Departamento de Estadística
Universidad Carlos III de Madrid

BIOESTADISTICA (55 - 10536)

Estudios de casos y controles

CONCEPTOS CLAVE

1) Características del diseño en un estudio de casos y controles.

2) Elección del tamaño muestral.

3) Estrategias para el análisis de estudios de casos y controles: la Razón de Odds como medida de asociación.

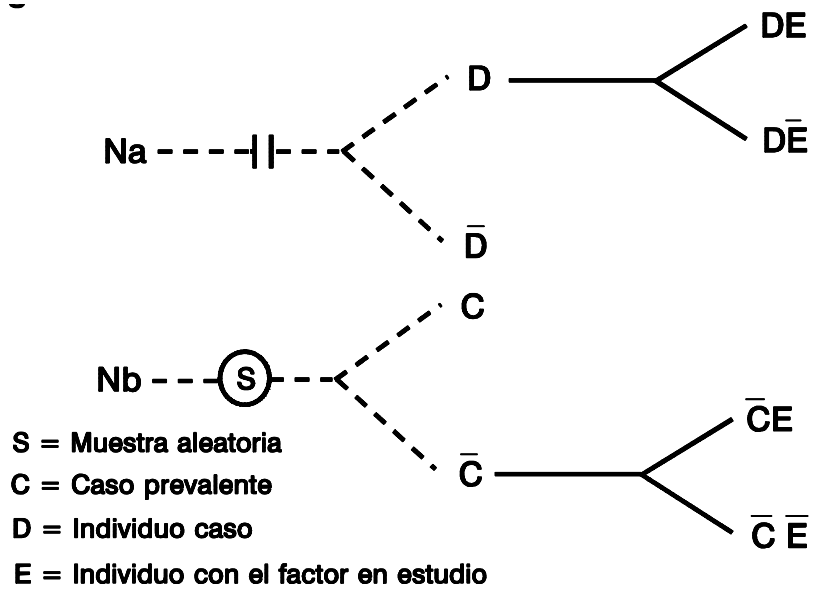
4) Análisis estratificados y sesgo de confusión: el método de Mantel-Haenszel

1. INTRODUCCION

Un **estudio de casos y controles** comienza con la identificación de personas con la enfermedad u otro tipo de característica y un grupo adecuado de personas de control (comparación, referencia) sin la enfermedad. Se examinan las relaciones entre un atributo y la enfermedad, mediante la comparación de los enfermos con los sanos, con respecto a la frecuencia con que el atributo se halla presente (o si es de carácter cuantitativo, qué niveles alcanza) en cada uno de los grupos.

Un estudio de esta naturaleza se puede calificar de *retrospectivo*, ya que comienza después del inicio de la enfermedad y busca en el pasado los supuestos factores causales. Los casos y controles de un estudio caso control pueden acumularse *prospectivamente*; en ese caso, a medida que se diagnostican, se incorporan al estudio.

En la figura siguiente muestra una representación gráfica de un estudio de casos y controles.



Tal y como indica la figura, el investigador selecciona por separado los grupos de poblaciones de casos y controles disponibles, lo que lo diferencia de los estudios transversales en los cuales la selección se realiza de una única población. Otra diferencia es que un estudio de caso control puede incluir casos incidentes, esto es, los casos y controles pueden ser tomados *prospectivamente* en el tiempo.

2. DISEÑO DE UN ESTUDIO DE CASOS Y CONTROLES

Los principales puntos metodológicos a tener en cuenta para la ejecución de un estudio de casos y controles son:

a) Definición precisa de la variable dependiente.

b) Definición de las variables independientes o de la exposición de interés.

c) Fuente y criterios de selección de los casos.

La manera "ideal" de selección de los casos se plantea que sea: en un área geográfica limitada, todos los casos que aparezcan en un tiempo determinado, o seleccionar una muestra representativa de éstos. Sin embargo, existen limitantes para proceder de esta manera: la necesidad de registros, y aún teniéndolos, no podemos asegurar que se captan todos los casos. Tomar sólo casos incidentes puede ser un problema si la enfermedad es poco frecuente.

d) Definición, fuente y criterios de selección de los controles.

El grupo control debe estar integrado por individuos que no tienen la enfermedad, por tanto, debe emplearse procedimientos diagnósticos similares a los utilizados por los casos. En relación a la fuente, se deben tomar los controles con el mismo criterio de selección que los casos. La idea fundamental que debe seguirse es la de establecer la mayor comparabilidad posible entre ambos grupos, con relación a los factores distintos de la exposición en estudio.

e) Obtención de la información.

f) Determinación del número de casos y controles a incluir en el estudio.

g) Determinar el tipo de análisis epidemiológico y estadístico de los datos.

