

# Curso de Estadística básica para Data Scientists

Dae-Jin Lee < [lee.daejin@gmail.com](mailto:lee.daejin@gmail.com) >

## TEMA 3. Estadística descriptiva bivalente

### Índice

<b>1. Estadística descriptiva bivalente</b>	<b>2</b>
1.1. Distribuciones de frecuencias . . . . .	2
1.2. Representaciones gráficas . . . . .	5
1.3. Medidas de dependencia lineal . . . . .	6

[Regresar a la página principal](#)

## 1. Estadística descriptiva bivalente

En el tema anterior hemos visto cómo describir una muestra de datos de una variable mediante:

- Representaciones gráficas
- Estadísticos (de posición, dispersión, etc ...)

En este tema consideraremos el caso de 2 variables. Nuestro interés será por tanto, saber si  $Y$  es otra variable definida sobre la misma población que  $X$ , ¿será posible determinar si existe alguna relación entre  $X$  e  $Y$ ?

Ahora tomamos dos medidas a cada individuo de la muestra. Consideramos una población de  $n$  individuos, donde cada uno de ellos presenta dos características que representamos mediante las variables  $X$  e  $Y$ .

Las variables pueden ser cuantitativas, discretas o continuas o cualitativas, dando lugar a muchas combinaciones:

- cualitativa/cualitativa, discreta/continua, continua/continua, etc.

El tipo de análisis dependerá de la combinación que tengamos.

Tomamos una muestra de la población y medimos las variables ( $X$  e  $Y$ ), esa muestra estará dividida en clases para cada una de las variables y existirán elementos que pertenezcan a las distintas combinaciones de las clases.

### 1.1. Distribuciones de frecuencias

Llamamos **distribución conjunta de frecuencias** de dos variables ( $X$  e  $Y$ ) a la tabla que representa los valores observados y las frecuencias relativas/absolutas de cada par.

#### 1.1.1. Distribución conjunta

Consideramos una población de  $n$  individuos, donde cada uno de ellos presenta dos caracteres que representamos mediante las variables  $X$  e  $Y$ . Representamos mediante:+

$$X \rightarrow x_1, x_2, \dots, x_k$$

las  $k$  modalidades que presenta la variable  $X$  y mediante

$$Y \rightarrow y_1, y_2, \dots, y_p$$

las  $p$  modalidades de  $Y$ .

Con la intención de reunir en una sola estructura toda la información disponible, creamos una tabla formada por  $k \times p$  casillas, organizadas de forma que se tengan  $k$  filas y  $p$  columnas. La casilla denotada de forma general mediante el subíndice  $_{ij}$  hará referencia a los elementos de la muestra que presentan simultáneamente las modalidades  $x_i$  e  $y_j$ .

$X \setminus Y$	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_p$	
$x_1$	$n_{11}$	$n_{12}$	$\dots$	$n_{1j}$	$\dots$	$n_{1p}$	$n_{1\cdot}$
$x_2$	$n_{21}$	$n_{22}$	$\dots$	$n_{2j}$	$\dots$	$n_{2p}$	$n_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_i$	$n_{i1}$	$n_{i2}$	$\dots$	$n_{ij}$	$\dots$	$n_{ip}$	$n_{i\cdot}$
$\dots\dots\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_k$	$n_{k1}$	$n_{k2}$	$\dots$	$n_{kj}$	$\dots$	$n_{kp}$	$n_{k\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$\dots$	$n_{\cdot j}$	$\dots$	$n_{\cdot p}$	$n_{\cdot \cdot}$

De este modo, para  $i = 1, \dots, k$ , y para  $j = 1, \dots, p$ , se tiene que  $n_{ij}$  es el número de individuos o frecuencia absoluta, que presentan a la vez las modalidades  $x_i$  e  $y_j$ .

