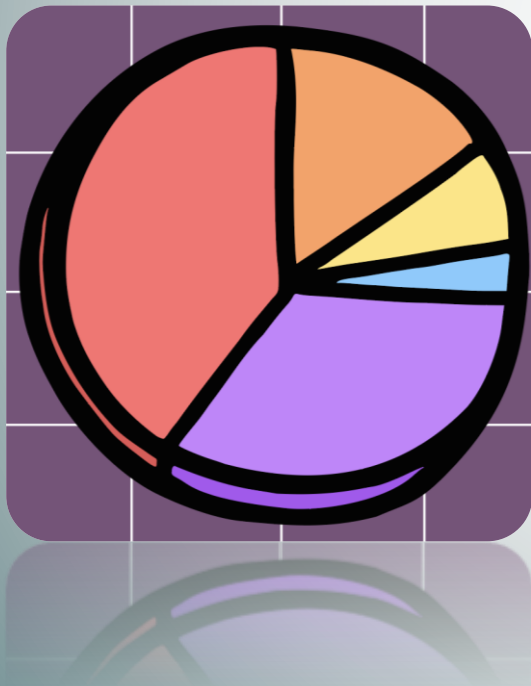


Estadística I

TEMA 3. ANÁLISIS DE ATRIBUTOS



Marta Guijarro Garvi

DEPARTAMENTO DE ECONOMÍA

Este material se publica bajo licencia:

[Creative Commons BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)



Tema 3: Análisis de atributos

Índice de contenidos

- Concepto de atributo.
- Distribución de frecuencias unidimensional.
- Distribución de frecuencias bidimensional: distribuciones marginales y condicionadas.
- Independencia entre atributos y tipo de asociación entre modalidades.
- Tablas 2 x 2: coeficiente de asociación H .
- Medidas de asociación: coeficiente ji-cuadrado de Pearson y coeficiente de rangos de Spearman.

Tema 3: Análisis de atributos

Resultados de aprendizaje

- Comprender el concepto de atributo.
- Entender los conceptos de distribuciones de frecuencias unidimensional y bidimensional.
- Saber obtener distribuciones marginales y condicionadas.
- Establecer la diferencia entre asociación entre dos atributos y tipo de asociación entre sus modalidades.
- Saber estudiar la asociación en tablas de contingencia de dimensión 2 x 2.

Concepto de atributo

Un **atributo** es cualquier característica cualitativa (no numérica) asociada a las unidades de una población. Las observaciones distintas de un atributo A son sus **modalidades** y se denotan por $A_1, \dots, A_h$.

Ejemplo 3.1 Una empresa de servicios informáticos cuenta con 200 personas trabajadoras. De ellas, 100 trabajan en desarrollo de software, 20 en soporte técnico, 50 en administración y las 30 restantes forman parte del equipo directivo. La población está constituida por las 150 personas de la empresa sobre quienes se analiza la característica relativa al departamento en el que realizan sus funciones. Esta característica no es numérica sino cualitativa. Se trata, por tanto, de un atributo cuyas modalidades son *desarrollo de software, soporte técnico, administración y dirección*.

Distribución de frecuencias unidimensional

La **frecuencia absoluta** n_i de la modalidad A_i de un atributo A es el número de observaciones iguales a dicha modalidad.

Si N es el número de observaciones y h el número de modalidades del atributo, se cumple:

$$\sum_{i=1}^h n_i = N$$

La **frecuencia relativa** f_i de la modalidad A_i de un atributo A es la proporción de observaciones iguales a dicha modalidad.

Se cumple:

$$f_i = \frac{n_i}{N}$$
$$\sum_{i=1}^h f_i = 1$$

Distribución de frecuencias unidimensional

La **distribución de frecuencias** del atributo A es el conjunto de sus modalidades y de sus respectivas frecuencias absolutas o relativas. Se denota por $(A_i; n_i)$ o por $(A_i; f_i)$ según el tipo de frecuencia utilizado.

Cuando todas las frecuencias absolutas son iguales a la unidad, la distribución de frecuencias de un atributo es una distribución de frecuencias *unitaria*.

