



Bioestadística

Práctica 1

Epidemiología

Según el libro «Vocabulario científico y técnico», de la Real Academia de Ciencias Exactas, Física y Naturales, la Bioestadística es la *Parte de la estadística que se refiere a la aplicación de los métodos estadísticos a problemas de la biología*. Y según el diccionario de la Real Academia Española (DRAE) la Bioestadística es la *Ciencia que aplica el análisis estadístico a los problemas y objetos de estudio de la biología*. Aunque éste es el nombre de la asignatura, el área a la cual se restringe el temario es el de la Epidemiología. Este término no se incluye en el primer libro mencionado (!), y en el segundo se define escuetamente como *Tratado de las epidemias*. Quizá aclaren algo las siguientes definiciones del DRAE:

endemia.

(Del gr. ἐνδημία, que afecta a un país).

1. f. *Med.* Enfermedad que reina habitualmente, o en épocas fijas, en un país o comarca.

epidemia.

(Del gr. ἐπιδημία).

1. f. Enfermedad que se propaga durante algún tiempo por un país, acometiendo simultáneamente a gran número de personas.

Y, como casos particulares:

pandemia.

(Del gr. πανδημία, reunión del pueblo).

1. f. *Med.* Enfermedad epidémica que se extiende a muchos países o que ataca a casi todos los individuos de una localidad o región.

epizootia.

(De *epi-* y el gr. ζωότης, naturaleza animal, con el infl. de *epidemia*).

1. f. *Veter.* Enfermedad que acomete a una o varias especies de animales, por una causa general y transitoria. Es como la epidemia en el hombre.

2. f. *Chile.* Glosopeda.

Una definición de Epidemiología puede ser la del primer artículo del material de la asignatura («Desarrollo histórico de la epidemiología: su formación como disciplina científica»; Sergio López-Moreno, Francisco Garrido-Latorre y Mauricio Hernández-Ávila; salud pública de México, vol. 42, no.2, 2000): *La epidemiología es la rama de la salud pública que tiene como propósito describir y explicar la dinámica de la salud poblacional, identificar los elementos que la componen y comprender las fuerzas que la gobiernan.* Respecto a los niveles de actuación de la Epidemiología, pueden ser uno o varios de los siguientes: simplemente describir una enfermedad, estudiar sus causas (etiología) o, además, intervenir en la población.

Merece también la pena leer la introducción incluida en la enciclopedia libre *Wikipedia*:

La epidemiología es la parte de la [medicina](#) que se dedica al estudio de la [distribución](#), [frecuencia](#), [determinantes](#), [relaciones](#), [predicciones](#) y [control](#) de factores relacionados con la [salud](#) y [enfermedad](#) en poblaciones humanas determinadas, así como la aplicación de este estudio a los problemas de salud. Por lo tanto la epidemiología estudia la salud de los grupos humanos en relación con su medio.

La epidemiología se considera la ciencia básica para la [medicina preventiva](#) y una fuente de información para la formulación de políticas de [salud pública](#). La epidemiología estudia, sobre todo, la relación causa-efecto entre exposición y enfermedad. Las enfermedades no se producen de forma aleatoria; tienen causas, muchas de ellas de origen humano, que pueden evitarse. Por tanto, muchas enfermedades podrían prevenirse si se conocieran sus causas. Los métodos epidemiológicos han sido cruciales para identificar numerosos factores etiológicos que, a su vez, han justificado la formulación de políticas sanitarias encaminadas a la prevención de enfermedades, [lesiones](#) y muertes prematuras.

Inicialmente la epidemiología surgió del estudio de las [epidemias](#) de las [enfermedades infecciosas](#). En el siglo XX los estudios de la epidemiología se basan en el estudio [demográfico](#) de cualquier enfermedad con la ayuda de la [estadística](#).

<http://es.wikipedia.org/>

Epidat

Entre los distintos programas que existen específicamente diseñados para hacer los cálculos que la Epidemiología requiere, muchos de los cuales se podrían hacer con la calculadora o incluso a mano, nosotros utilizaremos el programa gratuito *Epidat*. Es un programa de libre distribución desarrollado por instituciones públicas y dirigido a epidemiólogos y otros profesionales de la salud para el manejo de datos tabulados. Este programa es resultado de la colaboración entre la [Consejería de Salud](#) de la [Junta de Galicia](#) y la [Organización Panamericana de Salud](#). Para saber qué tiene implementado el programa, para descargarlo o para saber cuáles son los errores que se han encontrado en la última versión del programa, visitar:

<http://dxsp.sergas.es/ApliEdatos/Epidat/cas/default.asp>

<http://www.paho.org/spanish/sha/epidat.htm>

Aunque nosotros no utilizaremos todas las prestaciones del programa, con *Epidat* se

pueden hacer:

■ **Ajuste de Tasas**

- Método directo
- Método indirecto

■ **Demografía**

- Pirámides e indicadores demográficos
- Tablas de mortalidad abreviadas
- APVP
- Descomposición del cambio en la esperanza de vida
- Años de esperanza de vida perdidos

■ **Muestreo**

- Cálculo de tamaños de muestra
- Selección muestral
- Asignación de sujetos a tratamientos

■ **Distribuciones de probabilidad**

- Cálculo de probabilidades
- Generación de distribuciones

■ **Concordancia y consistencia**

- *Concordancia*
 - Dos observadores, dos o más categorías
 - Tres o más observadores
 - Comparación de kappas
- *Consistencia*
 - Alfa de Cronbach

■ **Pruebas diagnósticas**

- Pruebas simples
- Pruebas múltiples
- Prueba de referencia imperfecta
- Curvas ROC
- Curva de Lorenz

■ **Tablas de contingencia**

- Exposición-enfermedad
 - Tablas 2x2
 - Tablas 2xN
- Tablas generales
 - Tablas MxN
 - Regresión logística

■ **Inferencia sobre parámetros**

- Una población
- Dos poblaciones

■ **Análisis bayesiano**

- Proporción
- Media
- Tablas de contingencia
- Valoración bayesiana de pruebas convencionales

■ Vigilancia en salud pública

- Captura-Recaptura
- Detección de clusters
- Gráficos
- Ondas epidémicas
- Efectividad vacunal

■ Meta-análisis

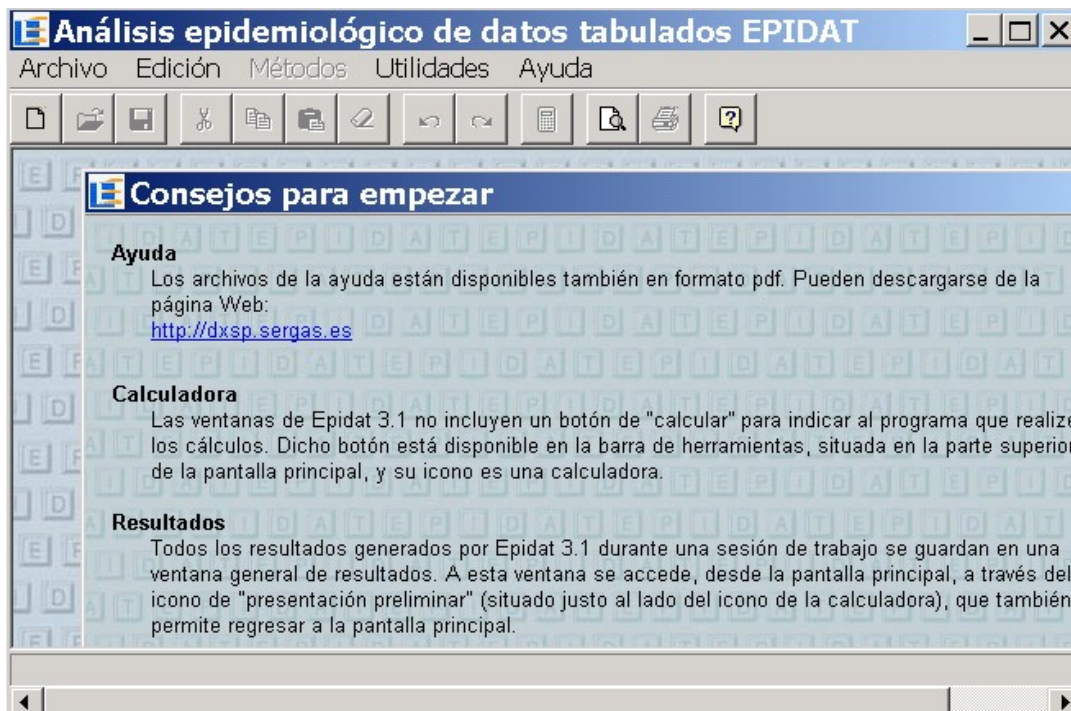
- Odds ratios
- Riesgos relativos
- Diferencia de riesgos
- Diferencia estandarizada de medias

■ Jerarquización

- Coeficiente de Gini y curva de Lorenz
- Índice de concentración y curva de concentración
- Índice de necesidad en salud
- Índice de desarrollo en salud comunitaria
- Índice de inequidades en salud
- Índice de disimilitud
- Índice de desarrollo humano
- Índice de desarrollo relativo al género

En la segunda de las URL dadas arriba están disponibles, en formato PDF y por temas, los archivos de ayuda del programa.

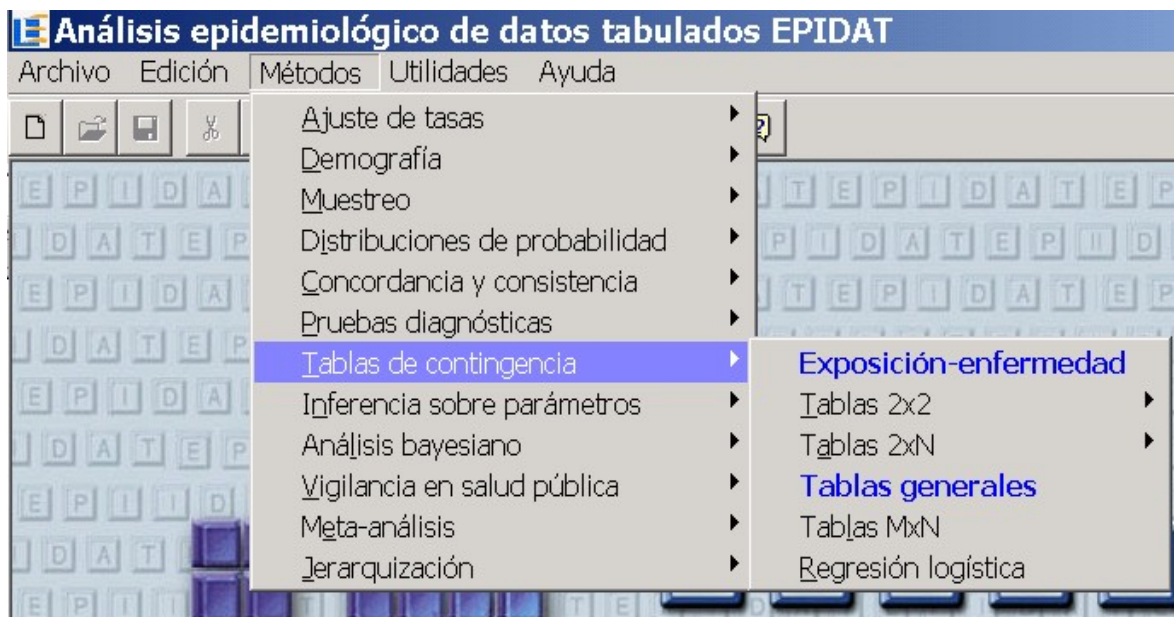
Interfaz de *Epidat*



Una vez descargado (es necesario rellenar un pequeño registro) e instalado el programa, lo primero que aparece al abrir el programa es la imagen anterior. Como la mayoría de programas, en la parte superior de la ventana se muestran tanto una barra de menús como una

barra de herramientas. En ese cuadro de ayuda nos informa de dónde está el botón que ordena al programa hacer los cálculos, una vez metidos los datos, y cómo presentará el programa los resultados.

