Departamento de Estadística
Universidad Carlos III de Madrid

BIOESTADISTICA (55 - 10536)

Estudios de cohortes

CONCEPTOS CLAVE

1) Características del diseño en un estudio de cohortes.

2) Elección del tamaño muestral.

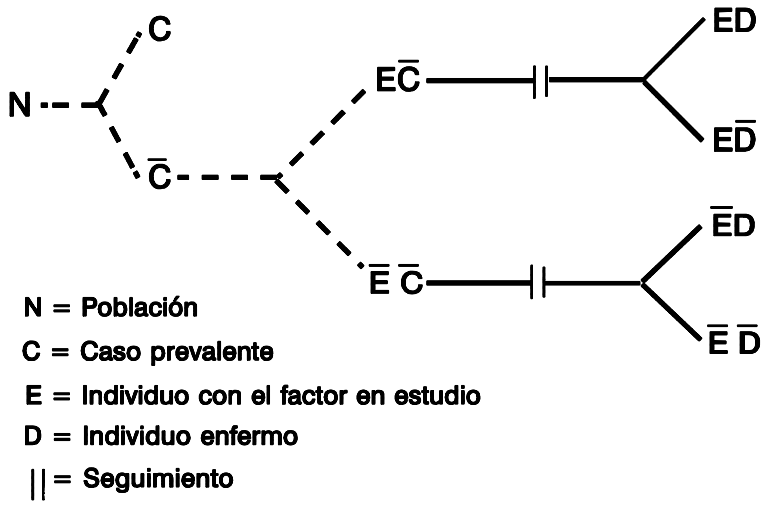
3) Estrategias para el análisis de estudios de cohortes: elección de Razón de Densidades de Incidencia o Incidencia acumulada como medida de asociación.

1. INTRODUCCIÓN

Un **estudio de cohortes** es aquel en el que se pueden identificar subclases de una población determinada, que están, han estado o pueden estar en el futuro expuestas o no expuestas, o expuestas en diferente grado a un factor o factores que supuestamente pueden influir sobre la probabilidad de aparición de una determinada enfermedad, u otro tipo de atributo.

Los términos alternativos de estudio de cohortes, es decir, estudio de seguimiento, longitudinal o prospectivo, describen una característica esencial del método, a saber, la observación de la población durante un número de años-persona como para estimar la tasa de incidencia en cada subclase de la población sometida a seguimiento.

En la Figura siguiente se muestra una representación gráfica de un estudio de cohortes.



Obsérvese que, a diferencia con un estudio de transversal, o de prevalencia, queda registrada formalmente la incidencia de una enfermedad y la secuencia temporal de causa a efecto queda necesariamente determinada en un estudio de este tipo.

Después del proceso se selección, todos los sujetos participantes son examinados acerca de la enfermedad descartándose los casos prevalentes, se determina su nivel actual o pasado del factor en estudio y otras variables de interés y son objeto de seguimiento durante el período que dura el estudio.

2. DISEÑO DE UN ESTUDIO DE COHORTES

Los principales puntos metodológicos a considerar en el diseño de un estudio de cohortes son:

a) Planteamiento de una hipótesis en forma precisa y operacional.

b) Definición de la variable independiente y de la exposición a la misma.

c) Definición y validación de los instrumentos destinados a medir la exposición y los efectos de interés.

d) Fuente y criterios de elección de las cohortes que se van a comparar.

Si la exposición al factor de riesgo objeto de estudio es relativamente común (consumo de cigarrillos, uso de determinados fármacos), las cohortes pueden obtenerse mediante el muestreo de la población. Si la exposición no es frecuente, se deben buscar grupos muy expuestos (grupos ocupacionales expuestos a riesgos industriales)

e) Obtención de la información.

El objetivo fundamental en este punto es conseguir información no sesgada de los grupos en estudio. La posibilidad de sesgo puede deberse a:

- Exámenes más minuciosos en la cohorte expuesta.
- Determinaciones de la exposición más acuciosas en el grupo expuesto.
- No considerar el cambio en los niveles de exposición de los individuos.
- No considerar correctamente los tiempos que han permanecido los individuos dentro de cada cohorte.

f) Determinación del tamaño muestral.

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

Nos centraremos en estos últimos puntos.

