Tema 16: Contraste de dos variables cualitativas. Odds ratios.

En el contraste de dos variables cualitativas hay que ver 1) si se trata de datos independientes o apareados 2) el número de modalidades de las variables (dos o más de dos). Ya que se utilizan fórmulas distintas

A) Contraste de 2 variables cualitativas con datos independientes

Como en toda prueba con datos independientes los problemas de comparación y de relación se resuelven por las mismas fórmulas, ya que son dos formas distintas de enfocar el mismo problema..

Responden a las preguntas:

----la frecuencia (absoluta, relativa o porcentaje) de una característica ¿es similar en los grupos o muestras contrastados?.

En caso afirmativo se trata de una **prueba de comparación**. H_0 : no hay diferencias significativas entre las frecuencias contrastadas, las diferencias observadas se deben a las variaciones normales por el azar.

----¿hay relación o dependencia entre las muestras contrastadas?

En caso afirmativo es una **prueba de relación**. H_0: NO hay relación o dependencia.

Fórmulas

En función del nº de modalidades y de los datos aplicaremos una de las fórmulas siguientes:

- Cuando ambas variables tienen dos modalidades:
  *** **Fórmula nº 1** : para contraste de proporciones o porcentajes
  *** **Fórmula nº 2** : para el contraste de frecuencias absolutas
- Si una o ambas variables tienen más de dos modalidades:
  *** **Fórmula nº 3** : en la práctica sólo se utilizan frecuencias absolutas

(es más fácil utilizar porcentajes que proporciones)

Fórmula nº 1

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_0 q_0}{N_1} + \frac{p_0 q_0}{N_2}}} \qquad \text{siendo } \mathbf{p_0} = \frac{\mathbf{N_1 p_1 + N_2 p_2}}{\mathbf{N_1 + N_2}}$$

Valoración: *si N1 y N2* ≥30, por la DN

si N1 ó N2 <30

a) si p_0N_1 , q_0N_1, p_0N_2 y q_0N_2 ≥5 (ó 500 si es %) por DN

b) ***si algún producto <5 pero > 3 : por t (N_1+N_2-2)***

si algún producto < 3 : por p exacta de Fisher

***Ejercicio 1.1

En una muestra de 100 varones encontramos un 70% de fumadores. En una muestra de 200 mujeres hay 80 fumadoras. ¿Hay diferencias importantes en el hábito de fumar entre ambos sexos?

---- Se trata de un problema de contraste entre dos variables CL (Sexo , Hábito de fumar) con dos modalidades cada una (Hombre, Mujer y Sí, No) con datos independientes. H_0 : no hay diferencias significativas entre los variables contrastadas.

Se puede resolver por la fórmula nº 1 (contraste de dos porcentajes) o por la fórmula nº 2 (contraste de 2 frecuencias). Lo haremos por ambas, pero si podemos elegir, es preferible la nº 2. En la nº 1 es mejor utilizar % que proporciones.

Empezamos por la nº 1 : Hemos medido el hábito de fumar en hombres y en mujeres.
Por el enunciado o mediante un pequeño cálculo sabemos que
p_1=70 ; N_1 = 100 ; p_2 = 40 ; N_2 = 200 ; p_0 = 50 ; q_0 = 50

$$Z=\frac{70-40}{\sqrt{\frac{50*50}{100}+\frac{50*50}{200}}}=4,90$$

Como N_1 y N_2 son > 30, se valora por c de la DN
$Z > c_{0,001} = 3,30 \rightarrow$ rechazo de H_0 : y aceptación de H_1 al nivel de significación de 0,001. ; **p < 0,001**
Sentido: el % de hombres fumadores es más alto..
Y contestando a la pregunta: **Sí** hay diferencias importantes, los hombres fuman más

Fórmula nº 2

$$Z=\frac{N\left(a_1b_2-a_2b_1\right)^2}{N_aN_bN_1N_2}$$

tabla:

a_1	a_2	N_a
b_1	b_2	N_b
N_1	N_2	N

Condición de aplicación: todas las E ≥ 5

Valoración: por χ^2_1

***Si* alguna E<5 , pero ≥ 3: usar fórmula de Yates**

Si alguna E<3: calcular p exacta de Fisher

Fórmula de Yates:

$$Z=\frac{N\left(\left|a_1b_2-a_2b_1\right|-\frac{N}{2}\right)^2}{N_aN_bN_1N_2}$$

***Ejercicio 1.2 : vamos a resolver el ejercicio anterior por la fórmula nº 2

1---Se construye la tabla de 2x2:

	Fuma	No fuma	
Hombre	70	30	*100*
Mujer	80	120	*200*
Total	*150*	*150*	300

2--- se comprueba condición de aplicación:
cumple, pues la E más baja (en a_1 y a_2) vale 50 y es > 5
3---se calcula Z

$$Z=\frac{300*(70*120-80*30)}{100*200*150*150}=24$$

4---Se valora por χ^2 con gl = 1 $Z > \chi^2$ (1, 0,001) = 10,83 ; **p < 0,001**
Por tanto se rechaza H_0 y se acepta H_1 : hay diferencias significativas a nivel de 0,001 entre las frecuencias de fumadores en hombres y mujeres. Sentido: los hombres fuman más. **Sí** hay diferencias importantes.

***Ejercicio 1.3 En una muestra de 20 personas de la tercera edad de la ciudad A el 30% tiene un colesterol alto. En la ciudad B lo tienen el 50% de una muestra de 30. ¿Es importante esa diferencia?

