

Tema 2: Distribuciones bidimensionales

Variable Bidimensional (X,Y) Sobre una población se observan simultáneamente dos variables X e Y.

La distribución de frecuencias bidimensional de (X,Y) es el conjunto de valores

$\{(x_i, y_j); n_{ij}\}$ $i=1, \dots, p; j=1, \dots, q$ tal que

$$\sum_i^p \sum_j^q n_{ij} = N \quad \text{O equivalente:} \quad \sum_i^p \sum_j^q f_{ij} = 1$$

donde n_{ij} es la frecuencia absoluta conjunta o total de elementos en la población que presenta el valor bidimensional (x_i, y_j) .

La frecuencia relativa conjunta f_{ij} es la proporción de elementos en la población que presenta el valor (x_i, y_j) .

$$f_{ij} = \frac{n_{ij}}{N}$$

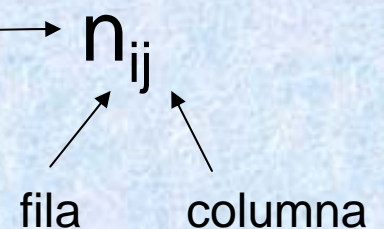
Tema 2: Distribuciones bidimensionales

La distribución de frecuencias bidimensional de (X,Y) se puede expresar en una tabla bidimensional:

	y_1	y_2	...	y_j	...	y_q	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	n_{1*}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	n_{2*}
...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	n_{i*}
...
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	n_{p*}
	n_{*1}	n_{*2}	...	n_{*j}	...	n_{*q}	N

Columna de frecuencias marginales

Frecuencia absoluta



Fila de frecuencias marginales

Total de elementos en la población

Tema 2: Distribuciones bidimensionales

La distribución de frecuencias bidimensional de (X,Y) se puede expresar en una tabla bidimensional (frecuencias absolutas):

	y_1	y_2	...	y_j	...	y_q	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	n_{1*}
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	n_{2*}
...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	n_{i*}
...
x_p	n_{p1}	n_{p2}	...	n_{pj}	...	n_{pq}	n_{p*}
	n_{*1}	n_{*2}	...	n_{*j}	...	n_{*q}	N

Total fila 1

$$n_{1*} = \sum_{j=1}^q n_{1j}$$

Total de elementos que presentan el valor x_i

$$n_{i*} = \sum_{j=1}^q n_{ij}$$

Total de elementos que presentan x_i e y_j

Total fila p

$$n_{p*} = \sum_{j=1}^q n_{pj}$$

Total de elementos en la población

$$N = \sum_{j=1}^q \sum_{i=1}^p n_{ij}$$

Total de elementos que presentan el valor y_j

Total columna j

$$n_{*j} = \sum_{i=1}^p n_{ij}$$

Tema 2: Distribuciones bidimensionales

La distribución de frecuencias bidimensional de (X,Y) se puede expresar en una tabla bidimensional (frecuencias relativas):

	y_1	y_2	...	y_j	...	y_q	
x_1	f_{11}	f_{12}	...	f_{1j}	...	f_{1q}	f_{1*}
x_2	f_{21}	f_{22}	...	f_{2j}	...	f_{2q}	f_{2*}
...
x_i	f_{i1}	f_{i2}	...	f_{ij}	...	f_{iq}	f_{i*}
...
x_p	f_{p1}	f_{p2}	...	f_{pj}	...	f_{pq}	f_{p*}
	f_{*1}	f_{*2}	...	f_{*j}	...	f_{*q}	1

Proporción de elementos que presenta el valor x_i

$$f_{i*} = \sum_{j=1}^q f_{ij}$$

Total fila 1

Total fila 2

Proporción de elementos que presenta x_i e y_j

$$1 = \sum_{j=1}^q \sum_{i=1}^p f_{ij}$$

Proporción de elementos que presenta el valor y_j

Total columna j

Total columna q

$$f_{*j} = \sum_{i=1}^p f_{ij}$$

Tema 2: Distribuciones bidimensionales

- Uno de los objetivos del análisis de distribuciones bidimensionales es estudiar si existe **asociación o relación** entre las variables X e Y.
- A partir de una distribución bidimensional se obtendrán distribuciones **unidimensionales** de dos tipos: **marginales** y **condicionadas**.
- Dos distribuciones marginales:
 - Marginal de X
 - Marginal de Y
- Condicionadas:
 - q distribuciones condicionadas de los valores de X a los q valores de Y
 - p distribuciones condicionadas de los valores de Y a los p valores de X

Tema 2: Distribuciones bidimensionales

A partir de una distribución bidimensional se pueden obtener 2 distribuciones unidimensionales MARGINALES: Marginal de X y Marginal de Y.

MARGINAL DE X

X	n_{i^*}	f_{i^*}
x_1	n_{1^*}	f_{1^*}
x_2	n_{2^*}	f_{2^*}
...
x_i	n_{i^*}	f_{i^*}
...
x_p	n_{p^*}	f_{p^*}
	N	1

$$f_{i^*} = \frac{n_{i^*}}{N}$$

Marginal de X: expresa cómo se distribuye X en la población total, al margen de la otra variable

Marginal de Y: expresa cómo se distribuye Y en la población total, al margen de la otra variable

MARGINAL DE Y

Y	y_1	y_2	...	y_j	...	y_q	
n_{*j}	n_{*1}	n_{*2}	...	n_{*j}	...	n_{*q}	N
f_{*j}	f_{*1}	f_{*2}	...	f_{*j}	...	f_{*q}	1

$$f_{*j} = \frac{n_{*j}}{N}$$

Tema 2: Distribuciones bidimensionales

A partir de una distribución bidimensional se pueden obtener distribuciones unidimensionales **CONDICIONADAS**: de X y de Y.

