EACS-Estadística descriptiva bivariante: Dos variables cualitativas

Índice

1. Introducción

Si queremos describir el comportamiento conjunto de dos variables de tipo cualitativo y analizar la posible relación estadística entre ellas, las herramientas más utilizadas para realizar dicho estudio son las tablas de contingencia, los gráficos de barras y los contrastes de independencia y homogeneidad. Las tablas de contingencia y los gráficos de barras permiten describir el comportamiento conjunto de las dos variables y también el comportamiento de una de las variables condicionada a los diferentes valores de la otra. Los contrastes de independencia permiten analizar si hay entre ellas alguna relación estadística. Los contrastes de homogeneidad permiten analizar si la distribución de una variable es distinta en diferentes poblaciones.

La tabla de contingencia de dos variables cualitativas se obtiene a través la opción de R Commander

Estadísticos > Tablas de contingencia > Tabla de doble entrada

Desde este menú, podemos calcular la distribución conjunta (expresada en frecuencias absolutas o en porcentajes) y las distribuciones condicionadas (expresada en porcentajes por filas o columnas). También se pueden obtener desde este menú los contrastes de independencia y homogeneidad.

Los gráficos de barras se obtienen a través de la opción

Gráficas > Gráfica de barras

2. Tabla de distribución conjunta

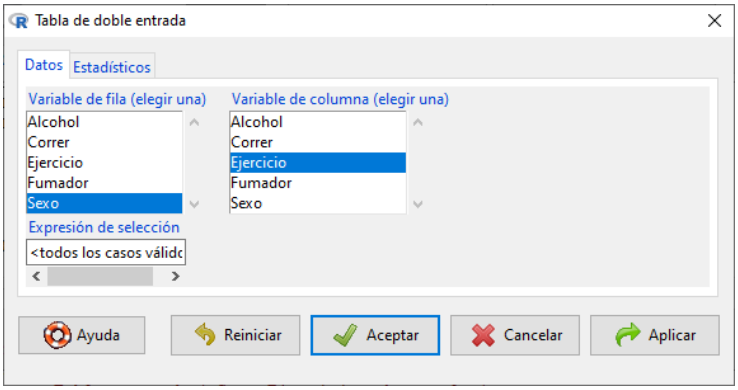
Queremos estudiar la relación entre las variables *Sexo* y *Ejercicio*, que se encuentran en el conjunto de datos *Pulso.RData*. Recordad que la variable *Ejercicio* toma tres modalidades, baja, moderada y alta, en función de la regularidad con la que se hace ejercicio físico.

Para obtener la tabla de doble entrada de frecuencias (la tabla de contingencia) de las variables *Sexo* y *Ejercicio*, los pasos a seguir son los siguientes:

1) Con el conjunto de datos *pulso* activo (*Pulso.RData*), selecciona en el menú principal

   Estadísticos > Tablas de contingencia > Tabla de doble entrada

2) En el cuadro de diálogo que aparece, selecciona una de las variables como *Variable de fila* (por ejemplo, *Sexo*) y la otra como *Variable de columna* (*Ejercicio*). La elección de las variables fila y columna podía haber sido la opuesta.
3) Si no modificamos las opciones por defecto de la pestaña *Estadísticos* esta opción proporciona una tabla de frecuencias y realiza un contraste de independencia, el test Chi-cuadrado (χ^2) de Pearson, que describiremos y comentaremos más adelante.



Al pulsar *Aceptar*, obtendrás la tabla de frecuencias bidimensional que se muestran a continuación:

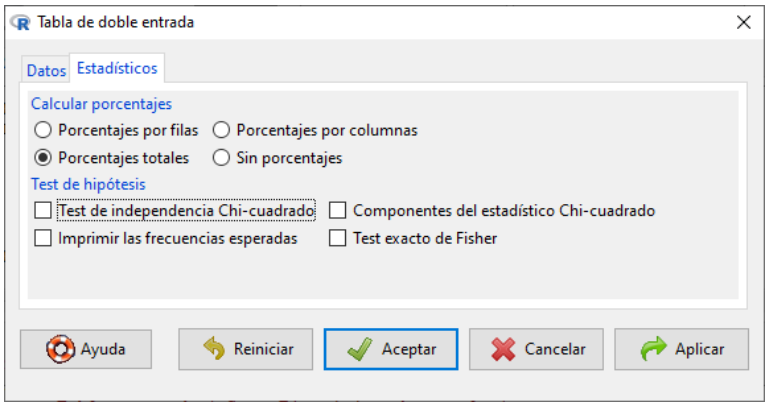
```
Frequency table:
        Ejercicio
Sexo     baja moderada alta
  hombre   17       31   11
  mujer    20       28    3

        Pearson's Chi-squared test

data:  .Table
X-squared = 4.4087, df = 2, p-value = 0.1103
```

Para obtener la tabla de porcentajes sobre el total de la muestra.

1) Como antes, selecciona en el menú principal

Estadísticos > Tablas de contingencia > Tabla de doble entrada

2) En el cuadro de diálogo que aparece, selecciona la variable *Sexo* como variable fila y la variable *Ejercicio* como variable columna.
3) En la pestaña *Estadísticos*, elige la opción *Porcentajes totales*.
4) Desmarca la casilla *Test de independencia Chi-cuadrado*. Por defecto, R Commander siempre tiene marcada esta casilla.



Al pulsar *Aceptar*, obtendrás la tabla de frecuencias absolutas y la tabla con los **porcentajes respecto del total**.

```
Frequency table:
        Ejercicio
Sexo     baja moderada alta
  hombre   17       31   11
  mujer    20       28    3

Total percentages:
       baja moderada alta Total
hombre 15.5     28.2 10.0  53.6
mujer  18.2     25.5  2.7  46.4
Total  33.6     53.6 12.7 100.0
```

Ejercicio: Indica **cuántos individuos de la muestra son mujeres y hacen ejercicio con regularidad alta**. Indica también cuántos son **hombres y hacen ejercicio con regularidad baja**. Indica **qué porcentaje de la muestra son mujeres y hacen ejercicio con regularidad alta**. Indica también el porcentaje de la muestra que son **hombres y hacen ejercicio con regularidad baja**.

Las 3 mujeres que hacen ejercicio con regularidad alta son un 2.7% de la muestra. Los 17 hombres que hacen ejercicio con regularidad baja son un 15.5% de la muestra.

Observa que R Commander no proporciona las frecuencias absolutas marginales en la primera tabla, pero sí los **porcentajes marginales** en la segunda tabla. Con esta segunda tabla podemos saber que el 46,4% de la muestra son mujeres y que hay un 33,6% de la muestra (incluyendo hombres y mujeres) que hacen ejercicio con regularidad baja.

