Dra. Teresa Peralta

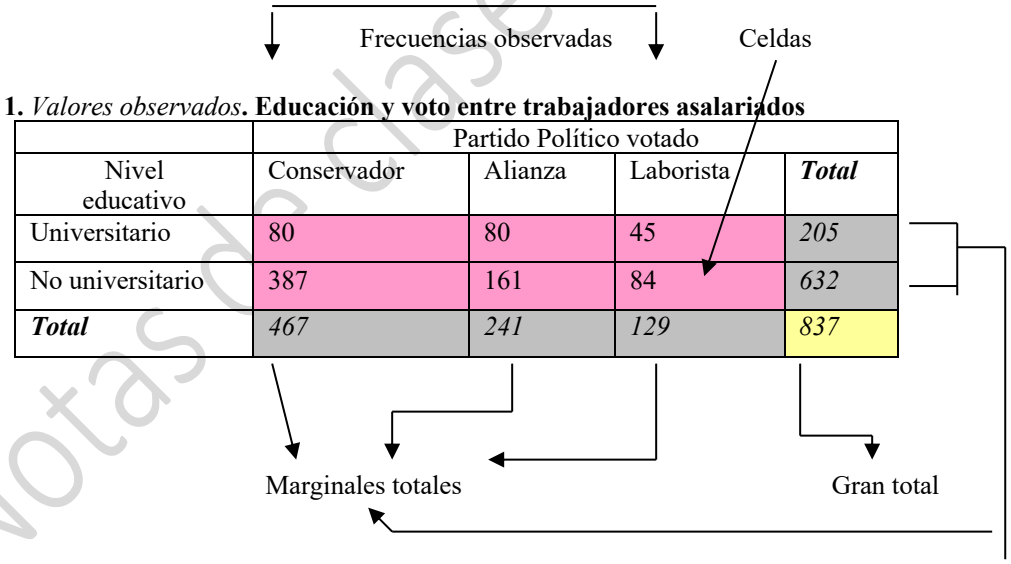
Pruebas para analizar datos categóricos
Técnicas no paramétricas

Prueba de la Ji Cuadrada de *Pearson* [χ^2] de asociación o independencia

Permite saber si existe una relación significativa entre dos variables categóricas, o si las variables son independientes entre sí. Si la respuesta es negativa entonces se tiene evidencia de que las variables están *asociadas* de alguna forma.

En otras palabras, se utiliza para analizar datos categóricos cuando dos (o más) muestras son independientes y sirve para evaluar la hipótesis de *no asociación* entre variables presentadas en una **tabla de contingencia** o de frecuencia[1]

Ejemplo
Estudio con 852 trabajadores asalariados que podían votar en las elecciones inglesas en 1987. Los sujetos fueron clasificados en función de si tenían educación universitaria o no y además se les preguntó por cuál de los tres partidos políticos principales votaron. 15 personas que votaron por otros partidos fueron excluidas del estudio.



1. *Valores observados*. **Educación y voto entre trabajadores asalariados**

	Partido Político votado			
Nivel educativo	Conservador	Alianza	Laborista	***Total***
Universitario	80	80	45	*205*
No universitario	387	161	84	*632*
Total	*467*	*241*	*129*	*837*

Lo que se quiere averiguar es si existe alguna relación entre educación universitaria y voto emitido.

[1] También conocida como una tabla de contingencia 2 x 2 (de doble entrada) o (*r* x *c*).

Primer paso
Escribir las hipótesis nula y alterna.

H_0 = No existen ninguna asociación entre la educación universitaria y el partido (político) por el cual se vota. H_1 = Existe una asociación entre educación universitaria y el partido (político) por el cual se vota.

En este estudio la H_1 supondría que la variable independiente (el nivel educativo de los trabajadores) está *relacionada* con la variable dependiente (el voto que le otorgan a determinado partido político). El análisis estadístico permite saber si, efectivamente, hay una *asociación* entre la variable independiente y la variable dependiente. En otras palabras, si hubo una asociación entre las respuestas proporcionadas por los sujetos y su pertenencia a un grupo y no a otro.

En lo que respecta a la H_0, en ésta, a diferencia de la anterior, se plantea que la variable explicativa y la variable de respuesta son en realidad independientes, es decir *no hay una asociación* entre la primera y la segunda; es la negación del punto que se está intentando probar.

El procedimiento es rechazar H_0 en favor de H_1, si la prueba estadística (en este caso la prueba de χ^2) proporciona un valor cuya probabilidad de ocurrencia asociada de acuerdo con H_0 es igual o menor que alguna probabilidad pequeña, generalmente denotada por **[α].**

A esta probabilidad se le conoce como el **nivel de significación**.

Segundo paso
Las respuestas del estudio son los **valores observados** que se muestran en la tabla de contingencia (2 x 3) sólo incluye los valores observados.

Para la prueba de la Ji Cuadrada es necesario calcular los **valores esperados.** Estos son los valores que uno esperaría encontrar en cada una de las celdas si no hubiera asociación entre las dos variables en cuestión.

La prueba compara los valores observados y los valores esperados para ver que si son significativamente diferentes.