CONDICIONAL DE X / Y=y_j

X	n _{ij}	f _{i/j}
x ₁	n _{1j}	n _{1j} /n _{*j} =f _{1/j}
x ₂	n _{2j}	n _{2j} /n _{*j} =f _{2/j}
...
x _i	n _{ij}	n _{ij} /n _{*j} =f _{i/j}
...
x _p	n _{pj}	n _{pj} /n _{*j} =f _{p/j}
	n_{*j}	1

Total de elementos en la subpoblación

Condicional de X dado Y=y_j: expresa cómo se distribuye X en la subpoblación que cumple la condición de presentar el valor Y=y_j

Condicional de Y dado X=x_i: expresa cómo se distribuye Y en la subpoblación que cumple la condición de presentar el valor X=x_i

CONDICIONAL DE Y / X=x_i

Y	y ₁	y ₂	...	y _j	...	y _q	
n _{ij}	n _{i1}	n _{i2}	...	n _{ij}	...	n _{iq}	n_{i*}
f _{j/i}	n _{i1} /n _{i*} =f _{1/i}	n _{i2} /n _{i*} =f _{2/i}	...	n _{ij} /n _{i*} =f _{j/i}	...	n _{iq} /n _{i*} =f _{q/i}	1

Total de elementos en la subpoblación

Tema 2: Distribuciones bidimensionales

Ejemplo distribución bidimensional (en frecuencias absolutas y en relativas):
Un grupo de 91 niños se clasifica según su edad (X) y puntuación en un test (Y)

Frecuencias absolutas

Edad	TEST			
	120	125	130	
5	10	8	2	20
6	7	8	6	21
7	2	10	13	25
8	1	4	20	25
	20	30	41	91

$$f_{ij} = \frac{n_{ij}}{N}$$



Frecuencias relativas

Edad	TEST			
	120	125	130	
5	0,110	0,088	0,022	0,220
6	0,077	0,088	0,066	0,231
7	0,022	0,110	0,143	0,275
8	0,011	0,044	0,220	0,275
	0,220	0,330	0,451	1,000

$$0,110 = \frac{10}{91}$$

$$0,220 = \frac{20}{91}$$

¿Cómo se expresa la distribución bidimensional en frecuencias relativas a partir de la de frecuencias absolutas?

¡Es muy fácil! Se divide cada casilla (frecuencia absoluta) entre N (91)

Observa que la fila y columna marginales (sombreadas) representan las frecuencias marginales (las absolutas en tabla de la derecha y las relativas en la de la izquierda).

Tema 2: Distribuciones bidimensionales

Ejemplo distribución bidimensional (en frecuencias absolutas y en relativas):
Un grupo de 91 niños se clasifica según su edad (X) y puntuación en un test (Y)

En frecuencias absolutas

Edad	TEST			Marginal ↓
	120	125	130	
5	10	8	2	20
6	7	8	6	21
7	2	10	13	25
8	1	4	20	25
Marginal →	20	30	41	91

En frecuencias relativas

Edad	TEST			Marginal ↓
	120	125	130	
5	0,110	0,088	0,022	0,220
6	0,077	0,088	0,066	0,231
7	0,022	0,110	0,143	0,275
8	0,011	0,044	0,220	0,275
Marginal →	0,220	0,330	0,451	1,000

¿Cómo se interpretan los valores 10 y 20?

Hay 10 niños que tienen 7 años y puntuación 125 en el test. Hay 20 niños con puntuación igual a 120.

¿Cómo se interpretan los valores 0,110 y 0,220?

Hay una proporción de 0,11 niños que tiene 7 años y puntuación 125 en el test. El 22% de los niños tiene puntuación igual a 120.

Tema 2: Distribuciones bidimensionales

Ejemplo (continuación)

Distribuciones marginales de la Edad y Test

Distribución marginal de la Edad

<i>Edad</i>	Número alumnos	Proporción de alumnos
5	20	0,220
6	21	0,231
7	25	0,275
8	25	0,275
	91	1

Distribución marginal Del Test

<i>TEST</i>	número de alumnos	proporción de alumnos
120	20	0,220
125	30	0,330
130	41	0,451
	91	1

Observa que el total de individuos observados en cada marginal es 91. Todos.

¿qué porcentaje de niños tiene edad igual 5?

¿qué proporción de alumnos obtiene en el test más de 125 puntos?

Tema 2: Distribuciones bidimensionales

Ejemplo (continuación)

Distribuciones condicionadas de la Edad a los valores del test

Distribución bidimensional

		TEST			
Edad	120	125	130		
5	10	8	2	20	
6	7	8	6	21	
7	2	10	13	25	
8	1	4	20	25	
	20	30	41	91	

Distribuciones condicionadas de la Edad

		TEST			
Edad	120	125	130		
5	0,500	0,267	0,049	0,220	
6	0,350	0,267	0,146	0,231	
7	0,100	0,333	0,317	0,275	
8	0,050	0,133	0,488	0,275	
	1,000	1,000	1,000	1,000	

¿Cómo se hace?

Se divide cada casilla de la bidimensional (tabla izquierda) entre el total de columna.

Las flechas de la tabla indican la dirección en que se han de hacer los cálculos

Por ejemplo, para obtener la distribución condicionada de la Edad / test =120 se divide cada casilla de la columna encabezada por 120 por el total de columna (20). Observa que la población que cumple esa condición es de 20 niños.

Observa que la última fila está formada por unos. Hay 3 distribuciones condicionadas. Una marginal.

Tema 2: Distribuciones bidimensionales

Ejemplo (continuación)

Distribuciones condicionadas de la Edad a los valores del test

Distribución bidimensional

Edad	TEST			
	120	125	130	
5	0,110	0,088	0,022	0,220
6	0,077	0,088	0,066	0,231
7	0,022	0,110	0,143	0,275
8	0,011	0,044	0,220	0,275
	0,220	0,330	0,451	1,000

Distribuciones condicionadas de la Edad

Edad	TEST			
	120	125	130	
5	0,500	0,267	0,049	0,220
6	0,350	0,267	0,146	0,231
7	0,100	0,333	0,317	0,275
8	0,050	0,133	0,488	0,275
	1,000	1,000	1,000	1,000

¿Cómo se hace si la distribución bidimensional está en frecuencias relativas?