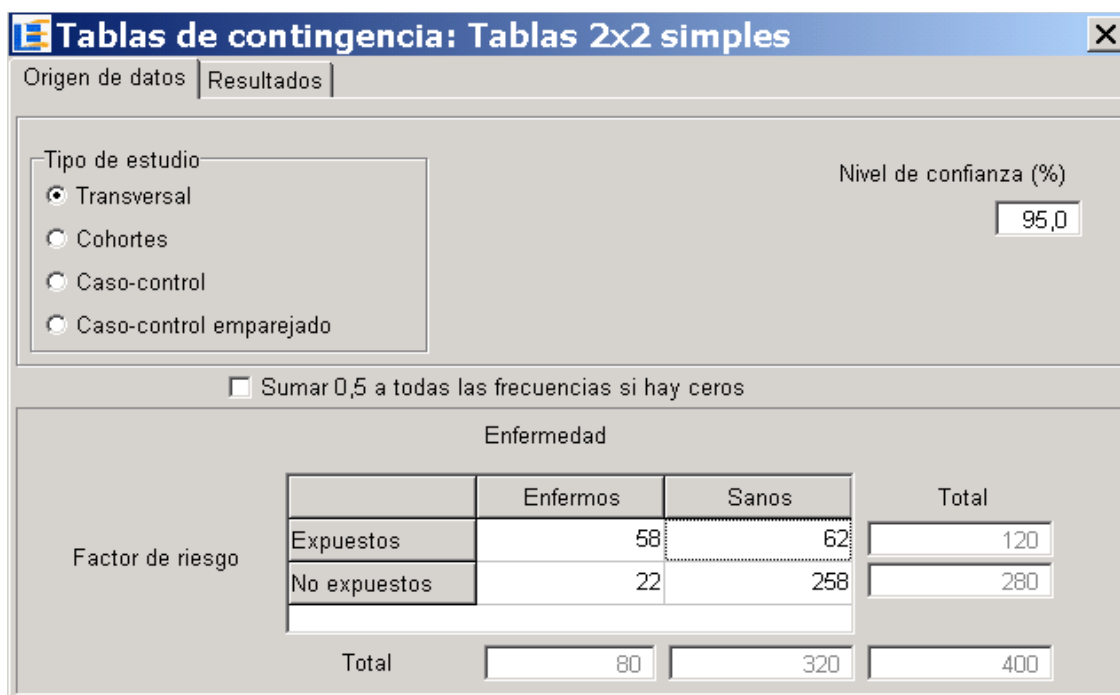
Prácticamente todos los cálculos se hacen a través del menú *Métodos*:



Sin preocuparnos todavía de la interpretación de lo que vamos a hacer, veamos con un ejemplo cómo se introducen los datos y cómo proporciona el programa los resultados. En concreto, analizemos los resultados de un estudio transversal (tomado del archivo de ayuda de *Epidat* para el tema de tablas de contingencia). Para ello entremos en el menú:

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

Introduzcamos los datos como se indica en la siguiente imagen. Para pasar de una celda a la siguiente pulsamos la tecla *Intro*, utilizamos las flechas del teclado o pulsamos con el cursor en ella.



La anterior ventana tiene dos pestañas: la primera para introducir los datos y elegir el tipo de estudio a que pertenecen, la segunda para mostrar los resultados. Al pulsar el botón de calcular, la pestaña de resultados viene al primer plano de la ventana:

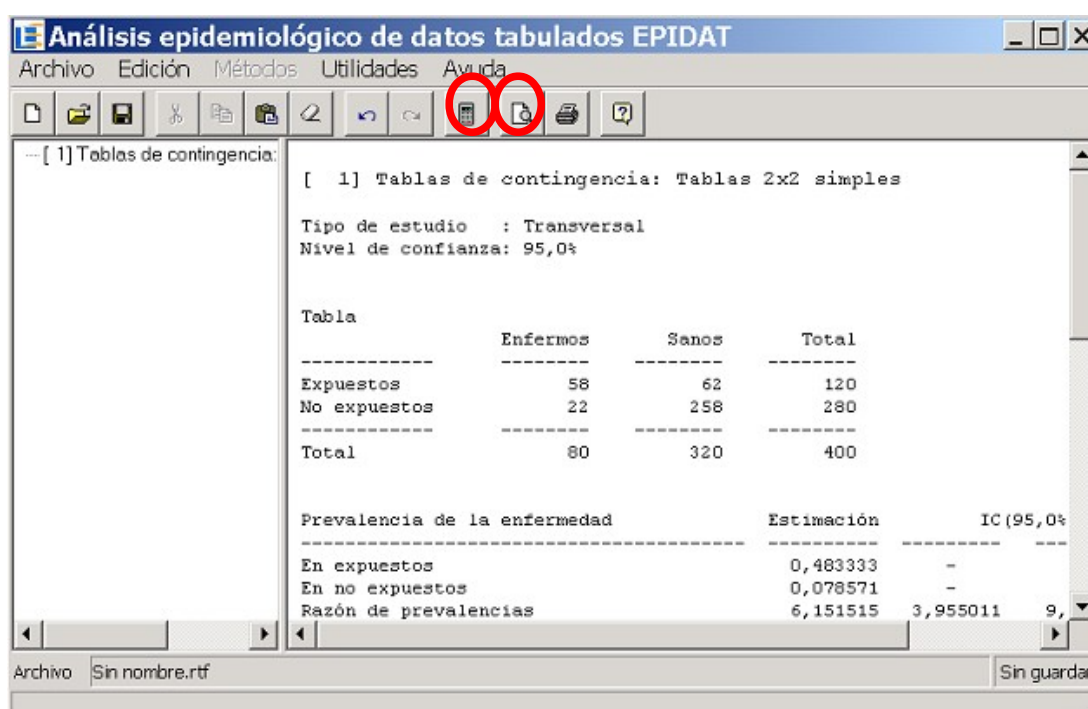
Origen de datos | Resultados

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	6,151515	3,955011	9,56789

En las siguientes prácticas nos centraremos en la interpretación de los resultados; los archivos de ayuda del programa están dedicados principalmente a ello. En ésta nos dedicaremos a aprender a manejar el programa. Desde la pestaña anterior no es posible guardar los resultados. Para ello es necesario antes pasar a la presentación preliminar de los resultados, que es una ventana en la que se van guardando todos los resultados de toda una sesión, no sólo del último análisis, y en la que se pueden editar los resultados. Para pasar de la pestaña de resultados a la ventana de resultados —y viceversa— basta pulsar el botón de la barra de herramientas que tiene el dibujo de la vista preliminar.



Para **guardar resultados**, desde esta última ventana de resultados, no desde la pestaña de resultados, hacer:

Archivo --> Guardar como

El formato en el que se guarda la información es un formato de texto no propietario, de extensión «.rtf» (de *rich text format*), que se puede editar después con cualquier procesador de textos (*Word, WordPad, Abiword*, etcétera).

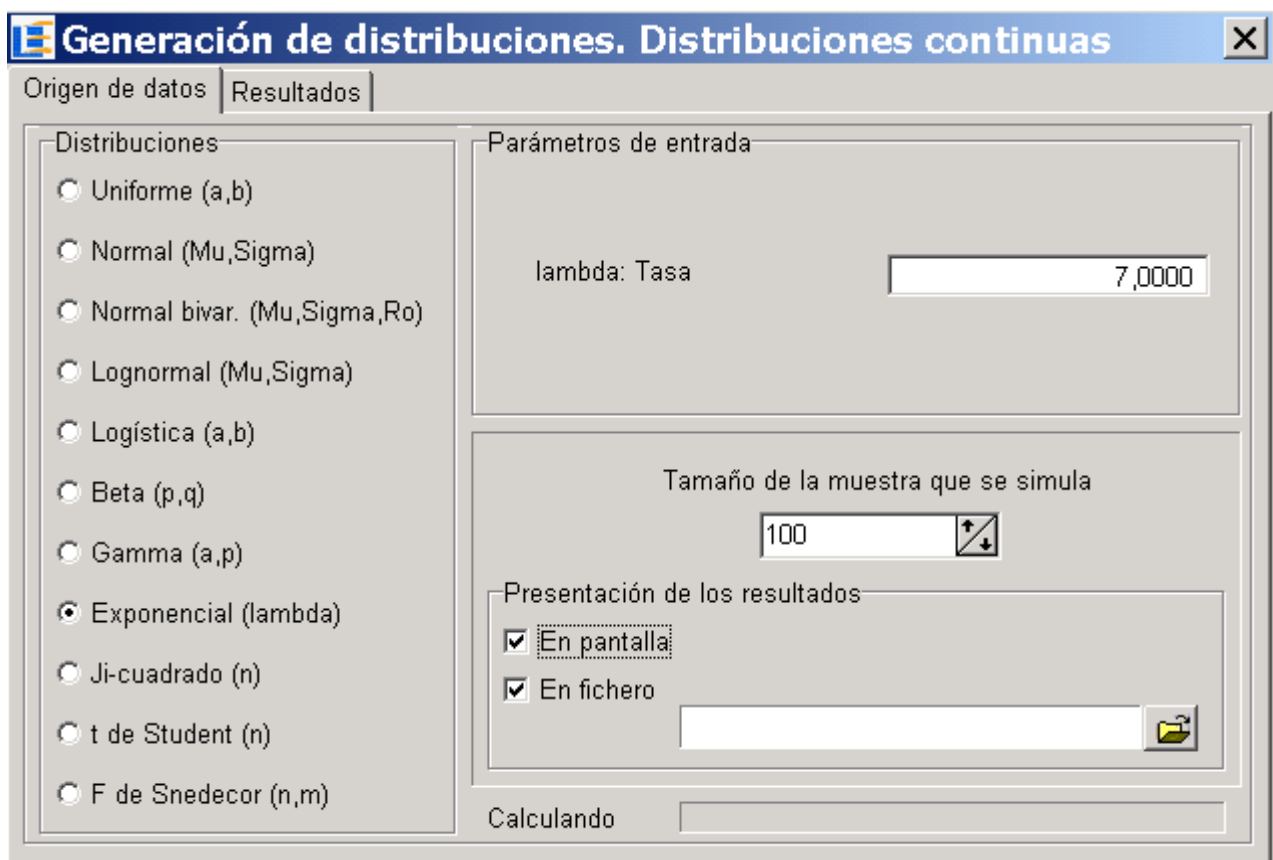
Para **cargar resultados** que ya existen en un archivo de extensión «.rtf» o «.txt», también desde esta última ventana de resultados (es decir, abriéndola desde el botón de presentación preliminar) hacer:

Archivo --> Abrir

y buscar el archivo que queremos cargar.

Como último ejercicio, practicaremos cómo guardar datos. Para ello generaremos y guardaremos una muestra de una variable aleatoria:

Métodos --> Distribuciones de probabilidad --> Generación de distribuciones --> Entonces se elige una distribución y los valores de sus parámetros. Hay que elegir también si queremos que los resultados los muestre por pantalla, los guarde en un archivo o ambas cosas. Los formatos de datos que puede manejar *Epidat* son los de *Excel, Access* y *DBase*.



En mis pruebas no he podido —o sabido— guardar los resultados en el formato de *Excel*,

porque me da un tipo de error; pero sí los muestra si selecciono sólo *En pantalla*, y sí los guarda en el formato de *Access*, por ejemplo.

Por último, el programa *Epidat* tiene incorporada también la ayuda de los archivos de su sitio web mencionados antes. Para **acceder a la ayuda**:

Ayuda --> Contenido



Bioestadística

Práctica 2

En esta práctica vamos a aprender a analizar (con *Epidat*) los datos provenientes de distintos tipos de estudios epidemiológicos. En cada caso vamos a identificar en la salida que proporciona el programa los distintas clases de medidas vistas en teoría, y reproduciremos algunas de las medidas con la calculadora.

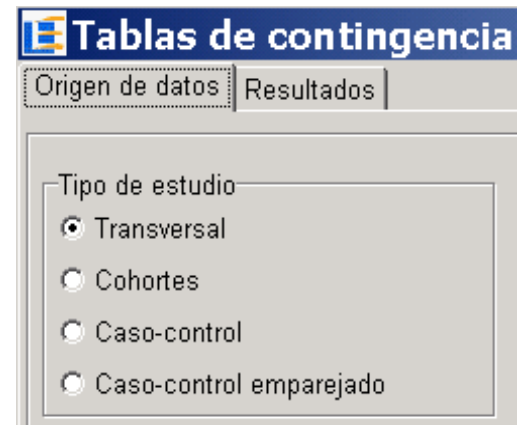
Introducción

Después de ver un resumen histórico de la Epidemiología, se han definido los conceptos básicos —pero cruciales— en que se apoya esta disciplina: relación de causalidad, diferencia entre asociación estadística y causalidad, características de las relaciones causales, tipos de causas y factor de riesgo. Recordemos que *la existencia de una asociación estadísticamente significativa entre la causa y su efecto es uno de los criterios para proponer una relación causal; aunque hay que tener en cuenta que no es el único.*

En el tema 2 se han clasificado los estudios epidemiológicos según distintos criterios. El programa permite analizar cómodamente los resultados cuando se encuentran ya en tablas de contingencia. Los menús del programa hacen una primera clasificación de los análisis por el tamaño de las tablas y por la existencia o no de estratos en la población del estudio



Una vez seleccionada la forma de la tabla en que se encuentran nuestros datos, *Epidat* trabaja con una clasificación de los tipos de estudio que corresponde a la que en los apuntes denomina «según la selección de la población». El programa va a solicitarnos que especifiquemos si la tabla pertenece a un **estudio transversal**, de **cohortes** o de **caso-control (normal o emparejado)**. Para recordar estos conceptos, vamos a buscar las definiciones en el artículo «Diseño de estudios epidemiológicos»; Mauricio Hernández-Ávila, Francisco Garrido-Latorre y Sergio López-Moreno; salud pública de México; vol. 42, no.2, 2000.



Por otro lado, se han clasificado también las distintas medidas que se utilizan en Epidemiología en **medidas de frecuencia**, **medidas de asociación** y **medidas de impacto**. Para recordar estas definiciones, vamos a utilizar el artículo «Principales medidas en epidemiología»; Alejandra Moreno-Altamirano, Sergio López-Moreno y Alexánder Corcho-Berdugo; salud pública de México; vol. 42, no. 4, 2000. Recordemos también la relación que hay entre los niveles de actuación de la Epidemiología —descriptivo, de conocimiento etiológico y de intervención— y los tipos de medidas —de frecuencia, de asociación y de impacto—, respectivamente.

