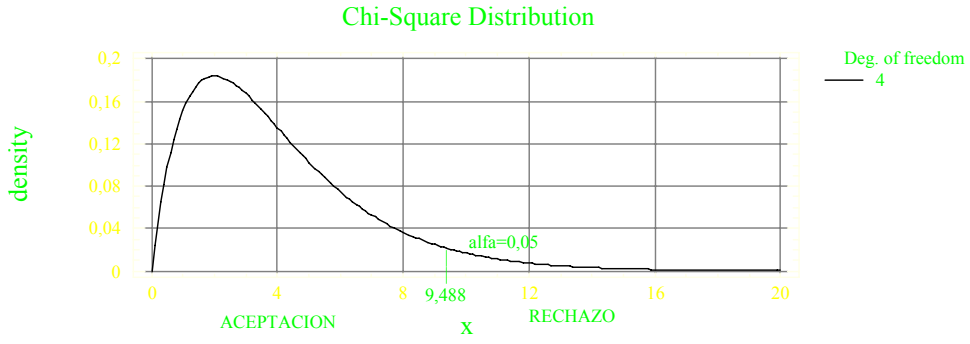
El gráfico siguiente muestra la función de densidad de dicha distribución. Observemos que los valores más probables están comprendidos entre 0 y 9, la cola de la derecha se va estrechando cuanto más nos alejamos a la derecha. La probabilidad de que la variable tome valores superiores a 69,083 es prácticamente nula 0,000.

La decisión de aceptar o rechazar la hipótesis nula se toma en función del valor obtenido para chi-cuadrado en la tabla. Si éste cae en la zona de rechazo se decide rechazarla; caso contrario, se acepta. El valor crítico que define la frontera de rechazo y aceptación es el punto 9,488, ya que puede comprobarse que $P(\chi^2>9,488)=0,05$. Es evidente que el valor 69,083 cae en la zona de rechazo, pues está a la derecha de 9,488. Si H_0 es cierta, es muy improbable que la variable tome el valor 69,083. Por tanto, decidimos rechazar la hipótesis de independencia.

R proporciona la probabilidad: $P(\chi^2>69,083)=0,000$ denominado **p-valor**. Si éste es menor que el nivel de significación elegido para contrastar la hipótesis, es porque cae en la zona de rechazo. En resumen, la decisión se toma comparando el p-valor con el nivel de significación alfa:

Si p-valor < α RECHAZO H_0
Si p-valor > α NO RECHAZO H_0

En el ejemplo el p-valor asociado a 69,083 es igual a 3.544e-14 < 0,05. Por tanto, se rechaza la hipótesis de independencia.



8.Clasificación múltiple: Análisis de Tablas multidimensionales

El control de las variables como una emulación de la experimentación:
En el campo social es difícil la manipulación de las variables que interesan. El investigador asigna categorías a las variables independientes pero no controla la asignación de los sujetos a esas categorías, esto hace más difícil los estudios de causalidad. En un experimento, donde se sospecha que la dosis de un fármaco (variable independiente) es importante para explicar la evolución de una enfermedad (variable dependiente), el experimentador puede asignar aleatoriamente grupos de individuos a las distintas categorías de la dosis.

En la investigación no experimental, las técnicas multivariantes y la tabulación múltiple (empleada fundamentalmente para tratamiento de datos cualitativos) ofrecen la posibilidad de suplir, en cierto grado, las limitaciones señaladas anteriormente.

La lógica del tratamiento se basa en desglosar la relación original entre dos variables X e Y en relaciones **condicionadas**, considerando una tercera variable denominada factor test (o variable de control). Es decir, estudiar una misma relación en diferentes contextos. Es posible que las variables X e Y que se manifiestan aparentemente relacionadas, respondan realmente a la convergencia de dos hechos. También podemos encontrarnos con la situación de que la relación original desaparezca, o que se intensifique, o que emerjan relaciones de naturaleza distinta para cada valor de la variable test.

Para generar una tabla multidimensional con R se utilizan las mismas funciones que para tablas bidimensionales, salvo que habrá de especificar las variables de control.

9.Ejemplo 3

La tabla siguiente representa la distribución bidimensional de un grupo de 11137 trabajadores clasificados según la EDAD y el SALARIO que perciben (estos mismos datos se analizarán teniendo en cuenta otra variable de clasificación, *tipo de trabajo*, en el ejemplo 4)

SALARIO (Miles)	EDAD		
	18-25	25-35	35-65
20-50	335	1022	2132
50-100	402	1429	2427
100-150	38	841	2511

A) Marginales
B) Condicionadas del Salario a la Edad
C) Estudio de la independencia mediante Chi-cuadrado

Usaremos tres columnas de nombres salario, edad y numperso para introducir los datos de la tabla en un archivo.