Igual que antes. Se divide cada casilla de la bidimensional (tabla izquierda) entre el total de columna.

Las flechas de la tabla indican la dirección en que se han de hacer los cálculos

Por ejemplo, para obtener la distribución condicionada de la Edad / test =120 se divide cada casilla de la columna encabezada por 120 por el total de columna (0,220). Observa que la población que cumple esa condición es de una proporción igual a 0,022 niños.

Observa que la última fila está formada por unos. Hay 3 distribuciones condicionadas de la Edad. Una marginal de la Edad.

Tema 2

Ejemplo (continuación)

Distribuciones condicionadas del Test a los valores de la edad

Distribución bidimensional

	TEST			
Edad	120	125	130	
5	0,110	0,088	0,022	0,220
6	0,077	0,088	0,066	0,231
7	0,022	0,110	0,143	0,275
8	0,011	0,044	0,220	0,275
	0,220	0,330	0,451	1,000

Distribuciones condicionadas del Test

	TEST			
Edad	120	125	130	
5	0,500	0,400	0,100	1
6	0,333	0,381	0,286	1
7	0,080	0,400	0,520	1
8	0,040	0,160	0,800	1
	0,220	0,330	0,451	1

¿Cómo se hace?

Las flechas de la tabla indican la dirección en que se han de hacer los cálculos

Por ejemplo, para obtener la distribución condicionada del test /Edad=6 años se divide cada casilla de la fila encabezada por 6 entre el total de fila (0,231). Observa que la población que cumple esa condición es de una proporción igual a 0,231 niños.

Observa que la última columna está formada por unos. Hay 4 distribuciones condicionadas del test. Y la marginal del test.

Tema 2

- Uno de los objetivos del análisis de distribuciones bidimensionales es estudiar si son ***independientes*** o por el contrario, existe ***asociación o relación*** entre las variables X e Y.
- Las variables X e Y se dicen que son ***independientes*** si los valores de una de ellas no afecta a la distribución de la otra. Esto equivale a decir que **todas las distribuciones condicionadas sean iguales**.
- De modo equivalente se dice que las variables X e Y son independientes si se cumple que la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales.
- Si las variables no son independientes se dice que están relacionadas o asociadas. Las distribuciones condicionadas NO son iguales.

Tema 2

Ejemplo:

Comprueba si son o no independientes las variables X e Y de la distribución bidimensional (X, Y) siguiente:

	y1	y2	
x1	23	69	92
x2	12	36	48
x3	15	45	60
x4	7	21	28
	57	171	228

↓
Cálculo

Basta ver que las distribuciones condicionadas son iguales. Por ejemplo, las condicionadas de X/Y

Condicionadas de X a los valores de Y: X/Y

	y1	y2	
x1	0,404	0,404	0,404
x2	0,211	0,211	0,211
x3	0,263	0,263	0,263
x4	0,123	0,123	0,123
	1	1	1

¿Cómo se hacen los cálculos?

Verticalmente: Dividiendo cada casilla (frecuencia) entre el total de columna

Observa que la variable X se distribuye igual en el conjunto de individuos que presenta la condición $Y=y1$, que en el grupo que cumple $Y=y2$.

La lectura de la tabla de condicionadas se hace en sentido contrario al que se hayan realizado los cálculos; es decir, en el ejemplo la lectura es horizontal: Fila 1: $0,404 = 0,404$; Fila 2: $0,211=0,211$; Fila 3: $0,263=0,263$; Fila 4: $0,123=0,123$.

Todas las condicionadas son iguales. Por tanto las variables X e Y son INDEPENDIENTES

Tema 2

Ejemplo (Continuación):

Comprueba si son o no independientes las variables X e Y de la distribución bidimensional (X, Y) siguiente:

	y1	y2	
x1	23	69	92
x2	12	36	48
x3	15	45	60
x4	7	21	28
	57	171	228

Cálculo



Otro modo de ver que son independientes es comprobando que las distribuciones condicionadas de Y/X son todas iguales.

Condicionadas de Y a los valores de X: Y/X

¿Cómo se hacen los cálculos?

	y1	y2	
x1	0,250	0,750	1,000
x2	0,250	0,750	1,000
x3	0,250	0,750	1,000
x4	0,250	0,750	1,000
	0,25	0,75	1

Lectura

Horizontalmente: Dividiendo cada casilla (frecuencia) entre el total de fila

Observa que la variable Y se distribuye igual en el conjunto de individuos que presenta la condición $X=x1$, que en el grupo que cumple $X=x2, \dots$, y que en el grupo $X=x4$.

La lectura de la tabla de condicionadas se hace en sentido contrario al que se hayan realizado los cálculos; es decir, en el ejemplo la lectura es vertical: Columna 1: $0,250 = 0,250 = 0,250 = 0,250$; Columna 2: $0,750 = 0,750 = 0,750 = 0,750$.

Todas las condicionadas son iguales. Por tanto las variables X e Y son INDEPENDIENTES

Tema 2

Ejemplo (Continuación):

Comprueba si son o no independientes las variables X e Y de la distribución bidimensional (X, Y) siguiente: (Puedes hacerlo con frecuencias absolutas o con relativas)

	y1	y2	
x1	23	69	92
x2	12	36	48
x3	15	45	60
x4	7	21	28
	57	171	228

Otro modo de ver que son independientes es comprobando que las frecuencias relativas conjuntas verifican la ecuación:

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

O la equivalente

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

¿Cómo?

Comprueba que cada frecuencia absoluta verifica la ecuación. Por ejemplo,

$$15 = \frac{60 \cdot 57}{228}$$

	y1	y2	
x1	0,101	0,303	0,404
x2	0,053	0,158	0,211
x3	0,066	0,197	0,263
x4	0,031	0,092	0,123
	0,250	0,750	1,000

¿Cómo?

si prefieres usar la primera ecuación:

Se obtiene la distribución bidimensional en frecuencias relativas. Para ello divide cada casilla correspondiente a una frecuencia absoluta entre 228

Por ejemplo, $0,101 = 23/228$.