Nos centraremos en estos dos últimos puntos.

Veamos a continuación como se calculan los tamaños de muestras en los estudios de caso control para distintas situaciones.

1) **Si el objetivo es estimar la Razón de Odds (OR), o razón de disparidad, con una precisión relativa especificada se deberá "conocer":**

   a) Dos de los siguientes elementos:
      - Probabilidad anticipada de la exposición al factor en individuos enfermos: $\boldsymbol{P_1} = \dfrac{a}{a+b}$
      - Probabilidad anticipada de la exposición en individuos sanos: $\boldsymbol{P_2} = \dfrac{c}{c+d}$
      - Razón de Odds anticipado: ***OR***

   b) Nivel de confianza: **100(1-α)%**

   c) Precisión relativa: $\boldsymbol{\varepsilon}$

Notemos que si conocemos el valor de P_1 y OR, podemos calcular, P_2 mediante: $P_2 = \frac{P_1}{OR(1-P_1)+P_1}$, análogamente podemos obtener OR por: $OR = \frac{P_1/(1-P_1)}{P_2/(1-P_2)}$ y P_1 por: $P_1 = \frac{P_2}{(1-P_2)/OR + P_2}$.

Se utiliza en este caso la siguiente fórmula para obtener el tamaño muestral:

$$n = z^2_{1-\alpha/2} \frac{1/[P_1(1-P_1)] + 1/[P_2(1-P_2)]}{\ln^2(1-\varepsilon)}$$

Ejemplo 1: En una región donde el cólera es un problema grave de salud, se supone que el 30% de la población utilizan agua contaminada. Se desea realizar un estudio para estimar el OR con una precisión relativa del 25% (OR anticipado igual a 2) con un 95% de confianza. ¿Cuál es el tamaño de la muestra en los casos de cólera y en los controles?

Tenemos que $P_2 = 0.3$, $OR = 2$, $\varepsilon = 0.25$ y $\alpha = 0.05$. Calculamos primero P_1 mediante $P_1 = \frac{P_2}{(1-P_2)/OR + P_2} = \frac{0.3}{0.7/2 + 0.3} \approx 0.46$ y sustituyendo en la fórmula obtenemos n = 408 individuos en cada grupo:

$$n = 3.8416 \frac{1/[0.46 \times 0.54] + 1/[0.3 \times 0.7]}{\ln^2(0.75)} \approx 3.8416 \frac{4.03 + 4.76}{0.083} \approx 408 .$$

2) **Si el objetivo es probar que la Razón de Odds (OR) o razón de disparidad difiere significativamente de 1, se deberá "conocer":**

   a) Hipótesis nula: $\mathbf{H_0: OR = 1}$

   b) Dos de los siguientes elementos:

      - Probabilidad anticipada de la exposición al factor en individuos enfermos $\boldsymbol{P_1}$
      - Probabilidad anticipada de la exposición en individuos sanos $\boldsymbol{P_2}$
      - Razón de Odds anticipado $\boldsymbol{OR_a}$

   c) Nivel de confianza: **100(1-α)%**

   d) Potencia del test: **100(1-β)%**

   e) Hipótesis alternativa: $\mathbf{H_a: OR_a \neq 1}$

Se utiliza en este caso la siguiente fórmula:

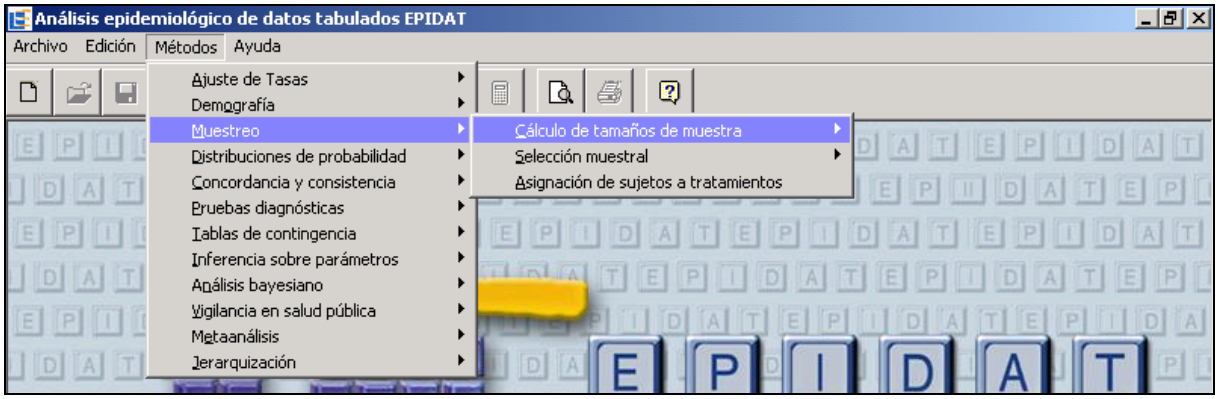
$$n' = \frac{[z_{1-\alpha/2}\sqrt{(r+1)P_M(1-P_M)} - z_{1-\beta}\sqrt{rP_1(1-P_1)+P_2(1-P_2)}]^2}{r(P_1-P_2)^2},$$

