No todo son números. Medición y visualización de la asociación entre variables cualitativas y modelos logarítmico-lineales.

Miguel Ángel Tarancón Morán & Consolación Quintana Rojo

Área de Estadística Económica y Empresarial

Departamento de Economía Aplicada 1

Universidad de Castilla – La Mancha

Contenidos.

1. Introducción.
2. Tablas de contingencia y asociación.
3. Modelos log-lineales.

Introducción.

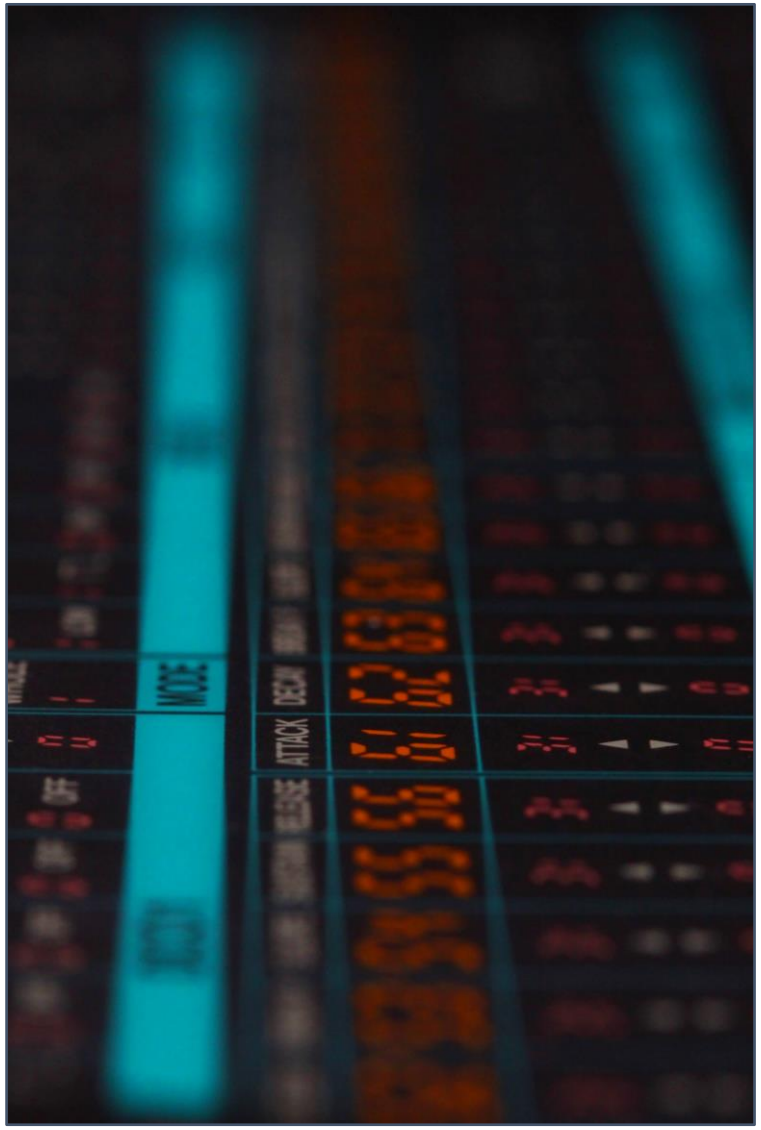
- En numerosas ocasiones la información con que el analista debe enfrentarse es de **naturaleza cualitativa**, esto es, la información se recoge en características no numéricas o atributos (o factores o variables categóricas o cualitativas).

- Cuando ocurre esto, es necesario recurrir a técnicas de explotación específicas para este tipo de datos.

- Uno de los aspectos más interesantes a tratar es el de la relación estadística o **asociación** entre atributos o factores.

- **Asociación:** existe cuando se da que el hecho de que los casos adopten ciertos niveles o categorías en unos factores hace que tiendan a tomar ciertos niveles o categorías de otro u otros factores.

Tablas de contingencia y asociación.



- La información de partida sobre los casos (categorías que adoptan en cada atributo o factor) viene sintetizada en **tablas de contingencia**.

- Por ejemplo, una tabla de contingencia que recoge las frecuencias conjuntas correspondientes a los atributos "situación laboral" y "situación académica":

	Aprobados	No Aprobados	ni.
Trabajan	23	43	66
No Trabajan	67	31	98
n.j	90	74	164

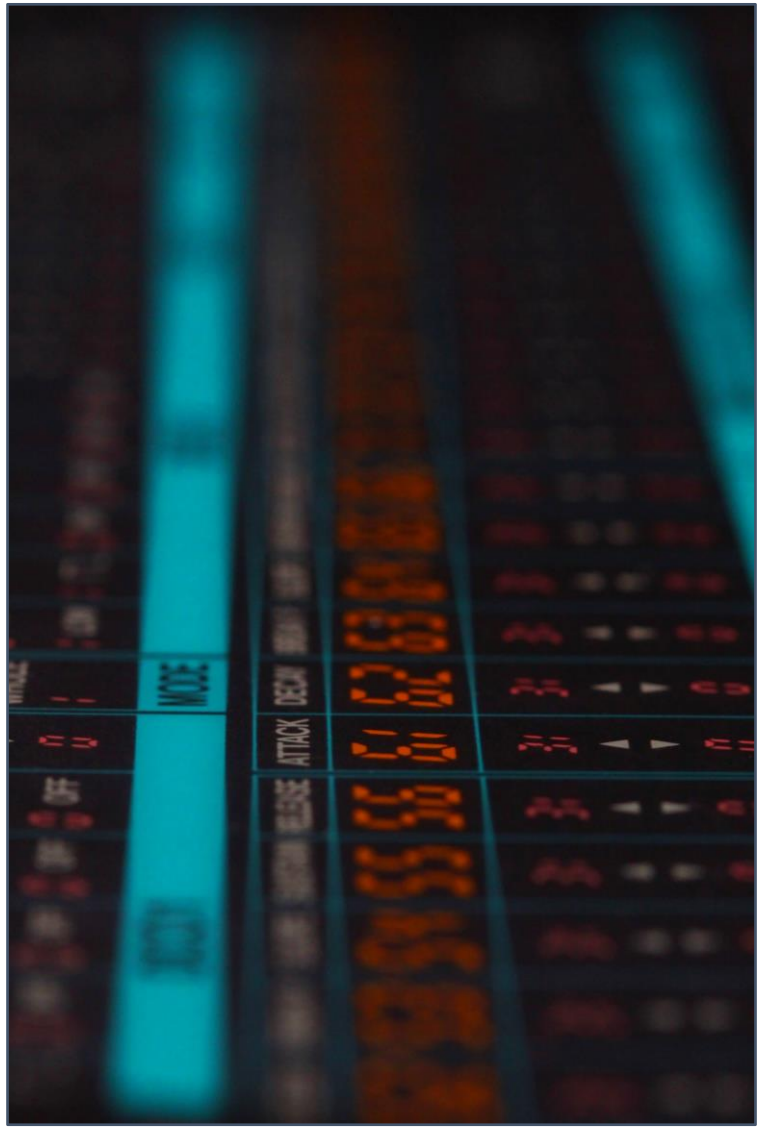
$\mathbf{n_{ij}}$: frecuencias conjuntas (número de casos para cada combinación de niveles de cada atributo)

$\mathbf{n_{.j}}$: frecuencias marginales del atributo "situación académica" (número de casos para cada nivel del atributo)

N: frecuencia total (número total de casos)

$\mathbf{n_{i.}}$: frecuencias marginales del atributo "situación laboral" (número de casos para cada nivel del atributo)

Tablas de contingencia y asociación.