NOTA 2.1: Observa que la suma de los porcentajes asociados a las categorías baja, moderada y alta (33.6, 53.6 y 12.7) es 99.9 y no 100. Eso es simplemente una cuestión de redondeo interno de R Commander.

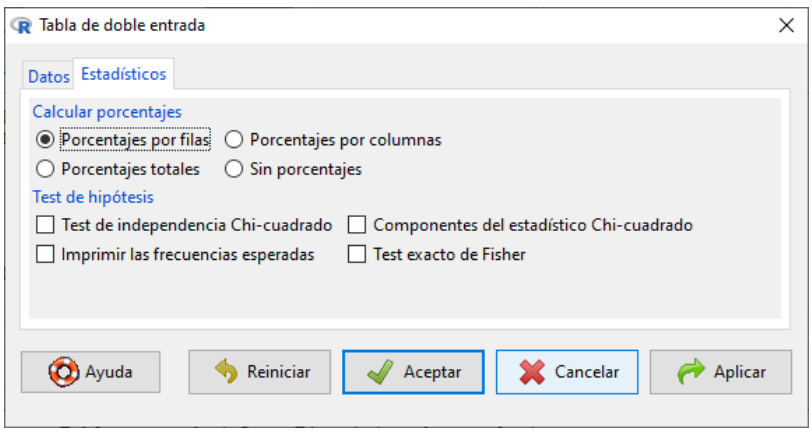
3. Tabla de distribuciones condicionadas, en porcentaje

Para obtener la distribución de una de las variables condicionada a los valores de la otra variable, basta elegir en la pestaña *Estadísticos* del cuadro de diálogo asociado a *Estadísticos* > *Tablas de contingencia* > *Tabla de doble entrada*, la opción *Porcentaje por filas* o la opción *Porcentaje por columnas.* Por ejemplo, supongamos que deseamos estudiar cómo se comporta la variable *Ejercicio* para hombres, por un lado, y para mujeres, por otro (esto es, condicionando a la variable *Sexo*). Para obtener estas distribuciones condicionales:

1) Con el conjunto de datos *pulso* activo, selecciona en el menú principal

   Estadísticos > *Tablas de contingencia* > *Tabla de doble entrada*

2) En el cuadro de diálogo que aparece, selecciona la variable *Sexo* como *Variable de fila* y la variable *Ejercicio* como *Variable de columna.*
3) En la pestaña *Estadísticos*, elige la opción *Porcentajes por filas* (ya que queremos condicionar a la variable *Sexo*, que hemos puesto en filas).
4) Desmarca la casilla *Test de independencia Chi-cuadrado*, ya que no vamos a utilizar ese test.



Al pulsar *Aceptar*, obtendrás la misma tabla de frecuencias, pero ahora los porcentajes proporcionan las distribuciones condicionadas (en porcentaje por filas) de la variable *Ejercicio* tanto para hombres como para mujeres. Observa que la suma de porcentajes por filas, salvo pequeños desajustes por redondeo, es 100.

```
Frequency table:
        Ejercicio
Sexo     baja moderada alta
  hombre   17       31   11
  mujer    20       28    3

Row percentages:
        Ejercicio
Sexo     baja moderada alta Total Count
  hombre 28.8     52.5 18.6  99.9    59
  mujer  39.2     54.9  5.9 100.0    51
```

Vemos que, **en el grupo de los hombres**, un 28,8% hace ejercicio con regularidad baja, un 52,5% lo hace de forma moderada, y un 18,6% con regularidad alta. **En las mujeres**, un 39,2% hace ejercicio con regularidad baja, un 54,9 lo hace con regularidad moderada y sólo un 5,9% con regularidad alta.

Si elegimos *Porcentajes por columnas,* en lugar de *Porcentajes por filas,* en el cuadro de diálogo anterior, el condicionamiento es sobre la variable *Ejercicio*, que es la que se ha puesto en columnas. Se obtiene la siguiente salida.

```
Frequency table:
        Ejercicio
Sexo     baja moderada alta
  hombre   17       31   11
  mujer    20       28    3

Column percentages:
        Ejercicio
Sexo      baja moderada  alta
  hombre  45.9     52.5  78.6
  mujer   54.1     47.5  21.4
  Total  100.0    100.0 100.0
  Count   37.0     59.0  14.0
```

En este caso, lo que estás estudiando es el comportamiento de la variable *Sexo* para cada una de las modalidades de la variable *Ejercicio*. Así, la tabla nos muestra que, **en el grupo de estudiantes que hacen ejercicio con regularidad alta**, un 78,6% son hombres y sólo un 21,4% son mujeres.

Ejercicio: En este mismo conjunto de datos, puedes analizar la relación entre las variables *Fumar* y *Ejercicio* ¿Qué porcentaje de los fumadores de esta muestra hacen ejercicio con regularidad baja? ¿Qué porcentaje de los estudiantes son fumadores y además hacen ejercicio con regularidad baja?

4. Gráficos de barras

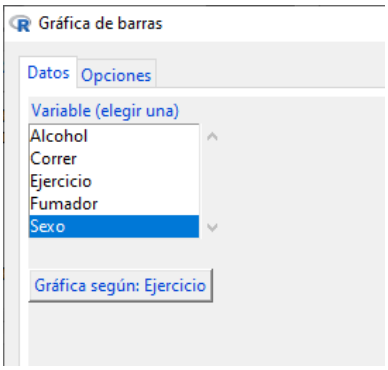
Los gráficos de barras de dos variables cualitativas nos permiten comparar gráficamente la distribución de una de las variables para los diferentes niveles o categorías de la otra, es decir, nos permiten representar gráficamente las distribuciones condicionadas (en porcentajes) de una tabla de contingencia.

Supongamos que queremos comparar gráficamente cómo se comporta la variable *Ejercicio* para hombres y mujeres por separado. Los pasos a seguir son los siguientes.

