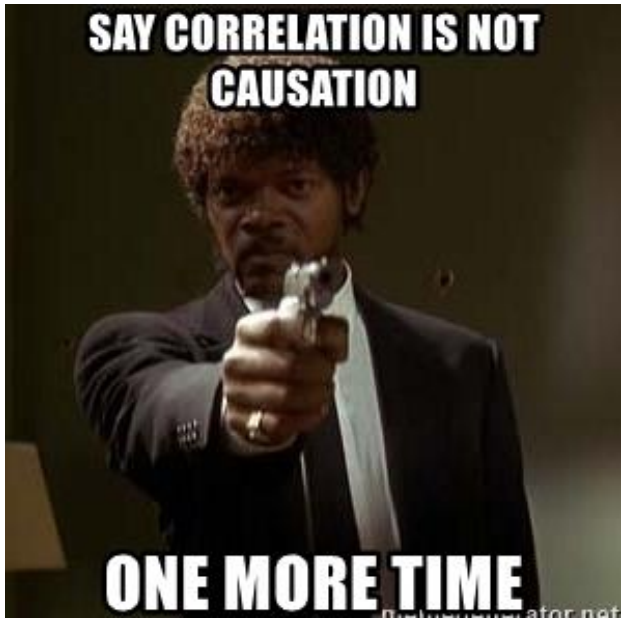

PI DATA STRATEGY & CONSULTING

“Say correlation is not causation … one more time”

Johanna Frau
Data Scientist
Pi Data Strategy & Consulting

“Say correlation is not causation … one more time”



Estadística para Data Science

4° encuentro

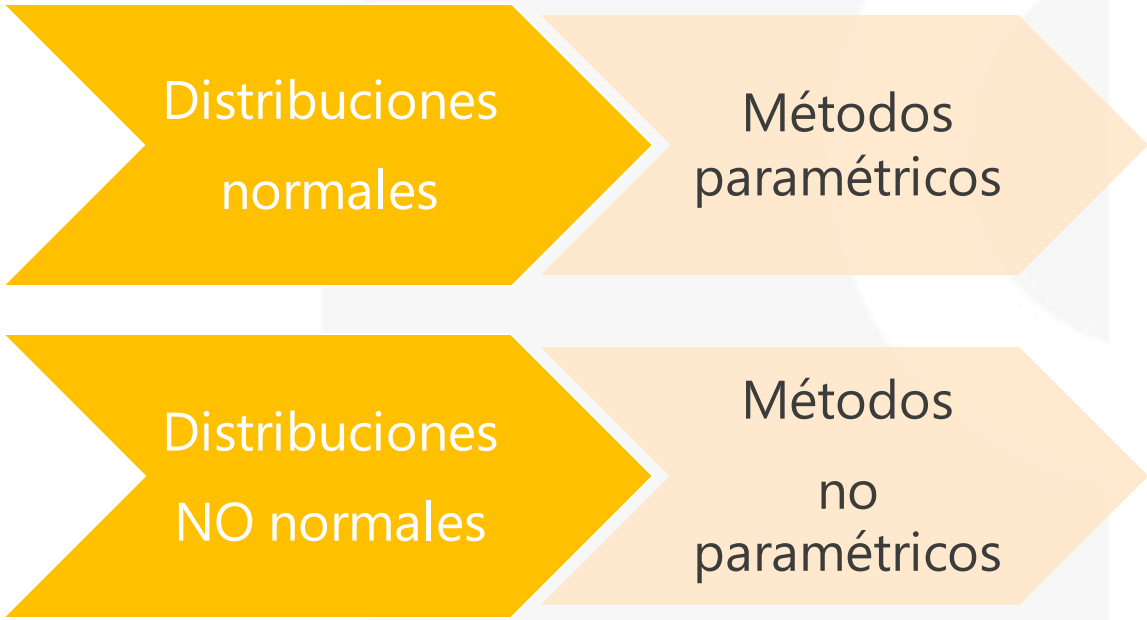
COEFICIENTES DE CORRELACIÓN

Agenda

- Introducción
- Coeficientes de correlación
  - Pearson
  - Spearman
  - Kendall
- Test de correlación