Para calcular los valores esperados se usa la siguiente fórmula:

$$\text{Valor esperado} = \frac{\text{total de la columna x total de la fila}}{\text{gran total}}$$

Por ejemplo, para los votantes conservadores con educación universitaria:

$$\text{Valor esperado} = \frac{467 \times 205}{837} = 114.38$$

2. *Valores esperados*. **Educación y voto entre trabajadores asalariados**

	Partido Político votado			
Nivel educativo	Conservador	Alianza	Laborista	***Total***
Universitario	114.38	59.03	31.59	*205*
No universitario	352.62	181.97	97.41	*632*
Total	*467*	*241*	*129*	*837*

Tercer paso
Ahora se puede calcular la prueba estadística. La fórmula para calcular la Ji Cuadrada es la siguiente:

$$\chi^2 = \Sigma \frac{(0 - E)^2}{E}$$

O	E	(O – E)	$(O - E)^2$	$\frac{(O - E)^2}{E}$
80	114.38	-34.38	1181.98	10.33
80	59.03	20.97	439.74	7.45
45	31.59	13.41	179.83	5.69
387	352.62	34.38	1181.98	3.35
161	181.97	-20.97	439.74	2.42
84	97.41	-13.41	179.83	1.85
			Total	31.09

Por lo tanto χ^2 en este caso es de **31.09**

Cuarto paso
A continuación es necesario encontrar un **valor crítico** (tomado de una tabla de χ^2). Estos valores se encuentran en una tabla, fácilmente asequible en cualquier libro de estadística o página de internet que discuta esta prueba, que contienen dichos valores.

Además, es necesario determinar cuál es el **nivel de significación** (los valores comunes de **α** son 0.05 y 0.01) que se quiere usar y los **grados de libertad**.

La manera como se distribuye la muestra de una Ji Cuadrada no está determinada por el número de categorías de una tabla, sino por una propiedad conocida como **grados de libertad (gl)** entendidos como *un índice de la cantidad de variabilidad que ocurre por azar y que puede presentarse en una situación dada.*
Computacionalmente, 1*gl* es igual número de filas menos 1 por el número de columnas menos 1.

$$gl = (\#\text{ de filas} - 1) \times (\#\text{ de columnas} - 1)$$

En el ejemplo presentado se tiene (2 – 1) x (3 – 1) = 1 x 2 = **2**.

Con esta información se determina si el resultado obtenido es estadísticamente significativo y en consecuencia se puede proceder a descartar o no la H_0. Se dice que un **valor es estadísticamente significativo** si es igual o mayor al valor crítico de la Ji Cuadrada en el nivel de significación previamente determinado.

Valores críticos de la Ji Cuadrada

	Nivel de significación				
gl	**.05**	**.025**	**.01**	**.005**	**.001**
1	3.84	5.02	6.63	7.88	10.83
2	5.99	7.38	9.21	10.60	13.82

Quinto paso

Conclusiones
En el ejemplo estudiado, el resultado obtenido es $\chi^2 = 31.09$ con 2*gl*. Este resultado se puede comparar contra los valores críticos de la Ji Cuadrada -mostrados parcialmente en la tabla anterior. Para que el valor observado de la Ji Cuadrada del ejemplo sea estadísticamente significativo debe ser ≥ a 5.99 en el nivel de significación mínimo de .05. Queda claro que 31.09 es extremadamente[2] estadísticamente significativo no sólo en nivel umbral P = .05 sino más allá del nivel más riguroso P = .001, y en consecuencia es posible sostener que hay evidencia muy fuerte en contra de la H_0.

[2] Esta denominación, sin embargo, no es siempre usada por los programas que calculan estas pruebas puesto que se parte del supuesto que, una vez que se ha decidido sobre el nivel umbral del valor de *p* para la significación estadística, si el valor de $p \leq \alpha$ entonces el resultado es o no es estadísticamente significativo. Este punto se retoma más adelante.

En otras palabras, se encontró que hay asociación entre la educación universitaria y el voto otorgado a determinado partido político entre trabajadores asalariados ingleses en la elección de 1987.

RESTRICCIONES

Por último, es necesario mencionar una restricción que tiene la prueba y se refiere a la necesidad de que el tamaño de la muestra sea grande y el número de frecuencias (esperadas) de cada celda de la tabla sea grande: siempre mayor a 5 en tablas de [2 x 2], o 5 o más en 80% de las celdas en tablas más grandes [r x c], pero nunca celdas con frecuencias 0[3].

La Corrección de Yates

Cuando el número de frecuencias de una o varias de las celdas es menor a 5, la aproximación de la Ji Cuadrada puede ser mejorada reduciendo el valor absoluto de las diferencias entre las frecuencias esperadas y observadas en 0.5 antes de hacer el cálculo al cuadrado. Este ajuste de 0.5, que hace la estimación más conservadora, es conocido como la **Corrección de Yates** y se aplica sólo a tablas de contingencia de [2 x 2].

Se dice que es conservador en el sentido de que hace que sea más difícil establecer que la distribución de las columnas y filas sea estadísticamente significativa, o dicho en otras palabras hace más difícil rechazar la hipótesis nula. En muchas investigaciones se prefiere usar la prueba de la Ji Cuadrada siempre con la Corrección de Yates, aún cuando no haya frecuencias menores a 5 en una o más celdas.

Fórmula de la Ji Cuadrada con corrección de Yates

$$\chi^2 = \sum \frac{((0 - E) - 0.5)^2}{E}$$

[3]La razón de ser de esta restricción es que la prueba de la Ji Cuadrada mide las probabilidades para cada celda y cuando las frecuencias esperadas caen por debajo de 5, dichas probabilidades no pueden ser calculadas con suficiente precisión.