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

1) Si el objetivo es estimar el riesgo relativo (RR) con una precisión relativa especificada se deberá "conocer":

a) Dos de los siguientes elementos:

- Probabilidad anticipada de enfermar en personas expuestas al factor de interés: $\mathbf{P_1}$
- Probabilidad anticipada de enfermar en personas no expuestas al factor de interés: $\mathbf{P_2}$
- Riesgo Relativo anticipado: **RR**

b) Nivel de confianza: $\mathbf{100(1-\alpha)\%}$

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

Notemos que en a) dada la relación $RR = P_1/P_2$ siempre que tengamos dos de los tres elementos podremos tener una estimación del tercero. Se utiliza entonces la siguiente fórmula:

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

Ejemplo 1: Supongamos que queremos estudiar la posible asociación de un factor ambiental en la incidencia de una enfermedad respiratoria específica, si deseamos una precisión del 10%, tenemos dos poblaciones semejantes una no expuesta con $P_2 = 0.2$ y una expuesta con $P_1 = 0.4$.

Sustituyendo en la fórmula obtenemos:

$$n = 1.96^2 \frac{[(1 - 0.2)/0.2 + (1 - 0.4)/0.4]}{\ln^2(1 - 0.1)} = 3.8416 \frac{4 + 1.5}{\ln^2(0.8)} \approx 1904.$$

2) **Si el objetivo es probar que el RR es estadísticamente diferente de 1 se deberá "conocer":**

a) Hipótesis nula H_0**: $RR_0 = 1$**

b) Dos de los siguientes elementos:

- Probabilidad anticipada de enfermar en personas expuestas al factor de interés: **P_1**
- Probabilidad anticipada de enfermar en personas no expuestas al factor de interés: **P_2**
- Riesgo Relativo anticipado: **RR_a**

b) Nivel de confianza: **100(1-α)%**

c) Potencia del test: **100(1-ß)%**

d) Hipótesis alternativa: **$RR_a \neq RR_0$**

e) Cantidad de no expuestos por cada expuesto: **r**

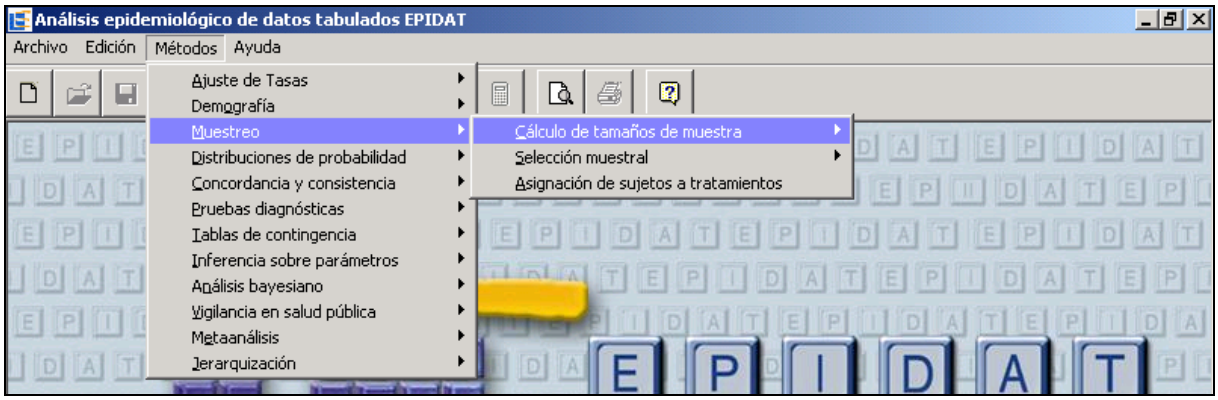
Se utiliza entonces la siguiente fórmula:

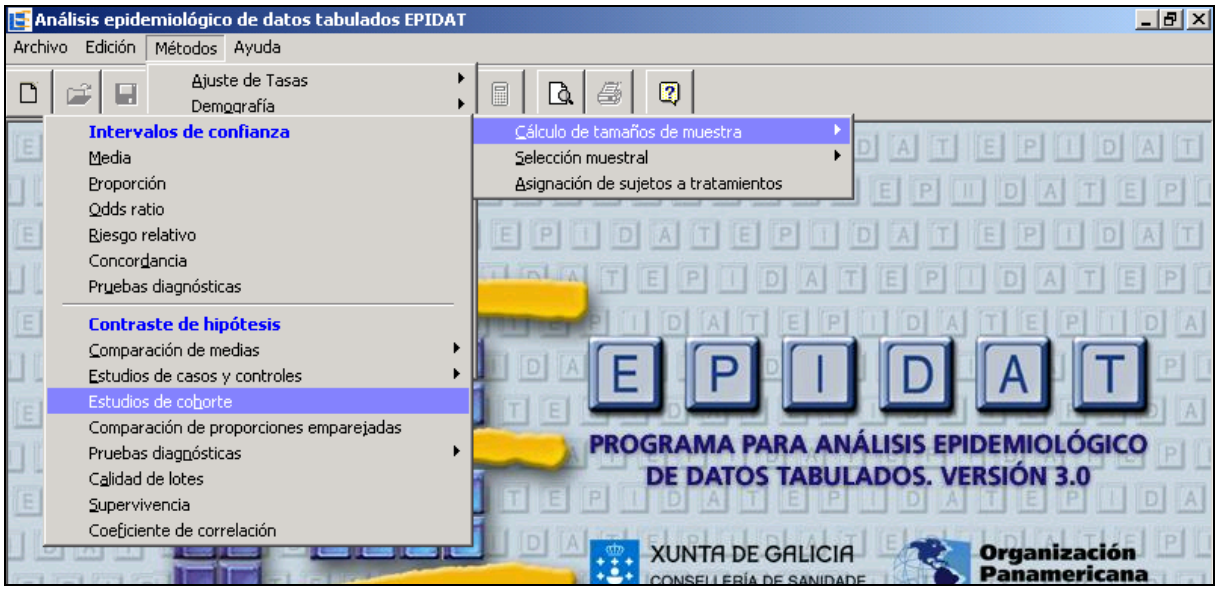
$$n' = \frac{[z_{1-\alpha/2}\sqrt{(r+1)P_M(1 - P_M)} - z_{1-\beta}\sqrt{rP_1(1 - P_1) + P_2(1 - P_2)}]^2}{r(P_1 - P_2)^2}$$