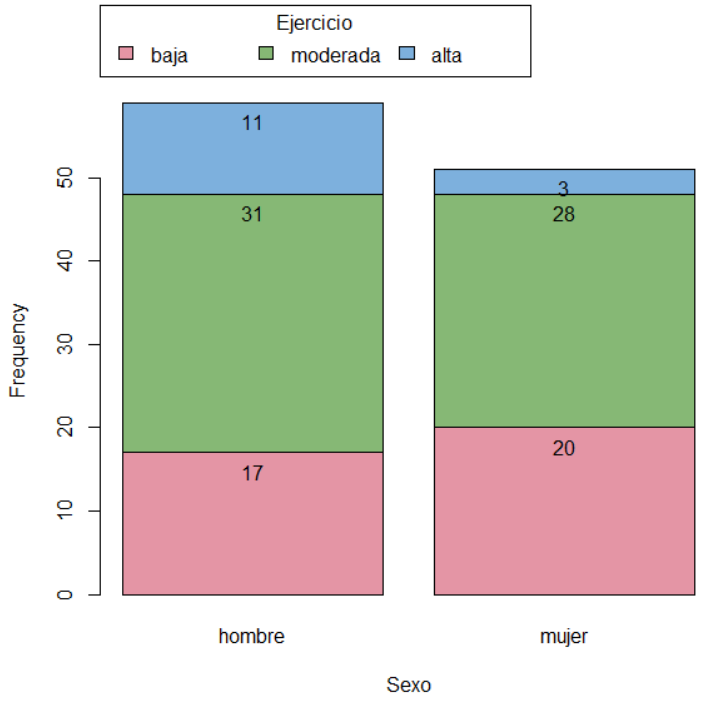
1) Con el conjunto de datos *pulso* activo, selecciona en el menú principal

Gráficas > Gráfica de barras.

2) En la pestaña *Datos*, en el campo *Variable (elegir una)* selecciona la variable *Sexo* (esta es la variable a la que vamos a condicionar) y, usando el botón *Gráfica por grupos*, selecciona la variable *Ejercicio* (esta es la variable de la que queremos obtener la distribución condicionada). El cuadro de diálogo te habrá quedado de la siguiente forma.



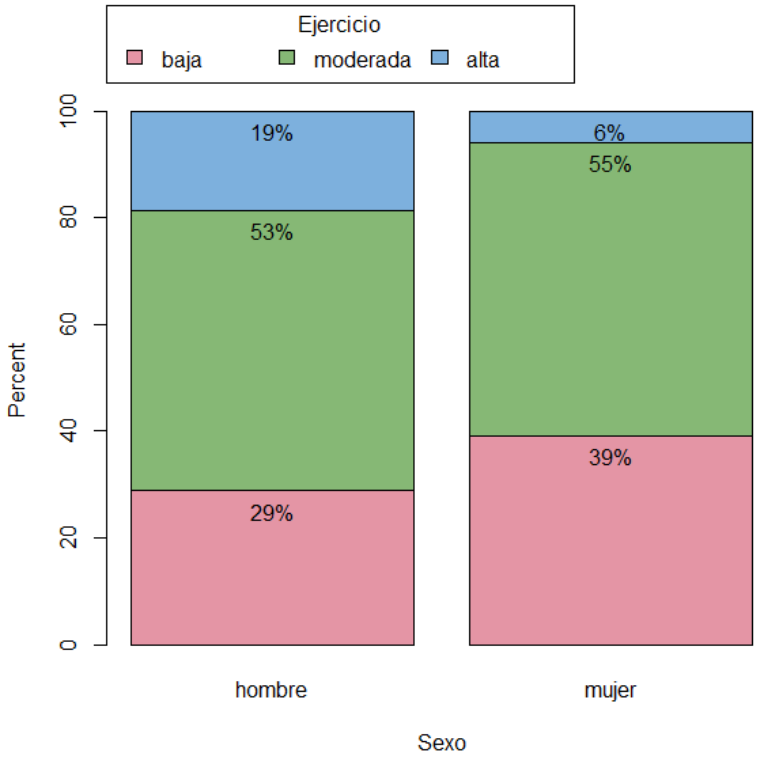
3) En la pestaña *Opciones*, observa que, para la *Escala de los ejes*, la opción por defecto es *Recuento de frecuencias* y para *Estilo…*, la opción por defecto es *Dividido (apilado)*. Si no modificas nada dentro de la pestaña de *Opciones*, el gráfico que obtendrás será el siguiente.



Este gráfico, que representa las **frecuencias**, no es adecuado para comparar el comportamiento de la variable *Ejercicio* entre hombres y mujeres. Hay más hombres que mujeres en la muestra y, como la altura de las barras es proporcional a la frecuencia, la barra vertical de los hombres es más alta que la de las mujeres.

Para comparar las distribuciones condicionadas, conviene representar **porcentajes sobre cada uno de los grupos**. Esto hace que el porcentaje total de cada grupo sea 100 y, por lo tanto, la altura de las dos barras sea la misma. Esto se consigue abriendo la pestaña *Opciones*, y seleccionando *Porcentajes* en la *Escala de los ejes*, y comprobando que está seleccionada la opción *Condicional* en *Porcentajes para Barras Agrupadas*.

Pulsando *Aceptar* obtenemos

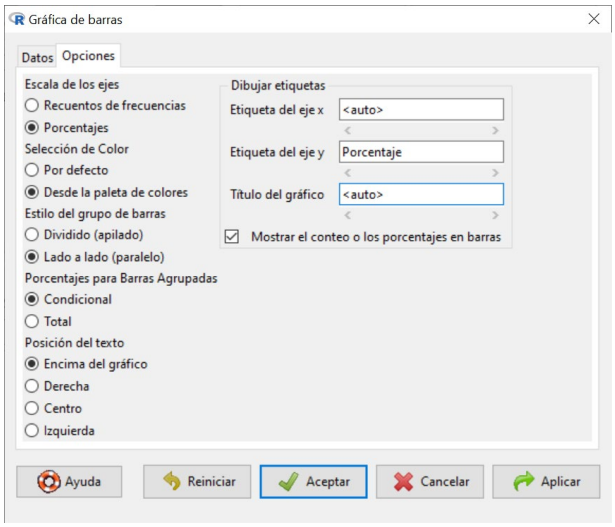


Sin embargo, el *Estilo* que suele ser más adecuado para comparar gráficamente las distribuciones condicionales, y que es **el que recomendamos**, es la representación en *paralelo*. Para obtenerla haremos lo siguiente

