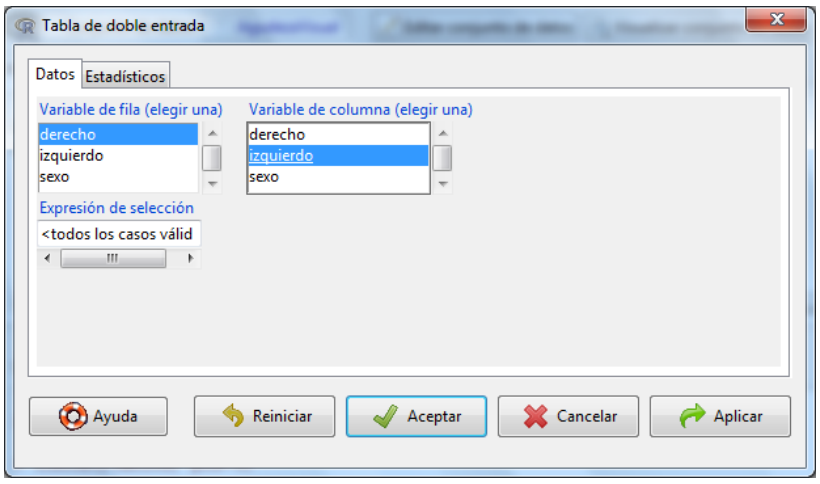
Práctica 4. Análisis de datos bidimensionales. Variables cualitativas.

1- Cálculo de tablas de frecuencias bidimensionales
2- Cálculo de porcentajes totales y distribuciones marginales
3- Cálculo de porcentajes por filas y por columnas (distribuciones condicionadas)
4- Generación de una tabla de frecuencias y cálculo de porcentajes. La paradoja de Simpson
5- Estudio conjunto de una variable cualitativa y una cuantitativa mediante agrupación en clases de ésta última
6- Ejercicios propuestos (para entrega voluntaria)
7- Ejercicios propuestos
8- Apéndice: información sobre ficheros

1. Cálculo de tablas de frecuencias.

En primer lugar, vamos a construir tablas de doble entrada con el fichero AgudezaVisual, que representa la agudeza visual de 10719 empleados en fábricas de munición inglesas durante la segunda guerra mundial. El fichero contiene, para cada individuo, las variables: agudeza visual en el ojo derecho (derecho) y agudeza visual en el ojo izquierdo (izquierdo), así como el sexo del individuo. Las agudezas visuales están clasificadas en cuatro niveles, siendo 1 el nivel más alto y 4 el más bajo.

En primer lugar, como el número de datos es muy grande, es conveniente presentar la información en tablas de frecuencias. Para simplificar, vamos a estudiar hombres y mujeres juntos. Construiremos una tabla de frecuencias donde las filas representarán la agudeza visual en el ojo dcho (y las columnas la del izdo). La tabla se obtiene a través de la opción: **Estadísticos>Tablas de contingencia>Tabla de doble entrada.**



Se obtiene el siguiente resultado:

Frequency table:

Izquierdo derecho	1	2	3	4
1	2339	378	209	101
2	350	2006	577	105
3	189	513	2355	292
4	79	116	285	823

Pearson's Chi-squared test

data: .Table
X-squared = 11474.87, df = 9, p-value < 2.2e-16

En la tabla se observa, de un primer vistazo, que ojo derecho y ojo izquierdo, no van por libre (hay dependencia entre las variables). De hecho, en cada fila el grupo más numeroso siempre es el de la misma agudeza visual en ambos ojos (las frecuencias de la diagonal).

Nota: De momento, a la información que va debajo de la tabla no le vamos a prestar mucha atención (porque no sabemos lo que significa), pero es interesante decir que el valor X-squared = 11474.87 es una medida de dependencia entre las dos variables (valor 0 implicaría que hay independencia y cuanto más grande es el valor, más dependencia hay entre las variables). De hecho, p-value < 2.2e-16 cuantifica esa dependencia (que es mucha, ya que p-values inferiores a 0.05 suelen considerarse como fuertes indicadores de dependencia entre las variables), pero de momento no vamos a entrar en ello.