----Es un problema de comparación entre dos variables CL con 2 modalidades cada una y datos independientes: COLESTEROL (alto, no alto) y CIUDAD (A , B)

H_0 : no hay diferencias significativas entre las variables contrastadas, las diferencias observadas se explican por las variaciones normales del azar.

La tabla guía nos indica que lo podemos resolver por la fórmula 1 ó la fórmula 2

Vamos a hacerlo a efectos didácticos, por ambas. Es más fácil, utilizar la nº 2.

1.3.1 Resolución por la fórmula nº 1

Por el enunciado o haciendo un pequeño cálculo se sabe que

$p_1=30$; $N_1= 20$; $p_2=50$; $N_2 = 30$; $p_0 = 42$; $q_0 = 58$

$$Z = \frac{30-50}{\sqrt{\frac{42*58}{20}+\frac{42*58}{30}}} = -1,4$$

Como una muestra es pequeña, hay que ver lo que valen los productos de ambas N por p_0 y q_0. Todos son > 500 (el menor: 20*42=840), por lo que la Z se valora por la DN.

$|Z| < 1,96 \rightarrow$ No puede rechazarse H_0 . p > 0,05 n.s.

Y contestando a la pregunta: la diferencia no es importante.

1.3.2 Resolución por la fórmula nº 2

1---se construye la tabla de 2x2:

		Ciudad		
		A	B	
Col	alto	6	15	*21*
	no alto	14	15	*29*
		20	*30*	50

2---cumple la condición de aplicación: la E más baja (a_1) vale 21*20/50 = 8,4 que es > 5

3---se calcula Z :

$$Z = \frac{50*(6*15-15*14)}{21*29*20*30} = 1,97$$

4---se valora por χ^2 con g l = 1 ; $Z < \chi^2$ (1, 0,05) = 3,84 $\rightarrow$ No puede rechazarse H_0 . p > 0,05 n.s. Contestando a la pregunta : la diferencia no es importante.

***Ejercicio 1.4 En un colegio se hace una encuesta en busca de miopes. Hay 4 entre 20 chicos y 7 entre 28 chicas. Valore la afirmación: la miopía es más frecuente entre las chicas.

--- Es un problema de contraste entre dos variables CL con 2 modalidades cada una : MIOPIA (sí, no) y SEXO (chico , chica). Datos independientes. A resolver por la fórmula nº 1 ó la nº 2.

H_0 : no hay diferencias significativas entre los variables contrastadas.

---no vemos en detalle la resolución por la fórmula nº 1. p0 vale 22,9% y qo 77,1%. Se obtiene una Z = -0,406, que hay que valorar por t(46 , 0'05) = 2,014 . $|Z| < t \rightarrow$No puede rechazarse H_0 . La afirmación no está justificada estadísticamente.

---resolución por la fórmula nº 2 :

1---construir la tabla:

		Miopía		
		Sí	No	
Sexo	Chico	4	16	*20*
	Chica	7	21	*28*
		11	*37*	48

2---Hay una E < 5 (la a1, que vale 4,6) $\rightarrow$ fórmula de Yates

3-- $$Z = \frac{48*\left(|(4*21)-(7*16)|-\frac{48}{2}\right)^2}{20*28*37*11} = 0,003$$

4--- $Z<\chi^2$ (1, 0'05) = 3,84 $\rightarrow$No puede rechazarse H_0. p>0,05 n.s. La afirmación no está justificada

***Ejercicio 1.5 Se estudia el efecto de la vacuna BCG en la prevención de la TBC (tuberculosis) en el pueblo X de un país en vías de desarrollo. Hay 10 enfermos entre 70 vacunados y 80 enfermos entre 120 no vacunados. ¿Tiene la vacuna efecto preventivo?

--Es un problema de contraste de 2 variables CL con dos modalidades cada una y datos independientes: BCG (sí , no) y TBC (sí , no). A resolver por la fórmula nº 1 o la nº 2. Lo haremos por la nº 2, pues es más fácil y por tanto preferible.

		BCG		
		SI	NO	
TBC	SI	10	80	90
	NO	60	40	100
		70	120	190

Cumple condición de aplicación: todas las E ≥5
Z= 48'66 >χ 2 (1 , 0'001)=10'83 y por tanto se rechaza H_0 al nivel de significación de 0'001: Sí hay diferencias. **p < 0,001** .
Sentido: los vacunados enfermas menos. "La vacuna tiene efecto preventivo".

B) Contraste de 2 variables CL con datos independientes y 3 ó + modalidades

Fórmula nº 3

Si todas las E ≥ 5 :

$$Z = \sum \frac{(O - E)^2}{E}$$

Si alguna E < 5 pero ≥ 3:

$$Z = \sum \frac{(|O - E| - 0,5)^2}{E}$$

Si alguna E < 3: no aplicable

Valoración: por $\chi^2_{(f-1)(k-1)}$

***Ejercicio 1.6 Se realiza un experimento de germinación con 3 tipos de semillas en un terreno abonado con la sal S al 5%. De 25 semillas de la especie A germinan 15, de 30 de la B germinan 25 y lo hacen 19 de las 25 de la especie C. ¿Se comportan las especies de forma distinta?

