Índices de riesgo

(ver 2019-2020)

Emilio Letón y Elisa M. Molanes-López

13-Oct-2019

Página legal

ÍNDICES DE RIESGO

Madrid, versión 2019-2020

ISBN electrónico: xxx.
Edición digital (epub): xxx.

Prefacio

Este material está diseñado con el paradigma del grupo de innovación docente miniXmodular de generar material mini y modular, según se puede ver en la página web www.minixmodular.ia.uned.es, donde se introduce, entre otros, el concepto de mini-libro electrónico modular.

En este mini-libro se contemplan dos variables dicotómicas X e Y donde X es la variable explicativa (también llamada independiente o factor de exposición) e Y es la variable respuesta (también llamada dependiente, resultado u objetivo). De forma esquemática se escribirá $D \leftarrow D$ para indicar que se está ante un modelo estadístico que establece que se está intentando explicar una variable dicotómica a través de otra variable dicotómica.

En esta situación los datos experimentales tendrán el patrón dado en la tabla 1, donde el número de individuos con perfil $Y = 1$ y $X = 1$ es a, con perfil $Y = 1$ y $X = 0$ es b, con perfil $Y = 0$ y $X = 1$ es c y con perfil $Y = 0$ y $X = 0$ es d.

Y	X
1	1
...	...
1	1
1	0
...	...
1	0
0	1
...	...
0	1
0	0
...	...
0	0

Tabla 1: Datos genéricos del esquema $D \leftarrow D$

A partir de los datos experimentales se construye la tabla de datos cruzados genérica dada en la tabla 2, donde los marginales son $r_1 = a + b$, $r_0 = c + d$, $s_1 = a + c$, $s_0 = b + d$ y $n = r_1 + r_0 = s_1 + s_0$ es el tamaño muestral.

	$X = 1$	$X = 0$	
$Y = 1$	a	b	r_1
$Y = 0$	c	d	r_0
	s_1	s_0	n

Tabla 2: Tabla con datos genéricos del esquema $D \leftarrow D$

Es conveniente señalar que a veces el diseño del estudio fija algunos marginales. Por ejemplo en los diseños prospectivos (hacia adelante) se fijan s_1 y s_0, en los retrospectivos (hacia atrás) r_1 y r_0 y en los transversales sólamente se fija n. En los diseños prospectivos se pueden utilizar como índices de riesgos la diferencia de riesgos, el riesgo relativo y el odds ratio. En los disesños retrospectivos y en los transversales el odds ratio.

Los mini-capítulos de este mini-libro se estructuran dependiendo de si los índices de riesgo son absolutos (diferencia de riesgos) o son relativos (riesgo relativo y odds ratio).

Se recomienda refrescar los conceptos básicos de intervalos de confianza (IC) y contraste de hipótesis (CH) en, por ejemplo, [6].

Por último, conviene mencionar que a lo largo de este mini-libro se presenta código en el lenguaje R [5] para realizar los cálculos estadísticos. Existen numerosas páginas web y libros que introducen este lenguaje; dos

referencias clásicas que combinan Estadística y R son [1] y [2]. No obstante, conviene señalar que el código que se utiliza es muy sencillo y autoexplicativo.

Mini-capítulo 1

Índices absolutos

En este mini-capítulo se estudia la diferencia de riesgos que es válida en estudios prospectivos.

1.1. Diferencia de riesgos

El parámetro poblacional θ diferencia de riesgos poblacionales está dado por

$$\theta = dr = \pi_{11} - \pi_{10} = P(Y = 1|X = 1) - P(Y = 1|X = 0)$$

y verifica que $dr \in (-1, 1)$. Para estimar el parámetro dr se considera el estadístico $\widehat{\Theta}$ dado por la variable aleatoria (v.a.) $\widehat{\Pi}_{11} - \widehat{\Pi}_{10}$ (es decir, $\widehat{\Theta} = \widehat{DR} = \widehat{\Pi}_{11} - \widehat{\Pi}_{10}$), siendo $\widehat{\Pi}_{11}$ y $\widehat{\Pi}_{10}$ las proporciones muestrales entendidas como v.a. ya que variarán de muestra a muestra y que aplicado a la muestra de estudio proporciona $\widehat{dr} = \frac{a}{s_1} - \frac{b}{s_0}$.

Al ser la diferencia de riesgos una diferencia de dos proporciones, se tiene que la media teórica de $\widehat{DR}$ está dada por

$$E[\widehat{DR}] = \pi_{11} - \pi_{10}$$

y la varianza teórica de $\widehat{DR}$ está dada por

$$V[\widehat{DR}] = \frac{\pi_{11}(1-\pi_{11})}{s_1} + \frac{\pi_{10}(1-\pi_{10})}{s_0},$$

con lo que el error estándar de $\widehat{DR}$ es

$$EE[\widehat{DR}] = \sqrt{\frac{\pi_{11}(1-\pi_{11})}{s_1} + \frac{\pi_{10}(1-\pi_{10})}{s_0}}.$$

Con la información anterior se define un nuevo estadístico dado por

$$\frac{\widehat{DR} - E[\widehat{DR}]}{EE[\widehat{DR}]} = \frac{\widehat{DR} - (\pi_{11} - \pi_{10})}{\sqrt{\pi_{11}(1-\pi_{11})\frac{1}{s_1} + \pi_{10}(1-\pi_{10})\frac{1}{s_0}}}$$

que sigue una $N(0, 1)$ y que servirá para construir un intervalo de confianza y un contraste de hipótesis para dr.