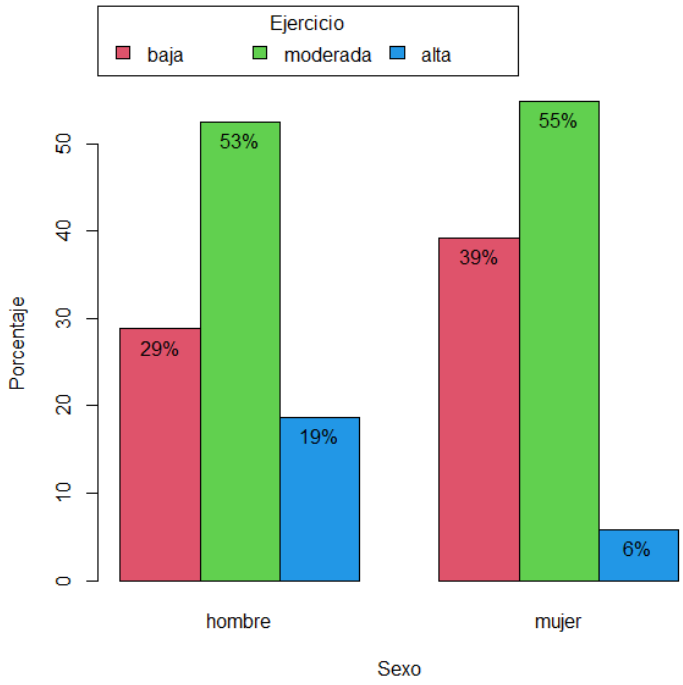
1) Como antes, con el conjunto de datos pulso activo, selecciona en el menú principal

Gráficas > *Gráfica de barras*

2) De la misma forma, en la pestaña *Datos*, en el campo *Variable (elegir una)* selecciona la variable *Sexo* y como la variable para la *Gráfica por grupos*, selecciona *Ejercicio*.
3) En la pestaña *Opciones*, elige *Porcentajes* en el campo *Escala de los ejes* y comprueba que esté seleccionada la opción *Condicional* en *Porcentajes para Barras Agrupadas*. **Pero ahora debes seleccionar la opción** *Lado a lado (paralelo)* dentro del *Estilo del grupo de barras*.
4) Si en *Selección de Color* marcas *Desde la paleta de colores* y escribes *Porcentajes* en la *Etiqueta del eje y*, la ventana queda



Al pulsar *Aceptar*, obtendrás el siguiente gráfico.



Observa que los porcentajes se han calculado sobre el total de individuos en cada grupo y por lo tanto la suma de los porcentajes en cada grupo es 100. Esta gráfica presenta los mismos porcentajes (salvo redondeos) que la tabla en la que habíamos calculado los porcentajes de la variable *Ejercicio* (en columnas) condicionada a la variable *Sexo* (en filas).

```
Frequency table:
        Ejercicio
Sexo     baja moderada alta
  hombre   17       31   11
  mujer    20       28    3

Row percentages:
        Ejercicio
Sexo     baja moderada alta Total Count
  hombre 28.8     52.5 18.6  99.9    59
  mujer  39.2     54.9  5.9 100.0    51
```

De la gráfica y de la tabla de distribuciones condicionales por filas se puede proporcionar el siguiente comentario, en el que hemos redondeado los porcentajes: **la proporción de hombres que hacen ejercicio de forma moderada (un 53%) y la proporción de mujeres que hacen ejercicio de forma moderada (un 55%) son muy similares. Sin embargo, se ven diferencias importantes en las proporciones de hombres y mujeres de las categorías alta y baja. Mientras que un 19% de los hombres hace ejercicio con regularidad alta, sólo un 6% de las mujeres hace ejercicio con regularidad alta. En la categoría regularidad baja tenemos a un 39% de las mujeres y a un 29% de los hombres**.

5. Contrastes de independencia

Uno de los objetivos de un estudio bivariante es analizar si las dos variables que se están estudiando están relacionadas. Si dos variables son independientes, la distribución condicional de una de ellas será la misma en cada uno de los niveles de la otra. En el ejemplo anterior, se apreciaban ciertas diferencias de comportamiento entre hombres y mujeres, en cuanto a la regularidad con la que realizan ejercicio físico, por lo que las variables *Ejercicio* y *Sexo* podrían no ser

independientes en la población de donde se ha extraído la muestra. Las tablas de contingencia o los gráficos de barras no permiten por sí solos realizar afirmaciones del tipo "estas dos variables son independientes" o "estas dos variables están relacionadas" en la población de la que se ha obtenido la muestra, por lo que hay que utilizar una de las técnicas de la inferencia estadística, el contraste de hipótesis.

Vamos a utilizar un contraste (o test) de hipótesis para analizar la independencia de esas dos variables categóricas. El test plantea una hipótesis (llamada **hipótesis nula o H_0**) que afirma que las **variables son independientes** en la población de la que se ha extraído la muestra, frente a la **hipótesis alternativa** (llamada H_1) de que **no lo son**. En el ejemplo que hemos estado estudiando, decir que las variables *Ejercicio* y *Sexo* son independientes significa decir que los hombres y las mujeres se comportan igual en cuanto a la realización del ejercicio o, lo que es lo mismo, el saber si un individuo de esa población es hombre o mujer no aporta información acerca de la regularidad en la realización de ejercicio y viceversa, su nivel de regularidad en la realización de ejercicio no da información sobre su sexo.

NOTA 5.1: Desde un punto de vista teórico este contraste asume que se tiene una única población y que se obtiene una muestra aleatoria de dicha población. Los individuos de la muestra son entonces clasificados de acuerdo a las dos variables. En nuestro ejemplo, la población es un conjunto de estudiantes que se clasifican según sean hombre o mujer y también según el ejercicio que realizan. Por lo tanto, la hipótesis nula es que las dos variables observadas, *Sexo* y *Ejercicio*, son independientes en la población de la que se ha extraído la muestra.

Para poder decidir en un contraste si se rechaza o no una hipótesis nula, se construye una medida de concordancia con la hipótesis nula de los datos recogidos. Esta medida toma valores entre 0 y 1 y se conoce con el nombre de p-valor. Un p-valor igual a 1 indicaría máxima concordancia de los datos con la hipótesis nula, un p-valor de 0 indicaría máxima discordancia. La regla de decisión del 5% (rechazar la hipótesis nula cuando el p-valor es menor que 0.05) implica que sólo rechazamos la hipótesis nula cuando los datos presentan mucha discordancia con dicha hipótesis. En otro caso, la hipótesis nula se retiene.

5.1 Contraste de independencia basado en la χ^2

El contraste Chi-cuadrado (χ^2) de Pearson, es el más utilizado cuando se desea estudiar la independencia de dos variables cualitativas. En este contraste, la medida de concordancia de los datos con la hipótesis de independencia está basada en un estadístico que mide las diferencias entre las frecuencias observadas para la muestra y las frecuencias que se esperarían bajo la hipótesis de independencia.