El número de individuos que presentan la modalidad  $x_i$ , es lo que llamamos frecuencia absoluta marginal de  $x_i$  y se representa como  $n_{i\cdot}$ , donde

$$n_{i\cdot} = n_{i1} + n_{i2} + \dots + n_{ip} = \sum_{j=1}^p n_{ij}$$

El símbolo  $\cdot$  en  $n_{i\cdot}$  simboliza que estamos considerando los elementos de  $x_i$  independientemente de los valores de  $Y$ . De forma análoga, se define la frecuencia absoluta marginal de la modalidad  $y_j$  como:

$$n_{\cdot j} = n_{1j} + n_{2j} + \dots + n_{kj} = \sum_{i=1}^k n_{ij}$$

### Propiedad

$$n_{\cdot \cdot} = \sum_{i=1}^k \sum_{j=1}^p n_{ij} = n$$

### 1.1.2. Distribución marginal

A la proporción de elementos (tanto por uno) que presentan simultáneamente las modalidades  $x_i$  e  $y_j$  la llamamos frecuencia relativa  $f_{ij}$

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

siendo las frecuencias relativas marginales:

$$f_{i\cdot} = \sum_{j=1}^p f_{ij} \frac{n_{i\cdot}}{n}$$

$$f_{\cdot j} = \sum_{i=1}^k f_{ij} \frac{n_{\cdot j}}{n}$$

$X \setminus Y$	$y_1$	$y_2$	$\dots$	$y_j$	$\dots$	$y_p$	
$x_1$	$f_{11}$	$f_{12}$	$\dots$	$f_{1j}$	$\dots$	$f_{1p}$	$f_{1\cdot}$
$x_2$	$f_{21}$	$f_{22}$	$\dots$	$f_{2j}$	$\dots$	$f_{2p}$	$f_{2\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_i$	$f_{i1}$	$f_{i2}$	$\dots$	$f_{ij}$	$\dots$	$f_{ip}$	$f_{i\cdot}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
$x_k$	$f_{k1}$	$f_{k2}$	$\dots$	$f_{kj}$	$\dots$	$f_{kp}$	$f_{k\cdot}$
	$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot j}$	$\dots$	$f_{\cdot p}$	1

**Propiedad**

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

### 1.1.3. Distribución condicionada

De todos los elementos de la población,  $n$ , podemos estar interesados, en un momento dado, en un conjunto más pequeño que está formado por aquellos elementos que han presentado la modalidad  $y_j$ , para algún  $j = 1, \dots, p$ . El número de elementos de este conjunto sabemos que es  $n_{\cdot j}$ . La variable  $X$  definida sobre este conjunto se denomina variable condicionada y se suele denotar mediante  $X|y_j$  o bien  $X|Y = y_j$ .

La distribución de frecuencias absolutas de esta nueva variable es exactamente la columna  $j$  de la tabla. Por tanto sus frecuencias relativas, que denominaremos frecuencias relativas condicionadas son

$$f_j^i = \frac{n_{ij}}{n_{.j}} \text{ para todo } i = 1, \dots, k$$

Del mismo modo, la variable condicionada  $Y|x_i$  cuya distribución de frecuencias relativas condicionadas es:

$$f_i^j = \frac{n_{ij}}{n_{i.}} \text{ para todo } j = 1, \dots, p$$

## 1.2. Representaciones gráficas

Dependerá del tipo de variables con las que estemos trabajando y de si están agrupadas o no.

- Cualitativas , Cuantitativas discretas:
- Cuantitativas continuas:

### 1.2.1. Diagramas de barras en 3 dimensiones para datos agrupados

**Ejemplo (continua/continua), en datos agrupados:** A continuación se muestran los datos recogidos de medir la longitud (X en mm) y el peso (Y en gr) de una muestra de 117 tornillos producidos por una máquina.

x/y	40-60	60-80	80-100	100-120	Total
140-160	4	0	0	0	4
160-180	14	60	2	0	76
180-200	0	20	16	1	37
<b>Total</b>	18	80	18	1	117

### 1.2.2. Diagrama de dispersión

La representación gráfica más útil para mostrar el tipo de relación entre dos variables continuas sin agrupar es el diagrama de dispersión, que representa cada par de puntos  $(x_i, y_i)$ ,  $i = 1, \dots, n$ , en un plano cartesiano.

