Departamento de Estadística
Universidad Carlos III de Madrid

BIOESTADISTICA (55 - 10536)

Estudios de prevalencia (transversales)

CONCEPTOS CLAVE

1) Características del diseño en un estudio de prevalencia, o transversal.

2) Importancia del mecanismo de muestro y selección del tamaño muestral.

3) Tipos y características de las medidas de prevalencia.

4) Estrategias para el análisis de estudios transversales: Estimación vs. Comparación de prevalencias.

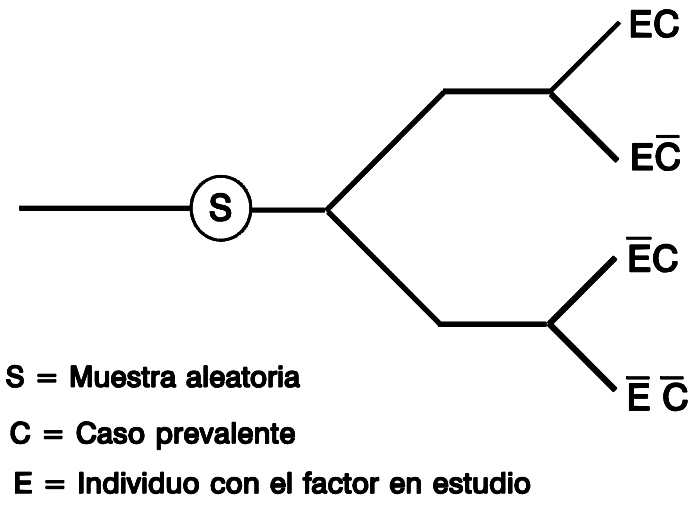
1. INTRODUCCION

Un **Estudio de Prevalencia** es aquel en el que se examinan las relaciones entre las enfermedades o entre las características relacionadas con la salud y otras variables de interés, del modo en que existen en una población y momento determinados. La presencia o ausencia de la enfermedad y de las otras variables (o, si son de tipo cuantitativo, su nivel) se determinan en cada miembro de la población estudiada o en una muestra representativa en un momento dado.

La relación entre una variable y la enfermedad puede examinarse:

1) En términos de la prevalencia de la enfermedad en diferentes subgrupos de población, definidos de acuerdo con la presencia o ausencia (o nivel) de las variables.
2) En términos de la presencia o ausencia (o nivel) de las variables en los individuos enfermos, en comparación con los sanos.

En la figura siguiente se muestra una representación gráfica de un estudio transversal.



Obsérvese que, en un estudio de prevalencia (o estudio transversal), lo que queda registrado formalmente no es la incidencia de una enfermedad, sino su prevalencia. La secuencia temporal de causa a efecto no queda necesariamente determinada en un estudio de este tipo.

Después del proceso se selección, todos los sujetos participantes son examinados, observados, y/o encuestados acerca de la enfermedad, su nivel actual o pasado del factor en estudio y otras variables de interés. En ocasiones los estudios transversales no utilizan muestreo probabilístico, en esos casos su valor es limitado para describir la frecuencia de la enfermedad y de otras características en la población objeto.

2. MEDIDAS DE PREVALENCIA

Hemos revizado las medidas fundamentales para cuantificar la prevalencia: **prevalencia puntual y prevalencia lápsica**.

1) La más utilizada de estas medidas es la **Prevalencia Puntual (P)**, que no es más que la probabilidad de que un individuo en una población presente una determinada característica (por ejemplo, enfermedad) en el tiempo *t*. Así, se estima por: $\hat{P}_t = \frac{C_t}{N_t}$, donde C_t es el número de casos prevalentes y N_t es la población encuestada.

2) La **Prevalencia Lápsica (PL)** no es más que la probabilidad de que un individuo de una población sea un caso en cualquier momento del período (t_0, t). Se estima por: $\hat{P}P_{(t_0,t)} = \frac{C_{(t_0,t)}}{N} = \frac{C_0 + I_{(t_0,t)}}{N}$. Donde $C_{(to,t)}$ incluye los casos prevalentes C_0 y los casos incidentes $I_{(to,t)}$.

Raras veces la prevalencia tiene interés directo en aplicaciones etiológicas de la investigación epidemiológica. Puesto que la probabilidad de sobrevivir (o de curación) afecta a la prevalencia, los estudios transversales o estudios basados en casos prevalentes obtienen asociaciones que reflejan los determinantes de la supervivencia (la cura) una vez que se padece la enfermedad, así como las causas de tal enfermedad. Una supervivencia mejor, y por tanto, una prevalencia más alta, podrían estar realmente relacionadas con la acción de factores preventivos que mitigasen de alguna manera la enfermedad, una vez que se produjese.

2.1. Relación entre Incidencia y Prevalencia

Aunque es evidente la prevalencia depende en parte de la incidencia, la relación funcional que se establece entre ellas es bastante compleja. Sin embargo, si asumimos que la población está en posición de equilibrio, o sea, es estable y las tasas de prevalencia e incidencia permanecen constantes, puede obtenerse una relación bastante sencilla.

Bajo las condiciones de equilibrio la cantidad de casos incidentes I en un período (t_0,t) de duración Δt es igual a los casos salientes TC (casos que mueren por la enfermedad, o por otra causa, y casos que curan) durante el mismo período, o sea, $I = TC$.

Tenemos que $I = DI(N\text{-}C)\Delta t$, donde N es la cantidad de individuos de la población, y $TC = TD \times C \times \Delta t$, donde TD es la tasa o densidad de salida y C los casos prevalentes, de donde obtenemos: $\frac{C}{N - C} = \frac{DI}{TD}$.
Si representamos las salidas de la enfermedad por un proceso Poisson, tenemos que TD es igual al inverso de la duración de la enfermedad $\overline{D}$. Dividiendo por N y sustituyendo TC, obtenemos:

$$P = \frac{DI\,\overline{D}}{DI\,\overline{D} + 1}$$

Si la prevalencia es pequeña ($P < 0.1$) se tiene una fórmula simplificada: $P = DI\,\overline{D}$.

