Bioestadística

Práctica 3

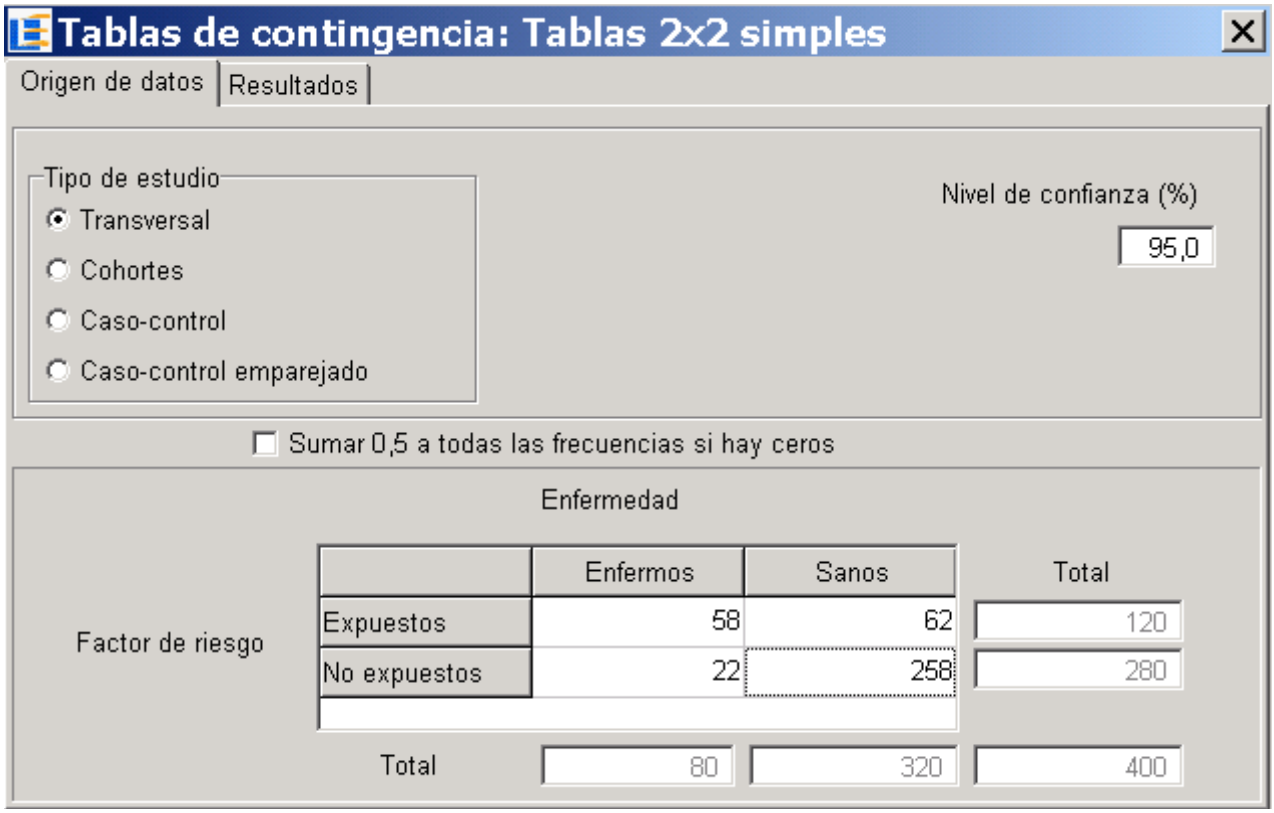
En esta práctica vamos a continuar calculando «a mano» algunas de las medidas epidemiológicas que proporciona *Epidat*. Aunque se puede hacer también con una calculadora, esta vez lo haremos con el lenguaje de programación gratuito R. Al hacer esto seguiremos descubriendo cómo trabaja *Epidat*, aprenderemos algo de teoría estadística y practicaremos el uso de un lenguaje de programación para hacer este tipo de cálculos. En el segundo ejercicio aprenderemos a buscar información para interpretar los resultados de los estudios de casos y control, transversales y de cohortes.