------Problema de comparación de dos Vbles. CL : ESPECIE, con 3 modalidades - A, B y C- y GERMINACION, con 2 modalidades –sí , no. Datos independientes.
A resolver por la fórmula nº 3.
H_0 : no hay diferencias significativas ; germinan de forma similar

		Germinación		
		SI	NO	
Esp.	A	15	10	25
	B	25	5	30
	C	19	6	25
		59	21	80

Se calculan las E y se añaden a la tabla . Cumple la condición de aplicación: todas las E ≥5

Germinación

Esp.		SI	NO
	A	**15** ; *18'43*	**10** *; 6'56*
	B	**25** ; *22'12*	**5** ; *7'87*
	C	**19** ; *18'43*	**6** ; *6'56*

Se aplica la fórmula nº 3 : Z=3'93 < χ^2 (2 ; 0'05)= 5,99
No se puede rechazar H_0. , p > 0,05
"No , el comportamiento es similar"

***Ejercicio 1.7 En 250 personas, elegidas al azar, encontramos las siguientes combinaciones de color de ojos y de pelo : (A=azul, G=gris, N=negro, R=rubio, C=castaño). En 65 A+R, en 20 A+C, en 8 A+N, en 32 G+R, en 40 G+C, en 30 G+N, en 5 N+R, en 10 N+C y en 40 N+N
¿Hay relación entre el color del pelo y el de los ojos?

Es un problema de contraste entre dos variables CL:
- COLOR OJOS con 3 modalidades (A, G y N)
- COLOR PELO con 3 modalidades (R, C y N)

y datos independientes, que se resuelve por la fórmula nº 3
H_0: no hay relación entre el color de los ojos y el color del pelo.
1) construir una tabla de 3x3:

		PELO			
		R	C	N	
OJOS	A	65	20	8	93
	G	32	40	30	102
	N	5	10	40	55
		102	70	78	250

2) calcular los E de cada casilla. (= total de su fila * total de su columna / total general). Vemos que todos son ≥5 y por tanto se cumple la condición de aplicación. Completamos la parte de la tabla que nos interesa, añadiendo al lado de los valores observados, los esperados (E). Los valores esperados son los que se deberían encontrar si no hubiera relación entre las variables, es decir, si H_0 fuera verdadera.

		PELO		
		R	C	N
OJOS	A	**65** ; *37'94*	**20** ; *26'04*	**8** ; *29'02*
	G	**32** ; *41'62*	**40** ; *28'56*	**30** ; *31'82*
	N	**5** ; *22'44*	**10** ; *15'4*	**40** ; *17'16*

3) aplicar la fórmula nº 3 : $\mathbf{Z = \sum \frac{(O-E)^2}{E}}$

Z = 19'30 + 1'40 + 15'23 + 2'22 + 4'58 + 0'10 + 13'55 + 1'89 + 30'40 = **88'67**
4) Z > χ2 (4 ; 0'001)=18'47 y por tanto se rechaza H_0 y se acepta H_1: hay relación entre el color de ojos y pelo al nivel de significación < 0'001. **p < 0,001** . Sentido: (lo vemos comparando las O y las E, nos lo dan los sumandos de Z) : los ojos negros se asocian con el pelo negro y, en menor medida, los ojos azules con el pelo rubio.

C) Contraste de 2 variables cualitativas con datos apareados

Veremos únicamente el caso de que cada variable tenga dos modalidades. Cada individuo proporciona dos datos, forma parte de ambos grupos.
Al igual que en el caso de datos independientes se plantean dos tipos de problemas:
----de comparación: ¿las frecuencias o porcentajes observados son similares en ambas muestras?

H_0: son similares, no hay diferencias significativas, las observadas se deben al azar

----de relación: ¿las variables están relacionadas entre sí?. ¿Hay dependencia entre ellas?

H_0: no hay relación o dependencia

Al ser los datos apareados, comparación y relación son dos cosas distintas, que deben ser resueltas de forma distinta, con fórmulas distintas. Para los problemas de comparación veremos dos fórmulas nuevas: la nº 4 y la nº 5. Para los problemas de relación se usan las ya vistas: nº 1 y nº 2

Pruebas de comparación

Se construye siempre una tabla de 2x2 , de forma un poco distinta a lo visto anteriormente ("se entrelazan" las variables ; los ejemplos mostrarán cómo). Sólo se tienen en cuenta los datos discordantes, aquellos en que no coinciden las variables: a uno se le llama N_1 y al otro N_2 , a la suma de ambos N

fórmula nº 4 : contraste de proporciones (si se utilizan % hay que dividir por 100)

$$Z = (p_1 - 0{,}5)\sqrt{4N}$$

¡esta N no es la N de la tabla!

siendo $N=N_1+N_2$; $N_1 =$ **nº de A+ B-** ; $N_2 =$ **nº de A- B+** ; $p_1 = \frac{N_1}{N}$

Valoración **: si** $N \geq 10$ **por DN ; si <10 pero** ≥ 5 **por** t_{N-1}**; si <5: p Fisher**

fórmula nº 5 : contraste de frecuencias (más sencilla que la anterior)

los símbolos son los mismos de la fórmula nº 4

Si $N \geq 10$**:** $$Z = \frac{(N_1 - N_2)^2}{N}$$

Si $N<10$ **y** ≥ 5**:** $$Z = \frac{(|N_1 - N_2| - 1)^2}{N}$$

si $N<5$ **: p exacta de Fisher**

Valoración **: por** χ_1^2

Ejercicio 2.1 En el diagnóstico de la enfermedad F se utilizan los análisis A y B. Aplicamos ambos análisis a 100 enfermos. Hay un 30% de resultados positivos con A y un 20% con B. Una cuarta parte de los positivos a B fueron negativos a A. En un 65% ambas pruebas fueron negativas. ¿Cual de los dos análisis es mejor?