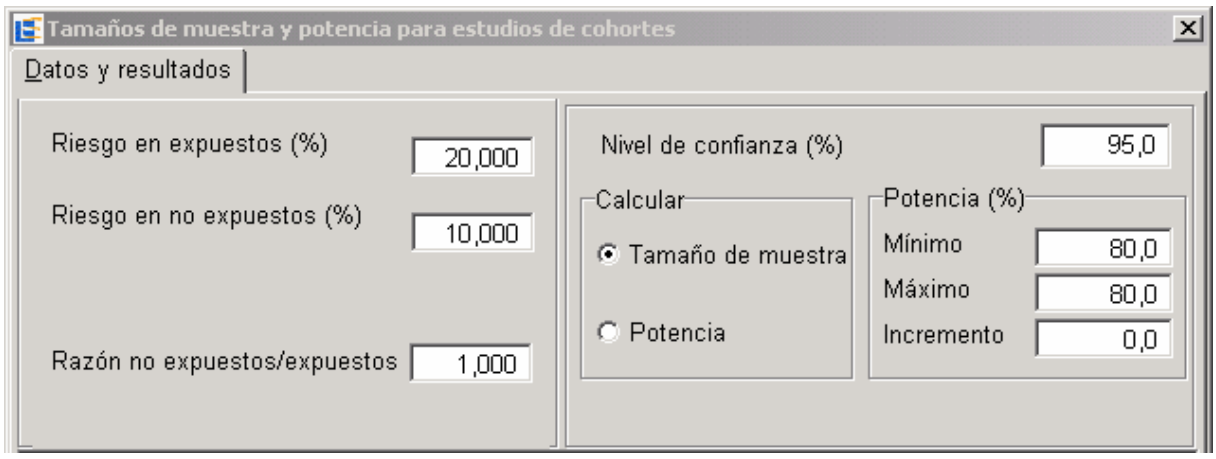
donde $P_M = (P_1 + rP_2)/(r + 1)$. Asimismo La mayoría de software estadístico, como EpiDat, propone la corrección de Yates para el cálculo del tamaño muestral:

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

Ejemplo 2: Salida de EpiDat, para calcular el tamaño de muestra suponiendo que $P_1 = 0.2$ y $P_2 = 0.1$, $\alpha = 0.05$, $\beta = 0.2$ y $r = 1$, o sea un no expuesto por cada expuesto.







```
[  1] Tamaños de muestra y potencia para estudios de cohortes


          Riesgo en expuestos:   20,000%
        Riesgo en no expuestos:   10,000%
  Razón no expuestos/expuestos:    1,000
           Nivel de confianza:     95,0%


                                                    Tamaño de muestra
    Potencia (%)    Ji-cuadrado                Expuestos      No expuestos
---------------    ------------------------- ----------    ---------------
            80,0    Sin corrección                   199                199
                    Corrección de Yates              219                219
```

Así, el tamaño de muestra calculado es $n = 219$ en ambos grupos.

Por otra parte, el número de individuos expuestos necesarios para el estudio puede reducirse aumentando el número r de individuos no expuestos por cada expuesto. Esto es de gran importancia cuando el factor de riesgo en estudio es poco frecuente.