Distribución de frecuencias unidimensional

Ejemplo 3.2 (continuación del Ejemplo 3.1). La distribución de frecuencias del atributo se compone de sus diferentes modalidades (primera columna) y de las frecuencias de cada una de ellas. En este caso, el enunciado aporta las frecuencias absolutas de cada modalidad (segunda columna).

Secciones	Número de trabajadores	Proporción de trabajadores
Desarrollo de software	100	0,50
Soporte técnico	20	0,10
Administración	50	0,25
Dirección	30	0,15

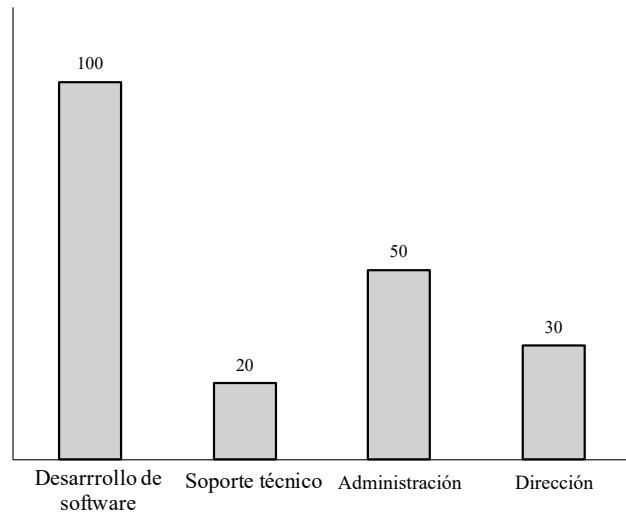
Así, por ejemplo, como 50 personas trabajadoras de la empresa pertenecen a la sección de administración, entonces, $n_3 = 50$.

En la tercera columna se muestra el peso de cada sector en el conjunto de la empresa, es decir, la frecuencia relativa de cada modalidad del atributo. Los valores de esta columna se obtienen como cociente entre la frecuencia absoluta y el número total de datos. Por ejemplo, $f_4 = 30/200 = 0,15$, con lo cual, el 15 por ciento de las personas trabajadoras de la empresa son personal de dirección.

Distribución de frecuencias unidimensional

Las representaciones gráficas más habituales de las distribuciones de frecuencias unidimensionales son el **diagrama de barras** y el **diagrama de sectores**. En el primero, cada modalidad del atributo se coloca en el eje horizontal y se le asocia una barra cuya altura refleja su frecuencia absoluta o relativa. El segundo es en un círculo dividido en sectores, cuyas áreas son proporcionales a las frecuencias absolutas o relativas.

Ejemplo 3.3 (continuación del Ejemplo 3.1). Las representaciones del atributo del ejemplo son:



Distribución de frecuencias unidimensional

Como un atributo es una característica no numérica, la **moda**, modalidad más frecuente, es la única medida de resumen de la información de la distribución de frecuencias unidimensional de un atributo.

No obstante, cuando las modalidades pueden ordenarse según la intensidad de la característica, también es posible calcular la **mediana**, modalidad que tiene el mismo número de observaciones “mayores” y “menores” que ella.

Ejemplo 3.4 (continuación del Ejemplo 3.1). La moda es *desarrollo de software*, ya que es la modalidad con mayor frecuencia.

Como este atributo no admite una ordenación de sus modalidades, no es posible calcular la mediana.

Distribución de frecuencias bidimensional

La observación conjunta de dos atributos A y B en las unidades de una población conduce a pares de datos cuyas componentes son cualitativas. La frecuencia absoluta o **frecuencia absoluta conjunta** n_{ij} de la modalidad (A_i, B_j) es el número de observaciones iguales a dicho par. La frecuencia relativa o **frecuencia relativa conjunta** f_{ij} es la proporción de observaciones iguales a dicho par.

Se cumple:
$$\sum_{i=1}^h \sum_{j=1}^k n_{ij} = N$$

La distribución de frecuencias de (A, B) o **distribución de frecuencias bidimensional** es el conjunto de pares de modalidades junto con sus frecuencias. Cuando las frecuencias son absolutas se denota por $(A_i, B_j; n_{ij})$ y por $(A_i, B_j; f_{ij})$ cuando son relativas.