---Es un problema de comparación entre 2 Vbles. CL con dos modalidades cada una y datos apareados: ANALISIS (A , B) y RESULTADO (+ , -)

Si no se ve claro que es un problema de comparación, hay que preguntarse: ¿que me piden? ¿que averigüe si los análisis diagnostican igual o uno es mejor que otro (comparación) o si los resultados de uno están relacionados con los del otro (relación)?

H_0 : no hay diferencias significativas entre los variables contrastadas. Diagnostican igual

1---construimos la tabla. Nos dan los datos de una forma un tanto enrevesada, pero con un poco de reflexión es fácil hacerlo:

		A		
		+	-	
B	+	15	5	*20*
	-	15	65	*80*
		30	*70*	100

Los datos discordantes son 15 y 5. Por tanto N_1 =15 y N_2 = 5

---2.1.1 resolución por fórmula nº 4

N_1=15 , N_2=5 , N=20 , p_1=15/20 = 0,75 $\mathbf{Z = (0,75 - 0,5) * \sqrt{4 * 20} = 2,24}$

Z > $c_{0,05}$=1,96 , por lo que se rechaza H_0 y se acepta H_1 al nivel de significación de 0,05.
p<0,05 Sentido: el análisis A es positivo con más frecuencia que B.
Contestando a la pregunta: sí, A es mejor.

---2.1.2 resolución por la fórmula nº 5

$\mathbf{Z = \frac{(15-5)^2}{20} = 5}$ > χ^2 (1 , 0'05) = 3,84 → rechazo de H0 y aceptación de H1 a ese nivel de significación. **p < 0,05** . La misma conclusión que antes.

Prueba de relación

Como ya hemos visto en la página 16-5 , estos problemas se resuelven como en el caso de datos independientes por las fórmulas 1 ó 2. Y por tanto se tienen en cuenta todos los valores de la tabla.

Ejercicio 2.1.3 ¿Están relacionados los análisis del ejercicio anterior?

Está claro por la pregunta que se trata de un problema de relación. Entre dos variables CL con dos modalidades cada una y datos apareados.
Veamos la resolución por la fórmula nº 2 :
H_0 :no hay relación significativa ; no hay dependencia
Cumple la condición de aplicación: todas las E ≥ 5

$$\mathbf{Z = \frac{100 * (15 * 65 - 15 * 5)^2}{20 * 80 * 70 * 30} = 24,11}$$

Z > χ^2 (1, 0'001)=10,83 → rechazo de H_0 a ese nivel de significación y aceptación de H_1:hay una relación significativa. p < 0,001
Sentido: la relación es positiva

Si se aplica la fórmula nº 1, se obtiene una Z = 4,91 , que es mayor que la $c_{0,001}$ = 3,30, lo que lleva a las mismas conclusiones.

Ejercicio 3 Se prueban dos avisadores de radar, X e Y, colocados ambos en 33 vehículos, que pasan ante un radar. El X avisó en 23 casos, el Y en 25 y en 5 ocasiones no avisó ninguno. ¿Es el Y de más confianza? ¿Hay dependencia entre ellos?

Nos plantean un problema de comparación y otro de relación.

Problema de comparación (resuelto por fórmula nº 5):
Es un problema de comparación entre 2 Vbles. CL con 2 modalidades cada una y datos apareados: AVISADOR (X - Y) y AVISO (sí – no).
H_0: no hay diferencias significativas entre las frecuencias o porcentajes de las variables contrastadas, ambos aparatos avisan igual, son de igual confianza

		X		
		SI	NO	
Y	SI	20	5	25
	NO	3	5	8
		23	10	33

Sólo interesan los datos discordantes : 5 y 3 : N1=5 , N2=3 , N=8
Como N está entre 5 y 10 se aplica la fórmula nº 5 corregida:
$Z= (|5-3|-1)^2 / 8 = 0'125$, a valorar por $\chi 2\ (1, 0'05)$: $Z < \chi 2$
y por tanto no se puede rechazar la hipótesis nula.
Conclusión: avisan igual .... El Y no es de más confianza

Problema de relación (a resolver por la fórmula 1 ó 2)
En ambos casos se comprueba que no cumplen la condición de aplicación.
Si elegimos la fórmula nº 1 : $N_1=23$, $N_2=10$, $p_1=20/23=0{,}8696$, $p_2=5/10=0'5$, $p_0=0'758$, $q_0=0'242$. al ser muestras pequeñas hay que comprobar la condición de aplicación: N2*q0=2'42 que es <3. Hay que calcular la p exacta de Fisher (pF).
Si elegimos la fórmula nº 2 : Hay una E (la que corresponde a la casilla b2) que vale 10*8/33=2,42 y también obliga a calcular la p exacta de Fisher

p exacta de Fisher (pF)

$$p_F = \sum_{a_1=a_1}^{a_1=0} \frac{N_1!\,N_2!\,N_a!\,N_b!}{a_1!\,b_1!\,a_2!\,b_2!\,N!}$$

nos da la p directamente; no hay que consultar tablas. Para que sea significativa debe ser < 0,05
Esta p es para prueba unilateral, que es la que se utiliza en la práctica. Para prueba bilateral, multiplicar por 2

Los programas estadísticos la calculan fácilmente y de un tirón.
Manualmente, con la ayuda de una calculadora científica se hace siguiendo estos pasos:

1) remodelar la tabla de tal forma que en a_1 quede el valor más bajo.