3) Si el objetivo es comparar dos tasas de incidencia en un estudio de seguimiento, se deberá "conocer":

a) Hipótesis nula H_0: $\boldsymbol{\mu_1 - \mu_2 = 0}$

b) Valores anticipados de $\boldsymbol{\mu_1}$ y $\boldsymbol{\mu_2}$

c) Nivel de confianza: **100(1-α)%**

d) Potencia del test: **100(1-ß)%**

e) Hipótesis alternativa H_1: $\boldsymbol{\mu_1 - \mu_2 < 0}$, $\boldsymbol{\mu_1 - \mu_2 > 0}$, ó $\boldsymbol{\mu_1 - \mu_2 \neq 0}$.

f) Duración del estudio: **T**

Si el estudio no tiene duración fija (o sea los sujetos se siguen durante toda su vida) se utiliza la siguiente fórmula:

Para pruebas de una cola: $n_1 = \dfrac{\left[z_{1-\alpha}\sqrt{(1+k)\mu_m^2} + z_{1-\beta}\sqrt{k\mu_1^2 + \mu_2^2} \right]^2}{k(\mu_1 - \mu_2)^2}$, $\mu_m=(\mu_1+\mu_2)/2$ y k es el cociente de la cantidad de individuos del segundo grupo (n_2) entre la cantidad del primer grupo (n_1). Para pruebas de dos colas tenemos: $n_1 = \dfrac{\left[z_{1-\alpha/2}\sqrt{(1+k)\mu_m^2} + z_{1-\beta}\sqrt{k\mu_1^2 + \mu_2^2} \right]^2}{k(\mu_1 - \mu_2)^2}$.

Si el estudio termina antes que todos los individuos experimenten el suceso de interés, se dice que las observaciones están censuradas. Los valores de μ deben modificarse de acuerdo a la siguiente fórmula: $f(\mu) = \dfrac{\mu^3 T}{\mu T - 1 + \exp(-\mu T)}$. Notemos que cuando T tiende a infinito entonces $f(\mu)=\mu^2$.

La fórmula apropiada para el tamaño de muestra en pruebas de dos colas es: $n_1 = \frac{\left[z_{1-\alpha/2}\sqrt{(1+k)f(\mu_m)} + z_{1-\beta}\sqrt{kf(\mu_1)+f(\mu_2)} \right]^2}{k(\mu_1 - \mu_2)^2}$.

Ejemplo 3: Como parte de un estudio de los efectos a largo plazo del ruido, se diseña un estudio de seguimiento de trabajadores de una industria ruidosa y de una ocupación menos ruidosa. Los sujetos serán seguidos de por vida, y periódicamente serán examinados en cuanto a disfunciones auditivas. Una encuesta previa sugiere una tasa de incidencia anual del 25% en dicha industria. ¿Cuántas personas se deberán seguir si queremos saber que esta tasa difiere del promedio nacional de disfunciones auditivas = 10%, con un nivel de significación del 5% y una potencia del 80%?

Tenemos que $\mu_1 = 0.25$, $\mu_2 = 0.10$, $\alpha = 0.05$ y $\beta = 0.2$. Utilizando la fórmula para pruebas de dos colas:

$$n_1 = \frac{\left[1.96\sqrt{2\times 0.175^2} + 0.84\sqrt{0.25^2 + 0.1^2} \right]^2}{0.15^2} \approx 23 \text{ en cada grupo.}$$

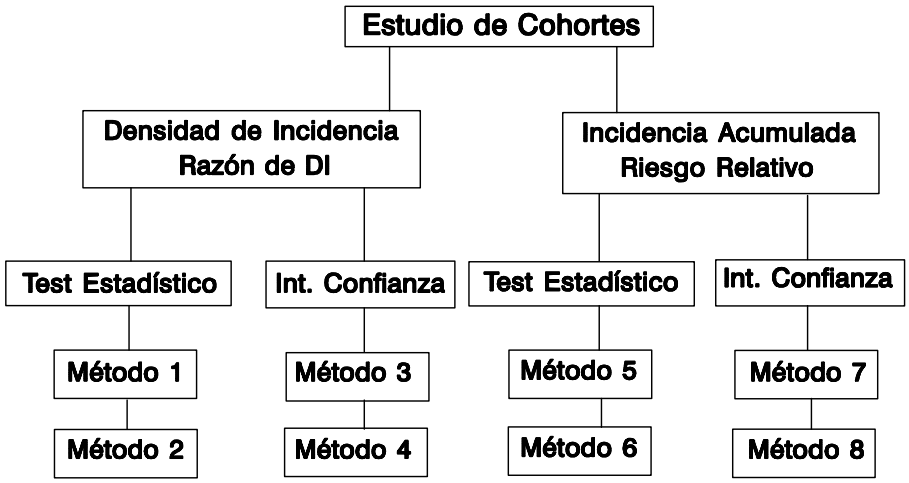
Si en este ejemplo consideramos que la duración del estudio es de 5 años, entonces $f(\mu_M=0.175)=0.0918$, $f(\mu_1=0.25)=0.1456$, y $f(\mu_2=0.1)=0.0469$. Sustituyendo, obtenemos:

$$n_1 = \frac{\left[1.96\sqrt{2\times 0.0918} + 0.84\sqrt{0.1456 + 0.0469} \right]^2}{0.15^2} \approx 65 \text{ en cada grupo.}$$

