

Prueba de Independencia

- También es conocida como tablas de contingencia (Dependencia)
- Es una de las aplicaciones más usadas del test de Bondad de Ajuste, que consiste en contrastar si dos maneras de clasificar n objetos son independientes o no.
- Se tienen dos variables cualitativas: X e Y
- Sean x_1, x_2, \dots, x_m los distintos atributos o niveles del criterio X y y_1, y_2, \dots, y_p los distintos atributos o niveles para el criterio Y .

Prueba de independencia

Se toma una muestra de tamaño n y se registran las frecuencias observadas que llamaremos o_{ij} , donde o_{ij} representa los individuos que tienen los atributos (niveles) x_i e y_j , con $i = 1, 2, \dots, m$ y $j = 1, 2, \dots, p$

$X \backslash Y$	y_1	y_2	\dots	y_p	TOTALES
x_1	o_{11}	o_{12}	\dots	o_{1p}	Tx_1
x_2	o_{21}	o_{22}	\dots	o_{2p}	Tx_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_m	o_{m1}	o_{m2}	\dots	o_{mp}	Tx_m
TOTALES	Ty_1	Ty_2	\dots	Ty_p	n

Tx_i : total de individuos observados que poseen el atributo x_i

Ty_j : total de individuos observados que poseen el atributo y_j

Prueba de independencia

Recordemos que al ser, la prueba de independencia una prueba de Bondad de Ajuste, se requiere contrastar lo observado con el valor estimado según modelo teórico donde se asuma independencia.

Para el modelo teórico, debemos considerar lo siguiente:

$P(X = x_i \wedge Y = y_j) = \frac{e_{ij}}{n}$, donde e_{ij} es la frecuencia real esperada de individuos que poseen simultáneamente los atributos x_i e y_j .

Note que en los datos de la tabla anterior lo que ocurre es que o_{ij} es una estimación de e_{ij}

Cola derecha

TODAS son de cola derecha

Si se cumple

$\chi^2_{obs} < \chi^2_c$, NO se rechaza H_0

Si se cumple

$p > \alpha$, NO se rechaza

Si se cumple

$\chi^2_{obs} > \chi^2_c$, se rechaza H_0

$p < \alpha$, se rechaza

Pruebas de independencia

Son un caso especial de bondad de ajuste

Ejemplo 4:

Los empleados de una empresa trabajan una jornada de 8 horas distribuidas en dos tandas de horario: el primero corresponde al bloque de la "mañana" en horario 8:00-12:00, luego se da una hora para el almuerzo y posteriormente llega el bloque de la "tarde" en horario 13:00-17:00. Se valoró si su desempeño, en cada uno de los bloques (mañana o tarde), era bueno, regular o malo. Con un nivel de significancia del 5%, ¿se puede asegurar que el desempeño de los empleados en ambos bloques de horarios son independientes?

		Y_1	Y_{TARDE}	Y_3	TOTAL X_i
		Bueno	Regular	Malo	
X_i MAÑANA	Bueno	56 O_{11}	71 O_{12}	12 O_{13}	139 T_{X1}
	Regular	47 O_{21}	163 O_{22}	38 O_{23}	248 T_{X2}
	Malo	14 O_{31}	42 O_{32}	85 O_{33}	141 T_{X3}
TOTAL Y_j		117 T_{Y1}	276 T_{Y2}	135 T_{Y3}	528

Sea X_i : Desempeño de los de la mañana
Sea Y_j : Desempeño de los de la tarde

Planteamiento de hipótesis:

H_0 : X_i, Y_j son independientes

H_1 : X_i, Y_j no son independientes

Al igual que en bondad de ajuste, se usan observados y esperados

Se calculan con:

$$e_{ij} = \frac{T_{X_i} \cdot T_{Y_j}}{n}, \quad T = \text{Total}$$

$n = \text{Total de elementos}$

		Y1	Y TARDE Y2	Y3	TOTAL X
		Bueno	Regular	Malo	
X MAÑANA	Bueno	56 O_{11}	71 O_{12}	12 O_{13}	139 T_{X1}
	Regular	47 O_{21}	163 O_{22}	38 O_{23}	248 T_{X2}
	Malo	14 O_{31}	42 O_{32}	85 O_{33}	141 T_{X3}
	TOTAL Y	117 T_{Y1}	276 T_{Y2}	135 T_{Y3}	528