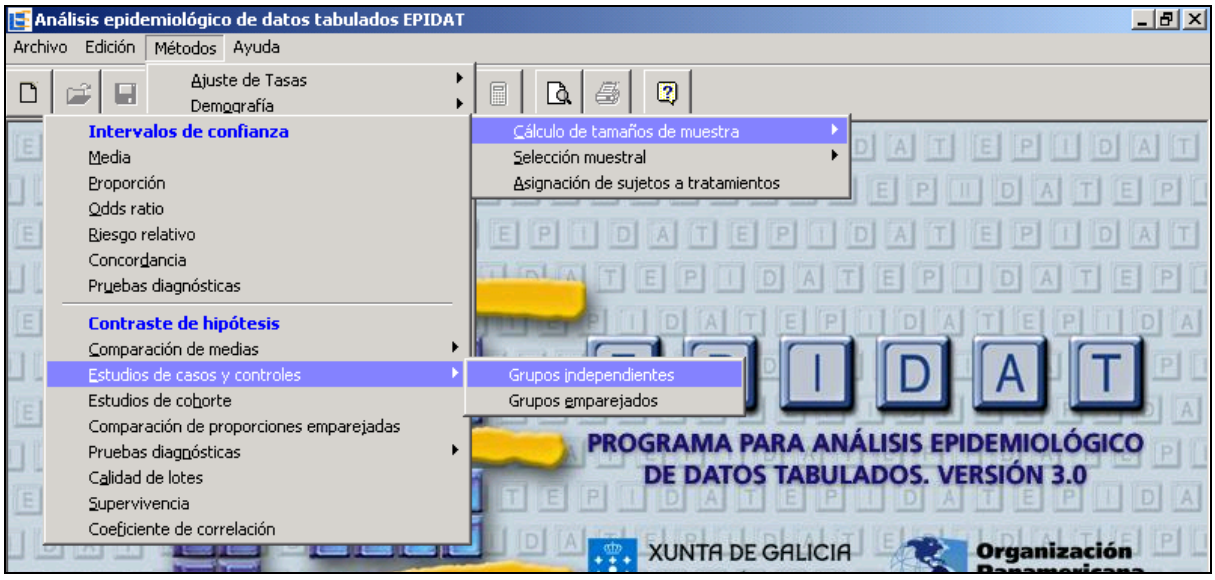
donde $P_M = (P_1 + rP_2)/(r+1)$.

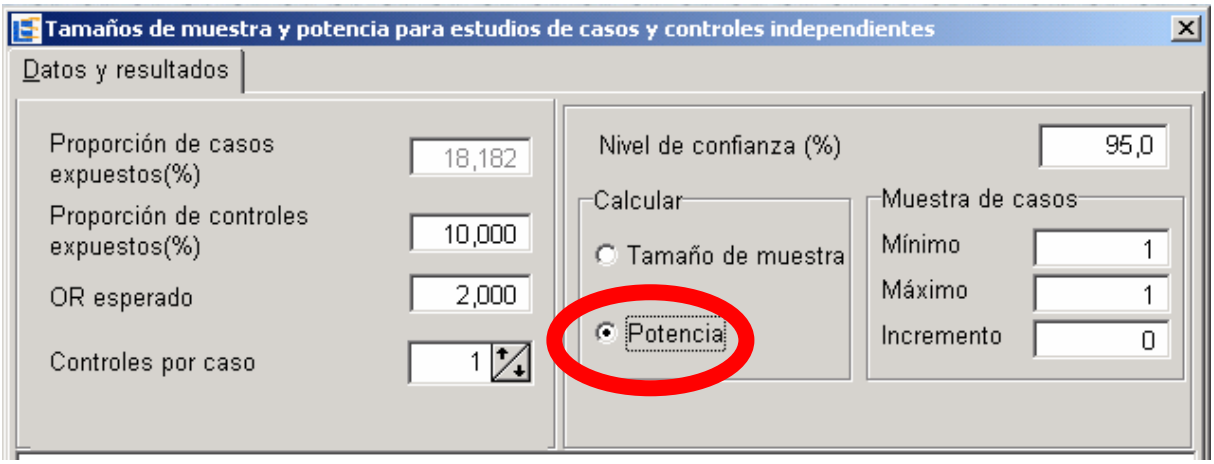
La mayoría de software estadístico, como EpiDat, propone la corrección de Yates para el cálculo del tamaño muestral:

$$n = \frac{n'}{4}\left[1+\sqrt{1+\frac{2(r+1)}{n'r|P_2-P_1|}}\right]^2$$

Ejemplo 2: A continuación se muestra la salida del programa EpiDat, para el cálculo del tamaño muestral suponiendo que $P_2 = 0.1$ y OR = 2.0, $\alpha = 0.05$, $\beta = 0.2$ y r = 1, o sea un control por cada caso. El tamaño de muestra calculado es n = 307 en ambos grupos.





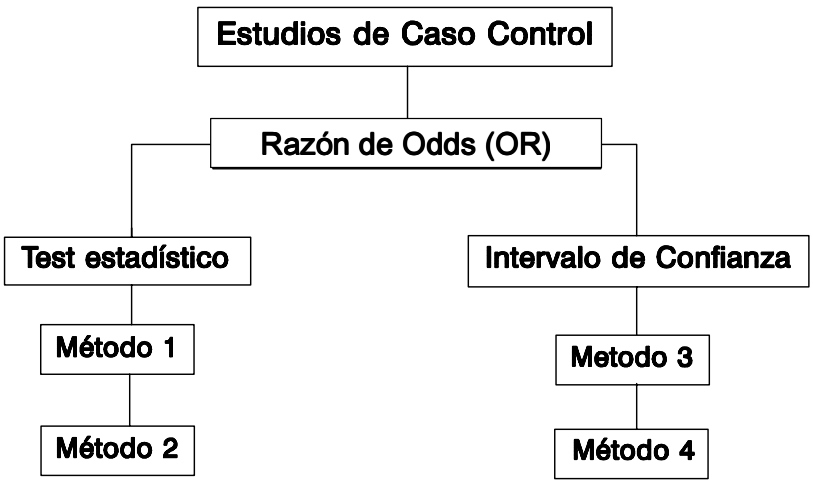



```
[  1] Tamaños de muestra y potencia para estudios de casos y controles independientes

 Proporción de casos expuestos:    18,182%
Proporción de controles expuestos: 10,000%
                  OR esperado:     2,000
             Controles por caso:         1
             Nivel de confianza:      95,0%


                                                  Tamaño de muestra
    Potencia (%)     Ji-cuadrado                      Casos     Controles
---------------     -------------------------  ----------     ----------
            80,0     Sin corrección                      283            283
                     Corrección de Yates                 307            307
```

3. ANÁLISIS DE ESTUDIOS DE CASOS Y CONTROLES

3.1. Plan de análisis estadístico para estudios de casos y controles

Un posible esquema de plan de análisis para estudios de casos y controles es el siguiente:



A continuación se presentan los principales métodos de análisis, a partir de la siguiente disposición de los resultados de un estudio de casos y controles:

	Casos	Controles	Total
Expuestos	a	b	N_1
No Expuestos	c	d	N_0
Total	M_1	M_0	T

***Método 1*: Modelo hipergeométrico**

La probabilidad de obtener ***a*** o más expuestos en el grupo de los casos está dada por:

$$\Pr(K \geq a) = \sum_{k=a}^{\min(M_1,N_1)} \frac{\binom{N_1}{k}\binom{N_0}{M_1 - k}}{\binom{T}{M_1}}$$

La regla de decisión es: rechazar H_0 si $\Pr(K \geq a) \leq \alpha$.

Este método se debe utilizar cuando la frecuencia esperada de alguna de las casillas es menor que 5.

***Método 2*: Aproximación normal a la hipergeométrica**

Cuando el valor esperado de todas las casillas es mayor que 5, se puede utilizar una aproximación normal de la distribución hipergeométrica con: $\mu = \frac{M_1 N_1}{T}$, y $\sigma = \sqrt{\frac{N_1 N_0 M_1 M_0}{T^2(T-1)}}$ con el siguiente test estadístico: $z = \frac{a - \mu}{\sigma}$, y como regla de decisión: $z \geq z_\alpha$. Este procedimiento es equivalente al estadístico χ^2 de Mantel-Haenszel.

