

Análisis Exploratorio de Datos, 2009-2010

Repaso de Estadística Descriptiva

1.1 Variables

1. Tipos de Variable según su significado:

- (a) **Cualitativas:** expresan cualidades no numéricas aunque puedan etiquetarse con números.
- (b) **Ordinales:** expresan orden.
- (c) **Cuantitativas:** expresan cantidades.

2. Soporte de una variable x : conjunto A_x donde toma valores. Ejemplos:

- (a) Variable **cualitativa**: $A_x = \{A, B, C\}$, $A_x = \{\text{mujer, hombre}\}$, $A_x = \{1, 2, 3\}$ (entendido como "tipo 1", "tipo 2", "tipo 2").
- (b) Variable **ordinal**: $A_x = \{1, 2, 3\}$ (entendido como "primero", "segundo", "tercero").
- (c) Variable **cuantitativa**: $A_x = \{1, 2, 3\} \subset \mathbb{N}$ (valores entendidos como cantidades), $A_x = \mathbb{N}$, $A_x = [3, 2) \subset \mathbb{R}$, $A_x = \mathbb{R}$, etc.

3. Tipos de Variable según su soporte A_x

- (a) **Discretas:** si A_x es finito, $\#A_x < \infty$, o infinito numerable ($\exists \phi : A_x \rightarrow \mathbb{N}$ inyectiva: los elementos de A_x pueden contarse o numerarse). Ejemplos: $A_x = \{\text{rojo, azul}\}$, $A_x = \{1, 2, 3\}$, $A_x = \mathbb{N}$, $A_x = \{n\pi : n \in \mathbb{N}\}$.
- (b) **Continuas:** si A_x es infinito no numerable. Ejemplos: $A_x = [a, b] \subset \mathbb{R}$, $A_x = \mathbb{R}$.
- (c) **Mixtas:** poseen parte discreta y parte continua. Ejemplo: $A_x = [a, b] \cup \{1\}$.

1.2 Distribución Unidimensional de Frecuencias

1. Distribución Unidimensional de Frecuencias: es la expresión tabulada de los valores que ha tomado una variable (**datos tabulados**).

2. Tipología según la presentación de los datos:

(a) Datos no agrupados:

$$X \equiv \{(x_i, n_i)\}_{i=1}^v$$

Caso particular: $n_i = 1, \forall i = 1, \dots, v$

(b) Datos agrupados:

$$X \equiv \{(I_i, n_i)\}_{i=1}^v$$

con $I_1 = [L_0, L_1]$, $I_i = (L_{i-1}, L_i]$, $i = 2, \dots, v$, intervalos.

(Cuando el número de valores que puede tomar x es muy elevado (en general variables continuas), se agrupan en intervalos $I_i = (L_{i-1}, L_i]$, $i = 1 : v$. (No se conocen o no son relevantes los verdaderos valores x_i)).

3. **Conceptos** asociados a una distribución de frecuencias:

- (a) **Número total de datos:** $n = \sum_{i=1}^v n_i$.
- (b) **Frecuencia absoluta:** n_i .
- (c) **Frecuencia relativa:** $f_i = n_i/n$, $i = 1, \dots, v$. Entonces: $\sum_{i=1}^v f_i = 1$. Se expresa en tantos por uno o en %.
- (d) **Frecuencia absoluta acumulada:** $N_i = \sum_{j=1}^i n_j$, $i = 1, \dots, v$. Entonces: $N_v = n$.
- (e) **Frecuencia relativa acumulada:** $F_i = \sum_{j=1}^i f_j$, $i = 1, \dots, v$. Entonces, $F_v = 1$.
(Pensar por qué las frecuencias acumuladas no se utilizan para variables categóricas).
- (f) **Recorrido o rango:** $R = \max_i \{x_i\} - \min_i \{x_i\}$. Si los datos están agrupados $\min_i \{x_i\} \equiv L_0$ y $\max_i \{x_i\} \equiv L_v$.
- (g) Sólo para datos agrupados:
1. **Amplitud** de un intervalo: $\ell_i = L_i - L_{i-1}$, $i = 1, \dots, v$
 2. **Marca de clase** de I_i : $\bar{x}_i = (L_{i-1} + L_i)/2$, $i = 1, \dots, v$.
 3. **Densidad de frecuencia** en I_i : $d_i = n_i/\ell_i$, $i = 1, \dots, v$.