En el mismo artículo se explican tanto las principales escalas de medición para esas medidas, **escalas nominal, ordinal, de intervalo y de razón**, como los conceptos matemáticos con que se definen las medidas, **proporción, tasa y razón**. Comprender bien la diferencia entre estos tres conceptos ayuda mucho a saber por qué las medidas se definen como se definen y cómo interpretarlas. Sacamos del artículo las primeras palabras de cada definición:

Proporciones: *Las proporciones son medidas que expresan la frecuencia con la que ocurre un evento en relación con la población total en la cual éste puede ocurrir.*

Tasas: *Las tasas expresan la dinámica de un suceso en una población a lo largo del tiempo. Se pueden definir como la magnitud del cambio de una variable (enfermedad o muerte) por unidad de cambio de otra (usualmente el tiempo) en relación con el tamaño de la población que se encuentra en riesgo de experimentar el suceso.*

Razones: *Las razones pueden definirse como magnitudes que expresan la relación aritmética existente entre dos eventos en una misma población, o un solo evento en dos poblaciones.*

Es interesante ahora que dediquemos unos minutos a pensar en las siguientes cuestiones:

- ¿Entre qué valores estará, por su propia definición, cada uno de estos conceptos?
- ¿Qué unidades tendrán?
- ¿Cuáles serán adecuadas y tendrán sentido para cada tipo de estudio?

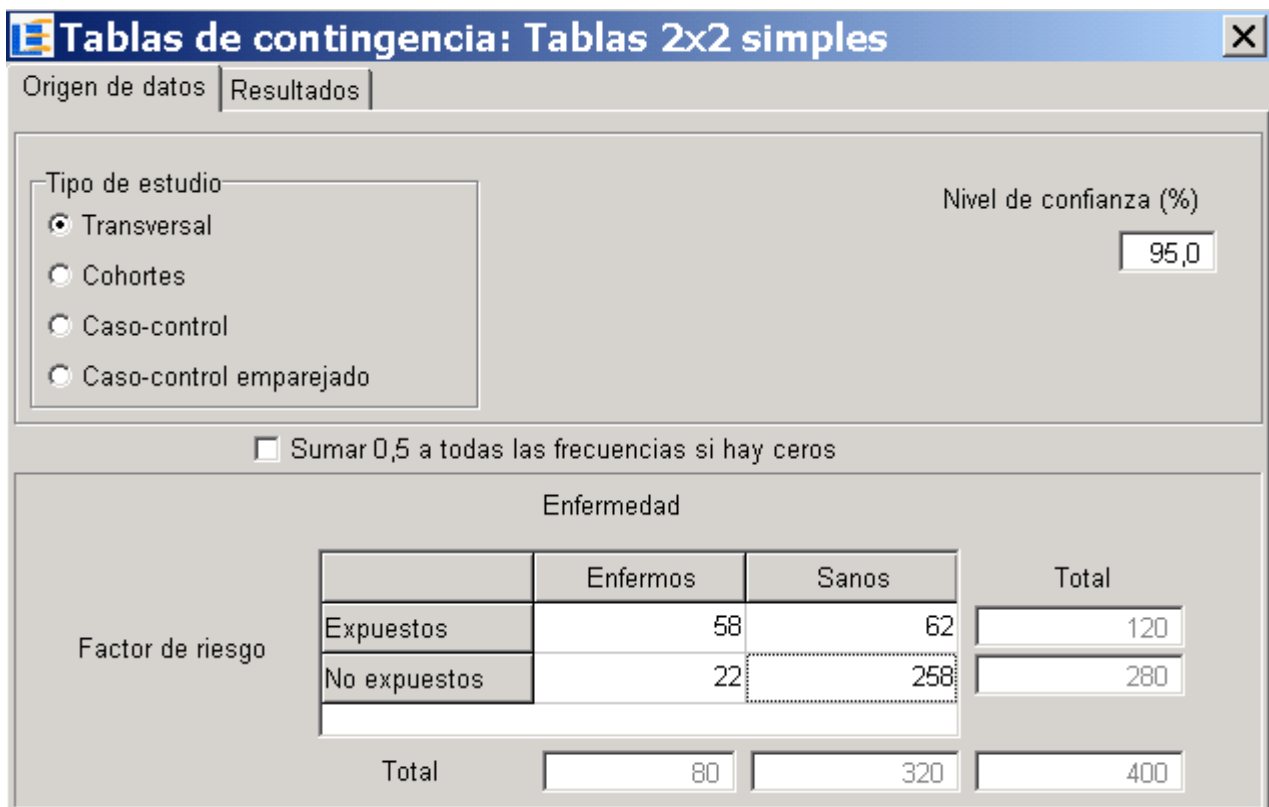
Ejercicio 1

Como primer ejercicio de la práctica, y después de esta larga introducción, vamos a estudiar la salida que *Epidat* proporciona para cada tipo de estudio. Además, para poder comparar las distintas salidas, vamos a hacer seguidamente cinco análisis. Los datos están tomados del archivo de ayuda de *Epidat* del tema de tablas de contingencia, donde se explican e interpretan los resultados. Nosotros insistiremos en la interpretación de estos resultados en la práctica siguiente, después de estudiar teóricamente este tipo de estudios.

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



		Enfermedad		Total
		Enfermos	Sanos	
Factor de riesgo	Expuestos	58	62	120
	No expuestos	22	258	280
Total		80	320	400

Para analizar datos de un **estudio de cohortes** (utilizando la **incidencia acumulada**) introducimos los datos como en la figura siguiente y pulsamos el botón de cálculo

Tablas de contingencia: Tablas 2x2 simples

Origen de datos | Resultados

Tipo de estudio:
 Transversal
 Cohortes
 Caso-control
 Caso-control emparejado

Tipo de datos:
 Tasa de incidencia
 Incidencia acumulada

Nivel de confianza (%)

Sumar 0,5 a todas las frecuencias si hay ceros

Enfermedad

		Enfermos	Sanos	Total
Factor de riesgo	Expuestos	76	6169	6245
	No expuestos	28	7867	7895
Total		104	14036	14140

Obsérvese que en este caso el programa nos ofrece trabajar con una de las dos medidas de frecuencia siguientes: la tasa de incidencia o la incidencia acumulada. Para analizar datos de un **estudio de cohortes** (utilizando la **tasa de incidencia**) introducimos los datos como en la figura siguiente y pulsamos el botón de cálculo

Tablas de contingencia: Tablas 2x2 simples

Origen de datos | Resultados

Tipo de estudio:
 Transversal
 Cohortes
 Caso-control
 Caso-control emparejado

Tipo de datos:
 Tasa de incidencia
 Incidencia acumulada

Nivel de confianza (%)

Sumar 0,5 a todas las frecuencias si hay ceros

		Casos	Personas-Tiempo
Factor de riesgo	Expuestos	76	116157
	No expuestos	28	177636
Total		104	293793

Para analizar datos de un **estudio de casos y controles** introducimos los datos como se indica en la figura siguiente y después pulsamos el botón de cálculo

Tablas de contingencia: Tablas 2x2 simples

Origen de datos | Resultados

Tipo de estudio

- Transversal
- Cohortes
- Caso-control
- Caso-control emparejado

Nivel de confianza (%)

Sumar 0,5 a todas las frecuencias si hay ceros

Enfermedad

		Casos	Controles	Total
Factor de riesgo	Expuestos	255	487	742
	No expuestos	500	268	768
Total		755	755	1510

Por último, para analizar los datos del **estudio de casos y controles emparejados**

Tablas de contingencia: Tablas 2x2 simples

Origen de datos | Resultados

Tipo de estudio

- Transversal
- Cohortes
- Caso-control
- Caso-control emparejado

Nº de controles por caso Nivel de confianza (%)

Sumar 0,5 a todas las frecuencias si hay ceros

Controles

		Expuestos	No expuestos	Total
Casos	Expuestos	160	95	255
	No expuestos	327	173	500
Total		487	268	755

Una vez que hemos hecho los cuatro análisis, pulsando el botón de vista preliminar, accedemos a la ventana en la que se han guardado los resultados de toda la sesión (no sólo del

último análisis, que es lo que muestra la pestaña de resultados).

El ejercicio ahora consiste en:

- ➔ Identificar las medidas de la salida de cada uno de los análisis por separado. ¿Da el programa las medidas ordenadas en medidas de frecuencia, de asociación y de impacto? ¿Cuáles son de cada tipo?
- ➔ Comparar entre sí las salidas de los distintos tipos de estudio.

Ejercicio 2

Los conceptos involucrados en Epidemiología suelen tener **definiciones bastante sencillas**. De hecho puede parecer que algunas de estas definiciones «coinciden», en el sentido de que se calculan de la misma forma a partir de los datos de las celdas de las tablas de contingencia. Las diferencias hay que buscarlas en **lo que los valores de las celdas significan** según el tipo de estudio del que procedan. En este ejercicio vamos a reproducir los cálculos de algunas de las medidas que proporciona el programa: así por un lado repasamos sus definiciones y por otro vemos que muchos de esos cálculos pueden hacerse fácilmente con una calculadora, o incluso a mano. La salida del programa para el estudio transversal anterior es la siguiente:

```
[ 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

Este ejercicio consiste en buscar las definiciones de los conceptos de esta salida que resaltados y, aplicando esa definición, reproducir los cálculos que ha hecho el ordenador con una calculadora o a mano. Un ejercicio un poco más difícil —pero no mucho más— consistiría en hacer lo mismo para las medidas del estudio de cohortes, donde habría que calcular las cantidades persona-tiempo.



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 transversales, de cohortes y de casos-controles.

Ejercicio 1

En uno de los ejercicios de la práctica anterior estuvimos hallando con la 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

		Enfermedad		
		Enfermos	Sanos	Total
Factor de riesgo	Expuestos	58	62	120
	No expuestos	22	258	280
Total		80	320	400

De la siguiente salida, vamos a calcular ahora las cantidades que están resaltadas

[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	6,151515	3,955011	9,567897	(Katz)

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

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 la 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

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

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 :

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

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

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

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:

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

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:

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

El código es:

```
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 la 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**:

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

```
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 este autor mostró que la probabilidad de obtener unos determinados valores de una tabla de contingencia venía dada por una distribución hipergeométrica:

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

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)
```