En el caso del intervalo de confianza hay que estimar $EE[\widehat{DR}]$ a través de

$$\widehat{EE}[\widehat{DR}] = \sqrt{\widehat{\pi}_{11}(1-\widehat{\pi}_{11})\frac{1}{s_1} + \widehat{\pi}_{10}(1-\widehat{\pi}_{10})\frac{1}{s_0}}$$

con $\widehat{\pi}_{11} = \frac{a}{s_1}$ y $\widehat{\pi}_{10} = \frac{b}{s_0}$, con lo que el intervalo de confianza al $(1-\alpha)\,\%$ está dado por

$$\begin{aligned} IC(1-\alpha)\,\%(dr) &= \Big(\widehat{\pi}_{11} - \widehat{\pi}_{10} \mp z_{1-\alpha/2}\sqrt{\widehat{\pi}_{11}(1-\widehat{\pi}_{11})\frac{1}{s_1} + \widehat{\pi}_{10}(1-\widehat{\pi}_{10})\frac{1}{s_0}}\Big) \\ &= \Big(\frac{a}{s_1} - \frac{b}{s_0} \mp z_{1-\alpha/2}\sqrt{\frac{a}{s_1}\Big(1-\frac{a}{s_1}\Big)\frac{1}{s_1} + \frac{b}{s_0}\Big(1-\frac{b}{s_0}\Big)\frac{1}{s_0}}\Big) \\ &= \Big(\frac{a}{s_1} - \frac{b}{s_0} \mp z_{1-\alpha/2}\sqrt{\frac{ac}{s_1^3} + \frac{bd}{s_0^3}}\Big). \end{aligned}$$

En lo que respecta al contraste de hipótesis

$$\begin{aligned} H_0: &\quad dr = 0 \\ H_1: &\quad dr \neq 0 \end{aligned}$$

hay que realizarlo a través del test z de diferencia de proporciones, de Fisher o del test χ^2.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa dicotómica en una variable respuesta dicotómica. Han participado en el estudio 76 individuos. Los datos experimentales se dan en la tabla 1.1, donde el número de individuos con perfil $rta = 1$ y $exp = 1$ es 21, con perfil $rta = 1$ y $exp = 0$ es 16, con perfil $rta = 0$ y $exp = 1$ es 8 y con perfil $rta = 0$ y $exp = 0$ es 31. Se pide calcular la diferencia de riesgos, su IC y su contraste de hipótesis asociado.

rta	exp
1	1
...	...
1	1
1	0
...	...
1	0
0	1
...	...
0	1
0	0
...	...
0	0

Tabla 1.1: Datos observados del esquema $D \leftarrow D$

SOLUCIÓN:

En primer lugar se muestran los cálculos detallados para el intervalo de confianza al 95 % dado por

$$\begin{aligned} IC95\,\%(dr) &= \left(\frac{21}{29} - \frac{16}{47} \mp 1.96\sqrt{\frac{21\cdot 8}{29^3} + \frac{16\cdot 31}{47^3}}\right) \\ &= (0.3837 \mp 1.96 \cdot 0.1080) \\ &= (0.1720, 0.5954), \end{aligned}$$

En segundo lugar el contraste de hipótesis dado por

$$z = \frac{\frac{21}{29} - \frac{16}{47}}{\sqrt{\frac{37}{76}\left(1 - \frac{37}{76}\right)\left(\frac{1}{29} + \frac{1}{47}\right)}} = \frac{0.3837}{\sqrt{0.0139}} = 3.2511$$

con lo que al ser $|z| = 3.2511 > 1.96$ se rechaza la hipótesis nula de que la diferencia de riesgos sea cero y se concluye que la variable explicativa influye en la variable respuesta de forma significativa con un p-valor$= 2 \cdot (1 - pnorm(abs(3.2511)) = 0.0011$.

Y en tercer lugar el contraste de hipótesis utilizando el test χ^2, donde los datos observados se muestran en la tabla 1.2 y los datos esperados en la tabla 1.3 con lo que

$$\begin{aligned} \chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \\ &= \frac{(21 - 14.12)^2}{14.12} + \frac{(16 - 22.88)^2}{22.88} + \frac{(8 - 14.88)^2}{14.88} + \frac{(31 - 24.12)^2}{24.12} \\ &= \frac{6.88^2}{14.12} + \frac{(-6.88)^2}{22.88} + \frac{(-6.88)^2}{14.88} + \frac{6.88^2}{24.12} \\ &= 10.5695 \end{aligned}$$

que tiene asociado un p-valor$= 1 - pchisq(10.5695, 1) = 0.0011$. El hecho de que ambos tests den el mismo resultado es debido a que se puede demostrar que ambos tests son equivalentes en el esquema $D \leftarrow D$.