2. Cálculo de porcentajes totales y distribuciones marginales.

Ahora vamos a construir tablas de porcentajes. En primer lugar, la tabla de porcentajes respecto del total (o distribución conjunta). Se vuelve al menú anterior y se activa (entrando en el botón **"Estadísticos"**) *porcentajes totales*, obteniendo lo siguiente:

Total percentages:

Izquierdo Derecho	1	2	3	4	Total
1	21.8	3.5	2.0	0.9	28.2
2	3.3	18.7	5.4	1.0	28.3
3	1.8	4.8	22.0	2.7	31.2
4	0.7	1.1	2.7	7.7	12.2
Total	27.6	28.1	32.0	12.3	100.0

Aquí se nos muestran los porcentajes asociados a cada pareja de agudezas visuales. Observamos, por ejemplo, que el grupo más numeroso es el que presenta agudeza visual 3 en los dos ojos (el 22% de la población) y el menos numeroso el que presenta AV 4 en el dcho y 1 en el izdo (0.7% de la población). Lógicamente, lo menos común es que alguien tenga una agudeza visual excelente en un ojo y una agudeza visual muy pobre en el otro.

Nótese que los totales de cada fila nos muestran los porcentajes relativos al ojo derecho (esto es, los porcentajes marginales). De ahí se observa que las agudezas visual más frecuente en el ojo derecho es la 3 (31.2%), y la menos frecuente la 4 (12.2%). Algo muy similar sucede en el ojo izquierdo (los porcentajes marginales por columnas aparecen en los totales por columnas).

Ejercicio: A partir de la tabla de frecuencias obtenida en el apartado 1, comprueba el cálculo del porcentaje asociado a una agudeza visual 3 en los dos ojos. Comprueba, asimismo el porcentaje marginal correspondiente a una agudeza visual 3 en el ojo derecho.

3. Cálculos de porcentajes por filas y por columnas (distribuciones condicionadas).

Ahora vamos a ver los porcentajes por filas. Estos se obtienen eligiendo la opción *porcentajes por filas* en el apartado "Estadísticos" del cuadro de diálogo anterior. El resultado que aparece es el siguiente:

Frequency table:

Izquierdo derecho	1	2	3	4
1	2339	378	209	101
2	350	2006	577	105
3	189	513	2355	292
4	79	116	285	823

Row Percentages:

Izquierdo Derecho	1	2	3	4	Total	Count
1	77.3	12.5	6.9	3.3	100.0	3027
2	11.5	66.0	19.0	3.5	100.0	3038
3	5.6	15.3	70.3	8.7	99.9	3349
4	6.1	8.9	21.9	63.2	100.1	1303

Por ejemplo, la primera fila representa a las personas con AV1 (alta) en el ojo dcho (OD). A estas personas se les clasifica según su AV en el ojo izdo (OI), resultando que:

- La mayor parte de ellas (el 77.3%) tienen AV 1 en el ojo izdo. Esto es, de entre las personas con AV 1 en el OD, más del 75% tienen esa misma AV en el OI.
- La menor parte de ellas (el 3.3%) tiene una AV de 4 en el OI. Como se indicó antes, resulta raro que las AV en cada ojo sean muy diferentes.

En todas las filas se observa una pauta similar: el grupo más numeroso es siempre el de la diagonal, que corresponde a una AV en el ojo izdo igual a la del dcho. El porcentaje de personas decrece a medida que nos alejamos de la diagonal (o sea, consideramos grupos de AV más alejados del correspondiente al OD).