Para realizar este contraste de independencia con las variables *Sexo* y *Ejercicio*, no tienes más que repetir los pasos descritos en la sección 2. Selecciona en el menú principal

Estadísticos > Tablas de contingencia > Tabla de doble entrada

y elige una de las variables como variable fila (por ejemplo, *Sexo*) y la otra como variable columna (por ejemplo, *Ejercicio*). Por defecto, la casilla *Test de independencia Chi-cuadrado* de la pestaña de *Opciones*, que es la que realiza el contraste de independencia basado en la χ^2, está marcada, por lo que al pulsar *Aceptar* obtendrás la siguiente salida.

```
Frequency table:
        Ejercicio
Sexo     baja moderada alta
  hombre   17       31   11
  mujer    20       28    3

        Pearson's Chi-squared test

data:  .Table
X-squared = 4.4087, df = 2, p-value = 0.1103
```

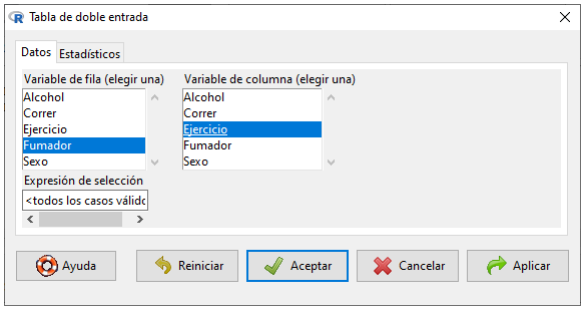
El p-valor obtenido en el contraste de independencia entre el *Sexo* y el E*jercicio* es de 0,1103. Es un valor bajo, que indica bastante discordancia de los datos con la hipótesis de independencia. Sin embargo, aplicando la regla del 5%, esta discordancia no es suficiente como para rechazar la hipótesis de independencia, puesto que p = 0,1103 > 0,0500. Por tanto, con esta regla no podemos rechazar la independencia entre las variables *Sexo* y *Ejercicio*.

NOTA 5.2: Si elegimos *Ejercicio* como fila y *Sexo* como columna, el test Chi-cuadrado da el mismo p-valor.

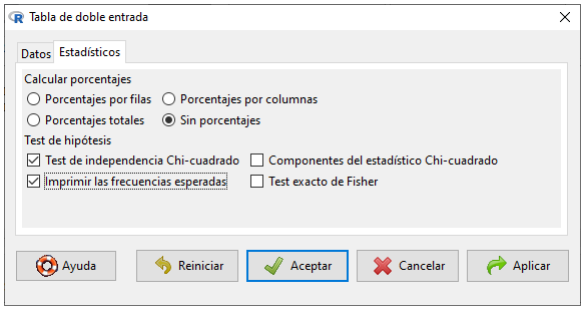
Con todo lo visto hasta ahora, una forma de resumir el estudio conjunto entre las variables *Sexo* y *Ejercicio* podría ser la siguiente**. En el estudio conjunto de las variables *Ejercicio* y *Sexo* hemos observado, a través de las distribuciones condicionadas y del gráfico de barras, que el porcentaje de hombres y mujeres que realizan ejercicio de forma moderada es muy similar (un 52,5% y 54.9%, respectivamente). Sin embargo, parece haber más diferencias en las modalidades alta y baja, donde parece que los hombres de esa población tienden a hacer más ejercicio que las mujeres (el 18,6% de los hombres hacen ejercicio con mucha regularidad, frente a un 5,9% en las mujeres, y el 28,8% de los hombres hacen ejercicio con poca regularidad, frente a un 34,2% en las mujeres). Estas diferencias observadas en la muestra no son estadísticamente significativas al 5%, puesto que el p-valor obtenido en el contraste de independencia entre las dos variables es 0,1103, mayor que 0,05. Por tanto, debemos retener la hipótesis de que la distribución de la variable que mide la regularidad en la realización de ejercicio físico es la misma en hombres y en mujeres o, equivalentemente, que el sexo y la regularidad con la que hacen ejercicio son independientes.**

NOTA 5.3: La aplicación de este contraste basado en la χ^2 requiere que el tamaño muestral sea relativamente grande. En concreto, el contraste que realiza R Commander comprueba que ninguna de las casillas tenga un valor esperado menor que 1, y que las casillas con valores esperados menores que 5 sean, como mucho, un 20% del total. Cuando estas condiciones no se cumplen R Commander avisa con una advertencia en el panel de Mensajes. En este caso no debemos utilizar este contraste y tendremos que plantearnos alguna de las siguientes alternativas: o tomar más datos, si es posible, o agrupar categorías de alguna de las dos variables, si tiene sentido esa agrupación, o bien utilizar algún contraste válido para tamaños muestrales pequeños, como el test exacto de Fisher.

Veamos una situación en la que no se debería utilizar el test Chi-cuadrado. Supongamos que deseamos estudiar la independencia entre las variables *Fumador* y *Ejercicio* con la información contenida en el conjunto de datos *pulso*. Siguiendo los pasos descritos en la sección anterior, seleccionaremos en *Datos* esas dos variables



En Estadísticos marcamos *Test de independencia Chi-cuadrado* y también la opción de *Imprimir las frecuencias esperadas*.



Al pulsar Aceptar, veremos el siguiente aviso en la ventana de Mensajes

```
[5] AVISO: Warning in chisq.test(.Table, correct = FALSE) :
Chi-squared approximation may be incorrect
[6] AVISO:
2 las frecuencias esperadas son inferiores a 5
```

En este mensaje se nos avisa de que, al realizar el contraste de independencia basado en la χ^2, alguna de las condiciones necesarias para su aplicación no se cumple (en este caso, se indica que hay 2 celdas con frecuencias esperadas inferiores a 5). Por tanto, las conclusiones que obtengamos a través de contraste χ^2 podrían no ser válidas.

Si analizamos la salida vemos primero el resultado del Test Chi-cuadrado

```
Frequency table:
       Ejercicio
Fumador baja moderada alta
     no   32       53   14
     sí    5        6    0

      Pearson's Chi-squared test

data:  .Table
X-squared = 2.0649, df = 2, p-value = 0.3561
```

Como el p-valor asociado al contraste es 0,3561, mayor que 0,0500, la conclusión que sacaríamos es que no podemos rechazar la hipótesis de que las variables *Fumador* y *Ejercicio* son independientes. Pero el aviso nos indica que no podemos usar este contraste.