	$X=1$	$X=0$	
$Y=1$	21	16	37
$Y=0$	8	31	39
	29	47	76

Tabla 1.2: Tabla con datos observados del esquema $D \leftarrow D$

	$X=1$	$X=0$	
$Y=1$	14.12	22.88	37
$Y=0$	14.88	24.12	39
	29	47	76

Tabla 1.3: Tabla con datos esperados del esquema $D \leftarrow D$

Por último se muestra la sintaxis en R para el cálculo del IC y del p-valor del test z de diferencia de dos proporciones y del test χ^2.

Si se supone que los datos sin agrupar están en el fichero d_d_3.txt la sintaxis es

```
rm(list=ls())

datos=read.table('d_d_3.txt',header=T)
attach(datos)

ind1=which(exp==1);
ind0=which(exp==0);
ind11=which(rta==1 & exp==1);
ind10=which(rta==1 & exp==0);
a=length(rta[ind11]); a
b=length(rta[ind10]); b
s1=length(rta[ind1]); s1
s0=length(rta[ind0]); s0
c=s1-a; c
d=s0-b; d
r1=a+b; r1
r0=c+d; r0
n=a+b+c+d; n

alfa=0.05
dr=a/s1-b/s0

ee=sqrt(a*c/s1^3+b*d/s0^3); ee

ic1=dr-qnorm(1-alfa/2)*ee; ic1
ic2=dr+qnorm(1-alfa/2)*ee; ic2

ee0=sqrt((r1/n)*(r0/n)*(1/s1+1/s0))
z=(dr)/ee0; z
z*z

p_valor= 2*(1-pnorm(abs(z))); p_valor

ab=c(a,b); ab
s1s0=c(s1,s0); s1s0
prop.test(ab,s1s0,correct=F)

tabla=table(-rta,-exp); tabla
```

```
chisq.test(tabla)\$expected

chisq.test(tabla,correct=FALSE)
```

En el caso de que se introduzcan directamente los datos agrupados, hay que sustituir las primeras líneas de la sintaxis anterior por

```
rm(list=ls())

tabla <- matrix(c(21,16,8,31),ncol=2,byrow=TRUE)
colnames(tabla) <- c("exp=1","exp=0")
rownames(tabla) <- c("rta=1","rta=0")
tabla <- as.table(tabla); tabla
chisq.test(tabla,correct=FALSE)
```

Mini-capítulo 2

Índices relativos

En este mini-capítulo se estudia el riesgo relativo y el odds ratio. El riesgo relativo es válido en estudios prospectivos y el odds ratio en prospectivos, retrospectivos y transversales. A la hora de calcular intervalos de confianza para estos índices habrá que utilizar el teorema de Taylor para v.a. (ver, por ejemplo, [4]) que afirma que si X es una v.a y h una función, se verifica que

$$\begin{aligned} E[h(X)] &\approx h(E[X]) \\ V[h(X)] &\approx h(V[X])(h'(E[X]))^2 \end{aligned}$$

2.1. Riesgo relativo

El parámetro poblacional θ riesgo relativo poblacional está dado por

$$\theta = rr = \frac{\pi_{11}}{\pi_{10}} = \frac{P(Y=1|X=1)}{P(Y=1|X=0)}$$

y verifica que $rr \in (0, \infty)$. Para estimar el parámetro rr se considera el estadístico $\widehat{\Theta}$ dado por la v.a. $\frac{\widehat{\Pi}_{11}}{\widehat{\Pi}_{10}}$ (es decir, $\widehat{\Theta} = \widehat{RR} = \frac{\widehat{\Pi}_{11}}{\widehat{\Pi}_{10}}$), siendo $\widehat{\Pi}_{11}$ y $\widehat{\Pi}_{10}$ las proporciones muestrales entendidas como v.a. ya que variarán de muestra a muestra y que aplicado a la muestra de estudio proporciona $\widehat{rr} = \frac{a/s_1}{b/s_0}$.

El rr es un índice relativo de asociación entre dos variables dicotómicas y admite, por su definición, una interpretación directa. Por ejemplo:

- Si $rr = 3 > 1$, la probabilidad de presencia de Y en los individuos con $X = 1$ es el triple que la probabilidad de presencia de Y en los individuos con $X = 0$.
- Si $rr = 1$, hay la misma probabilidad de presencia de Y en los individuos con $X = 1$ que en lo individuos con $X = 0$.
- Si $rr = \frac{1}{3}$, la probabilidad de presencia de Y en los individuos con $X = 1$ es un tercio de la probabilidad de presencia de Y en los individuos con $X = 0$.

En el caso de que Y sea la variable "Enfermedad" codificada con 1 en el caso de "Sí enfermo" y con 0 en el caso de "No enfermo" y que X sea la variable "Exposición" codificada con 1 en el caso de "Sí expuesto" y con 0 en el caso de "No expuesto", la interpretación del rr se puede explicitar de forma más clara. Por ejemplo:

- Si $rr = 3 > 1$, el hecho de estar expuesto multiplica por 3 el riesgo (la probabilidad) de enfermar (es decir, X es un factor de riesgo).
- Si $rr = 1$, hay la misma probabilidad de enfermar en los expuestos que en los no expuestos (es decir, X no es ni un factor de riesgo ni un factor protector).
- Si $rr = \frac{1}{3}$, el hecho de estar expuesto divide por 3 el riesgo (la probabilidad) de enfermar (es decir, X es un factor protector).