Comprueba luego que se verifica $0,101 = 0,0404$ por $0,250$; $0,303 = 0,404$ por $0,750$;, $0,092 = 0,123$ por $0,750$.

Tema 2: Distribuciones bidimensionales

- Resumiendo, habrás observado que una tabla bivariante para una bidimensional (X, Y) puede expresarse en frecuencias absolutas y relativas.
- Cuando las variables X o Y son cualitativas se denomina **tabla de contingencia**
- Una tabla en proporciones puede indicar que hay una sola distribución bidimensional o que hay varias distribuciones unidimensionales condicionadas.
- ¿Cómo puedo saber si hay una sola distribución de carácter bidimensional o varias condicionadas (unidimensionales)?
 - La respuesta es fácil. Si la suma de todas las frecuencias de la tabla es 1, hay una sola distribución bidimensional. Estas proporciones se obtienen dividiendo cada frecuencia absoluta n_{ij} entre el total de elementos N .
 - Si la suma de cada columna es 1, hay tantas distribuciones como columnas. Las proporciones se han obtenido dividiendo cada casilla por el total columna.
 - Si la suma de cada fila es 1, hay tantas distribuciones como filas. Las proporciones se han obtenido dividiendo cada casilla por el total de fila.
- Vamos a repasar un ejemplo que ya vimos.

Tema 2

Ejemplo (repasso)

<i>Edad</i>	<i>TEST</i>			
	120	125	130	
5	0,500	0,400	0,100	1
6	0,333	0,381	0,286	1
7	0,080	0,400	0,520	1
8	0,040	0,160	0,800	1
	0,220	0,330	0,451	1

Observa que la suma de las frecuencias de cada fila es 1

Hay 5 distribuciones UNIDIMENSIONALES: 4 condicionadas y una marginal

¿Cómo se interpreta la frecuencia 0,100 de la fila 1?

El 10% de los niños que tienen 5 años ha obtenido una puntuación de 130 en el test

¿Cuál es la distribución condicionada del Test para el grupo que tiene 8 años?

<i>Edad</i>	<i>TEST</i>			
	120	125	130	
8	0,040	0,160	0,800	1

¿Cómo se distribuye la edad?

No se puede saber con la información que hay en la tabla bidimensional

Tema 2: Distribuciones bidimensionales

- Cuando las variables X o Y son cualitativas se denomina **tabla de contingencia**.
- Un análisis típico de una tabla de contingencia es el estudio de la posible asociación o relación entre las variables X e Y.
- Una medida muy importante de asociación es el estadístico **Chi-cuadrado**:

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(t_{ij} - n_{ij})^2}{t_{ij}}$$

Con

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

Donde t_{ij} es el valor de la frecuencia que teóricamente se observaría si las variables X e Y fueran independientes

Tema 2

Ejemplo:

Vamos a calcular este estadístico en los dos ejemplos anteriores.

	y1	y2	
x1	23	69	92
x2	12	36	48
x3	15	45	60
x4	7	21	28
	57	171	228

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(t_{ij} - n_{ij})^2}{t_{ij}}$$

Con

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

Observa que:

$$t_{11} = \frac{92 \cdot 57}{228} = 23; \quad t_{12} = \frac{92 \cdot 171}{228} = 69; \dots; t_{42} = \frac{28 \cdot 171}{228} = 21$$

Observa que todo t_{ij} coincide con lo observado realmente (n_{ij}) y los numeradores de la expresión de Chi-cuadrado son todos nulos, y por tanto la suma y Chi-cuadrado es cero.

Tema 2

Ejemplo: Veamos el valor de chi-cuadrado en la tabla siguiente:

		TEST			
		120	125	130	
Edad	5	10	8	2	20
	6	7	8	6	21
	7	2	10	13	25
	8	1	4	20	25
		20	30	41	91

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(t_{ij} - n_{ij})^2}{t_{ij}}$$

Con

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

$$t_{11} = \frac{20 \cdot 20}{91} = 4,396; \quad t_{12} = \frac{20 \cdot 30}{91} = 6,593; \dots; t_{43} = \frac{25 \cdot 41}{91} = 11,264$$

Para realizar los cálculos es cómodo colocar columnas que indiquen los pasos sucesivos a realizar para obtener el estadístico:

Tema 2

Ejemplo: Veamos el valor de chi-cuadrado en la tabla siguiente:

Valores observados (n_{ij})

		TEST			
		120	125	130	
Edad	5	10	8	2	20
	6	7	8	6	21
	7	2	10	13	25
	8	1	4	20	25
		20	30	41	91

Valores teóricos bajo independencia (t_{ij})

		TEST			
		120	125	130	
Edad	5	4,396	6,593	9,011	20
	6	4,615	6,923	9,462	21
	7	5,495	8,242	11,264	25
	8	5,495	8,242	11,264	25
		20	30	41	91

$$t_{11} = \frac{20 \cdot 20}{91} = 4,396; \quad t_{12} = \frac{20 \cdot 30}{91} = 6,593; \dots; t_{43} = \frac{25 \cdot 41}{91} = 11,264$$

Para realizar los cálculos es cómodo colocar columnas que indiquen los pasos sucesivos a realizar para obtener el estadístico:

Tema 2

Ejemplo: Cálculo chi-cuadrado (continuación):

La tabla siguiente indica los cálculos necesarios

n_{ij}	t_{ij}	$n_{ij}-t_{ij}$	$(n_{ij}-t_{ij})^2$	$(n_{ij}-t_{ij})^2 / t_{ij}$
10	4,3956	5,6044	31,4093	7,1456
7	4,6154	2,3846	5,6864	1,2321
2	5,4945	-3,4945	12,2116	2,2225
1	5,4945	-4,4945	20,2006	3,6765
8	6,5934	1,4066	1,9785	0,3001
8	6,9231	1,0769	1,1598	0,1675
10	8,2418	1,7582	3,0914	0,3751
4	8,2418	-4,2418	17,9925	2,1831
2	9,0110	-7,0110	49,1540	5,4549
6	9,4615	-3,4615	11,9822	1,2664
13	11,2637	1,7363	3,0146	0,2676
20	11,2637	8,7363	76,3223	6,7759