Ejemplo 3: Resultados de en un estudio de casos y controles para evaluar el efecto del consumo de cigarrillos (exposición) sobre el cáncer de bucofaringe (casos).

	Casos	Controles	Total
Expuestos	352	228	580
No Expuestos	48	122	170
Total	400	350	750

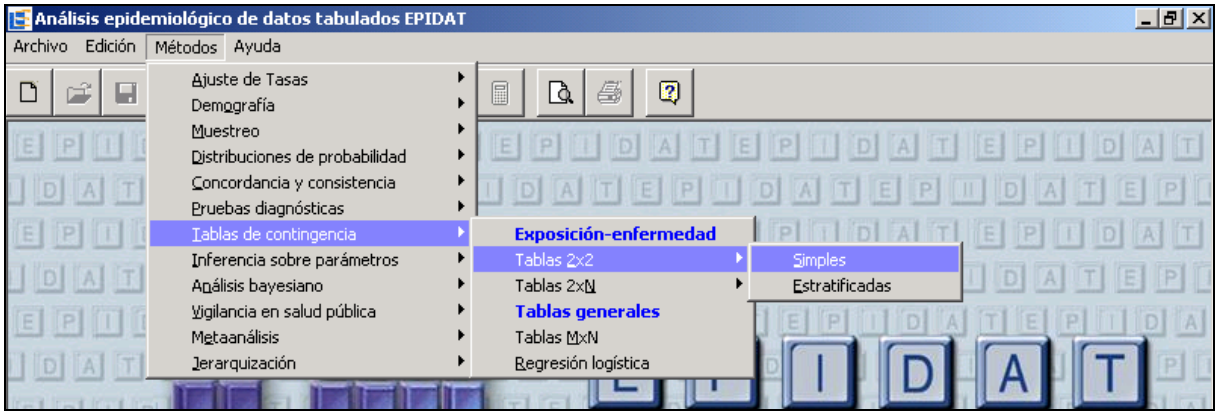
Si aplicamos el Método 2, entonces:

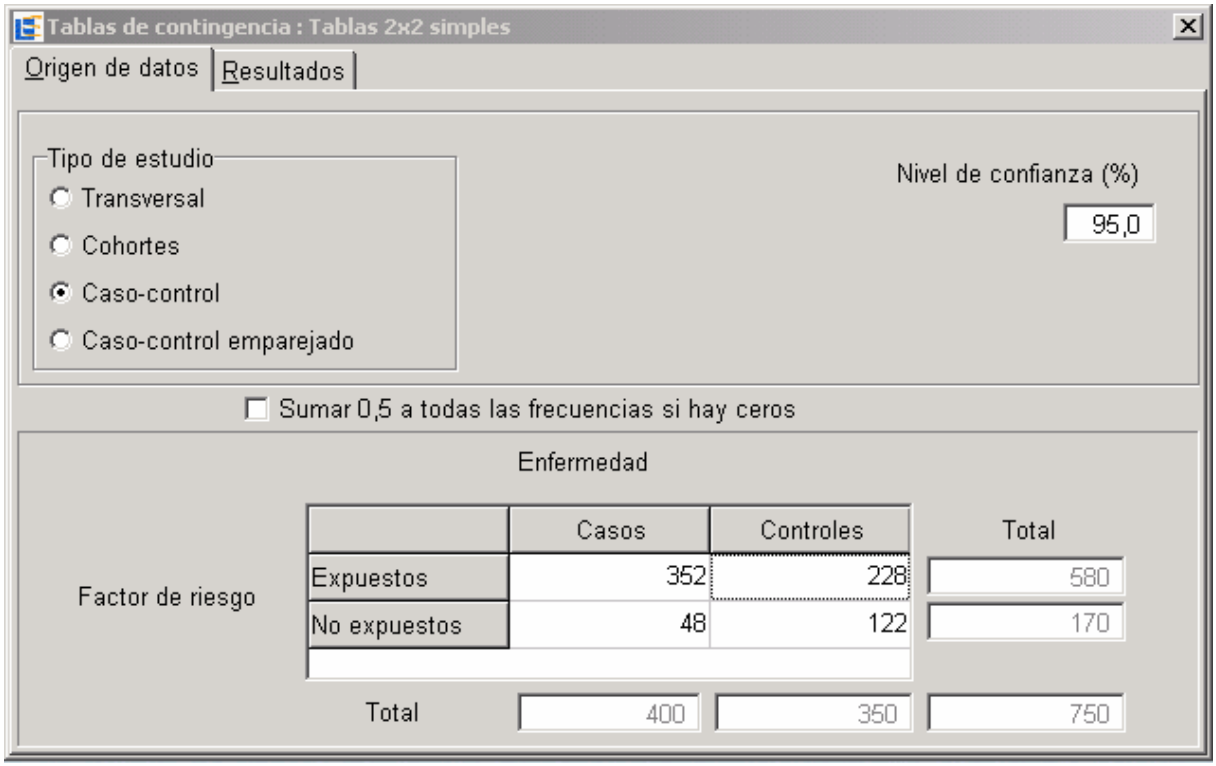
$$\mu = \frac{M_1 N_1}{T} = \frac{400 \times 580}{750} \approx 309.33$$