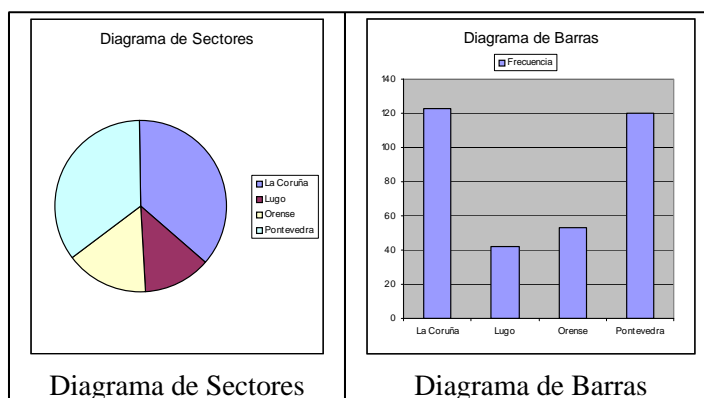
4. **Disposición** de la tabla de frecuencias. Por ejemplo, datos agrupados:

$I_i = (L_{i-1}, L_i]$	n_i	N_i	f_i	F_i	ℓ_i	\bar{x}_i	d_i
$[L_0, L_1]$	n_1	n_1	n_1/n	N_1/n	$L_1 - L_0$	$\bar{x}_1 = \frac{L_0 + L_1}{2}$	$d_1 = n_1/\ell_1$
$(L_1, L_2]$	n_2	$n_1 + n_2$	n_2/n	N_2/n	$L_2 - L_1$	$\bar{x}_2 = \frac{L_1 + L_2}{2}$	$d_2 = n_2/\ell_2$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$(L_{v-1}, L_v]$	n_v	n	n_v/n	1	$L_v - L_{v-1}$	$\bar{x}_v = \frac{L_{v-1} + L_v}{2}$	$d_v = n_v/\ell_v$
Total	n		1				

1.3 **Representaciones Gráficas de Distribuciones de Frecuencias**

1. Variables Categóricas.

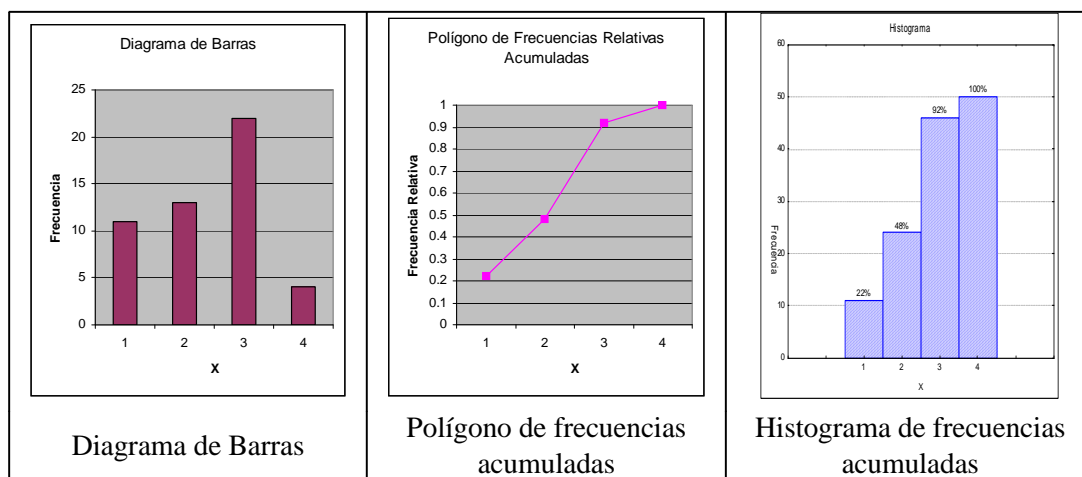
- (a) Diagrama de Sectores (Tarta): el área es proporcional a la frecuencia.
- (b) Diagrama de Rectángulos (Barras): la altura es la frecuencia (absoluta o relativa).