3. ANÁLISIS DE ESTUDIOS DE COHORTES

3.1. Plan de análisis estadístico para estudios de cohortes

Un posible esquema de plan de análisis para estudios de cohortes es el siguiente:



3.2. Métodos relacionados con Razón de Densidades de Incidencia (RDI)

A continuación presentaremos resumidamente los principales métodos. Presentamos primero los métodos relacionados con la razón de densidades de incidencia. La disposición de los resultados de un estudio de cohortes en este caso es la siguiente:

	Casos nuevos	Personas-año (tiempo)
Expuestos	a	N_1
No Expuestos	b	N_0
Total	M_1	T

***Método 1*: Modelo binomial.**

La probabilidad de obtener exactamente *k* casos está dada por la fórmula siguiente:

$$p(k) = \binom{n}{k} p^k q^{n-k}$$

donde n = M_1, y bajo la hipótesis nula de no asociación entre el factor y la enfermedad se tiene que p = N_1/T, y q = N_0/T

Regla de decisión:

a) Prueba unilateral H_0: RDI ≤ 1 vs. H_a: RDI > 1: Pr(K ≥ a) ≤ α

b) Prueba bilateral H_o: RDI = 1 vs. H_a: RDI≠1: Pr(K ≥ a) + Pr(K ≤ b) ≤ α.

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

Cuando se considera una muestra que cumpla las condiciones siguientes $np \geq 5$, y $nq \geq 5$, se puede utilizar para los contrastes de hipótesis anteriores una aproximación normal a la distribución binomial con: $\mu = np = \frac{M_1 N_1}{T}$, y $\sigma = \sqrt{\frac{M_1 N_1 N_0}{T^2}}$, el test estadístico queda como: $z = \frac{a - \mu}{\sigma} = \chi_{M-H}$, y la regla de decisión es: $z \geq z_\alpha$.

***Método 3*: Intervalo de confianza para la RDI aplicando transformación logarítmica**

Cuando la muestra es suficientemente grande se tiene que el logaritmo de la *RDI* distribuye aproximadamente normal con media $\ln(RDI)$ y varianza $\frac{1}{a} + \frac{1}{b}$. El intervalo de confianza está dado por: $\ln(RDI) \pm z_{1-\alpha}\sqrt{\frac{1}{a} + \frac{1}{b}}$.

***Método 4*: Intervalo de confianza basado en el test estadístico (Miettinen).**

El intervalo de confianza está dado por: $RDI^{(1 \pm z_{1-\alpha}/\chi)}$ donde χ es el valor del test estadístico obtenido con el Método 2.

3.3. Métodos relacionados con Incidencia Acumulada (IA)

A continuación presentaremos los métodos relacionados con la incidencia acumulada y el riesgo relativo. La disposición de los resultados de un estudio de cohortes en este caso es la siguiente:

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

***Método 5*: Modelo hipergeométrico**

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

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

Con lo cual, 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 6*: 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.

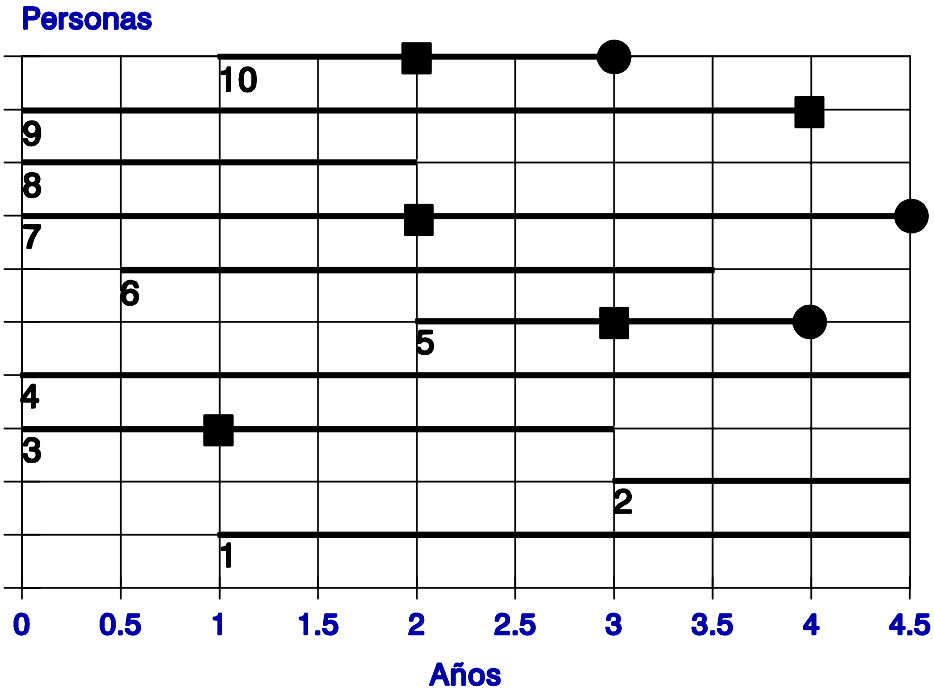
***Método 7*: Intervalo de confianza aproximado para el RR**

