Contrastes χ^2

Estudiaremos tres tipos de contrastes, cada uno de los cuales trata de responder a la pregunta correspondiente:

- Bondad del ajuste
  - ¿Proceden los datos de una determinada distribución?
- Homogeneidad
  - ¿Proceden las k colecciones de datos de variables con la misma distribución?
- Independencia
  - En una muestra de n elementos se observan dos variables (discretas) ¿Son estas variables independientes?

Contrastes de bondad del ajuste

Se conjetura que una m.a.s. procede de una variable X, con distribución discreta, que toma valores $a_1, a_2, \ldots, a_k$, con probabilidades respectivas $p_1, p_2, \ldots, p_k$. La muestra tiene un total de n elementos.

El número esperado de observaciones del valor a_i en la muestra es de $E_i = p_i \cdot n$, $i=1, 2, \ldots, k$.

En la muestra se han obtenido un total de O_i observaciones del valor a_i, $i=1, 2, \ldots, k$.

¿Es este resultado compatible con las probabilidades p_i conjeturadas?

El contraste a efectuar tiene como hipótesis nula que la distribución tiene las probabilidades conjeturadas.

$H_0 \equiv P(X = a_i) = p_i \quad ; \quad i=1, 2, \ldots, k.$

¿Es el resultado obtenido en la muestra probabilísticamente compatible con H_0?

Para construir el contraste se utiliza la propiedad siguiente:

El estadístico

$$X^2 = \Sigma_i \, (O_i - E_i)^2 / E_i$$

Tiene aproximadamente (cuando n es "grande") una distribución χ^2_{k-1}

[En realidad, bajo la hipótesis H_0, X^2 converge a una χ^2_{k-1} cuando $n \to \infty$]

Cuando H_0 es cierta, el valor de X^2 será pequeño. La región de rechazo de H_0 es de la forma

$$\mathcal{R} = \{ X^2 > \chi^2_{k-1;\,\alpha} \},$$

es decir, para X^2 lo suficientemente grande.

Ejemplo

Se quiere determinar si un dado está o no equilibrado, para ello se realizan 60 lanzamientos. Se anota el número de resultados obtenidos

1	2	3	4	5	6
14	7	10	15	9	5

Estos son los valores denotados por O_i

La hipótesis nula H_0 es $p_i = 1/6$, por tanto los valores esperados son todos iguales: $E_i = 60 \cdot 1/6 = 10$.

El valor del estadístico de contraste será:

$X^2 = ((14-10)^2 + (10-10)^2 + (7-10)^2 + (15-10)^2 + (9-10)^2 + (5-10)^2)/10 = 6'76$

Dado que k=6, si queremos contrastar la hipótesis nula al nivel 0'05, tenemos que comparar X^2 con

$$\chi^2_{5;\,0'05} = 11'07$$

No se rechaza H_0.

En las tablas también podemos ver que

$$\chi^2_{5;\,0'25} = 6'63$$

por tanto, el *p*–valor de $X^2 = 6'76$ es aproximadamente 0'25.

Observaciones

1. El cálculo del estadístico X^2 también se puede realizar mediante la fórmula:

$$X^2 = \sum_i (O_i^2 / E_i) - n$$

Ya que

$$\sum_i O_i = \sum_i E_i = n$$

Esta fórmula permite un cálculo más rápido de X^2 pero con ella se pierde la información sobre qué término de los k sumados contribuye más al posible valor alto de X^2.

2. El número de grados de libertad debe reducirse en el número de parámetros que haya que estimar para determinar exactamente la distribución.

Aplicación a variables continuas

Un contraste X^2 del tipo descrito puede aplicarse a una distribución continua por medio de una «discretización», es decir dividiendo el rango de la variable en un número finito de intervalos

$$I_1, I_2, \ldots, I_k$$

Y contando las incidencias observadas y esperadas en cada intervalo.