Se cumple: $f_{ij} = \frac{n_{ij}}{N}$ y, por tanto,
$$\sum_{i=1}^h \sum_{j=1}^k f_{ij} = 1$$

Distribución de frecuencias bidimensional

En el caso de atributos, la distribución de frecuencias bidimensional se presenta en una tabla de doble entrada llamada **tabla de contingencia** que, además de las modalidades de cada atributo, contiene las frecuencias conjuntas (absolutas o relativas):

A	B	B_1	$\dots$	B_k
A_1		n_{11}	$\dots$	n_{1k}
$\vdots$		$\vdots$		$\vdots$
A_h		n_{h1}	$\dots$	n_{hk}

A	B	B_1	$\dots$	B_k
A_1		f_{11}	$\dots$	f_{1k}
$\vdots$		$\vdots$		$\vdots$
A_h		f_{h1}	$\dots$	f_{hk}

Distribución de frecuencias bidimensional

Ejemplo 3.5 Tras finalizar la temporada de verano, la asociación de heladeros artesanales de una región realiza un estudio sobre los hábitos de consumo de helado. La tabla muestra los litros de helado, en millones, vendidos en la última temporada estival, clasificados en tres variedades (vainilla, chocolate y fresa) y diferenciando las ventas por zona costera y zona interior:

Zona	Interior	Costera
Variedad de helado		
Vainilla	1000	2000
Chocolate	600	500
Fresa	400	1500

La tabla de contingencia recoge la distribución de frecuencias bidimensional de estos dos atributos, siendo la población los litros de helado. El primer atributo, A , corresponde a la variedad de helado, con tres modalidades: *vainilla*, *chocolate* y *fresa*. El segundo, B , hace referencia a la zona de venta, con dos modalidades: *zona costera* y *zona interior*. En las casillas de la tabla aparecen las frecuencias absolutas conjuntas, Así, por ejemplo, el número de litros de helado de fresa y vendidos en la zona interior es millones. $n_{32} = 1500$

Distribución de frecuencias bidimensional: distribuciones marginales

Partiendo de una distribución de frecuencias bidimensional se obtienen las **distribuciones marginales** de A y B , esto es, $(A_i; n_{i.})$ y $(B_j; n_{.j})$ cuando las frecuencias son absolutas y $(A_i; f_{i.})$ y $(B_j; f_{.j})$ cuando son relativas.

Cuando las frecuencias son absolutas, las correspondientes frecuencias marginales genéricas son

$$n_{i.} = \sum_{j=1}^k n_{ij} \qquad n_{.j} = \sum_{i=1}^h n_{ij}$$

Y, cuando son relativas, las frecuencias marginales genéricas son

$$f_{i.} = \sum_{j=1}^k f_{ij} \qquad f_{.j} = \sum_{i=1}^h f_{ij}$$

Además,

$$f_{i.} = n_{i.} / N \qquad f_{.j} = n_{.j} / N$$

Distribución de frecuencias bidimensional: distribuciones marginales

Ejemplo 3.6 (continuación del Ejemplo 3.5) A partir de la tabla del Ejemplo 3.5 se calculan las frecuencias marginales. Así, sumando cada valor de la primera columna con el correspondiente de la segunda, se obtienen las frecuencias del atributo A , variedad de helado (segunda columna). Por ejemplo, los litros de helado de chocolate vendidos en la última temporada, $n_{2.} = 1100$, se calculan sumando los vendidos en la zona costera y en la zona interior, $n_{21} + n_{22} = 600 + 500$.

Análogamente, sumando cada valor de la primera fila con los correspondiente de la segunda y tercera, se obtienen las frecuencias del atributo B , zona de venta. Así, por ejemplo, el número de litros de helados vendidos en la zona costera, $n_{.2} = 4000$, es la suma de los litros de helado de vainilla, chocolate y fresa comercializados en dicha zona $n_{12} + n_{22} + n_{32} = 2000 + 500 + 1500$.