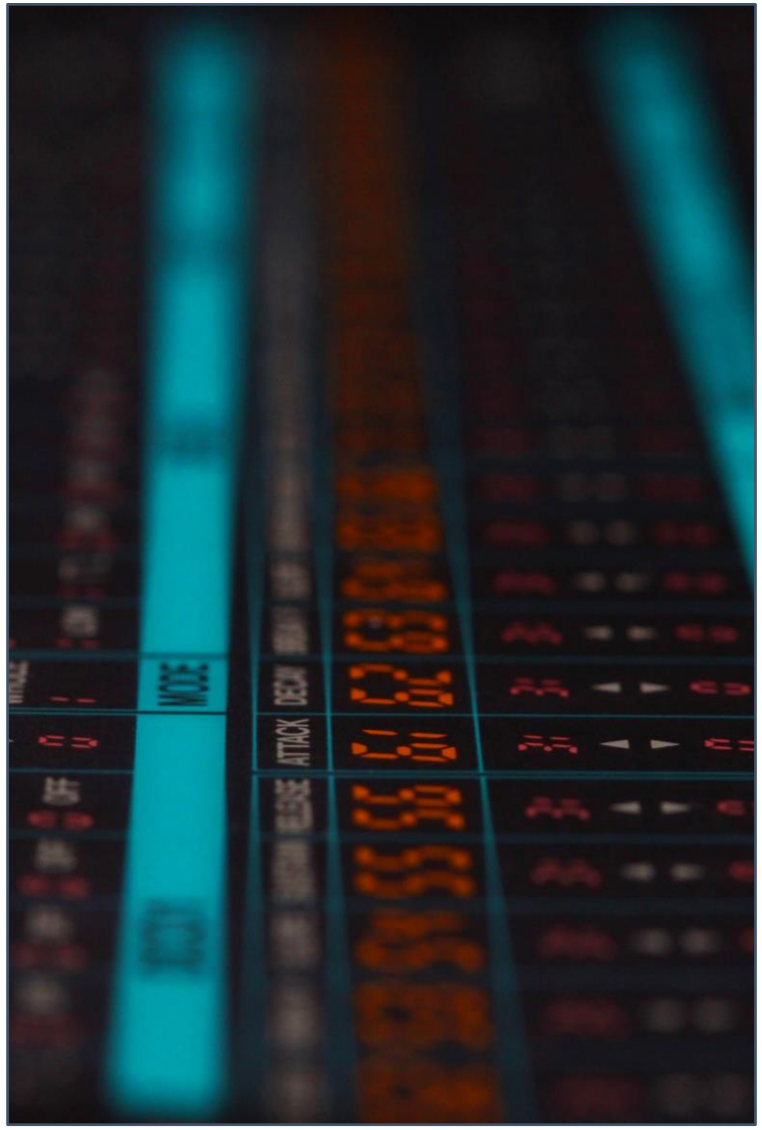
- En el caso de atributos, la condición de independencia (no asociación) es que la frecuencia relativa conjunta coincida con el producto de las correspondientes frecuencias marginales para todas las frecuencias de la tabla de contingencia:

$$f_{ij} = \frac{n_{ij}}{N} = \frac{n_{i\bullet}}{N} \cdot \frac{n_{\bullet j}}{N} = f_{i\bullet} \cdot f_{\bullet j} \quad \forall i, j$$

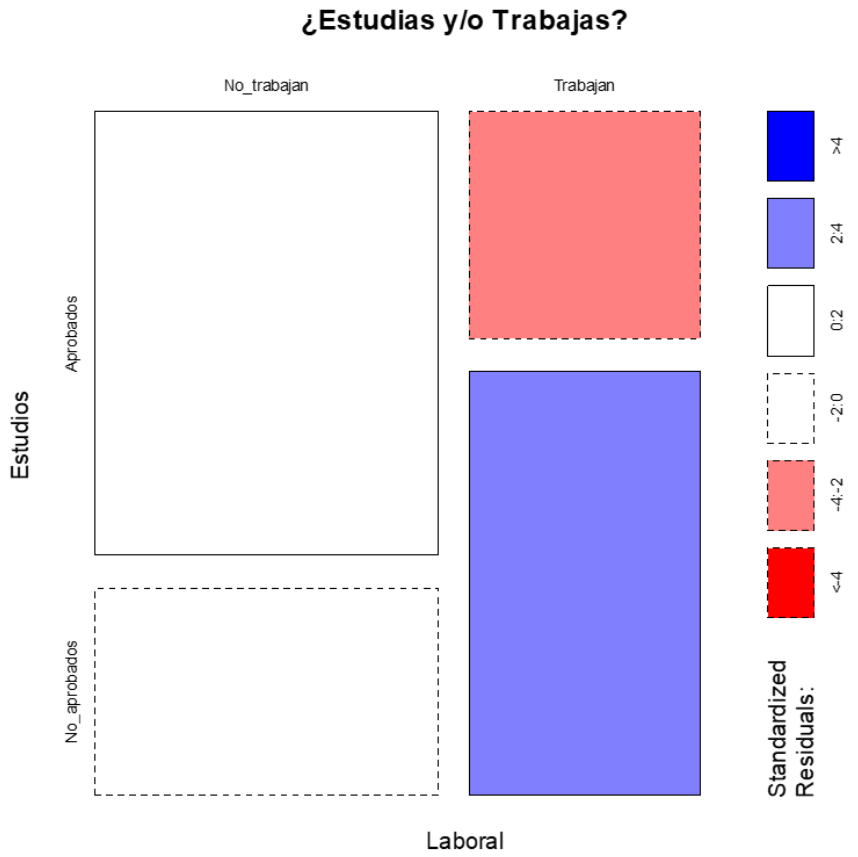
- Operando, las frecuencias esperadas (teóricas) en caso de independencia serían:

$$\frac{n_{ij}}{N} = \frac{n_{i\bullet}}{N} \cdot \frac{n_{\bullet j}}{N} \Rightarrow E_{ij} = n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{N}$$