Salario	Edad	Numperso
20-50	18-25	335
50-100	18-25	402
100-150	18-25	38
20-50	25-35	1022
50-100	25-35	1429
100-150	25-35	841
20-50	35-65	2132
50-100	35-65	2427
100-150	35-65	2511

```
> d=read.table('eje3.dat',header=T)
> d
  salario edad numperso
1      35 21,5      335
2      75 21,5      402
3     125 21,5       38
```

```
4      35   30     1022
5      75   30     1429
6     125   30      841
7      35   50     2132
8      75   50     2427
9     125   50     2511
> d$salario=factor(d$salario, labels=c("bajo", "medio", "alto"))
> d$edad=factor(d$edad, labels=c("joven","medio",'mayor'))
> d
  salario  edad numperso
1    bajo joven      335
2   medio joven      402
3    alto joven       38
4    bajo medio     1022
5   medio medio     1429
6    alto medio      841
7    bajo mayor     2132
8   medio mayor     2427
9    alto mayor     2511

> tabla=xtabs(numperso ~ ., d)

> tabla
    edad
salario joven medio mayor
  bajo    335  1022  2132
  medio   402  1429  2427
  alto     38   841  2511
```

La **distribución marginal** del salario está formada por las clases salariales (20-50, 50-100, 100-150) (etiquetadas con bajo medio y alto) y las correspondientes frecuencias en la columna Sum.

La distribución marginal está formada por las clases de la edad (18-25, 25-35, 35-65) (etiquetadas como joven, medio y mayor) y las correspondientes frecuencias en la fila Sum.

Distribución condicional del salario a la edad de 18-25 años:
Está formada por los valores del salario y los porcentajes de la primera columna (18-25 años): 43,2, 51,9 y 4,9 que representan las frecuencias relativas, multiplicadas por 100, correspondientes a los valores salariales.
De modo similar se obtienen las condicionadas del salario a los otros valores de la edad.
Observe que las frecuencias se obtienen dividiendo cada casilla por el total de columna.

Se rechaza la hipótesis de independencia del salario y la edad. Según la tabla siguiente el p-valor asociado al estadístico Chi-cuadrado es 0,000 altamente significativo.

```
> a=chisq.test(tabla)
> a

        Pearson's Chi-squared test

data:  tabla
X-squared = 378.9477, df = 4, p-value < 2.2e-16
```

Clasificación múltiple:

10.Ejemplo 4

Supongamos que los datos del ejemplo anterior se han clasificado ahora atendiendo a 3 variables. Nos interesa estudiar el salario y su relación con otros factores que ayuden a

interpretar la relación que se puso de manifiesto entre edad y salario. Introducimos el factor de control tipo de trabajo (manual e intelectual).

SALARIO (Miles)	manual			intelectual		
	18-25	18-25	25-35	18-25	18-25	25-35
20-50	165	644	1800	170	378	332
50-100	168	672	1763	234	757	664
100-150	17	84	187	21	757	2234

A) Condicionadas del Salario a la Edad, controlando por tipo de trabajo.
B) Estudio de la independencia del Salario y Edad, mediante Chi-cuadrado

---

El archivo de datos contendrá las siguientes columnas:

Salario	**Edad**	**tipotra**	**numperso**
20-50	25-35	manual	644
50-100	25-35	manual	672
100-150	25-35	manual	84
20-50	35-65	manual	1800
50-100	35-65	manual	1763
100-150	35-65	manual	187
20-50	18-25	intelectual	170
50-100	18-25	intelectual	234
100-150	18-25	intelectual	21
20-50	25-35	intelectual	378
50-100	25-35	intelectual	757
100-150	25-35	intelectual	757
20-50	35-65	intelectual	332
50-100	35-65	intelectual	664
100-150	35-65	intelectual	2234

```
> d=read.table('eje4.dat',header=T)
> d
   salario edad tipotrab numperso
1       35 21,5        1      165
2       75 21,5        1      168
3      125 21,5        1       17
4       35   30        1      644
5       75   30        1      672
6      125   30        1       84
7       35   50        1     1800
8       75   50        1     1763
9      125   50        1      187
10      35 21,5        2      170
11      75 21,5        2      234
12     125 21,5        2       21
13      35   30        2      378
14      75   30        2      757
15     125   30        2      757
16      35   50        2      332
17      75   50        2      664
18     125   50        2     2234
```

```
> d$tipotrab=factor(d$tipotrab, labels=c("manual","intelectual"))
```

```
> d
   salario edad     tipotrab numperso
1       35 21,5       manual      165
2       75 21,5       manual      168
3      125 21,5       manual       17
4       35   30       manual      644
5       75   30       manual      672
6      125   30       manual       84
7       35   50       manual     1800
8       75   50       manual     1763
9      125   50       manual      187
10      35 21,5 intelectural      170
11      75 21,5 intelectural      234
12     125 21,5 intelectural       21
13      35   30 intelectural      378
14      75   30 intelectural      757
15     125   30 intelectural      757
16      35   50 intelectural      332
17      75   50 intelectural      664
18     125   50 intelectural     2234
```

```
> tab1=ftable(xtabs(numperso ~ edad+salario, subset=tipotrab==1 ,data = d))
> chisq.test(tab1)

        Pearson's Chi-squared test

data:  tab1
X-squared = 3.2136, df = 4, p-value = 0.5227
```

```
> tab1=ftable(xtabs(numperso ~ edad+salario, subset=tipotrab==2 ,data = d))
> chisq.test(tab1)

        Pearson's Chi-squared test

data:  tab1
X-squared = 882.5047, df = 4, p-value < 2.2e-16
```

