TABLAS DE CONTINGENCIA

INDICE

TABLAS DE CONTINGENCIA

CONCEPTOS GENERALES

Analizar la distribución de una variable con relación a otra u otras es una tarea corriente en Salud Pública, vinculada, la mayoría de las veces, a la búsqueda de un patrón que indique la relación, (o la falta de ella) entre las variables estudiadas. Este es un proceso clave en la identificación de las posibles causas de los problemas de salud, y también de factores que, aun cuando no puedan ser finalmente considerados causales, resulten estar asociados a estos daños y constituyan importantes elementos prácticos para la identificación de grupos con mayores riesgos de padecer determinado daño.

El estudio de la influencia de una variable (variable independiente) sobre la forma en que se modifica otra (variable dependiente) es conocido como análisis bivariado; y será multivariado cuando el estudio evalúe de forma simultánea el efecto sobre una variable dependiente de dos o más variables independientes.

Las tablas de contingencia (tablas de doble entrada) son una herramienta fundamental para este tipo de análisis. Están compuestas por filas (horizontales), para la información de una variable y columnas (verticales) para la información de otra variable. Estas filas y columnas delimitan *celdas* donde se vuelcan las frecuencias de cada combinación de las variables analizadas. En su expresión más elemental, las tablas tienen solo 2 filas y 2 columnas (tablas de 2x2); en Epidat 3.1 estas tablas presentan la siguiente estructura tipo:

	Enfermedad		
Factor de riesgo	**Sí**	**No**	**Total**
Expuestos	**...**	...	...
No expuestos	...	...	...
Total	...	...	...

donde las filas identifican el nivel de exposición a la variable en estudio y las columnas la categoría en relación a la enfermedad (y las personas-tiempo en el caso de utilizar tasas de incidencia).

En general, las tablas pueden abarcar varias filas (M) y columnas (N). El análisis puede ocasionalmente involucrar más variables; por ejemplo, puede considerarse una tercera variable, cada una de cuyas clases dé lugar a una tabla de MxN.

En muchos análisis subyacen dos hipótesis. Un ejemplo típico es el caso en que se tiene una que afirma la existencia de cierta asociación entre las variables estudiadas. Ocasionalmente, por su estructura teórica, el problema encarado permite hablar de una variable dependiente y de variables independientes. Otras veces, sin embargo, el examen de la asociación no está previamente direccionado. En cualquier caso, la otra hipótesis afirma que no existe tal relación y que ambas variables tienen completa independencia (hipótesis nula). Salvo en situaciones muy especiales, la simple inspección de la información contenida en las tablas de contingencia no permite ser concluyente sobre cuál de las dos hipótesis es la que ha de elegirse como válida.

Para examinar este problema, un primer paso puede ser calcular la frecuencia relativa de cada celda (las medidas de frecuencia pueden ser diversas: prevalencia, incidencia acumulada, densidad de incidencia, odds, etc.).

Sin embargo, el análisis de la relación entre las variables estudiadas es más directo cuando se computan medidas de asociación. Estas medidas, basadas en la comparación entre las frecuencias del daño en diferentes grupos, pueden realizarse a través de razones (razón de prevalencias, riesgo relativo, *odds ratio*) o de sus diferencias (riesgo atribuible y fracción atribuible). Finalmente, para evaluar la posibilidad de que los resultados observados sean solo producto del azar, la estadística clásica aporta una serie de métodos y pruebas que permiten pronunciarse al respecto. Dichas pruebas computan la probabilidad de haber obtenido los datos empíricamente observados, calculada bajo el supuesto de que la hipótesis de nulidad es correcta (la cual se denota como "p"). En general, la mayoría de los investigadores trabajan con un *nivel de significación* del 5% (equivalentemente, con un nivel de confianza del 95%), por lo que aceptan que existe asociación entre las variables estudiadas cuando el valor de p es menor que 0,05.

Tanto las medidas de efecto como las pruebas estadísticas a utilizar, dependerán del diseño del estudio del que proceden los datos, del tipo de variables y categorías consideradas y de que se haya considerado o no más de un estrato (una tercera variable).

Las pruebas de significación estadística que acompañan el análisis basan su examen en comparar los resultados observados con los esperados (bajo el supuesto de que no hay asociación). Cuanto mayor sea la diferencia entre la distribución observada y la esperada, menos razonable es suponer que la distribución observada sea solo producto del azar.

El cálculo de los valores esperados se realiza usando los valores marginales de la tabla, asumiendo que la probabilidad para cada categoría es la misma que la de todo el grupo y que no existe asociación entre las variables estudiadas.

Por ejemplo, en una situación en la que se conoce que hay 300 individuos y que los valores marginales son, por una parte, que 100 están enfermos y 200 sanos y, por otra parte, que 60 de los 300 estuvieron expuestos a cierto factor de riesgo y 240 no, entonces los valores esperados para cada celda pueden calcularse obteniendo el producto de los dos valores marginales de la celda y dividiéndolo por el gran total. Por ejemplo, el número esperado de enfermos entre los expuestos será de 60x100/300 = 20, y el de no enfermos no expuestos: 240x200/300 = 160. El resultado de ese cómputo de valores esperados para las celdas se muestra en la tabla siguiente.

	Enfermedad		
Factor de riesgo	**Sí**	**No**	**Total**
Expuestos	20	40	**60**
No expuestos	80	160	**240**
Total	**100**	**200**	**300**

En la práctica, en las tablas de 2x2 solo sería necesario calcular el valor esperado de una celda, porque los valores de las restantes se podrán deducir del que ésta asuma y de los valores marginales. En el ejemplo, si se espera que haya 20 enfermos expuestos, los 80 enfermos restantes serán no expuestos. Y como de los 60 expuestos solo 20 están enfermos, los restantes 40

serán sanos. Así, el valor de la celda correspondiente a los no enfermos no expuestos no puede ser otro que 160 para completar los 200 no enfermos.

Esta dependencia e independencia de las celdas se conoce como *grados de libertad* y, como se vio, en las tablas 2x2 solo hay un grado de libertad. El cálculo de los grados de libertad resulta de multiplicar el número de columnas menos 1 por el número de filas menos 1:

Grados de libertad = (nº de columnas–1) x (nº de filas–1)

Las pruebas de independencia basadas en las discrepancias entre frecuencias observadas y esperadas sólo son válidas en el caso de muestras grandes. Si la muestra es pequeña, se recomienda utilizar métodos exactos, como la prueba de Fisher, que calcula la probabilidad exacta de obtener los resultados observados si las dos variables son independientes y los totales marginales son fijos.

Finalmente, es importante considerar que para poder realizar estos cálculos, las categorías de las variables deberán ser excluyentes y exhaustivas. Es decir, deberá evitarse que su definición permita que algunas observaciones puedan pertenecer a 2 ó más categorías así como que, por el contrario, algunas observaciones no sean incluibles en categoría alguna.

Este módulo está integrado por 4 submódulos:

Tablas de asociación: exposición–enfermedad

- Tablas de 2x2 (simples y estratificadas)
- Tablas de 2xN (simples y estratificadas)

Tablas generales

- Tablas de MxN
- Regresión logística

Epidat 3.1 permite una entrada manual de los datos en las tablas 2x2, 2xN y MxN, y calculará las medidas de frecuencia, de asociación o efecto y las pruebas específicas para cada diseño de estudio, tipo de variables y estructura de la tabla.

En el caso de tablas 2x2, Epidat 3.1 permite, de forma opcional, sumar 0,5 a todas las frecuencias de la tabla en el caso de que alguna de ellas sea igual a cero. Si no se activa esta opción, el programa sólo presentará aquellos resultados susceptibles de ser computados.

Para el cálculo de la regresión logística, los datos podrán ser ingresados en forma manual o desde un archivo.

TABLAS DE ASOCIACIÓN: EXPOSICIÓN–ENFERMEDAD

TABLAS 2X2 SIMPLES

Las tablas 2x2 simples (de un único estrato) permiten el análisis de 2 variables dicotómicas: típicamente, una variable independiente (exposición) y una variable dependiente (enfermedad). Debe advertirse que esta es la situación más común y que es por ello que se usan las denominaciones exposición y enfermedad, pero podría tratarse de otra situación como la de un

ensayo clínico, por ejemplo, en la cual, en lugar de dos niveles de exposición tuviéramos dos tratamientos y en lugar de enfermedad tuviéramos dos posibles desenlaces.

Hay cuatro opciones de tablas 2x2 destinadas a cuatro diseños de estudios epidemiológicos:

- Estudio transversal
- Estudio de cohortes
  - Para tasas de incidencia
  - Para incidencia acumulada
- Estudio de casos y controles
- Estudio de casos y controles emparejados

Estudio transversal

Los estudios transversales examinan la prevalencia de enfermedades y problemas de salud y también de conocidos o potenciales factores de riesgo y/o protección. Se tratan básicamente de una imagen "fotográfica" de la población, o de una muestra de ella, en la que se explora, a nivel individual y en forma simultánea, la presencia o ausencia (o niveles) de una o más variables independientes y de una o más variables potencialmente dependientes de las primeras. Si bien la imagen de una fotografía da la sensación de que en un estudio de este tipo la información se recolecta en un "instante" (un día o pocos días), la recolección de datos puede ser más prolongada (semanas o meses). Sin embargo, la información de cada individuo seguirá siendo "una foto".

Entre sus mayores ventajas están, en general, su bajo costo, su relativa facilidad de ejecución y la posibilidad de obtener estimaciones puntuales de las prevalencias de varias enfermedades e información de varios factores potencialmente determinantes en un mismo momento. Entre sus mayores limitaciones están la dificultad (y con frecuencia, imposibilidad) para establecer la relación temporal entre lo que serían las exposiciones y los daños, la limitación para obtener incidencias y la vulnerabilidad a diferentes tipos de sesgos.

El formato que se usará de la tabla para el análisis bivariado de variables dicotómicas de los estudios transversales presentará a la variable independiente (exposición) en las filas y la variable dependiente (enfermedad o daño) en las columnas:

Factor de riesgo o factor de protección	**Enfermedad o daño**		
	Sí	**No**	**Total**
Expuestos	a	b	a + b
No expuestos	c	d	c + d
Total	a + c	b + d	a + b + c + d

Las salidas previstas en Epidat 3.1 son:

- Medidas de frecuencia
  - Prevalencia de enfermedad en expuestos y no expuestos.

    - Prevalencia de exposición en enfermos y no enfermos.
- Medidas de asociación
    - Razón de prevalencias de exposición e intervalo de confianza (Katz).
    - Razón de prevalencias de enfermedad e intervalo de confianza (Katz).
    - Odds ratio e intervalo de confianza (Woolf y Cornfield).
- Medidas de significación estadística
    - Test Ji-cuadrado de asociación, con y sin corrección.
    - Prueba exacto de Fisher y valor de p unilateral y bilateral.

Ejemplo

Un estudio transversal para conocer la prevalencia de osteoporosis y su relación con algunos factores de riesgo potenciales incluyó a 400 mujeres con edades entre 50 y 54 años. A cada una se le realizó una densitometría de columna y en cada caso se completó un cuestionario de antecedentes.

Para el ejemplo se consideran solo las variables dicotómicas osteoporosis y antecedentes de dieta pobre en calcio. De las 80 pacientes que presentaban osteoporosis 58 presentaban antecedentes de dieta pobre en calcio, en tanto que entre las 320 que no tenían osteoporosis, el número de mujeres con este antecedente era de 62.