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 sencillo (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)
```

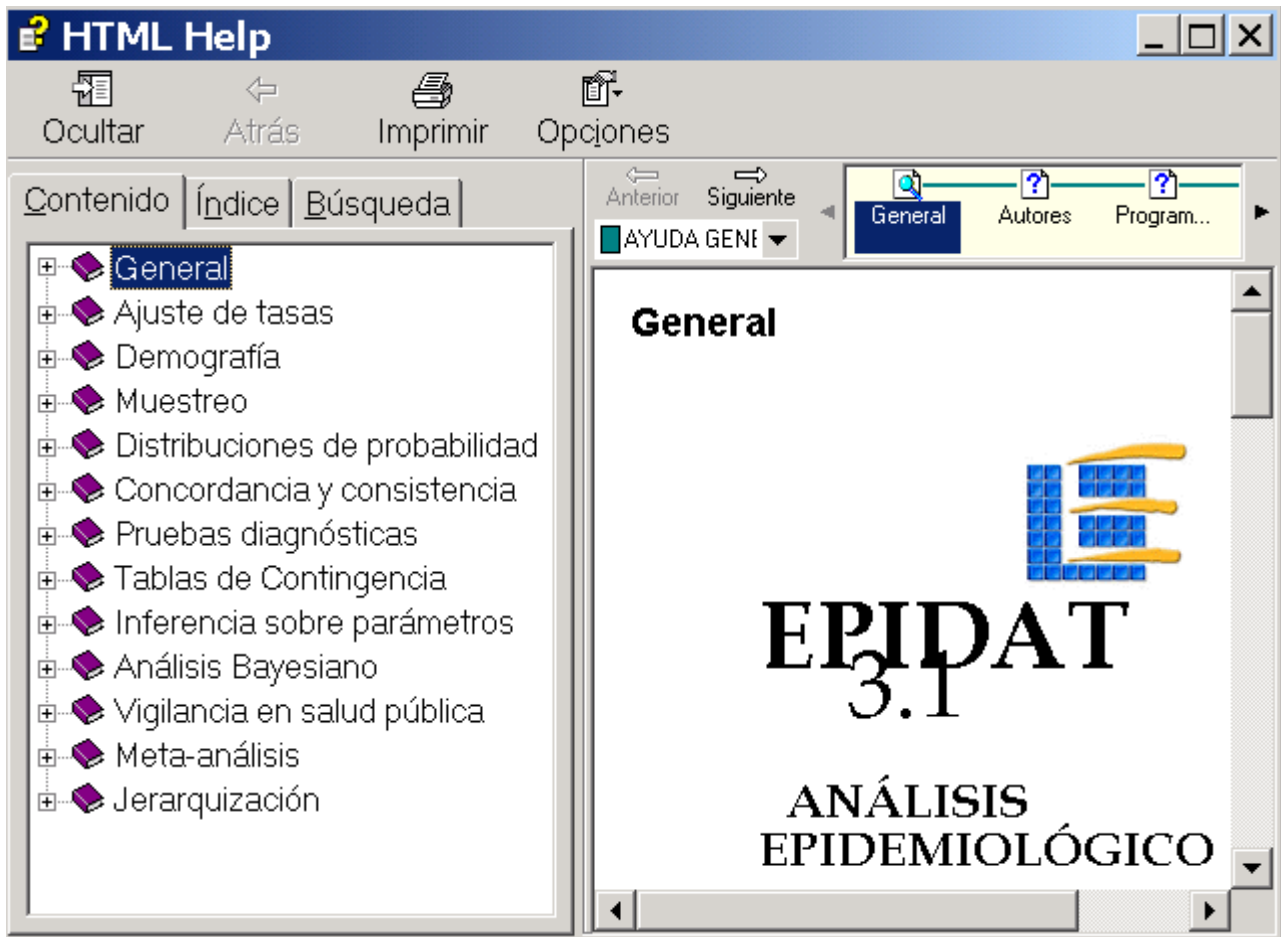
0

Ejercicio 2

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

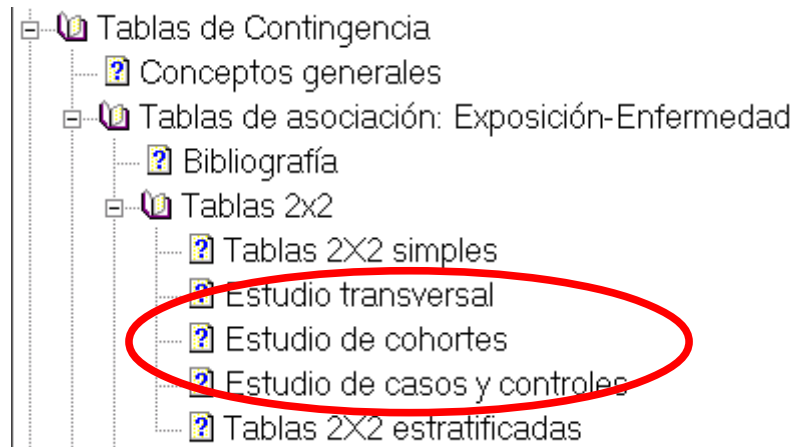
Ayuda --> Contenido

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.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://www.seepidemiologia.es/>

Bioestadística en internet:

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



Bioestadística

Práctica 4

Vamos a insistir en esta práctica en los conceptos de *variable modificadora de efecto* y *variable de confusión*. Veremos también la *regresión logística*. Aplicaremos a unos mismos datos tanto el análisis propio de las tablas de contingencia como la regresión logística; esto nos permitirá resaltar algunas propiedades de la regresión logística, a la vez que relacionar los resultados de ambos métodos.

Ejercicio 1

En uno de los últimos capítulos de teoría de la asignatura se han introducido los conceptos de *modificación de efecto* y *confusión*, cruciales en Epidemiología a la hora de interpretar los resultados de un estudio. Hay ocasiones en las que alguna variable distinta a las que hemos incluido como causa y efecto es más o menos «responsable» de esos resultados. Podemos pensar entonces que se trata de una variable que modifica el efecto o de una variable que confunde. Veamos qué diferencias hay entre unas y otras (de «¿Qué es una variable modificadora de efecto?»; Jokin de Irala, Miguel Ángel Martínez-González y Francisco Guillén Grima; Medicina Clínica; 117, 297-302, 2001; y de «¿Qué es una variable de confusión?»; —; Medicina Clínica; 117, 377-385, 2001):

Variable de confusión: Siempre que se estima una medida de asociación, el resultado es suma de dos factores: el efecto real y la «confusión». Esta última se debe a que es imposible, en general, estudiar exactamente la misma población en las situaciones de expuesta y no expuesta a una determinada causa. Una variable (o factor) de confusión es una variable que distorsiona la medida de la asociación entre otras dos variables. Las condiciones que debe verificar una variable para ser de confusión son: 1) Estar asociada con el desenlace, tanto en expuestos como en no expuestos; 2) Estar asociada con la exposición pero no ser un resultado de la misma; 3) No ser un eslabón causal intermedio entre la exposición y el desenlace. La edad y el sexo son las principales candidatas a ser variables de confusión. Se habla de confusión cuando existen diferencias importantes entre los resultados brutos y los ajustados por los posibles factores de confusión. Entre los métodos para evitarla están: aleatorización de la muestra, restricción en la admisión de sujetos para la muestra y el emparejamiento de los datos. Entre los métodos para identificarla están: la estratificación y el análisis multivariante. Dentro del primer método, el estimador de Mantel-Haenszel combina las medidas de los distintos estratos mediante un promedio ponderado. Se suele considerar que una variable es confusora si el sesgo entre las razones cruda y ponderada es superior al 10%.

Variable modificadora de efecto: Es un concepto complejo que debe distinguirse claramente de la confusión, ya que su identificación determinará una actitud radicalmente opuesta por parte del investigador: así como en presencia de confusión el objetivo es eliminar una distorsión de la medida de asociación objeto de la investigación, ante la presencia de interacción el objetivo es describir mejor un fenómeno, una riqueza existente en los datos. Se puede concluir que hay modificación de efecto o lo contrario según la escala —aditiva o multiplicativa— escogida para evaluar el fenómeno. Es preciso, por tanto, especificar la escala en que se está midiendo. Además, para hablar del efecto de una exposición es necesario tener en cuenta el valor de la otra variable o el estrato de la otra variable en el que se evalúa

dicho efecto. Cuando no hay interacción se puede explicar el efecto de cada variable independientemente de las otras. En cuanto a la identificación: Se evalúa la relación causa-efecto, se excluyen las variables de confusión, se hace un análisis por estratos.

El siguiente cuadro, tomado de los mismos trabajos, aclara bien la diferencia entre ambos tipos de variables.

Características de las variables de confusión y de los modificadores del efecto

Características	Confusión	Interacción
Significado biológico	No corresponde a un fenómeno biológico Es una distorsión de la asociación entre una exposición y un desenlace debido a una tercera variable que es el factor de confusión	Puede corresponder a un fenómeno biológico, sobre todo cuando existe una modificación de la aditividad de las medidas de efecto Corresponde al cambio del verdadero valor de la asociación entre una exposición y un desenlace, en los diversos niveles de una tercera variable que es la «modificadora del efecto»
Consecuencia de su presencia	Introduce un error o distorsión en la estimación de la medida de asociación	Enriquece la información que se puede dar de la medida del efecto
Reproducibilidad	No se reproduce necesariamente en el tiempo ni en otros estudios	En el caso de representar un fenómeno biológico, debería reproducirse en el tiempo o en otros estudios
Identificación	Elaboración de gráficas causales (DAG) Comparar asociaciones brutas con un promedio de las ajustadas (Mantel-Haenzel) Análisis estratificado Análisis multivariable+	Análisis del efecto de una variable en subgrupos de la otra Utilizar escalas aditivas y multiplicativas Análisis por subgrupos Análisis multivariable (términos de producto)
Actuación del investigador	Eliminar el efecto de confusión Prevenirlo en el diseño, controlarlo en el análisis, ajustando por cada factor de confusión	Describir en detalle este fenómeno Tablas con medidas del efecto de la exposición separados para cada subgrupo del modificador del efecto Su existencia no depende del diseño del estudio
Metodología analítica	Análisis estratificado, análisis multivariable, ajuste de tasas, otros procedimientos	Estimación manual o automática de medidas de asociación en subgrupos a partir de los datos de un modelo
Presentación científica	Estimación del efecto de la exposición ajustando por los factores de confusión (En todo caso, comparación entre medidas del efecto brutas y ajustadas)	No se puede presentar un solo valor de la medida de efecto. Valores de medidas de asociación en cada subgrupo de interés Ayuda de gráficos para una mejor comprensión

En los dos ejemplos desarrollados en los apuntes de la asignatura (con *Epidat*, no los repetimos aquí), pueden verse los siguientes pasos para la identificación y actuación de una variable que sospechamos que puede ser modificadora de efecto o confusora:

Ejemplo 1: Se estratifica por la variable sospechosa (sexo) -> Para estudiar al capacidad modificadora de efecto de la variable, por un lado se comparan los valores de las razones de puntos de ambos estratos y por otro lado se observa el resultado del contraste de homogeneidad -> Como las razones son cercanas y no hay evidencia para rechazar la hipótesis de homogeneidad, se considera que la variable no es modificadora de efecto -> Se evalúa si la variable es confusora: para ello se unen los datos en una única muestra (son homogéneos) y se comparan las razones de puntos cruda y combinada (Mantel-Haenzel) -> Como son bastante distintas y el sesgo debido a la variable es del 155%, se concluye que es una variable de confusión -> El resultado que se considera correcto es el de la ponderación de Mantel-Haenzel, que ajusta por la variable.

Ejemplo 2: Se estratifica por la variable sospechosa (obesidad) -> Para estudiar al capacidad modificadora de efecto de la variable, por un lado se comparan los valores de las razones de puntos de ambos estratos y por otro lado se observa el resultado del contraste de homogeneidad -> Como las razones son bastante distintas en los estratos y hay evidencia para rechazar la hipótesis nula, se considera que se trata de una variable modificadora de efecto -> No es necesario pasar a estudiar la confusión, y los resultados que se proporcionan son los del análisis en cada estrato.

Vamos a hacer ahora uno de los ejemplos incluido en la ayuda de *Epidat*.

Se estudia la infección hospitalaria posquirúrgica en pacientes operados de la cadera. El resultado se mide a través de la variable INFEC (INFEC=1 cuando el paciente se infecta a

lo largo de la primera semana, $INFEC=0$ si no se infecta). Se desea evaluar un nuevo régimen técnico-organizativo de la atención de enfermería que se dispensa a estos pacientes. Se define la variable *RÉGIMEN*, de naturaleza dicotómica, que vale 1 si el sujeto estuvo ingresado bajo el nuevo régimen y 0 en caso de que haya estado atendido bajo el régimen convencional. Imagínese que se han estudiado 80 pacientes de diferentes edades, 36 de los cuales se han ubicado en el régimen convencional y 44 en el régimen en estudio, y que los resultados son los que se recogen en la siguiente tabla.

Régimen	Infección		OR=0,30
	Sí (1)	No (0)	
Nuevo (1)	7	37	
Convencional (0)	14	22	

Considérese, además, que se quiere evaluar si la edad del paciente (se nombrará *EDAD* a esta variable) constituye una variable de confusión en la relación que pudiera existir entre el régimen organizativo y el hecho de desarrollar una infección.