2. Variables Cuantitativas.

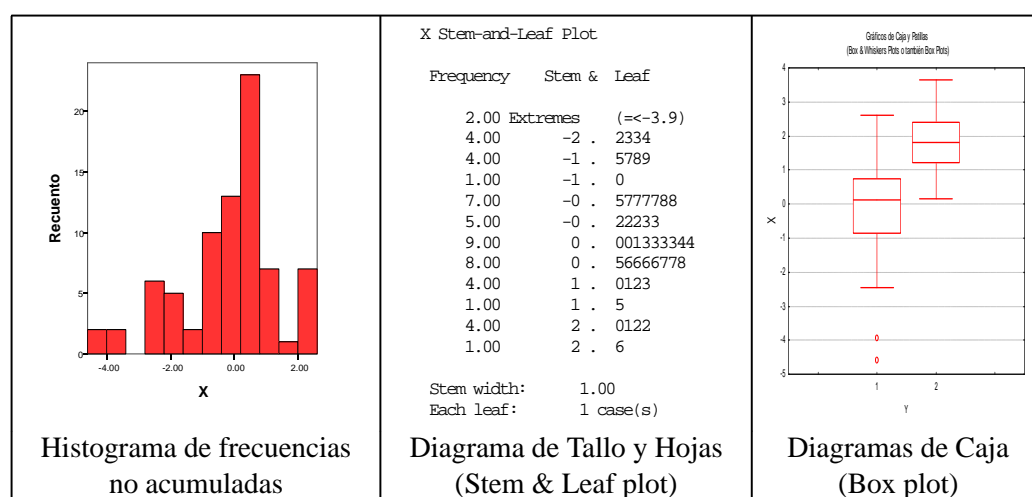
(a) Datos No Agrupados.

1. Diagrama de Barras: la altura es la frecuencia (absoluta o relativa).
2. Polígono de frecuencias (absolutas o relativas).
3. Diagrama o histograma acumulativo de frecuencias.



(b) Datos Agrupados.

1. Histograma de frecuencias (absolutas o relativas). La base de cada barra es ℓ_i (centrada en la marca de clase \bar{x}_i) y la altura suele ser la densidad de frecuencia d_i con lo que el área de la barra es $base \times altura = n_i$ (aunque Excel no lo hace así).
2. Histograma de frecuencias acumuladas (absolutas o relativas).
3. Polígono de frecuencias absolutas o relativas.
4. Otros Gráficos: Diagramas de Caja y Diagrama de Tallo y Hojas.



1.4 Resumen de Datos. Propiedades de una Distribución de Frecuencias

1. En adelante, se suponen variables numéricas: $A_x \subset \mathbb{R}$.

2. Medidas o propiedades de una Distribución de Frecuencias:

- (a) Medidas de Posición.
- (b) Medidas de Dispersión.
- (c) Medidas de Forma.

En todo lo que sigue, para datos agrupados, sustituir x_i por las marcas de clase \bar{x}_i .

1.4.1 Medidas de Posición

1. Media aritmética:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^v x_i n_i \text{ ó } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ si } n_i = 1 \forall i$$

- (a) En general, **media aritmética ponderada**:

$$\bar{x}_p = \frac{1}{\sum_{i=1}^v w_i} \sum_{i=1}^v x_i w_i$$

- (b) Propiedades de la media aritmética:

1. Transformación lineal (traslación y cambio de escala): si $Y = aX + b \Rightarrow \bar{y} = a\bar{x} + b$.
Demostrar.
2. $\sum_{i=1}^v (x_i - \bar{x}) n_i = 0$
3. $\bar{x} = \arg \min_C \sum_{i=1}^n (x_i - C)^2$
4. Muy sensible a datos extremos o atípicos.

2. Media geométrica:

$$G = \sqrt[n]{x_1^{n_1} \times \dots \times x_v^{n_v}} = \sqrt[n]{\prod_{i=1}^v x_i^{n_i}}$$

- (a) Propiedades:

1. $\log G = \frac{1}{n} \sum_{i=1}^v n_i \log x_i = \overline{\log x}$
2. Para valores sólo positivos.

- (b) **Ejemplo:** un capital C_0 colocado a n períodos con tipos de interés i_1, \dots, i_n . Calcular el tipo medio, es decir, el tipo i tal que $C_n = C_0(1+i)^n$.