En consecuencia, las distribuciones de frecuencias de los atributos A y B son, respectivamente:

Variedad de helado	Nº litros vendidos $n_{i.}$	$f_{i.}$
Vainilla	3000	0,500
Chocolate	1100	0,183
Fresa	1900	0,317

Zona	Nº litros vendidos $n_{.j}$	$f_{.j}$
Interior	2000	0,33
Costera	4000	0,67

Distribución de frecuencias bidimensional: distribuciones condicionadas

A partir de la distribución de frecuencias bidimensional se obtienen las **distribuciones de frecuencias condicionadas** del atributo A por las modalidades de B , y del atributo B por las modalidades de A .

Distribución de A condicionada por B_j de B

Notación: $(A_i / B_j; n_{i/j})$ o bien $(A_i / B_j; f_{i/j})$

A_i / B_j	$n_{i/j}$	$f_{i/j}$
x_1	$n_{1/j}$	$f_{1/j}$
$\vdots$	$\vdots$	$\vdots$
x_h	$n_{h/j}$	$f_{h/j}$

$$n_{i/j} = n_{ij} \quad f_{i/j} = n_{i/j} / n_{.j} = n_{ij} / n_{.j}$$

Distribución de B condicionada por A_i de A

Notación: $(B_j / A_i; n_{j/i})$ o bien $(B_j / A_i; f_{j/i})$

B_j / A_i	$n_{j/i}$	$f_{j/i}$
y_1	$n_{1/i}$	$f_{1/i}$
$\vdots$	$\vdots$	$\vdots$
y_k	$n_{k/i}$	$f_{k/i}$

$$n_{j/i} = n_{ij} \quad f_{j/i} = n_{j/i} / n_{i.} = n_{ij} / n_{i.}$$

Cuando las frecuencias de la distribución bidimensional son relativas, entonces:

$$f_{i/j} = f_{ij} / f_{.j}$$

$$f_{j/i} = f_{ij} / f_{i.}$$

Distribución de frecuencias bidimensional: distribuciones condicionadas

Ejemplo 3.6 (continuación del Ejemplo 3.5) De la distribución del Ejemplo 3.3 resultan las distribuciones del atributo A condicionadas por cada modalidad de B . Así, la distribución de ventas de helado *dentro* de la zona interior, esto es, $(A_i / B_1; n_{i/1})$, tiene como modalidades, *vainilla*, *chocolate* y *fresa*, y como frecuencias las de la primera columna de la tabla de contingencia, esto es, $n_{i/1} = n_{i1}$ para todo i .

Análogamente, la distribución condicionada $(A_i / B_2; n_{i/2})$, es decir, la distribución de ventas de helado *dentro* de la zona costera tiene las mismas modalidades y como frecuencias las de la segunda columna de la tabla de contingencia, esto es, $n_{i/2} = n_{i2}$ para todo i .

Las distribuciones condicionadas del atributo B por cada modalidad del atributo A , se calculan de igual modo: la distribución de ventas por zonas del helado de vainilla $(B_j / A_1; n_{j/1})$, del helado de chocolate, $(B_j / A_2; n_{j/2})$ y del helado de fresa $(B_j / A_3; n_{j/3})$ se obtienen de las filas de la tabla de contingencia.

Las siguientes tablas presentan estas cuatro distribuciones de frecuencias unidimensionales; se incluyen, además, las frecuencias relativas calculadas de la forma habitual.

Variedad de helado	Nº litros (zona interior) n_{i1}	$f_{i/1}$	Nº litros (zona costera) n_{i2}	$f_{i/2}$
Vainilla	1000	0,5	2000	0,5
Chocolate	600	0,3	500	0,125
Fresa	400	0,2	1500	0,375

Zona	Nº litros (vainilla) n_{1j}	$f_{j/1}$	Nº litros (chocolate) n_{2j}	$f_{j/2}$	Nº litros (fresa) n_{3j}	$f_{j/3}$
Interior	1000	0,33	600	0,55	400	0,21
Costera	2000	0,67	500	0,45	1500	0,79

Independencia entre atributos y tipo de asociación entre modalidades

Sea una distribución de frecuencias bidimensional $(A_i, B_j; f_{ij})$. Los atributos A y B son **independientes**, si se cumple la *condición de independencia* para todos los pares (i, j) , es decir, si para cualesquiera i y j , se cumple que

$$f_{ij} = f_{i\cdot} \cdot f_{\cdot j}$$

o, equivalentemente,

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

La condición de independencia entre las modalidades A_i y B_j también puede expresarse como:

$$f_{i/j} = f_{i\cdot} \quad \text{o bien} \quad f_{j/i} = f_{\cdot j}$$

En consecuencia, la condición necesaria y suficiente para que dos atributos sean independientes es que las distribuciones condicionadas coincidan con la correspondiente distribución marginal.