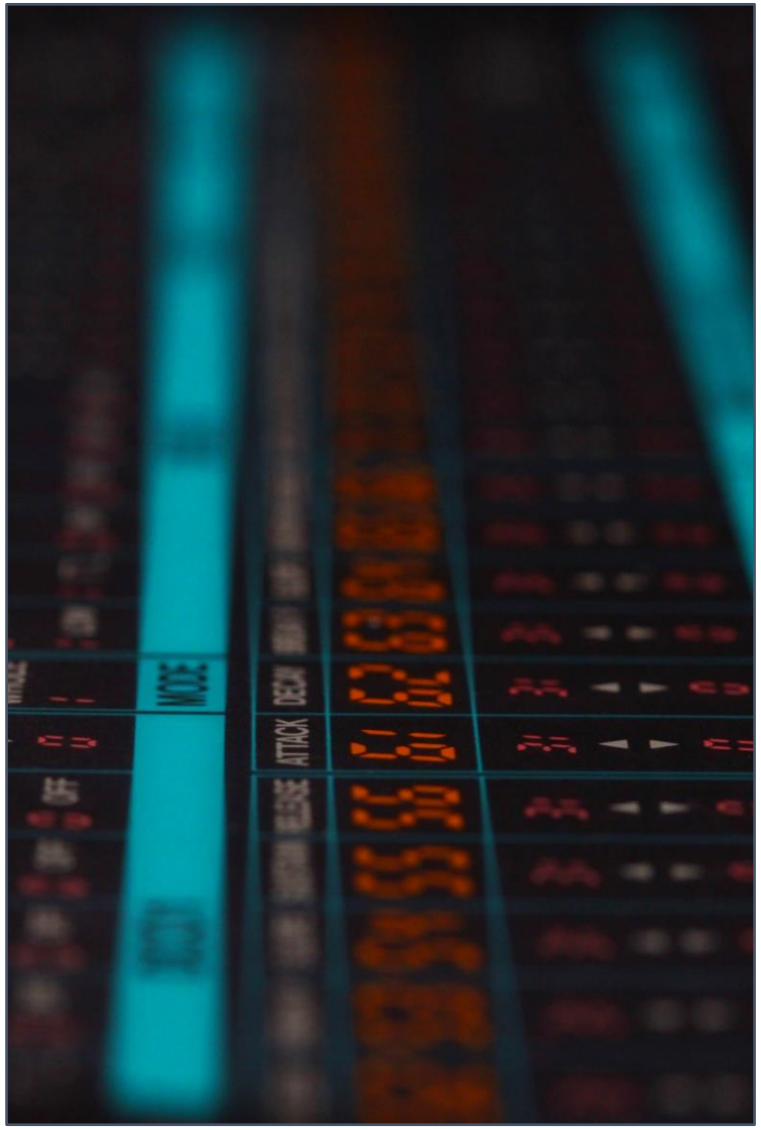
Tablas de contingencia y asociación.



- En nuestro ejemplo:



El color rojo/salmón indica frecuencias observadas menores a las esperadas en caso de independencia. El color azul indica frecuencias observadas mayores a las esperadas en caso de independencia.



- Para contrastar estadísticamente la existencia de independencia entre dos atributos o factores (o, por el contrario, admitir asociación), existe la prueba basada en el **coeficiente de contingencia ji-cuadrado**.

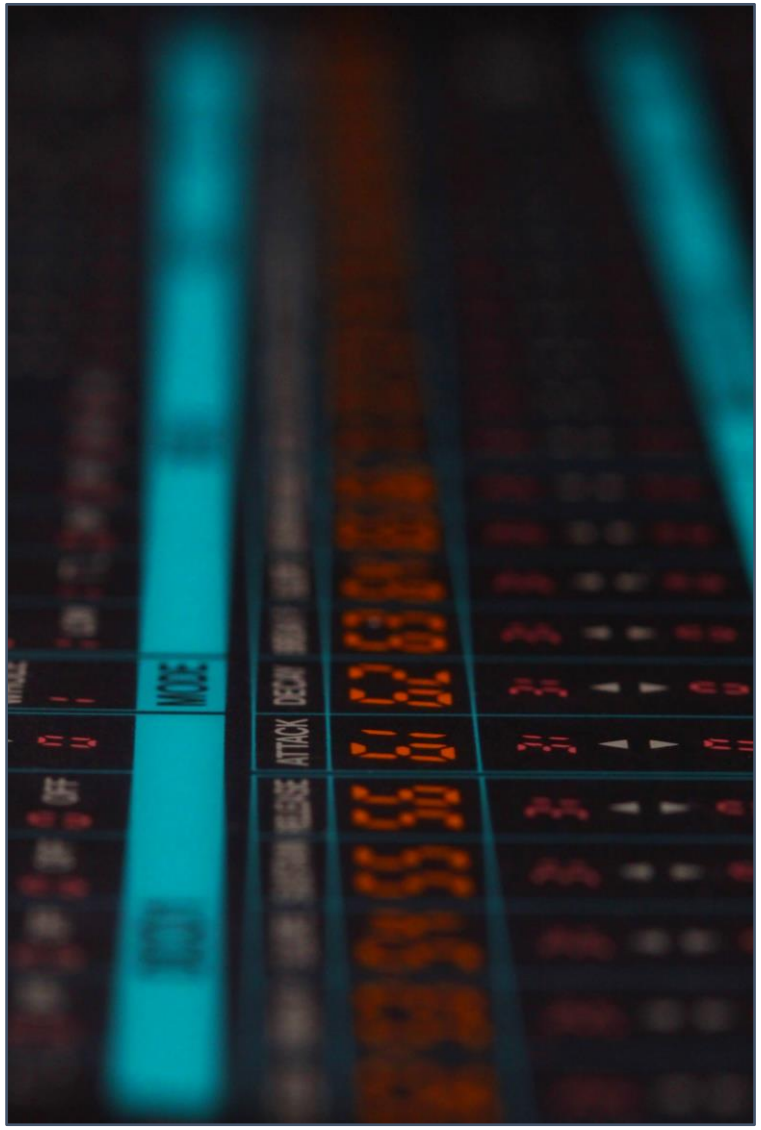
- El coeficiente es una medida de la diferencia existente entre la tabla de contingencia observada y la que debería de darse en caso de independencia (teórica):

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}}$$

- Bajo H0 (independencia) , el coeficiente se distribuye:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}} \rightarrow \chi^2_{(r-1)(c-1)}$$

con: nij, frecuencias observadas; Eij, frecuencias teóricas; r, número de filas; c



- En nuestro ejemplo:

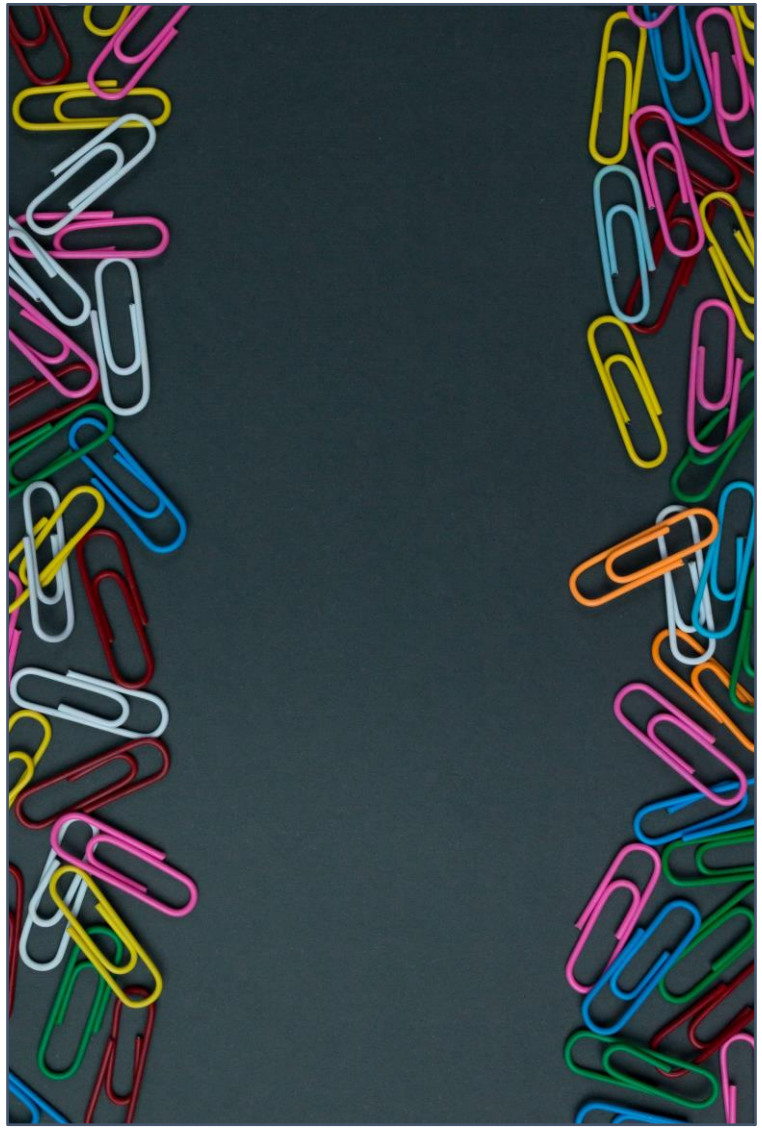
$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = 17{,}8944$$

$$\chi^2_{(r-1)(c-1)} = \chi^2_{(2-1)(2-1)} = \chi^2_1$$

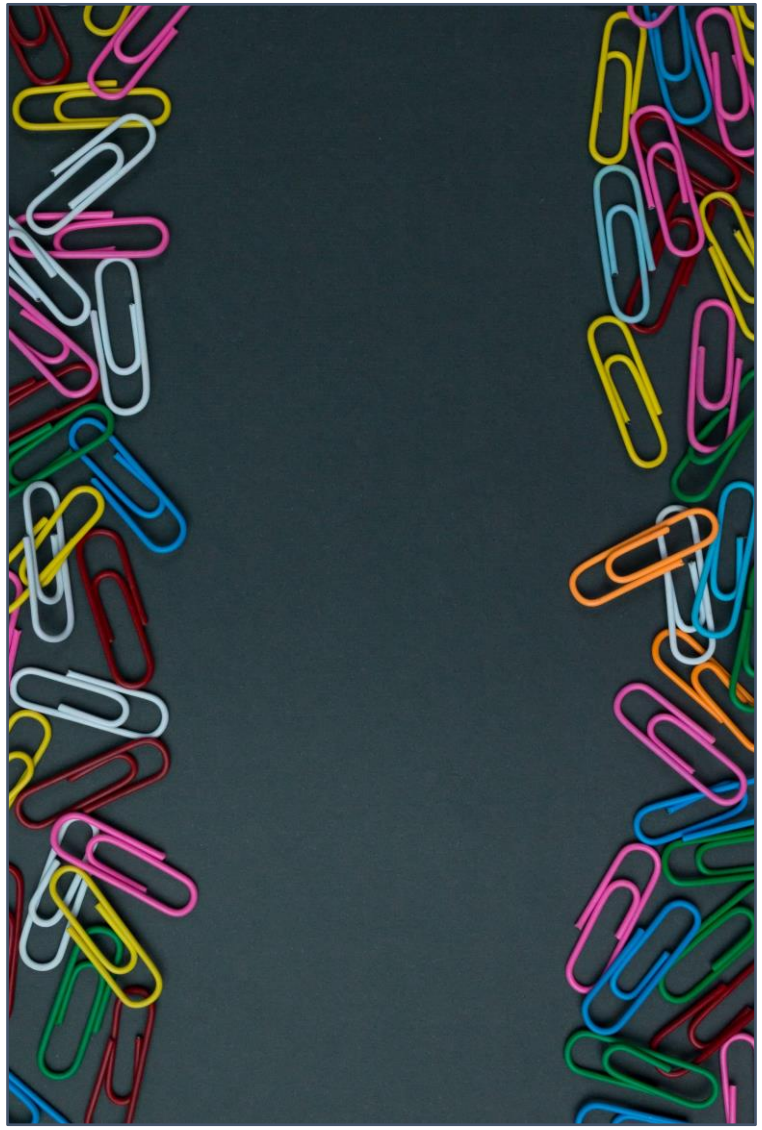
P-valor = 0,00002335

Luego **se rechaza H0** (independencia) para una significación de 0,05, y **se admite asociación** entre los atributos o factores "situación laboral" y situación académica.

Modelos log-lineales.



- Los **modelos log-lineales** intentan resolver cuestiones como:
  - ¿Cuál es la **influencia individual** que cada factor o atributo ejerce a través de cada nivel o categoría sobre la distribución de frecuencias?
  - Los factores, ¿son independientes o existe algún grado de **asociación** entre ellos?
  - De existir asociación, ¿cuál es la influencia o **efecto conjunto** de los distintos factores sobre las frecuencias observadas?



- Vamos a plantear por simplicidad el caso bidimensional con dos atributos A y B (tabla de contingencia rxc)

- Si los dos atributos o factores son **independientes**, acabamos de ver que sus frecuencias observadas deberían coincidir con las teóricas o esperadas:

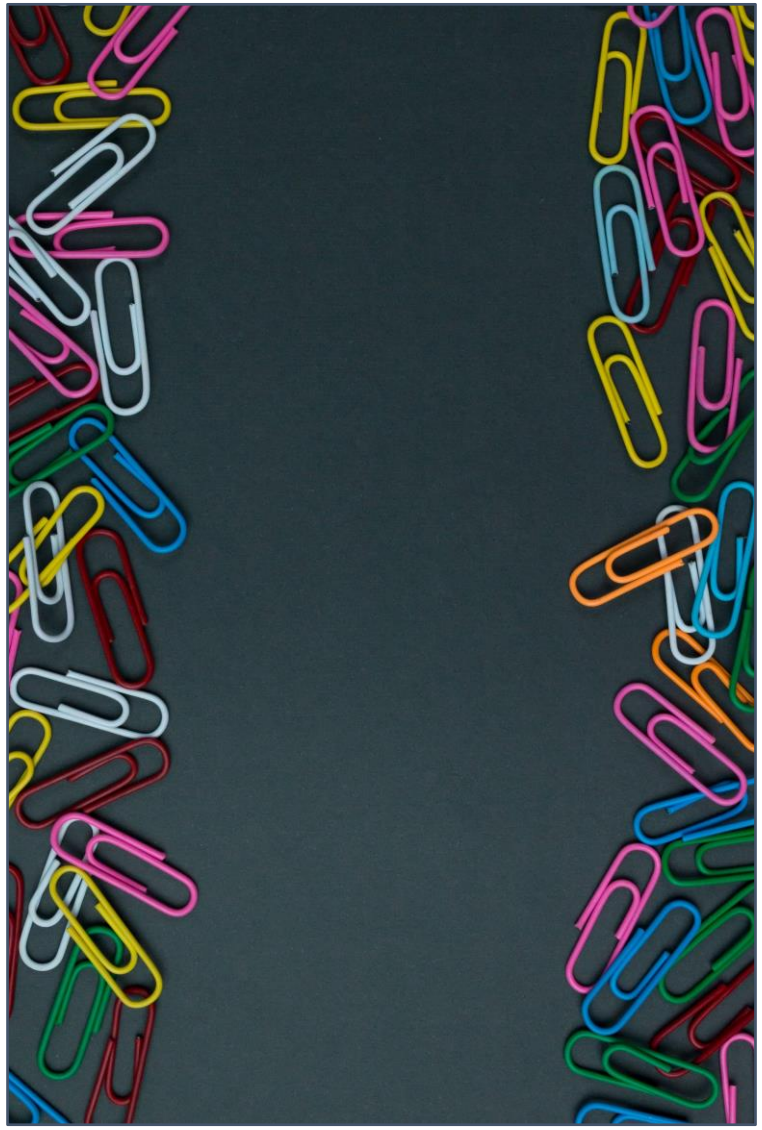
$$n_{ij} = E_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{N} = \frac{N \cdot n_{i\bullet} \cdot n_{\bullet j}}{N \cdot N} = N \cdot f_{i\bullet} \cdot f_{\bullet j}$$

- Tomando logaritmos:

$$\ln n_{ij} = \ln N + \ln f_{i\bullet} + \ln f_{\bullet j}$$

- Renombrando términos:

$$\ln n_{ij} = \lambda + \lambda_i^A + \lambda_j^B$$



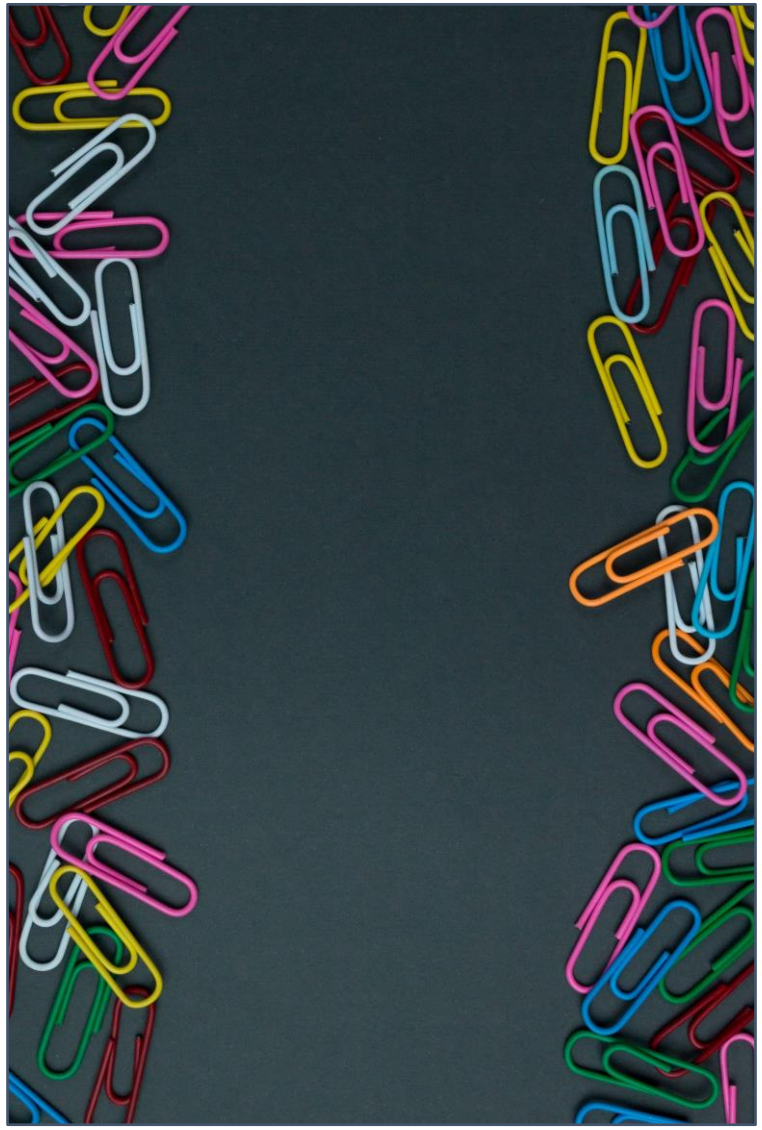
- **Si no existe independencia** entre ambas variables o factores, tendremos:

$$\ln n_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \text{ con } \lambda_{ij}^{AB} \neq 0$$

- Los términos λ_i^A y λ_{ij}^B se llaman **efectos directos o principales**.

- El término λ_{ij}^{AB} recoge el **efecto conjunto o interacción entre ambos factores** sobre el valor de la frecuencia conjunta correspondiente al nivel i del primer factor A y al nivel j del segundo factor B. **Bajo la hipótesis de independencia, ese término tomaría valor 0**.

Modelos log-lineales.