Está claro que la variable *EDAD* cumple con los tres criterios convencionalmente admitidos para ser considerada como variable de confusión. Primero, el riesgo de infección aumenta con la edad. Segundo la proporción de pacientes mayores de 40 años es mayor en el grupo que recibió el régimen de atención convencional. Por último, es inverosímil creer que el efecto protector del régimen de intervención sobre el hecho de desarrollar una infección se produzca a través de la edad.

Para valorarlo [debía poner *evaluarlo*], los datos se dividen en dos categorías de edad (menores e iguales o mayores de 40 años). En este caso, se codifica la variable del modo siguiente: $EDAD=1$ si el sujeto es menor de 40 años y $EDAD=2$ si no lo es, lo que produce la configuración que recoge la siguiente tabla.

		Infección		OR ₁ =0,41
		Sí (1)	No (0)	
Edad<40 (1)	Régimen nuevo (1)	2	22	
	Régimen convencional (0)	2	9	
Edad≥40 (2)	Régimen nuevo (1)	5	15	OR ₂ =0,36
	Régimen convencional (0)	12	13	

Hacemos el análisis estratificado siguiendo los menús:

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

e introducimos los datos como se indica en la siguiente figura

Tablas de contingencia: Tablas 2x2 estratificadas

Origen de datos | Resultados

Tipo de estudio

Transversal

Cohortes

Caso-control

Nivel de confianza (%)

Número de estratos

Estrato	Expuestos		No expuestos	
	Enfermos (a)	Sanos (b)	Enfermos (c)	Sanos (d)
1	2	22	2	9
2	5	15	12	13

Los resultados de este análisis son

Tipo de estudio : Transversal
 Número de estratos: 2
 Nivel de confianza: 95,0%

Tabla global

	Enfermos	Sanos	Total
Expuestos	7	37	44
No expuestos	14	22	36
Total	21	59	80

RAZÓN DE PREVALENCIAS DE ENFERMEDAD (RP)

Estrato	RP	IC (95,0%)	
1	0,458333	0,073860	2,844155 (Katz)
2	0,520833	0,220003	1,233018 (Katz)
Cruda	0,409091	0,185072	0,904273 (Katz)
Combinada (M-H)	0,508049	0,233023	1,107675
Ponderada	0,508839	0,233416	1,109250

Prueba de homogeneidad			
	Ji-cuadrado	gl	Valor p
Combinada (M-H)	0,0154	1	0,9012
Ponderada	0,0154	1	0,9012

RAZÓN DE PREVALENCIAS DE EXPOSICIÓN (RP)

Estrato	RP	IC (95,0%)		
1	0,704545	0,257763	1,925743	(Katz)
2	0,549020	0,243466	1,238049	(Katz)
Cruda	0,531532	0,281380	1,004071	(Katz)
Combinada (M-H)	0,596818	0,315939	1,127406	
Ponderada	0,605924	0,321976	1,140284	

Prueba de homogeneidad			
	Ji-cuadrado	gl	Valor p
Combinada (M-H)	0,1451	1	0,7033
Ponderada	0,1429	1	0,7054

ODDS RATIO (OR)

Estrato	OR	IC (95,0%)		
1	0,409091	0,049706	3,366928	(Woolf)
2	0,361111	0,100339	1,299602	(Woolf)
Cruda	0,297297	0,104080	0,849210	(Woolf)
Combinada (M-H)	0,372585	0,124791	1,112414	
Ponderada	0,373463	0,125006	1,115749	

Prueba de homogeneidad			
	Ji-cuadrado	gl	Valor p
Combinada (M-H)	0,0098	1	0,9210
Ponderada	0,0098	1	0,9210

PRUEBA DE ASOCIACIÓN DE MANTEL-HAENSZEL

Ji-cuadrado	gl	Valor p
3,1470	1	0,0761

Como vemos, las razones de puntos por estratos toman valores cercanos y el contraste de homogeneidad refleja que los datos dan un apoyo grande a la hipótesis nula, por lo que no podemos considerar que la variable es modificadora de efecto y pasamos a evaluar su sesgo de confusión, según la fórmula (tomada de los apuntes de la asignatura):

$$\left(\frac{OR_{MH}}{OR_{crudo}} - 1\right) \cdot 100 = \left(\frac{0,372585}{0,297297} - 1\right) \cdot 100 = 25.32417$$

Como el sesgo de confusión es superior al 10%, podemos afirmar que es una variable de confusión.

Observación: Por la definición de la variable, que recoge el resultado de un proceso semanal, podrían haberse analizado como si se tratase de un estudio de cohortes, en vez de un estudio transversal; entonces habría que considerar la razón de tasas de incidencia en lugar de la razón de puntos.

Ejercicio 2

En este ejercicio vamos a practicar los ejemplos de regresión logística que se incluyen en la ayuda de *Epidat*, e insistiremos en algunas características de esta regresión comparando los resultados de este ejercicio con los del anterior.

Entre los propósitos de muchas investigaciones epidemiológicas se halla el establecimiento de las leyes que rigen los fenómenos que se examinan. El examen se realiza típicamente en un marco complejo, donde la coexistencia de factores mutuamente relacionados determina el comportamiento de otros. Para sondear o incluso desentrañar la naturaleza de tales relaciones, el investigador puede auxiliarse, entre otras alternativas, del análisis de regresión. La regresión logística (RL) forma parte del conjunto de métodos estadísticos que caen bajo tal denominación y es la variante que corresponde al caso en que se valora la contribución de diferentes factores en la ocurrencia de un evento simple.

En general, la RL es adecuada cuando la variable de respuesta Y es politómica (admite varias categorías de respuesta, tales como MEJORA MUCHO, MEJORA, SE MANTIENE IGUAL, EMPEORA, EMPEORA MUCHO), pero es especialmente útil en particular cuando sólo hay dos posibles respuestas (cuando la variable de respuesta es dicotómica), que es el caso más común. Tal es el caso, por ejemplo, de las siguientes situaciones: el paciente muere o sobrevive en las primeras 48 horas de su ingreso, el organismo acepta o no un trasplante, se produjo o no un intento suicida antes de los 60 años, etc.) y lo que se quiere es construir un modelo que exprese la probabilidad de ocurrencia del evento de que se trate en función de un conjunto de variables independientes. Y se codifica como 1 (si se produce cierto desenlace) y como 0 en caso opuesto, de modo que la RL expresa $P(Y=1)$ en función de ciertas variables relevantes a los efectos del problema que se haya planteado. La finalidad con que se construye ese modelo no es única: básicamente, puede tratarse de un mero esfuerzo descriptivo de cierto proceso, puede hacerse en el contexto de la búsqueda de explicaciones causales o puede desearse la construcción de un modelo para la predicción.

La RL es una de las técnicas estadístico-inferenciales más empleadas en la producción científica contemporánea.

El modelo logístico. El problema que resuelve la regresión logística es el de expresar la probabilidad de cierto desenlace ($Y=1$) en función de r variables X_1, X_2, \dots, X_r . Concretamente, lo que hace el programa es hallar los coeficientes β_i que mejor se ajustan a la siguiente representación funcional:

$$P(Y = 1) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_1 - \dots - \beta_r X_r)}$$

donde $\exp(\cdot)$ representa la función exponencial.

Una idea que merece la pena resaltar sobre la regresión logística es la utilización de las variables ficticias (*dummy*) donde se tengan variables cualitativas nominales:

Variables ficticias (*dummy*). Las variables explicativas de tipo nominal con más de dos categorías deben ser incluidas en el modelo definiendo variables *dummy*. Epidat 3.1 permite indicar que una variable independiente sea tratada de este modo y, en tal caso, construye automáticamente las *dummy* correspondientes.

Brevemente dicho, el sentido de las variables *dummy* es el siguiente: supóngase que cierta variable es nominal (raza, religión profesada, grupo sanguíneo, etc.) y consta de k categorías; deben crearse entonces k-1 variables dicotómicas que son las llamadas variables *dummy* asociadas a esta variable nominal. Se denotarán por Z_1, Z_2, \dots, Z_{k-1} . A cada categoría o clase de la variable nominal le corresponde un conjunto de valores de los Z_i con el cual se identifica dicha clase.

La manera más usual de definir estas k-1 variables es la siguiente: si el sujeto pertenece a la primera categoría, entonces las k-1 variables *dummy* valen 0: se tiene $Z_1 = Z_2 = \dots = Z_{k-1} = 0$; si el sujeto se halla en la segunda categoría, entonces $Z_1 = 1$ y las restantes valen 0; Z_2 vale 1 solo para aquellos individuos que están en la tercera categoría, en cuyo caso las otras variables asumen el valor 0, y así sucesivamente hasta llegar a última categoría, para la cual Z_{k-1} es la única que vale 1.

Por ejemplo, si la variable nominal de interés es el grupo sanguíneo, la cual tiene k=4 categorías (sangre tipo O, tipo A, tipo B y tipo AB); entonces se tendrían los siguientes valores de las variables *dummy* para cada grupo sanguíneo:

Variable nominal (grupo sanguíneo)	Z_1	Z_2	Z_3
O	0	0	0
A	1	0	0
B	0	1	0
AB	0	0	1

Repitamos el análisis de los datos del ejercicio anterior pero ahora aplicando regresión logística.

Al usar este submódulo hay que teclear los datos de una tabla de contingencia de 3 entradas con 8 celdas, o prepararla en EXCEL, Dbase o ACCESS para que el programa la lea automáticamente según la siguiente estructura:

INFEC	REGIMEN	EDAD	FREQ
0	0	1	9
0	0	2	13
0	1	1	22
0	1	2	15
1	0	1	2
1	0	2	12
1	1	1	2
1	1	2	5

El archivo CADERA.xls que se incluye en Epidat 3.1 contiene la tabla arriba expuesta. Al emplear el programa, el usuario puede elegir cuántas y cuáles variables independientes incorporar al modelo. A continuación se exponen los resultados que se obtienen cuando se pone una sola variable (RÉGIMEN), y luego los que se producen cuando se añade la variable EDAD.

Parece que algunas versiones de *Epidat* dan un error al intentar cargar ese archivo. Una solución es crear otro archivo de un formato del que también puedan importarse los datos; abriendo el archivo con *Excel* y guardándolo con el formato de *Dbase* (4) se soluciona el problema, puesto que *Epidat* sí lo carga.

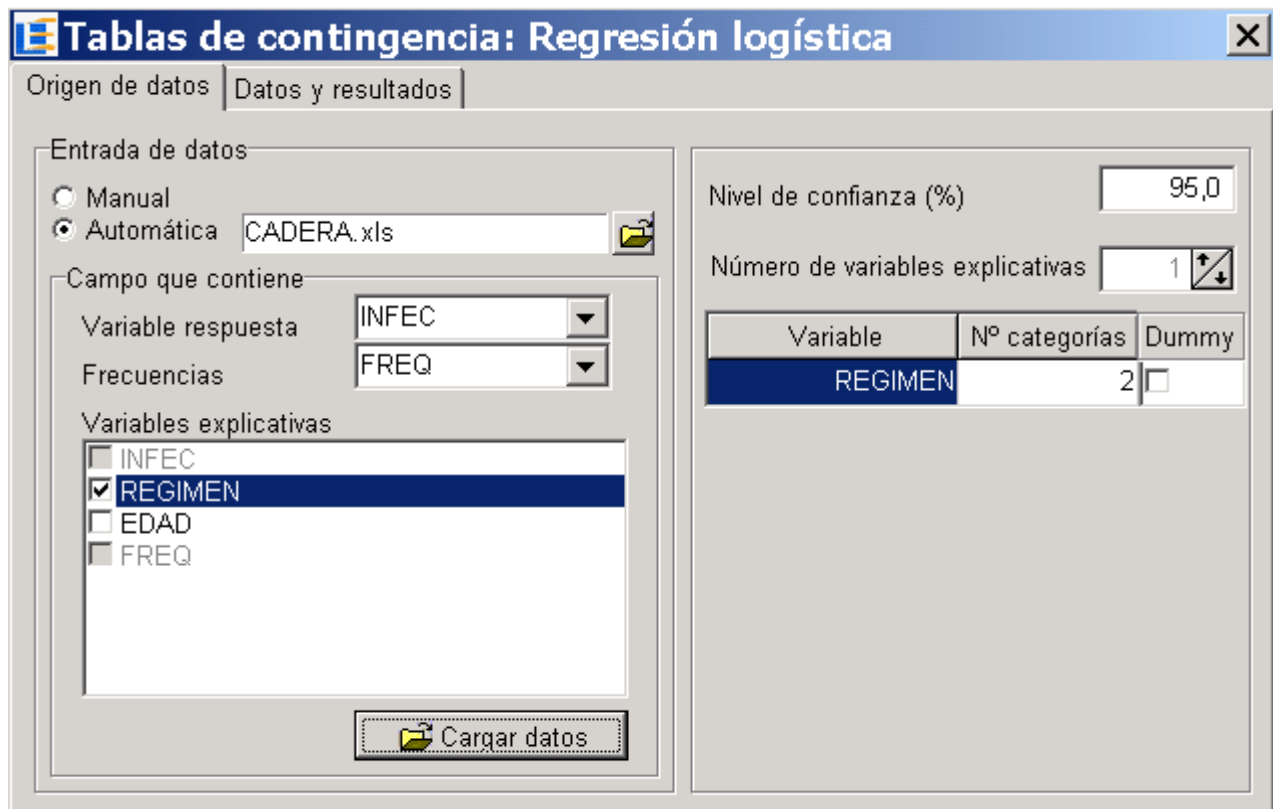
Hacemos el análisis siguiendo los menús:

Métodos --> Tablas de contingencia --> Regresión logística

e introducimos los datos de forma automática (desde el archivo que los contiene), en

C:\Archivos de programa\Epidat 3.1\Ejemplos\Tablas de contingencia

Marcamos primero sólo RÉGIMEN como variable independiente



Al pulsar el botón *Cargar datos* y el de hacer los cálculos, obtenemos:

```
Archivo de trabajo: C:\Archivos de programa\Epidat 3.1\Ejemplos\Tablas de
contingencia\CADERA.xls
Campo que contiene:
Variable respuesta: INFEC
Frecuencias: FREQ
Variables explicativas: RÉGIMEN

Nivel de confianza: 95,0%
```

Variable respuesta:

Valor N° sujetos

0	59
1	21

Total	80

La sucesión de estimadores ha convergido

N° iteraciones necesarias 3

-2 ln Verosimilitud inicial: 92,104901

-2 ln Verosimilitud final : 86,671970

Cociente de verosimilitud

Estadístico	gl	Valor p
-----	-----	-----
5,4329	1	0,0198

Coefficiente de determinación: 0,0675

Variable	Coefficiente	EE	Valor de Z	Valor p
-----	-----	-----	-----	-----
Constante	-0,451473			
RÉGIMEN	-1,210425	0,535158	-2,261807	0,0237

Variable	Odds ratio	IC (95,0%)	
-----	-----	-----	-----
RÉGIMEN	0,298071	0,104422	0,850838

PRUEBA DE BONDAD DE AJUSTE DE HOSMER Y LEMESHOW

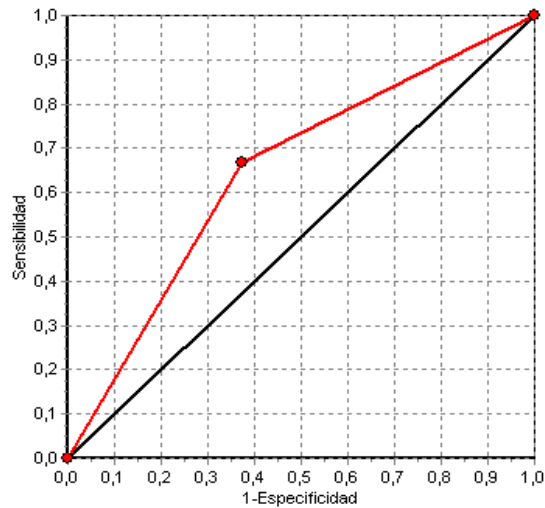
Grupos basados en los deciles

Grupo de Probabilidad	Respuesta = 0	Respuesta = 1	
	Valor observado	Valor observado	Valor esperado
-----	-----	-----	-----
1	37	7	7,02
2	22	14	14,00

Ji-cuadrado	gl	Valor p
-----	-----	-----
0,0001	0	NAN

Compara los resultados encerrados en las elipses con los de la razón de puntos cruda de los resultados de la página 5.

Curva ROC



Área ROC	EE	IC (95%)		
0,6469	0,0615	0,5263	0,7675	DeLong
	0,0731	0,5036	0,7902	Hanley & McNeil

Si ahora se repite la regresión incluyendo además la variable EDAD como independiente (además de RÉGIMEN)



Se tiene:

Archivo de trabajo: C:\Archivos de programa\Epidat 3.1\Ejemplos\Tablas de contingencia\CADERA.xls

Campo que contiene:

Variable respuesta: INFECC

Frecuencias: FREC

Variables explicativas: RÉGIMEN EDAD

Nivel de confianza: 95,0%

Variable respuesta:

Valor N° sujetos

0	59
1	21

Total	80

La sucesión de estimadores ha convergido

N° iteraciones necesarias 3

-2 ln Verosimilitud inicial: 92,104901

-2 ln Verosimilitud final : 81,324308

Cociente de verosimilitud

Estadístico	gl	Valor p
-----	-----	-----
10,7806	2	0,0046

Coefficiente de determinación: 0,1308

Variable	Coefficiente	EE	Valor de Z	Valor p
-----	-----	-----	-----	-----
Constante	-2,759493			
RÉGIMEN	-0,974758	0,554901	-1,756635	0,0790
EDAD	1,332184	0,622533	2,139941	0,0324

Variable	Odds ratio	IC (95,0%)	
-----	-----	-----	-----
RÉGIMEN	0,377284	0,127156	1,119438
EDAD	3,789310	1,118560	12,836931

PRUEBA DE BONDAD DE AJUSTE DE HOSMER Y LEMESHOW

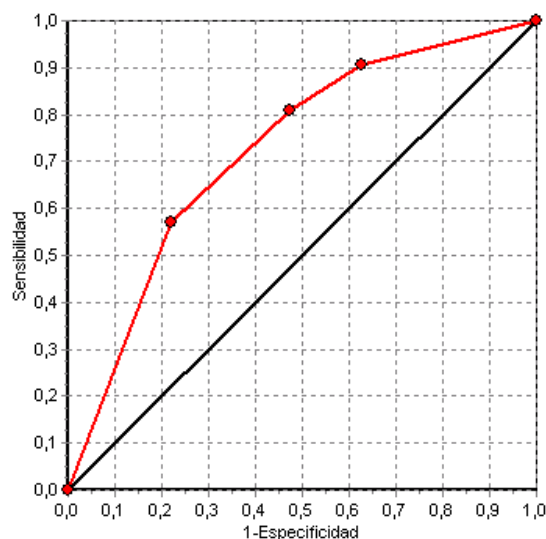
Grupos basados en los deciles

Grupo de Probabilidad	Respuesta = 0		Respuesta = 1	
	Valor observado	Valor esperado	Valor observado	Valor esperado
-----	-----	-----	-----	-----
1	22	22,01	2	1,99
2	9	8,87	2	2,13
3	15	14,89	5	5,11
4	13	13,09	12	11,91

Compara los resultados encerrados en las elipses con los de la razón de puntos cruda de los resultados de la página 5.

Ji-cuadrado	gl	Valor p
0,0142	2	0,9929

Curva ROC



Área ROC	EE	IC (95%)		
0,7244	0,0608	0,6052	0,8436	DeLong
	0,0692	0,5888	0,8599	Hanley & McNeil

Obsérvese que ahora el área encerrada bajo la curva ROC es mayor que en el análisis sin la variable EDAD, lo que demuestra que ha mejorado el poder de predicción de la regresión.

Varias cosas que merece la pena resaltar sobre la regresión logística son:

a) El análisis de la RL suple al análisis estratificado. Nótese que, en el caso de los pacientes operados de la cadera, el odds ratio (0,298) coincide con la razón de productos cruzados correspondiente a la tabla. El intervalo de confianza que produce la RL [0,10 ; 0,85] es también coincidente con el que se obtiene mediante el análisis no paramétrico que arroja el análisis hecho a través de tablas de 2x2 incluido en otro submódulo del presente módulo. Por otra parte, el OR=0,377 que se obtiene a través del exponencial del coeficiente que corresponde a RÉGIMEN en el modelo que incluye las dos variables independientes, no es otra cosa que la estimación de Mantel-Haenszel (lo mismo ocurre con el intervalo de confianza).

b) La valoración [debía poner *evaluación*] sobre el posible papel confusor de un factor se desarrolla de manera ágil. Basta correr el modelo con y sin el factor y comparar los coeficientes de la variable independiente. En el ejemplo de los operados de la cadera, se compara (puede utilizarse la fórmula que se incluye al principio de la página 6) 0,298 con 0,377 lo cual permite pensar que sí hay efecto confusor. El OR correspondiente a RÉGIMEN tiene, en el primer caso, un intervalo de confianza que no contiene al 1 (significativo al nivel 0,05) mientras que el que se obtiene cuando se controla la edad sí lo contiene (pierde la significación).

c) El ajuste suele ser bueno. El resultado que se ha obtenido en estos ejemplos, donde los valores esperados y observados son muy parecidos, es típico.

d) Si el contexto del problema es predictivo, la probabilidad del suceso para un perfil de entrada dado ha de computarse independientemente empleando los coeficientes estimados.

Y algunas recomendaciones que también se dan en la ayuda de *Epidat* son:

- Las variables explicativas deben tener una relación monótona con la probabilidad del evento que se estudia.
- Las variables independientes involucradas en el modelo no deben estar correlacionadas entre sí. Si la correlación entre dos variables es alta, entonces los resultados de la RL son poco confiables. Concretamente, los errores estándares se incrementan apreciablemente y suele ocurrir que los coeficientes no son significativamente diferentes de cero, aunque la aportación global de las variables sí lo sea.
- Debe recordarse que el conjunto de variables *dummy* constituye un todo indisoluble con el cual se suple a una variable nominal. Cualquier decisión que se adopte o valoración que se haga concierne al conjunto íntegro.
- Es muy importante distinguir entre un contexto explicativo y un contexto predictivo. En el primer caso, el modelo para cada posible factor de riesgo o protector se ajusta con los factores que pueden ser confusores para él. Solo en los estudios predictivos se ajusta el mejor modelo. Debe tenerse en cuenta, en este caso, que una variable puede tener valor predictivo aunque no sea parte del mecanismo causal que produce el fenómeno en estudio.



Bioestadística

Práctica 5: Estudio del sesgo de autoselección

Introducción

Para estudiar un ejemplo del **efecto de un tipo de sesgo sobre los resultados de un estudio epidemiológico**, podemos hacer el siguiente ejercicio de simulación. En concreto, vamos a estudiar el sesgo de autoselección, es decir, el que se produce en la población del estudio cuando dejamos que los sujetos puedan presentarse voluntarios a formar parte de la muestra. Vamos a hacer las siguientes suposiciones:

- Consideraremos un *estudio epidemiológico transversal*, y estudiaremos las *prevalencias*, las *razones de prevalencias* y las *razones de puntos* (este último es el nombre que prefiere la [Real Academia de Ciencias Exactas, Física y Naturales](#) para traducir «odds ratios». Un «odds ratio» es un cociente entre la probabilidad de un suceso y la de su complementario. Por ejemplo, la razón de puntos de enfermedad será la probabilidad de enfermedad dividida por la de no enfermedad).
- En una población general hay siempre un tanto por ciento de personas que se prestarían voluntarias para formar parte de la muestra del estudio. Su proporción sería la *probabilidad de que, eligiendo una persona al azar, fuese una persona de las que se hubiese autoseleccionado*. Si nosotros elegimos la muestra totalmente al azar, habrá en ella esta misma proporción de personas que se hubiesen seleccionado voluntariamente. Vamos a estudiar qué pasaría si fuésemos dejando gradualmente libertad a los sujetos para formar parte de la muestra. Una observación interesante a la hora de hacer la simulación es que *desde el punto de vista de los investigadores del estudio, esto equivale a ir aumentando la probabilidad mencionada anteriormente (desde el valor real hasta uno), manteniendo la condición de que ellos siguiesen eligiendo la muestra totalmente al azar*. Obviamente los resultados de las medidas para cada valor de esta probabilidad se comparan con los resultados «reales», es decir, los correspondientes al valor real de esa probabilidad.
- Suponemos que entre las personas que se hubiesen prestado voluntarias, por sus características, hay diferencias con respecto a las otras personas. Concretamente, *vamos a suponer que el grado de exposición a la causa que se está estudiando es distinto en las personas autoseleccionadas*. Esto se va a traducir en que los autoseleccionados tendrían asociada una tabla de contingencia, *A*, distinta de la de los no autoseleccionados, *B*. Más precisamente, *A* tendrá distinto número de sujetos repartidos en sus dos filas, aunque la suma total será igual, el tamaño de la población.
- Vamos a dejar constante la relación exposición-enfermedad en todo momento; es decir, movemos el número de personas que están o no expuestas a la causa (como

motivo de dejar que se autoseleccionen), pero *dejamos invariante la proporción de sujetos que, en expuestos y no expuestos, enferman o no*. Recordemos que queremos estudiar sólo el efecto de la composición de la muestra. La matriz A tendrá sus columnas con las mismas proporciones entre sí que las de la matriz B .

- Lo que los investigadores observan, en función de la proporción de sujetos autoseleccionados, es una suma ponderada entre los valores de los sujetos autoseleccionados y los no autoseleccionados. Por las propiedades de los sucesos de probabilidad, y *si vemos cada tabla como una matriz*, podríamos pensar que la tabla que los observadores anotan es:

$$M = P(A)*A + [1-P(A)]*B$$

Cálculos (con R)

Lo que vamos a hacer en los cálculos es ir cambiando el valor de $P(A)$ en la anterior ecuación y calculando las medidas epidemiológicas, que representaremos gráficamente después en función de $P(A)$.

Podríamos hacer los cálculos para algunos valores de $P(A)$ con *Epidat*, pero es mejor hacer muchos más cálculos con algún lenguaje de programación, como por ejemplo R , *Octave* o *Matlab*. El siguiente código es de R , que es un lenguaje gratuito que puedes bajar e instalar en unos minutos desde <http://www.r-project.org>. Lo que hace el código está explicado en él. Puedes ir copiándolo del archivo y pegándolo en la consola de R .

```
# Tabla/matriz para los que no se autoseleccionarían.
B <- matrix(c(58,22,62,258), nrow=2)

# Tamaño de la muestra.
N <- sum(B)

# Definimos la matriz de los que se autoseleccionarían.
# Proporción de expuestos que elegimos para la población
# de autoseleccionados.
proporcionExp <- 0.20
numExp <- ceiling(N*proporcionExp)

# Construimos la matriz de los autoseleccionados
# manteniendo constante la proporción entre las columnas
# que hay en B, pero haciendo que la primera fila sume
# numExp y la segunda N-numExp.
a11 <- ceiling(numExp*B[1,1]/sum(B[1,]))
a21 <- ceiling((400-numExp)*B[2,1]/sum(B[2,]))
A <- matrix(c(a11,a21,numExp-a11,(400-numExp)-a21),nrow=2)
```

```

# Proporción de los que se autoseleccionarían en la
# población.
proporcionA <- 0.02
n <- ceiling(N*proporcionA)

vecP_a <- (n:N)/N
numP_a <- length(vecP_a)