3. DISEÑO DE UN ESTUDIO DE PREVALENCIA

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

a) Definir la población de referencia.

b) Determinar si el estudio se realizará sobre el total de la población o en una muestra.

c) Determinar el tamaño de la muestra poblacional y las formas de selección de la misma.

d) Elaborar y validar los instrumentos y técnicas mediante los cuales se determinará la presencia o ausencia de las características de interés.

e) Asegurar la comparabilidad de la información obtenida en los diferentes grupos.

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

g) Determinar la conducta a seguir con los datos detectados.

Nos centraremos en los puntos **c)** y **f)**

Veamos a continuación como se calculan los tamaños de muestra en los estudios de prevalencia para distintas situaciones:

1) Si el objetivo es estimar una proporción (P) en una población con una precisión absoluta especificada se deberá "conocer":

a) Proporción esperada en la población ***P***

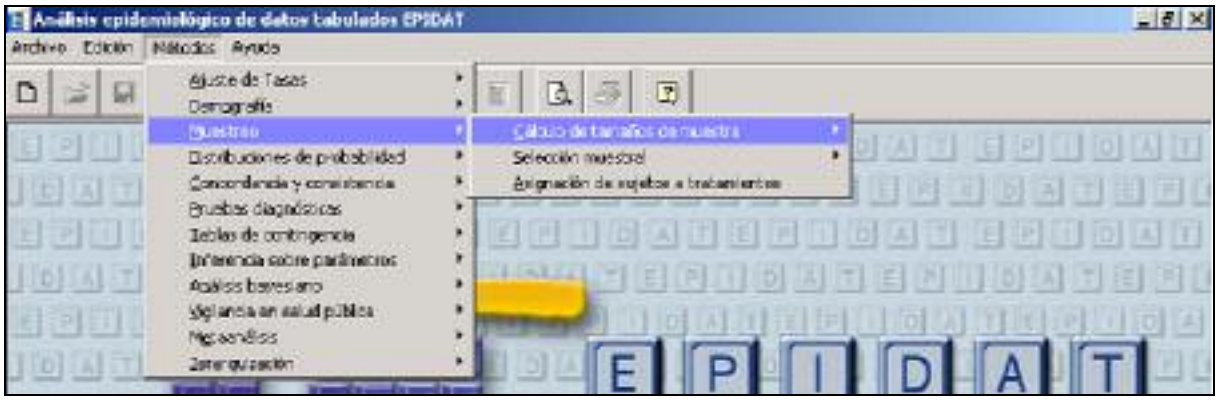
b) Nivel de confianza **100(1-*α*)%**

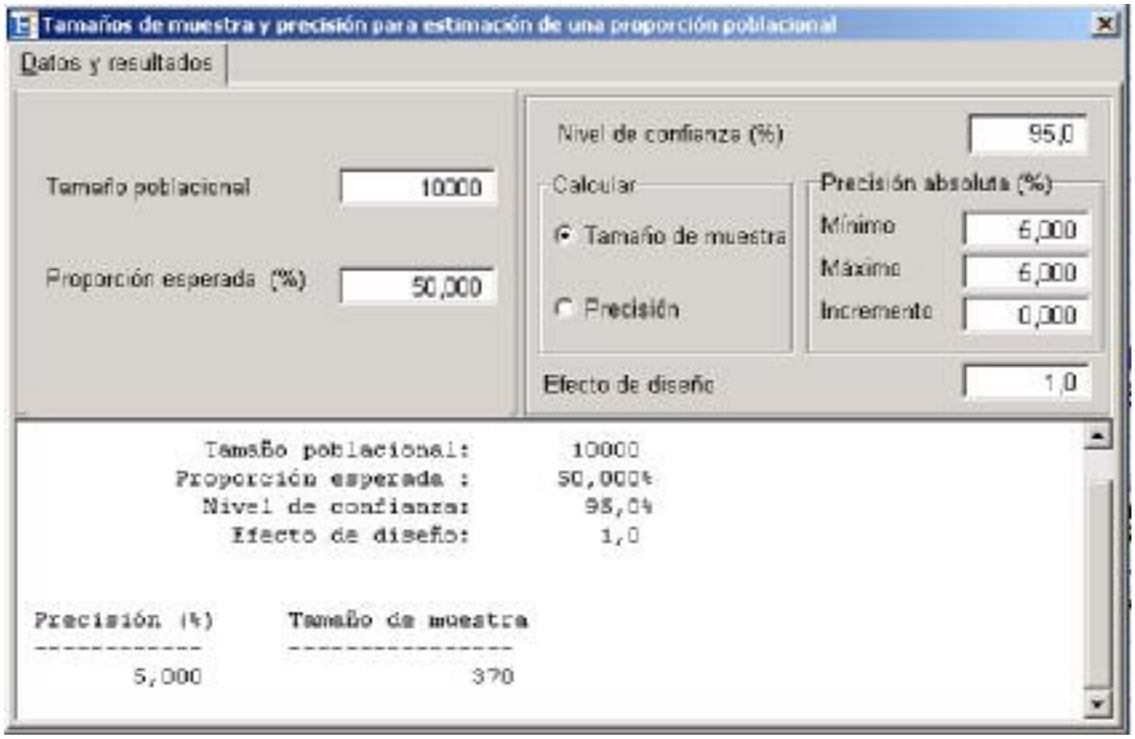
c) Precisión absoluta requerida ***d***

Se utiliza en este caso la siguiente fórmula: $n = \frac{z^2_{1-\alpha/2}\, P(1-P)}{d^2}$

La mayoría de software estadístico, entre ellos el programa **EpiDat** utilizan además la siguiente corrección: $n = \frac{n_0}{1 + \frac{n_0}{N}}$, donde n_0 se obtiene por la fórmula anterior y N es el tamaño de la población.