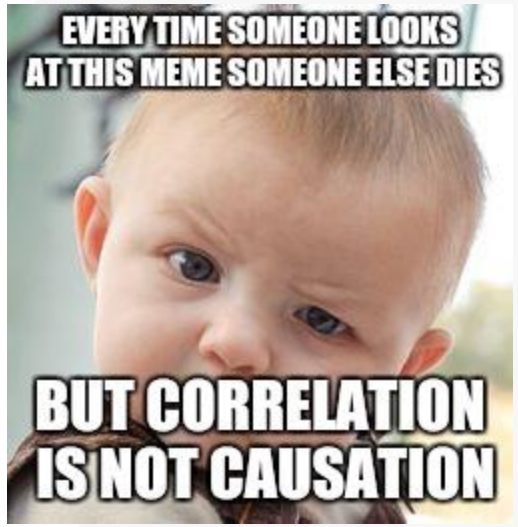
¿Qué es la correlación?

La correlación es el área dentro de estadística que estudia la relación sistemática entre dos o más variables.



Coeficientes de correlación

Los ***coeficientes de correlación*** miden el grado en el cual dos variables tienden a cambiar juntas y determinan la fortaleza y la dirección de la relación entre ambas.



Coeficientes de correlación

Pearson

- Cada variable debe ser continua y estar normalmente distribuida.
- Asume que existe una relación lineal entre ambas variables y que los datos se distribuyen por igual alrededor de una línea recta.
- Es sensible a outliers.

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \cdot \sum(Y - \overline{Y})^2}}$$

$$Where,\ \overline{X} = mean\ of\ X\ variable$$
$$\overline{Y} = mean\ of\ Y\ variable$$

Coeficientes de correlación

Spearman

- Las variables pueden ser ordinales, continuas o de razón.
- Asume que existe una relación monótona entre ambas variables.
- No asume distribución normal de las variables.
- Sensible a errores o discrepancias en los datos.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

donde

$d_i = \mathrm{rg}(X_i) - \mathrm{rg}(Y_i)$ Diferencia entre los dos rangos de cada observación

n: Número de observaciones

Coeficientes de correlación

Kendall

- Las variables deben estar medidas en una escala ordinal, de intervalo o de razón.
- Asume que existe una relación monótona entre ambas variables.
- Menos sensible a errores.

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$$

$$Where,\ n_c = number\ of\ concordant\ pairs$$
$$n_d = number\ of\ discordant\ pairs$$
$$n = number\ of\ pairs$$

Pasos para realizar un test estadístico

Planteamiento de hipótesis:

Ho: Hipótesis nula - H1: Hipótesis alternativa

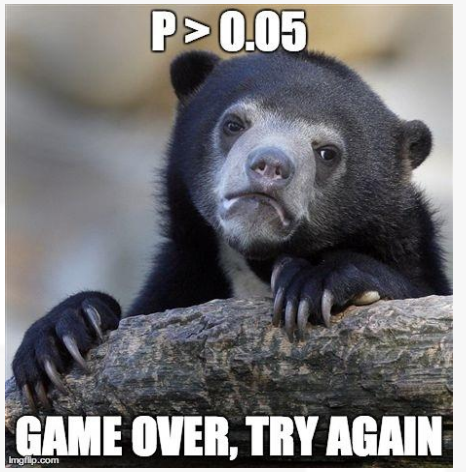
Elegir un nivel de significancia (α: alpha)

Selección del test estadístico

(Selección del estadístico de prueba)

Comparar el p-valor con α

Realizar la conclusión correspondiente



Test de correlación: H0 y H1

Test

- H0: No existe correlación entre pares de variables
- H1: Existe correlación entre pares de variables.