		X		
		SI	NO	
Y	NO	3	5	8
	SI	20	5	25
		23	10	33

2) quedando fijos Na, Nb, N1, N2 y N , se disminuye a_1 en 1 unidad y se cambian los otros valores del interior de la tabla para que las sumas marginales fijas sean correctas. Se sigue haciendo lo mismo hasta que a_1 sea 0 Así:

2 6	1 7	0 8
21 4	22 3	23 2

3) se aplica la fórmula de la pF para cada una de las tablas y al final se suman todos los resultados parciales obtenidos.

Nota: Como Na, Nb, N1, N2 y N no cambian , recomiendo calcular y dejar en la memoria Na!Nb!N1!N2!/N! .En cada tabla dividiremos este valor almacenado entre el producto $a_1!b_1!a_2!b_2!$ y así obtendremos las p parciales, que sumadas nos dan la pF

En este problema : Na!Nb!N1!N2!/N! = $6'75675^{21}$

p parciales : para a_1 =3 0’032143978
para a_1 =2 0’00382666
para a_1 =1 0’00019879
para a_1 =0 3’2411 $^{-06}$
---
pF = 0’03617267 **p<0,05**

que al ser < 0’05 se rechaza H0 y se acepta H1 : hay relación entre los avisadores, no son independientes. Sentido: bastante coincidencia en el aviso, cuando avisa uno lo suele hacer el otro.

Odds ratio (OR)

Otros nombres: razón de probabilidades, razón de desigualdades

Es el parámetro típico de los estudios caso-control (pero la OR vale para todo tipo de estudios, que queden reflejados en una tabla de 2x2)). Se comparan dos variables CL. Un grupo de individuos que presentan una características determinada (generalmente una enfermedad : casos o “afectados”) se compara con otro grupo de individuos que no la presentan (controles o “no afectados”) para investigar el nivel de exposición a determinados factores que podrían ser causales. A cada caso le corresponden uno o más controles, que deben ser lo más parecidos posible a los casos, excepto en la característica en cuestión.

Se parte de la hipótesis nula: la presencia de la característica no está relacionada con la exposición. El investigador determina el tamaño muestral de los dos grupos, casos y controles, pero ignora como se reparte la exposición entre ellos. La asociación entre exposición y resultado se estima por la razón de probabilidades, más conocida con el nombre de Odds Ratio (OR), que se obtiene dividiendo las probabilidades de casos y controles. Valores posibles: $0 \leq OR \leq \infty$

En vez de la exposición se pueden estudiar los resultados de un análisis en casos y controles para ver su eficacia en el diagnóstico de la enfermedad. O se puede vigilar la aparición de una enfermedad después de haber introducido una vacuna contra la misma, etc.

:

		Enfermedad		
		+	-	
Exposición o resultado	+	**a_1**	**a_2**	Na
	-	**b_1**	**b_2**	Nb
		N_1	N_2	N

Fórmulas:

a) datos independientes : (lo más frecuente)

$$\mathbf{OR} = \frac{a_1}{a_2} : \frac{b_1}{b_2} = \frac{a_1 b_2}{a_2 b_1}$$

b) datos apareados: $\mathbf{OR} = \frac{a_2}{b_1}$ (son los datos discordantes)

Si alguna casilla vale 0 , la OR y su intervalo de confianza pueden ser incalculables.
Solución : sumar 0,5 al valor de cada casilla

Si la OR es >1 , la asociación es positiva, tanto más intensa, cuanto más alta es. La exposición favorece la aparición de la enfermedad. No hay límite superior para el valor que puede alcanzar la OR. El valor de la casilla a_1 es mayor de lo esperado.

Si la OR es < 1 , la asociación es negativa, tanto más cuanto más baja sea (aunque el número siempre es positivo, ya que por la estructura de la fórmula no puede ser < 0). La exposición dificulta la aparición de la enfermedad, protege contra la misma (p. e. una vacuna eficaz). El valor de la casilla a1 es menor de los esperado.

Si la OR es = 1 , no hay asociación; la exposición no influye nada en la aparición de la enfermedad. Es la que corresponde a H_0 .

Para interpretar una OR se toma como referencia el significado de la casilla a_1, que generalmente es la conjunción de enfermedad + y exposición +. Si se cambia el orden de las filas o de las columnas, sale otra OR, ya que hay otra confluencia de modalidades en la casilla a_1. Si los datos son apareados, la casilla de referencia es la a_2 , comparada con la b_1
La hipótesis nula, H_0 , presupone que la OR vale 1 . Pero la OR sola es un valor puntual y no sirve para la valoración estadística; hay que calcular el intervalo de confianza, que veremos enseguida. **Si el intervalo no incluye el 1, se rechaza la hipótesis nula** y se concluye que hay una asociación significativa al nivel de significación que hayamos elegido para c ó t y en el sentido que indique la casilla de referencia. Si el intervalo incluye el 1, no puede rechazarse H_0

Cálculo del intervalo de confianza de una OR

El método más sencillo utiliza logaritmos. Se halla el logaritmo neperiano de la OR y a éste se le suma y resta el error muestral E, que tiene una fórmula fácil (habitualmente se toma un nivel de significación alfa para c ó t de 0,05).
Así tendremos los límites del intervalo, cuyos antilogaritmos son los límites del IC de la OR

a) **DATOS INDEPENDIENTES**

$$\textbf{IC del ln OR} = \ln OR \pm c\sqrt{\frac{1}{a_1}+\frac{1}{a_2}+\frac{1}{b_1}+\frac{1}{b_2}} \quad ; \quad \textbf{si N<30, en vez de c se toma } t_{N-2}$$

hallado el intervalo se calculan los antilogaritmos (e^x) de los límites del intervalo: son los límites del IC de OR

Ejemplo: Se estudia en una comarca la mortalidad precoz (antes de los 60 años) en fumadores y no fumadores.