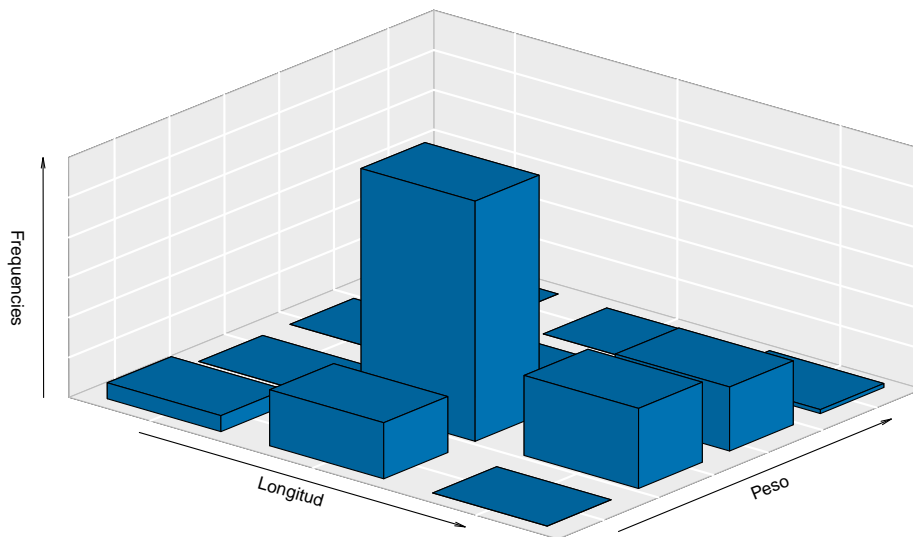


Figura 1: Histograma en 3 dimensiones

Supongamos que tenemos las 117 mediciones de los tornillos sin agrupar. La siguiente figura muestra un gráfico de dispersión, que nos permite intuir la relación entre el Peso y la Longitud.

### 1.3. Medidas de dependencia lineal

Las dos medidas más utilizadas para cuantificar el grado y el sentido de la dependencia lineal son: la covarianza y la correlación.

#### 1.3.1. Covarianza

Nos indica si la relación entre las variables es positiva o negativa. Su magnitud depende de las unidades.

Cuando los datos están agrupados en forma de tabla:

$$S_{xy} = \sum_i \sum_j f_{ij} (x_i - \bar{x})(y_j - \bar{y}) = \sum_i \sum_j f_{ij} x_i y_j - \bar{x} \bar{y}$$

*Ejercicio:* Supongamos la siguiente tabla donde  $X = N^\circ$  de hermanos e  $Y = N^\circ$  de suspensos

<https://www.youtube.com/watch?v=zWXwwFJwCZE>

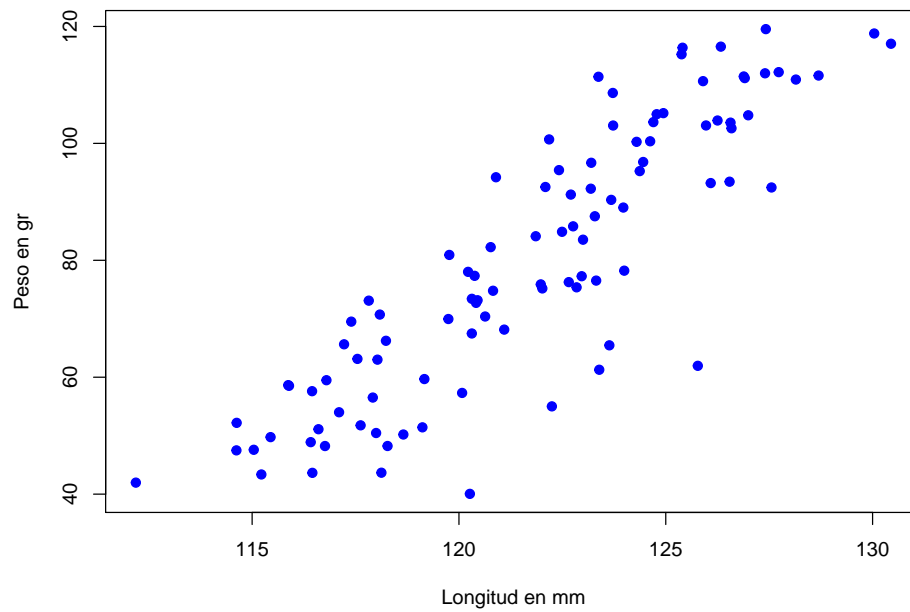


Figura 2: Gráfico de dispersión Longitud/Peso

$X$	$Y$			
	0	1	2	3
0	4	5	2	1
1	2	5	4	2
2	3	5	3	3
3	2	4	4	1

Cálculo de marginales

$X$	$Y$				$f_i$
	0	1	2	3	
0	4	5	2	1	<b>12</b>
1	2	5	4	2	<b>13</b>
2	3	5	3	3	<b>14</b>
3	2	4	4	1	<b>11</b>
$f_j$	<b>11</b>	<b>19</b>	<b>13</b>	<b>7</b>	<b>n=50</b>