En símbolos

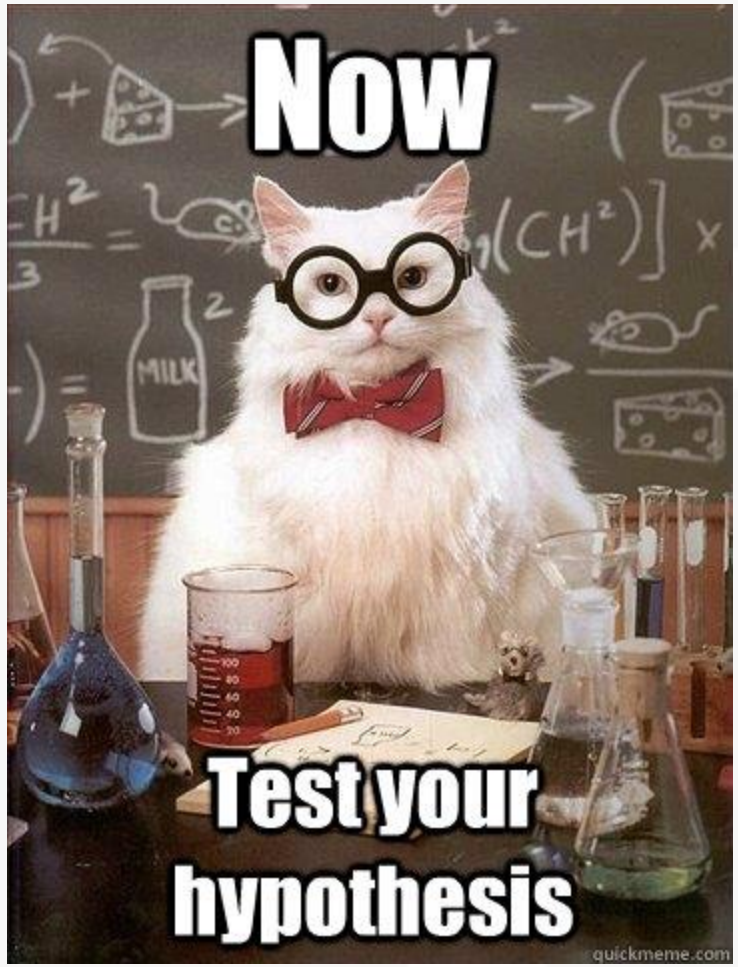
$$\boldsymbol{H0}: \rho = 0$$
$$\boldsymbol{H1}: \rho \neq 0$$

Manos a la obra

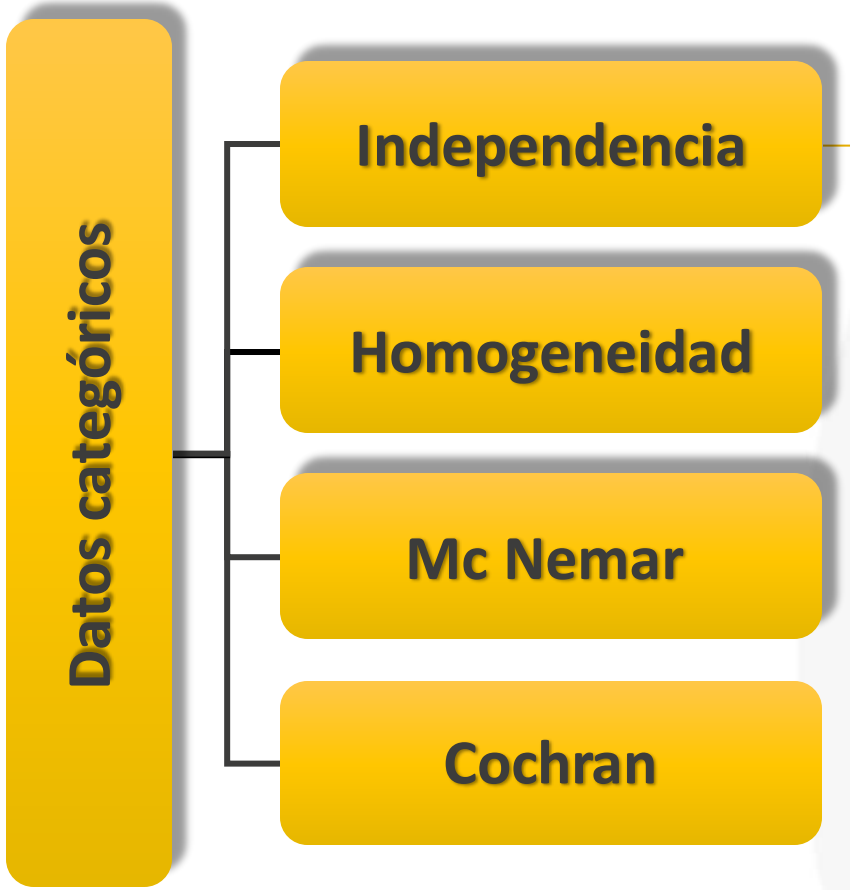
PRUEBAS PARA VARIABLES CATEGÓRICAS

Agenda

- Test de Independencia
- Test de Homogeneidad
- Test de McNemar
- Test Q de Cochran



Tipos de Test: Independencia



Se utiliza para mostrar si *dos* conjuntos de datos categóricos son *independientes* entre sí sobre una misma población.