Ejemplo 1. Cálculo tamaño muestral con EpiDat, para estimar una prevalencia del 50% (P=0.5) con una precisión del 5% (d = 0.05) en una población con 10,000 habitantes (N=10000).





2) **Si el objetivo es probar que la proporción poblacional difiere de un valor dado P_0 se deberá "conocer":**

   a) Valor de la proporción bajo la hipótesis nula, P_0

   b) Valor anticipado de la proporción P_a

   c) Nivel de significación $100\alpha\%$

   d) Potencia del test $100(1-\beta)\%$

   e) Hipótesis alternativas: $P_a > P_0$, $P_a < P_0$ ó $P_a \neq P_0$.

Se utilizan en este caso las siguientes fórmulas:

Para pruebas de una sola cola: $n = \dfrac{[z_{1-\alpha}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_a(1-P_a)}\ ]^2}{(P_0 - P_a)^2}$, y para pruebas de dos colas: $n = \dfrac{[z_{1-\alpha/2}\sqrt{P_0(1-P_0)} + z_{1-\beta}\sqrt{P_a(1-P_a)}\ ]^2}{(P_0 - P_a)^2}$.

Ejemplo 2: La proporción de pacientes curados de cáncer después de 5 años de tratamiento se reporta en la literatura como del 50%. Un investigador desea probar que esta tasa de cura es válida en su distrito sanitario. ¿Qué tamaño de muestra necesita si está interesado en rechazar la hipótesis si la tasa verdadera es menor que 50% y desea con un 90% de seguridad detectar una tasa verdadera del 40% a un nivel de confianza del 95%?

En este caso $P_0 = 0.5$, $P_a = 0.4$, $\alpha = 0.05$, $\beta = 0.1$ y H_a: $P_a < P_0$, sustituyendo obtenemos que n=211.

3) **Si el objetivo es estimar la diferencia entre dos proporciones poblacionales con una precisión absoluta especificada, se deberá "conocer":**

   a) Proporciones esperadas de las poblaciones P_1 y P_2

   b) Nivel de confianza $100(1-\alpha)\%$

   c) Precisión absoluta requerida d

Se utiliza en este caso la siguiente fórmula:

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

Ejemplo 3: ¿Qué tamaño de muestra es necesario seleccionar de dos poblaciones para estimar una diferencia de riesgo con un nivel de confianza del 95% y una precisión de 0.05, cuando no se conocen estimaciones de P_1 y P_2?

En este caso la selección más aconsejable es $P_1 = P_2 = 0.5$, pues el valor $V = P_1(1-P_1) + P_2(1-P_2)$ es máximo para esos valores. Tenemos además $\alpha = 0.05$ y $d = 0.05$, de donde utilizando la fórmula anterior obtenemos, $n = 769$ para cada una de las poblaciones.

4) **Si el objetivo es probar que dos proporciones poblacionales son diferentes se deberá "conocer":**

   a) Hipótesis nula: $\boldsymbol{P_1}$-$\boldsymbol{P_2}$ = 0.

   b) Proporciones esperadas de las poblaciones $\boldsymbol{P_1}$ y $\boldsymbol{P_2}$

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

   d) Potencia del test **100(1-β)%**

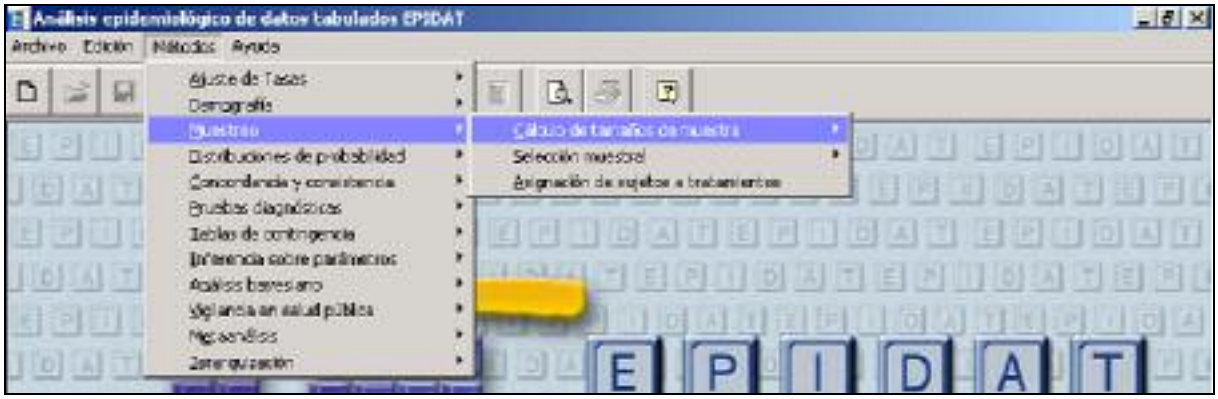
   e) Hipótesis alternativas: $\boldsymbol{P_1}$-$\boldsymbol{P_2} > 0$, $\boldsymbol{P_1}$-$\boldsymbol{P_2} < 0$ ó $\boldsymbol{P_1}$-$\boldsymbol{P_2} \neq 0$