Al ser $\widehat{RR}$ una v.a. asimétrica (su rango es $(0, \infty)$), su distribución es no normal. Es conveniente, por tanto, considerar una transformación que consiga normalidad con media teórica y varianza teórica que se puedan estimar fácilmente. Se puede demostrar que la transformación logaritmo neperiano (Ln) consigue este propósito. A efectos teóricos, ahora es $\theta = Ln(rr)$ y su estimador es $\widehat{\Theta} = Ln(\widehat{RR})$.

En este contexto se puede construir para $Ln(rr)$, ver por ejemplo [3], un intervalo de confianza al $(1-\alpha)\,\%$ dado por

$$IC(1-\alpha)\,\%(Ln(rr)) \quad = \quad \Big(\widehat{E}[Ln(\widehat{RR})] \mp z_{1-\alpha/2}\widehat{EE}[Ln(\widehat{RR})]\Big)$$

con

$$\widehat{E}[Ln(\widehat{RR})] = Ln(rr) = Ln\Big(\frac{a/s_1}{b/s_0}\Big)$$

y

$$\widehat{EE}[Ln(\widehat{RR})] = \sqrt{\frac{1}{a}-\frac{1}{s_1}+\frac{1}{b}-\frac{1}{s_0}}.$$

Deshaciendo la transformación del logaritmo neperiano, se tiene que

$$IC(1-\alpha)\,\%(rr) \quad = \quad \Big(exp\Big(Ln\Big(\frac{a/s_1}{b/s_0}\Big) \mp z_{1-\alpha/2}\sqrt{\frac{1}{a}-\frac{1}{s_1}+\frac{1}{b}-\frac{1}{s_0}}\Big)\Big)$$

En lo que respecta al contraste de hipótesis

$$\begin{aligned} H_0: &\quad rr = 1 \\ H_1: &\quad rr \neq 1, \end{aligned}$$

dado que $rr = 1 \Leftrightarrow dr = 0$ hay que realizarlo a través del test z de diferencia de proporciones, de Fisher o del test χ^2.

Ejemplo

1. Se ha realizado un estudio para ver si influye una variable explicativa dicotómica en una variable respuesta dicotómica. Han participado en el estudio 76 individuos. Los datos experimentales se dan en la tabla 1.1, donde el número de individuos con perfil $rta = 1$ y $exp = 1$ es 21, con perfil $rta = 1$ y $exp = 0$ es 16, con perfil $rta = 0$ y $exp = 1$ es 8 y con perfil $rta = 0$ y $exp = 0$ es 31. Se pide calcular el riesgo relativo, su IC y su contraste de hipótesis asociado.

   SOLUCIÓN:

   En primer lugar se muestran los cálculos detallados para el intervalo de confianza al 95 % dado por

$$\begin{aligned} IC(95\,\%(Ln(rr)) &= \Big(exp\Big(Ln\Big(\frac{21/29}{16/47}\Big) \mp 1.96\sqrt{\frac{1}{21}-\frac{1}{29}+\frac{1}{16}-\frac{1}{47}}\Big)\Big) \\ &= (0.7548 \mp 1.96 \cdot 0.2332) \\ &= (0.2978, 1.2118), \end{aligned}$$

   y deshaciendo la transformación se tiene que el $IC(95\,\%)(rr) = (exp(0.2978), exp(1.2118)) = (1.3469, 3.3594)$.

   En segundo lugar el contraste de hipótesis dado por el test χ^2, es

$$\chi^2 \quad = \quad \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-e_{ij})^2}{e_{ij}} = 10.5695$$

   que tiene asociado un p-valor$= 1 - pchisq(10.5695, 1) = 0.0011$.

   Por último se muestra la sintaxis en R para el cálculo del riesgo relativo, su IC y su contraste de hipótesis asociado mediante el test χ^2.

   Si se supone que los datos sin agrupar están en el fichero d_d_3.txt la sintaxis es

```
rm(list=ls())

datos=read.table('d_d_3.txt',header=T)
attach(datos)
```

```
ind1=which(exp==1);
ind0=which(exp==0);
ind11=which(rta==1 & exp==1);
ind10=which(rta==1 & exp==0);
a=length(rta[ind11]); a
b=length(rta[ind10]); b
s1=length(rta[ind1]); s1
s0=length(rta[ind0]); s0
c=s1-a; c
d=s0-b; d
r1=a+b; r1
r0=c+d; r0
n=a+b+c+d; n

alfa=0.05
rr=(a/s1)/(b/s0)
ee.lnrr=sqrt(1/a-1/s1+1/b-1/s0)
ic1=exp(log(rr)-qnorm(1-alfa/2)*ee.lnrr); ic1
ic2=exp(log(rr)+qnorm(1-alfa/2)*ee.lnrr); ic2

tabla=table(-rta,-exp); tabla
chisq.test(tabla,correct=FALSE)
```