Las incidencias esperadas se calculan por medio de las probabilidades

$$p_i = \mathrm{P}(X \in I_i) \;:\; E_i = n \cdot p_i$$

Es conveniente que al utilizar este procedimiento se elijan al menos 5 clases I (es decir, $k \geq 5$) y que el número esperado de ocurrencias en cada clase sea de al menos 3.

Ejemplo

Contrastemos si los datos del primer ejercicio de la lista (Cavendish) proceden de una distribución normal.

Estimamos su media y su varianza por *Ave* = 5'45; s^2=0'0484 (s = 0'22). Dado que n=29 tomaremos 5 intervalos equiprobables de la Normal, así el número de casos de la muestra esperado en cada uno será de 29/6=4'8.

Cinco intervalos equiprobables de una N(0,1) tienen extremos ±0'25, ±0'84, por tanto, para nuestros datos los extremos serán 5'45±(0'22)·(0'25); 5'45±(0'22)·(0'84) y por tanto

	X<5'265	<X<5'395	<X<5'505	<X<5'635	<X
O	4	8	5	7	5
E	5'8	5'8	5'8	5'8	5'8

Se tiene

$$\mathcal{R} = \{ X^2 > \chi^2_{2;\, 0'05} \}$$

$$\chi^2_{2;\, 0'05} = 5'99 \text{ y } X^2 = 1'86.$$

No se rechaza la normalidad de la variable.

Importante: El número de grados de libertad se ha reducido en 2 (5−1−2), ya que ha habido que estimar 2 parámetros (media y varianza) con los datos.

Contrastes de homogeneidad

Se tienen dos o más m.a.s. Se desea saber si proceden o no de variables aleatorias con la misma distribución. La distribución toma valores en k clases $A_1, A_2, \ldots, A_k$. Las probabilidades serían $p_1, p_2, \ldots, p_k$, en principio desconocidas.

$X_{11}, X_{12}, \ldots, X_{1n_1}$
$X_{21}, X_{22}, \ldots, X_{2n_2}$
$\ldots, \quad \ldots, \quad \ldots, \ldots$
$X_{p1}, X_{p2}, \ldots, X_{pn_p}$

Hipótesis nula: las p muestras proceden de variables con la misma distribución.

Valores observados de cada muestra (índice i) en cada clase (índice j) se denotan: O_{ij}.
Los valores p_j se estiman mediante todas las muestras, bajo la hipótesis nula: $O_j = \sum_i O_{ij}$; $j = 1, 2, \ldots, k$. Por tanto, $p_j = O_j / n$; $n = n_1 + n_2 + \ldots + n_p$.

Los valores esperados E_{ij} se estiman entonces por medio de:

$$E_{ij} = n_i \cdot p_j$$

El contraste se efectúa comparando los valores Oij con los Eij. Si son «muy» diferentes se rechaza la hipótesis nula. El estadístico a utilizar es

$$X^2 = \Sigma_{ij} \, (O_{ij} - E_{ij})^2 / E_{ij}$$

cuya distribución aproximada es una χ^2 con $p(k-1) - (k-1)$ grados de libertad, es decir $(p-1)(k-1)$.

Por tanto

$$\mathcal{R} = \{ X^2 > \chi^2_{(k-1)(p-1);\,\alpha} \}.$$

Se experimentan dos métodos I y II de cultivo sobre 100 parcelas cada uno. Se obtienen los resultados:

O_{ij}	Muy bueno (MB)	Bueno (B)	Intermedio (I)	Regular (R)	Malo (M)
I	15	25	32	17	11
II	9	18	29	28	16
Total	24	43	61	45	27
p_j	0’120	0’215	0’305	0’225	0’135

A partir de estas pi se obtienen

E_{ij}	Muy bueno (MB)	Bueno (B)	Intermedio (I)	Regular (R)	Malo (M)
I	12	21’5	30’5	22’5	13’5
II	12	21’5	30’5	22’5	13’5

Así: $X^2 = 6'4$; $\chi^2_{4;\,0'05} = 9'5$ Por tanto, no se rechaza la hipótesis nula.