Test de Independencia: H0 y H1

Test de Independencia

- H0: Las dos variables son independientes (no existe relación entre ambas).
- H1: Las dos variables son dependientes.

Tablas de contingencia

Material	Location 1	2	3	Total
Glass	8	23	29	60
Cardboard	28	61	91	180
Plastic	39	85	116	240
Metal	25	31	64	120
Total	100	200	300	600

- Presenta el número o la proporción de observaciones que caen en cada combinación posible de valores para cada una de las variables.
- Proporciona una vista rápida de la interacción entre las variables.

Frecuencias marginales

El test de independencia compara los datos de la muestra en la tabla de contingencia con la distribución de valores que se esperaría si la hipótesis nula (H0) es correcta.

Valores esperados y grados de libertad

Valor esperado para cada elemento de la matriz de contingencia:

$$\text{Valor esperado} = \frac{\text{Total de las filas} \times \text{Total de las columnas}}{\text{Total}}$$

Grados de libertad para el test:

$$\text{grados\_libertad} = (\text{nº de filas} - 1) \times (\text{nº de columnas} - 1)$$

Veamos un ejemplo...

¿Las posibilidades de sobrevivir a un accidente son independientes en cuanto a ser hombre, mujer o niño?

	MUJER	HOMBRE	NIÑO	Total de filas
NO SOBREVIVIÓ	7	9	20	40
SOBREVIVIÓ	129	46	215	390
Total de columnas	136	55	239	N=430

40*136/430 = 12.65

	MUJER	HOMBRE	NIÑO	Total de filas
NO SOBREVIVIÓ	7 (12,65)	9 (5,12)	20 (22,23)	40
SOBREVIVIÓ	129 (123,35)	46 (49,88)	215 (216,77)	390
Total de columnas	136	55	239	N=430

Estadístico de chi-cuadrado

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Donde

- O_i: es el número de ***casos observados*** en la categoría i.
- E_i: es el número de ***casos esperados*** en la categoría i.

Tablas de contingencia 2x2

Frecuencias esperadas entre 5 y 10

Correción de Yates para continuidad

$$\chi^2(\text{corregida}) = \sum_i \frac{(|o_i - e_i| - 0.5)^2}{e_i}.$$

Frecuencias esperadas menores que 5.

Prueba exacta de Fisher

Podria evitarse tomando una muestra mas grande.

Tipos de Test: Homogeneidad

Datos categóricos

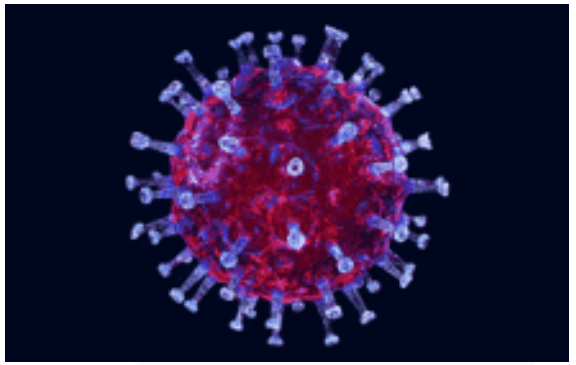
- Independencia
  - Se utiliza para mostrar si *dos* conjuntos de datos categóricos son *independientes* entre sí sobre una misma población.
- **Homogeneidad**
  - Determina si *dos o más poblaciones* (o grupos) tienen la misma distribución de *una variable categórica*.
- **McNemar**
- **Q de Cochran**

Ejemplo

La UNC quiere saber como reciben los estudiantes de diversas carreras la enseñanza virtual durante la pandemia del año 2020.

Se recopilan datos de las siguientes carreras:

- Psicología
- Ingeniería
- Odontología
- Arquitectura



¿La proporción de estudiantes de las 4 carreras y su conformidad con la enseñanza virtual es significativamente diferente?