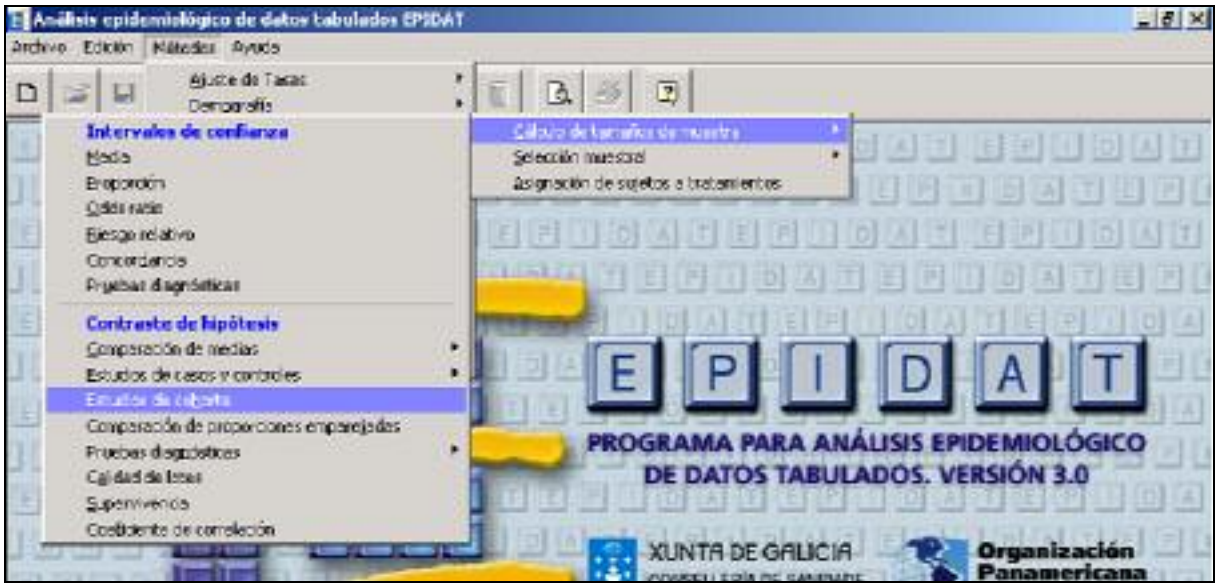
Se utilizan en este caso las siguientes fórmulas:

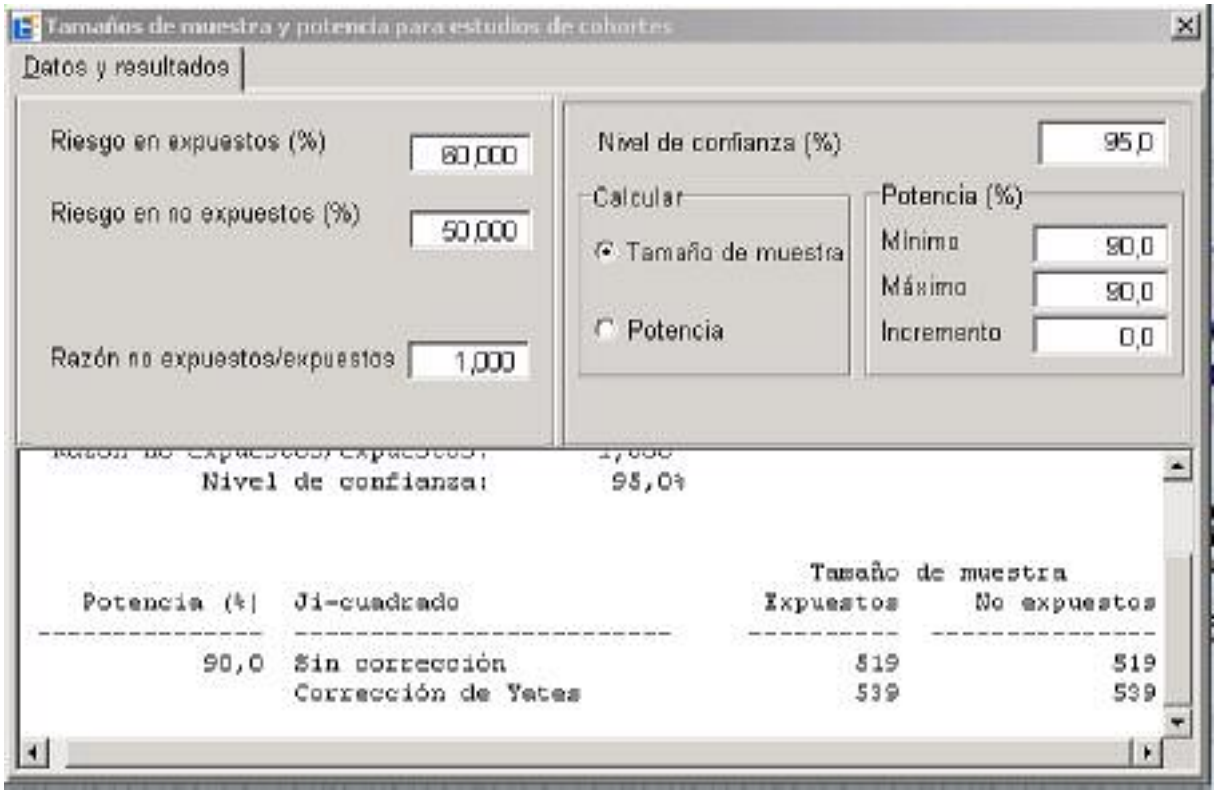
Para pruebas de una sola cola: $n = \dfrac{\left[z_{1-\alpha}\sqrt{2P_m(1-P_m)} + z_{1-\beta}\sqrt{V} \right]^2}{(P_1 - P_2)^2}$, donde $Pm = (P_1+P_2)/2$ y $V = P_1(1-P_1) + P_2(1-P_2)$. Para pruebas de dos colas: $n = \dfrac{\left[z_{1-\alpha/2}\sqrt{2P_m(1-P_m)} + z_{1-\beta}\sqrt{V} \right]^2}{(P_1 - P_2)^2}$.

De nuevo en la mayoría de software estadístico, como el programa EpiDat, permite el cálculo para una prueba de dos colas utilizando la corrección de Yates: $n = \dfrac{n'}{4}\left[1 + \sqrt{1 + \dfrac{4}{n'|P_1 - P_2|}} \right]^2$.