Nota 1: Como información, nos han proporcionado también la tabla de frecuencias para que nos aclaremos mejor. Como se explicó en teoría, los porcentajes por filas se obtienen dividiendo la frecuencia absoluta de cada casilla entre el total de su fila (y multiplicando por 100). De esta manera, personas con AV 1 en el ojo dcho hay 3027 (la información nos la dan en la columna Count de la segunda tabla, que se obtiene sumando toda la primera fila en la primera tabla). De ellos, 2339 presentan AV 1 en el ojo izdo (primera casilla de la primera tabla). Por tanto, $(2339/3027)\cdot 100=77.3\%$ es el porcentaje de personas con AV1 en el ojo izdo, de entre las personas con AV1 en el ojo dcho.

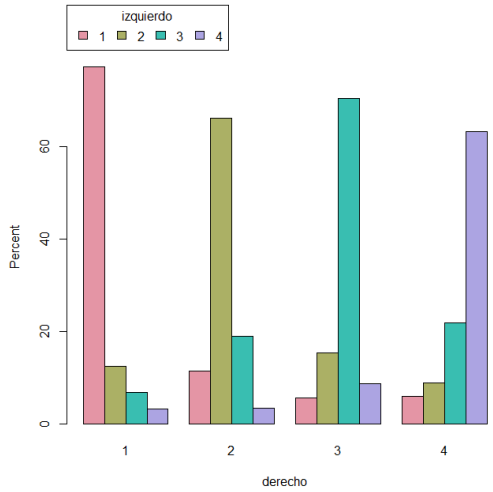
***Ejercicio*:** A partir de la tabla de frecuencias, comprueba el cálculo del porcentaje por filas asociado a agudeza visual 3 en ambos ojos.

Nota 2: Los porcentajes por filas deben sumar 100 (y siempre lo hacen). Nótese que en las dos últimas filas, debido a error de redondeo al primer decimal, sale 99.9 y 100.1.

Nota 3: En muchas situaciones, es de interés obtener una visualización de cómo se comportan los porcentajes de una de las variables para cada uno de los grupos de la otra. El siguiente gráfico nos muestra los porcentajes por fila obtenidos en la tabla anterior. Esto es, nos muestra cómo se distribuye la agudeza visual del ojo izquierdo fijada la agudeza visual del ojo derecho.

Para realizar este gráfico iremos a Gráficas/gráfica de barras. Seleccionaremos como variable "derecho" y en el botón "gráfica según" pondremos "izquierdo" (el lenguaje que utiliza R-commander no es

muy afortunado, pero probablemente se corrija en futuras versiones). En opciones seleccionaríamos porcentajes, en paralelo y condicionales



Nota 4: El estudio de porcentajes por columnas sería muy similar. En este caso observaríamos, en cada columna, la distribución de porcentajes de AV en el OD, una vez fijada la AV en el OI.

***Ejercicio*:** A partir de la tabla de frecuencias, comprueba el cálculo del porcentaje por columnas asociado a agudeza visual 3 en ambos ojos.

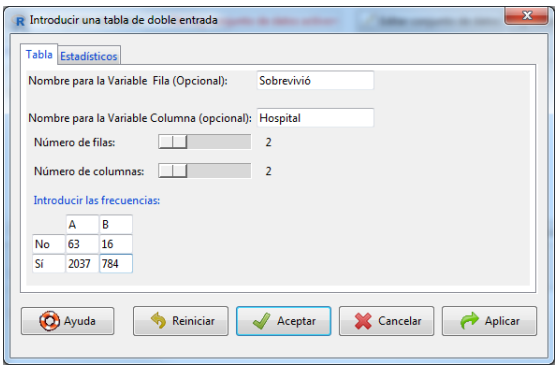
4. Generación de una tabla de frecuencias y cálculo de porcentajes. La paradoja de Simpson.

A veces ya tenemos la tabla de frecuencias, y lo que nos interesa es calcular los porcentajes. Por ejemplo, tenemos un estudio realizado en dos hospitales (hospital A y hospital B) donde se clasificó a los pacientes operados en un periodo de tiempo según si sobrevivieron a la operación pasados 6 meses (si) o no (no).