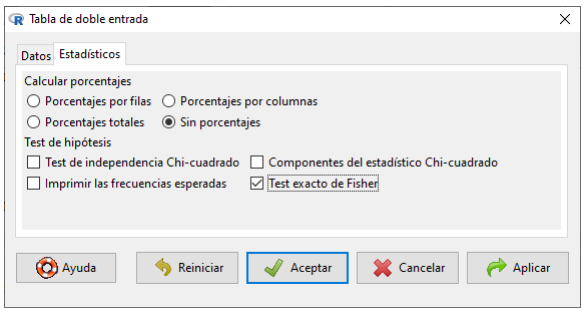
También le hemos pedido que nos proporcione las **frecuencias esperadas**. Puede verse (las hemos marcado en rojo) que hay dos casillas en las que el valor esperado es menor que 5.

```
Expected counts:
       Ejercicio
Fumador baja moderada alta
     no 33.3     53.1 12.6
     sí  3.7      5.9  1.4
```

5.2 Test exacto de Fisher

El test exacto de Fisher permite contrastar la independencia entre dos variables categóricas y se puede aplicar para cualquier tamaño muestral. Por lo tanto, es una buena alternativa cuando no se puede utilizar el test Chi-cuadrado. Las hipótesis nula y alternativa son las mismas que en el contraste χ^2, es decir, la hipótesis nula es la de independencia y la hipótesis alternativa es la de no independencia. El test de Fisher también proporciona un p-valor que mide la concordancia con la hipótesis nula y con el que se puede decidir si se rechaza o no la independencia aplicando la regla del 5%.

Vamos a aplicarlo para analizar si hay independencia entre las variables *Fumador* y *Ejercicio*. En la pestaña *Datos*, seleccionaremos estas variables. En la pestaña *Estadísticos*, marcaremos *Test Exacto de Fisher*.



La salida que obtendríamos se muestra a continuación.

```
Frequency table:
       Ejercicio
Fumador baja moderada alta
     no   32       53   14
     sí    5        6    0

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.4484
alternative hypothesis: two.sided
```

El p-valor que se obtiene en el test de Fisher es 0,4484. Como ese p-valor es mayor que 0,0500 tenemos que retener la hipótesis de independencia entre las variables *Fumador* y *Ejercicio*. En este ejemplo, la conclusión a la que llegamos con los dos contrastes, el basado en la χ^2 y el de Fisher, es la misma (no rechazar la independencia), pero no siempre es así.

6. Contraste de homogeneidad de proporciones

En el contraste de homogeneidad de proporciones se asume que se han obtenido muestras independientes de la misma variable Y en k poblaciones distintas. La hipótesis nula que se desea contrastar es que la variable Y se comporta de la misma forma en cada una de esas poblaciones. En concreto, que la proporción con la que se puede observar cada una de sus categorías es la misma en las k poblaciones muestreadas (las proporciones son homogéneas en todas las poblaciones). La hipótesis alternativa es que la variable Y se comporta de forma distinta en alguna de esas poblaciones.

Este esquema de muestreo se puede encajar en una tabla de contingencia si se considera una "variable ficticia" X tal que sus k niveles indican las poblaciones de las que se ha extraído cada muestra. Aunque este método de muestreo es muy distinto del utilizado para el contraste de independencia, tanto el estadístico del contraste como su distribución son similares, por lo que también se puede usar el contraste χ^2 para tablas de contingencia para contrastar la homogeneidad de proporciones de la variable Y en todas las poblaciones indicadas por la "variable" X.

Veamos un ejemplo de aplicación. Queremos comparar la tasa de fracaso quirúrgico en 3 hospitales (A, B y C). Para ello, se eligen 100 operaciones realizadas en el hospital A, 100 realizadas en el B y 150 realizadas en el C, procurando que estas muestras sean independientes. Los datos se encuentran en el archivo *Hospitales.RData*, que contiene la variable *Operación*, que es la variable que se desea estudiar y que toma dos modalidades, *Éxito* o *Fracaso*, en función de si la operación se realiza con éxito o no, y la "variable ficticia" *Hospital*, que toma los valores *A*, *B* o *C* en función del hospital del que se ha obtenido la muestra de resultados de las operaciones.

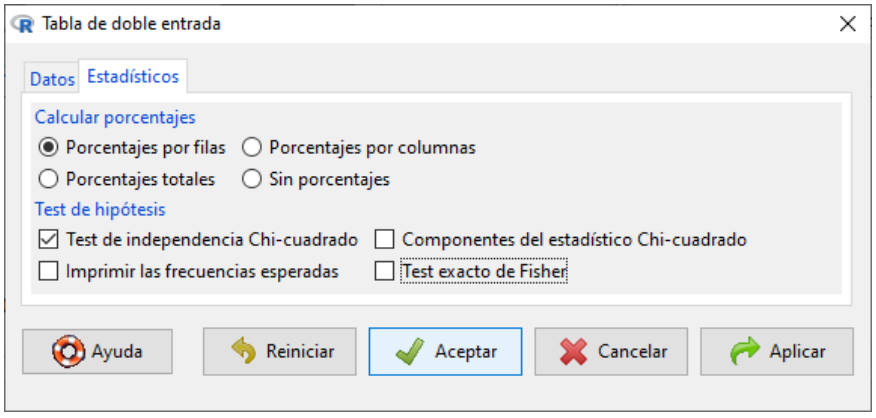
NOTA 6.1: La diferencia con el contraste de independencia está en el método de muestreo. Si se hubiera elegido al azar una única muestra de 350 individuos y se hubieran asignado aleatoriamente a uno de los tres hospitales, para cada individuo tanto el resultado de la operación como el hospital en el que se ha realizado serían verdaderas variables estadísticas y podríamos analizar si esas dos características son independientes en la única población muestreada. Pero en este caso se han obtenido tres muestras, con tamaño prefijado, una en cada uno de los hospitales. Por lo tanto, la "variable" *Hospital* no se ha muestreado. Sólo nos indica en qué hospital se ha obtenido cada una de las muestras. Lo único que podemos analizar es si la tasa (proporción) de éxitos es la misma en los tres hospitales.

Los pasos a seguir para realizar el contraste χ^2 con estos datos serían los siguientes:

1) Con el conjunto de datos *Hospitales* activo, selecciona

   Estadísticos > Tablas de contingencia > Tabla de doble entrada.

2) En el cuadro de diálogo, elige *Hospital* como variable fila, y *Operación* como variable columna.
3) En la pestaña *Estadísticos*, elige *Porcentajes por filas*, para que la tabla proporcione el porcentaje de éxitos y de fracasos en cada hospital, y deja seleccionada también la casilla *Test de independencia Chi-cuadrado*.