Ejemplo 4: Volvamos al ejemplo anterior, y supongamos ahora que $P_1 = 0.6$, $P_2 = 0.5$, α=0.05, y ß=0.1, se podría calcular el tamaño muestral utilizando EpiDat




Análisis epidemiológico de datos tabulados EPIDAT
Archivo Edición Métodos Ayuda
Ajuste de Tasas
Demografía
Intervalos de confianza
Media
Proporción
Odds ratio
Riesgo relativo
Concordancia
Pruebas diagnósticas
Contraste de hipótesis
Comparación de medias
Estudios de casos y controles
Estudios de cohorte
Comparación de proporciones emparejadas
Pruebas diagnósticas
Calidad de tests
Supervivencia
Coeficiente de correlación
Cálculo de tamaños de muestra
Selección muestral
Asignación de sujetos a tratamientos
E P I D A T
PROGRAMA PARA ANÁLISIS EPIDEMIOLÓGICO
DE DATOS TABULADOS. VERSIÓN 3.0
XUNTA DE GALICIA
Organización Panamericana


Tamaños de muestra y potencia para estudios de cohortes
Datos y resultados
Riesgo en expuestos (%) 60,000
Riesgo en no expuestos (%) 50,000
Razón no expuestos/expuestos 1,000
Nivel de confianza (%) 95,0
Calcular
Tamaño de muestra
Potencia
Potencia (%)
Mínimo 90,0
Máximo 90,0
Incremento 0,0
Nivel de confianza: 95,0%
Tamaño de muestra
Potencia (%) Ji-cuadrado Expuestos No expuestos
90,0 Sin corrección 519 519
Corrección de Yates 539 539

4. ANÁLISIS DE ESTUDIOS TRANSVERSALES

La disposición de los resultados de un estudio de prevalencia se presenta en la siguiente tabla:

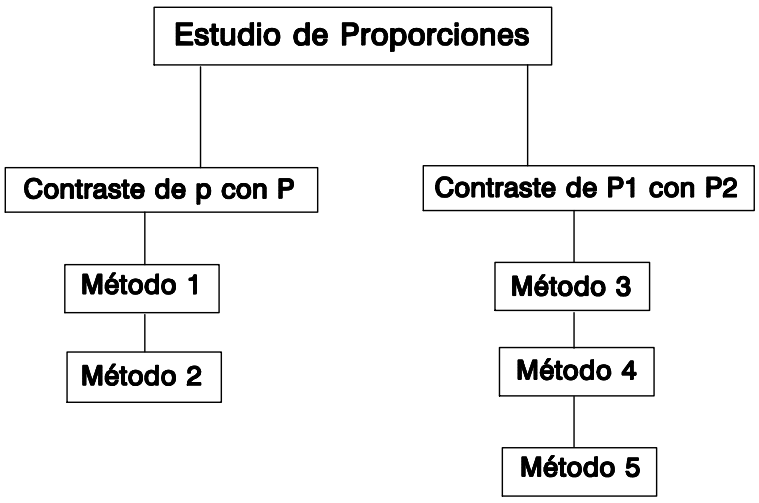
	Variable Dependiente			
Variable Independiente	Presente	Ausente	Total	Tasa
Presente	A	b	m_1	a/m_1
Ausente	C	d	m_2	c/m_2
Total	n_1	n_2	n	n_1/n

En los estudios transversales puede calcularse la **Razón de Prevalencias** por $RR = \frac{a/m_1}{c/m_2}$.

Cuando se tiene una población en estado de equilibrio se puede estimar la razón de densidad de incidencia ***RDI***, utilizando la relación entre ***DI*** y ***P***: $RDI = \frac{ID_1}{ID_2} = \frac{\overline{T}_2}{\overline{T}_1}\left[\frac{P_1/(1-P_1)}{P_2/(1-P_2)}\right]$, o equivalentemente: $ORP = RDI\,\frac{\overline{T}_1}{\overline{T}_2}$ que puede estimarse por $\frac{ad}{bc}$.

4.1. Plan de análisis estadístico para estudios de proporciones

Un posible esquema del plan de análisis estadístico para estudios de proporciones.



Método 1: Estimación de una proporción (Modelo binomial)

Supongamos que una experiencia se repite n veces, y cada vez la probabilidad p de ocurrencia de un suceso dado (por ejemplo, enfermedad, exposición) sea siempre la misma. Si las experiencias sucesivas son independientes, podemos asumir un modelo binomial.

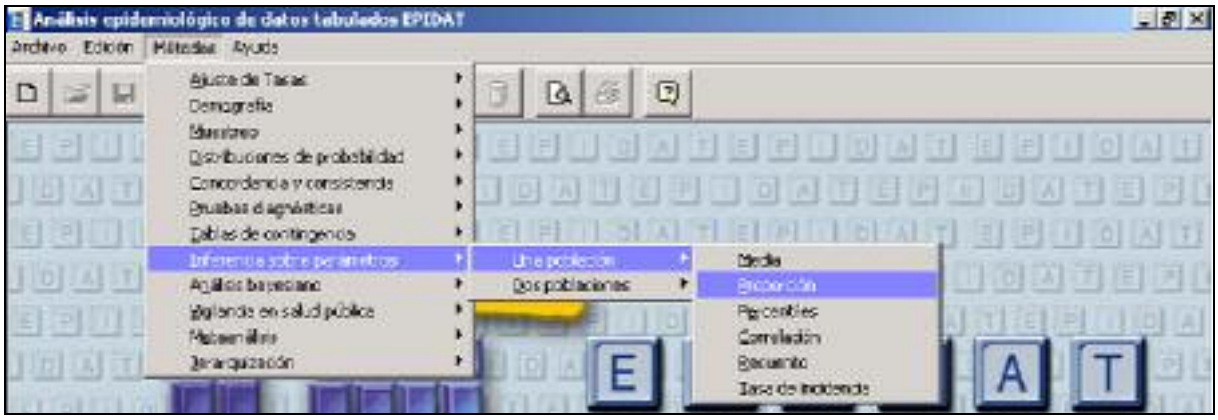
Si en n repeticiones independientes se observa X veces el suceso de interés, se tienen los siguientes resultados:

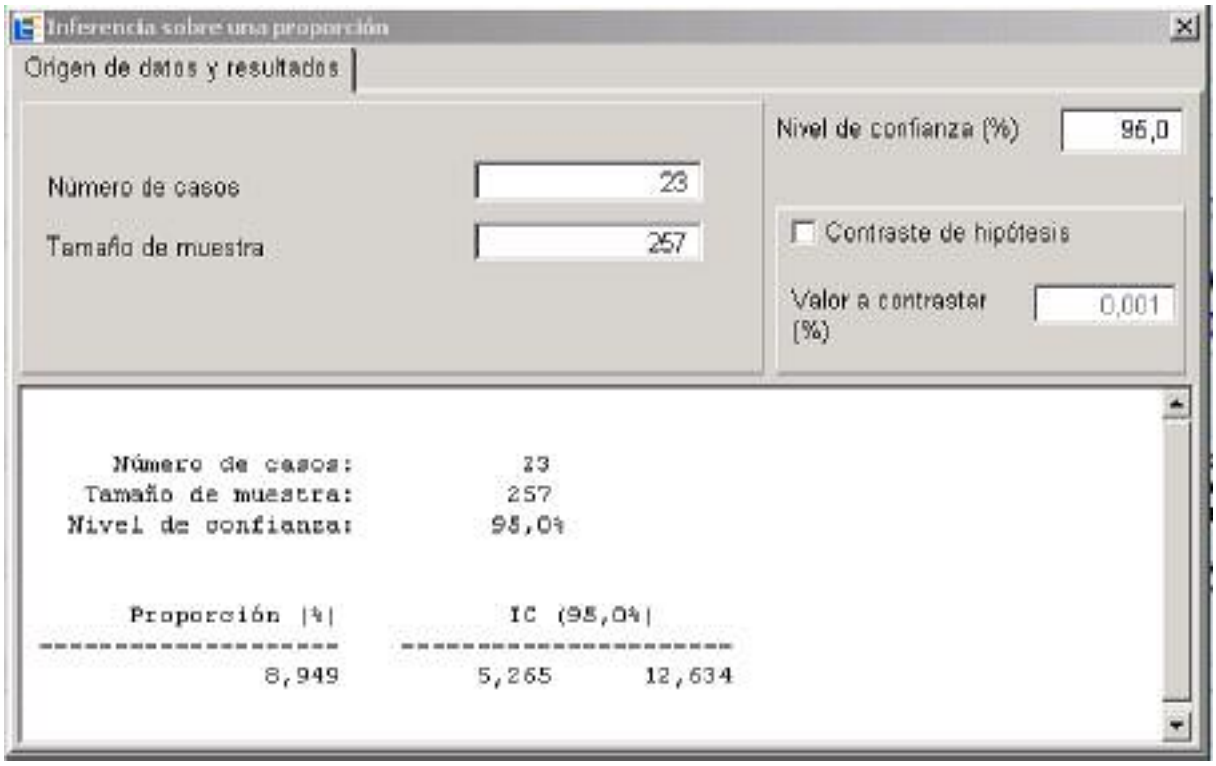
- $\hat{p} = \dfrac{X}{n}$ es un estimador sin sesgo de p.
- Su varianza es igual a $\dfrac{p(1-p)}{n}$ y se estima por $\dfrac{\hat{p}(1-\hat{p})}{n}$.
- El intervalo exacto de para p es de difícil cálculo y por tanto se utiliza un intervalo de confianza aproximado definido por $(\underline{p}, \overline{p})$:

$$\underline{p} = \frac{n}{n+z^2}\left[\hat{p} + \frac{z^2}{2n} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}\right] \text{ y } \overline{p} = \frac{n}{n+z^2}\left[\hat{p} + \frac{z^2}{2n} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}\right]$$

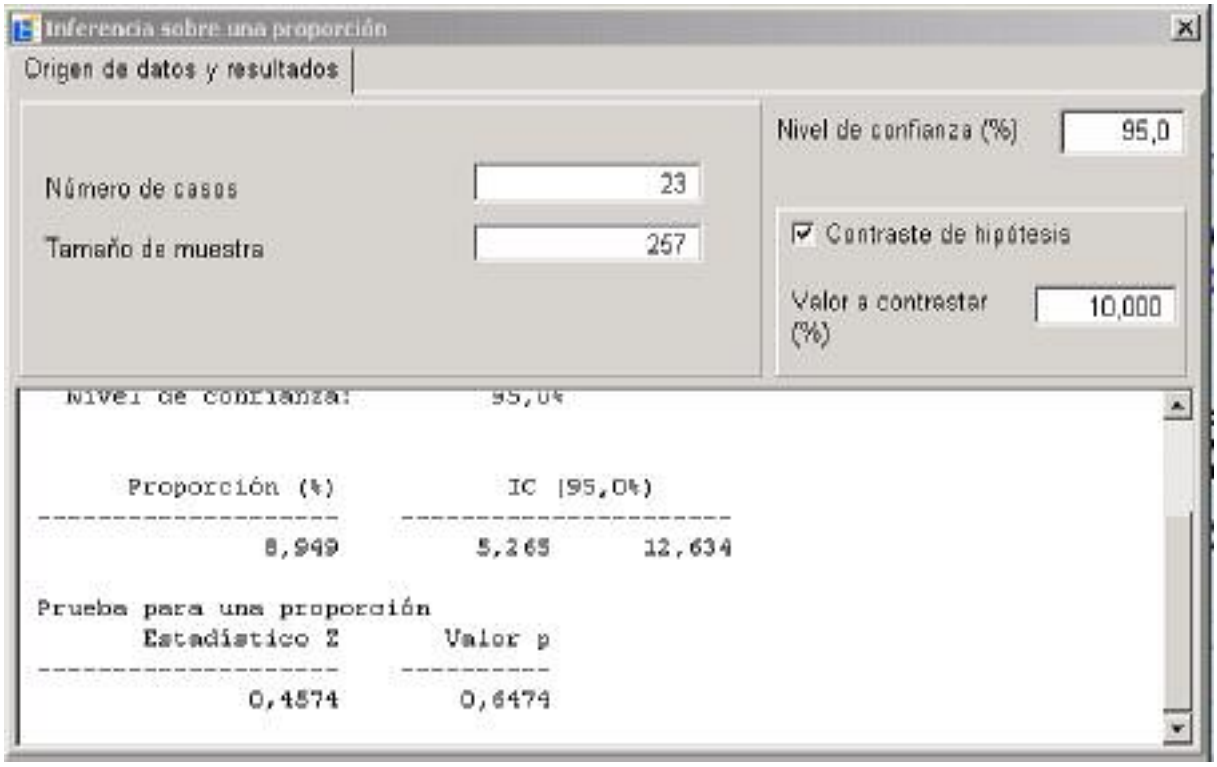
- Si $n \geq 30$ el intervalo aproximado para p será con: $\underline{p} = \hat{p} - z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ y $\overline{p} = \hat{p} + z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, donde $z = z_{1-\alpha/2}$.
- También podemos encontrar este método en forma de test estadístico de p respecto a una proporción P dada: $z = \dfrac{|p-P| - \dfrac{1}{2n}}{\sqrt{\dfrac{P(1-P)}{n}}}$.