Ejercicio 1

En uno de los ejercicios de la práctica anterior estuvimos calculando con una calculadora algunas cantidades de la salida que proporciona *Epidat* en el caso de una tabla de contingencia de un estudio transversal. En este ejercicio vamos a calcular la mayoría del resto de cantidades. Para **analizar una tabla 2x2 simple** entramos en el menú:

Métodos --> Tablas de contingencia --> Tablas 2x2 --> Simples

Una vez dentro del menú, para analizar datos de un **estudio transversal** introducimos los datos como se indica en la figura siguiente y después pulsamos el botón de cálculo



De la siguiente salida, vamos a calcular ahora las cantidades que están en letra negrita

[1] Tablas de contingencia: Tablas 2x2 simples

Tipo de estudio : Transversal
Nivel de confianza: 95,0%

Tabla

	Enfermos	Sanos	Total
Expuestos	58	62	120
No expuestos	22	258	280
Total	80	320	400

Prevalencia de la enfermedad	Estimación	IC(95,0%)	
En expuestos	0,483333	-	-
En no expuestos	0,078571	-	-
Razón de prevalencias (Katz)	**6,151515**	**3,955011**	**9,567897**

Prevalencia de exposición	Estimación	IC(95,0%)	
En enfermos	0,725000	-	-
En no enfermos	0,193750	-	-
Razón de prevalencias (Katz)	**3,741935**	**2,882081**	**4,858324**

OR	IC(95,0%)		
10,970674	**6,243768**	**19,276133**	(Woolf)
	6,264300	19,204815	(Cornfield)

Prueba Ji-cuadrado de asociación	Estadístico	Valor p
Sin corrección	**86,0119**	**0,0000**
Corrección de Yates	**83,5007**	**0,0000**

Prueba exacta de Fisher	Valor p
Unilateral	**0,0000**
Bilateral	**0,0000**

Utilizaremos las siguientes letras para definir algunas cantidades teóricas:

	Enfermos	Sanos	
Expuestos	a	b	$a+b$
No expuestos	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d$

En el lenguaje de programación la matriz de nuestros datos se introduce como:

```
tabla <- matrix(c(58, 22, 62, 258), nrow=2)
```

Conviene mencionar el camino que se suele seguir para **deducir teóricamente los intervalos de confianza** de las estimaciones que nos interesan, que en este caso son razones. Los siguientes pasos se pueden encontrar en cualquier libro en que se traten las tablas de contingencia con un contenido teórico mínimo.

1) Supongamos que R es una razón. Como las razones toman valores —para tablas con valores positivos— en $(0,+\infty)$, convendría aplicarle una transformación que le hiciese distribuirse simétricamente y, si es posible, siguiendo una normal, aunque sea asintóticamente. La transformación que se suele tomar es el logaritmo neperiano.

2) Ahora, para esta distribución asintóticamente normal, se construye el intervalo de confianza

$$\log(R) \mp z_{0,025} \cdot \sqrt{V(\log(R))}$$

Un paso no trivial es calcular la expresión de la varianza.

3) Una vez construido este intervalo, se deshace la transformación que se hizo en 1) para obtener un intervalo para la razón R:

$$R \cdot \exp(\mp z_{0,025} \cdot \sqrt{V(\log(R))})$$

Ahora, **para calcular la razón de prevalencias por exposición (de enfermedad) y el intervalo de confianza**, tendremos en cuenta que teóricamente se tiene que:

$$V(\log(R)) = \frac{1}{a} - \frac{1}{(a+b)} + \frac{1}{c} - \frac{1}{(c+d)}$$

El código es entonces:

```
Prev_Exp <- tabla[1,1]/sum(tabla[1,])
```

```
Prev_nExp <- tabla[2,1]/sum(tabla[2,])
Raz_Exp <- (tabla[1,1]*sum(tabla[2,]))/
           (tabla[2,1]*sum(tabla[1,]))

6.151515

EE <- sqrt(1/tabla[1,1]-1/sum(tabla[1,])+
          1/tabla[2,1]-1/sum(tabla[2,]))
IClog <- c(log(Raz_Exp)-1.96*EE, log(Raz_Exp)+1.96*EE)
IC <- exp(IClog)

3.954979 9.567975
```