Una vez ingresados estos datos, la tabla se presentará de la siguiente manera:

Antecedente de dieta pobre en calcio	**Osteoporosis**		
	Sí	**No**	**Total**
Expuestos	58	62	120
No expuestos	22	258	280
Total	80	320	400

Y los resultados que proporciona Epidat 3.1 serán:

```
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

Prevalencia en expuestos y no expuestos. Por tratarse de estudios transversales, las frecuencias del daño se presentan como tasas de prevalencia puntualmente estimadas. Estas tasas miden el número de personas que presentaban el daño en el momento del estudio en cada grupo (expuestos y no expuestos) en comparación con el total de la población en ambos grupos.

Si la información recolectada en el estudio transversal registrase la ocurrencia de una determinada enfermedad o daño durante un período determinado (por ejemplo, se ha indagado: ¿Ha presentado al menos un episodio convulsivo en los últimos 6 meses?), los datos obtenidos han de interpretarse como incidencias o riesgos. Nótese que en tal caso el estudio es transversal porque la pregunta se formula en el momento actual, pero de hecho es una pregunta que, por su naturaleza, contempla la precedencia temporal de los acontecimientos.

En el ejemplo, la prevalencia en los expuestos fue de 48,3% (58 de las 120 mujeres con antecedentes de dieta pobre en calcio tenían osteoporosis), en tanto entre los no expuestos la prevalencia fue de 7,9% (22 de 280).

Razón de prevalencias. Esta razón permite comparar la prevalencia de expuestos con la de los no expuestos. Cuando la prevalencia en expuestos es más alta, la razón es superior a 1 y esto estaría indicando que la exposición aumenta el riesgo de tener ese daño. Si la razón es menor que 1, estaría indicando lo opuesto (sería un factor de protección). Si fuera igual a 1, entonces la prevalencia en ambos es similar, lo que sugeriría que la exposición no está relacionada con el daño.

El intervalo de confianza de la razón de prevalencias representa un recorrido de valores dentro del cual hay una determinada confianza (generalmente 95%) de que se encuentre el verdadero valor de la razón.

El resultado del ejemplo muestra que la razón de las prevalencias fue de 6,15 (IC 95%: [3,96 ; 9,57]), indicando que existiría una asociación entre el antecedente y el daño. El IC 95% sugiere que el verdadero valor estaría dentro de ese recorrido. Un enfoque a veces empleado sugiere examinar si el valor 1 se halla dentro del intervalo o no y hacer de hecho por esa vía una prueba de significación. En este caso, como el extremo inferior del intervalo está por arriba de 1, se rechazaría la hipótesis de que no hay asociación y de que la que se ha observado sea solo producto del azar.

Medidas de significación estadística. Finalmente, los resultados presentan las medidas de significación estadística que se resumen en el valor de p, la probabilidad de haber hallado estos resultados en el supuesto que no hay asociación. Valores de p menores a 0,05 implican que tal probabilidad es menor al 5%. Este valor está relacionado con la amplitud y posición del intervalo de confianza. Cuando ambos extremos del intervalo de confianza estén a uno u otro lado de 1, el valor de p será menor que 0,05, en tanto que si un extremo está por debajo de 1 y el otro por arriba, p será superior a 0,05. Pese a esta estrecha relación entre ambos enfoques, para una mejor representación del fenómeno estudiado es recomendable considerar el intervalo de confianza, que resulta más informativo.

Estudio de cohortes

Los estudios de cohortes sustentan su estrategia de análisis en el seguimiento en el tiempo de dos o más grupos de individuos que han sido divididos según el grado de exposición a un determinado factor (corrientemente en 2 grupos: expuestos y no expuestos).

Al inicio, ninguno de los individuos incluidos en ambos grupos tiene la enfermedad o daño en estudio y para responder a la pregunta acerca de si la exposición influye en el desenlace habrá de compararse la incidencia de "nuevos casos" entre ambos grupos. Estas incidencias pueden ser calculadas de dos formas:

- Como número de casos nuevos en relación a la población que integra la cohorte (incidencia acumulada);
- Considerando el período que cada individuo permaneció en el grupo (tasa de incidencia o densidad de incidencia).

La incidencia acumulada es más sencilla de calcular porque como denominador solo se requiere el número de individuos que se incluyó en cada grupo. Sin embargo, la tasa de incidencia es una medida más precisa, ya que considera el momento en que se producen los casos y los períodos de seguimiento de los individuos, que tipícamente no son iguales para todos los sujetos. Por ejemplo, si el daño en un grupo aparece más tempranamente que en otro, aunque al final del período ambos grupos hayan acumulado igual número de casos, la incidencia acumulada en ambos grupos será la misma, pero la tasa de incidencia en el grupo donde los casos fueron más prematuros será más elevada. De forma similar, si se decide seguir a un grupo durante 4 años y algunos individuos abandonan el estudio al finalizar el año 2, el "peso" de estos individuos en el denominador debería ser la mitad del de aquellos individuos que sí permanecieron bajo observación los 4 años, ya que los restantes pudieron adquirir la enfermedad en los años 3 y 4.

La mayor ventaja de este tipo de estudios es su capacidad analítica para aceptar o rechazar hipótesis. Permiten estudiar incidencias y están poco expuestos a los sesgos de selección. Su mayor eficiencia se logra cuando se valoran exposiciones *raras*, que no podrían ser estudiadas con otro tipo de estudios, y para dolencias de cortos períodos entre el comienzo de la exposición y la aparición de la enfermedad. Su mayor limitación viene dada por sus costos, en general muy elevados, en especial para el estudio de daños poco frecuentes o de largos períodos de latencia.

El formato de la tabla para el análisis de los estudios de cohorte es similar a las otras tablas de contingencia, solo que para el cálculo de las tasas de incidencia se considerará el período, (personas–tiempo):

Factor de riesgo o factor de protección	**Enfermedad o daño**	**Personas–tiempo**
Expuestos	A	T_1
No expuestos	C	T_0
Total	A+C	$T=T_1+T_0$

Los resultados que Epidat 3.1 brinda son:

- Medidas de frecuencia
  - Riesgo en expuestos y no expuestos (incidencia acumulada) o,
  - Tasa de incidencia en expuestos y no expuestos (densidad de incidencia).
- Medidas de asociación
  - Riesgo relativo o razón de las tasas de incidencia e intervalo de confianza (Katz).
  - Diferencia de riesgos o diferencia de tasas de incidencia e intervalo de confianza.
  - Fracción atribuible o prevenible para la población expuesta e intervalo de confianza.
  - Odds ratio e intervalo de confianza (Woolf y Cornfield), para incidencia acumulada.
- Medidas de impacto
  - Fracción atribuible o prevenible para la población.
- Medidas de significación estadística

  Para incidencia acumulada:
  - Prueba Ji-cuadrado de asociación, con y sin corrección.
  - Prueba exacta de Fisher y valor de p, unilateral y bilateral.

  Para tasa de incidencia:
  - Prueba de asociación.

Ejemplo

Para evaluar el efecto de la exposición a asbesto sobre el riesgo de fallecer por cáncer de pulmón, un estudio comparó un grupo de 6.245 trabajadores expuestos a este agente con otro grupo de 7.895 trabajadores sin exposición a este factor.

A lo largo de 22 años de seguimiento, en el primer grupo se presentaron 76 defunciones por cáncer en el aparato respiratorio, en tanto que en el grupo no expuesto el número de defunciones por esta causa fue 28. El tiempo total de seguimiento del grupo expuesto fue de 116.157 personas-año, mientras que en el segundo grupo fue de 177.636.

Ingresados estos datos en las tablas se tendrá:

Tabla para incidencia acumulada

	Defunción por cáncer		
Exposición a asbesto	**Sí**	**No**	**Total**
Expuestos	76	6.169	6.245
No expuestos	28	7.867	7.895
Total	104	14.036	14.140

Tabla para tasas de incidencia

Exposición a asbesto	Defunciones	Personas-año
Expuestos	76	116.157
No expuestos	28	177.636
Total	104	293.793

Y los resultados para incidencia acumulada serán:

```
Tablas de contingencia : Tablas 2x2 simples

Tipo de estudio   : Cohortes
Tipo de datos     : Incidencia acumulada
Nivel de confianza: 95,0%
Tabla
             Enfermos     Sanos     Total
------------ -------- -------- --------
Expuestos          76     6169      6245
No expuestos       28     7867      7895
------------ -------- -------- --------
Total             104    14036     14140

                                    Estimación           IC(95,0%)
---------------------------------   ----------    --------- ---------
Riesgo en expuestos                   0,012170        -         -
Riesgo en no expuestos                0,003547        -         -
Riesgo relativo                       3,431431     2,227679  5,285644 (Katz)
Diferencia de riesgos                 0,008623     0,005604  0,011642
Odds ratio                            3,461385     2,241395  5,345416
    (Woolf)
                                                   2,248792  5,327744
(Cornfield)
```

Fracción atribuible en expuestos	0,708576	0,551102	0,810808
Fracción atribuible poblacional	0,517806	0,338412	0,648556

Prueba Ji-cuadrado de asociación	Estadístico	Valor p
Sin corrección	35,5135	0,0000
Corrección de Yates	34,3422	0,0000

Para tasas de incidencia se tiene lo siguiente:

Tablas de contingencia : Tablas 2x2 simples

Tipo de estudio : Cohortes
Tipo de datos : Tasa de incidencia
Nivel de confianza: 95,0%

Tabla

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

	Estimación	IC(95,0%)	
Tasa de incidencia en expuestos	0,000654	-	-
Tasa de incidencia en no expuestos	0,000158	-	-
Razón de tasas de incidencia	4,150889	2,691321	6,402016
Diferencia de tasas de incidencia	0,000497	0,000338	0,000655
Fracción atribuible en expuestos	0,759088	0,628435	0,843799
Fracción atribuible poblacional	0,554718	0,388850	0,675569

Prueba de asociación

Estadístico Z	Valor p
6,8954	0,0000

Riesgo en expuestos y no expuestos. El riesgo se cuantifica mediante la incidencia acumulada y se presenta como riesgo absoluto para el total del período estudiado. El riesgo de 0,01217 de los expuestos puede entenderse como una incidencia de 1,2% ó 12,17 por mil (durante todo el período). Este riesgo, considerablemente más alto que el de los no expuestos (3,55 por mil), señala que la exposición al asbesto sí estaría causando un mayor riesgo de enfermar y morir por cáncer del aparato respiratorio.

Tasa de incidencia en expuestos y no expuestos. Las tasas de incidencia, al considerar el tiempo real de seguimiento, corrigen los errores que se pueden introducir por diferencias en el tiempo de seguimiento entre los grupos. De hecho, en el ejemplo presentado el tiempo promedio de seguimiento de los expuestos (18,6 años) fue menor que el de los no expuestos (22,5 años).

La tasa de incidencia en expuestos de 0,000654 señala que la incidencia anual en este grupo fue de 0,654 por mil.

Riesgo relativo y razón de tasas de incidencia. Tienen una interpretación similar a la razón de prevalencias. Señalan la relación entre las incidencias de ambos grupos. El riesgo relativo de 3,43 indica que en los expuestos la incidencia es 3,43 veces la de los no expuestos, lo que también puede leerse como que en los expuestos hay 2,43 veces más riesgo que en los no expuestos. El valor obtenido para el *odds ratio* (3,46) es muy próximo al del riesgo relativo por tratarse de una enfermedad rara.

La razón de las tasas de incidencia resultó mayor (4,15). Esta diferencia es consecuencia del diferente tiempo en promedio de seguimiento en uno y otro grupo, y pone de manifiesto la conveniencia de considerar esta medida y no solo el riesgo relativo.

Fracción atribuible o prevenible entre los expuestos. Representa la fracción del daño que podría ser evitada entre los expuestos si se eliminara enteramente esa exposición. Este tipo de análisis asume causalidad. Esto significa que, efectivamente, la exposición es un factor responsable del exceso de daño en el grupo de expuestos y que, por lo tanto, si no hubiera existido tal exposición, esa fracción de sujetos que padecen el daño no hubiera ocurrido.

La fracción atribuible es aplicable a un análisis de tipo prospectivo. Responde a la pregunta ¿cuánto daño se podrá evitar si esta población no se expusiera en absoluto a tal factor? Pero cuando la exposición de hecho existe y se pretende estimar la reducción del daño al eliminar la exposición, esto es solo aplicable en caso de que la exposición sea totalmente reversible.

Por ejemplo, si un grupo de personas tiene un exceso de riesgo por no usar cinturón de seguridad al conducir, y se elimina la exposición (todos comienzan a usar cinturón), mediante la fracción atribuible se podrá estimar el monto relativo del daño que se evitará. Sin embargo, esto no podrá aplicarse al hábito de fumar porque se trata de una exposición no reversible en un 100% (el riesgo de los nunca fumadores no es similar al de los ex–fumadores). En cualquier caso, este indicador puede tener una virtualidad teórica en la medida que cuantifica, supuestamente, el peso etiológico de determinado factor en términos de la salud pública.

En el caso del ejemplo, un 70,8% (IC 95%: [55,1% ; 81,1%]) de los casos de cáncer de pulmón entre los trabajadores expuestos, podría ser atribuido al asbesto.

Fracción atribuible o prevenible en la población. Esta es una medida del impacto potencial que tendría la eliminación de una exposición en toda la población. Representa la fracción del daño total de enfermos que podría ser evitada y, como en el caso anterior, se asume causalidad y solo es aplicable para exposiciones totalmente reversibles, o para la construcción de escenarios prospectivos.

Siguiendo con el ejemplo, una fracción atribuible poblacional de 0,518 significa que un 51,8% de los casos de cáncer respiratorio en la población de trabajadores, podría atribuirse a la exposición a asbestos y por ende evitarse si tal exposición fuera enteramente suprimida.

Las medidas de significación estadística tienen una interpretación similar a la de las tablas para estudios transversales.

Estudio de casos y controles

En los estudios de casos y controles los sujetos incluidos proceden típicamente de dos grupos, según sean casos (con la enfermedad o daño en estudio) o controles (sin el daño en cuestión). Este tipo de diseño hizo su aparición a mediados del siglo XX cuando, en los países desarrollados, el interés de la Salud Pública comenzó a centrarse en las enfermedades crónicas. La idea básica es comparar los antecedentes de los "enfermos" de una población con los de los "sanos" de la misma población. Se trata de poner de manifiesto posibles diferencias en las exposiciones que expliquen, al menos parcialmente, la razón por la que unos enfermaron y otros no.

En el análisis se comparan las exposiciones de los casos con las de los controles, y los resultados son presentados usando los llamados *odds* (cociente entre la probabilidad de enfermar y la probabilidad de no enfermar) y la razón de odds de adquirir una enfermedad entre expuestos y entre no expuestos (*odds ratio*, OR).

	Casos	**Controles**	**Total**
Expuestos	a	b	a + b
No expuestos	c	d	c + d
Total	a + c	b + d	a + b + c + d

El *odds ratio* estimado (OR = (axd)/(cxb)) será mayor cuanto mayor sea el número de casos expuestos y el de controles no expuestos, y menor cuanto mayor sea el número de casos no expuestos y el de controles expuestos.

El número de controles por cada caso puede diferir entre un estudio y otro, pero en general oscila entre uno y tres; a lo sumo, se toman cuatro controles por cada caso. No tiene mayor interés tomar más de cuatro controles por caso, ya que la potencia de la prueba no crece de manera apreciable, mientras que sí lo hacen los costos. Por esta razón, excepto que se cuente con los datos a un bajo costo, superar los 4 controles por caso no es recomendable. Por otro lado, cuando existe un gran número de casos, y quizás dificultades para obtener controles, es posible también diseñar un estudio donde la relación caso/control sea 2 a 1 ó 3 a 1.

Entre las principales ventajas de este tipo de diseño comparado con los estudios de seguimiento está su eficiencia en términos de costo y tiempo, en especial para enfermedades poco comunes y/o de largos períodos de incubación. Esta eficiencia deriva del hecho de que, una vez diagnosticada la enfermedad o el evento, solo es necesario incluir en el estudio un número relativamente pequeño de casos, y en especial de controles. Esto lo diferencia significativamente de los estudios de seguimiento donde, por ejemplo en las enfermedades raras, deberá *seguirse* la evolución de una enorme cantidad de individuos para obtener unos pocos casos.

Otra ventaja de los estudios de casos y controles, comparados con los de seguimiento, es la posibilidad de estudiar varias exposiciones en forma simultánea.

La mayores desventajas de los estudios de casos y controles son, por un lado, el *sesgo de selección* que pueda introducirse al elegir los controles y, por otro, el hecho de que a la hora de la inclusión de los individuos en el estudio, tanto las exposiciones como el daño ya han ocurrido. Esto dificulta establecer la precisión y la similitud de criterio con que exposiciones y daños han sido medidos en los participantes. Existe incluso el potencial problema que se presenta en los estudios transversales, donde la secuencia exposición-daño podría no conocerse adecuadamente

e incluso estar invertida en algunos casos (esto es, que la exposición se haya modificado como consecuencia del daño, o de estadios subclínicos de la dolencia) sin que el investigador tenga la posibilidad siquiera de enterarse.

Esto hace que este tipo de estudios esté particularmente expuesto a errores de clasificación, tanto en la evaluación de las exposiciones como en la de los resultados.

Entre estos errores es importante destacar el *sesgo del recuerdo*, que surge de un recuerdo "modificado" en los casos respecto de los controles y el sesgo en la recolección de los datos (*sesgo del observador*) inducido por el hecho de que el observador realiza un esfuerzo diferente a la hora de evaluar cada sujeto en dependencia de que sea un caso o un control.

Los estudios de casos y controles no permiten estimar directamente las medidas de riesgo dentro de cada grupo, ya que la proporción de enfermos en el grupo de expuestos y en el de no expuestos dependerá de la decisión del investigador en cuanto al número de casos y de controles involucrados en el estudio. Dicho de otra manera, la muestra típicamente no es representativa de la población en cuanto a la proporción enfermos/no enfermos y ello cancela la posibilidad de estimar adecuadamente las tasas de enfermos entre expuestos y de enfermos entre los que están libres de la exposición.

Con el objetivo de "controlar" diferentes factores de confusión posibles, tales como edad, género, consumo de tabaco, etc., es corriente que los casos y los controles sean emparejados según estas características. Cuando esto se realiza durante el análisis, los datos pueden ser procesados como si este emparejamiento no se hubiera realizado o, por el contrario, a través de una tabla especial que busca comparar las diferencias entre estos "pares".

En un estudio de casos y controles, Epidat 3.1 presenta los siguientes resultados:

- Medidas de frecuencia
  - Proporción de casos expuestos.
  - Proporción de controles expuestos.
- Medidas de asociación
  - Odds ratio e intervalo de confianza (Woolf y Cornfield), para incidencia acumulada.
  - Fracción atribuible o prevenible para la población expuesta e intervalo de confianza.
- Medidas de impacto
  - Fracción atribuible o prevenible para la población.
- Medidas de significación estadística

  Para datos no emparejados:
  - Prueba Ji-cuadrado de asociación, con y sin corrección.
  - Prueba exacta de Fisher y valor de p, unilateral y bilateral.

  Para datos emparejados:
  - Prueba de asociación de McNemar.

Ejemplo

Con el objetivo de investigar si la lactancia materna constituye un factor de protección para el cáncer de mama, un estudio incluyó a 755 mujeres menores de 36 años de 11 regiones sanitarias del Reino Unido, a las que se les diagnosticó cáncer de mama durante el período 1982 a 1985. Para cada caso se eligió un control al azar de la lista de pacientes del mismo médico general que asistía al caso. Estos controles debían tener una diferencia de edad con los casos menor a 6 meses. Cada caso y su correspondiente control fueron entrevistados por el mismo encuestador. Los resultaron mostraron que en el grupo de casos, 255 mujeres realizaron una lactancia plena de al menos 3 meses, mientras que entre los controles este antecedente estaba presente en 487 mujeres (de los 255 controles de los casos que tuvieron una lactancia plena, 160 lactaron y 95 no, en tanto de los 500 controles de los casos que no lactaron, 327 si lo habían hecho y 173 no).

Ingresados estos datos en las tablas, los datos quedan resumidos del modo siguiente:

Casos y controles

	Casos	Controles	Total
Exp.	255	487	742
No exp.	500	268	768
Total	755	755	1.510

Casos y controles emparejados

	Controles		
Casos	**Exp.**	**No exp.**	**Total**
Exp.	160	95	255
No exp.	327	173	500
Total	487	268	755

Nota: adviértase de que se está llamando "exposición" a la práctica de lactancia materna; obviamente, esto constituye cierto abuso de lenguaje, pero no dará lugar a confusión siempre que el usuario comprenda que en este caso la persona ha estado "expuesta" (o no) a un factor presuntamente protector.

Los resultados de la tabla de contingencia para casos y controles serán:

```
Tablas de contingencia : Tablas 2x2 simples

Tipo de estudio   : Caso-control
Nivel de confianza: 95,0%

Tabla
                Casos  Controles    Total
------------ -------- -------- --------
Expuestos         255       487       742
No expuestos      500       268       768
------------ -------- -------- --------
Total             755       755      1510

                                    Estimación           IC(95,0%)
---------------------------------- ----------    --------- ---------
Proporción de casos expuestos        0,337748        -         -
Proporción de controles expuestos    0,645033        -         -
Odds ratio                           0,280657     0,227028  0,346954
    (Woolf)
                                                  0,227051  0,346920
(Cornfield)
---------------------------------- ----------    --------- ---------
```

```
Fracción prevenida en expuestos        0,719343     0,653046  0,772972
Fracción prevenida poblacional         0,392876     0,323084  0,455473


Prueba Ji-cuadrado de asociación          Estadístico     Valor p
----------------------------------------   ---------    ---------
Sin corrección                              142,6224       0,0000
Corrección de Yates                         141,3956       0,0000


Prueba exacta de Fisher                     Valor p
----------------------------------------  ----------
Unilateral                                   0,0000
Bilateral                                    0,0000
```

En el análisis emparejado los resultados que se presentan son:

```
Tablas de contingencia : Tablas 2x2 simples

Tipo de estudio          : Caso-control emparejado
Nº de controles por caso :   1
Nivel de confianza       : 95,0%

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


                                          Estimación           IC(95,0%)
----------------------------------------  ----------    --------- ---------
Proporción de casos expuestos               0,337748        -         -
Proporción de controles expuestos           0,645033        -         -
Odds ratio                                  0,290520     0,220645  0,381744
----------------------------------------  ----------    --------- ---------


Prueba de asociación de McNemar
Ji-cuadrado           Valor p
--------------------   ---------
            126,4479      0,0000
```

Estos resultados presentan la proporción de casos y controles con antecedentes de exposición y el *odds ratio*. En este caso, una proporción de expuestos mayor entre los controles estaría indicando un efecto de protección atribuible a la lactancia, algo que se evidencia en un *odds ratio* menor que 1, con un intervalo de confianza cuyos dos extremos están por debajo de 1. La interpretación del *odds ratio* es un poco más complicada que la del riesgo relativo. Mientras que un riesgo relativo igual a 2 indicaría que en los expuestos la tasa de incidencia es el doble que en los no expuestos (1 vez más frecuente o un 100% mayor), un *odds ratio* de 2 indica que el *odds* de

enfermar es el doble para expuestos que para quienes no lo están, lo que constituye otra medida del riesgo de padecer la enfermedad. Cuanto menos frecuente es la enfermedad o daño, más cercanos estarán entre sí el *odds ratio* y el riesgo relativo.

Las fracciones prevenidas en expuestos y en la población tienen la misma interpretación que la fracción atribuible, en tanto que las pruebas estadísticas con valores de p pequeños (en particular, menores a 0,05) indican que se puede descartar el azar como explicación de la asociación observada con una reducida probabilidad de cometer el error de primer tipo (hacer un rechazo indebido). De hecho, en el ejemplo, el valor de p fue muy inferior a 0,05; concretamente, inferior al 1 por 10.000.

Los resultados del análisis emparejado presentan, además, el cálculo de un *odds ratio* a partir de los pares desiguales. Tras excluir a los pares donde casos y controles son los dos expuestos o ambos no expuestos, el *odds ratio* se calcula a través de un cociente entre los casos expuestos con controles no expuestos (a favor de que la exposición aumenta la incidencia) y los casos no expuestos con controles expuestos (situación que indicaría lo contrario).

TABLAS 2X2 ESTRATIFICADAS

La relación entre un factor de riesgo (supuesto o real) y un daño es en ocasiones "modificada" por la presencia de un tercer factor. Esta situación, conocida como *efecto de confusión*, podría definirse como la que producen aquellos factores que, estando relacionados con el factor de riesgo en estudio, condicionan la aparición del daño (siempre que no se trate de un factor que se halle en el trayecto causal que va del factor de riesgo al daño). Así, por ejemplo, la relación directa del consumo diario de comprimidos de β-carotenos y la prevención de las enfermedades coronarias será usualmente distorsionada por la presencia de otros factores que se encuentran vinculados a la "actitud" preventiva de quien toma suplementos. Seguramente, entre quienes toman esta medicación, habrá una menor proporción de fumadores y desarrollarán mayor actividad física que los que no la toman. Como estos factores tienen un efecto protector sobre la enfermedad coronaria, el efecto en la reducción del daño será resultado de la acción combinada de estos factores. De no repararse en esto, se estaría atribuyendo solo al consumo de β-carotenos una acción protectora mayor a la real.

Existen diferentes estrategias para "controlar" este efecto y una de ellas es la *estratificación*. Por ejemplo, supongamos que un estudio de casos y controles arrojó una asociación positiva entre consumo de café y cáncer de páncreas con los siguientes datos:

Café	Casos	Controles
Sí	196	104
No	89	106

OR = 2,24

Sin embargo, al considerar un tercer factor como el tabaco y dividir los individuos del estudio en dos estratos (fumadores y no fumadores) no parece existir relación entre café y cáncer de páncreas en los no fumadores y tampoco en los fumadores (OR = 1 en ambos grupos).

	No fumadores		Fumadores	
Café	**Casos**	**Controles**	**Casos**	**Controles**
Sí	32	64	164	40
No	48	96	41	10
	OR = 1,0		**OR = 1,0**	

El análisis por estratos hace evidente que el consumo de tabaco ha distorsionado la relación entre el consumo de café y el cáncer de páncreas. En esta relación es el tabaco el que estaría incrementando el riesgo de cáncer de páncreas, y como entre los fumadores el consumo de café es más frecuente, la tabla simple mostraba una asociación entre café y cáncer de páncreas.

El análisis individual de cada estrato debe ser complementado con un análisis que estime el efecto general considerando los valores de cada estrato. El método de Mantel-Haenszel es uno de los más útiles para este análisis. La existencia de diferencias entre los resultados de un análisis no estratificado y uno estratificado estará mostrando que el factor por el que se estratificó ejerce un efecto de confusión en la relación que exhiben los factores estudiados.

Si bien es posible la estratificación conjunta de varios factores con el objetivo de controlarlos o ajustarlos simultáneamente (por ejemplo,varones fumadores, varones no fumadores, mujeres fumadoras, mujeres no fumadoras), la generación de varios estratos disminuye notablemente el tamaño muestral de cada estrato, lo que hace en extremo inestables las estimaciones realizadas al interior de cada estrato.

Epidat 3.1 permite la realización de tablas 2x2 estratificadas para estudios transversales, de cohortes (con incidencia acumulada o con tasas de incidencia), y de casos y controles.

Ejemplo

En el análisis estratificado arriba descrito, donde un estudio de casos y controles busca analizar el efecto del consumo de café en la incidencia de cáncer de páncreas, pero considerando el posible efecto de confusión del consumo de tabaco, los resultados del análisis de las tablas 2x2, previa estratificación, serían los siguientes:

```
Tablas de contingencia : Tablas 2x2 estratificadas

Tipo de estudio   : Caso-control
Número de estratos: 2
Nivel de confianza: 95,0%


Tabla global
                       Casos     Controles         Total
---------------   ----------    ----------    ----------
Expuestos                196           104           300
No expuestos              89           106           195
---------------   ----------    ----------    ----------
Total                    285           210           495
```

```
ODDS RATIO (OR)


Estrato                    OR           IC(95,0%)
---------------    ----------    ----------    ----------
1                    1,000000      0,578205      1,729490    (Woolf)
2                    1,000000      0,461694      2,165934    (Woolf)
---------------    ----------    ----------    ----------
Cruda                2,244598      1,552439      3,245358    (Woolf)
Combinada (M-H)      1,000000      0,639586      1,563510
Ponderada            1,000000      0,639586      1,563510


Prueba de homogeneidad
                  Ji-cuadrado            gl       Valor p
---------------    ----------    ----------    ----------
Combinada (M-H)       0,0000             1        1,0000
Ponderada             0,0000             1        1,0000


PRUEBA DE ASOCIACIÓN DE MANTEL-HAENSZEL

    Ji-cuadrado            gl       Valor p
---------------    ----------    ----------
         0,0000             1        1,0000
```

Estos resultados incluyen una tabla global (suma de los estratos), el *odds ratio* e intervalo de confianza (calculado según el método de Woolf) para cada estrato, el *odds ratio* de la tabla global (*odds ratio* crudo) y el *odds ratio* combinado (método de Mantel-Haenszel) y ponderado por el método del inverso de la varianza. Además, se presentan las pruebas de homogeneidad entre estratos y de asociación de Mantel-Haenszel.

La diferencia entre el *odds ratio* crudo (2,24) y el combinado de Mantel-Haenszel (1,00) confirma el efecto de confusión que ejerce la variable por la que se estratifica.

La prueba de homogeneidad permite examinar las diferencias entre los *odds ratio* de los estratos. En el caso presentado, el Ji-cuadrado es bajo y el valor de p, superior a un 5% (p>0,05),lo que hace pensar que no hay diferencias apreciables entre los OR en los estratos y que, por ende, los resultados ajustados pueden considerarse para el conjunto. Un resultado que indique lo contrario marcará la necesidad de presentar por separado los resultados de cada estrato.

Finalmente, la Prueba de asociación de Mantel-Haenszel, con un valor de p por arriba de 0,05, señala la falta de asociación entre la exposición y el daño (café y cáncer de páncreas), una vez controlado el efecto del tabaco.

TABLAS 2XN SIMPLES

Las tablas 2xN simples (de un único estrato) permiten el análisis de una variable categórica (variable independiente que mide los niveles de exposición) y una variable dicotómica (variable dependiente que señala la presencia o no del daño).

Como en el caso de las tablas 2x2, se podrá optar por tres formatos de tablas según se esté analizando un estudio transversal, de cohorte, o de casos y controles.

- Estudio transversal
- Estudio de cohortes
    - Para tasas de incidencia
    - Para incidencia acumulada
- Estudio de casos y controles

Este tipo de tablas permite calcular las prevalencias, incidencia u odds (según el tipo de estudio) para cada nivel de exposición y calcula la razón de las prevalencias, tasas de incidencia u *odds ratio*, utilizando por defecto como valor de referencia el nivel 1 de exposición.

El nivel de referencia puede ser seleccionado y, si bien en general la elección es "natural", se deberá considerar que es más fácil analizar las razones y las tasas cuando se utiliza como nivel de referencia al nivel con menor prevalencia o incidencia.

Finalmente, se deberá considerar que ciertas exposiciones presentan una asociación con un daño determinado en forma de "J" o de "U", como por ejemplo el peso al nacer, el consumo de alcohol y el índice de masa corporal, todos con relación a la mortalidad. Una asociación en forma de "U" significa que ambos extremos en los niveles de exposición presentan mayor mortalidad que alguno de los valores intermedios. Los niños de bajo peso al nacer, y también los de alto peso, tienen mayor mortalidad que los de peso adecuado. En estos casos se buscará usar como referencia aquel nivel que represente la situación de menor riesgo.

En forma adicional, se podrá dar un peso a cada categoría de exposición para el cálculo de la prueba de tendencia lineal, que permite valorar la hipótesis de ausencia de tendencia lineal en el crecimiento del riesgo a medida que aumenta la exposición. El método usual para definir las puntuaciones consiste en asignar los valores 1, 2, ..., N, respectivamente, a los N niveles; si la exposición está medida en escala continua, otra posibilidad es asignar a cada categoría de exposición el punto medio del intervalo. Más que de las puntuaciones asignadas a cada nivel, la prueba de tendencia depende de la distancia entre los valores numéricos definidos. Por ejemplo, en el caso de 3 niveles de exposición, la prueba produce el mismo resultado con puntuaciones 1, 2 y 3 que con 10, 20 y 30, porque en ambos casos la distancia entre valores es constante; sin embargo, se obtendría un valor diferente si se asignaran los valores 1, 10 y 100.

Cuando el resultado de esta prueba genera una p con un valor pequeño (típicamente menor que 0,05) se considera que hay una alta posibilidad de que exista una tendencia lineal en la que a mayor exposición aumenta el riesgo. La modificación de la puntuación de cada categoría permite cambiar el peso relativo de los diferentes niveles de exposición.

Ejemplo

En un análisis del riesgo de morir en el primer año de vida con relación al peso al nacer, un estudio de cohorte realizado en dos hospitales permitió establecer lo siguiente:

Peso al nacer	Número de nacidos vivos	Defunciones antesdel 1er año
Menos de 1.500 g	65	45

1.500 a 2.499 g	370	34
2.500 a 4.199 g	6.400	57
4.200 g o más	89	8
Total	6.924	144

El análisis en Epidat 3.1 de estos mismos datos en una tabla de contingencia 2xN para estudios de cohorte (incidencia acumulada) muestra los siguientes resultados:

```
Tablas de contingencia : Tablas 2xN simples


Tipo de estudio      : Cohortes
Tipo de datos        : Incidencia acumulada
Niveles de exposición: 4
Nivel de confianza   : 95,0%


Tabla
                  Nivel 1   Nivel 2   Nivel 3   Nivel 4     Total
---------------   --------  --------  --------  --------  --------
Enfermos                45        34        57         8       144
Sanos                   20       336      6343        81      6780
---------------   --------  --------  --------  --------  --------
Total                   65       370      6400        89      6924


RIESGO RELATIVO (RR)

Nivel de exposición     Riesgo
-------------------    -------
            Nivel 1     0,6923
            Nivel 2     0,0919
            Nivel 3     0,0089
            Nivel 4     0,0899


Nivel de exposición          RR       IC(95,0%)
-------------------    --------  --------------------
            Nivel 1     77,7328   57,2954  105,4603  (Katz)
            Nivel 2     10,3177    6,8365   15,5714  (Katz)
Ref.->      Nivel 3      1,0000         -         -
            Nivel 4     10,0926    4,9630   20,5242  (Katz)

PRUEBA DE HOMOGENEIDAD ENTRE NIVELES

 Ji-cuadrado        gl   Valor p
------------  --------  --------
   1596,1653         3    0,0000

PRUEBA DE TENDENCIA LINEAL

 Ji-cuadrado        gl   Valor p
------------  --------  --------
```

```
    816,1199         1    0,0000
```

En la tabla con los datos se han ingresado como enfermos a las defunciones y como sanos a los nacidos en cada nivel de peso que sobrevivieron el primer año. El nivel 1 representa a los que pesaron menos de 1.500 gramos, el nivel 2 a los que pesaron de 1.500 a 2.499, el nivel 3 de 2.500 a 4.199 y el nivel 4 a los que pesaron 4.200 ó más.

Luego de la tabla se presentan los riesgos (que equivalen a las tasas de la tabla anterior), y luego el riesgo relativo y sus intervalos de confianza (IC 95%). Puesto que se señaló como nivel de referencia al nivel 3 (peso adecuado), el RR del nivel 3 es 1, en tanto los restantes RR se deben interpretar como exceso de riesgo en relación a los niños que nacieron con peso adecuado.

La prueba de homogeneidad con un Ji-cuadrado muy elevado y un valor de p<0,0001, señala que existe un riesgo distinto en los diferentes niveles de exposición, comparados con el de referencia.

Finalmente, la prueba de tendencia lineal hará pensar que existe una relación lineal en la que, a menor peso, mayor es el riesgo de morir antes del año. Sin embargo, como el patrón de los riesgos muestra una curva en forma de "J", al eliminar el nivel de exposición 4 (niños de alto peso al nacer), la prueba de tendencia lineal arroja resultados aún más significativos:

```
Tablas de contingencia : Tablas 2xN simples

Tipo de estudio      : Cohortes
Tipo de datos        : Incidencia acumulada
Niveles de exposición: 3
Nivel de confianza   : 95,0%

Tabla
                   Nivel 1   Nivel 2   Nivel 3     Total
---------------   --------  --------  --------  --------
Enfermos                45        34        57       136
Sanos                   20       336      6343      6699
---------------   --------  --------  --------  --------
Total                   65       370      6400      6835

RIESGO RELATIVO (RR)

Nivel de exposición     Riesgo
-------------------    -------
            Nivel 1     0,6923
            Nivel 2     0,0919
            Nivel 3     0,0089

Nivel de exposición          RR       IC(95,0%)
-------------------    --------  --------------------
            Nivel 1     77,7328   57,2954  105,4603  (Katz)
            Nivel 2     10,3177    6,8365   15,5714  (Katz)
Ref.->      Nivel 3      1,0000         -         -


PRUEBA DE HOMOGENEIDAD ENTRE NIVELES

 Ji-cuadrado        gl   Valor p
```

```
------------  --------  --------
   1644,7339         2    0,0000

PRUEBA DE TENDENCIA LINEAL

 Ji-cuadrado        gl   Valor p
------------  --------  --------
   1123,8274         1    0,0000
```

TABLAS 2XN ESTRATIFICADAS

La estratificación de las tablas 2xN permite incorporar otra variable o factor para analizar si la relación entre la exposición y el daño cambia según las diferentes categorías de la variable por la que se está estratificando.

También aquí se podrá optar por tres formatos de tablas según se esté analizando un estudio transversal, de cohortes, o de casos y controles, y deberá definirse un nivel de referencia para el cálculo de las razones de prevalencia, riesgos relativos u *odds ratio*, respectivamente.

Ejemplo

Si en el estudio del epígrafe precedente se quisiera considerar por separado los datos de uno y otro hospital, se podrían presentar los datos en una tabla como la que sigue:

	Número de nacidos vivos			**Defunciones antes del 1er año**		
Peso al nacer	**Total**	**Hosp A**	**Hosp B**	**Total**	**Hosp A**	**Hosp B**
Menos de 1.500 g	65	40	25	45	21	24
1.500 a 2.499 g	370	220	150	34	18	16
2.500 a 4.199 g	6.400	3.390	3.010	57	25	32
4.200 g o más	89	60	29	8	5	3
Total	**6.924**	**3.710**	**3.214**	**144**	**69**	**75**

El análisis en Epidat 3.1 de estos mismos datos muestra estos resultados:

```
Tablas de contingencia : Tablas 2xN estratificadas

Tipo de estudio      : Cohortes
Tipo de datos        : Incidencia acumulada
Nivel de confianza   :  95,0%
Niveles de exposición: 4
Número de estratos   : 2


Tabla global
                   Nivel 1   Nivel 2   Nivel 3   Nivel 4     Total
---------------   --------- --------- --------- --------- ---------
Enfermos                45        34        57         8       144
Sanos                   20       336      6343        81      6780
---------------   --------- --------- --------- --------- ---------
```

```
Total                     65       370      6400        89      6924


RESULTADOS CRUDOS

Nivel de exposición          RR       IC(95,0%)
-------------------    --------  --------------------
            Nivel 1     77,7328   57,2954  105,4603  (Katz)
            Nivel 2     10,3177    6,8365   15,5714  (Katz)
Ref.->      Nivel 3      1,0000         -         -
            Nivel 4     10,0926    4,9630   20,5242  (Katz)
RR: Riesgo relativo


RESULTADOS AJUSTADOS

Nivel de exposición          RR         IC(95,0%)
-------------------    --------  --------------------
            Nivel 1     80,2639   59,4295  108,4022  (Mantel-Haenszel)
            Nivel 2     10,5647    6,9896   15,9685  (Mantel-Haenszel)
Ref.->      Nivel 3      1,0000         -         -
            Nivel 4     10,6525    5,2134   21,7661  (Mantel-Haenszel)