Se verifica que:

$$\begin{aligned} C_1 &= C_0(1+i_1) \\ &\vdots \\ C_n &= C_0(1+i_1) \cdots (1+i_n) \end{aligned}$$

Luego, $C_n = C_0(1+i)^n = C_0(1+i_1) \cdots (1+i_n) \Rightarrow (1+i) = \sqrt[n]{(1+i_1) \cdots (1+i_n)}$ media geométrica

3. Media Armónica: Para valores tabulados:

$$H = \frac{1}{\overline{(1/x)}} = \frac{1}{\frac{1}{n} \sum_{i=1}^v \frac{1}{x_i} n_i} = \frac{1}{\sum_{i=1}^v \frac{1}{x_i} f_i}$$

y para valores no tabulados:

$$H = \frac{1}{\overline{(1/x)}} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}}$$

- (a) Problemas con valores pequeños. Y si algún $x_i = 0$, sale cero.
- (b) Vale para medias de rendimientos, tasas, velocidades.
- (c) **Ejemplo 1:** n fábricas que producen cada una m chips con una productividad de $x_i = m/t_i$ chips por trabajador, $i = 1, \dots, n$, donde t_i es el número de trabajadores de la fábrica i . Obtener la productividad media.

1. Si quisiésemos calcular la media aritmética de la productividad:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n \frac{m}{t_i} = \frac{m}{n} \sum_{i=1}^n \frac{1}{t_i} = m \overline{(1/t)}$$

2. Mediante la media armónica:

$$H_x = \frac{1}{(1/x)} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{t_i}{m}} = \frac{mn}{\sum_{i=1}^n t_i} = \text{Prod. media}$$

- (d) **Ejemplo 2:** n fábricas con t trabajadores que producen m_i chips con una productividad de $x_i = m_i/t$, $i = 1, \dots, n$. Obtener la productividad media.

1. Si quisiésemos calcular la media aritmética de la productividad:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n \frac{m_i}{t} = \frac{\sum_{i=1}^n m_i}{nt} = \text{Prod. media}$$

2. Mediante la media armónica:

$$H_x = \frac{1}{(1/x)} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{t}{m_i}} = \frac{nt}{\sum_{i=1}^n \frac{1}{m_i}}$$

- (e) **Ejemplo 3:** n fábricas que producen m_i chips con una productividad de $x_i = m_i/t_i$ chips por trabajador, $i = 1, \dots, n$, donde t_i es el número de trabajadores de la fábrica i . Obtener la productividad media.

1. La productividad media será ahora:

$$\text{prod. media} = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n t_i}$$

2. Si quisiésemos calcular la media aritmética de la productividad:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n \frac{m_i}{t_i}$$

3. Mediante la media armónica:

$$H_x = \frac{1}{(1/x)} = \frac{1}{\frac{1}{n} \sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\sum_{i=1}^n \frac{t_i}{m_i}}$$

4. Mientras que la productividad media es:

$$\text{prod. media} = \frac{\sum_{i=1}^n m_i}{\sum_{i=1}^n t_i} = \frac{M}{\sum_{i=1}^n t_i} = \frac{1}{\frac{1}{M} \sum_{i=1}^n \frac{m_i}{x_i}} = \frac{1}{\frac{1}{M} \sum_{i=1}^n \frac{1}{x_i} m_i}$$

que es una **media armónica ponderada** por los valores m_i .

4. **Relación Media Aritmética, Geométrica y Armónica:**

$$H \leq G \leq \bar{x}$$

5. **Mediana.** El valor M_e que deja a su izquierda la misma frecuencia que a su derecha. Método de cálculo:

(a) **Datos No Agrupados.** Sea F_i tal que $i = \min_{j=1:v} \{j : F_j \geq 1/2\}$:

$$\text{a) } F_i > 1/2 \Rightarrow M_e \doteq x_i$$

$$\text{b) } F_i = 1/2 \Rightarrow \begin{cases} M_e \doteq \frac{x_i + x_{i+1}}{2} & \text{si } x \text{ es continua} \\ M_e \doteq \{x_i, x_{i+1}\} & \text{si } x \text{ es discreta} \end{cases}$$

(b) **Datos Agrupados.** Sea F_i tal que $i = \min_{j=1:v} \{j : F_j \geq 1/2\}$:

$$\text{a) } F_i = 1/2 \Rightarrow M_e \doteq L_i \text{ (límite superior intervalo } I_i)$$

$$\text{b) } F_i > 1/2 \Rightarrow M_e = L_{i-1} + l \text{ donde:}$$

$$\begin{aligned} \frac{\text{longitud de } I_i}{\text{frecuencia en } I_i} &= \frac{L_i - L_{i-1}}{f_i} = \frac{\text{incremento longitud}}{\text{incremento frecuencia}} = \frac{l}{1/2 - F_{i-1}} \\ \Rightarrow l &= (1/2 - F_{i-1}) \frac{(L_i - L_{i-1})}{f_i} \end{aligned}$$

(c) Ventajas: insensible a valores extremos (más robusta).