		Fumador		
		Si	No	
Muerte precoz	Si	700	200	900
	No	300	300	600
		1000	500	1500

La **OR vale 3,5** ; la probabilidad de muerte precoz de un fumador es 3,5 veces mayor que la un no fumador.
El ln de la OR es 1,252762968 (hay que seguir trabajando al menos con 6 decimales)

IC del ln OR = 1,252763 ± 1,96 √ 1/700 + 1/200 + 1/300 + 1/300 =
1.252763 ± 0,224291 = ∈ (1,028472÷1,477054)
Calculando los antilogaritmos de ambos límites (y redondeando a dos decimales) :
IC de OR = € (2,80 ÷ 4,38) , que es significativo al no estar el 1 en el intervalo. nivel de significación 0,05 (hemos tomado para c el valor de 1,96). Asociación positiva entre fumar y muerte precoz.

b) DATOS APAREADOS

La fórmula es la misma, excepto lo que va dentro de la raíz cuadrada: $\sqrt{\frac{1}{a_2}+\frac{1}{b_1}}$

Sólo se tienen en cuenta los datos discordantes. Al ser datos apareados N = a2 + b1.
La OR va referida a la casilla a_2 (comparada con la b_1)

Ejemplo: Se comparan en 62 pacientes la eficiencia de dos análisis distintos (A y B) en el diagnóstico de una enfermedad.

		A		
		+	-	
B	+	**20**	**12**	32
	-	**15**	**15**	30
		35	27	62

OR = 0,8 ; ln 0,8 = -0,223144 ; **N** = 12 + 15 = 27 (los discordantes!)
IC lnOR = -0,223144 ± 2,060 √ 1/12 + 1/15 = -0,223144 ± 0,797835 = Є (-1,020979 ÷ 0,574691) . Sus antilogaritmos son los límites de OR (redondeamos a dos decimales) : **IC de OR = Є (0,36 ÷ 1,78)**
La OR no es significativa al incluir al 1 en su intervalo. Es n.s. p>0,05 . Ambos análisis son igual de eficientes, aunque B parezca algo inferior, ya que la OR de 0,8 indica según las casillas a_2 y b1 que es inferior en acertar cuando el otra análisis falla.

Riesgo relativo (RR)

Es el parámetro típico de los estudios de cohortes, que son estudios prospectivos en los que se siguen durante años a personas expuestas y no expuestas a un determinado riesgo o condición para ver si enferman o no. Por ejemplo, el seguimiento durante años de personas que toman un determinado medicamento para prevenir enfermedades graves y de un grupo control que no lo toma. En vez de medicamentos el objeto de estudio puede ser el ejercicio físico u otros hábitos saludables, psicoterapia, etc. Aunque se habla de riesgo, a veces se trata de un beneficio. Problemas del lenguaje.
Matemáticamente es siempre posible calcular el RR, con independencia de que sea un estudio caso-control o de cohortes. Pero cada uno tiene su parámetro adecuado. Si el riesgo es escaso (< 0,1 ó 10%) OR y RR toman valores muy parecidos, pero a medida que el evento se hace más frecuente empiezan a separarse cada vez más. En muchos estudios se usa la OR como equivalente del RR, lo que no es correcto.

El RR es el cociente de los riesgos de expuestos y no expuestos.. Se expresa como proporción o porcentaje.

Se parte de la tabla de 2x2 :

		Enfermedad o evento negativo		
		+	-	
Exposición o factor a estudio	+	$\mathbf{a_1}$	$\mathbf{a_2}$	Na
	-	$\mathbf{b_1}$	$\mathbf{b_2}$	Nb
		N_1	N_2	N

$$\mathbf{RR} = \frac{\mathbf{a_1}}{\mathbf{N_a}} : \frac{\mathbf{b_1}}{\mathbf{N_b}} = \frac{\mathbf{a_1 N_b}}{\mathbf{b_1 N_a}}$$

La hipótesis nula H_0 es que RR = 1. **La valoración es similar a la de la OR**. Para ver si la asociación es significativa, es preciso calcular el intervalo de confianza de RR.