La tabla es la siguiente:

	Hospital A	Hospital B
No sobrevivieron	63	16
Sí sobrevivieron	2037	784

Para saber los porcentajes de supervivientes en cada hospital vamos a generar la tabla de los porcentajes por hospital (en **Estadísticos>Tablas de Contingencia>Introducir y analizar una tabla de doble entrada,** activando la opción porcentajes por columnas en el botón "estadísticos"**)**



Al aceptar en el cuadro de diálogo anterior, se obtendría la siguiente salida:

> colPercents(.Table) # Column Percentages

Hospital Sobrevivió	A	B
No	3	2
Si	97	98
Total	100	100
Count	2100	800

Parece que el hospital A tuvo un menor número de supervivientes que el B. ¿Deberíamos concluir que la atención sanitaria es peor en el A? Todavía no, hay que tener en cuenta que no todas las operaciones son igual de complejas. De hecho, se tiene la siguiente información adicional acerca del estado general de los pacientes antes de la operación (buena salud o salud delicada).

Buena salud:

	Hospital A	Hospital B
No sobrevivieron	6	8
Sí sobrevivieron	594	592

Salud delicada:

	Hospital A	Hospital B
No sobrevivieron	57	8
Sí sobrevivieron	1443	192

Al calcular los porcentajes por hospital asociados a cada tabla, comprobamos que el hospital A tiene mejores porcentajes en ambos grupos de pacientes. Lo que sucede es que muchos más pacientes con salud delicada (con menos probabilidades de supervivencia) fueron operados en el hospital A. Por tanto, en el hospital A hay mayor mortalidad porque se opera un número mayor de pacientes de riesgo, y no porque la atención sanitaria sea peor que en el B (es justo al revés).

El ejemplo que hemos visto tiene que ver con la llamada paradoja de Simpson. Los resultados que proporciona el estudio de dos variables pueden interpretarse de modo erróneo, si no se tiene en cuenta las variables latentes que pueden intervenir (e influir) en el estudio. Es un ejemplo parecido al que vimos con los establecimientos ópticos y el número de prisioneros.

***Ejercicio*:** En el fichero AgudezaVisual, tenemos la variable sexo en el estudio, resultaría interesante realizar el cálculo de porcentajes de las agudezas visuales en cada ojo, separadas según el sexo. Para ello, ir a **Estadisticos>Tablas de contingencia>Tabla de entrada múltiple** y elegir ojo derecho en filas, ojo izquierdo en columnas y sexo como variable control. Calcular, por ejemplo, los porcentajes por fila. Comparar los resultados

5. Estudio conjunto de una variable cualitativa y una cuantitativa mediante agrupación en clases de esta última.

En ocasiones tenemos una variable cualitativa, y otra cuantitativa, y nos interesa compararlas. Una manera sería clasificar la variable cuantitativa según categorías. Como ejemplo, vamos a usar el fichero Wesdr. En las dos primeras prácticas observamos que el comportamiento de la variable hemoglobina

glicosilada (gly) se comportaba de forma diferente según si el individuo desarrolló o no retinopatía. Para estudiar ambas variables conjuntamente, vamos a codificar la variable gly según cuartiles.

- Calcula las estadísticas de la variable gly, para obtener los cuartiles.
- Crea una nueva variable a partir de gly (qgly, por ejemplo) que indique en qué cuartil de dicha variable se encuentra cada individuo.
- Crea una tabla de frecuencias para las variables ret y qgly. Crea la(s) tabla(s) de porcentajes que consideres adecuadas para el estudio conjunto de dichas variables. Extrae conclusiones al respecto.

6. Ejercicios propuestos (para entrega voluntaria)

1) El hecho de que un acusado de asesinato sea condenado o no a muerte, parece estar influenciado por la raza de la víctima. Tenemos datos de 326 casos en los que el acusado fue declarado culpable de asesinato. Las variables de estudio son raza del acusado, raza de la víctima y si se produjo condena a muerte o no. Los datos se encuentran en el fichero Condenados.rda.

