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 \ Y	y_1	y_2	$\cdots$	y_p	TOTALES
x_1	o_{11}	o_{12}	$\cdots$	o_{1p}	Tx_1
x_2	o_{21}	o_{22}	$\cdots$	o_{2p}	Tx_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	o_{m1}	o_{12}	$\cdots$	o_{mp}	Tx_m
TOTALES	Ty_1	Ty_2	$\cdots$	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_ie 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

Si se cumple $V_{obs} < V_C$, NO se rechaza H_0

Si se cumple $V_{obs} > V_C$, Se rechaza H_0

TODAS son de cola derecha

Si se cumple $P > \alpha$, No se rechaza

$P < \alpha$, 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_2	Y_3	
		Bueno	Regular	Malo	TOTAL X
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

$\to n$

Sea X: Desempeño de los de la mañana

Sea Y: Desempeño de los de la tarde

Planteamiento de hipotesis:

H_0: X, Y son independientes

H_1: X, Y no son independientes

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

Se calculan con:

$$e_{ij} = \frac{T_{x_i} \cdot T_{y_j}}{n}, \quad T = \text{Total}, \quad n = \text{Total de elementos}$$

		TARDE Y_1 Bueno	TARDE Y_2 Regular	TARDE Y_3 Malo	TOTAL X
MAÑANA X	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

		Y_1	Y TARDE Y_2	Y_3	
		Bueno	Regular	Malo	TOTAL
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

Tarde

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

Tabla nueva j

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

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

Para el χ^2_{obs}, se suman todos

$$\frac{(56-30,80)^2}{30,80} + \frac{(71-72,66)^2}{72,66} + \frac{(12-35,54)^2}{35,53} +$$

$$\frac{(47-54,95)^2}{54,95} + \frac{(163-129,64)^2}{129,63} + \frac{(38-63,41)^2}{63,40} +$$

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

$\chi^2_{obs} = 145,7812$ $\quad \alpha = 0,05$ $\quad \nu = (f-1)(c-1)$,

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

$\chi^2_c = \chi^2_{0.05,\,4} = 9,48773$ $\quad 3 \times 3 \to 2 \times 2$

$4 \times 4 \to 3 \times 3$

$\boxed{\therefore\ \chi^2_{obs} = 145,7812 > 9,48773 \quad \therefore \text{ se rechaza Ho, NO es independiente}}$

Con valor p

$\chi^2_{obs} = 145{,}7812$ $\quad v = 4$ $\quad \alpha = 0{,}05$

$P(\chi^2 > 145{,}7812) = 0{,}00000$

$$\boxed{\begin{array}{l} p = 00000 < \alpha = 0{,}05 \\ \text{Se rechaza } H_0\text{, No es independrente} \end{array}}$$

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 = Vacunados NV = No Vacunados
H = Hepatitis NH = No Hepatitis

	H	NH	$Total_x$
V	11	538	549
NV	70	464	534
$Total_y$	81	1002	1083 $\longrightarrow n$

Calcular esperados $\dfrac{T_{x_i} \cdot T_{y_j}}{n}$, $n = 1083$

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

$$e_{12} = \frac{T_{x1} \cdot T_{y2}}{n} = \frac{549 \cdot 1002}{1083} \simeq 507{,}94$$

$$e_{21} = \frac{T_{x2} \cdot T_{y1}}{n} = \frac{534 \cdot 81}{1083} \simeq 39{,}94$$

$$e_{22} = \frac{T_{x2} \cdot T_{y2}}{n} = \frac{534 \cdot 1002}{1083} \simeq 494{,}06$$

Tabla vieja

	H	NH	Total x
V	11	538	549
NV	70	464	534
Total y	81	1002	1083

$\} i$

Tabla nueva

	H	NH	Total x
V	41,06	507,94	549
NV	39,94	494,06	534
Total y	81	1002	1083

$\} j$

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 $\nu \geq 2$.

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

$$\frac{(70-39,94)^2}{39,94} + \frac{(464-494,06)^2}{494,06}$$

$\chi^2_{obs} = 48,2388$ $\qquad V = (2-1)(2-1) = 1$

$\alpha = 0,05$

$\chi^2_c = \chi^2_{0.05,1} = 3,84146$

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

Con valor P

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

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

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 estres son independientes
H_1: El trabajo y estres no son independientes

Tabla inicial   A = Admistrativos   D = Docentes

	I	II	III	Total x
A	42	24	30	96
D	54	78	72	204
Total y	96	102	102	300 → n

Calcular esperados $\frac{T_{x_i} \cdot T_{y_j}}{n}$, $n = 300$

$$e_{11} = \frac{96 \cdot 96}{300} = 30{,}72 \qquad e_{21} = \frac{204 \cdot 96}{300} = 65{,}28$$

$$e_{12} = \frac{96 \cdot 102}{300} = 32{,}64 \qquad e_{22} = \frac{204 \cdot 102}{300} = 69{,}36$$

$$e_{13} = \frac{96 \cdot 102}{300} = 32{,}64 \qquad e_{23} = \frac{204 \cdot 102}{300} = 69{,}36$$

Tabla vieja

	I	II	III	Total x
A	42	24	30	96
D	54	78	72	204
Total y	96	102	102	300

i

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	32,64	32,64	96
D	65,28	69,36	69,36	204
Total y	96	102	102	300

j

$$\chi^2_{obs} = \frac{(42-30{,}72)^2}{30{,}72} + \frac{(24-32{,}64)^2}{32{,}64} + \frac{(30-32{,}64)^2}{32{,}64} + \frac{(54-65{,}28)^2}{65{,}28} + \frac{(78-69{,}36)^2}{69{,}36} + \frac{(72-69{,}36)^2}{69{,}36}$$

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

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

$\alpha = 0{,}05$

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

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

Con valor P

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

$\boxed{\text{Como } P = 0{,}007566 < 0{,}05 \text{ se rechaza Ho, NO son independientes}}$