Independencia entre atributos y tipo de asociación entre modalidades

Cuando los atributos *no* son independientes, se puede hablar de **tipo de asociación** entre sus modalidades. Así, cuando

$$n_{ij} > \frac{n_{i.} \cdot n_{.j}}{N}$$

es decir, cuando la frecuencia absoluta conjunta de las modalidades A_i y B_j es mayor que la esperada en el caso de independencia entre los atributos A y B , entonces existe **asociación positiva** entre las modalidades A_i y B_j .

Recíprocamente, cuando

$$n_{ij} < \frac{n_{i.} \cdot n_{.j}}{N}$$

existe **asociación negativa** entre estas dos modalidades.

Independencia entre atributos y tipo de asociación entre modalidades

Ejemplo 3.7 La dirección de un restaurante sospecha que no existe relación entre el sexo de las personas que acudieron el último mes y el tipo de menú que eligieron. La tabla recoge la distribución de frecuencias de los tres tipos de menú ofrecidos (vegetariano, tradicional y gourmet) y el sexo de los 1000 comensales elegidos. En la última fila y en la última columna se muestran, además, las respectivas distribuciones marginales:

Tipo de menú	Hombres	Mujeres	$n_{i\cdot}$
Vegetariano	20	80	100
Tradicional	14	56	70
Gourmet	6	24	30
$n_{\cdot j}$	40	160	200

Se cumple que: $\frac{20}{40} = \frac{80}{160} = \frac{100}{200}$, esto es, $f_{1/1} = f_{1/2} = f_{1\cdot}$; además, $\frac{14}{40} = \frac{56}{160} = \frac{70}{200}$, es decir, $f_{2/1} = f_{2/2} = f_{2\cdot}$, y por último, $\frac{6}{40} = \frac{24}{160} = \frac{30}{200}$, o lo que es igual, $f_{3/1} = f_{3/2} = f_{3\cdot}$. Por tanto, los atributos son independientes.

Se cumple, asimismo: $f_{1/1} = f_{1/2} = f_{1/3} = f_{1\cdot}$ y $f_{2/1} = f_{2/2} = f_{2/3} = f_{2\cdot}$.

Tablas 2 x 2: coeficiente de asociación H

Cuando la tabla de contingencia es de dimensión 2×2 , resulta suficiente comprobar la condición de independencia con una sola pareja de modalidades. Además, cuando los atributos no son independientes, el estudio del tipo de asociación de cualquier par de modalidades de la tabla determina el tipo de asociación del resto de los pares.

Así, por ejemplo, si se cumple que

$$n_{12} > \frac{n_{1.} \cdot n_{.2}}{N},$$

es decir, si hay asociación positiva entre A_1 y B_2 , también la asociación es positiva entre A_2 y B_1 , y la asociación es negativa entre A_1 y B_1 , y entre A_2 y B_2 .

Tablas 2 x 2: coeficiente de asociación H

El hecho de que en una tabla de contingencia de dimensión 2×2 el estudio de la independencia entre atributos o, en su caso, del tipo de asociación entre sus modalidades se realice a partir de cualquier par de modalidades¹ permite el estudio de este tipo de tablas mediante el **coeficiente de asociación H** :

$$H = n_{11} - \frac{n_{1.} \cdot n_{.1}}{N}$$

- Cuando $H = 0$ se cumple la condición de independencia entre las modalidades A_1 y B_1 y, como la tabla es de dimensión 2×2 , los atributos son independientes.
- Cuando $H > 0$, los atributos no son independientes, existiendo asociación positiva entre las modalidades A_1 y B_1 y entre A_2 y B_2 , y asociación negativa entre A_1 y B_2 y entre A_2 y B_1 .
- Cuando $H < 0$, los atributos no son independientes y entre A_1 y B_1 y entre A_2 y B_2 hay asociación negativa y asociación positiva entre A_1 y B_2 y entre A_2 y B_1 .