Al pulsar *Aceptar*, obtendrás la siguiente salida.

```
Frequency table:
        Operación
Hospital Éxito Fracaso
       A    92       8
       B    85      15
       C   119      31

Row percentages:
        Operación
Hospital Éxito Fracaso Total Count
       A  92.0     8.0   100   100
       B  85.0    15.0   100   100
       C  79.3    20.7   100   150

        Pearson's Chi-squared test

data:  .Table
X-squared = 7.3975, df = 2, p-value = 0.02475
```

Las tasas de éxito en los tres hospitales A, B y C son del 92.0%, del 85.0% y del 79.3%, respectivamente.

Como el p-valor asociado al contraste de homogeneidad de proporciones es 0,02475, menor que 0,05000, podemos rechazar la hipótesis de homogeneidad de proporciones. Por tanto, hay evidencias de que la tasa de éxito no es la misma en los tres hospitales.

7. Cómo introducir y analizar una tabla de doble entrada con R Commander

En los ejemplos anteriores hemos visto el análisis de dos variables cualitativas cuando se dispone de un conjunto de datos con todos los individuos de la muestra, es decir, con una fila por cada individuo y una columna por cada variable. Pero en algunas ocasiones solo se nos facilita la tabla de doble entrada, como en el siguiente ejemplo.

Un grupo de investigadores quiere estudiar la relación entre fumar y padecer hipertensión arterial en la población formada por los habitantes mayores de 40 años que residen en una cierta ciudad. Para ello seleccionan una muestra aleatoria de 300 individuos de ese colectivo a los que clasifican teniendo en cuenta si fuman o no y si padecen hipertensión o no. Se han obtenido los resultados que se muestran a continuación.

	Hipertensión	
Fumador	Sí	No
Sí	42	60
No	23	175

Nos piden analizar si ser fumador conlleva más riesgo de ser hipertenso en la población muestreada.

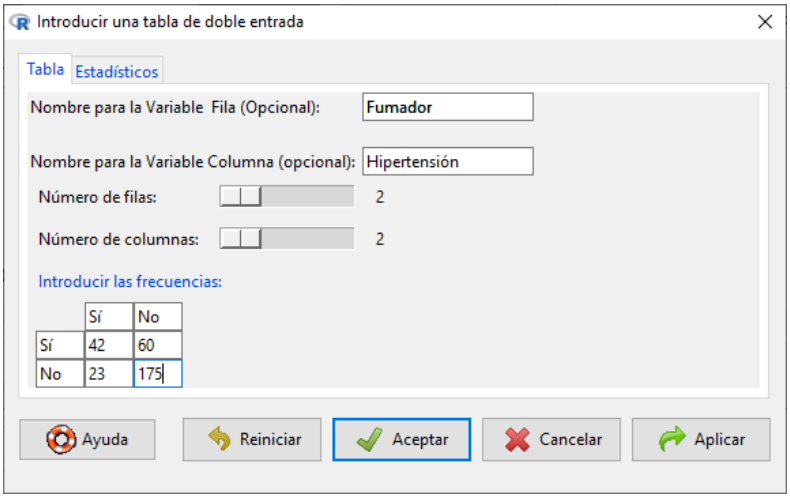
Para contestar a la pregunta, debemos calcular las distribuciones condicionadas para comparar la tasa de hipertensión entre fumadores y no fumadores. También sería útil hacer un contraste de independencia. Nótese que se ha elegido una única muestra con 300 individuos, seleccionados al azar de la población que se quiere estudiar, y se han observado ambas variables para cada uno de esos individuos.

Para introducir los datos en R Commander, obtener las distribuciones marginales (por filas o por columnas) y realizar y obtener la salida del contraste de independencia Chi-cuadrado, los pasos que seguirías serían los siguientes.

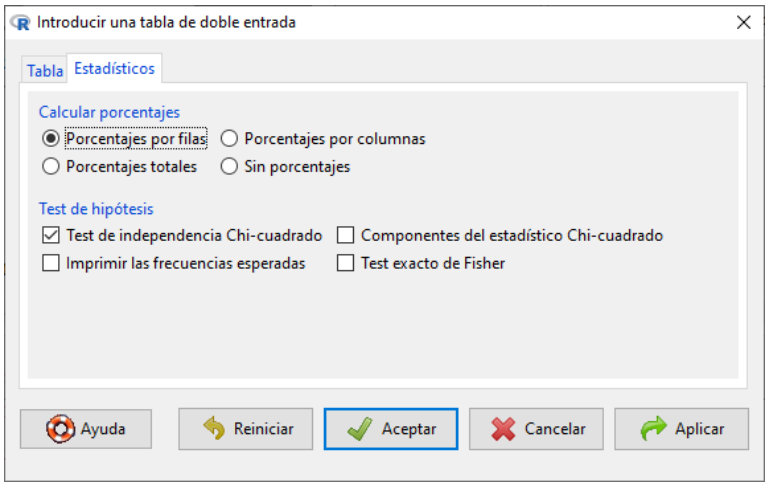
1) Selecciona

   Estadísticos > Tablas de contingencia > Introducir y analizar una tabla de doble entrada.

2) Rellena el cuadro de diálogo como aparece a continuación, donde se está eligiendo *Fumador* como variable fila, e *Hipertensión* como variable columna. Las modalidades de las variables, así como las frecuencias observadas, se introducen en las casillas que se encuentran debajo del texto *Introducir las frecuencias*.



3) En la pestaña *Estadísticos*, elige *Porcentajes por filas*, y marca la casilla *Test de independencia Chi-cuadrado*.



Al pulsar *Aceptar*, obtendrás los siguientes resultados.

```
> .Table <- matrix(c(42,60,23,175), 2, 2, byrow=TRUE)

> dimnames(.Table) <- list("Fumador"=c("Sí", "No"), "Hipertensión"=c("Sí", "No"))

> .Table  # Counts
       Hipertensión
Fumador Sí  No
     Sí 42  60
     No 23 175

> rowPercents(.Table) # Row Percentages
       Hipertensión
Fumador   Sí   No Total Count
     Sí 41.2 58.8   100   102
     No 11.6 88.4   100   198


> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 34.66, df = 1, p-value = 0.000000003927
```

Observamos, en la tabla de porcentajes condicionados, que es mucho más frecuente (41.2%) tener hipertensión si fumas que si no fumas (11.6%). El p-valor en el contraste de independencia es casi cero, lo que nos indica que podemos rechazar la hipótesis nula de independencia entre las variables *Fumador* e *Hipertensión*. Por tanto, las diferencias que vemos en el comportamiento de la hipertensión entre fumadores y no fumadores no son debidas al azar, sino a la relación entre las variables.

En tablas 2x2, se pueden plantear alternativas unilaterales. Podemos concluir que en la población muestreada hay una mayor tasa de hipertensión entre fumadores que entre no fumadores.