Para **calcular la razón de prevalencias por enfermedad (de exposición) y el intervalo de confianza**, tendremos en cuenta que teóricamente se tiene que:

$$V(\log(R))=\frac{1}{a}-\frac{1}{(a+c)}+\frac{1}{b}-\frac{1}{(b+d)}$$

El código es entonces:

```
Prev_Enf <- tabla[1,1]/sum(tabla[,1])
Prev_nEnf <- tabla[1,2]/sum(tabla[,2])
Raz_Enf <- (tabla[1,1]*sum(tabla[,2]))/
           (tabla[1,2]*sum(tabla[,1]))

3.741935

EE <- sqrt(1/tabla[1,1]-1/sum(tabla[,1])+
          1/tabla[1,2]-1/sum(tabla[,2]))
IClog <- c(log(Raz_Enf)-1.96*EE, log(Raz_Enf)+1.96*EE)
IC <- exp(IClog)

2.882067 4.858347
```

Por último, **para calcular la razón de puntos (*odds ratio*) y su intervalo de confianza**, tendremos en cuenta que teóricamente se tiene que:

$$V(\log(R))=\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$$

El código es entonces:

```
OR <- (tabla[1,1]*tabla[2,2])/(tabla[1,2]*tabla[2,1])

10.97067

EE <- sqrt(1/tabla[1,1] + 1/tabla[1,2] +
          1/tabla[2,1] + 1/tabla[2,2])
```

o, más elegantemente, esta última cantidad se puede calcular con

```
EE <- sqrt(sum(1/tabla))

IClog <- c(log(OR)-1.96*EE, log(OR)+1.96*EE)
IC <- exp(IClog)

6.243703 19.276332
```

Por otra parte, para obtener los resultados del **contraste ji-cuadrado de asociación (sin corrección)**, tenemos las expresión que compara, celda a celda, la tabla de contingencia que ha salido con la que tendría que haber salido (esperada). No entramos aquí en la teoría que explica cómo obtener esta segunda tabla esperada.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

```
O <- c(tabla[1,1], tabla[1,2], tabla[2,1], tabla[2,2])
E <- c(sum(tabla[1,])*sum(tabla[,1]),
       sum(tabla[1,])*sum(tabla[,2]),
       sum(tabla[2,])*sum(tabla[,1]),
       sum(tabla[2,])*sum(tabla[,2]))/sum(tabla)

Chi <- sum(((O-E)^2)/E)

86.0119

pValor <- 2*pchisq(Chi, 1, lower.tail=FALSE) # Caso bilateral

3.577113e-20
```

Las correcciones por continuidad pretenden enmendar el error que se comete al aproximar una distribución discreta (frecuencia) por una distribución continua (ji--cuadrado). La que tiene *Epidat* implementada es la de Yates, que es quizá la más utilizada. Existe polémica sobre el uso de la corrección, porque existen casos en los que al aplicarla se rechaza la independencia con bastante menor significatividad que sin ella. No obstante, su efecto es pequeño cuando el tamaño muestral es grande. Apliquemos el **contraste ji-cuadrado de asociación con la corrección de Yates**:

$$\chi^2_{Yates} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - .5)^2}{E_i}$$

```
ChiYates <- sum(((abs(O-E)-0.5)^2)/E)

83.50074

pValorY <- 2*pchisq(ChiYates, 1, lower.tail=FALSE) # Bilateral

1.27384e-19
```

Por último, para implementar la **prueba exacta de Fisher**, que intenta medir la asociación de dos variables de una tabla de contingencia, tendremos en cuenta que esta autor mostró que la probabilidad de obtener una tabla de contingencia venía dada por una distribución hipergeométrica:

$$p = \binom{a+b}{a}\binom{c+d}{c} \Big/ \binom{n}{a+c} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Para hacer estos cálculos hay que tener cuidado, porque las cantidades son tan grandes que el lenguaje de programación no puede calcularlas:

```
p <- prod(1:sum(tabla[1,]))*prod(1:sum(tabla[2,]))*
     prod(1:sum(tabla[,1]))*prod(1:sum(tabla[,2]))/
     prod(1:sum(tabla))*prod(1:tabla[1,1])*prod(1:tabla[1,2])*
     prod(1:tabla[2,1])*prod(1:tabla[2,2])
```

```
NaN
```