En el caso de que se introduzcan directamente los datos agrupados, hay que sustituir las primeras líneas de la sintaxis anterior por

```
rm(list=ls())

tabla <- matrix(c(21,16,8,31),ncol=2,byrow=TRUE)
colnames(tabla) <- c("exp=1","exp=0")
rownames(tabla) <- c("rta=1","rta=0")
tabla <- as.table(tabla); tabla
a=tabla[1,1]; b=tabla[1,2]; c=tabla[2,1]; d=tabla[2,2]
s1=a+c; s0=b+d
alfa=0.05
rr=(a/s1)/(b/s0)
ee.lnrr=sqrt(1/a-1/s1+1/b-1/s0)
ic1=exp(log(rr)-qnorm(1-alfa/2)*ee.lnrr); ic1
ic2=exp(log(rr)+qnorm(1-alfa/2)*ee.lnrr); ic2
chisq.test(tabla,correct=FALSE)
```

2. Utilizando el teorema de Taylor para v.a., se pide demostrar que

$$E[Ln(\widehat{\Pi}_{11})] \approx Ln(\pi_{11}),$$
$$E[Ln(\widehat{\Pi}_{10})] \approx Ln(\pi_{10}),$$
$$V[Ln(\widehat{\Pi}_{11})] \approx \frac{1-\pi_{11}}{\pi_{11}}\frac{1}{s_1}$$

y

$$V[Ln(\widehat{\Pi}_{10})] \approx \frac{1-\pi_{10}}{\pi_{10}}\frac{1}{s_0}.$$

SOLUCIÓN:

En el contexto del teorema de Taylor para v.a. se tiene que $h(\pi) = Ln(\pi)$, por tanto $E[Ln(\widehat{\Pi}_{11})] \approx Ln(E[\widehat{\Pi}_{11})] \approx Ln(\pi_{11})$ (de forma análoga $E[Ln(\widehat{\Pi}_{10})] \approx Ln(\pi_{10})$). Por otra parte, $h'(\pi)) = \frac{1}{\pi}$, con lo que

$$\begin{aligned} V[Ln(\widehat{\Pi}_{11})] &\approx V[\widehat{\Pi}_{11}](h'(E[\widehat{\Pi}_{11}])^2) \\ &= \pi_{11}(1-\pi_{11})\frac{1}{s_1}(h'(\pi_{11}))^2 = \pi_{11}(1-\pi_{11})\frac{1}{\pi_{11}^2} \\ &= \frac{1-\pi_{11}}{\pi_{11}}\frac{1}{s_1} \end{aligned}$$

y, de forma análoga, $V[Ln(\widehat{\Pi}_{10})] \approx \dfrac{1-\pi_{10}}{\pi_{10}}\dfrac{1}{s_0}$.

2.2. Odds ratio

En general, dada π una proporción (probabilidad) cualquiera, se define el parámetro poblacional o, odds de una proporción, como el cociente de las probabilidades complementarias. Es decir, $o = \frac{\pi}{1-\pi}$, con lo que $o \in (0, \infty)$. En el caso de que se esté considerando $\pi = P(Y = 1)$, se tiene que el odds de presencia de Y es $\frac{P(Y=1)}{1-P(Y=1)}$ y, por tanto, $P(Y = 1) = \frac{o}{o+1}$. En el caso de considerar el odds de enfermedad, su valor se interpreta de la forma siguiente:

- Si $o = 3 > 1$, se tiene que la probabilidad de estar enfermo es el triple de la de estar sano, o lo que es lo mismo que $P(Y = 1) = \frac{3}{3+1} = 0.75$ y $P(Y = 0) = 0.25$.
- Si $o = 1$, hay la misma probabilidad de estar enfermo que de estar sano.
- Si $o = \frac{1}{3}$, se tiene que la probabilidad de estar enfermo es un tercio de la de estar sano, o lo que es lo mismo que $P(Y = 1) = \frac{1/3}{1/3+1} = 0.25$ y $P(Y = 0) = 0.75$.

Si además se incorpora el conocimiento extra de X, se puede hablar del odds de presencia de Y en los individuos $X = 1$, $o_{X=1}$, y del odds de presencia de Y en los individuos $X = 0$, $o_{X=0}$, que vienen dados por

$$o_{X=1} = \frac{\pi_{11}}{1-\pi_{11}} = \frac{P(Y = 1|X = 1)}{1 - P(Y = 1|X = 1)}$$

$$o_{X=0} = \frac{\pi_{10}}{1-\pi_{10}} = \frac{P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)}.$$

Con estos preliminares, se define el parámetro poblacional or_X, el odds ratio de presencia de Y respecto a X como la razón de los odds $o_{X=1}$ y $o_{X=0}$, es decir

$$or_X = \frac{o_{X=1}}{o_{X=0}} = \frac{\frac{\pi_{11}}{1-\pi_{11}}}{\frac{\pi_{10}}{1-\pi_{10}}} = \frac{\dfrac{P(Y = 1|X = 1)}{1 - P(Y = 1|X = 1)}}{\dfrac{P(Y = 1|X = 0)}{1 - P(Y = 1|X = 0)}}$$

y el parámetro poblacional or_Y, el odds ratio de presencia de X respecto a Y como la razón de los odds $o_{Y=1}$ y $o_{Y=0}$, es decir

$$or_Y = \frac{o_{Y=1}}{o_{Y=0}} = \frac{\frac{\pi'_{11}}{1-\pi'_{11}}}{\frac{\pi'_{10}}{1-\pi'_{10}}} = \frac{\dfrac{P(X = 1|Y = 1)}{1 - P(X = 1|Y = 1)}}{\dfrac{P(X = 1|Y = 0)}{1 - P(X = 1|Y = 0)}}$$