Aplicando una transformación logarítmica se obtiene: $\ln(RR) \pm z_{1-\alpha}\text{e.e.}(\ln(RR))$ donde $\text{e.e.}(\ln(RR)) = \sqrt{\frac{c}{aN_1} + \frac{d}{bN_2}}$.

***Método 8*: Intervalo de confianza basado en el test estadístico (Miettinen).**

El intervalo de confianza está dado por: $RR^{1\pm z_{1-\alpha}/\chi}$ donde χ es el valor del test estadístico obtenido con el Método 6.

Ejemplo 4: En la Figura 2 se representa una población de diez personas seguidas durante 4.5 años y que al inicio del estudio estaban libres de enfermedad, ■ indica el punto en que la enfermedad es diagnosticada, ● indica el punto en que la persona muere y - representa la observación de una persona. Obtenemos en la siguiente tabla de resultados para el cálculo de la *RDI*.



Así, resumiendo los datos en una tabla tenemos que:

	Casos nuevos	Años-Personas
Expuestos	4	11.5
No Expuestos	1	12
Total	5	23.5

Aplicando la regla de decisión del Método 1:

$$\Pr(K \geq a) = \sum_{k=4}^{5} \binom{5}{k} 0.49^{k} 0.51^{5-k} = \binom{5}{4} 0.49^{4} 0.51^{1} + \binom{5}{5} 0.49^{5} 0.51^{0}$$

y efectuando los cálculos obtenemos:

$$\Pr(K \geq 4) = 5 \times 0.49^4 \times 0.51 + 1 \times 0.49^5 \times 1 = 0.147 + 0.028 = 0.175$$

que es mayor que $\alpha = 0.05$. Notemos que si en lugar de $a = 4$ y $b = 1$ hubiésemos tenido los resultados $a = 5$ y $b = 0$ entonces $\Pr(K \geq 5) = 1 \times 0.49^5 \times 1 = 0.028$, y el resultado hubiese sido que *RDI* es significativamente diferente de 1.

En este mismo ejemplo si calculamos la *IA* y *RR*, tenemos los siguientes resultados:

	Casos	No Casos	Total
Expuestos	4	1	5
No Expuestos	1	4	5
Total	5	5	10

Aplicando la regla de decisión del Método 5:

$$\Pr(K \geq 4) = \sum_{k=4}^{5} \frac{\binom{5}{k}\binom{5}{5-k}}{\binom{10}{5}} = \frac{\binom{5}{4}\binom{5}{1}}{\binom{10}{5}} + \frac{\binom{5}{5}\binom{5}{0}}{\binom{10}{5}} = \frac{25}{252} + \frac{1}{252} = 0.1031$$

como $\Pr(K \geq 4) \geq 0.05$ no se rechaza la hipótesis H_0 de no asociación.

Ejemplo 5: En la tabla siguiente se presentan los resultados de un estudio de seguimiento durante 4 años de 120 bebedores-problema y 380 bebedores-moderados y la aparición de disfunciones hepáticas.

	Casos	No Casos	Total	Tiempo en observación
Expuestos	40	80	120	420
No Expuestos	60	320	380	1406
Total	100	400	500	1826