Ejemplo 5: Supongamos que el tamaño de la muestra es n = 257 y se observan 23 casos, entonces con EpiDat se obtiene:





Ejemplo 6: Siguiendo con el ejemplo anterior también podríamos contrastar si H_0: P=0.10, entonces con EpiDat se obtiene:



Lo cual nos lleva a concluir que hay evidencias suficientes (p=0.6474) para aceptar H_0.

Método 2: Estimación de una proporción (Modelo hipergeométrico)

El modelo hipergeométrico parte de suponer que en una población de tamaño N, hay N_1 enfermos y N-N_1 individuos sanos. Se toma una muestra de n personas que contiene n_1 enfermos y n-n_1 sanos. Se desea estimar la proporción N_1/N, o sea, la tasa de prevalencia de cierta población. Bajo el modelo hipergeométrico se tienen los siguientes resultados:

- n_1/n es un estimador sin sesgo de N_1/N.
- La varianza de n_1/n se estima por: $\frac{N-n}{N-1}\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\frac{1}{n}$, y si $n/N \le 0.10$ entonces la varianza se aproxima por: $\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\frac{1}{n}$.
- El intervalo exacto para p es de difícil cálculo y por tanto utilizaremos un intervalo de confianza aproximado definido por $(\underline{p}, \overline{p})$:

$$\underline{p} = \frac{n_1}{n} - z\sqrt{\frac{N-n}{N-1}\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\frac{1}{n}}, \text{ y } \overline{p} = \frac{n_1}{n} + z\sqrt{\frac{N-n}{N-1}\frac{n_1}{n}\left(1-\frac{n_1}{n}\right)\frac{1}{n}}.$$

Método 3: Modelos discretos (Dos-binomial e hipergeométrico)

a) Modelo Dos-binomial: Consideremos un estudio transversal en el que de los n_1 sujetos expuestos a un factor están enfermos ***a*** individuos, y de los n_2 no expuestos ***b*** son enfermos, entonces de acuerdo al modelo binomial la probabilidad de tener ***a*** casos expuestos y ***b*** casos no expuestos es igual a: $\binom{n_1}{a} p_1^a (1-p_1)^c \binom{n_2}{b} p_2^b (1-p_2)^d$, que es el producto de las probabilidades binomiales para cada uno de los grupos expuestos y no expuestos. El cálculo de probabilidades para la comprobación de hipótesis es difícil y ambiguo, lo cual exige uso de paquetes estadísticos.

b) Modelo Hipergeométrico: Si consideramos que los marginales de la Tabla 2×2 son fijos podemos utilizar el modelo hipergeométrico, entonces la probabilidad de obtener a o más casos en los expuestos esta dada por:

$$\Pr(K \ge a) = \sum_{k=a}^{\min(m_1, n_1)} \frac{\binom{n_1}{k}\binom{n_2}{m_1 - k}}{\binom{n}{m_1}}$$

La regla de decisión es rechazar que existe asociación entre el factor y la enfermedad H_0 si la probabilidad $\Pr(K \ge a) \le \alpha$.

Método 4: Aproximaciones a los modelos discretos.

Debido al número de cálculos que requieren los modelos del Método 3, se utilizan aproximaciones (prueba estadística asintótica) para calcular el valor de *P*:

a) Modelo Hipergeométrico: Bajo H_0, la variable aleatoria ***a*** (número de casos expuestos) tiene media $\mu = \frac{m_1 n_1}{n}$ y $\sigma^2 = \frac{m_1 m_2 n_1 n_2}{n^2(n-1)}$, y el test estadístico que se utiliza es: $z = \frac{a-\mu}{\sigma}$ y como regla de decisión $z \geq z_{1-\alpha}$.

b) Modelo Dos-binomial: El test estadístico que se obtiene es similar al anterior con la única diferencia que $\sigma^2 = \frac{m_1 m_2 n_1 n_2}{n^3}$.

Ejemplo 6: En la siguiente tabla se muestran los resultados de un estudio sobre consumo de clorodiazepóxido en la fase inicial de la gestación, en madres con niños nacidos con defecto cardíacos congénitos y madres con niños normales.