Se tiene:

$$e_{11} = \frac{T_{X1} \cdot T_{Y1}}{n} = \frac{139 \cdot 117}{528} = 30,80$$

$$e_{12} = \frac{T_{X1} \cdot T_{Y2}}{n} = \frac{139 \cdot 276}{528} = 72,66$$

$$e_{13} = \frac{T_{X1} \cdot T_{Y3}}{n} = \frac{139 \cdot 135}{528} = 35,54$$

$$e_{21} = \frac{T_{X2} \cdot T_{Y1}}{n} = \frac{248 \cdot 117}{528} = 54,95$$

$$e_{22} = \frac{T_{X2} \cdot T_{Y2}}{n} = \frac{248 \cdot 276}{528} = 129,64$$

$$e_{23} = \frac{T_{X2} \cdot T_{Y3}}{n} = \frac{248 \cdot 135}{528} = 63,41$$

$$e_{31} = \frac{T_{X3} \cdot T_{Y1}}{n} = \frac{141 \cdot 117}{528} = 31,24$$

$$e_{32} = \frac{T_{X3} \cdot T_{Y2}}{n} = \frac{141 \cdot 276}{528} = 73,70$$

$$e_{33} = \frac{T_{X3} \cdot T_{Y3}}{n} = \frac{141 \cdot 135}{528} = 36,05$$

En base a esos resultados debe hacerse una nueva tabla similar pero con los nuevos datos

		Y1	Y2 TARDE	Y3	TOTAL
		Bueno	Regular	Malo	
X MAÑANA	Bueno	56 O_{11}	71 O_{12}	12 O_{13}	139 T_{x1}
	Regular	47 O_{21}	163 O_{22}	38 O_{23}	248 T_{x2}
	Malo	14 O_{31}	42 O_{32}	85 O_{33}	141 T_{x3}
	TOTAL	117 T_{y1}	276 T_{y2}	135 T_{y3}	528

} Tabla
vieja i

Tar de

	Bueno	Regular	Malo	Total	} Tabla nueva j
Bueno	30,80	72,66	35,54	139	
Regular	59,95	129,64	63,41	248	
Malo	31,24	73,70	36,05	141	
Total	117	276	135	528	

Para probar independencia cuando cada $e_{ij} \geq 5$

Estadístico de prueba: $\chi^2 = \sum_{i=1}^f \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
con $\nu = (f-1)(c-1)$, si $\nu \geq 2$.

Para el $\chi^2_{0,05}$,
se suman todos

$$\frac{(56 - 30,80)^2}{30,80} + \frac{(71 - 72,66)^2}{72,66} + \frac{(12 - 35,54)^2}{35,54} +$$

$$\frac{(47 - 59,95)^2}{59,95} + \frac{(163 - 129,64)^2}{129,64} + \frac{(38 - 63,41)^2}{63,41} +$$

$$\frac{(14 - 31,24)^2}{31,24} + \frac{(42 - 73,70)^2}{73,70} + \frac{(85 - 36,05)^2}{36,05}$$

$$\chi^2_{0,05} = 175,7812 \quad \alpha = 0,05 \quad \nu = (f-1)(c-1),$$

$$\nu = (3-1)(3-1) = 4$$

$$\chi^2_c = \chi^2_{0,05, 4} = 9,48773$$

$$3 \times 3 \rightarrow 2 \times 2$$

$$4 \times 4 \rightarrow 3 \times 3$$

$$N \quad \chi^2_{0,05} = 175,7812 > 9,48773$$

∴ se rechaza H_0 , NO es independiente

Con valor p

$$\chi^2_{obs} = 195,7812 \quad V = 9 \quad \alpha = 0,05$$

$$P(\chi^2 > 195,7812) = 0,00000$$

$$p = 0,00000 < \alpha = 0,05$$

Se rechaza H_0 , No es independiente

Ejemplo 3:

En un estudio de una vacuna de hepatitis participan **1083 voluntarios**. De éstos, se eligen aleatoriamente **549 y son vacunados**. Los otros, **534, no son vacunados**. Después de un cierto tiempo, se observa que 70 de los 534 **no vacunados** han contraído la hepatitis, mientras que sólo **11 de los 549 vacunados** la han contraído.

¿Es el hecho de contraer hepatitis independiente de haber sido vacunado contra la dolencia?