Test de Homogeneidad: H0 y H1

Test de Homogeneidad

- H0: La proporción de la variable en cuestión no es significativamente diferente en todas las poblaciones estudiadas.
- H1: Al menos una de las proporciones de la variable es significativamente diferente.

En símbolos

$$\boldsymbol{H0}: p_1 = p_2 = \ \ldots = \ p_n$$
$$\boldsymbol{H1}: \exists p_i \neq p_j \text{ para algún j=1,2, } \ldots, n$$

Ejemplo: H0 y H1

H0: Para cada una de opciones en la encuesta anterior, las proporciones de estudiantes de Psicología, Ingeniería, Odontología y Arquitectura ***no son significativamente diferentes***.

H1: Para al menos una de las opciones, las proporciones de estudiantes de Psicología, Ingeniería, Odontología y Arquitectura ***son significativamente diferente***.

Consideraciones importantes

Test de independencia

Las unidades de observación se recogen al azar de ***UNA POBLACIÓN*** y se observan ***DOS VARIABLES CATEGÓRICAS***.

Test de Homogeneidad

Los datos se recopilan mediante muestro aleatorio de cada subgrupo ***(VARIAS POBLACIONES)*** por separado.

La HO es que cada subgrupo comparte la misma distribución de una ***VARIABLE CATEGÓRICA ÚNICA***.

Tipos de Test: McNemar

Datos categóricos

- **Independencia**
  - Se utiliza para mostrar si *dos* conjuntos de datos son *independientes* entre sí sobre una misma población.
- **Homogeneidad**
  - Determina si *dos o más poblaciones* (o grupos) tienen la misma distribución de *una variable categórica*.
- **Mc Nemar**
  - *Determina si existe una diferencia significativa en proporciones entre datos pareados sobre dos grupos.*
- **Cochran**

McNemar: Ingredientes

- ***Una variable nominal*** dependiente con dos categorías (variables dicótomicas).
- ***Una variable independiente*** con dos grupos relacionados.
- Los grupos en la variable dependiente deben ser mutuamente excluyentes.
- La muestra debe ser aleatoria sobre la población de interés.

Test de McNemar: H0 y H1

Test de McNemar

- H0: No hay cambios significativos entre los dos grupos de interés.
- H1: Hay cambios significativos entre los dos grupos de interés.

En símbolos

$$\boldsymbol{H0}: p_b = p_c$$
$$\boldsymbol{H1}: p_b \neq p_c$$

¿b, c?

Los definiremos pronto!

Algunos ejemplos

- Comparar el rendimiento en un examen (Aprobado ó Desaprobado) de dos grupos.
- Comparar la respuesta a un tratamiento médico de dos grupos (Tratamiento/Control).
- En machine learning, comparar la precisión predictiva de dos modelos.