Utilizando el teorema de Bayes, es fácil ver que $or_X = or_Y$, es decir que el odds ratio de presencia de Y respecto a X es el mismo que el odds ratio de presencia de X respecto a Y. Por ello, se puede hablar simplemente de odds ratio, or y se puede calcular en todo tipo de estudios. El parámetro poblacional $\theta = or$ verifica que $or \in (0, \infty)$. Para estimar el parámetro or se considera el estadístico $\widehat{\Theta}$ dado por la v.a.

$$\widehat{\Theta} = \widehat{OR} = \frac{\frac{\widehat{\Pi}_{11}}{1-\widehat{\Pi}_{11}}}{\frac{\widehat{\Pi}_{10}}{1-\widehat{\Pi}_{10}}},$$

siendo $\widehat{\Pi}_{11}$ y $\widehat{\Pi}_{10}$ las proporciones muestrales entendidas como v.a. ya que variarán de muestra a muestra. Aplicando este estadístico a la muestra de estudio se obtiene $\widehat{or} = \dfrac{\frac{a/s_1}{c/s_1}}{\frac{b/s_0}{d/s_0}} = \frac{ad}{bc}$, motivo por el que a veces al odds ratio también se le denomina razón de productos cruzados (aunque en algunos textos también se le llama razón de disparidades o razón de posibilidades).

El or es un índice relativo de asociación entre dos variables dicotómicas y se puede demostrar que

- $or > 1 \Leftrightarrow P(Y = 1|X = 1) > P(Y = 1|X = 0)$, por lo que la presencia de X favorece la presencia de Y.
- $or = 1 \Leftrightarrow P(Y = 1|X = 1) = P(Y = 1|X = 0)$, por lo que la presencia de X ni favorece ni desfavorece la presencia de Y.

- $or < 1 \Leftrightarrow P(Y=1|X=1) < P(Y=1|X=0)$, por lo que la presencia de X desfavorece la presencia de Y.

En el caso de que Y sea la variable "Enfermedad" codificada con 1 en el caso de "Sí enfermo" y con 0 en el caso de "No enfermo" y que X sea la variable "Exposición" codificada con 1 en el caso de "Sí expuesto" y con 0 en el caso de "No expuesto", la interpretación anterior del *or* se puede hacer de forma más clara:

- Si $or > 1 \Leftrightarrow X$ es un factor de riesgo.
- Si $or = 1 \Leftrightarrow X$ no es ni un factor de riesgo ni un factor protector
- Si $or < 1 \Leftrightarrow X$ es un factor protector.

Al ser $\widehat{OR}$ una v.a. asimétrica (su rango es $(0,\infty)$), su distribución es no normal. Es conveniente, por tanto, considerar una transformación que consiga normalidad con media teórica y varianza teórica que se puedan estimar fácilmente. Se puede demostrar que la transformación logaritmo neperiano (Ln) consigue este propósito. A efectos teóricos, ahora es $\theta = Ln(or)$ y su estimador es $\widehat{\Theta} = Ln(\widehat{OR})$.

En este contexto se puede construir para $Ln(or)$, ver por ejemplo [3], un intervalo de confianza al $(1-\alpha)\,\%$ dado por

$$IC(1-\alpha)\,\%(Ln(or)) \quad = \quad \Big(\widehat{E}[Ln(\widehat{OR})] \mp z_{1-\alpha/2}\widehat{EE}[Ln(\widehat{OR})]\Big)$$

con

$$\widehat{E}[Ln(\widehat{OR})] = Ln(or) = Ln\Big(\frac{ad}{bc}\Big)$$

y

$$\widehat{EE}[Ln(\widehat{OR})] = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}.$$

Deshaciendo la transformación del logaritmo neperiano, se tiene que

$$IC(1-\alpha)\,\%(or) \quad = \quad \Big(exp\Big(Ln\Big(\frac{ad}{bc}\Big) \mp z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\Big)\Big)$$

En lo que respecta al contraste de hipótesis

$$\begin{aligned} H_0: &\quad or = 1 \\ H_1: &\quad or \neq 1, \end{aligned}$$

dado que $or = 1 \Leftrightarrow rr = 1 \Leftrightarrow dr = 0$ hay que realizarlo a través del test z de diferencia de proporciones, de Fisher o del test χ^2.