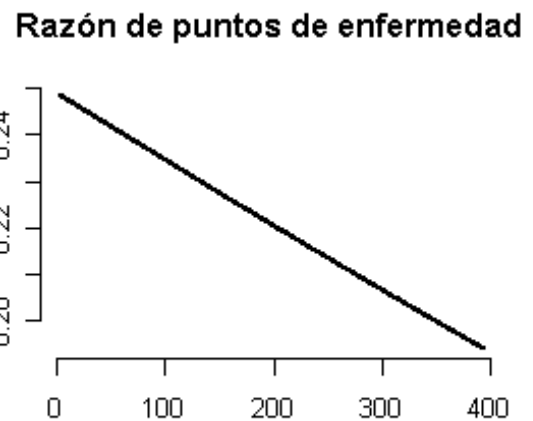
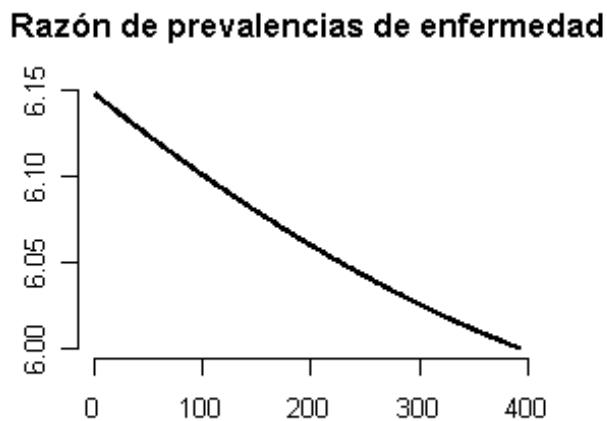
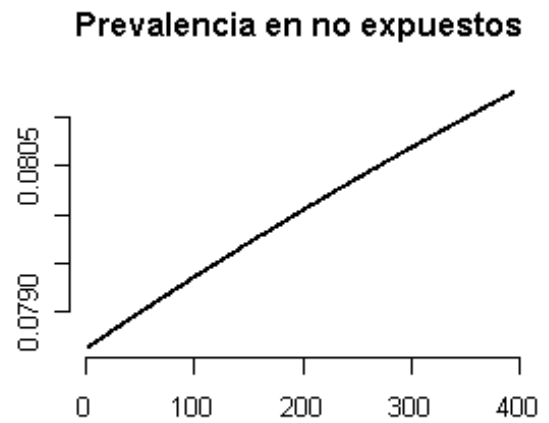
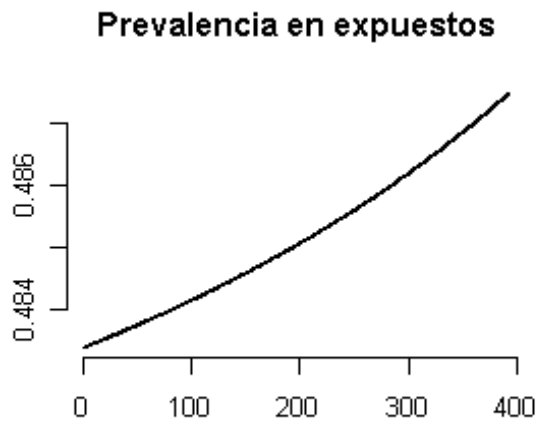
# Hacemos los cálculos necesarios. En el código de MATLAB
# (apartado siguiente) se incluye además de ésta otra forma
# más clara, pero más lenta, de hacer estos cálculos.
vec11 <- vecP_a*A[1,1]+(1-vecP_a)*B[1,1]
vec12 <- vecP_a*A[1,2]+(1-vecP_a)*B[1,2]
vec21 <- vecP_a*A[2,1]+(1-vecP_a)*B[2,1]
vec22 <- vecP_a*A[2,2]+(1-vecP_a)*B[2,2]

# Medidas por exposición (filas)
vecPrev_Exp <- vec11/(vec11+vec12)
vecPrev_nExp <- vec21/(vec21+vec22)
vecRaz_Exp <- (vec11*(vec21+vec22))/(vec21*(vec11+vec12))
vecRaz_Puntos_Exp <- (vec11+vec21)/(vec12+vec22)

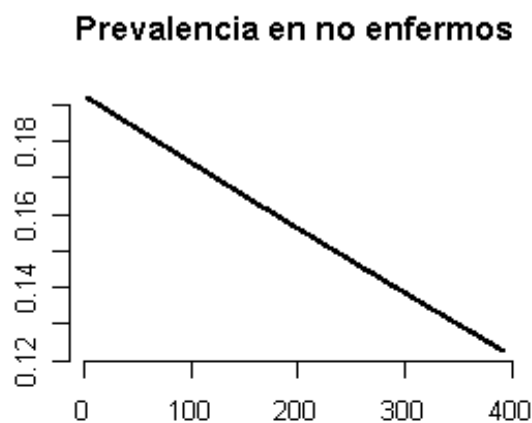
# Medidas por enfermedad (columnas)
vecPrev_Enf <- vec11/(vec11+vec21)
vecPrev_nEnf <- vec12/(vec12+vec22)
vecRaz_Enf <- (vec11*(vec12+vec22))/(vec12*(vec11+vec21))
vecRaz_Puntos_Enf <- (vec11+vec12)/(vec21+vec22)

# Hacemos los gráficos de las medidas de enfermedad
x11()
par(mfcol=c(2,2), lty=1, lwd=2, bty="n")
plot(vecPrev_Exp, xlab="", ylab="",
      main="Prevalencia en expuestos", type="l")
plot(vecRaz_Exp, xlab="", ylab="",
      main="Razón de prevalencias de enfermedad", type="l")
plot(vecPrev_nExp, xlab="", ylab="",
      main="Prevalencia en no expuestos", type="l")
plot(vecRaz_Puntos_Exp, xlab="", ylab="",
      main="Razón de puntos de enfermedad", type="l")

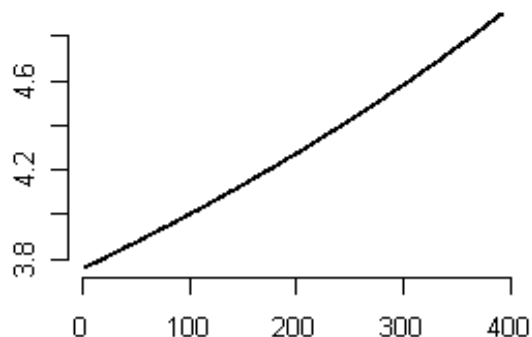
```



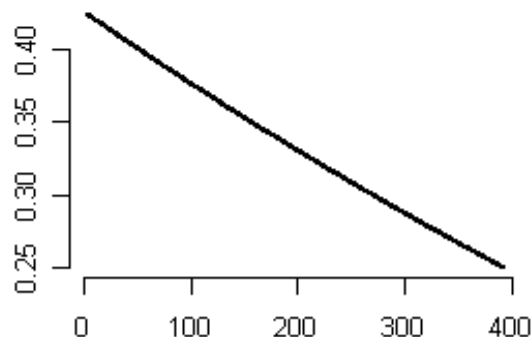
```
# Hacemos los gráficos de las medidas de exposición
x11()
par(mfcol=c(2,2), lty=1, lwd=2, bty="n")
plot(vecPrev_Enf, xlab="", ylab="",
      main="Prevalencia en enfermos", type="l")
plot(vecRaz_Enf, xlab="", ylab="",
      main="Razón de prevalencias de exposición", type="l")
plot(vecPrev_nEnf, xlab="", ylab="",
      main="Prevalencia en no enfermos", type="l")
plot(vecRaz_Puntos_Enf, xlab="", ylab="",
      main="Razón de puntos de exposición", type="l")
```



Razón de prevalencias de exposición

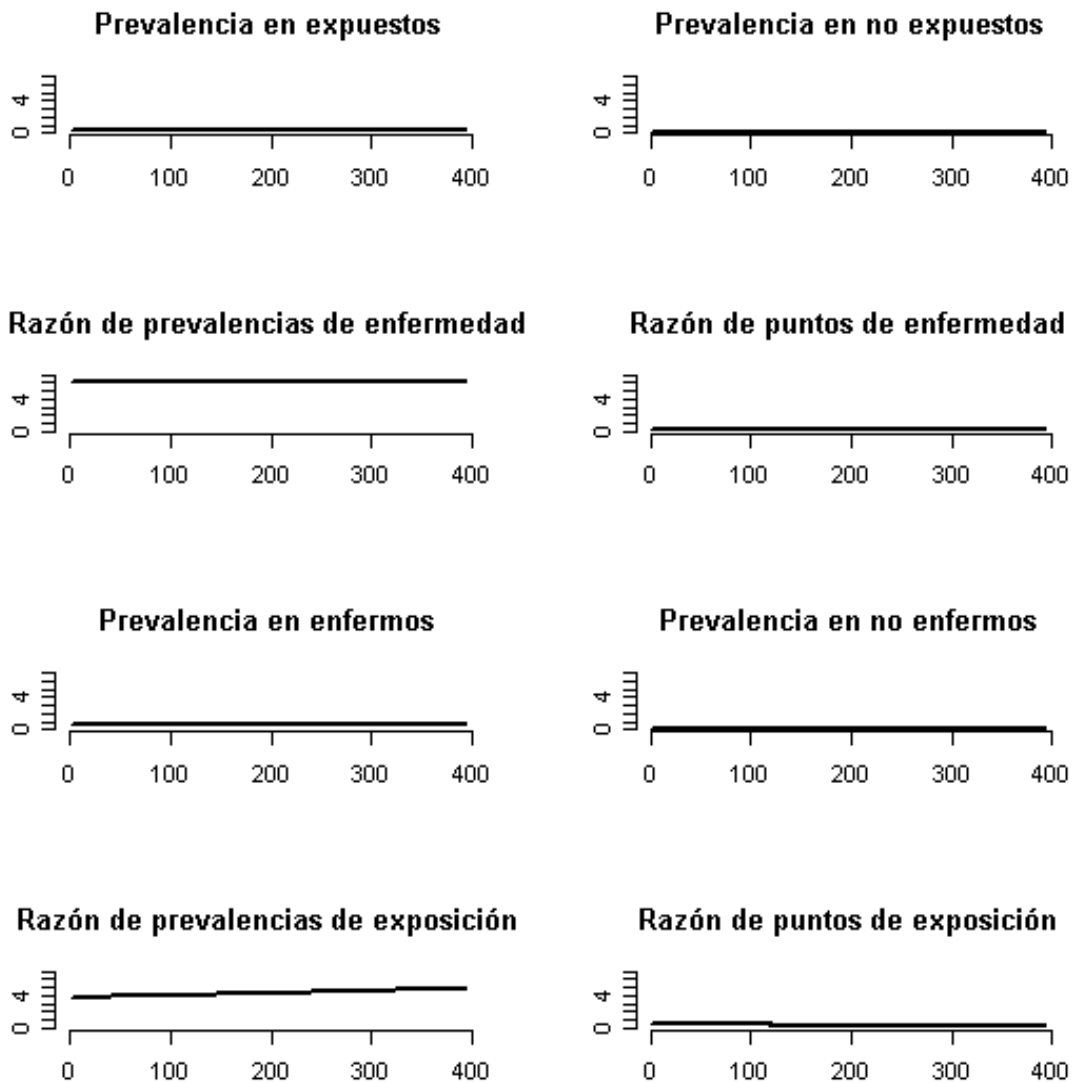


Razón de puntos de exposición



Hacemos los gráficos de todas las medidas

```
x11()
par(mfcol=c(4,2), lty=1, lwd=2, bty="n")
plot(vecPrev_Exp, ylim=c(0,7), xlab="", ylab="",
     main="Prevalencia en expuestos", type="l")
plot(vecRaz_Exp, ylim=c(0,7), xlab="", ylab="",
     main="Razón de prevalencias de enfermedad", type="l")
plot(vecPrev_Enf, ylim=c(0,7), xlab="", ylab="",
     main="Prevalencia en enfermos", type="l")
plot(vecRaz_Enf, ylim=c(0,7), xlab="", ylab="",
     main="Razón de prevalencias de exposición", type="l")
plot(vecPrev_nExp, ylim=c(0,7), xlab="", ylab="",
     main="Prevalencia en no expuestos", type="l")
plot(vecRaz_Puntos_Exp, ylim=c(0,7), xlab="", ylab="",
     main="Razón de puntos de enfermedad", type="l")
plot(vecPrev_nEnf, ylim=c(0,7), xlab="", ylab="",
     main="Prevalencia en no enfermos", type="l")
plot(vecRaz_Puntos_Enf, ylim=c(0,7), xlab="", ylab="",
     main="Razón de puntos de exposición", type="l")
```



```

# Calculamos los errores relativos (respecto a la estimación correcta,
# no respecto al valor que se está estimando con cada concepto)
erroresPrev_Exp <- abs((vecPrev_Exp-vecPrev_Exp[1])/vecPrev_Exp[1])
erroresPrev_nExp <- abs((vecPrev_nExp-vecPrev_nExp[1])/vecPrev_nExp[1])
erroresRaz_Exp <- abs((vecRaz_Exp-vecRaz_Exp[1])/vecRaz_Exp[1])
erroresRaz_Puntos_Exp <- abs((vecRaz_Puntos_Exp-vecRaz_Puntos_Exp[1])
                             /vecRaz_Puntos_Exp[1])
erroresPrev_Enf <- abs((vecPrev_Enf-vecPrev_Enf[1])/vecPrev_Enf[1])
erroresPrev_nEnf <- abs((vecPrev_nEnf-vecPrev_nEnf[1])/vecPrev_nEnf[1])
erroresRaz_Enf <- abs((vecRaz_Enf-vecRaz_Enf[1])/vecRaz_Enf[1])
erroresRaz_Puntos_Enf <- abs((vecRaz_Puntos_Enf-vecRaz_Puntos_Enf[1])
                              /vecRaz_Puntos_Enf[1])