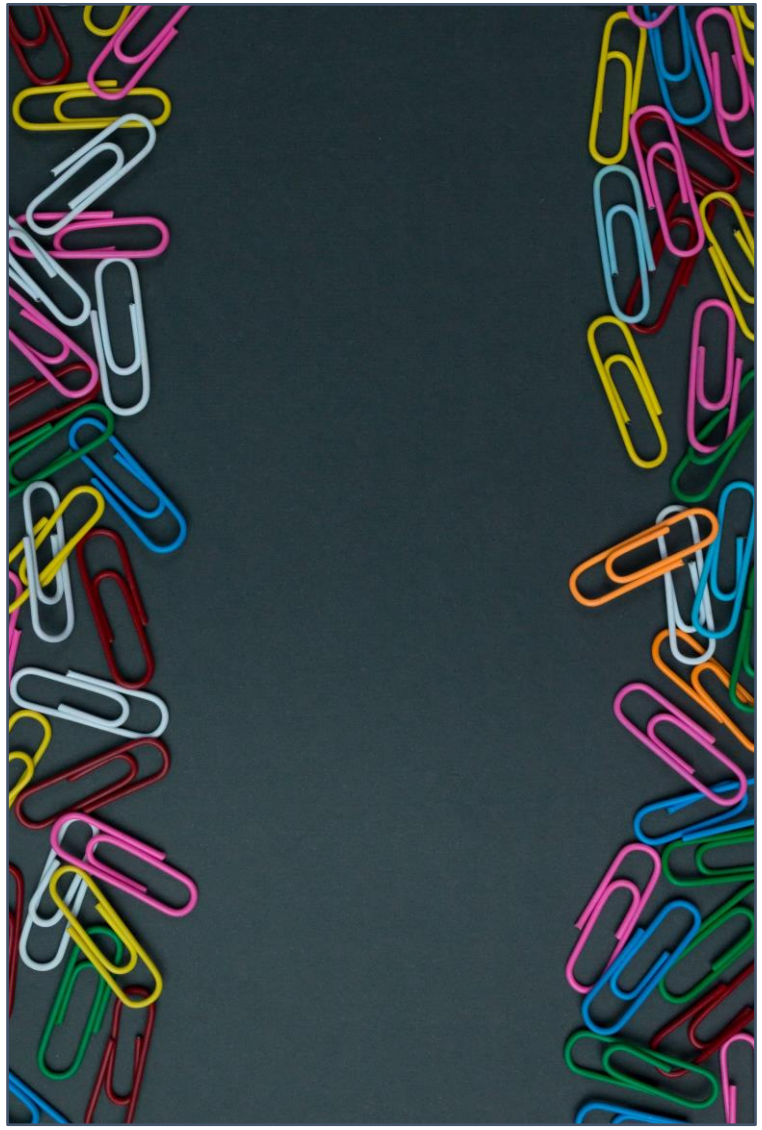
- Si en el modelo se especifican **solo los efectos directos**, se dirá que el modelo log-lineal es **de independencia**.

$$\ln n_{ij} = \lambda + \lambda_i^A + \lambda_j^B$$

- Si en el modelo se especifican **los efectos directos y todos los efectos conjuntos posibles**, se dirá que el modelo es **saturado**.

- El modelo saturado otorga un ajuste perfecto; pero es poco útil a la hora de extraer conclusiones relevantes. Se requiere un modelo que, aunque no ajuste al 100% las frecuencias, recoja solo los **efectos más importantes**.

- **Estimación** de modelos: máxima verosimilitud.



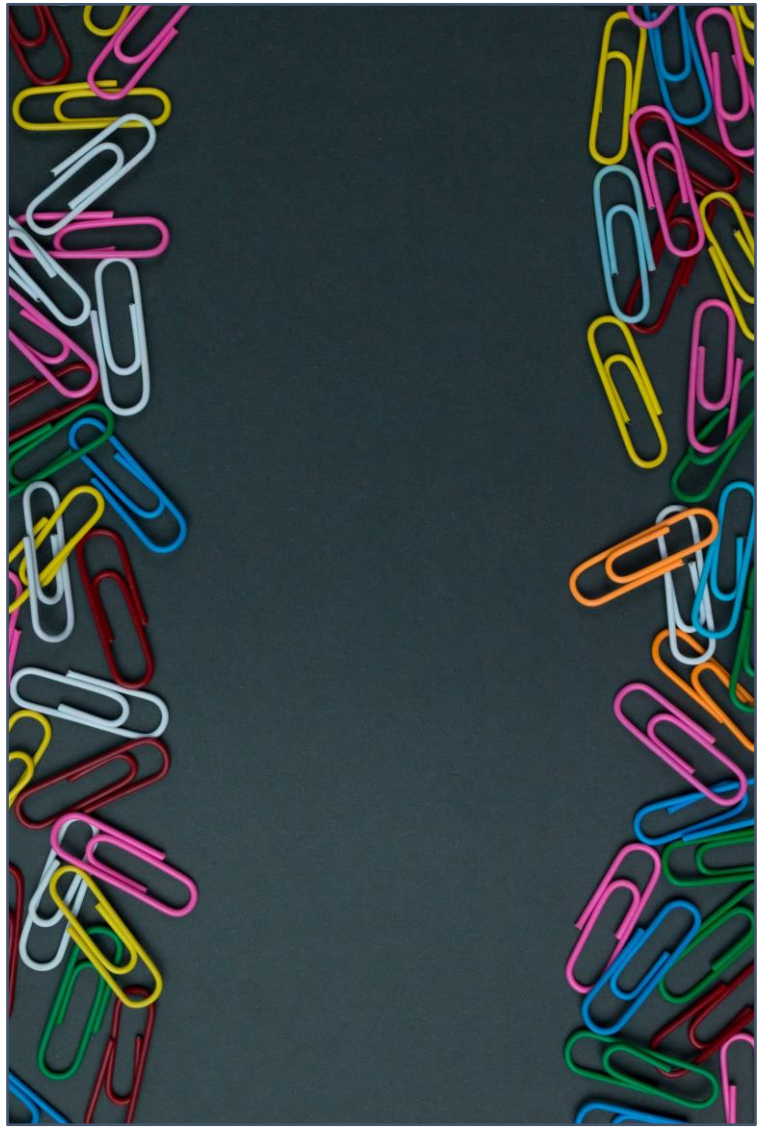
- **Ejemplo:** estimación del modelo de independencia del caso situación laboral / situación académica.

$$\ln n_{ij} = \lambda + \lambda_i^A + \lambda_j^B$$

```
$`(Intercept)`
[1] 3.689382

$Laboral
No_trabajan    Trabajan
  0.1976564  -0.1976564

$Estudios
   Aprobados No_aprobados
  0.09787229  -0.09787229
```

- Por ejemplo, la estimación del logaritmo neperiano de la frecuencia de los casos de no-trabajo y aprobado será:

$$\widehat{lnn_{(no-trabajan,aprobados)}} = 3{,}6894 + 0{,}1977 + 0{,}0979 = 3{,}985 \Rightarrow e^{3{,}985} = 53{,}7853$$: frecuencia estimada