En las condicionadas de **edad x salario,** dado el tipo de trabajo, en frecuencias absolutas, no permiten apreciar directamente la relación entre las variables:

```
> tabla=xtabs(numperso ~ ., d)
> tabla
, , tipotrab = manual

       edad
salario 21,5   30   50
    35   165  644 1800
    75   168  672 1763
    125   17   84  187

, , tipotrab = intelectual

       edad
salario 21,5   30   50
    35   170  378  332
    75   234  757  664
    125   21  757 2234
```

Estudiaremos para cada tipo de trabajo (manual, intelectual) si existe o no asociación entre sexo y salario:

```
> tab1=ftable(xtabs(numperso ~ edad+salario, subset=tipotrab=="manual" ,data = d))
> chisq.test(tab1)

        Pearson's Chi-squared test

data:  tab1
X-squared = 3.2136, df = 4, p-value = 0.5227
```

```
> tab1
     salario   35   75  125
```

```
edad
21,5          165  168   17
30            644  672   84
50           1800 1763  187
```

La tabla anterior muestra que para el grupo de trabajadores "manual" no existe asociación entre salario y edad. Tal como muestra el p-valor correspondiente al contraste de hipótesis de independencia de salario y edad.
La relación entre salario y edad presenta un nivel de significación igual a 0,523 > 0,05. No puede rechazarse la hipótesis de independencia del salario y edad para los trabajadores de la categoría manual.

Por el contrario, sí se aprecia fuerte relación entre las variables sexo y salario para el tipo de trabajador "intelectual", tal como muestra el resultado siguiente. La relación es altamente significativa:
El nivel de significación (2.2e-16) permite rechazar la hipótesis de independencia.

```
> tab1=ftable(xtabs(numperso ~ edad+salario, subset=tipotrab=="intelectual" ,data = d))
> chisq.test(tab1)

        Pearson's Chi-squared test

data:  tab1
X-squared = 882.5047, df = 4, p-value < 2.2e-16

> tab1
     salario   35   75  125
edad
21,5           170  234   21
30             378  757  757
50             332  664 2234
```

Para completar el análisis mostramos las condicionadas relativas a salario y sexo, distinguiendo por tipo de trabajador:

1. Condicionada del salario a la edad, para el grupo de trabajadores intelectuales:
La lectura de la tabla debe realizarse verticalmente, dado que los cálculos se han realizado dividiendo por los totales fila (redondeando a centésimas, comparamos, por ejemplo, 0,40 con 0,20, con 0,10, lo que demuestra tal como demuestra el test chi-cuadrado, la fuerte relación entre las variables en este colectivo)

```
> prop.table(tab1,1)
     salario         35         75        125
edad
21,5          0.40000000 0.55058824 0.04941176
30            0.19978858 0.40010571 0.40010571
50            0.10278638 0.20557276 0.69164087
```

De modo similar se puede ver la tabla condicionada de salario a edad para el grupo de trabajadores de tipo manual. Aquí la asociación entre sexo y salario no es importante, tal como muestra la tabla siguiente y el test chi-cuadrado, previamente realizado:

```
> prop.table(tab1,1)
     salario         35         75        125
edad
21,5          0.47142857 0.48000000 0.04857143
30            0.46000000 0.48000000 0.06000000
50            0.48000000 0.47013333 0.04986667
```

(comparamos, por ejemplo, 0,47 con 0,46 con 0,48; 0,48 con 0,48 con 0,47; y por último: 0,05 con 0,06 con 0,05)

En resumen, las condicionadas del salario a la edad muestran grandes diferencias si el trabajo es intelectual. No ocurre lo mismo para los trabajadores de la otra categoría.

11.FUNCIONES R USADAS EN ANÁLISIS DE TABLAS DE CONTINGENCIA

addmargins(); chisq.test(); factor(); fisher.test(); ftable(); prop.table(); table(); xtabs().