En el cuadro mensajes no vemos ningún aviso acerca de que no se cumplen las condiciones necesarias (las que hemos comentado en la nota 5.3) para la realización del contraste basado en la χ^2. Por lo tanto, los resultados obtenidos son válidos.

Podríamos también realizar el test exacto de Fisher, que nos llevaría a la misma conclusión, pues el p-valor que se obtiene es 0,000000015.

NOTA 7.1: El menú utilizado en este apartado, *Estadísticos > Tablas de contingencia > Introducir y analizar una tabla de doble entrada,* no permite la representación gráfica conjunta de las variables.

Una aproximación alternativa sería construir un fichero a partir de la tabla de doble entrada. A partir de ese fichero, se pueden obtener las distribuciones condicionales, obtener los resultados de los contrastes y también obtener los gráficos de barras de distribuciones condicionales. Una forma relativamente sencilla de hacerlo es construir un fichero Excel y luego importarlo desde R Commander.

Vamos a ilustrar el procedimiento con los datos de la tabla de doble entrada de las variables *Fumador* e *Hipertensión.* En una hoja de cálculo Excel pondremos en una columna los valores de *Fumador* y en otra los de *Hipertensión*. Para cada par de posibles valores de esas variables añadiremos ese par el número de veces que indica la tabla de frecuencias. Esto es, repetiremos el par (Sí, Sí) 42 veces, el par (Sí, No) 60 veces, el par (No, Sí) 23 veces y el par (No, No) 175. Una vez introducidos los datos, los guardamos en un fichero con nombre *FumadorHipertension.xlsx*. Una vez creado ese fichero Excel con los 300 datos, ya podemos importarlo desde R Commander, y guardar los datos en un fichero con nombre *FumadorHipertension.RData*, que tendrá un formato propio de R. Utilizando este último fichero podemos aplicar todas las técnicas que hemos ido comentando en esta práctica.

8. Ejercicios

1. En el archivo Pulso.RData, los estudiantes de la muestra han indicado si fuman o no (variable Fumador, con niveles sí y no), si beben o no (variable Alcohol, con niveles sí y no), y su regularidad en la realización de ejercicio físico (variable Ejercicio, con niveles baja, moderada y alta). Contesta a las siguientes preguntas.

   a) ¿Qué porcentaje de los fumadores de esta muestra hacen ejercicio con regularidad baja? ¿Qué porcentaje de los estudiantes son fumadores y además hacen ejercicio con regularidad baja?
   b) Obtén el gráfico que compara la distribución de la variable Ejercicio entre los que beben y los que no beben.
   c) Obtén el p-valor del contraste de independencia χ^2 entre las variables Ejercicio y Alcohol. Indica si se puede rechazar, al 5%, la hipótesis de que estas variables son independientes en la población de la que se ha extraído la muestra.

2. Con objeto de estudiar la relación entre litiasis renal (cálculos en el riñón) e hiperuricemia (alta concentración de ácido úrico en sangre), se realizó un estudio de seguimiento de una muestra aleatoria de los habitantes adultos de una determinada población. Los datos obtenidos fueron:

	Ácido úrico	
Litiasis	Alto	Normal
Sí	250	450
No	1900	3800

   Obtén la tabla de porcentajes respecto del total. ¿Existe algún tipo de relación entre estas dos variables? Justifica tu respuesta.

3. Se cree que la proporción de personas con hiperlipemia (exceso de grasas en sangre) es distinta en dos ciudades. Para comprobarlo, se seleccionan dos muestras aleatorias, una de cada una de las ciudades, y se estudia el número de personas hiperlipémicas. Los datos se encuentran en el archivo Ciudad.RData. Obtén la tabla de porcentajes de hiperlipemia para cada una de las ciudades, el diagrama de barras agrupadas (paralelo) de la variable hiperlipemia y el contraste de homogeneidad de proporciones. ¿Se pueden considerar similares las proporciones de hiperlipémicos en ambas ciudades?

4. Para estudiar si existe asociación entre los valores elevados de pepsinógeno (enzima estomacal) y la presencia de úlceras gastroduodenales, se selecciona una muestra aleatoria entre los pacientes atendidos en un centro de salud. Los datos se encuentran en el fichero Ulcera.RData. Utilizando tablas de porcentajes, diagrama de barras y contrastes de hipótesis, ¿puedes indicar si existe algún tipo de asociación entre estas dos variables, nivel de pepsinógeno y presencia de úlcera? En caso afirmativo, ¿cómo es esa asociación?

5. Se quiere estudiar si los pacientes con cáncer de boca responden mejor al tratamiento de quimioterapia que los pacientes con cáncer de tráquea. Para ello se han seleccionado 60 pacientes con cáncer de boca y 58 con cáncer de tráquea y se ha valorado si el paciente mejoraba. Los datos se encuentran en el fichero Cancer.RData. Obtén la tabla de porcentajes por filas, el diagrama de barras agrupadas (paralelo) de la variable Mejora por cada tipo de cáncer (variable Cáncer) y realiza el contraste de hipótesis de homogeneidad de proporciones. ¿Qué proporción de pacientes con cáncer de boca mejoran después del tratamiento? ¿Y qué proporción de pacientes con cáncer de tráquea mejoran con el tratamiento? ¿Podríamos decir que los pacientes con cáncer de tráquea responden mejor al tratamiento?

6. Se quieren comparar tres terapias diferentes para aliviar cefaleas vasculares, identificadas por A, B y C. Se asignan aleatoriamente tres muestras de pacientes, que son seguidos hasta seis horas después de aplicada la terapia. Al cabo de ese tiempo, se evalúa a los pacientes y se anota si el dolor persiste o ha desaparecido. Los datos se encuentran en el fichero Terapia.RData. Obtén la tabla de porcentajes por columnas, el diagrama de barras agrupado para la variable Dolor en las diferentes terapias (variable Terapia) y el contraste de homogeneidad de proporciones. ¿Se pueden considerar las tres terapias igual de eficientes a la hora de aliviar el dolor?

7. Un grupo de 800 personas con una determinada enfermedad ha sido tratado con dos medicamentos distintos (A y B). De las 800 personas, 400 fueron elegidas aleatoriamente para ser tratadas con el medicamento A y el resto con el medicamento B. Los datos se encuentran en el fichero Medicamento.RData. Obtén la tabla de porcentajes por filas, el diagrama de barras agrupado para la variable Curación en los diferentes medicamentos y el contraste de homogeneidad de proporciones ¿Se pueden considerar los dos medicamentos igual de eficientes en cuanto a la curación de la enfermedad?