¹ Como consecuencia, el coeficiente puede definirse con cualquier pareja de modalidades.

Tabla 2 x 2: coeficiente de asociación H

Ejemplo 3.8 Una cadena de supermercados ha encargado un estudio para conocer la relación entre el sexo de las personas consumidoras y el hecho de comprar o no productos ecológicos. Los resultados obtenidos para 1000 personas encuestadas se recogen en la siguiente tabla:

Productos ecológicos	Sexo	
	Hombre	Mujer
Sí	200	300
No	150	350

Tomando, por ejemplo, la segunda modalidad de primer atributo, *no comprar productos ecológicos*, y la primera del segundo atributo, *ser hombre*, se cumple:

$$n_{21} = 150 \neq 175 = \frac{n_{2\cdot} \cdot n_{\cdot 1}}{N}.$$

En definitiva, no se cumple la condición de independencia para, al menos, una pareja de modalidades y, por tanto, los dos atributos no son independientes. Como, además, $150 < 175$, existe asociación negativa entre las modalidades *no comprar productos ecológicos* y *ser hombre*, y entre *comprar productos ecológicos* y *ser mujer*; mientras que la asociación es positiva entre *no comprar productos ecológicos* y *ser mujer* y entre *comprar productos ecológicos* y *ser hombre*.

Medidas de asociación: coeficiente ji-cuadrado de Pearson

El coeficiente χ^2 de Pearson es una medida del grado de asociación entre atributos:

$$\chi^2 = \sum_{i=1}^h \sum_{j=1}^k \frac{\left(n_{ij} - \frac{n_{i.} \cdot n_{.j}}{N} \right)^2}{\frac{n_{i.} \cdot n_{.j}}{N}}$$

Este coeficiente compara cada frecuencia absoluta conjunta, n_{ij} , con la teórica que correspondería en caso de independencia : $n_{i.} \cdot n_{.j} / N$. Si el coeficiente es cero, los atributos son independientes y, si es distinto de cero, son dependientes; además, cuanto mayor sea su valor, mayor será el grado de asociación entre ambos.

De este coeficiente derivan diversas medidas de asociación como el *coeficiente de contingencia*, acotado entre 0 y 1, lo que facilita su interpretación:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Medidas de asociación: coeficiente de rangos de Spearman

Se dispone de N unidades ordenadas según el rango o posición que ocupan respecto a dos atributos A y B . El **coeficiente de rangos de Spearman** mide el grado de concordancia o discordancia entre ambas ordenaciones:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^N d_i^2}{N^3 - N},$$

donde $d_i = x_i - y_i$ es la diferencia genérica entre los rangos de ambos atributos.

Cuando $\rho = 0$ no hay relación entre ambas ordenaciones; si $\rho = 1$, la concordancia entre ambas ordenaciones es perfecta, y, si $\rho = -1$, la discordancia es perfecta.

Medidas de asociación: coeficiente de rangos de Spearman

Ejemplo 3.9. Se presentaron cinco agrupaciones estudiantiles a las elecciones del consejo de estudiantes. La tabla recoge los resultados de la encuesta previa realizada por una empresa especializada en estudios sociológicos y los relativos al escrutinio de las votaciones definitivas (ambos en porcentaje):

	E.U.D	J.V.C	A.L.E.	U.A.C	F.I.E
Encuesta	30	20	9	40	1
Votaciones	15	30	5	10	40

La tabla recoge los cálculos necesarios para obtener el coeficiente de rangos de Spearman:

	E.U.D	J.V.C	A.L.E.	U.A.C	F.I.E
1ª ordenación (encuesta)	2	3	4	1	5
2ª ordenación (votaciones)	3	2	5	4	1
d_i	-1	1	-1	-3	4
d_i^2	1	1	1	9	16

En definitiva:

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^N d_i^2}{N^3 - N} = 1 - \frac{6 \cdot 28}{5^3 - 5} = -0,4$$

Este valor indica que existe una ligera discordancia entre ambas ordenaciones.