Cálculo de medias marginales

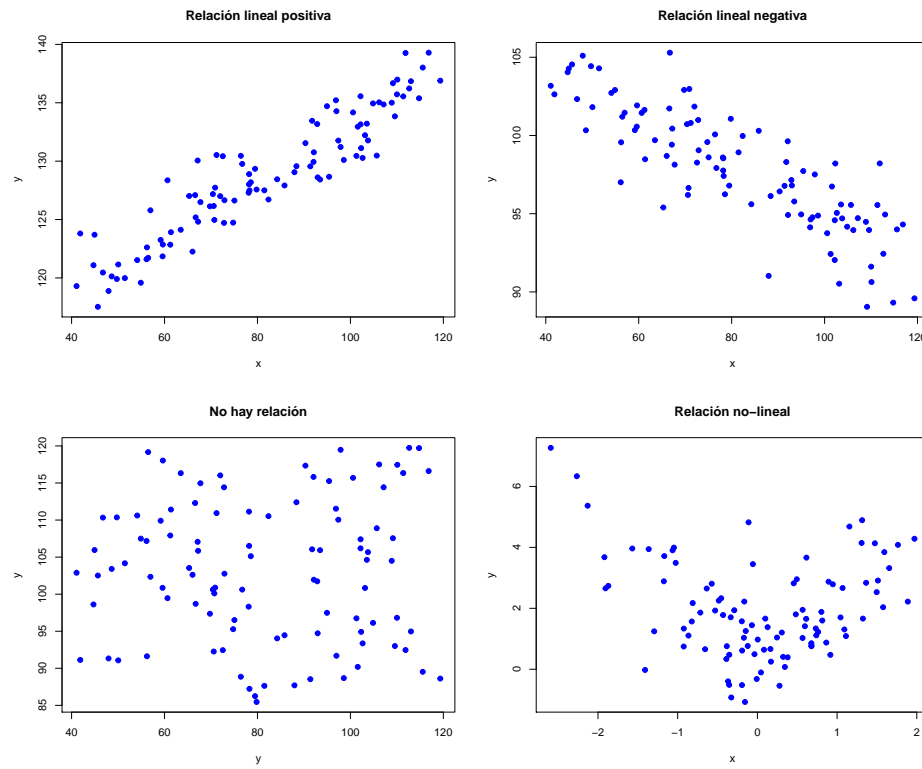


Figura 3: Tipos de relaciones X/Y en diagramas de dispersión



Media de  $X$ 

$x_i$	$f_i$	$x_i f_i$
0	12	0
1	13	13
2	14	28
3	11	33
$n = 50$		$\sum_i x_i f_i = 74$

Media de  $Y$ 

$y_j$	$f_j$	$y_j f_j$
0	11	0
1	19	19
2	13	26
3	7	21
$n = 50$		$\sum_j y_j f_j = 66$

$$\bar{x} = \frac{\sum_i x_i f_i}{n} = 74/50 = 1.48 \quad \bar{y} = \frac{\sum_j y_j f_j}{n} = 66/50 = 1.32$$

$$S_{xy} = \frac{\sum f_{ij} x_i y_j}{n} - \bar{x} \bar{y} = \frac{0 \cdot 0 \cdot 4 + 0 \cdot 0 \cdot 5 + 0 \cdot 2 \cdot 2 + 0 \cdot 3 + \dots + 3 \cdot 3 \cdot 1}{50} - 1.48 \times 1.32 =$$

$$= \frac{104}{50} - 1.9536 = 0.1264$$

Varianza de  $X$ 

$x_i$	$f_i$	$x_i^2 f_i$
0	12	0
1	13	13
2	14	56
3	11	99
$n = 50$		$\sum_i x_i^2 f_i = 168$

$$s_x^2 = \frac{\sum_i x_i^2 f_i}{n} - \bar{x}^2 = 168/50 - 1.48^2 = 1.1696$$

Varianza de  $Y$ 

$y_j$	$f_j$	$y_j^2 f_j$
0	11	0
1	19	19
2	13	52
3	7	63
$n = 50$		$\sum_j y_j^2 f_j = 134$

$$s_y^2 = \frac{\sum_i y_i^2 f_i}{n} - \bar{y}^2 = 134/50 - 1.32^2 = 0.93786$$