Le ayudamos calculando a mano los valores de la fórmula, que darían

```
p <- prod(1:80)*prod(1:320)*
     prod(1:120)*prod(1:280)/
     prod(1:400)*prod(1:58)*prod(1:62)*
     prod(1:22)*prod(1:258)
```

y haciendo a mano algunas simplificaciones. Podemos agrupar las cantidades más grandes del numerador con las del denominador, y viceversa, y hacer simplificaciones:

```
p <- prod(80:59)*prod(120:63)*prod(280:259)/
     prod(400:321)* prod(1:22)
```

```
7.089034e+23
```

Vemos que ya es suficiente con este artificio rudimentario (hay otros mucho más avanzados) para que el ordenador pueda calcular la cantidad deseada. Si no fuese suficiente, podríamos ir haciendo los cálculos por pasos agrupando los términos de manera que fuesen comparables en numerador y denominador. Hallamos el nivel crítico para el caso unilateral y bilateral:

```
pValorUnilat <- phyper(p, sum(tabla), sum(tabla[1,]),
                       sum(tabla[,1]), lower.tail = FALSE)
```

```
0
```

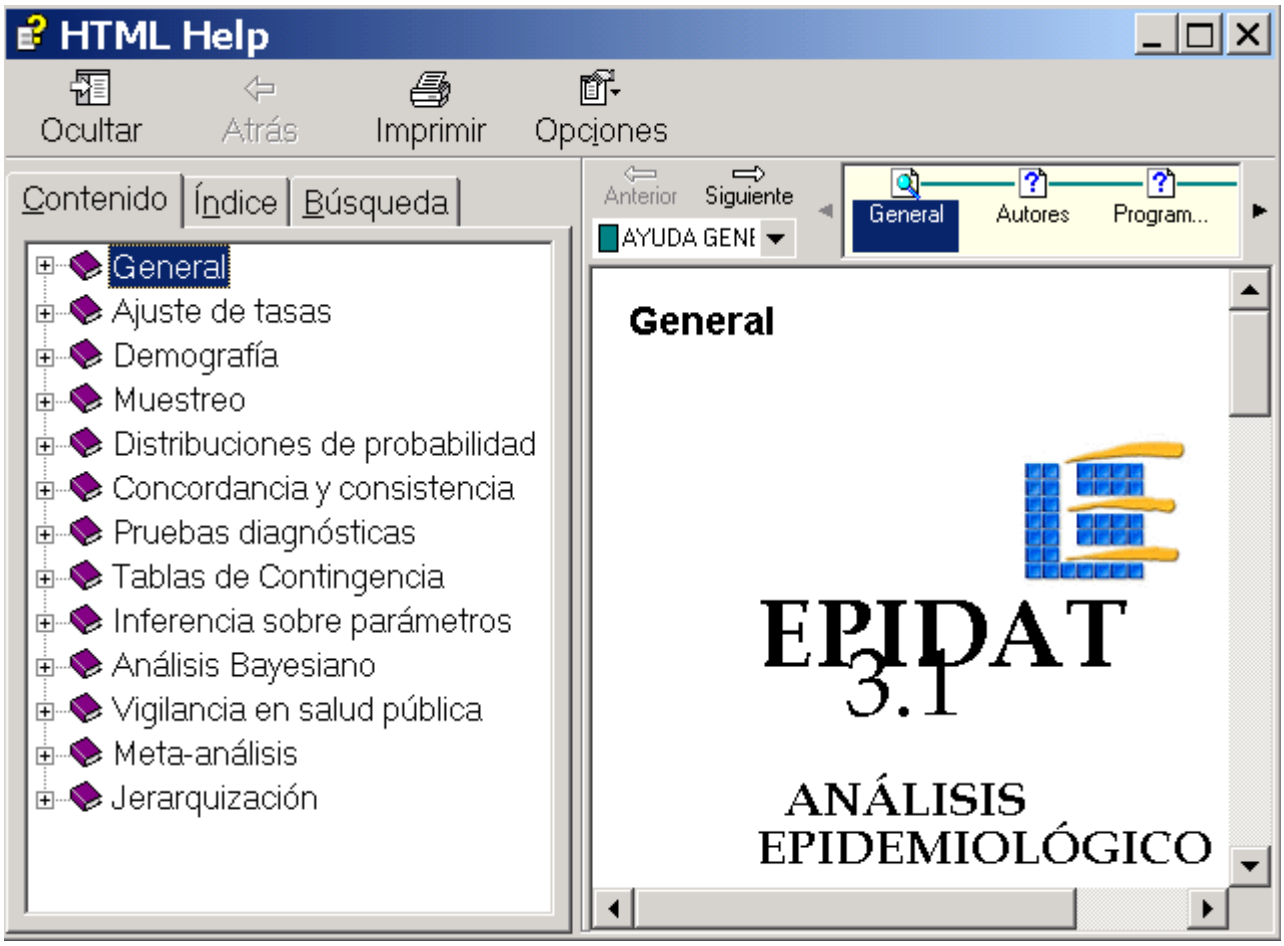
```
pValorBilat <- 2*phyper(p, sum(tabla), sum(tabla[1,]),
                        sum(tabla[,1]), lower.tail = FALSE)
```