A la hora de deducir propiedades acerca del *or*, conviene introducir el concepto de transforamción logit de una proporción π que viene dada por

$$h(\pi) = logit(\pi) = Ln\Big(\frac{\pi}{1-\pi}\Big)$$

y que verifica que

$$\pi = \frac{e^{logit(\pi)}}{1+e^{logit(\pi)}}.$$

En el caso del parámetro poblacional $logit(\pi)$, su IC viene dado por

$$IC(1-\alpha)\,\%(logit(\pi)) \quad = \quad \Big(\widehat{E}[logit(\widehat{\Pi})] \mp z_{1-\alpha/2}\widehat{EE}[logit(\widehat{\Pi})]\Big)$$

con $\widehat{E}[logit(\widehat{\Pi})] = Ln(\frac{r_1}{r_0})$ y $\widehat{EE}[logit(\widehat{\Pi})] = \sqrt{\frac{n}{r_1 r_0}}$.

Por último, hay que señalar que en el caso de que la $P(Y=1)$ sea pequeña, el *or* es próximo al rr.

Ejemplo

1. Utilizando el teorema de Bayes se pide demostrar que $or_X = or_Y$.

   SOLUCIÓN:

   En primer lugar se observa utilizando el teorema de Bayes que:

   - $\pi_{11} = P(Y=1|X=1) = \dfrac{P(X=1|Y=1)P(Y=1)}{P(X=1|Y=1)P(Y=1)+P(X=1|Y=0)P(Y=0)}$.
   - $1-\pi_{11} = P(Y=0|X=1) = \dfrac{P(X=1|Y=0)P(Y=0)}{P(X=1|Y=0)P(Y=0)+P(X=1|Y=1)P(Y=1)}$.
   - $\pi_{10} = P(Y=1|X=0) = \dfrac{P(X=0|Y=1)P(Y=1)}{P(X=0|Y=1)P(Y=1)+P(X=0|Y=0)P(Y=0)}$.
   - $1-\pi_{10} = P(Y=0|X=0) = \dfrac{P(X=0|Y=0)P(Y=0)}{P(X=0|Y=0)P(Y=0)+P(X=0|Y=1)P(Y=1)}$.
   - $\pi'_{11} = P(X=1|Y=1) = \dfrac{P(Y=1|X=1)P(X=1)}{P(Y=1|X=1)P(X=1)+P(Y=1|X=0)P(X=0)}$.
   - $1-\pi'_{11} = P(X=0|Y=1) = \dfrac{P(Y=1|X=0)P(X=0)}{P(Y=1|X=0)P(X=0)+P(Y=1|X=1)P(X=1)}$.
   - $\pi'_{10} = P(X=1|Y=0) = \dfrac{P(Y=0|X=1)P(X=1)}{P(Y=0|X=1)P(X=1)+P(Y=0|X=0)P(X=0)}$.
   - $1-\pi'_{10} = P(X=0|Y=0) = \dfrac{P(Y=0|X=0)P(X=0)}{P(Y=0|X=0)P(X=0)+P(Y=0|X=1)P(X=1)}$.

   Por tanto

   $$\begin{aligned} or_X &= \frac{\frac{\pi_{11}}{1-\pi_{11}}}{\frac{\pi_{10}}{1-\pi_{10}}} = \frac{\dfrac{P(X=1|Y=1)P(Y=1)}{P(X=1|Y=0)P(Y=0)}}{\dfrac{P(X=0|Y=1)P(Y=1)}{P(X=0|Y=0)P(Y=0)}} \\ &= \frac{\dfrac{P(X=1|Y=1)}{P(X=1|Y=0)}}{\dfrac{P(X=0|Y=1)}{P(X=0|Y=0)}} = \frac{\dfrac{P(X=1|Y=1)}{P(X=0|Y=1)}}{\dfrac{P(X=1|Y=0)}{P(X=0|Y=0)}} = \frac{\dfrac{P(X=1|Y=1)P(X=1)}{P(X=0|Y=1)P(X=0)}}{\dfrac{P(X=1|Y=0)P(X=1)}{P(X=0|Y=0)P(X=0)}} \\ &= \frac{\frac{\pi'_{11}}{1-\pi'_{11}}}{\frac{\pi'_{10}}{1-\pi'_{10}}} = or_Y \end{aligned}$$

2. Utilizando el teorema de Taylor para v.a., se pide demostrar que

   $$E[logit(\widehat{\Pi}_{11})] \approx logit(\pi_{11}),$$
   $$E[logit(\widehat{\Pi}_{10})] \approx logit(\pi_{10}),$$
   $$V[logit(\widehat{\Pi}_{11})] \approx \frac{1}{\pi_{11}(1-\pi_{11})}\frac{1}{s_1}$$

   y

   $$V[logit(\widehat{\Pi}_{10})] \approx \frac{1}{\pi_{10}(1-\pi_{10})}\frac{1}{s_0}.$$

   SOLUCIÓN:

   En el contexto del teorema de Taylor para v.a. se tiene que $h(\pi) = logit(\pi) = Ln\left(\frac{\pi}{1-\pi}\right)$, por tanto $E[logit(\widehat{\Pi}_{11})] \approx logit(E[\widehat{\Pi}_{11}]) = logit(\pi_{11})$ (de forma análoga $E[logit(\widehat{\Pi}_{10})] \approx logit(\pi_{10})$). Por otra parte, $h'(\pi) = \dfrac{1}{\frac{\pi}{1-\pi}}\dfrac{(1-\pi)-\pi(-1)}{(1-\pi)^2} = \dfrac{1}{\pi(1-\pi)}$, con lo que

   $$\begin{aligned} V[logit(\widehat{\Pi}_{11})] &\approx V[\widehat{\Pi}_{11}](h'(E[\widehat{\Pi}_{11}])^2) = \pi_{11}(1-\pi_{11})\frac{1}{s_1}(h'(\pi_{11}))^2 \\ &= \pi_{11}(1-\pi_{11})\frac{1}{s_1}\frac{1}{(\pi_{11}(1-\pi_{11}))^2} = \frac{1}{\pi_{11}(1-\pi_{11})}\frac{1}{s_1} \end{aligned}$$

   y, de forma análoga, $V[logit(\widehat{\Pi}_{10})] \approx \dfrac{1}{\pi_{10}(1-\pi_{10})}\dfrac{1}{s_0}$.

Referencias

[1] Davies, T.M., *The Book of R: A first course in programming and statistics*, No Starch Press, 2016.

[2] García-Pérez, A., *Estadística básica con R*, Universidad Nacional de Educación a Distancia (UNED) (Madrid), 2010.

[3] Letón, E., Pedromingo, A., *Introducción al análisis de datos en meta-análisis*, Díaz de Santos (Madrid), 2001.

[4] Peña, D., *Fundamentos de Estadística*, Alianza editorial (Madrid), 2014.

[5] R Core Team, *R: A Language and Environment for Statistical Computing. R*, Foundation for Statistical Computing, Vienna, Austria, 2017. URL https://www.R-project.org/.

[6] Ramos, E., Vélez, R., Hernández, V., *Modelos probabilísticos y optimización*, Sanz y Torres (Madrid), 2019.

Acerca de

Emilio Letón

Figura 2.1: Emilio Letón

Emilio Letón nace en Madrid en 1966. Es Licenciado en Matemáticas por la UCM en 1989 y doctor en Matemáticas en la misma universidad en 2002. En la actualidad es profesor contratado doctor de la UNED en el departamento de Inteligencia Artificial, al que se incorporó en 2009. Anteriormente fue profesor del departamento de Estadística de la UC3M durante 5 años. Asimismo, ha trabajado durante 15 años en departamentos de Planificación y Estadística dentro del sector bancario y de la industria farmacéutica. Sus líneas de investigación incluyen el Análisis de Supervivencia, tests no paramétricos, PLS, Meta-Análisis, Bioestadística y B-Learning. Ha participado en más de 30 proyectos de innovación docente (siendo coordinador en más de 10 de ellos) colaborando con distintas universidades: UNED, UC3M, UCM y UPM. Ha recibido 1 premio en excelencia en publicaciones científicas (UC3M) y 5 premios en excelencia docente (1 en UC3M, 3 en UNED y 1 en OCW Consortium). En @emilioleton se pueden encontrar sus tweets y su página web personal con información ampliada de su curriculum. https://twitter.com/emilioleton

Elisa M. Molanes-López

Figura 2.2: Elisa M. Molanes-López

Elisa M. Molanes López nace en Vigo en 1976. Se licencia en Matemáticas por la Universidad de Santiago de Compostela en 2000 y consigue el grado de doctora, con acreditación europea, por la Universidad de A Coruña en 2007. Entre 2003 y 2007, al amparo de una beca FPI, realiza varias estancias de investigación

en las siguientes universidades extranjeras: Universiteit Hasselt (Bélgica), Université Catholique de Louvain (Bélgica) y The University of Texas (EE.UU.). Durante 8 años, entre 2007 y 2015, es profesora visitante en el departamento de Estadística de la Universidad Carlos III de Madrid. Actualmente, desde octubre de 2015, es profesora ayudante doctora en la Unidad Departamental de Bioestadística del Departamento de Estadística e Investigación Operativa de la Universidad Complutense de Madrid. Durante su etapa docente ha participado en 10 proyectos de innovación docente. En diciembre de 2013, recibe un accésit a la mejor práctica docente en los Premios del Consejo Social de la UNED por su participación en el MOOC "Mini-vídeos docentes modulares: un elemento crítico en el diseño de un MOOC", y en abril de 2014, el OpenCourseWare Consortium le otorga un premio de excelencia por su participación en el curso OCW "Mini-vídeos docentes modulares para diseñar un MOOC". Sus líneas de investigación incluyen la estadística no paramétrica, el análisis de supervivencia, las curvas ROC y las funciones cópula. En la página web personal de Elisa M. Molanes-López, se puede encontrar información más detallada. https://elisamariamolanes.wixsite.com/emolanes

Índice general

1. Índices absolutos **1**
1.1. Diferencia de riesgos 1

2. Índices relativos **5**
2.1. Riesgo relativo 5
2.2. Odds ratio 8

Índice de figuras

2.1. Emilio Letón 13
2.2. Elisa M. Molanes-López 13

Índice de tablas

1. Datos genéricos del esquema $D \leftarrow D$ III
2. Tabla con datos genéricos del esquema $D \leftarrow D$ III

1.1. Datos observados del esquema $D \leftarrow D$ 2
1.2. Tabla con datos observados del esquema $D \leftarrow D$ 3
1.3. Tabla con datos esperados del esquema $D \leftarrow D$ 3