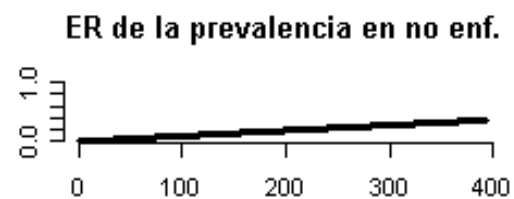
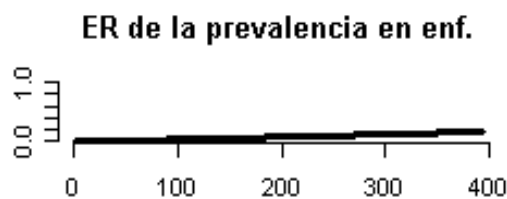
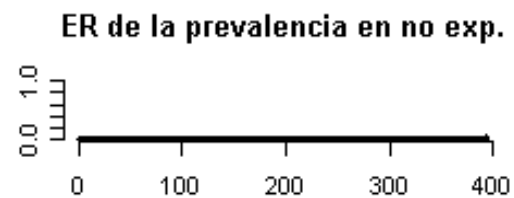
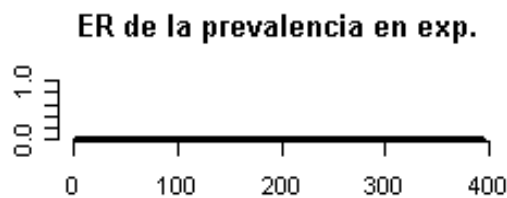
# Hacemos los gráficos de todos los errores relativos
x11()
par(mfcol=c(4,2), lty=1, lwd=3, bty="n")
plot(erroresPrev_Exp, ylim=c(0,1), xlab="", ylab="",

```

```

main="ER de la prevalencia en exp.", type="l")
plot(erroresRaz_Exp, ylim=c(0,1), xlab="", ylab="",
     main="ER de la razón de prev. de enf.", type="l")
plot(erroresPrev_Enf, ylim=c(0,1), xlab="", ylab="",
     main="ER de la prevalencia en enf.", type="l")
plot(erroresRaz_Enf, ylim=c(0,1), xlab="", ylab="",
     main="ER de la razón de prev. de exp.", type="l")
plot(erroresPrev_nExp, ylim=c(0,1), xlab="", ylab="",
     main="ER de la prevalencia en no exp.", type="l")
plot(erroresRaz_Puntos_Exp, ylim=c(0,1), xlab="", ylab="",
     main="ER de la razón de puntos de enf.", type="l")
plot(erroresPrev_nEnf, ylim=c(0,1), xlab="", ylab="",
     main="ER de la prevalencia en no enf.", type="l")
plot(erroresRaz_Puntos_Enf, ylim=c(0,1), xlab="", ylab="",
     main="ER de la razón de puntos de exp.", type="l")

```



Cálculos (con *Matlab*)

Estos cálculos son los mismos que los del apartado anterior, pero con este lenguaje de programación.

```
% Tabla/matriz para los que no se autoseleccionarían.
B = [58 62; 22 258];

% Tamaño de la muestra.
N = sum(sum(B));

% Definimos la matriz de los que se autoseleccionarían.
% Proporción de expuestos que elegimos para la población
% de autoseleccionados.
proporcionExp = 0.20;
numExp = ceil(N*proporcionExp);
% Construimos la matriz de los autoseleccionados
% manteniendo constante la proporción entre las columnas
% que hay en B, pero haciendo que la primera fila sume
% numExp y la segunda N-numExp.
a11 = ceil(numExp*B(1,1)/sum(B(1,:)));
a21 = ceil((400-numExp)*B(2,1)/sum(B(2,:)));
A = [a11 numExp-a11; a21 (400-numExp)-a21];

% Proporción de los que se autoseleccionarían en la
% población.
proporcionA = 0.02;
n = ceil(N*proporcionA);
vecP_a = (n:N)/N;
numP_a = length(vecP_a);

% -- Forma 1 de hacer los cálculos -----
% Vamos generando las observaciones perturbadas y
% calculamos las medidas epidemiológicas que nos interesan.
% Para guardar las medidas
vecPrev_Exp = zeros(1,numP_a);
vecPrev_nExp = zeros(1,numP_a);
vecRaz_Exp = zeros(1,numP_a);
vecRaz_Puntos_Exp = zeros(1,numP_a);
vecPrev_Enf = zeros(1,numP_a);
vecPrev_nEnf = zeros(1,numP_a);
vecRaz_Enf = zeros(1,numP_a);
vecRaz_Puntos_Enf = zeros(1,numP_a);
for i1 = 1:numP_a
    pseudoP_a = vecP_a(i1);
    C = [pseudoP_a*A(1,1)+(1-pseudoP_a)*B(1,1),...
        pseudoP_a*A(1,2)+(1-pseudoP_a)*B(1,2);...
        pseudoP_a*A(2,1)+(1-pseudoP_a)*B(2,1),...
        pseudoP_a*A(2,2)+(1-pseudoP_a)*B(2,2)];
    % Medidas por exposición (filas)
    vecPrev_Exp(i1) = C(1,1)/sum(C(1,:));
```

```

vecPrev_nExp(i1) = C(2,1)/sum(C(2,:));
vecRaz_Exp(i1) = (C(1,1)*sum(C(2,:)))/(C(2,1)*sum(C(1,:)));
vecRaz_Puntos_Exp(i1) = sum(C(:,1))/sum(C(:,2));
% Medidas por enfermedad (columnas)
vecPrev_Enf(i1) = C(1,1)/sum(C(:,1));
vecPrev_nEnf(i1) = C(1,2)/sum(C(:,2));
vecRaz_Enf(i1) = (C(1,1)*sum(C(:,2)))/(C(1,2)*sum(C(:,1)));
vecRaz_Puntos_Enf(i1) = sum(C(1,:))/sum(C(2,:));
end
% -----

% -- Forma 2 de hacer los cálculos -----
% Otra forma más elegante y rápida (pero oscura) de hacer
% los cálculos anteriores es la siguiente, que aprovecha
% la capacidad de trabajar con vectores del lenguaje,
% evitando los lentos bucles «for».
vec11 = vecP_a*A(1,1)+(1-vecP_a)*B(1,1);
vec12 = vecP_a*A(1,2)+(1-vecP_a)*B(1,2);
vec21 = vecP_a*A(2,1)+(1-vecP_a)*B(2,1);
vec22 = vecP_a*A(2,2)+(1-vecP_a)*B(2,2);
% Medidas por exposición (filas)
vecPrev_Exp = vec11./(vec11+vec12);
vecPrev_nExp = vec21./(vec21+vec22);
vecRaz_Exp = (vec11.*(vec21+vec22))./(vec21.*(vec11+vec12));
vecRaz_Puntos_Exp = (vec11+vec21)./(vec12+vec22);
% Medidas por enfermedad (columnas)
vecPrev_Enf = vec11./(vec11+vec21);
vecPrev_nEnf = vec12./(vec12+vec22);
vecRaz_Enf = (vec11.*(vec12+vec22))./(vec12.*(vec11+vec21));
vecRaz_Puntos_Enf = (vec11+vec12)./(vec21+vec22);
% -----

% Hacemos los gráficos de las medidas de enfermedad
figure()
subplot(2,2,1)
plot(vecPrev_Exp)
title('Prevalencia en expuestos')
subplot(2,2,2)
plot(vecPrev_nExp)
title('Prevalencia en no expuestos')
subplot(2,2,3)
plot(vecRaz_Exp)
title('Razón de prevalencias de enfermedad')
subplot(2,2,4)
plot(vecRaz_Puntos_Exp)
title('Razón de puntos de enfermedad')

% Hacemos los gráficos de las medidas de exposición
figure()
subplot(2,2,1)
plot(vecPrev_Enf)
title('Prevalencia en enfermos')

```

```

subplot(2,2,2)
plot(vecPrev_nEnf)
title('Prevalencia en no enfermos')
subplot(2,2,3)
plot(vecRaz_Enf)
title('Razón de prevalencias de exposición')
subplot(2,2,4)
plot(vecRaz_Puntos_Enf)
title('Razón de puntos de exposición')

% Hacemos los gráficos de todas las medidas
figure()
subplot(4,2,1)
plot(vecPrev_Exp)
axis([0 numP_a 0 7])
title('Prevalencia en expuestos')
subplot(4,2,2)
plot(vecPrev_nExp)
axis([0 numP_a 0 7])
title('Prevalencia en no expuestos')
subplot(4,2,3)
plot(vecRaz_Exp)
axis([0 numP_a 0 7])
title('Razón de prevalencias de enfermedad')
subplot(4,2,4)
plot(vecRaz_Puntos_Exp)
axis([0 numP_a 0 7])
title('Razón de puntos de enfermedad')
subplot(4,2,5)
plot(vecPrev_Enf)
axis([0 numP_a 0 7])
title('Prevalencia en enfermos')
subplot(4,2,6)
plot(vecPrev_nEnf)
axis([0 numP_a 0 7])
title('Prevalencia en no enfermos')
subplot(4,2,7)
plot(vecRaz_Enf)
axis([0 numP_a 0 7])
title('Razón de prevalencias de exposición')
subplot(4,2,8)
plot(vecRaz_Puntos_Enf)
axis([0 numP_a 0 7])
title('Razón de puntos de exposición')

% Calculamos los errores relativos
erroresPrev_Exp = ...
    abs((vecPrev_Exp-vecPrev_Exp(1))./vecPrev_Exp(1));
erroresPrev_nExp = abs((vecPrev_nExp-vecPrev_nExp(1))./vecPrev_nExp(1));
erroresRaz_Exp = abs((vecRaz_Exp-vecRaz_Exp(1))./vecRaz_Exp(1));
erroresRaz_Puntos_Exp = ...
    abs((vecRaz_Puntos_Exp-vecRaz_Puntos_Exp(1))./vecRaz_Puntos_Exp(1));
erroresPrev_Enf = abs((vecPrev_Enf-vecPrev_Enf(1))./vecPrev_Enf(1));
erroresPrev_nEnf = abs((vecPrev_nEnf-vecPrev_nEnf(1))./vecPrev_nEnf(1));

```

```

erroresRaz_Enf = abs((vecRaz_Enf-vecRaz_Enf(1))./vecRaz_Enf(1));
erroresRaz_Puntos_Enf =...
    abs((vecRaz_Puntos_Enf-vecRaz_Puntos_Enf(1))./vecRaz_Puntos_Enf(1));

% Hacemos los gráficos de todos los errores relativos
figure()
subplot(4,2,1)
plot(erroresPrev_Exp)
axis([0 numP_a 0 1])
title('ER de la prevalencia en exp.')
subplot(4,2,2)
plot(erroresPrev_nExp)
axis([0 numP_a 0 1])
title('ER de la prevalencia en no exp.')
subplot(4,2,3)
plot(erroresRaz_Exp)
axis([0 numP_a 0 1])
title('ER de la razón de prev. de enf.')
subplot(4,2,4)
plot(erroresRaz_Puntos_Exp)
axis([0 numP_a 0 1])
title('ER de la razón de puntos de enf.')
subplot(4,2,5)
plot(erroresPrev_Enf)
axis([0 numP_a 0 1])
title('ER de la prevalencia en enf.')
subplot(4,2,6)
plot(erroresPrev_nEnf)
axis([0 numP_a 0 1])
title('ER de la prevalencia en no enf.')
subplot(4,2,7)
plot(erroresRaz_Enf)
axis([0 numP_a 0 1])
title('ER de la razón de prev. de exp.')
subplot(4,2,8)
plot(erroresRaz_Puntos_Enf)
axis([0 numP_a 0 1])
title('ER de la razón de puntos de exp.')

```

Interpretación

Salvo algún error en mi planteamiento del problema o en la implementación, una posible forma de interpretar estos resultados es la siguiente: A primera vista, después de observar los dos primeros gráficos, podríamos pensar que hay una variación no lineal en las medidas epidemiológicas debida al efecto del sesgo, y que esta variación es de magnitud similar en las medidas de enfermedad que en las de exposición. Sin embargo, haciendo cuentas algebraicas es fácil demostrar rigurosamente qué cantidades deben variar en nuestro experimento y cuáles no. No vamos a hacer tales demostraciones aquí, pero sí vamos a razonar sobre lo que sucede.

Si las dos matrices A y B se diferencian en el número de sujetos que suman en cada fila, pero —dentro de cada fila— se conserva la proporción entre las columnas (enfermos y no enfermos), sólo deben cambiar las cantidades que se miden por columnas, puesto que ha variado la distribución proporcional de sujetos en expuestos y no expuestos; las cantidades que se miden por filas no deben cambiar, puesto que se conserva la proporción de sujetos en sus celdas. Otra cosa distinta es que haya otros efectos que causen una ligera tendencia a que la estimación mejore o empeore; por ejemplo, por el mero hecho de que hay más o menos sujetos para estimar.

De hecho, cuando nos damos cuenta de que los valores del eje vertical no están en el mismo rango en las dos primeras gráficas, repetimos los gráficos con un mismo rango. Lo que se observa ahora es lo que la teoría preveía. Las prevalencias de enfermedad, sus razones y sus errores se mantienen estables, mientras que las prevalencias de exposición, sus razones y sus errores se ven afectadas en la dirección esperada: las prevalencias de exposición disminuyen ligeramente y su razón aumenta, los errores relativos de las cantidades que se miden por columnas (prevalencias de exposición y sus razones) tienen un aumento no lineal. Por otro lado, también debemos aprender indirectamente de este ejercicio la importancia de tener el cuidado de hacer con un mismo patrón las comparaciones entre cantidades: en este caso, la importancia de hacer comparables entre sí todos los gráficos.



Universidad Complutense de Madrid

└ Facultad de Ciencias Económicas y Empresariales

└ Departamento de Estadística e Investigación Operativa II

└ David Casado de Lucas

15 de febrero del 2012