PRUEBA DE TENDENCIA LINEAL

 Ji-cuadrado        gl   Valor p
------------  --------  --------
    831,9166         1    0,0000
```

Los RR crudos son iguales a los del análisis simple, ya que de hecho se calculan sin considerar los estratos; en cambio, los RR ajustados consideran el efecto de la variable hospital (lugar de realización del parto). A pesar de que los riesgos para los diferentes grupos de peso difieren entre un hospital y otro (en el hospital B las tasas de mortalidad resultaron más elevadas en cada grupo), la escasa diferencia entre los RR crudos y ajustados de los niveles 2 y 4 señala que el riesgo que implica nacer con un peso de 1.500 g a 2.499 g, o de 4.200 g o más no está significativamente influenciado por el hospital donde se produjeron los nacimientos. Sin embargo, en el caso de los niños nacidos con menos de 1.500 g (nivel 1) la diferencia entre los RR (crudo y ajustado) estaría indicando que el lugar del parto modifica el efecto del riesgo del bajo peso. En otras palabras, las tasas de mortalidad en el hospital "B" son mayores, pero en el caso particular de los niños con muy bajo peso el riesgo de morir es mayor en el hospital "B".

El resto de los resultados deben interpretarse como en el análisis simple.

BIBLIOGRAFÍA

1. Breslow NE, Day NE. *Statistical methods in cancer research I. The analysis of case-control studie*s. Lyon: IARC; 1980.

2. Everitt BS. *The analysis of contingency tables*. London: Chapman and Hall; 1977.

3. Fleiss JL. *Statistical methods for rates and proportions*. New York: John Wiley & Sons; 1981.

4. Rothman KJ, Greelnland S. *Modern epidemiology.* 2nd ed. Philadelphia: Lippincott-Raven; 1998.

5. Schlesselman JJ. *Cases-control studies. Design, conduct, analysis*. New York: Oxford University Press; 1982.

TABLAS GENERALES

TABLAS MXN

Una tabla de contingencia MxN se obtiene cuando se clasifican los individuos de una muestra con respecto a dos variables cualitativas con M y N categorías respectivamente. Esta clasificación debe ser exhaustiva y mutuamente exclusiva, lo que significa que cada individuo ha de poder asignarse a una de estas MxN categorías, y solo a una. El submódulo de Tablas MxN permite analizar 2 variables nominales u ordinales en tablas de hasta 20 columnas por 20 filas.

En general, la cuestión más importante que se plantea ante una tabla de contingencia es si las variables son independientes o no. Para resolver esta cuestión se han propuesto diversas pruebas de hipótesis; las incluidas en Epidat 3.1 son:

- La prueba Ji-cuadrado de Pearson
- La prueba de razón de verosimilitudes
- La prueba Ji-cuadrado con corrección de Yates (para tablas 2x2)
- La prueba exacta de Fisher (para tablas 2x2)

La prueba Ji-cuadrado de Pearson se basa en la hipótesis de que no hay discrepancias entre las frecuencias observadas en la tabla y las esperadas en caso de independencia o no asociación entre las variables. El estadístico de esta prueba sigue, aproximadamente, una distribución Ji-cuadrado con (M-1)x(N-1) grados de libertad. Cochran ha estudiado la validez de la aproximación y recomienda que sólo se utilice esta prueba cuando se cumplan las siguientes condiciones: menos de un 20% de las celdas de la tabla tienen frecuencia esperada menor que 5 y ninguna celda tiene frecuencia esperada menor que 1.

Para tablas 2x2, existe una versión del estadístico Ji-cuadrado de Pearson que, para mejorar la aproximación, incorpora la llamada corrección de Yates; pero hay gran discrepancia en la literatura en cuanto a la validez de este procedimiento, que muchos autores cuestionan. Para tablas basadas en tamaños de muestra grandes, se obtienen resultados similares con y sin corrección; y en el caso de muestras pequeñas, la recomendación es utilizar métodos exactos, como es el caso de la prueba exacta de Fisher. Esta prueba calcula la probabilidad exacta de obtener los resultados observados si las dos variables son independientes y los totales marginales son fijos.

La prueba de razón de verosimilitudes es una alternativa a la prueba Ji-cuadrado para contrastar la hipótesis nula de que las dos variables son independientes, y está basado en la teoría de la máxima verosimilitud. El estadístico de esta prueba, que sigue también una distribución

Ji-cuadrado con (M-1)x(N-1) grados de libertad, se basa en comparar la probabilidad de los datos observados con la probabilidad de los datos esperados en caso de ser cierta la hipótesis de independencia. Por tanto, valores altos del estadístico son indicativos de asociación entre las variables. La distribución del estadístico es también aproximada, por lo que puede no ser apropiado si el tamaño de la muestra es pequeño.

Por otra parte, Epidat 3.1 calcula varias medidas que cuantifican la intensidad de la asociación entre las dos variables de la tabla de contingencia. Algunas de estas medidas son válidas en general para variables nominales; otras son específicas de variables ordinales. En el primer caso, se incluyen:

Coeficiente de contingencia C. Toma valores entre 0 y 1. Vale 0 en caso de independencia completa; sin embargo, no siempre toma el valor máximo de 1, porque incluso en el caso de asociación completa, el valor de C depende del número de filas y columnas de la tabla.

Coeficiente V de Cramer. Corrige el problema de la dependencia que tiene el coeficiente de contingencia del número de filas y columnas. Para tablas 2×2 toma valores entre -1 y 1 y, en otro caso, varía entre 0 y 1, alcanzando el –1 ó 0, respectivamente, en caso de independencia completa, y el 1 en caso de asociación completa.

Estas dos medidas de asociación están basadas en el estadístico Ji-cuadrado de la prueba de Pearson y son simétricas, es decir, no dependen de cual es la variable de filas y cual la de columnas; si se intercambian las variables se obtiene el mismo resultado para estos coeficientes.

Tau de Goodman y Kruskal. Es una medida de asociación asimétrica, que permite considerar a una de las variables como independiente y a la otra como dependiente, y ocasionalmente valorar en qué medida la primera predice a la segunda. Epidat 3.1 presenta el coeficiente Tau de Goodman y Kruskal en dos situaciones: tomando las filas como categorías de la variable dependiente (Filas/Columnas) o considerando las columnas como tales (Columnas/Filas). Ambos coeficientes toman valores entre 0 y 1, y serán más cercanos a 1 cuanto mayor sea la capacidad de predecir sin error la variable dependiente, mientras que un valor de 0 significa que la variable independiente no tiene capacidad para predecir la dependiente

Cuando las variables son ambas ordinales, existen medidas de asociación específicas que tienen en cuenta la ordenación, y toman valores positivos cuando una de las variables tiende a aumentar a medida que lo hace la otra, y valores negativos en la situación inversa, es decir, cuando los valores altos de una variable se asocian con valores bajos de la otra. Las incluidas en Epidat 3.1 son:

Tau-b de Kendall. Está basado en el número de concordancias, discordancias y empates entre pares de casos. Un par es concordante si los valores de ambas variables para un caso son menores/mayores que los valores correspondientes para el otro caso, y discordante si ocurre lo contrario. Considérese, por ejemplo, la relación entre las variables clase social (CS: 3 categorías en orden creciente) y diagnóstico (D: 4 categorías de menor a mayor gravedad) en una muestra de pacientes psiquiátricos; un par de pacientes con valores CS=2, D=2 y CS=3, D=4 es concordante, mientras que un par discordante sería por ejemplo el formado por dos pacientes con valores CS=2, D=2 y CS=3, D=1.

El coeficiente Tau-b de Kendall puede tomar valores entre -1 y 1, aunque solo alcanza estos extremos en el caso de tablas cuadradas. Si el predominio de los pares es concordante, el valor es próximo a 1 y se dice que la asociación es positiva; si la mayoría de los pares es discordante, la

asociación será negativa y el valor se acercará a –1. El valor 0 indica que no hay relación entre las dos variables y ocurre cuando los pares concordantes y discordantes son igualmente probables.

Tau-c de Stuart. Es una variante del coeficiente anterior, y se diferencia de él en que puede alcanzar los valores mínimo y máximo, -1 y 1, en tablas de cualquier dimensión, salvo pequeñas discrepancias cuando el tamaño de la muestra no es un múltiplo del mínimo entre M y N (número de filas y columnas, respectivamente).

Gamma de Goodman y Kruskal. También basada en pares concordantes y discordantes, toma valores entre -1 y 1; el valor 0 se alcanza en caso de que las variables sean independientes y la asociación es tanto mayor cuanto más se aproxima gamma a –1 ó a 1.

D de Somers. Es una medida asimétrica que, por tanto, permite realizar un análisis de relación entre dos variables tomando una de ellas como dependiente, por lo que se obtendrán dos índices, igual que en el caso de la Tau de Goodman y Kruskal, uno cuando la variable independiente es la situada en las filas y otro en el caso de que dicha variable sea la de columnas. Los dos coeficientes de Somers también toman valores entre -1 y 1.

Para estas medidas de asociación con datos ordinales Epidat 3.1 presenta el error estándar y una prueba de significación.

Coeficiente de correlación por rangos de Spearman. Es una medida de correlación utilizada habitualmente para variables ordinales. Los valores de cada una de las variables se clasifican de menor a mayor y se calcula el coeficiente de correlación de Pearson en base a los rangos. Los valores del coeficiente de correlación por rangos de Spearman varían entre –1 y 1, y un valor 0 indica que no existe ninguna relación lineal entre las variables.

Por último, en el caso de tablas 2xN con variables ordinales, Epidat 3.1 realiza una prueba de tendencia lineal, que contrasta la hipótesis de que los porcentajes, calculados para cada columna, tienden a aumentar o disminuir a lo largo de la primera fila o, equivalentemente, de la segunda fila.

Ejemplo 1

Para analizar si la distribución de los motivos de consulta en 4 centros de atención ambulatoria pediátrica eran similares, se clasificaron las consultas en 6 grupos: (1) Medicina preventiva; (2) Infecciones respiratorias altas; (3) Otras enfermedades agudas; (4) Enfermedades crónicas; (5) Traumatismos e intoxicaciones; y (6) Problemas sociales.

La tabla resultante fue la siguiente:

Centro de Salud	**Grupo de motivos de consulta**						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Centro A	350	87	65	12	23	23	**560**
Centro B	120	43	38	6	10	12	**229**

Centro C	426	67	34	7	45	67	**646**
Centro D	267	49	35	5	18	18	**392**
Total	**1.163**	**246**	**172**	**30**	**96**	**120**	**1.827**

El análisis en Epidat 3.1 de estos datos mostraría estos resultados:

```
Tablas de contingencia : Tablas MxN

Número de filas   : 4
Número de columnas: 6
Filas y columnas  : Nominales

Frecuencias observadas
               1         2         3         4         5         6     Total
-----    -------   -------   -------   -------   -------   -------   -------
    1        350        87        65        12        23        23       560
    2        120        43        38         6        10        12       229
    3        426        67        34         7        45        67       646
    4        267        49        35         5        18        18       392
-----    -------   -------   -------   -------   -------   -------   -------
Total       1163       246       172        30        96       120      1827

Porcentajes (Por filas)
               1         2         3         4         5         6       Total
-------  -------   -------   -------   -------   -------   -------   ----------
      1    62,50     15,54     11,61      2,14      4,11      4,11      100,00
      2    52,40     18,78     16,59      2,62      4,37      5,24      100,00
      3    65,94     10,37      5,26      1,08      6,97     10,37      100,00
      4    68,11     12,50      8,93      1,28      4,59      4,59      100,00
-------  -------   -------   -------   -------   -------   -------   ----------
  Total    63,66     13,46      9,41      1,64      5,25      6,57      100,00

% de celdas con frecuencia esperada <5: 4,2%


Prueba Ji-cuadrado de Pearson

Ji-cuadrado         gl   Valor p
-----------   -------- --------
    76,9442         15    0,0000


Prueba de razón de verosimilitudes

Ji-cuadrado         gl   Valor p
   --------   -------- --------
    75,4224         15    0,0000



Medidas de asociación para variables nominales
                            Estimación
--------------------------- ----------
Coeficiente de contingencia     0,2010
Coeficiente V de Cramer         0,1185
```

```
Tau de Goodman y Kruskal
       Filas/Columnas           0,0098
       Columnas/Filas           0,0156
```

El porcentaje de celdas con frecuencia esperada menor que 5, que se informa inmediatamente debajo de la tabla, está vinculado a las exigencias para la utilización de la prueba Ji-cuadrado de Pearson. Estas exigencias son:

- Menos de un 20% de celdas con frecuencia esperada menor que 5.
- Ninguna celda con frecuencia esperada menor que 1.

Como ambas condiciones están presentes, puede considerarse válido el uso de la Ji-cuadrado de Pearson. Esta, por otra parte, indica que se puede descartar enfáticamente que la distribución observada diste de la uniformidad entre centros solo por producto del azar. Lo que, en otras palabras, indica que existe un patrón de distribución de los motivos de consulta que no es el mismo en todos los centros. De hecho, el porcentaje de motivos en la categoría 6 (Problemas sociales), por ejemplo, resultó mucho mayor en el centro C que en los restantes centros.

Por su parte, los valores de los coeficientes de contingencia y de Cramer obtenidos indican una asociación baja entre el centro de salud y el motivo de consulta.

El valor 0,0156 del coeficiente Tau de Goodman y Kruskal calculado considerando la variable "Centro de salud" (filas) como independiente tiene la siguiente interpretación: conociendo el centro donde se hizo la consulta, se reduce en un 1,56% la probabilidad de cometer un error al predecir el motivo de la consulta (columnas). Esto significa que el centro de salud no tiene capacidad predictiva sobre el motivo de la consulta.

Ejemplo 2

Se quiere estudiar la relación entre la edad de las mujeres y su aceptación de una ley sobre interrupción del embarazo. Para ello se ha llevado a cabo una encuesta sobre 400 mujeres cuyos resultados se recogen en la siguiente tabla:

	Aceptación		
Edad	**Baja**	**Media**	**Alta**
0-18	21	34	25
18-35	24	31	25
36-50	30	30	20
51-65	37	30	13
> 65	40	30	10

Dada la naturaleza ordinal de las dos variables, debe seleccionarse la opción de "Datos ordinales" en la pantalla de entrada de Epidat 3.1 Los resultados que muestra el programa son los siguientes:

```
Tablas de contingencia : Tablas MxN

Número de filas   : 5
Número de columnas: 3
```

Filas y columnas : Ordinales

Frecuencias observadas

	1	2	3	Total
1	21	34	25	80
2	24	31	25	80
3	30	30	20	80
4	37	30	13	80
5	40	30	10	80
Total	152	155	93	400

Porcentajes (Total)

	1	2	3	Total
1	5,25	8,50	6,25	20,00
2	6,00	7,75	6,25	20,00
3	7,50	7,50	5,00	20,00
4	9,25	7,50	3,25	20,00
5	10,00	7,50	2,50	20,00
Total	38,00	38,75	23,25	100,00

% de celdas con frecuencia esperada <5: 0,0%

Prueba Ji-cuadrado de Pearson

Ji-cuadrado	gl	Valor p
19,2828	8	0,0134

Prueba de razón de verosimilitudes

Ji-cuadrado	gl	Valor p
19,9445	8	0,0105

Medidas de asociación para variables nominales

	Estimación
Coeficiente de contingencia	0,2145
Coeficiente V de Cramer	0,1553

Tau de Goodman y Kruskal

Filas/Columnas	0,0224
Columnas/Filas	0,0121

Medidas de asociación para variables ordinales

	Estimación	EE	Estadístico Z	Valor p

Tau-b de Kendall	-0,1799	0,0400	-4,4852	0,0000
Tau-c de Stuart	-0,1948	0,0434	-4,4852	0,0000
Gamma de Goodman y Kruskal	-0,2478	0,0545	-4,4852	0,0000
D de Sommers				
Filas/Columnas	-0,1994	0,0443	-4,4852	0,0000
Columnas/Filas	-0,1623	0,0362	-4,4852	0,0000

Coeficiente de correlación por rangos de Spearman

R	Estadístico t	gl	Valor p
-0,2134	-4,3585	398	0,0000

Cuando las dos variables son ordinales, como en este ejemplo, Epidat 3.1 muestra todos los resultados que se presentan para variables nominales, válidos también en este caso y, además, las medidas de asociación específicas para variables ordinales.

Las medidas nominales cuantifican el grado de asociación, mientras que las ordinales indican además si la asociación es monótona en el sentido de que la clasificación en una variable tiende a aumentar cuando lo hace la otra (asociación positiva) o a disminuir (asociación negativa).

En el ejemplo, la prueba Ji-cuadrado de Pearson y la prueba de razón de verosimilitudes indican que hay evidencia de asociación entre el grado de aceptación del aborto y la edad de las mujeres, en tanto que las medidas de asociación informan que ésta es negativa, es decir, que el grado de aceptación disminuye al aumentar la edad.

REGRESIÓN LOGÍSTICA

Conceptos generales

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 solo 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. Surge en la década del 60 con la aparición del trabajo de Cornfield, Gordon y Smith[1] sobre el riesgo de padecer una enfermedad coronaria que constituye su primera aplicación práctica trascendente. Su generalización dependía de la solución que se diera al problema de la estimación de los coeficientes. El algoritmo de Walker-Duncan[2] para la obtención de los estimadores de máxima verosimilitud vino a solucionar en parte este problema, pero era de naturaleza tal que el uso de computadoras era imprescindible.

De su amplio y creciente empleo han dado cuenta varias revisiones. Silva, Pérez y Cuellar[3] consignan que ésta fue la técnica estadística más usada entre los 1.045 artículos publicados por *American Journal of Epidemiology* entre 1986 y 1990 (casi 3 de cada 10 trabajos allí publicados). Levy y Stolte[4] llevaron a cabo un estudio para caracterizar la tendencia en el uso de métodos estadísticos surgidos recientemente (entre los 60 y los 70) y que, además, hubieran tenido un impacto considerable en el análisis de datos biomédicos; entre ellos figura la regresión logística. Las propias *American Journal of Public Health* y *American Journal of Epidemiology* han puesto de manifiesto que la tendencia en el uso de la RL fue creciente en los artículos de ambas revistas. El porcentaje de artículos publicados en la década de los 70 que hicieron uso de este recurso fue 0,7%; ya en los 80, ascendió espectacularmente a 17,0% y a lo largo de la década de los 90 alcanzó 28,1%.

En MEDLINE, base de datos que contiene referencias bibliográficas y resúmenes de 4.500 revistas biomédicas de la literatura de habla inglesa, usando PUBMED se encontró que el crecimiento en los últimos cinco años ha sido sostenido: los resúmenes publicados que hacen mención del término *logistic regression* son para 1997, 1998, 1999, 2000 y 2001, respectivamente, los siguientes: 3.394, 3.654, 3.972, 4.397 y 5.218.

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 $X_1,X_2,...,X_r$. Concretamente, lo que hace el programa es hallar los coeficientes β_0, β_1, ..., β_r que mejor se ajustan a la siguiente representación funcional:

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

donde exp(.) representa la función exponencial (antilogaritmo neperiano).

Regresión logística para datos tabulados. Epidat 3.1 se ocupa del análisis de datos tabulados; la regresión logística es el único método de regresión que puede emplearse legítimamente con datos recogidos en tablas de contingencia (véase Silva[5]). Este hecho, unido al intenso empleo contemporáneo de este recurso, determinó que la presente versión de Epidat incluyera la técnica de RL, para lo cual se empleó el algoritmo sugerido por Jones[6].

Debe enfatizarse, sin embargo, que con frecuencia los datos disponibles no son susceptibles de ser recogidos en una tabla de contingencia por involucrar variables continuas, en cuyo caso ha de emplearse un programa que contemple esta posibilidad. Por otra parte, al emplear la RL, como ocurre en rigor con cualquier otra técnica estadística, se debe ser cauteloso. Si bien el modelo no

tiene restricciones en cuanto a la distribución de las variables independientes (eso es lo que hace posible, precisamente, que se pueda emplear con datos tabulados), para que el análisis tenga sentido pleno, debe aplicarse con fines predictivos solo en los estudios prospectivos, cuando se tenga certeza de que los acontecimientos registrados por las variables independientes ocurrieron antes que los desenlaces. Por otra parte, si se emplea para el tratamiento de estudios retrospectivos (estudios de casos y controles), entonces debe tenerse en cuenta que no se puede emplear el modelo para hacer predicciones.

La variable de respuesta se codificará siempre como 1 si el evento se produce, y como 0 en caso opuesto. Cada variable independiente se codificará como 1, 2, ⋯, k, donde k es el número de categorías que contiene. Si hay r variables independientes con k_1, k_2, ..., k_r categorías respectivamente, se tendrán en total $2\times k_1\times k_2\times...\times k_r$ configuraciones posibles; para cada una de ellas hay que consignar el número de sujetos de la muestra que se encuadran en dicha configuración (la frecuencia). En Epidat 3.1, el número máximo de variables independientes es 10.

Variables 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, ..., 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 $Z_1= Z_2= ...= Z_{k-1}=0$; 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

En cualquier caso, si se ajusta un modelo que incluya una variable nominal con **k** clases, ésta debe ser sustituida por las **k-1** variables *dummy*, y a cada una de ellas corresponderá su respectivo coeficiente.

Como se ilustra más adelante, una de las razones que confiere especial interés a la regresión logística consiste en que suple en buena medida al análisis basado en la estratificación. Al igual que el análisis estratificado, la RL permite la evaluación y control del efecto de confusión, así como evaluar y describir el de interacción.

Bondad de ajuste del modelo. Siempre que se ajusta un modelo de regresión, de cualquier tipo, una precaución importante a los efectos de sacar conclusiones es la de corroborar que este modelo se ajusta efectivamente a los datos usados. La RL no es una excepción. Epidat 3.1 permite evaluar la calidad del ajuste del modelo estimado mediante el test de bondad de ajuste de Hosmer y Lemeshow[7]. El estadístico que ellos proponen se calcula definiendo 10 grupos mediante los deciles de las probabilidades predichas por el modelo, y comparando las frecuencias observadas en dichos grupos con las esperadas.

Es bien conocido que, en el contexto de la regresión lineal múltiple, se suele emplear el llamado *coeficiente de determinación* (R^2) para cuantificar mediante una única medida, con cotas interpretables, el grado de "explicación de la variabilidad de la variable de respuesta" conseguido con el modelo por parte de las variables independientes. Varias sugerencias se han hecho para obtener algo similar en el marco de la RL. Sin embargo, no hay una opinión unánime sobre cuál podría ser la mejor. Epidat 3.1 ha incorporado una, preferida por Mittlböck y Schemper[8] (quienes examinan 12 posibles mediciones) a la que se denomina aquí, análogamente, *coeficiente de determinación*. R^2 es un número que se halla necesariamente entre 0 y 1. Alcanza el valor 1 cuando el vaticinio es perfecto (esto quiere decir, que R^2 alcanzaría el valor máximo solo si el modelo atribuyera probabilidad 1 a aquellos sujetos de la muestra que efectivamente tuvieron el evento, y valores iguales a 0 a quienes no lo tuvieron) y R^2 se aproxima a 0 en la medida que las probabilidades atribuidas por el modelo disten más, respectivamente, de 1 y 0.

Cabe advertir, no obstante, que este coeficiente no mide la bondad del ajuste (un concepto diferente al de "variabilidad explicada por el modelo"), la cual debe valorarse a través de las pruebas específicamente diseñadas con ese fin (en particular, la prueba de Hosmer y Lemeshow).

Cociente de verosimilitud. Para que un modelo sea considerado adecuado, éste debe atribuir una alta probabilidad de enfermar a aquellos sujetos para los cuales Y=1 (o sea, a los que padecen la enfermedad) y una baja probabilidad de enfermar (o una alta probabilidad de no padecerla) a quienes no manifiestan la enfermedad. Por tanto, una medida razonable para valorar el grado en que el modelo arroja resultados coherentes con los datos usados para su construcción sería el producto de todas las probabilidades (predichas por el modelo) de que los n sujetos de la muestra tengan la condición que realmente tienen. Si se llama $\hat{P}_i$ a la probabilidad estimada por el modelo de que el i-ésimo sujeto tenga la condición, se puede computar la expresión siguiente:

$$V = \underbrace{\hat{P}_1 \times \hat{P}_2 \times \cdots \times \hat{P}_d}_{\substack{d \text{ sujetos con la} \\ \text{condición}}} \times \underbrace{(1 - \hat{P}_{d+1}) \times (1 - \hat{P}_{d+2}) \times \cdots \times (1 - \hat{P}_n)}_{\substack{n - d \text{ sujetos sin la} \\ \text{condición}}}$$

conocida como la *verosimilitud del modelo*. Un modelo completamente exitoso, el cual atribuya una probabilidad de enfermar igual a 1 a cada sujeto enfermo y de 0 a cada sujeto libre de la enfermedad, tendría una verosimilitud máxima de 1; por el contrario, un modelo relativamente no exitoso tendría una verosimilitud pequeña. En consecuencia, la proximidad de la verosimilitud a 1 expresa cuán eficiente ha sido este recurso para modelar la realidad.

Debido a que la función de verosimilitud mide la plausibilidad de un modelo de regresión logística, no debe sorprender que para valorar su capacidad predictiva sea central la consideración de la verosimilitud; es decir, de la magnitud V antes introducida. Concretamente, se suele emplear la expresión:

$$L=-2\ln V$$

A esta transformación se le conoce como *lejanía* del modelo (*deviance* en inglés). Nótese que, siendo $V<1$, su logaritmo siempre será negativo; de modo que la lejanía L siempre será un número positivo. El grado de ajuste de un modelo es mejor cuanto más próxima a 1 es la *verosimilitud* y, en consecuencia, cuanto más se aproxima a cero la *lejanía*.

Siempre que se ajusta un modelo, el algoritmo de la regresión logística computa dos lejanías: la que corresponde propiamente al modelo que se ha ajustado (L), y la que corresponde al "modelo nulo" (L_0) que es aquel en que no se ha incorporado ninguna variable independiente.

La lejanía del modelo nulo es más grande que la de cualquier modelo ampliado. Esto es razonable debido a que se trata de un modelo mucho menos sofisticado y debe necesariamente tener una falta de ajuste mayor. La diferencia entre estas lejanías mide "el aporte" que hacen las variables incorporadas al modelo. Es decir, para valorar dicho aporte se puede calcular el cociente o razón de verosimilitudes:

$$RV = L_0 - L = -2\ln V_0 + 2\ln V = -2(\ln V_0 - \ln V) = -2\ln\left(\frac{V_0}{V}\right)$$

que se distribuye Ji-cuadrado con k grados de libertad, donde k es el número de variables presentes en el modelo ampliado.

En general, esta razón de verosimilitudes es útil, en fin, para determinar si hay una diferencia significativa entre incluir en el modelo todas las variables y no incluir ninguna; o, dicho de otro modo: RV sirve para evaluar si las variables X_1, X_2, ..., X_r tomadas en conjunto, contribuyen efectivamente a "explicar" las modificaciones que se producen en $P(Y=1)$.

Curva ROC. En un contexto predictivo debe seleccionarse el mejor modelo entre todos los posibles. El área bajo la curva ROC es una forma de comparar diferentes modelos, ya que da una medida de la capacidad predictiva de los mismos. Cuanto mayor sea esa área, más eficiente es el modelo. Para un modelo concreto, la curva ROC se construye del modo siguiente:

Las probabilidades predichas por el modelo permiten, definiendo un punto de corte, clasificar a los sujetos en dos grupos: los que presentan el evento (respuesta 1) y los que no lo presentan (respuesta 0). Desde esta perspectiva, puede considerarse el modelo de regresión logística como un medio para definir una prueba diagnóstica cuantitativa. Fijando un umbral para hacer el diagnóstico (por ej, diagnosticar enfermedad si P(Y=1)>0,8 y declarar sano en caso contrario) en una situación en que se conozcan los verdaderos desenlaces, para la que es posible calcular la sensibilidad (porcentaje de sujetos con el evento que son clasificados correctamente por el modelo) y la especificidad (porcentaje de sujetos sin el evento que son clasificados correctamente por el modelo). Si se toman varios puntos de corte o umbrales sucesivamente, se tendrán sucesivas parejas de sensibilidad-especificidad. La curva ROC se obtiene representando, en un cuadrado de lado 1, los valores de 1-especificidad frente a sensibilidad para todos los posibles puntos de corte en las probabilidades predichas.

La curva empieza en el punto (0,0), que corresponde al punto de corte 1, y termina en (1,1) que se obtiene al considerar el 0 como punto de corte. Si el modelo tiene capacidad predictiva nula, la curva coincide con la diagonal principal del cuadrado, y el área bajo la curva toma su valor mínimo de 0,5. Por el contrario, un modelo perfecto tiene una curva ROC con área 1.

Nota: En los modelos múltiples puede ser interesante incorporar la interacción entre dos variables explicativas; esto significa que la influencia de una variable sobre la respuesta puede ser diferente en función de los valores que tome otra variable incluida en el modelo. Epidat 3.1 no contempla la posibilidad de definir interacciones de forma automática, pero esta posibilidad se puede encarar por parte del usuario definiendo previamente el producto de las dos variables cuya interacción se desea evaluar, e incluyéndola en el modelo como una explicativa más. En tal caso, el usuario debe emplear la lectura automática de datos, tras haber construido un archivo (por ejemplo, en formato Excel) que incluya esta información. Por otra parte, si al menos una de las variables cuya interacción se quiere valorar se fuera a tratar como *dummy*, entonces no se debe emplear para dicha variable la construcción automática de variables *dummy* que realiza Epidat 3.1 (ya que produciría un resultado erróneo). En tal caso, las variables *dummy* han de ser construidas por el usuario y éste ha de incorporar a la tabla de contingencia incluida en el archivo antes mencionado k-1 productos (los de dichas variables dummy por la otra variable considerada en la interacción). Véase el ejemplo 1 para comprender mejor este problema. Naturalmente, esta idea puede extenderse a más variables. Podrían incorporarse términos que involucren a tres o más de ellas. Una regla general que se ha dado es que, si en un ajuste se incluye un término de cierto orden, se incluyan entonces todos los de orden inferior.

Ejemplo 1

Supóngase que se quiere modelar a través de la RL la relación entre el hecho de tener anticuerpos a cierto VIRUS (variable de respuesta: 1-SI, 0-NO) y dos variables independientes: ZONA DE RESIDENCIA (con 4 categorías: 1-NORTE, 2-SUR, 3-ESTE y 4-OESTE) y FACTOR RH (con dos categorías: 1-POSITIVO y 2-NEGATIVO). El archivo VIRUS.xls, que se incluye en el paquete de Epidat 3.1, contiene una tabla de contingencia con los datos de los 6.305 sujetos que constituyen la muestra, y que tiene la siguiente estructura:

VIRUS	ZONA	RH	FRECUENCIA
0	1	1	445
0	2	1	729
0	3	1	32
0	4	1	242
0	1	2	464

0	2	2	757
0	3	2	67
0	4	2	284
1	1	1	463
1	2	1	772
1	3	1	82
1	4	1	290
1	1	2	483
1	2	2	789
1	3	2	90
1	4	2	316

Al realizar el ajuste con VIRUS en función de ZONA y RH, se declara ZONA como dummy, y se obtiene lo siguiente:

```
Tablas de contingencia : Regresión logística

Archivo de trabajo: C:\Archivos de programa \Epidat 3.1 \Ejemplos \Tablas de
contingencia\VIRUS.xls
Campo que identifica:
  Variable respuesta: VIRUS
  Frecuencias: FRECUENCIA
  Variables explicativas: ZONA RH

Nivel de confianza:  95,0%

Variable respuesta:
Valor    N° sujetos
-----    ----------
0              3020
1              3285
-----    ----------
Total          6305




Variables Dummy:
ZONA
Categoría ZONA-1    ZONA-2    ZONA-3
--------------    ------    ------
        1     0         0         0
        2     1         0         0
        3     0         1         0
        4     0         0         1
La sucesión de estimadores ha convergido
N° iteraciones necesarias: 3


-2 ln Verosimilitud inicial: 8729,444680
-2 ln Verosimilitud final  : 8711,623122


Cociente de verosimilitud
Estadístico   g.l. Valor p
```

Ji-cuadrado	g.l.	Valor p
17,8216	2	0,0001

Coeficiente de determinación: 0,0028

Variable	Coeficiente	S.E.	Valor de Z	Valor p
Constante	0,108645			
ZONA-1	0,009200	0,058917	0,156153	0,8759
ZONA-2	0,515650	0,134489	3,834145	0,0001
ZONA-3	0,102583	0,075566	1,357522	0,1746
RH	-0,045509	0,050541	-0,900438	0,3679

Variable	Odds ratio	IC(95,0%)	
ZONA-1	1,009243	0,899177	1,132781
ZONA-2	1,674727	1,286669	2,179824
ZONA-3	1,108029	0,955496	1,284912
RH	0,955511	0,865395	1,055010

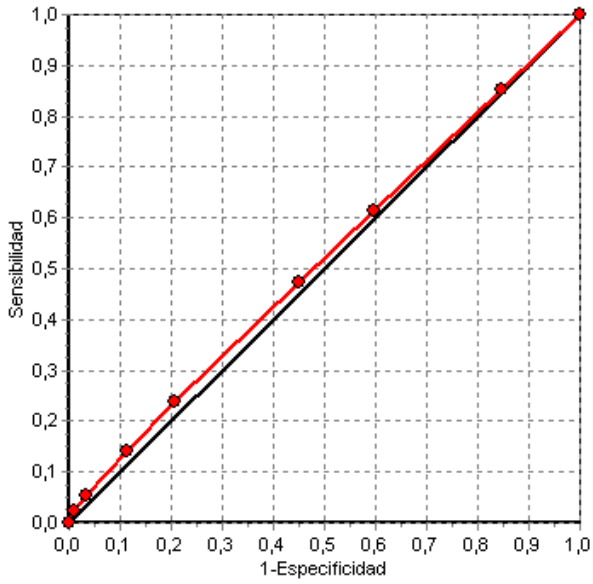
PRUEBA DE BONDAD DE AJUSTE DE HOSMER Y LEMESHOW

Grupos basados en los deciles

Grupo de probabilidad	Respuesta = 0 Valor observado	Respuesta = 0 Valor esperado	Respuesta = 1 Valor observado	Respuesta = 1 Valor esperado
1	464	469,33	483	477,67
2	757	762,63	789	783,37
3	445	439,67	463	468,33
4	729	723,37	772	777,63
5	284	281,99	316	318,01
6	242	244,01	290	287,99
7	99	99,00	172	172,00

Ji-cuadrado	g.l.	Valor p
0,4693	5	0,9932

CURVA ROC



Area ROC	EE	IC(95%)		
0,5180	0,0071	0,5040	0,5319	Delong
	0,0073	0,5037	0,5322	Hanley & McNeil

El modelo ajustado resultó ser el siguiente:

$$p(anticuerpos) = \frac{1}{1 + \exp(-0{,}108 - 0{,}009Z_1 - 0{,}516Z_2 - 0{,}103Z_3 + 0{,}046RH)}$$

La bondad del ajuste fue excelente; nótese la similitud entre valores esperados y observados en el procedimiento de Hosmer y Lemeshow.

Si quisiéramos estimar a través del modelo la tasa de anticuerpos entre sujetos del ESTE que tienen RH negativo, usaríamos la ecuación precedente con Z_1=0, Z_2=1, Z_3=0, RH=2. Esto arroja:

$$p(anticuerpos) = \frac{1}{1 + \exp(-0{,}532)} = 0{,}63$$

Si se computara la tasa de sujetos con anticuerpos en esta subpoblación usando la información de la tabla inicial, tendríamos que computar, simplemente, la razón:

$$\frac{82}{82 + 32} = 0{,}72$$

Ahora bien, si Ud. quiere que el modelo contemple la interacción (INTER) entre ZONA y RH, debe incluir como una variable más el producto de estas dos últimas. Si ZONA pudiera tratarse como si fuera una variable cuantitativa (tal como número de hijos) u ordinal (tal como una escala de satisfacción) la tabla que habría que crear sería la siguiente:

VIRUS	ZONA	RH	INTER	FRECUENCIA
0	1	1	1	445
0	2	1	2	729
0	3	1	3	32
0	4	1	4	242
0	1	2	2	464
0	2	2	4	757
0	3	2	6	67
0	4	2	8	284
1	1	1	1	463
1	2	1	2	772
1	3	1	3	82
1	4	1	4	290
1	1	2	2	483
1	2	2	4	789
1	3	2	6	90
1	4	2	8	316

Nota: Adviértase que, en este caso, la variable INTER no se codifica con valores 1, 2, 3, ..., sino que es simplemente el producto de las otras dos.

Sin embargo, puesto que ZONA ha de tratarse a través de las variables dummy, esto sería incorrecto en este caso. Pero para hacer el ajuste incorporando la interacción de ZONA y RH, no se debe indicar a Epidat que maneje la ZONA a través de variables dummy, sino que deben construirse las 3 variables dummy previamente y luego los tres productos procedentes de éstas con RH. La tabla de contingencia sería como sigue:

VIRUS	Z1	Z2	Z3	RH	Z1-RH	Z2-RH	Z3-RH	FRECUENCIA
0	0	0	0	1	0	0	0	445
0	1	0	0	1	1	0	0	729
0	0	1	0	1	0	1	0	32
0	0	0	1	1	0	0	1	242
0	0	0	0	2	0	0	0	464
0	1	0	0	2	2	0	0	757
0	0	1	0	2	0	2	0	67
0	0	0	1	2	0	0	2	284
1	0	0	0	1	0	0	0	463
1	1	0	0	1	1	0	0	772
1	0	1	0	1	0	1	0	82
1	0	0	1	1	0	0	1	290
1	0	0	0	2	0	0	0	483
1	1	0	0	2	2	0	0	789
1	0	1	0	2	0	2	0	90
1	0	0	1	2	0	0	2	316

El archivo VIRUS1.xls, que se incluye en el paquete de Epidat 3.1, contiene esta tabla de contingencia.

Al realizar el ajuste se obtiene lo siguiente:

Tablas de contingencia : Regresión logística

Archivo de trabajo: C:\Archivos de programa \Epidat 3.1 \Ejemplos \Tablas de contingencia\VIRUS1.xls
Campo que identifica:
  Variable respuesta: VIRUS
  Frecuencias: FRECUENCIA
  Variables explicativas: Z1 Z2 Z3 RH Z1-RH Z2-RH Z3-RH

Nivel de confianza: 95,0%

Variable respuesta:

Valor	N° sujetos
0	3020
1	3285
Total	6305

La sucesión de estimadores ha convergido
N° iteraciones necesarias: 3

-2 ln Verosimilitud inicial: 8729,444680
-2 ln Verosimilitud final : 8705,834318

Cociente de verosimilitud

Estadístico	g.l.	Valor p
23,6104	7	0,0013

Coeficiente de determinación: 0,0036

Variable	Coeficiente	S.E.	Valor de Z	Valor p
Constante	0,039173			
Z1	0,034045	0,187378	0,181693	0,8558
Z2	1,546052	0,470753	3,284208	0,0010
Z3	0,215945	0,242614	0,890078	0,3734
RH	0,000479	0,092912	0,005159	0,9959
Z1-RH	-0,016387	0,117848	-0,139053	0,8894
Z2-RH	-0,645534	0,279461	-2,309920	0,0209
Z3-RH	-0,074655	0,151323	-0,493344	0,6218

Variable	Odds ratio	IC(95,0%)	
Z1	1,034631	0,716621	1,493763
Z2	4,692907	1,865245	11,807229
Z3	1,241034	0,771384	1,996626
RH	1,000479	0,833915	1,200314
Z1-RH	0,983746	0,780856	1,239354
Z2-RH	0,524383	0,303229	0,906831
Z3-RH	0,928064	0,689876	1,248489

PRUEBA DE BONDAD DE AJUSTE DE HOSMER Y LEMESHOW

Grupos basados en los deciles

Grupo de probabilidad	Respuesta = 0 Valor observado	Respuesta = 0 Valor esperado	Respuesta = 1 Valor observado	Respuesta = 1 Valor esperado
1	445	445,00	463	463,00
2	464	464,00	483	483,00
3	757	757,00	789	789,00
4	729	729,00	772	772,00
5	284	284,00	316	316,00
6	242	242,00	290	290,00
7	99	99,02	172	171,98

Ji-cuadrado	g.l.	Valor p
0,0000	5	1,0000

Ejemplo 2

Supóngase que se evalúa la satisfacción con la atención primaria de 1.027 personas mediante la variable SATISF (0- Satisfecho, 1- Insatisfecho) y que la probabilidad de estar insatisfecho se quiere poner en función de r=3 variables, a saber:

RAZA, con $k_1 = 3$ categorías: 1- Negro, 2- Blanco, 3- Mestizo

GÉNERO, con $k_2 = 2$ categorías: 1- Hombre, 2- Mujer

EDAD, con $k_3 = 2$ categorías: 1- Adulto, 2- Anciano

Entonces, se tendrán 2×3×2×2=24 configuraciones y hay que informar, por tanto, las respectivas frecuencias. Esto quiere decir que hay que teclear los datos de una tabla de contingencia de 4 entradas, o prepararla de antemano en EXCEL, Dbase o ACCESS para que el programa la lea automáticamente:

SATISF	RAZA	GENERO	EDAD	FREQ
0	1	1	1	109
0	1	1	2	19
0	1	2	1	54
0	1	2	2	14
0	2	1	1	90
0	2	1	2	8
0	2	2	1	44
0	2	2	2	13
0	3	1	1	84
0	3	1	2	6
0	3	2	1	42
0	3	2	2	13
SATISF	**RAZA**	**GENERO**	**EDAD**	**FREQ**
1	1	1	1	54

1	1	1	2	9
1	1	2	1	27
1	1	2	2	7
1	2	1	1	45
1	2	1	2	2
1	2	2	1	20
1	2	2	2	5
1	3	1	1	211
1	3	1	2	33
1	3	2	1	97
1	3	2	2	21

El archivo SATISF.xls que se incluye en este paquete de programas contiene la tabla con los datos de este ejemplo. Puesto que la variable RAZA no es ordinal, es razonable plantear que sea manejada como una variable *dummy*. Sin embargo, no es menester hacerlo en la tabla de entrada sino que ello se menciona entre las entradas manuales cualquiera sea la vía de comunicar los datos.

Al aplicar el programa a los datos precedentes se obtiene:

```
Tablas de contingencia : Regresión logística

Archivo de trabajo: C:\Archivos de programa \Epidat 3.1 \Ejemplos \Tablas de
contingencia \SATISF.xls
Campo que identifica:
  Variable respuesta: SATISF
  Frecuencias: FREQ
  Variables explicativas: RAZA GENERO EDAD

Nivel de confianza:  95,0%

Variable respuesta:
Valor    N° sujetos
-----    ----------
0          496
1          531
-----    ----------
Total          1027

Variables Dummy:
RAZA
CategoríaRAZA-1    RAZA-2
---------------    ------
        1     0         0
        2     1         0
        3     0         1


La sucesión de estimadores ha convergido
N° iteraciones necesarias: 3

-2 ln Verosimilitud inicial: 1422,531283
```

-2 ln Verosimilitud final : 1261,790448

Cociente de verosimilitud

Estadístico	g.l.	Valor p
160,7408	3	0,0000

Coeficiente de determinación: 0,1522

Variable	Coeficiente	S.E.	Valor de Z	Valor p
Constante	-0,545640			
RAZA-1	-0,061063	0,189369	-0,322453	0,7471
RAZA-2	1,618271	0,158472	10,211732	0,0000
GENERO	-0,128420	0,143426	-0,895378	0,3706
EDAD	0,012738	0,193605	0,065795	0,9475

Variable	Odds ratio	IC(95,0%)	
RAZA-1	0,940764	0,649066	1,363554
RAZA-2	5,044360	3,697556	6,881728
GENERO	0,879484	0,663963	1,164963
EDAD	1,012820	0,693003	1,480231

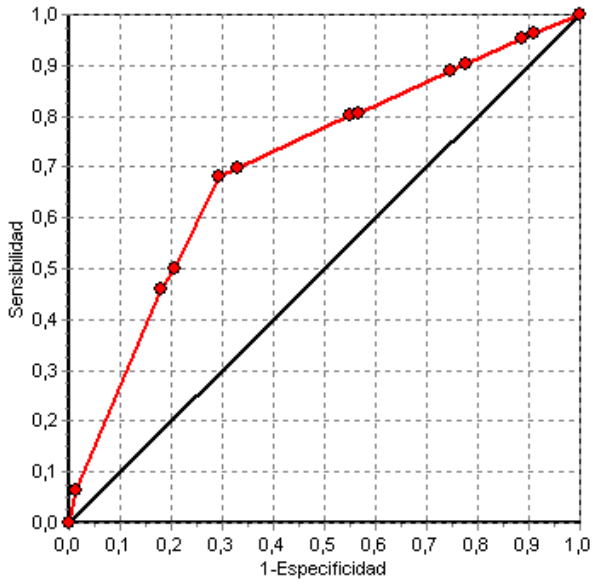
PRUEBA DE BONDAD DE AJUSTE DE HOSMER Y LEMESHOW

Grupos basados en los deciles

Grupo de probabilidad	Respuesta = 0 Valor observado	Respuesta = 0 Valor esperado	Respuesta = 1 Valor observado	Respuesta = 1 Valor esperado
1	111	113,12	52	49,88
2	104	105,26	52	50,74
3	117	114,21	56	58,79
4	19	18,39	9	9,61
5	42	42,25	97	96,75
6	97	92,10	232	236,90
7	6	10,72	33	28,28

Ji-cuadrado	g.l.	Valor p
3,6689	5	0,5980

CURVA ROC



Area ROC	EE	IC(95%)		
0,7032	0,0161	0,6717	0,7347	Delong
	0,0161	0,6717	0,7347	Hanley & McNeil

En este ejemplo, el ajuste es francamente bueno, lo cual se aprecia comparando frecuencias observadas y esperadas y se confirma al obtener una P muy superior a los niveles admitidos convencionalmente para declarar significación. El área bajo la curva ROC en este ejemplo es considerablemente alta, hecho coherente con que las 3 variables incorporadas consiguen una reducción significativa de la lejanía.

Ejemplo 3

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 REGIMEN, 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 Tabla 1.

Tabla 1. Distribución de pacientes según régimen de atención enfermera y condición respecto de la infección.

	Infección	
Régimen	**Sí (1)**	**No (0)**
Nuevo (1)	7	37
Convencional (0)	14	22

OR=0,30

Si a partir de los datos brutos se estima el efecto del régimen de atención de enfermería sobre el hecho de desarrollar una infección, el *odds ratio* resultante es de 0,30 (procedente de computar la llamada *razón de productos cruzados* (7×22)/(14×37).

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 (De Irala, Martínez y Guillén[9]) 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, 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 Tabla 2. Los estimados del *odds ratio* en las dos categorías son de 0,41 y 0,36 respectivamente.

Un método usual para valorar una confusión consiste en comparar de forma directa el estimado bruto del efecto y el estimado de éste una vez controlado el presunto factor de confusión. Para ello se debe obtener una estimación del efecto global a partir de los datos estratificados, mediante una media ponderada de las estimaciones de los efectos por estrato.

Tabla 2. Distribución de pacientes según régimen de atención enfermera, condición respecto de la infección y grupo de edad.

		Infección		
		Sí (1)	**No (0)**	
Edad<40 (1)	**Régimen nuevo (1)**	2	22	**OR_1=0,41**
	Régimen convencional (0)	2	9	
Edad≥40 (2)	**Régimen nuevo (1)**	5	15	**OR_2=0,36**
	Régimen convencional (0)	12	13	

Retomando nuevamente el ejemplo, ¿será posible que el *odds ratio* total de 0,30 refleje, en alguna dimensión, el efecto confusor que pudiera tener la edad en la relación entre el régimen de atención de enfermería y la infección?

Dentro de cada categoría o estrato formado por los dos grupos de edad (menores de 40 y no menores de 40) se puede calcular el *odds ratio* como única medida de la asociación entre el régimen y la infección. Una medida única global se obtiene como un promedio ponderado de los odds ratio dentro de los estratos. Esto es exactamente lo que provee el *odds ratio* de Mantel Haenszel que, en este caso, como puede corroborarse a través del análisis de tablas 2x2 estratificadas en este mismo módulo, arroja el valor 0,37.

Al usar el submódulo de regresión logística 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 (REGIMEN), y luego los que se producen cuando se adiciona la variable EDAD.

Caso en que solo se incluye la variable régimen como independiente:

```
Tablas de contingencia : Regresión logística

Archivo de trabajo: F:\Xestion e Calidade\Informacion Saude
Publica\Epidat\Epidat 3.1\Ayuda\Ejemplos\Tablas\CADERA.xls
Campo que identifica:
  Variable respuesta: INFEC
  Frecuencias: FREQ
  Variables explicativas: REGIMEN


Variable     Coeficiente      S.E.       Valor de Z    Valor p
------------ -----------    --------     ----------    -------
   Constante  -0,451473
     REGIMEN  -1,210425     0,535158      -2,261807     0,0237
```

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

Caso en que se incluyen régimen y edad como variables independientes:

```
Tablas de contingencia : Regresión logística

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

Nivel de confianza:  95,0%

Variable      Coeficiente      S.E.   Valor de Z    Valor p
------------  -----------    --------  ----------    -------
   Constante  -2,759493
     REGIMEN  -0,974758     0,554901  -1,756635     0,0790
        EDAD   1,332184     0,622533   2,139941     0,0324


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

Varias cosas procede subrayar a partir de los ejemplos desarrollados. Sucintamente, cabe llamar la atención sobre las siguientes:

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 1. 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 REGIMEN 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 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 0,298 con 0,377 lo cual permite pensar que sí hay efecto confusor. El OR correspondiente a REGIMEN 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. Por ejemplo, en el Ejemplo 2, si se quiere saber cuál es la probabilidad de que un sujeto esté insatisfecho, hay que aplicar la fórmula siguiente:

$$P(SATISF = 1) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 RAZA_1 - \beta_2 RAZA_2 - \beta_3 GENERO - \beta_4 EDAD)}$$

donde:

$\beta_0 = -0{,}5456 \quad \beta_1 = -0{,}0611 \quad \beta_2 = 1{,}6183 \quad \beta_3 = -0{,}1284 \quad \beta_4 = 0{,}0127$

Si se trata de una mujer mestiza y mayor de 40 años, los valores de las variables son: $RAZA_1 = 0$, $RAZA_2 = 1$, $GENERO = 2$ y $EDAD = 2$, lo cual arroja:

$$P(SATISF = 1) = 0{,}699$$

Para un hombre blanco de menos de 40 años, el perfil de entrada sería: $RAZA_1 = 1$, $RAZA_2 = 0$, $GENERO = 1$ y $EDAD = 1$ y, al aplicar la fórmula, se tendría:

$$P(SATISF = 1) = 0{,}327$$

Nota: El ejemplo que se acaba de desarrollar (y varios más de este epígrafe) ha sido tomado del libro "Regresión Logística" de Silva y Barroso[10], donde el usuario de Epidat hallará muchos más detalles conceptuales y prácticos.

Recomendaciones

- 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.

BIBLIOGRAFÍA

1. Cornfield J, Gordon T, Smith WN. Quantal response curves for experimentally uncontroled variables. *Bulletin of the International Statistical Institute* 1961; 38: 97-115.

2. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. *Biometrika* 1967; S4: 167-79.

3. Silva LC, Pérez C, Cuellar I. Uso de la estadística en la investigación de salud contemporánea. *Gac Sanit* 1994; 9(48): 189-95.

4. Levy PS, Stolte K. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Stat Methods Med Res* 2000; 9: 41-55.

5. Silva LC. *Excursión a la regresión logística en ciencias de la salud*. Madrid: Díaz de Santos; 1995.

6. Jones RH. Probability estimation using a multinomial logistic function. *Journal of Statistical and Computer Simulation* 1975; 3: 315-29.

7. Hosmer DW Jr, Lemeshow S. *Applied Logistic Regression*. New York: John Wiley & Sons; 1989.

8. Mittlböck M, Schemper M. Explained variation for logistic regression. *Stat Med* 1996; 15: 1987-97.

9. De Irala J, Martínez MA, Guillén F. ¿Qué es una variable de confusión? *Med Clin* (Barc) 2001; 117: 377-85.

10. Silva LC, Barroso J. Regresión Logística. Cuaderno 27. Madrid: La Muralla; 2004.