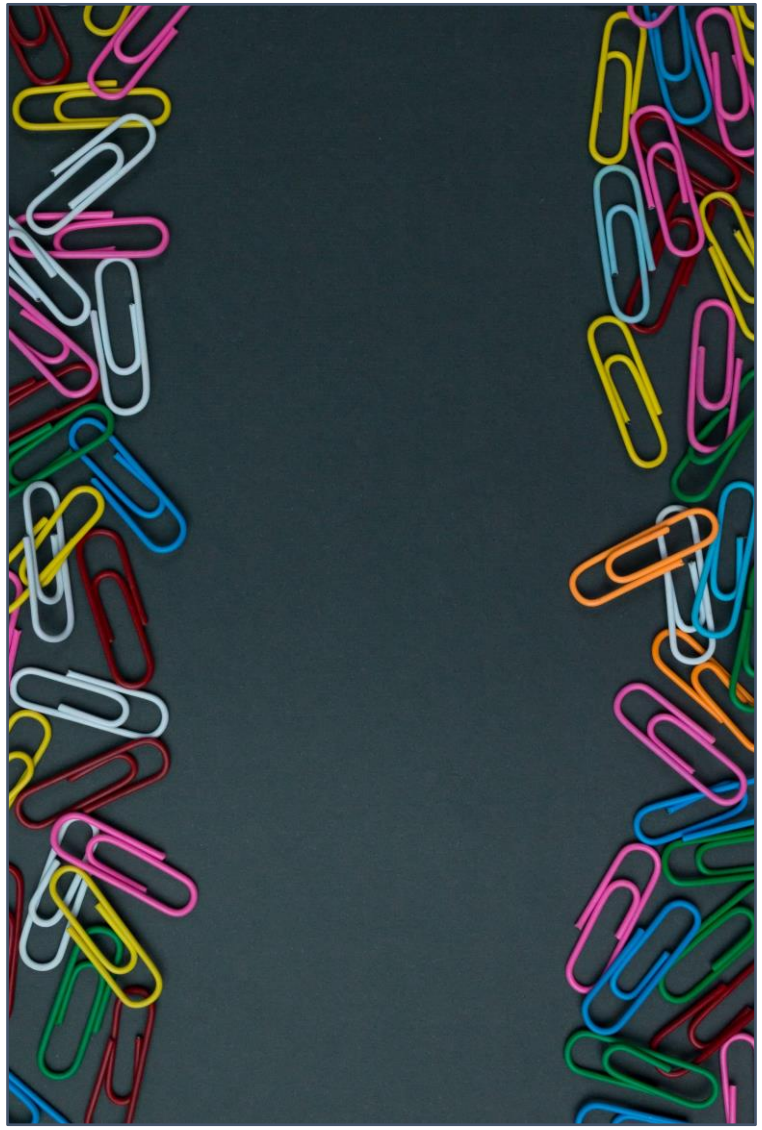
- **Ejemplo:** estimación del modelo saturado del caso situación laboral / situación académica.

$$\ln n_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

```
$`(Intercept)`
[1] 3.633844

$Laboral
No_trabajan    Trabajan
  0.1854964  -0.1854964

$Estudios
   Aprobados No_aprobados
  0.03624988  -0.03624988

$Laboral.Estudios
             Estudios
Laboral        Aprobados No_aprobados
  No_trabajan  0.3491028   -0.3491028
  Trabajan    -0.3491028    0.3491028
```

- La estimación del logaritmo neperiano de la frecuencia de los casos de no-trabajo y aprobado será:

$$\widehat{lnn_{(no-trabajan,aprobados)}} = 3{,}6338 + 0{,}1855 + 0{,}0363 + \\ + 0{,}3491 = 4{,}2047 \Rightarrow e^{3{,}985} = 67{,}0005$$: frecuencia estimada

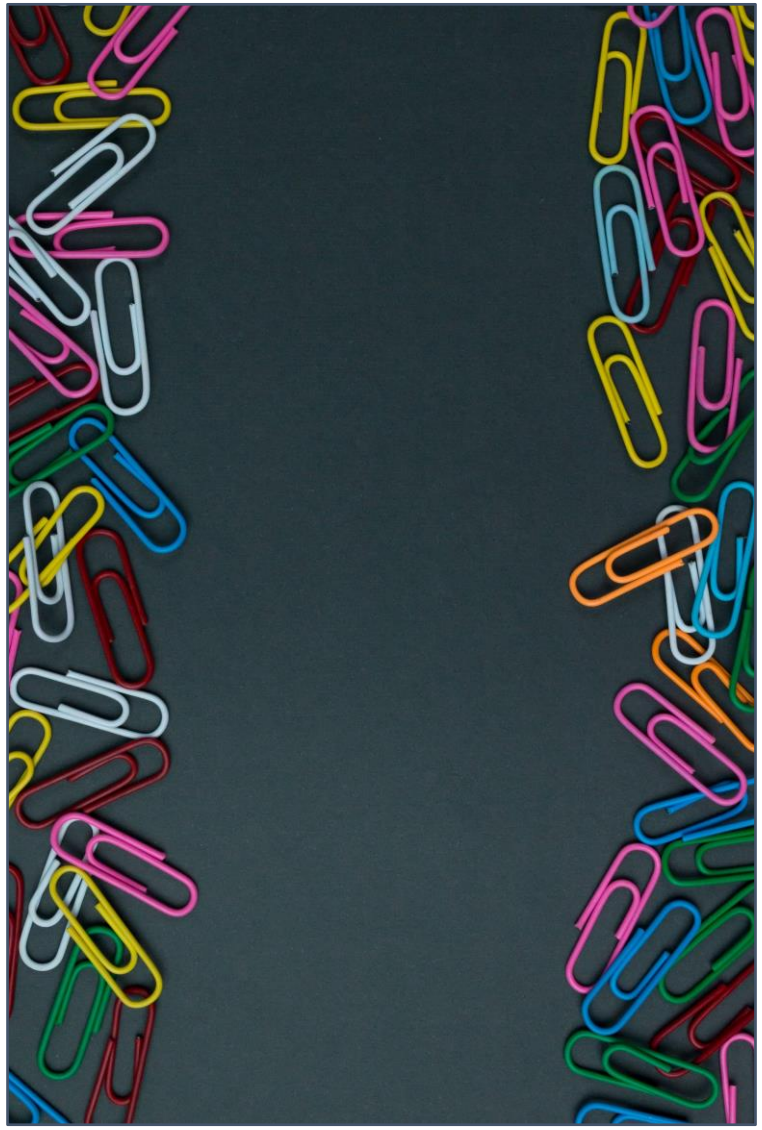


- ¿Cómo saber si es **aceptable** una especificación?
- Una opción es el estudio del **ratio de verosimilitud** (likelihood ratio), también denominado ***deviance***:

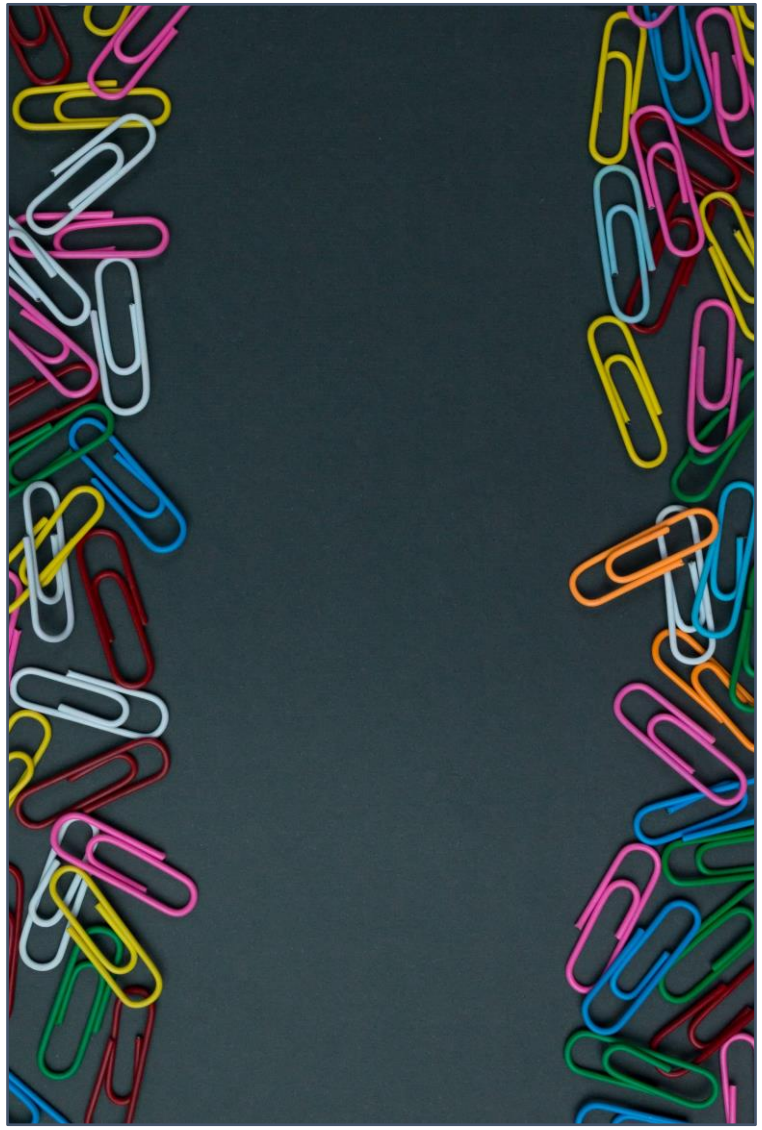
$$G^2 = 2 \cdot \sum_{ij} f_{ij} \cdot \ln\left(\frac{f_{ij}}{f_{ij}^*}\right)$$

f_{ij} son las frecuencias (relativas) conjuntas observadas, y f_{ij}^* son las estimadas por el modelo.

- G^2 se distribuye, bajo la H_0 de frecuencias iguales (buen ajuste), como una χ^2 cuyos grados de libertad dependen del número de parámetros de la especificación.
- Así, para que sea aceptable una especificación (buen ajuste, debe **no-rechazarse** H0, es decir, obtener un p-valor superior a 0.05 (valor de G^2 bajo).



- En el ejemplo anterior, los resultados de G^2 son:
  - Modelo de independencia: 18,13 (p-valor de 2.06*E-05)
  - Modelo saturado: 0 (p-valor de 1)
- El modelo saturado tiene un G^2 de 0, y un p-valor de 1, ya que predice las frecuencias observadas perfectamente. Pero no suele ser útil para tablas de contingencia más complejas, ya que lo que se busca es un modelo que recoja solo las relaciones entre atributos y frecuencias más importantes.
- El p-valor del modelo de independencia es muy bajo, luego se rechaza H0. Esto significa que **se rechaza que exista independencia entre los atributos**, y se admite asociación entre los atributos o factores.



- Si existieran **tres factores** A, B, C, el **modelo de independencia** sería:

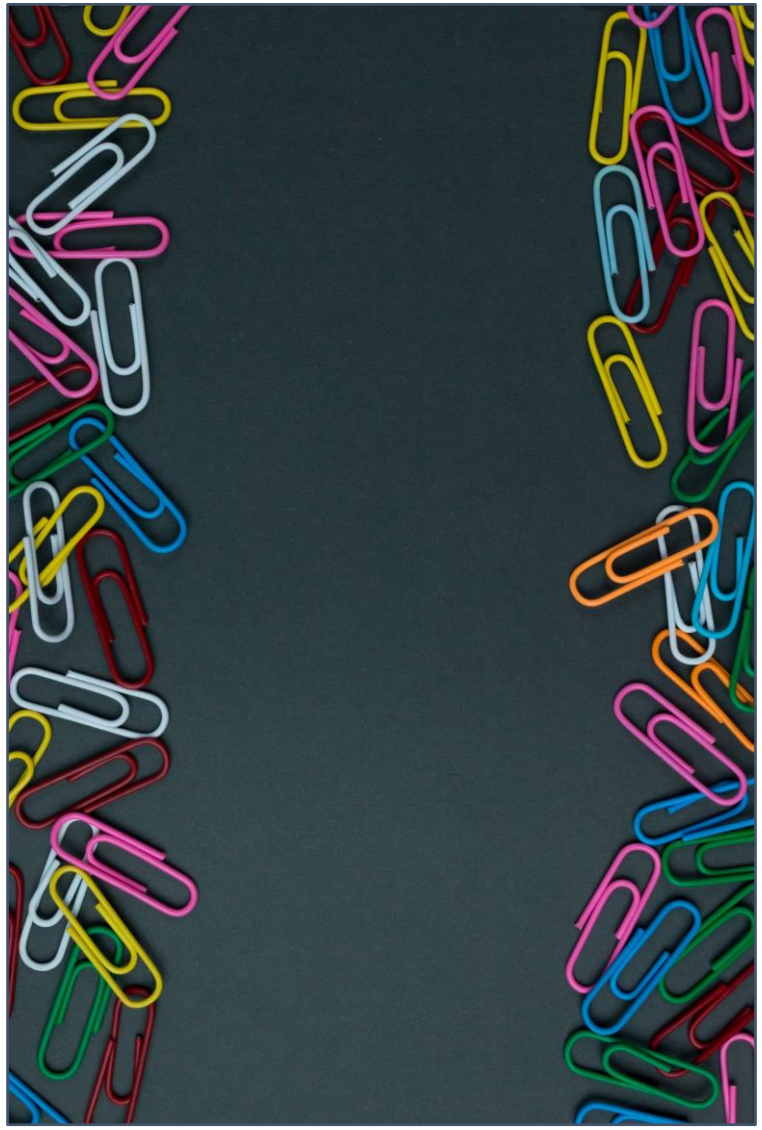
$$\ln n_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

- El modelo **saturado:**

$$\ln n_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC} + \lambda_{ik}^{AC} + \lambda_{ijk}^{ABC}$$

- Entre ambos modelos existe un abanico de posibles especificaciones en el que aparecen algunos efectos conjuntos; pero no todos.

Modelos log-lineales.



- Existen **procesos iterativos** para la selección de la especificación más apropiada, basados en el **AIC** (*criterio de información de Akaike*):

$$AIC = 2 \cdot k - 2 \cdot \ln(L)$$

con k=nº de parámetros estimados y ln(L) logaritmo de la función de verosimilitud de la muestra.

- El proceso más habitual parte del modelo saturado y va probando la eliminación de términos que hagan menor el valor de AIC, manteniendo la **"regla de jerarquía"**: por ejemplo, no puede aparecer una interacción entre el factor A y B si no están especificados los términos inferiores, esto es, los efectos independientes tanto de A como de B.

¡Muchas gracias!