El test de contingencia (χ^2 de contingencia)

Dadas las distribuciones de dos procesos A y B considerados individualmente, el test de contingencia trata de analizar su distribución conjunta. Entre otras cosas, sirve para establecer hasta qué punto dos fenómenos o procesos (cada uno de ellos con diferentes alternativas) ocurren o no de forma independiente. El test de contingencia tiene gran utilidad en muchas ramas de la Ciencia. Uno de los usos más frecuentes en Genética es en la determinación de la existencia de ligamiento entre dos genes (véanse ejemplos de aplicación del test de contingencia).

Imaginemos los dos procesos A y B (podrían ser la segregación del gen A,a y del gen B,b en una descendencia), cada uno de ellos con varias alternativas (A_1, A_2, ..., A_n y B_1, B_2 ..., B_m). Si se realizan observaciones de esos dos procesos conjuntamente, tendremos nxm clases con sus correspondientes frecuencias observadas (O_{11}, O_{12} ..., O_{nm}):

Aparición conjunta de las alternativas:	frecuencia
A_1 y B_1	O_{11}
A_1 y B_2	O_{12}
...	...
A_i y B_j	O_{ij}
...	...
A_n y B_m	O_{nm}

Estos datos pueden presentarse en forma de tabla de contingencia: →

Tabla de contingencia

Alternativas del proceso B ↓ / Alternativas del proceso A → A_1	A_2	...	A_n	Distribución del proceso B ↓
B_1	O_{11} $E_{11}=\frac{TA_1 \times TB_1}{T}$	O_{21} $E_{21}=\frac{TA_2 \times TB_1}{T}$	...	O_{n1} $E_{n1}=\frac{TA_n \times TB_1}{T}$	$TB_1=\sum_{i=1}^{n} O_{i1}$
B_2	O_{12} $E_{12}=\frac{TA_1 \times TB_2}{T}$	O_{22} $E_{22}=\frac{TA_2 \times TB_2}{T}$	...	O_{n2} $E_{n2}=\frac{TA_n \times TB_2}{T}$	$TB_2=\sum_{i=1}^{n} O_{i2}$
...	...	...	...	...	...
B_m	O_{1m} $E_{1m}=\frac{TA_1 \times TB_m}{T}$	O_{2m} $E_{2m}=\frac{TA_2 \times TB_m}{T}$	...	O_{nm} $E_{nm}=\frac{TA_n \times TB_m}{T}$	$TB_m=\sum_{i=1}^{n} O_{im}$
Distribución del proceso A →	$TA_1=\sum_{i=1}^{m} O_{1i}$	$TA_2=\sum_{i=1}^{m} O_{2i}$	...	$TA_n=\sum_{i=1}^{m} O_{ni}$	$T=\sum_{i=1}^{n} TA_i=\sum_{i=1}^{m} TB_i$

En la columna de la derecha de esta tabla, como resultado de la suma de las correspondientes filas, se indican las frecuencias de la distribución del proceso B, es decir, las frecuencias totales de las distintas alternativas B_1, B_2, ... B_m (sin tener en cuenta el proceso A). Al margen del test de contingencia, si se dispone de la hipótesis apropiada, podría compararse con la correspondiente distribución esperada mediante un test χ^2 normal (gl= m-1).

En la casilla inferior derecha figura el número total de observaciones (T).

En la fila inferior de esta tabla, como resultado de la suma de las correspondientes columnas, se indican las frecuencias de la distribución del proceso A, es decir, las frecuencias totales de las distintas alternativas A_1, A_2, ... A_n (sin tener en cuenta el proceso B). Al margen del test de contingencia, si se dispone de la hipótesis apropiada, podría compararse con la correspondiente distribución esperada mediante un test χ^2 normal (gl= n-1).

Las probabilidades ("a posteriori") de las distintas alternativas de cada proceso son: probabilidad de la alternativa $A_i = \frac{TA_i}{T}$ probabilidad de la alternativa $B_j = \frac{TB_j}{T}$

Por tanto, en el supuesto de que los dos procesos ocurran de forma independiente (hipótesis nula), el número esperado de casos (E_{ij}) en los que deberían ocurrir simultáneamente las alternativas A_i y B_j es: $E_{ij} = \frac{TA_i}{T} \times \frac{TB_j}{T} \times T = \frac{TA_i \times TB_j}{T}$

Estos valores esperados (E_{ij}) aparecen indicados en color rojo en la tabla de contingencia.