$V = \text{Vacunados}$ $NV = \text{No Vacunados}$
 $H = \text{Hepatitis}$ $NH = \text{No Hepatitis}$

	H	NH	Total _x
V	11	538	549
NV	70	464	534
Total _y	81	1002	1083 $\rightarrow n$

Calcular esperados $\frac{T_{xi} \cdot T_{yj}}{n}$, $n = 1083$

$$e_{11} = \frac{T_{x1} \cdot T_{y1}}{n} = \frac{549 \cdot 81}{1083} \approx 41,06$$

$$e_{12} = \frac{T_{x1} \cdot T_{y2}}{n} = \frac{549 \cdot 1002}{1083} \approx 507,99$$

$$e_{21} = \frac{T_{x2} \cdot T_{y1}}{n} = \frac{534 \cdot 81}{1083} \approx 39,99$$

$$e_{22} = \frac{T_{x2} \cdot T_{y2}}{n} = \frac{534 \cdot 1002}{1083} \approx 497,06$$

Tabla vieja

	H	NH	Total _x
V	11	538	549
NV	70	464	534
Total _y	81	1002	1083

Tabla nueva

	H	NH	Total _x
V	41,06	507,94	549
NV	39,94	497,06	534
Total _y	81	1002	1083

Para probar independencia cuando cada $e_{ij} \geq 5$

Estadístico de prueba: $\chi^2 = \sum_{i=1}^I \sum_{j=1}^C \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
con $\nu = (I-1)(C-1)$, si $\nu \geq 2$.

$$\chi^2_{obs} = \frac{(11 - 41,06)^2}{41,06} + \frac{(538 - 507,94)^2}{507,94} +$$

$$\frac{(70 - 39,94)^2}{39,94} + \frac{(464 - 497,06)^2}{497,06}$$

$$\chi^2_{obs} = 48,2388$$

$$\nu = (2-1)(2-1) = 1$$

$$\alpha = 0,05$$

$$\chi^2_c = \chi^2_{0,05, 1} = 3,84146$$

|| Como $\chi^2_{obs} = 48,2388 > \chi^2_c = 3,84146$,
se rechaza H_0 , NO son independientes

Con valor P

$$P(\chi^2 \geq 48,2388) = 0,0000 \quad \alpha = 0,05$$

$P = 0,0000 < \alpha = 0,05$, se rechaza H_0
No son independientes

Ejemplo

En una universidad se realiza un estudio para verificar si el tipo de trabajo (administrativo y docente) se relaciona con el grado de estrés (I, II y III) de los trabajadores. Para lo cual se elige una muestra aleatoria de 300 trabajadores y se clasifican en la tabla siguiente. (5 puntos)

	I	II	III
Administrativos	42	24	30
Docentes	54	78	72

Pruebe la hipótesis de que el tipo de trabajo afecta el grado de estrés del trabajador.

H_0 : El trabajo y estrés son independientes

H_1 : El trabajo y estrés no son independientes

Tabla inicial A = Administrativos D = Docentes

	I	II	III	Total x
A	42	24	30	96
D	54	78	72	207
Total y	96	102	102	300 $\rightarrow n$

Calcular esperados $\frac{T_{xi} \cdot T_{yj}}{n}$, $n = 300$

$$e_{11} = \frac{96 \cdot 96}{300} = 30,72$$

$$e_{21} = \frac{207 \cdot 96}{300} = 65,28$$

$$e_{12} = \frac{96 \cdot 102}{300} = 32,64$$

$$e_{22} = \frac{207 \cdot 102}{300} = 69,36$$

$$e_{13} = \frac{96 \cdot 102}{300} = 32,64$$

$$e_{23} = \frac{207 \cdot 102}{300} = 69,36$$

Tabla vieja

	I	II	III	Total x
A	92	29	30	96
D	59	78	72	209
Total y	96	102	102	300

Para probar independencia cuando cada $e_{ij} \geq 5$

Estadístico de prueba: $\chi^2 = \sum_{i=1}^f \sum_{j=1}^c \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$
con $\nu = (f-1)(c-1)$, si $\nu \geq 2$.

Tabla Nueva

	I	II	III	Total x
A	30,72	32,69	32,69	96
D	65,28	69,36	69,36	209
Total y	96	102	102	300

$$\chi^2_{obs} = \frac{(92 - 30,72)^2}{30,72} + \frac{(29 - 32,69)^2}{32,69} + \frac{(30 - 32,69)^2}{32,69} +$$

$$\frac{(59 - 65,28)^2}{65,28} + \frac{(78 - 69,36)^2}{69,36} + \frac{(72 - 69,36)^2}{69,36}$$

$$\chi^2_{obs} = 9,7683$$

$$\nu = (2-1) \cdot (3-1) = 2$$

$$\alpha = 0,05$$

$$\chi^2_c = \chi^2_{0,05,2} = 5,99146$$

R/ Como $\chi^2_{obs} = 9,7683 > \chi^2_c = 5,99146$
se rechaza H_0 , NO son independientes

Con valor P

$$P(\chi^2 > 9,7683) = 0,007566 \quad \alpha = 0,05$$

Como $P = 0,007566 < 0,05$ se rechaza H_0 ,
NO son independientes