   - Construye una tabla de contingencia donde se relacione la raza del acusado con el hecho de si fue condenado a muerte o no. Calcula los porcentajes de condenados a muerte según raza del acusado.
   - Construye tablas separadas de dichos porcentajes según si la víctima era de raza negra o no. Para ello, proceder como en el al último ejercicio de la seción 4: ir a Estadisticos>Tablas de contingencia>Tabla de entrada múltiple y elegir Raza.Victima como variable control.
   - Constata que se verifica la paradoja de Simpson: en conjunto, un mayor porcentaje de acusados blancos son condenados a pena de muerte, en cambio, considerando de manera independiente a las víctimas blancas y a las víctimas negras, el porcentaje de acusados negros condenados a muerte es mayor que el de blancos.
   - Da una explicación a la paradoja en este caso. Para ello, sería conveniente construir las tablas de frecuencias absolutas del apartado anterior.

2) Usando el fichero excel Opticos, selecciona dos variables cualitativas (o discretas). Calcula las tablas de los apartados 1, 2 y 3. Acompaña dichas tablas con sus representaciones gráficas asociadas. Comenta los resultados. En particular, extrae al menos un dato de cada una de las tablas e interpreta su significado. **Nota**: También puedes seleccionar una variable cuantitativa, si previamente la agrupas en intervalos.

7. Ejercicios propuestos

3) Se cree que la proporción de personas con miopía 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 con miopía. Los datos obtenidos fueron.

	Miopía	
	Sí	No
Ciudad A	62	138
Ciudad B	120	200

Calcula los porcentajes de miopía en cada una de las ciudades. A la vista de los resultados obtenidos, ¿dirías que pueden considerar similares los porcentajes de miopes en ambas ciudades?

4) 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 resultados obtenidos fueron:

	Curación		
Medicamento	En 2 semanas	En 4 semanas	Muertos
A	240	140	20
B	80	220	100

Calcula los porcentajes de curación para cada uno de los dos medicamentos. A la vista de los resultados obtenidos, ¿crees que los medicamentos son igual de efectivos a la hora de mejorar a los enfermos?

8. Apéndice. Información sobre ficheros.

Información sobre el fichero AgudezaVisual.

- **Descripción**: el fichero corresponde a un conjunto de datos estudiados en Kendall & Stuart [1] sobre visión sin corrección de 3.242 hombres y 7.477 mujeres, pertenecientes al grupo de edad 30-39 y empleados en fábricas de la U.K. Royal Ordnance 1943-1946 (fábricas de munición en el periodo de la II guerra mundial).

- **Formato**: este fichero contiene 10.719 filas con las siguientes variables:

    - Derecho: agudeza visual en el ojo derecho, estableciendo 4 niveles, del más alto (1) al más bajo (4).
    - Izquierdo: agudeza visual en el ojo derecho, estableciendo 4 niveles, del más alto (1) al más bajo (4).
    - Sexo: variable que indica el sexo del empleado.

Referencias

[1] Kendall, M., Stuart, A. (1961), The Advanced Theory of Statistics (vol. 2), Griffin.

Información sobre el fichero Condenados.

- **Descripción**: el ejemplo es de Moore [1], y ha sido construido a partir de datos de [2]. El fichero contiene 326 filas, que representan acusados declarados culpables de asesinato.

- **Formato**: este fichero contiene las siguientes variables:

    - Raza.Acusado: blanca o negra.
    - Raza.Victima: blanca o negra.
    - Condenado: condenado a muerte (si) o no (no).

Referencias

[1] Moore, D. (2000), Estadística aplicada básica, 2ª edición, Antoni Bosch.

[2] Radelet, M. (1981), Racial characteristics and imposition of the death penalty, American Sociological Review, 46, 918-927.