El RR es significativo si su IC no incluye al 1

Cálculo del intervalo de confianza de RR

--Se calcula el IC del logaritmo neperiano del RR y luego se vuelve a "números normales"... Así:

$$\mathbf{IC\ del\ ln\ de\ RR = ln\ R \pm c}\sqrt{\frac{1}{\mathbf{a_1}} + \frac{1}{\mathbf{b_1}} - \frac{1}{\mathbf{N_a}} - \frac{1}{\mathbf{N_b}}}$$

¡ojo a los dos signos menos!
si N es menor de 30, en vez de **c** se toma **t** con gl N-2

--luego se calculan los antilogaritmos (e^x) de los extremos del intervalo :
son los límites del IC del RR **IC = ∈ (límite inferior ÷ límite superior)**

Ejemplo:
En un hospital inglés se aplicó un programa destinado a incrementar la duración de la lactancia materna. A los 3 meses ya no daban el pecho 32 de las 51 mujeres del grupo de intervención y 52 de las 57 del grupo control. Concluyen que con el programa han reducido claramente el riesgo de abandono de la lactancia materna a los 3 meses.
Veamos:

Programa fomento Lactancia Materna (LM)

		Abandono		
		+	-	
Programa	+	32	19	51
	-	52	5	57
		84	24	108

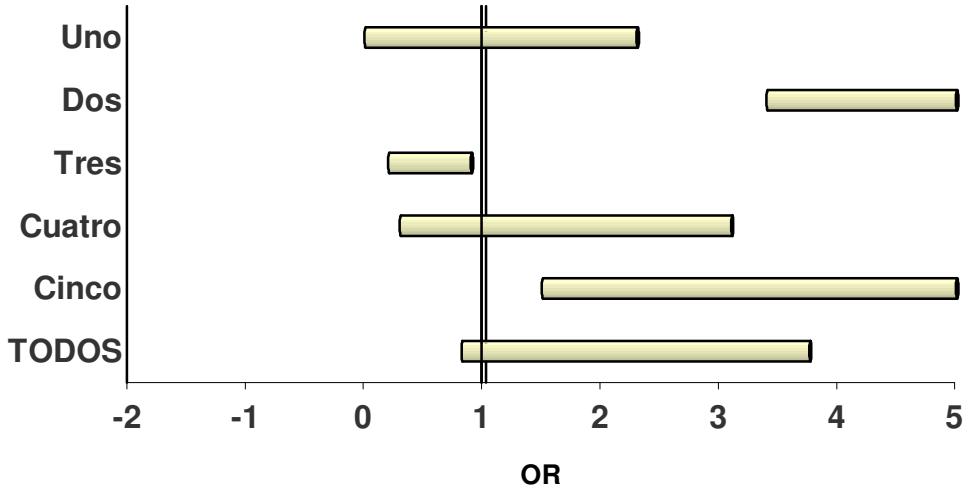
RR de abandono de la LM en las que han seguido el programa:

RR = (32 * 57)/(52 * 51) = **0,688**

Al ser la RR < 1 indica que el riesgo es menor para la situación que indica la casilla a1, es decir, abandono habiendo seguido el programa. Pero este riesgo menor ¿es significativo? . Para contestar a esta pregunta hay que calcular el IC de RR, que aplicando la fórmula resulta ser
$\in (0{,}55 \div 0{,}86)$, que al no incluir el 1 es significativo al nivel de significación empleado, que es 0,05 (ya que se ha tomado c = 1,96)

Metaanálisis

Con frecuencia se observa que estudios sobre un mismo tema dan resultados divergentes, incluso con grandes diferencias. En estos casos es de ayuda la técnica llamada Metaanálisis, que permite calcular un IC conjunto para todos los estudios y de él sacar la conclusión adecuada. Es un procedimiento muy complejo y laborioso, en el que no entramos (está muy bien descrito en el libro de Armitage/Berry). Como orientación se pueden hacer dos cosas: 1) pasar a un gráfico los IC de las diversas OR, lo que nos da una idea del conjunto 2) a partir de una tabla que englobe el total de los datos de todos los estudios, calcular la OR y su IC por el procedimiento ordinario en vez de por el complicado método ortodoxo.
El siguiente gráfico representa gráficamente un metaanálisis:



OR . Aclaraciones sobre la tabla y repaso de la valoración

Con los datos que se dan en el enunciado de un problema, la tabla se puede construir de 4 formas distintas, que nos dan dos OR diferentes, pero relacionadas. Cada OR es la inversa de la otra (1/OR). En los límites de confianza el inverso del Li de una OR es el Ls de la otra y viceversa.

Ejemplo

Pinto y col. han estudiado en una zona de México la relación entre malformaciones congénitas y consanguinidad en 33194 recién nacidos en un periodo de 6 años. Hubo 1117 neonatos con alguna anomalía congénita. Se tomó como control de cada caso al primer neonato sano del mismo sexo nacido después. 21 de los malformados tenían el antecedente de consanguinidad por 8 de los controles. Valore el resultado (por OR).

Se trata de un contraste de dos variables cualitativas con dos modalidades cada una : Malformación (Sí , No) y Consanguinidad (Sí , No) . Los datos son independientes. La hipótesis nula H_0 es que no hay diferencias significativas en las malformaciones que aparecen en niños con y sin antecedente de consanguinidad,o sea una OR =1 . Este problema se puede resolver por la fórmula nº 2 ó 1, pero se pide que se haga valorando la OR.

Pasos:

1---construir la tabla de 2x2 ; ocurre que podemos construir 4 tablas distintas. Calcularemos en cada una la OR y su IC (se ha tomado una c = 1,96 que corresponde a $\alpha = 0,05$))

1	Malformaciones		
		Sí	No	
Consanguinidad	Sí	21	8	29
	No	1096	1109	2205
		1117	1117	2234

OR = 2,66 (2,65613...)

$\in$ (1,171÷6,022) (1,171487... y 6,022306...)