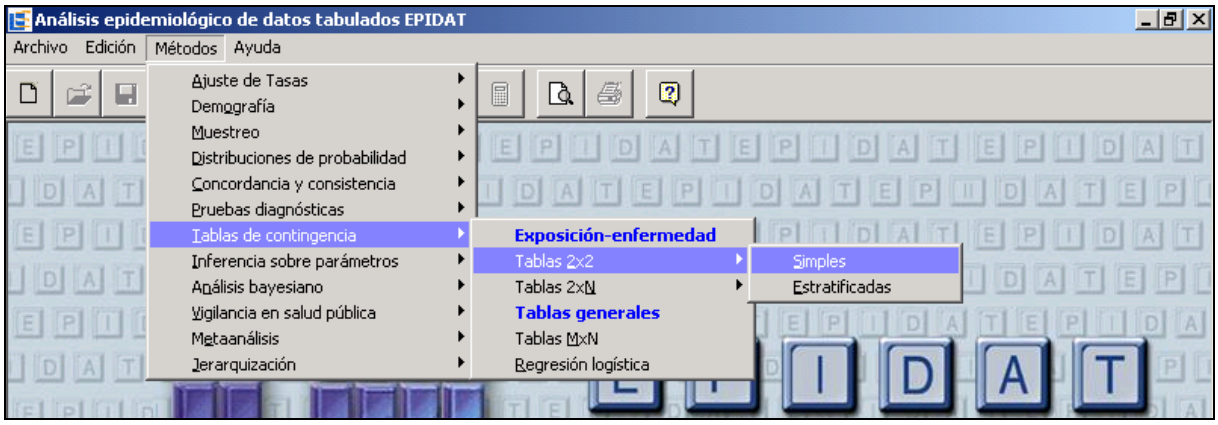
En este caso podemos utilizar el Método 2 para *RDI* y el Método 6 para *RR*.

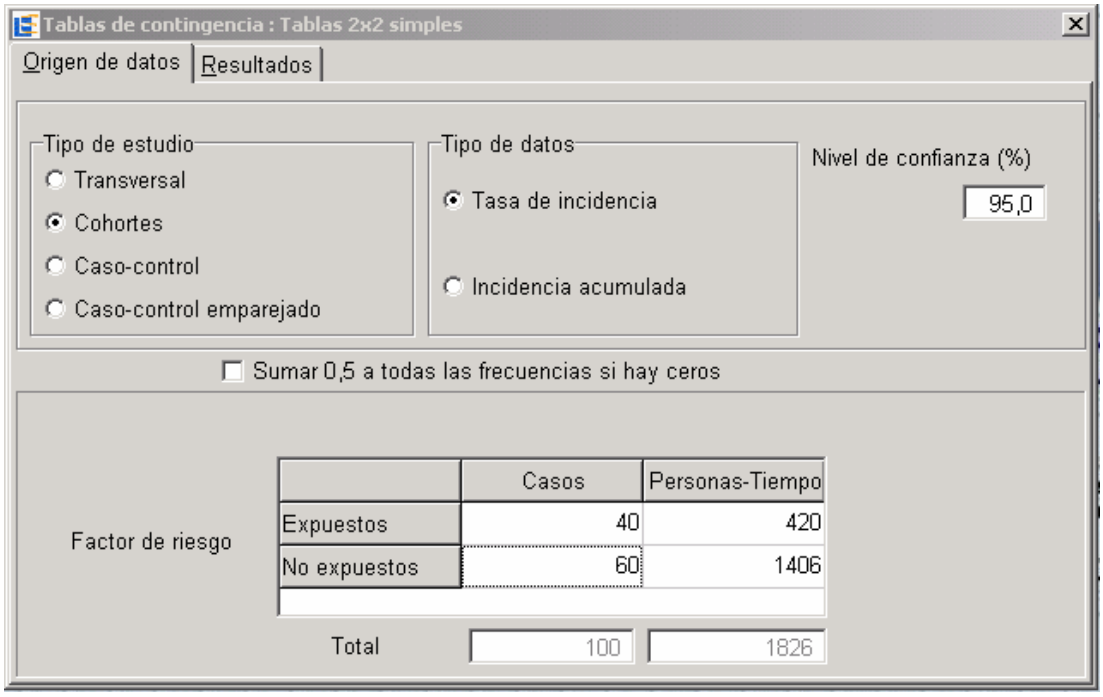
a) Método 2: $\mu = \frac{M_1 N_1}{T} = \frac{100 \times 420}{1826} \approx 23.0$ y

$\sigma = \sqrt{\frac{M_1 N_1 N_0}{T^2}} = \sqrt{\frac{100 \times 420 \times 1406}{1826^2}} \approx 4.21$, de donde obtenemos

$z = \frac{a - \mu}{\sigma} = \frac{40 - 23.0}{4.21} \approx 4.04$ que es mayor que $z = 1.96$ por tanto rechazamos H_0 de no asociación.