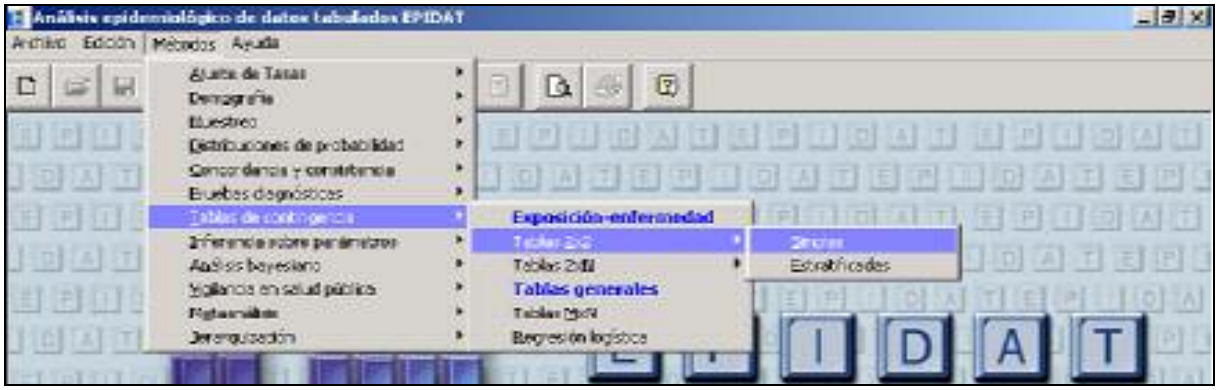
	Uso de clorodiazepóxido		
	Si	No	Total
Madre caso	4	386	390
Madre no caso	4	1250	1254
Total	8	1636	1644

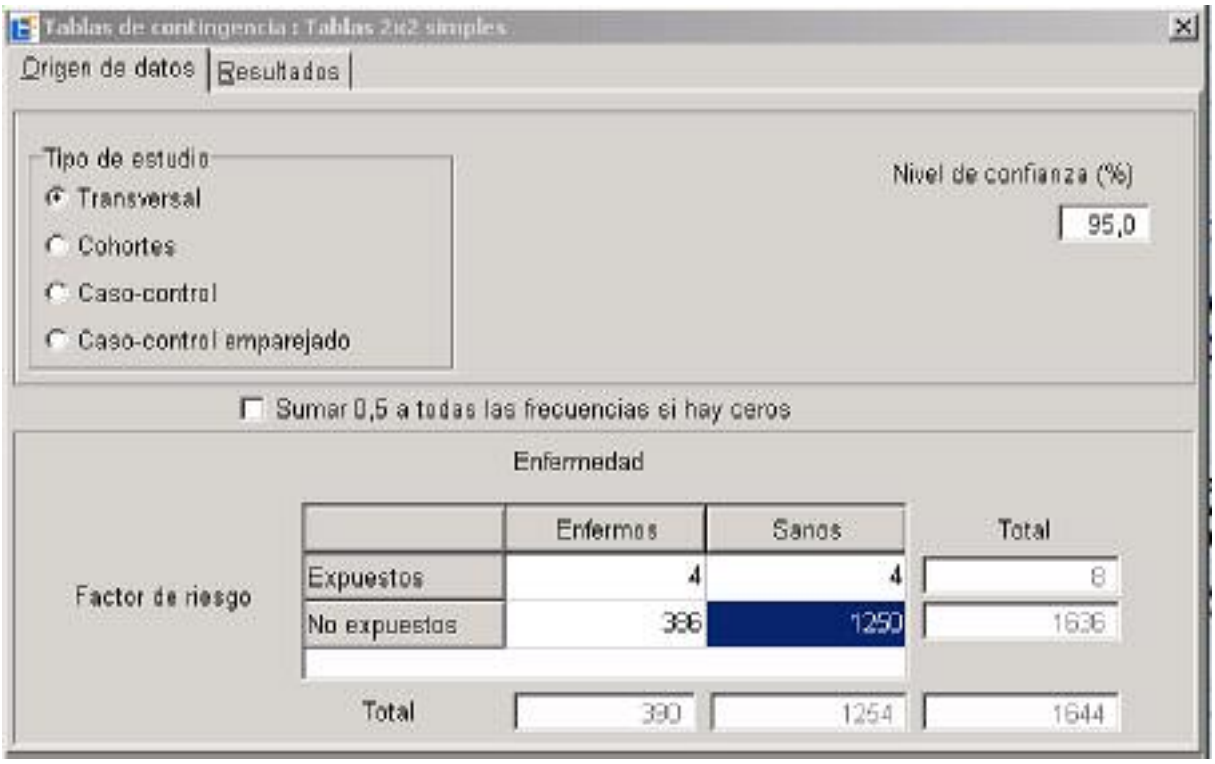
Obtenemos μ = 390×8/1644 ≈ 1.898, σ^2 = 8×1636×390×1254/(1644^2×1645) ≈ 1.44 y σ = 1.2, tenemos z = (4-1.898)/1.2 ≈ 1.75, que para una prueba de una cola p = 0.04.

Método 5: Test χ^2 (Ji-cuadrado)

Para los datos de la Tabla 2×2, se utiliza el test estadístico: $\chi^2 = \frac{(|ad-bc|-1/2n)^2 n}{n_1 n_2 m_1 m_2}$, que permite calcular los intervalos de confianza para el cociente de prevalencias mediante la expresión: $RR^{1\pm z/\chi}$, donde $z = z_{1-\alpha/2}$.

Ejemplo 7: Analizando la Tabla 2×2 del ejemplo anterior con EpiDat, obtenemos:





Tablas de contingencia : Tablas 2x2 simples

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

Tabla

	Enfermos	Sanos	Total
Expuestos	4	4	8
No expuestos	386	1250	1636
Total	390	1254	1644

Prevalencia de la enfermedad	Estimación	IC(95,0%)	
En expuestos	0,500000	-	-
En no expuestos	0,235941	-	-
Razón de prevalencias	2,119171	1,054016	4,260736

Prevalencia de exposición	Estimación	IC(95,0%)	
En enfermos	0,010256	-	-
En no enfermos	0,003190	-	-
Razón de prevalencias	3,215385	0,807922	12,796660

OR	IC(95,0%)		
3,238342	0,806111	13,009196	(Woolf)
	0,883539	11,869266	(Cornfield)

Prueba Ji-cuadrado de asociación	Estadístico	Valor p
Sin corrección	3,0677	0,0799
Corrección de Yates	1,7820	0,1819