2	Malformaciones		
		No	Sí	
Consanguinidad	Sí	8	21	29
	No	1109	1096	2205
		1117	1117	2234

OR = 0,38 (0,376486...)

$\in$ (0,166÷0,854) (0,166049... y 0,853615...)

3	Malformaciones		
		Sí	No	
Consanguinidad	No	1096	1109	2205
	Sí	21	8	29
		1117	1117	2234

OR = 0,38

$\in$ (0,166÷0,854)

4	Malformaciones		
		No	Sí	
Consanguinidad	No	1109	1096	2205
	Sí	8	21	29
		1117	1117	2234

OR = 2,66

$\in$ (1,171÷6,022)

Se obtiene pues dos OR distintas.

El nº inverso de la primera OR es 1/2,65613 ≈ 0,38 (la otra OR) y el inverso de la segunda OR es 1/0,3746486 ≈ 2,66

Y para el intervalo de confianza : 1/1,171487 ≈ 0,854 y 1/6,022306 ≈ 0,166
1/0,166049 ≈ 6,022 y 1/0,853615 ≈ 1,171

La valoración de la OR se hace por la casilla a_1 .

Recuerden la nomenclatura de las casillas:

	a_1	a_2	Na
	b_1	b_2	N_b
	N_1	N_2	N

En la tabla 1 la casilla a_1 es la confluencia de malformación y consanguinidad; como la OR (2,66) es >1 , interpretamos que cuando hay consanguinidad, se observan más malformaciones de lo esperado. Esta asociación es estadísticamente significativa al no estar el uno en el intervalo de confianza ($p<0,05$). La tabla 4 es lo mismo, pero visto desde el lado opuesto. Se asocian no malformación y no consanguinidad.

En la tabla 2 la casilla a_1 corresponde a consanguíneos no malformados; su OR = 0,38 , que es <1 , es decir que los niños consanguíneos sin malformación son menos de los esperados y además de forma significativa ($p<0,05$) al no incluir el 1 su intervalo de confianza. En la tabla 3 confluyen malformación y no consaguinidad, con valoración similar.

¿Cuál elegir?

La que mejor se corresponda al objetivo del problema, que en este caso es valorar una posible asociación entre consanguinidad y malformaciones congénitas. Por tanto la mejor tabla es la nº 1, que lo hace de forma directa, seguida de la 2 . Pero todas son buenas y nos llevarán a la misma conclusión, aunque por caminos más retorcidos y menos intuitivos.

Un razonamiento similar se puede hacer para el RR

Puntos débiles de las OR

La OR es otra forma de enfocar el contraste de frecuencias de dos variables cualitativas con dos modalidades cada una. La decisión estadística es la misma.

Es un parámetro que se puso de moda en el pasado decenio. Es muy útil, pero tiene también sus puntos débiles, **los mismos que el procedimiento clásico**. Recordémoslos:

--las muestras de casos y controles con frecuencia no son aleatorias. Siempre hay que preguntarse si todos los individuos de las poblaciones de casos y controles han tenido la misma probabilidad de salir elegidos para el estudio.

--Los criterios de exclusión del estudio a veces no son los mismos para casos y controles.

--Hay que vigilar los sesgos de recuerdo (“recall bias”) en la documentación clínica, pues los pacientes son reiteradamente preguntados sobre los factores de riesgo, cosa que no les ha sucedido a los controles.

--Hay que buscar la posible existencia de factores de confusión, que pueden simular asociación significativa entre exposición y enfermedad. Por ejemplo, un estudio puede sugerir que los alcohólicos tienen un riesgo elevado de padecer cáncer de pulmón, hasta que se descubre que prácticamente todos los alcohólicos eran fumadores. Otro ejemplo: en muchas ocasiones se prescriben estrógenos para las hemorragias vaginales. Si meses después se descubre un cáncer de útero, podría pensarse que es un efecto secundario de los estrógenos. Pero no hay que olvidar que las hemorragias son un síntoma de cáncer uterino.

Si se identifican “confundidores” hay que estratificar en subgrupos del confundidor. Los más frecuentes son: edad, sexo, nivel sociocultural, tabaco, alcohol, drogas....
--No se debe olvidar que una relación o asociación significativa sólo permite concluir causalidad si el estudio es experimental.

En los ejercicios que hemos realizado por los contrastes clásicos, se puede también calcular la OR, aunque no sea el parámetro más adecuado. Pero se llega a las mismas conclusiones:

Ejercicio	Variables	Datos	a1	a2	b1	b2	OR	IC- OR	¿rechazo de H_0?
1.1 y 1.2	Fumar (sí , no) Sexo (♂ , ♀)	Independientes	70	30	80	120	3’50	2’10 5’84	SI Hombres fuman más
1.3	Ciudad (X , Y) Colesterol (alto, bajo)	Independientes	6	14	15	15	0’43	0’13 1’42	NO
1.4	Miopía (si-no) Sexo (♂ - ♀)	Independientes	4	16	7	21	0,75	0’19 3’01	NO
1.5	BCG (si-no) TBC (sí , no)	Independientes	10	80	60	40	0’08	0’04 0’19	SI Si BCG, menos TBC
2.1	Análisis (A-B) Result. (+ -)	Apareados	15	5	15	65	0’33	0’11 0’99	SI A es mejor
3	Radar (X , Y) Aviso (sí , no)	Apareados	20	5	3	5	1’67	0’28 9’95	NO

La OR se debe reservar para los estudios caso-control, aunque siempre es calculable.