Veamos estos resultados obtenidos con el programa EpiDat:





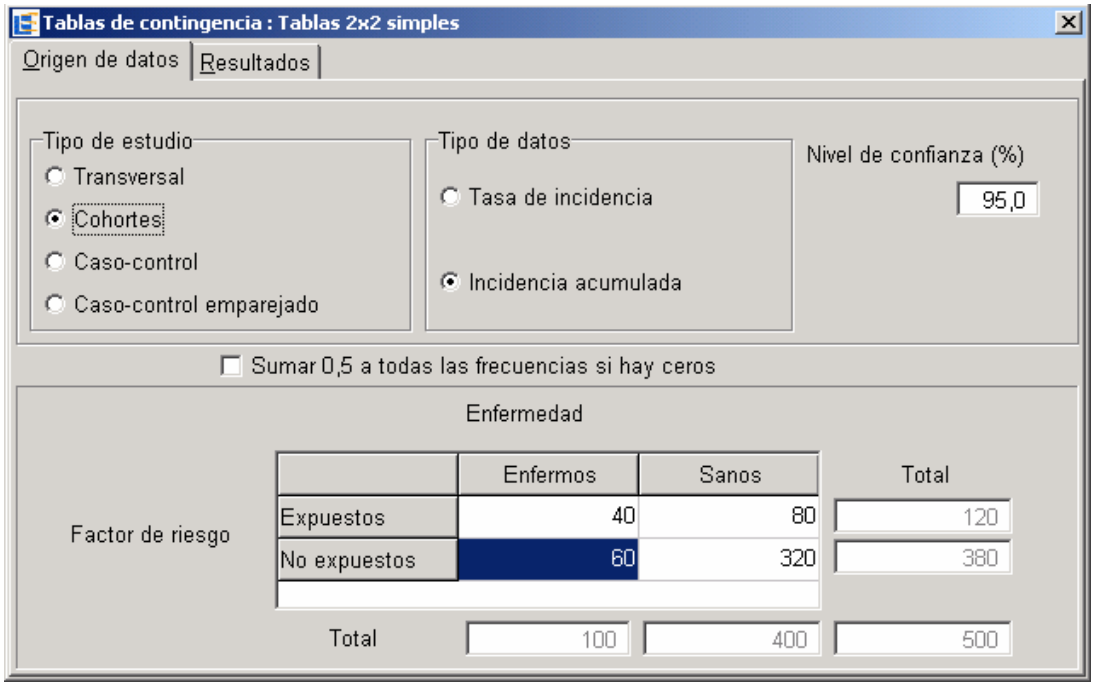
```
[  3] Tablas de contingencia : Tablas 2x2 simples

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

Tabla
                  Casos       Personas-Tiempo
------------   --------       ---------------
Expuestos            40                   420
No expuestos         60                  1406
------------   --------       ---------------
Total               100                  1826

                                          Estimación            IC(95,0%)
----------------------------------------  ----------  ---------  ---------
Tasa de incidencia en expuestos             0,095238          -          -
Tasa de incidencia en no expuestos          0,042674          -          -
Razón de tasas de incidencia                2,231746  1,495870   3,329627
Diferencia de tasas de incidencia           0,052564  0,021137   0,083991
----------------------------------------  ----------  ---------  ---------
Fracción atribuible en expuestos            0,551920  0,331493   0,699666
Fracción atribuible poblacional             0,220768  0,085535   0,336003

Prueba de asociación

       Estadístico Z    Valor p
--------------------  ---------
              3,9205      0,0001
```

b) **Método 6:** $\mu = \frac{M_1 N_1}{T} = \frac{100 \times 120}{500} = 24$,

y $\sigma = \sqrt{\frac{N_1 N_0 M_1 M_0}{T^2 (T-1)}} = \sqrt{\frac{120 \times 380 \times 100 \times 400}{500^2 \times 499}} \approx 3.82$,

donde obtenemos: $z = \frac{a - \mu}{\sigma} = \frac{40 - 24}{3.82} \approx 4.19$ que es mayor que $z = 1.96$ y por tanto rechazamos H_0.

Veamos de nuevo estos resultados obtenidos con los programas EpiDat:



```
[  4] 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           40             80            120
No expuestos        60            320            380
------------  --------       --------       --------
Total              100            400            500

                                          Estimación            IC(95,0%)
----------------------------------------  ----------     ---------      ---------
Riesgo en expuestos                        0,333333          -              -
Riesgo en no expuestos                     0,157895          -              -
Riesgo relativo                            2,111111       1,497491       2,976171
Diferencia de riesgos                      0,175439       0,083471       0,267406
Odds ratio                                 2,666667       1,668122       4,26294 (W)
                                                          1,671672       4,25515 (C)
----------------------------------------  ----------     ---------      ---------
Fracción atribuible en expuestos            0,526316      0,332216        0,663998
Fracción atribuible poblacional             0,210526      0,080686        0,322029


Prueba Ji-cuadrado de asociación          Estadístico      Valor p
----------------------------------------   ---------     ---------
Sin corrección                              17,5439         0,0000
Corrección de Yates                         16,4645         0,0000
```

Notemos que $z^2 = 4.19^2 \approx 17.55$ coincide con χ^2=17.54.