Ejemplo

Una asignatura se imparte en tres grupos distintos: A, B y C. En el examen final, común a los tres grupos, se ha obtenido la siguiente distribución de calificaciones:

	Suspenso	Aprobado	Notable	Sobresaliente	Total
A	30	45	15	5	95
B	24	31	18	7	80
C	22	33	16	6	77
Total	76	109	49	18	252

Se quiere contrastar, al nivel alfa = 0’05, la hipótesis nula $H_0 \equiv$ la distribución de notas es análoga en los tres grupos. Para calcular los valores esperados bajo la hipótesis nula calculamos primero las proporciones globales en cada calificación.

	Suspenso	Aprobado	Notable	Sobresaliente	Total
p	0’302	0’433	0’194	0’071	1
A	28’7	41’1	18’4	6’7	94’9
B	24’2	34’6	15’5	5’7	80’0
C	23’3	33’3	14’9	5’5	77’0
Total	76’2	109’0	48’8	17’9	251’9

Se obtiene: $X^2 = 2'9$; $\chi^2_{6;\,0'05} = 12'6$. No se rechaza H_0 (p-valor = 0’8).

Contrastes de independencia

Sobre una determinada población se analizan dos variables X e Y. Se miden las dos variables sobre n individuos de la población, elegidos al azar. Se obtiene la m. a. s. de vectores aleatorios (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n). Se quiere estudiar si las variables X e Y son independientes. Los valores de las variables se clasifican en k y l clases $A_1, A_2, \ldots, A_k$; $B_1, B_2, \ldots, B_l$; respectivamente. Se cuentan los casos O_{ij} ocurridos en el producto de clases $A_i \times B_j$. Para cada resultado posible $A_i \times B_j$ del vector aleatorio tendremos una probabilidad p_{ij} y por lo tanto, para una muestra de longitud n, tendremos un número esperado de casos E_{ij} igual a $n \cdot p_{ij}$. Bajo la hipótesis nula de independencia de las variables, podemos estimar p_{ij} como producto $p_i \cdot p_j$, donde p_i y q_j se estiman por medio de los resultados de la muestra.

Bajo la hipótesis nula, el estadístico

$$X^2 = \Sigma_i \Sigma_j \, (O_{ij} - E_{ij})^2 / E_{ij}$$

tiene aproximadamente una distribución

$$\chi^2_{(k-1)(l-1)}$$

Por tanto, la región de rechazo al nivel de significación alfa será

$$\mathcal{R} = \{ X^2 > \chi^2_{(k-1)(p-1);\,\alpha} \}.$$

Ejemplo

(citado por Moore & McCabe, Intro to the Practice of Statistics)

Un estudio sobre grupos sanguíneos y grupos raciales en Hawaii ha dado los siguientes resultados:

	I	II	III	IV	Total
0	1903	4469	2206	53759	62337
A	2490	4671	2368	50008	59537
B	178	606	568	16252	17604
AB	99	236	243	5001	5579
Total	4670	9982	5385	125020	145057

¿Son independientes las variables «Grupo sanguíneo» y «Grupo racial»?

Las proporciones estimadas de Grupos sanguíneos y de Grupos raciales son

0	A	B	AB
0,430	0,410	0,121	0,038

I	II	III	IV
0,032	0,069	0,037	0,862

Bajo la hipótesis nula de independencia, la probabilidad de cada combinación de valores será

	I	II	III	IV
0	0,014	0,030	0,016	0,370
A	0,013	0,028	0,015	0,354
B	0,004	0,008	0,005	0,105
AB	0,001	0,003	0,001	0,033

Los valores esperados son entonces:

E_{ij}	I	II	III	IV
0	2.007	4.290	2.314	53.726
A	1.917	4.097	2.210	51.313
B	567	1.211	654	15.172
AB	180	384	207	4.808

El valor del estadístico de contraste es X^2 =1078. El valor crítico al nivel de significación α= 0'05 es $\chi^2_{9;0'05}$ = 16'9. Conclusión: las variables no son independientes.