Prueba de Independencia χ^2
(Prueba de Pearson)

DANIEL LEFF YAFFE

INDICE – PRUEBA DE INDEPENDENCIA

INTRODUCCIÓN

- Pruebas de Pearson
  - Bondad y Ajuste: uso de datos muestrales para probar si la población tiene “x” distribución.
  - **Independencia: uso de datos muestrales para probar la independencia entre dos variables.**
- Éstas pruebas de hipótesis se basan en qué tan “cerca” se encuentran las frecuencias muestrales de las frecuencias esperadas.

EJEMPLO Y PASOS

- Corona tiene 3 tipos de cerveza; ligera, clara y oscura.
- Queremos saber si las preferencias de los consumidores por estos tipos de cerveza difieren entre hombres y mujeres.
  - Si las preferencias son independientes del género del consumidor → Corona hace una campaña publicitaria para todas sus cervezas.
  - Si las preferencias NO son independientes del género del consumidor → Corona ajusta la publicidad de acuerdo al tipo de cerveza.

EJEMPLO Y PASOS

- Para saber si las preferencias de los consumidores son independientes de su género, Corona hace una encuesta aleatoria y obtiene una “TABLA DE CONTINGENCIA”

		Cerveza Preferida			
		Ligera	Clara	Oscura	**Total**
Género	Hombre	20	40	20	**80**
	Mujer	30	30	10	**70**
	Total	**50**	**70**	**30**	**150**

EJEMPLO Y PASOS

- I. Establecer Hipótesis Nula y Alternativa
  - Ho: La variable "X" (de las columnas) es independiente de la variable "Y" (de los renglones).
    - La preferencia por un tipo de cerveza es independiente del género del consumidor.
  - Ha: La variable "X" (de las columnas) NO es independiente de la variable "Y" (de los renglones).
    - La preferencia por un tipo de cerveza NO es independiente del género del consumidor.

EJEMPLO Y PASOS

- 2. Seleccionar una muestra aleatoria y hacer una "Tabla de Contingencia" con las frecuencias observadas.

		Cerveza Preferida			
		Ligera	Clara	Oscura	**Total**
Género	Hombre	20	40	20	**80**
	Mujer	30	30	10	**70**
	Total	**50**	**70**	**30**	**150**

EJEMPLO Y PASOS

- 3. Suponiendo que la hipótesis nula es verdadera (las variables son independientes), determinar para cada contingencia las frecuencias esperadas.

OBSERVADA

	Ligera	Clara	Oscura	Total
Hombre	20	40	20	**80**
Mujer	30	30	10	**70**
Total	**50**	**70**	**30**	**150**

ESPERADA BAJO INDEPENDENCIA

	Ligera	Clara	Oscura	Total
Hombre	26.67	37.33	16.00	**80**
Mujer	23.33	32.67	14.00	**70**
Total	**50**	**70**	**30**	**150**

$$e_{\text{Hombre,Ligera}} = 26.67 = \left(\frac{50}{150}\right) * 80 = \frac{50 * 80}{150}$$

$$e_{i,j} = \frac{Total_i * Total_j}{Total\ Muestra}; \qquad fila\ "i", columna\ "j"$$

EJEMPLO Y PASOS

- 4. Calcular el estadístico de independencia Pearson:

$$\chi^2_{(n-1)(m-1)} = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

$$f_{ij} = Frecuencia\ Observada_{ij}$$
$$e_{ij} = Frecuencia\ Esperada_{ij}$$

n filas, m columnas

EJEMPLO Y PASOS

$$\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(f_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i=1}^{2}\left[\frac{(f_{i1}-e_{i1})^2}{e_{i1}}+\frac{(f_{i2}-e_{i2})^2}{e_{i2}}+\frac{(f_{i3}-e_{i3})^2}{e_{i3}}\right]$$

$$=\sum_{i=1}^{2}\frac{(f_{i1}-e_{i1})^2}{e_{i1}}+\sum_{i=1}^{2}\frac{(f_{i2}-e_{i2})^2}{e_{i2}}+\sum_{i=1}^{2}\frac{(f_{i3}-e_{i3})^2}{e_{i3}}$$

$$=\frac{(f_{11}-e_{11})^2}{e_{11}}+\frac{(f_{12}-e_{12})^2}{e_{12}}+\frac{(f_{13}-e_{13})^2}{e_{13}}+\frac{(f_{21}-e_{21})^2}{e_{21}}+\frac{(f_{22}-e_{22})^2}{e_{22}}+\frac{(f_{23}-e_{23})^2}{e_{23}}$$

EJEMPLO Y PASOS

i	j	Género	Preferida	f_{ij}	e_{ij}	$(f_{ij} - e_{ij})^2 / e_{ij}$
1	**1**	H	Ligera	20	26.67	1.67
1	**2**	H	Clara	40	37.33	0.19
1	**3**	H	Oscura	20	16.00	1.00
2	**1**	M	Ligera	30	23.33	1.90
2	**2**	M	Clara	30	32.67	0.22
2	**3**	M	Oscura	10	14.00	1.14
						6.12

$$\chi^2_{(n-1)(m-1)} = 6.12$$

EJEMPLO Y PASOS

- 5. Rechazar Ho si:

$$Pvalue \leq \alpha \quad \leftrightarrow \quad \chi^2_{(n-1)(m-1)} \geq \chi^2_{\alpha}$$

Alpha	0.05
Test Statistic	6.12
Filas (n)	2
Columnas (m)	3
Grados de Libertad (m-1)*(n-1)	2

P-value	0.0468
Alpha	0.05

Test Statistic	6.12
Valor Teórico	5.99

} Rechazar Ho

MINITAB

		Cerveza Preferida			
		Ligera	Clara	Oscura	**Total**
Género	Hombre	20	40	20	**80**
	Mujer	30	30	10	**70**
	Total	**50**	**70**	**30**	**150**

Ejercicio

- Considera las aerolíneas A, B y C. En la siguiente tabla de contingencia se muestra, en minutos, el retraso de 240 vuelos.
  - ¿El retraso es independiente de la aerolínea? Usa un nivel de significancia del 1%.

		Aerolínea		
		A	B	C
Minutos tarde	0	20	30	20
	$0 < x \leq 30$	30	60	25
	$30 < x$	10	15	30

¡Gracias!