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{(t_{ij} - n_{ij})^2}{t_{ij}}$$

La suma de la última columna es el valor de chi-cuadrado

31,067

Tema 2

- Análisis de regresión

El análisis de regresión consiste en la búsqueda de una función que exprese la forma en que se relaciona una variable dependiente (Y) con una o más variables independientes (X)

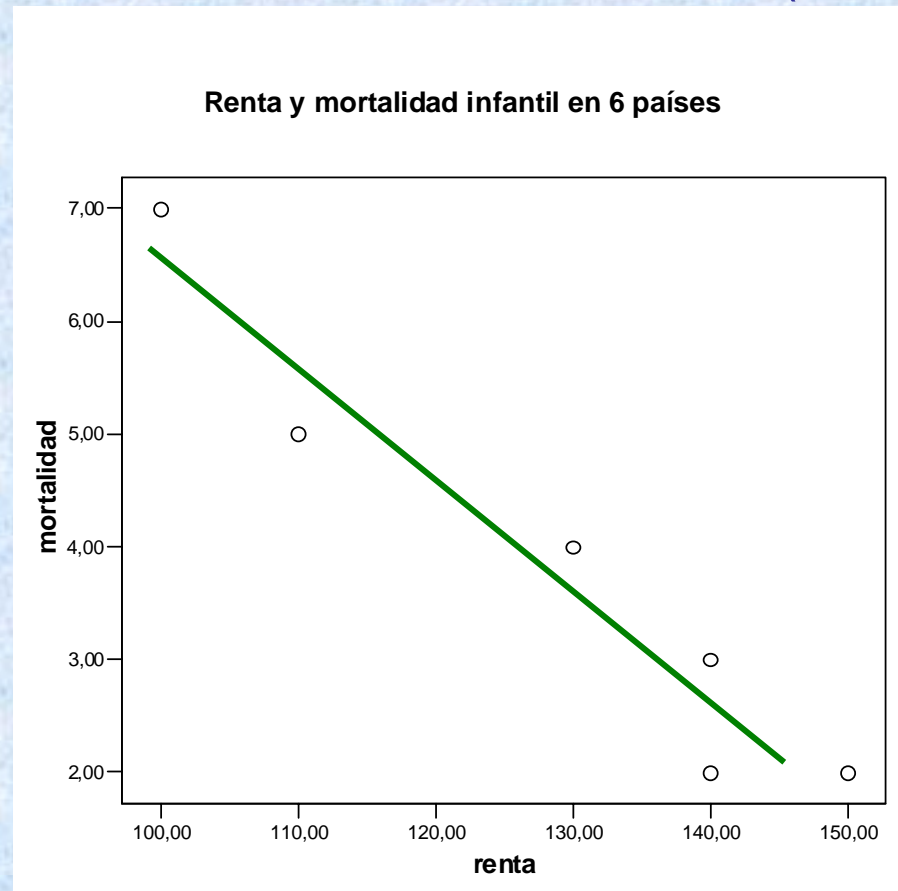
- Nos ocuparemos sólo del caso de regresión lineal simple: una variable dependiente y otra independiente.
- Se supone que la relación entre las variables es aproximadamente lineal (una recta). Una forma visual de comprobar si es o no lineal la trayectoria de la relación entre las variables es mediante el denominado diagrama de dispersión o nube de puntos.

Tema 2

- Gráfico de dispersión o Nube de puntos.
 - Es la representación gráfica en el plano del conjunto de puntos (x_i, y_i) que constituyen los valores bidimensionales de la variable bidimensional (X, Y) .

Renta	Mortalidad
100	7
110	5
130	4
140	3
140	2
150	2

Se observa una trayectoria casi lineal



Tema 2

- Recta de regresión de Y sobre X.
- La recta de regresión Y/X presenta la forma:

$$Y = a + bX$$

Diagram illustrating the components of the regression equation $Y = a + bX$:

- Y : Variable dependiente
- a : Ordenada en el origen
- b : Pendiente
- X : Variable independiente

El objetivo es encontrar los valores **a** y **b** que definen la recta que se encuentra a la mínima distancia de los puntos de la nube.

El procedimiento que permite encontrar dicha recta se denomina de **mínimos cuadrados**

Tema 2

- Recta de regresión de Y sobre X: $Y/X: Y=a+bX$

$$S = \sum_i d_i^2 n_i = \sum_i (y_i - y'_i)^2 n_i =$$

$$= \sum_i (y_i - a - bx_i)^2 n_i$$

Para obtener el mínimo de S se deriva la ecuación anterior respecto de a y b. El sistema de ecuaciones generado viene dado por:

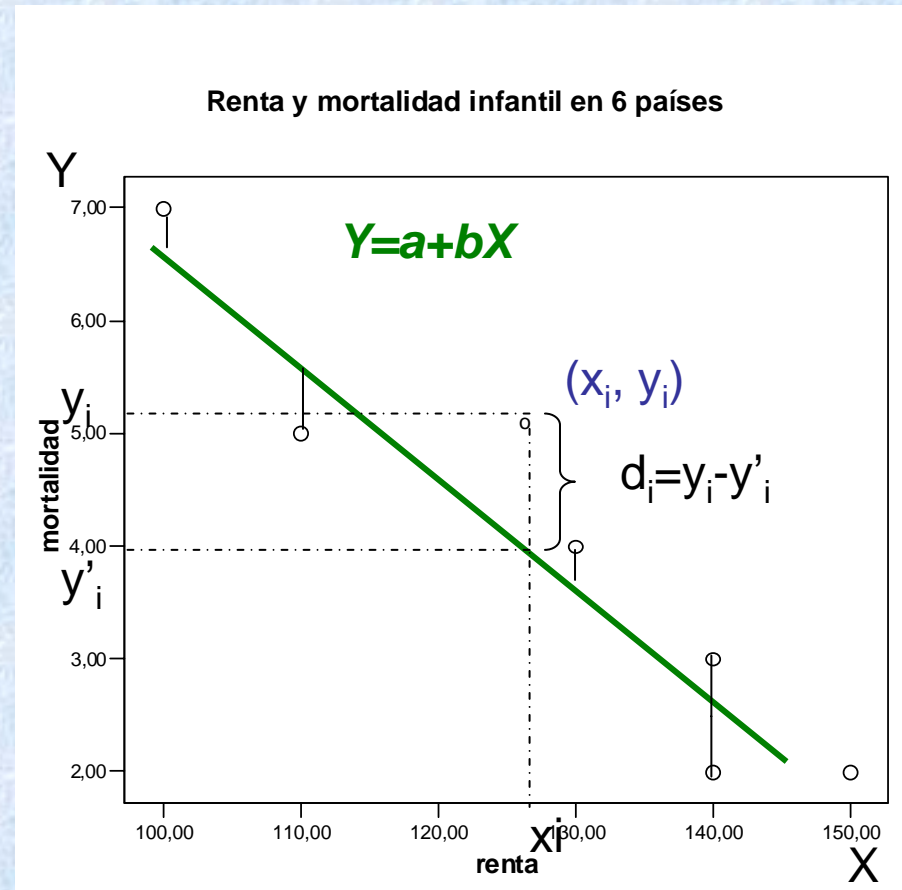
$$\sum_i y_i n_i = a \sum_i n_i + b \sum_i x_i n_i$$

$$\sum_i x_i y_i n_i = a \sum_i x_i n_i + b \sum_i x_i^2 n_i$$

Y la solución **a** y **b** es:

$$b = \frac{\frac{\sum_i x_i y_i n_i}{N} - \bar{X}\bar{Y}}{\frac{\sum_i x_i^2 n_i}{N} - \bar{X}^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$a = \bar{Y} - b\bar{X}$$



Nota: El estadístico $Cov(X,Y)$ se denomina covarianza de X e Y.

Tema 2

- Recta de regresión de X sobre Y.
- La recta de regresión X/Y presenta la forma:

$$X = a' + b'Y$$

Diagram illustrating the components of the regression equation $X = a' + b'Y$:

- Variable dependiente (Dependent variable) points to X .
- Ordenada en el origen (Intercept) points to a' .
- Pendiente (Slope) points to b' .
- Variable independiente (Independent variable) points to Y .

El objetivo es encontrar los valores a' y b' que definen la recta que se encuentra a la mínima distancia de los puntos de la nube.

El procedimiento que permite encontrar dicha recta se denomina de **mínimos cuadrados**

Tema 2

- Recta de regresión de X sobre Y: X/Y: $X=a'+b'Y$

$$S = \sum_i d_i^2 n_i = \sum_i (x_i - x'_i)^2 n_i =$$

$$= \sum_i (x_i - a' - b' y_i)^2 n_i$$

Para obtener el mínimo de S se deriva la ecuación anterior respecto de a' y b' . El sistema de ecuaciones generado viene dado por:

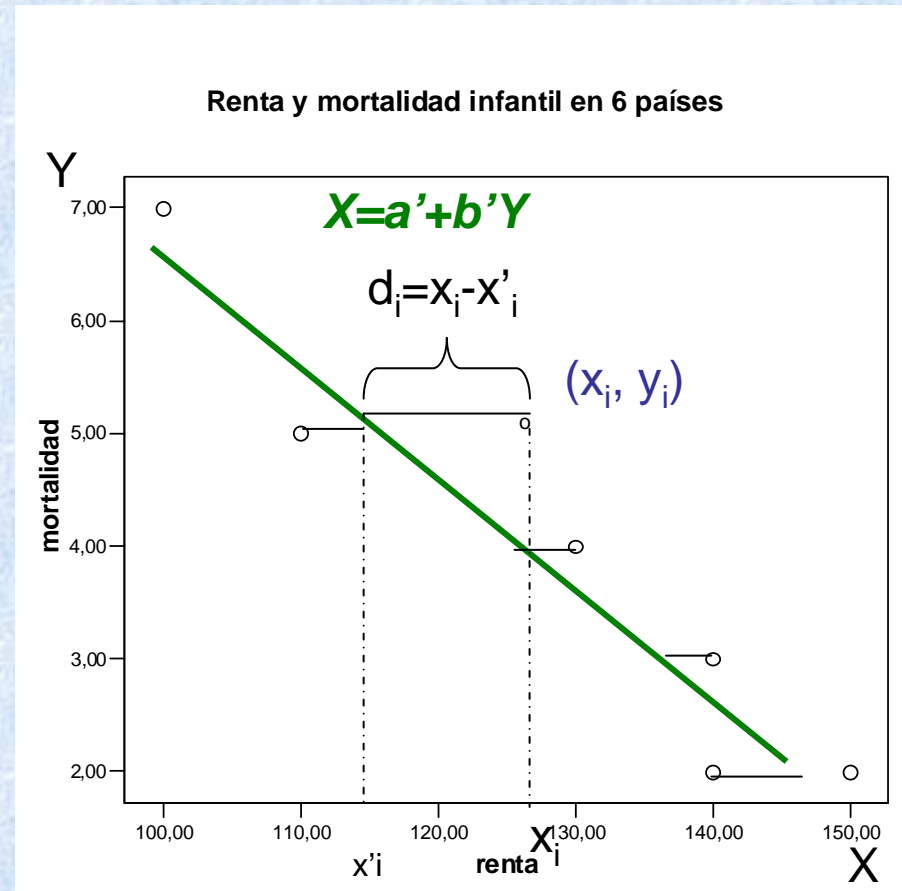
$$\sum_i x_i n_i = a' \sum_i n_i + b' \sum_i y_i n_i$$

$$\sum_i x_i y_i n_i = a' \sum_i y_i n_i + b' \sum_i y_i^2 n_i$$

Y la solución a' y b' es:

$$b' = \frac{\frac{\sum_i x_i y_i n_i}{N} - \bar{X}\bar{Y}}{\frac{\sum_i y_i^2 n_i}{N} - \bar{Y}^2} = \frac{Cov(X, Y)}{Var(Y)}$$

$$a' = \bar{X} - b'\bar{Y}$$



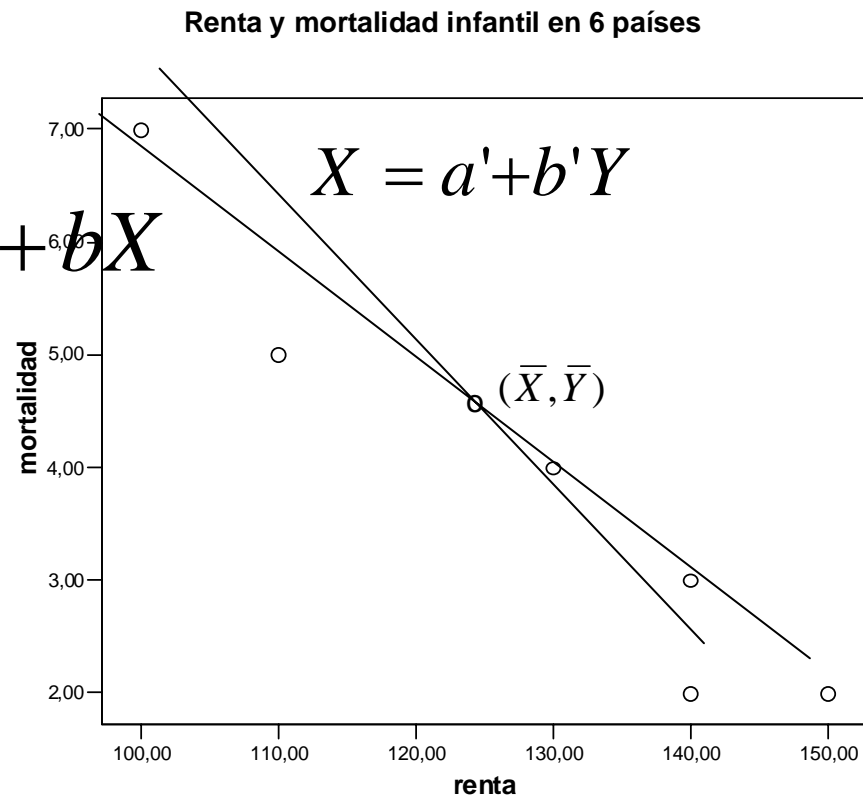
Nota: Observa que el procedimiento es el mismo salvo que se cambian los papeles de X por Y. Las distancias ahora son horizontales (paralelas al eje X).

Tema 2

- Las rectas de regresión de Y sobre X y de X sobre Y se cortan en el punto medio de las variables. Cuando el ajuste es perfecto, las dos rectas coinciden

$$Y = a + bX$$

$$X = a' + b'Y$$



Tema 2

- Ejemplo:

- Obtenga las rectas de regresión de Y sobre X y de X sobre Y.

- Y=Mortalidad infantil, X=Renta per cápita $Y = a + bX$ $X = a' + b'Y$

Renta	Mortalidad
100	7
110	5
130	4
140	3
140	2
150	2

Recta de regresión de Y sobre X: $Y = a + bX$

$$b = \frac{\frac{\sum_i x_i y_i n_i}{N} - \bar{X}\bar{Y}}{\frac{\sum_i x_i^2 n_i}{N} - \bar{X}^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$a = \bar{Y} - b\bar{X}$$

Para determinar a y b necesitamos los cálculos que expresamos por comodidad en las columnas de la tabla siguiente:

Renta(X)	Mortalidad (Y)	XY	X^2
100	7	700	10000
110	5	550	12100
130	4	520	16900
140	3	420	19600
140	2	280	19600
150	2	300	22500
770	23	2770	100700

$$\bar{X} = \frac{\sum_i x_i n_i}{N} = \frac{770}{6} = 128,333$$

$$\bar{Y} = \frac{\sum_i y_i n_i}{N} = \frac{23}{6} = 3,833$$

$$V(X) = \frac{\sum_i x_i^2 n_i}{N} - \bar{X}^2 = \frac{100700}{6} - 128,333^2 = 313,889$$

$$Cov(X,Y) = \frac{\sum_i x_i y_i n_i}{N} - \bar{X}\bar{Y} = \frac{2770}{6} - 128,333 \cdot 3,833 = -30,278$$

$$b = \frac{Cov(X,Y)}{Var(X)} = \frac{-30,278}{313,889} = -0,096$$

$$a = \bar{Y} - b\bar{X} = 3,833 - (-0,096 \cdot 128,333) = 16,212$$

Tema 2

- Ejemplo (continúa):
 - La ecuación de la recta de regresión de Y sobre X es:

$$Y = 16,212 - 0,096X$$

Obtenga la recta de regresión de X sobre Y: $X = a' + b'Y$

Renta(X)	Mortalidad(Y)	XY	Y ²
100	7	700	49
110	5	550	25
130	4	520	16
140	3	420	9
140	2	280	4
150	2	300	4
770	23	2770	107

$$b' = \frac{\frac{\sum x_i y_i n_i}{i} - \bar{X}\bar{Y}}{\frac{\sum y_i^2 n_i}{i} - \bar{Y}^2} = \frac{Cov(X, Y)}{Var(Y)}$$

$$a' = \bar{X} - b'\bar{Y}$$

$$V(Y) = \frac{\sum y_i^2 n_i}{N} - \bar{Y}^2 = \frac{107}{6} - 3,8333^2 = 3,139$$