Tabla de contingencia 2x2

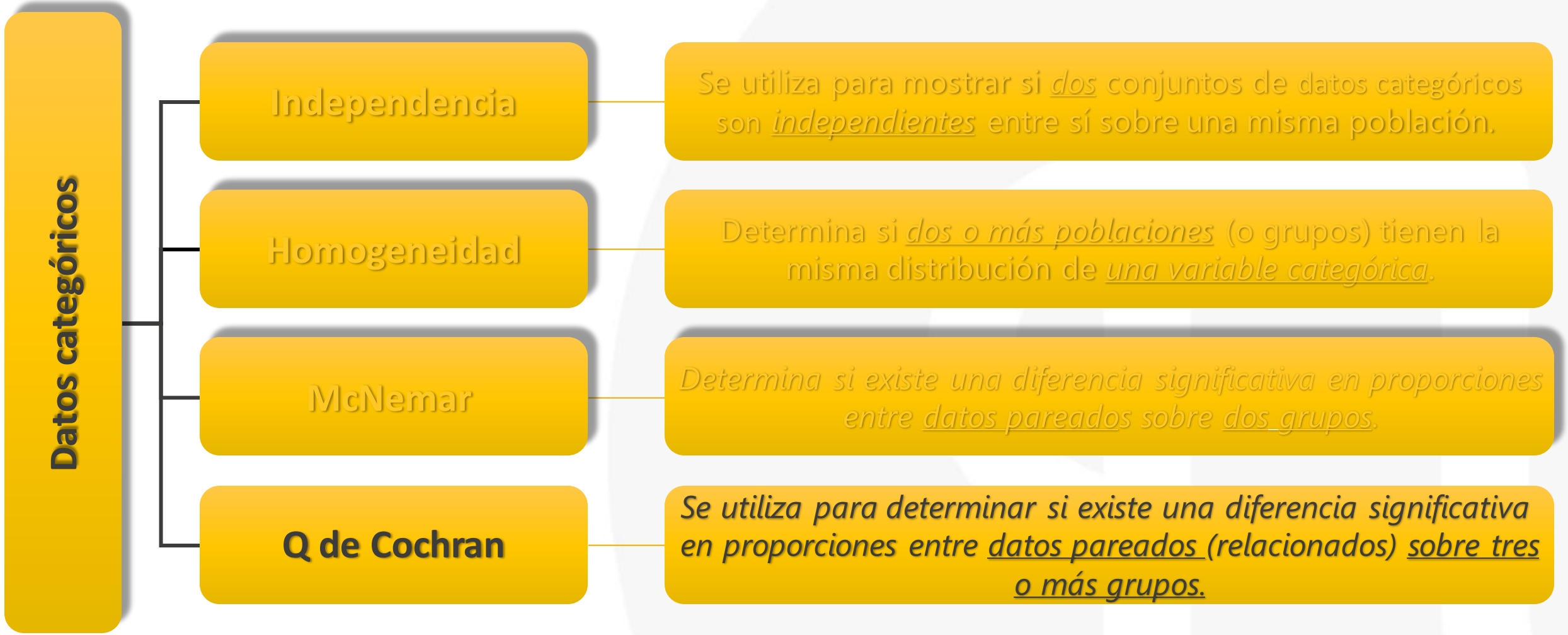
Tabla de contingencia y Estadístico de prueba

Después del Tratamiento

Antes del Tratamiento	**SI**	**NO**	
SI	80 → *a*	100 → *b*	180
NO	10 → *c*	110 → *d*	120
	90	210	300

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Tipos de Test: Test Q de Cochran



Test Q de Cochran: Ingredientes

- *Una variable nominal* dependiente con dos categorías (variables dicótomicas).
- *Una variable independiente* con tres o más grupos relacionados.
- Los dos grupos en la variable dependiente deben ser mutuamente excluyentes.
- La muestra debe ser aleatoria sobre la población de interés.

Test Q de Cochran: H0 y H1

Test Q de Cochran

- H0: No hay cambios significativos entre los grupos de interés.
- H1: Hay cambios significativos entre los grupos de interés.

En símbolos

$$\boldsymbol{H0}: p_1 = p_2 = \cdots = p_n$$
$$\boldsymbol{H1:}\ p_i \neq p_j \textit{ para al menos algún par } p_i, p_j$$

Tabla de contingencia y Estadístico de prueba

Tratamientos

Personas

	T1	T2	T3	T4
P1	1	0	1	1
P2	0	0	1	0
P3	1	1	1	1
P4	0	1	0	1
P5	1	0	0	1
P6	0	1	1	1
P7	1	0	0	1

$$T = k\,(k-1)\,\frac{\sum_{j=1}^{k}\left(X_{\bullet j} - \frac{N}{k}\right)^2}{\sum_{i=1}^{b} X_{i\bullet}\,(k - X_{i\bullet})}$$

- k es el número de columnas (tratamientos).
- $X_{\bullet j}$ es el total de la columna para el tratamiento j
- b es el número de bloques (filas).
- $X_{i\bullet}$ es el total de la fila para el bloque j
- N es el total

Tipos de Test: Resumen

Datos categóricos

Independencia	Se utiliza para mostrar si *dos* conjuntos de datos son *independientes* entre sí sobre una misma población.
Homogeneidad	Determina si *dos o más poblaciones* (o grupos) tienen la misma distribución de *una variable categórica*.
McNemar	*Determina si existe una diferencia significativa en proporciones entre datos pareados sobre dos grupos.*
Q de Cochran	*Se utiliza para determinar si existe una diferencia significativa en proporciones entre datos pareados (relacionados) sobre tres o más grupos.*

Manos a la obra