$$\sigma = \sqrt{\frac{N_1 N_0 M_1 M_0}{T^2 (T-1)}} = \sqrt{\frac{580 \times 170 \times 400 \times 350}{750^2 \times 749}} \approx 5.72$$

luego: $z = \frac{a - \mu}{\sigma} = \frac{352 - 309.33}{5.72} \approx 7.46$ que es mayor que z = 1.96, por lo tanto rechazamos la hipótesis nula de no asociación.

Veamos el resultado obtenido por el programa EpiDat:





```
[  2] Tablas de contingencia : Tablas 2x2 simples

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

Tabla
                 Casos       Controles         Total
------------   --------      --------       --------
Expuestos           352           228            580
No expuestos         48           122            170
------------   --------      --------       --------
Total               400           350            750

                                          Estimación            IC(95,0%)
----------------------------------------    ----------     ---------      ---------
Proporción de casos expuestos                 0,880000         -              -
Proporción de controles expuestos             0,651429         -              -
Odds ratio                                    3,923977      2,701761      5,699094 (Woolf)
                                                            2,704889      5,691788 (Cornfield)
----------------------------------------    ----------     ---------      ---------
Fracción atribuible en expuestos              0,745156      0,629871      0,824534
Fracción atribuible poblacional               0,655738      0,534569      0,745362


Prueba Ji-cuadrado de asociación            Estadístico      Valor p
----------------------------------------     ---------     ---------
Sin corrección                                 55,6360        0,0000
Corrección de Yates                            54,3397        0,0000


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

Notemos que $z^2 = 7.46^2 \approx 55.65$ que es similar a $\chi^2_{M\text{-}H} = 55.56$. Luego, de nuevo rechazamos la hipótesis nula de no asociación ($p<0.0001$).

Asimismo, recordemos del Tema 3, que la interpretación de este resultado en términos epidemiológicos vendría dada como que existe una asociación entre el consumo de cigarrillos y el cáncer de bucofaringe, donde el riesgo de desarrollar cáncer de bucofaringe en fumadores (expuestos) es 3.92 veces más elevado que el de los no fumadores (no expuestos).

***Método 3*: Intervalo de confianza aproximado para OR, Método de Woolf**

Aplicando una transformación logarítmica se obtiene: $\ln(OR) \pm z_{1-\alpha}\text{e.e.}(\ln(OR))$, donde $\text{e.e.}(\ln(OR)) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$. Con los datos del ejemplo anterior, obtenemos: $\text{e.e.}(\ln(OR)) = \sqrt{\frac{1}{352}+\frac{1}{228}+\frac{1}{48}+\frac{1}{122}} \approx 0.19$ y el intervalo de confianza del 95% está dado por: $\exp(\ln(3.92) - 1.96 \times 0.19) = 2.69$ y $\exp(\ln(3.92) + 1.96 \times 0.19) = 5.70$, muy similares de nuevo a los calculados por el programa EpiDat.

Cuando el tamaño de la muestra es inferior a 30 debe introducirse la siguiente corrección:

$$OR = \frac{(a+0.5)(d+0.5)}{(c+0.5)(b+0.5)}, \text{ y } \text{e.e.}(\ln(OR)) = \sqrt{\frac{1}{a+0.5}+\frac{1}{b+0.5}+\frac{1}{c+0.5}+\frac{1}{d+0.5}}$$

***Método 4*: Intervalo de confianza aproximado para el OR, basado en el estadístico χ^2**

Se utiliza la siguiente fórmula: $OR^{1\pm z_{1-\alpha}/\chi}$, donde χ^2 es valor del test chi-cuadrado sin corrección de Yates. En el ejemplo anterior χ^2=55.64, de donde obtenemos: $OR^{1-1.96/\sqrt{55.64}} = 2.74$, y $OR^{1+1.96/\sqrt{55.64}} = 5.61$. El programa EpiDat utiliza para este cálculo el método de Cornfield.

4. ESTRATIFICACION Y SESGO DE CONFUSION

Imaginemos que los datos correspondientes a **una población de origen, a** partir de la cual se quiere realizar un estudio de casos y controles, vienen definidos en la siguiente tabla:

	Hombres		Mujeres		Totales	
	Exp.	No Exp.	Exp.	No Exp.	Exp.	No Exp.
Casos	999	20	111	180	1110	200
Controles	89001	9980	9889	89820	98890	99800
Totales	90000	10000	10000	90000	100000	100000

Donde: $OR_H = 5.6$ $OR_M = 5.6$ $OR_{Total} = 5.6$

Retomando los conocimientos de los Temas 1 y 2, observamos en esta población una asociación directa entre la exposición y la enfermedad, donde los expuestos tiene un riesgo de desarrollar la enfermedad 5.6 veces superior a los no expuestos. En esta población podemos decir no existe sesgo confusión debido al sexo en la asociación entre la exposición y la enfermedad, dado que $OR_H = OR_M = OR_{Total}$.

Pero, veamos ahora un como queda el estudio de casos y controles, **seleccionando en la muestra del estudio todos los casos y una muestra aleatoria de controles estratificados** a partir de la variable sexo, extraído de esta población de origen, donde se obtienen los resultados siguientes:

	Hombres		Mujeres		Totales	
	Exp.	No Exp.	Exp.	No Exp.	Exp.	No Exp.
Casos	999	20	111	180	1110	200
Controles	916	103	29	262	945	365
Totales	1915	123	140	442	2055	565

$OR_H = 5.6$ $OR_M = 5.6$ $OR_{Total} = 2.1$

Nótese que en este caso la estratificación por sexos introduce un factor, o sesgo, de confusión ($OR_H = OR_M \neq OR_{Total}$). De hecho cuando la variable de estratificación no está relacionada con la enfermedad, pero si con la exposición en un estudio de casos y controles, se convertirá en un factor de confusión al estratificar los resultados por esa variable.

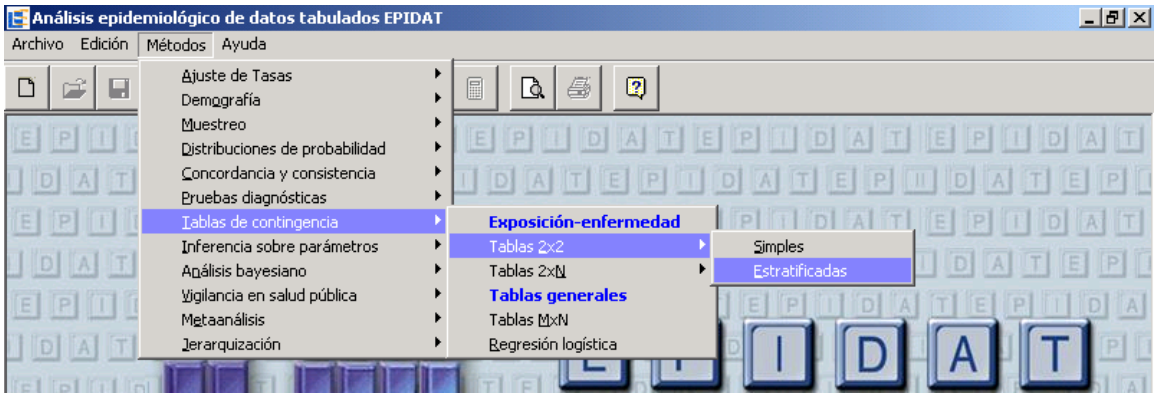
En esta situación debe utilizarse el ***OR* de Mantel-Haenszel**. Para calcular el OR de Mantel-Haenszel, se construye una tabla 2×2 para las parejas caso/control, que en el estudio de un control por cada caso debe corresponder a uno de estos patrones:

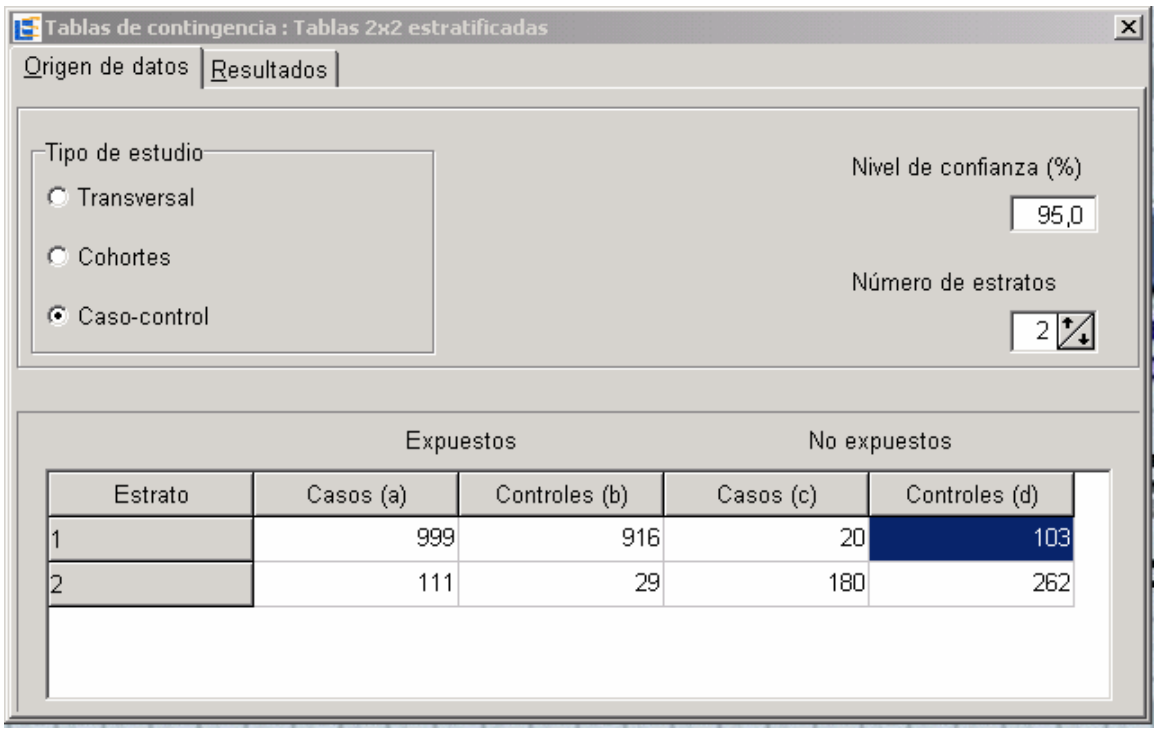
	Pareja A		Pareja B		Pareja C		Pareja D	
	E+	E-	E+	E-	E+	E-	E+	E-
Caso	1	0	1	0	0	1	0	1
Control	1	0	0	1	1	0	0	1

E+: Expuestos, E-: No Expuestos

Así, el *OR* de Mantel-Haenszel se obtiene por: $OR_{M\text{-}H} = \dfrac{\sum a_i d_i / T_i}{\sum c_i b_i / T_i}$, en estudios de un control por caso puede calcularse por $OR_{M\text{-}H} = T_{10}/T_{01}$ donde T_{10} es el número de tablas del tipo **B** y T_{01} es el número de tablas del tipo **C**.

Veamos el cálculo de $OR_{M\text{-}H}$ por el programa EpiDat.