$$b' = \frac{Cov(X, Y)}{Var(Y)} = \frac{-30,278}{3,139} = -9,646$$

$$a' = \bar{X} - b'\bar{Y} = 128,333 - (-9,646 \cdot 3,8333) = 165,310$$

$$X = 165,310 - 9,646Y$$

Tema 2

- **Coeficiente de correlación lineal de Pearson.**
- Un coeficiente muy usado para medir el grado de relación lineal entre las variables X e Y es el debido a Pearson, que notamos con r
- Se define como el cociente entre la covarianza y el producto de las desviaciones típicas de las variables
- Al coeficiente r al cuadrado se denomina coeficiente de determinación y expresa la proporción de variación de la variable dependiente que es explicada por la independiente.
- También se usa como medida de bondad de ajuste. Una propiedad interesante del coeficiente de correlación lineal de Pearson es que está comprendido entre los valores -1 y 1. El valor 0 indica ausencia de correlación lineal. Los valores -1 y 1 indican correlación lineal perfecta (todos los puntos están sobre las rectas de regresión), el negativo indican que cuando una variable crece (disminuye) la otra decrece (aumenta) y el positivo indica que cuando una aumenta (disminuye) la otra también aumenta (disminuye).
- Se dice que la correlación es más débil cuanto más se aproxima a cero. Y más fuerte cuanto más se aproxima a los extremos -1 ó 1.

$$r = \frac{Cov(X, Y)}{\sigma_x \sigma_y}$$

$$-1 \leq r \leq 1$$

$$r^2 = \frac{Cov^2(X, Y)}{V(X) \cdot V(Y)} = \frac{Cov(X, Y)}{V(X)} \cdot \frac{Cov(X, Y)}{V(Y)} = b \cdot b'$$

Tema 2

- **Ejemplo:** Con los datos del ejemplo anterior determina el coeficiente de correlación lineal de Pearson y Coeficiente de determinación. Interpretación.

$$r^2 = \frac{Cov^2(X, Y)}{V(X) \cdot V(Y)} = \frac{(-30,278)^2}{313,889 \cdot 3,139} = 0,930$$

El 93% de la variabilidad de la variable dependiente es explicada por la independiente

$$r = -\sqrt{0,930} = -0,965$$

El coeficiente de correlación lineal de Pearson presenta un valor negativo y próximo a -1 (-0,965), por tanto, las variables están relacionadas linealmente con fuerte grado de relación positiva. Es decir, cuanto mayor es la renta menor es la mortalidad.

Observa que el signo de la correlación es el signo de la covarianza

Tema 2

- **Ejemplo:** Con los datos del ejemplo anterior determina el coeficiente de correlación lineal de Pearson y Coeficiente de determinación. Interpretación.

$$r^2 = \frac{Cov^2(X, Y)}{V(X) \cdot V(Y)} = \frac{(-30,278)^2}{313,889 \cdot 3,139} = 0,930$$

El 93% de la variabilidad de la variable dependiente es explicada por la independiente

$$r = -\sqrt{0,930} = -0,965$$

El coeficiente de correlación lineal de Pearson presenta un valor negativo y próximo a -1 (-0,965), por tanto, las variables están relacionadas linealmente con fuerte grado de relación negativa. Es decir, cuanto mayor es la renta menor es la mortalidad.

Observa que el signo de la correlación es el signo de la covarianza

Tema 2

- Ejemplo:** Con los datos del ejemplo anterior determina la recta de regresión de Test sobre Edad, el coeficiente de correlación lineal de Pearson y Coeficiente de determinación. Interpretación. Determina el valor esperado o ajustado para el test para un niño de 10 años.

Edad	TEST			
	120	125	130	
5	10	8	2	20
6	7	8	6	21
7	2	10	13	25
8	1	4	20	25
	20	30	41	91

$$Test = a + bEdad$$

$$r = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

$$r^2 = \frac{Cov^2(X, Y)}{V(X) \cdot V(Y)}$$

$$Test = a + b \cdot 10$$

Vamos a expresar la tabla en un formato más cómodo para realizar los cálculos:
3 columnas

Edad	Test	Frecuencias
------	------	-------------

Nota: Observa que en el ejemplo que vimos anteriormente se omitió la columna frecuencias por valer 1

Tema 2

- **Ejemplo (continuación):** La tabla siguiente recoge los cálculos necesarios

<i>Edad=X</i>	<i>Test=Y</i>	<i>frecuencias =n</i>	<i>Xn</i>	<i>Yn</i>	<i>XYn</i>	<i>X²n</i>	<i>Y²n</i>
5	120	10	50	1200	6000	250	144000
6	120	7	42	840	5040	252	100800
7	120	2	14	240	1680	98	28800
8	120	1	8	120	960	64	14400
5	125	8	40	1000	5000	200	125000
6	125	8	48	1000	6000	288	125000
7	125	10	70	1250	8750	490	156250
8	125	4	32	500	4000	256	62500
5	130	2	10	260	1300	50	33800
6	130	6	36	780	4680	216	101400
7	130	13	91	1690	11830	637	219700
8	130	20	160	2600	20800	1280	338000
			601	11480	76040	4081	1449650

$$Test = a + bEdad \equiv Y = a + bX$$

Tema 2

- Ejemplo (continuación):

$$\bar{X} = \frac{601}{91} = 6,6044; \bar{Y} = \frac{11480}{91} = 126,1538$$

$$\text{Cov}(X, Y) = \frac{76040}{91} - 6,6044 \cdot 126,1538 = 2,4345$$

$$V(X) = \frac{4081}{91} - 6,6044^2 = 1,2281$$

$$V(Y) = \frac{1449650}{91} - 126,1538^2 = 15,4269$$

$$b = \frac{2,4345}{1,2281} = 1,9823$$

$$a = 126,1538 - 1,9823 \cdot 6,6044 = 113,0619$$

$$\text{Test} = a + b\text{Edad} \equiv Y = a + bX$$

$$\text{Test} = 113,06 + 1,98\text{Edad}$$

$$\text{Test} = 113,06 + 1,98\text{Edad}$$

$$132,86 = 113,06 + 1,98 \cdot 10$$

$$r^2 = \frac{\text{Cov}^2(X, Y)}{V(x) \cdot V(Y)} = \frac{2,4345^2}{1,2281 \cdot 15,4269} = 0,3128$$
$$r = 0,5593$$