Ejercicio 2

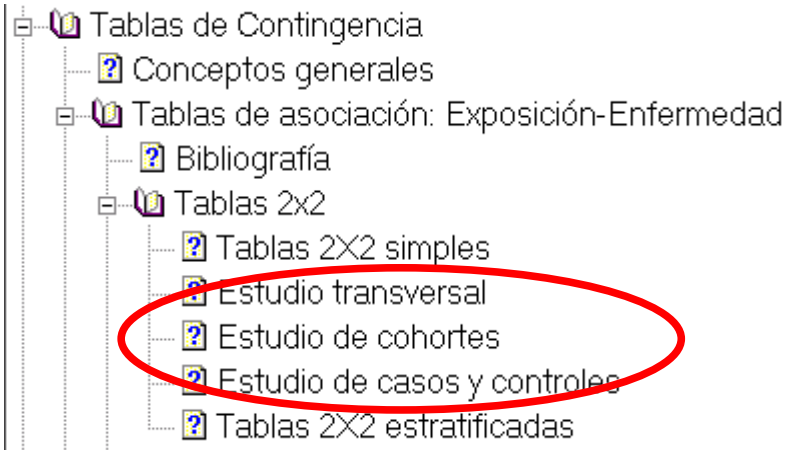
Para acceder a la ayuda del programa, se pulsa la tecla F1 o se entra en:

Ayuda --> Contenido

Como vemos, el contenido de la ayuda está ordenado por tema.



Para acceder a la ayuda específica para los estudios transversales, de los estudios de cohortes y de los estudios de casos y controles, dentro de la ventana anterior entramos en el apartado correspondiente.



Hay suficiente materia en la ayuda como para interpretar la salida de los análisis que hagamos. No obstante, se incluyen algunos otros recursos que pueden ser útiles.

Hay abundantes enlaces a **conceptos de Epidemiología** en la Wikipedia:

http://en.wikipedia.org/wiki/Category:Epidemiology

Relacionados también con la práctica de hoy están estas **herramientas en línea para calcular la razón de puntos (*odds ratio*)** de una tabla:

http://www.swin.edu.au/victims/resources/software/oddsratio/odds_ratio_generator.html

http://www.hutchon.net/ConfidOR.htm

Y esta otra para **calcular la prueba exacta de Fisher** (recordemos cómo nos las hemos tenido que apañar para hacer los cálculos con un lenguaje de programación):

http://www.physics.csbsju.edu/stats/exact2.html

Otras fuentes más generales que vamos a visitar son:

Centro Nacional de Epidemiología:

http://cne.isciii.es/

Sociedad Española de Epidemiología:

http://as-seepidemiologia.es/

Bioestadística en internet:

http://www.seh-lelha.org/webestad.htm

26/04/07

David Casado de Lucas: http://www.est.uc3m.es/dcasado/
Departamento de Estadística
Universidad Carlos III de Madrid