```
[  3] 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                   1110            945           2055
No expuestos                 200            365            565
---------------       ----------     ----------     ----------
Total                       1310           1310           2620

ODDS RATIO (OR)
Estrato                       OR            IC(95,0%)
---------------       ----------     ----------     ----------
1                       5,616648       3,450400       9,142923     (Woolf)
2                       5,571264       3,550126       8,743065     (Woolf)
---------------       ----------     ----------     ----------
Cruda                   2,143651       1,767362       2,600055     (Woolf)
Combinada (M-H)         5,593982       4,014059       7,795757
Ponderada               5,592142       4,016968       7,784989

Prueba de homogeneidad
                     Ji-cuadrado            gl        Valor p
---------------       ----------     ----------     ----------
Combinada (M-H)           0,0006              1         0,9808
Ponderada                 0,0006              1         0,9809

PRUEBA DE ASOCIACIÓN DE MANTEL-HAENSZEL
    Ji-cuadrado               gl        Valor p
---------------       ----------     ----------
       122,5582                1         0,0000
```

Obtenemos el OR_{Total} = 2.14 (sin considerar los estratos), mientras que $OR_{M\text{-}H}$ = 5.59 (considerando los estratos) es muy similar a la OR de la población de referencia, y por lo tanto libre de sesgo de confusión.