Los valores observados (O_{11}, O_{12}, ..., O_{nm}) y los correspondientes esperados (E_{11}, E_{12}, ..., E_{nm}) se comparan mediante un test χ^2 (χ^2 de contingencia), en el que el número de grados de libertad es: gl= (m-1) x (n-1), es decir, (número de filas -1) x (número de columnas - 1).

Grados de libertad

En temas relacionados con la comparación entre distribuciones observadas y esperadas, el término grados de libertad puede entenderse como la libertad de que se dispone para construir una distribución esperada, dentro de determinadas restricciones.

Por ejemplo, si tenemos una F2 con una distribución de fenotipos compuesta por 580 individuos de fenotipo A y 186 de fenotipo a, tenemos un total de 766 individuos. Es obvio que cualquier distribución esperada con la que queramos comparar esta distribución observada deberá tener la restricción de tener el mismo número total de individuos: 766. Como la distribución no tiene más que dos clases que deben sumar 766, si asignamos libremente un valor a una de estas clases, el valor de la otra clase estará tambien determinado. Es decir, cualquier hipótesis en la que basemos la construcción de la distribución esperada tendrá un grado de libertad. Si la hipótesis es que la clase con fenotipo A está constituida por 3/4 del total, la otra clase esperada tiene que estar constituida por el resto de los individuos: 1/4 del total. Supongamos ahora una F2 compuesta de 134 individuos de genotipo AA, 268 Aa y 125 aa. La distribución esperada deberá tener un total de 527 individuos, pero al tener tres clases, la hipótesis tiene dos grados de libertad ya que puede asignar libremente valores a dos de las clases antes de que todo quede determinado. En resumen, en distribuciones de este tipo en las que se toma un sólo parámetro de la muestra (el número total de observaciones) para construir la distribución esperada, el número de grados de libertad es: gl= número de clases - 1.

En el caso de un test de contingencia, la situación es diferente ya que las restricciones son mayores. En definitiva, dadas las distribuciones de dos procesos A y B considerados individualmente, el test de contingencia trata de analizar su distribución conjunta. Es decir, la hipótesis en la que se basa la distribución esperada no se refiere a las distribuciones individuales de los procesos A y B, sino a la relación existente entre ellas. Las dos distribuciones individuales deben cumplirse al construir la distribución esperada.

Gen A,a ↓ / Gen B,b → Fenotipo B	Fenotipo b	Segregación del gen A,a ↓
Fenotipo A	406 $\frac{580 \times 541}{766} = 409.6$	174 $\frac{580 \times 225}{766} = 170.4$	580
Fenotipo a	135 $\frac{186 \times 541}{766} = 131.4$	51 $\frac{186 \times 225}{766} = 54.6$	186
Segregación del gen B.b →	541	225	Total= 766

Por ejemplo, supongamos una F2 en la que están segregando los dos genes A,a y B,b, ambos con dominancia completa, de la siguiente manera: 406 AB, 174 Ab, 135 aB y 51 ab (véase tabla de contingencia a la derecha). La distribución individual para el gen A,a es: 580 A, 186 a. La distribución para el gen B,b es: 541 B, 225 b. Estas dos distribuciones se tienen que cumplir en la distribución conjunta esperada (en color rojo), por lo que sólo se dispone de un grado de libertad para construir tal distribución: al asignar un valor a una de las clases, el resto quedarán determinadas. En este ejemplo, para construir la distribución esperada es necesario tomar de la muestra (valores observados) un mínimo de tres parámetros: el número total de observaciones, una de las dos clases de la segregación del gen A,a (la otra clase viede dada restando del número total de observaciones), y una de las dos clases de la segregación del gen B,b (la otra clase viede dada restando del número total de observaciones). Por tanto, se cumple que el número de grados de libertad de la distribución conjunta (gl= 1) es igual al número de clases (4) menos el número de parámetros tomados de la muestra para construir la distribución esperada (3). Por la misma razón, a partir de una tabla de contingencia de m filas y n columnas como la que figura en la parte superior de esta página (mxn clases), sería necesario tomar de la muestra 1 + (m-1) + (n-1) parámetros: el número total de observaciones, m-1 clases de la distribución del suceso B y n-1 clases de la distribución del suceso A. El número de grados de libertad en ese caso sería: gl= (mxn) - (1 + (m-1) + (n-1))= (m-1)x(n-1); es decir, el número de filas menos 1 por el número de columnas menos uno.