6. **Moda.** El valor que más se repite. $M_o = x_i$ (o \bar{x}_i para datos agrupados) con $i = \arg \max_{j=1:v} \{n_j\}$. (Para datos agrupados, existen definiciones más complejas).

7. **Cuantiles.** Cuantil x_p con $p \in [0, 1]$: valor que deja una frecuencia relativa p a izquierda y a lo sumo una frecuencia $1 - p$ a derecha. Se tiene que $x_{1/2} = M_e$.

(a) **Cuantiles de uso frecuente:** Si se divide la frecuencia relativa (absoluta) en C partes iguales, los cuantiles son $\{x_{q/C} \text{ tal que } q = 1, 2, 3, \dots, C\}$.

1. $C = 4$: cuartiles.

2. $C = 10$: deciles.

3. $C = 100$: percentiles.

(b) Cálculo: Sea C el número de partes en que se divide la frecuencia y q el orden del cuantil, sea $i = \min_{j=1:v} \{j : F_j \geq p = q/C\}$:

1. Datos No Agrupados:

$$\text{a) } F_i > q/C \Rightarrow x_{q,C} \doteq x_i$$

$$\text{b) } F_i = q/C \Rightarrow \begin{cases} x_{q,C} \doteq \frac{x_i + x_{i+1}}{2} & \text{si } x \text{ es continua} \\ x_{q,C} \doteq \{x_i, x_{i+1}\} & \text{si } x \text{ es discreta} \end{cases}$$

2. Datos Agrupados:

$$\text{a) } F_i = q/C \Rightarrow x_{q,C} \doteq L_i \text{ (límite superior intervalo } I_i)$$

$$\text{b) } F_i > q/C \Rightarrow x_{q,C} = L_{i-1} + l :$$

$$\begin{aligned} \frac{\text{longitud de } I_i}{\text{frecuencia en } I_i} &= \frac{L_i - L_{i-1}}{f_i} = \frac{\text{incremento longitud}}{\text{incremento frecuencia}} = \frac{l}{q/C - F_{i-1}} \\ \Rightarrow l &= (q/C - F_{i-1}) \frac{L_i - L_{i-1}}{f_i} \end{aligned}$$

1.4.2 Medidas de Dispersión. Absolutas y relativas:

1. Absolutas:

(a) **Desviación absoluta** con respecto a la media:

$$D_{\bar{x}} = \frac{1}{n} \sum_{i=1}^v |x_i - \bar{x}| n_i$$

(b) **Varianza** (muestral) y **desviación típica** (muestral):

$$S_X^2 = \frac{1}{n} \sum_{i=1}^v (x_i - \bar{x})^2 n_i = \frac{1}{n} \sum_{i=1}^v x_i^2 n_i - \bar{x}^2$$

$$S_X = \sqrt{\frac{1}{n} \sum_{i=1}^v (x_i - \bar{x})^2 n_i}$$

Propiedades de S_X^2 :

1. $S_X^2 \geq 0$
2. $S_X^2 = \min_C \{f(C) = \frac{1}{n} \sum_{i=1}^v (x_i - C)^2 n_i\} = f(\bar{x})$

$$0 = f'(C) = -\frac{2}{n} \sum_{i=1}^v (x_i - C) n_i \Rightarrow C = \bar{x}$$

$$f''(\bar{x}) = \frac{2}{n} \sum_{i=1}^v n_i = 2 > 0 \Rightarrow \bar{x} \text{ mínimo}$$

(c) **Cuasivarianza** (muestral) y **cuasidesviación típica** (muestral):

$$s_X^2 = \frac{1}{n-1} \sum_{i=1}^v (x_i - \bar{x})^2 n_i = \frac{n}{n-1} S_X^2$$

$$s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^v (x_i - \bar{x})^2 n_i}$$

(d) **Recorrido, rango** o intervalo de variación: $R = \max_{i=1:v} \{x_i\} - \min_{i=1:v} \{x_i\}$.

(e) Intervalos intercuantílicos:

1. Intervalo intercuartílico: $IQ = Q_3 - Q_1$ con $Q_j = x_{j/4}$
2. Intervalo intercuartílico relativo: $IQR = (Q_3 - Q_1)/M_e$.

2. Diagrama de Caja y Patillas (Box & whiskers plot).

(a) Límite inferior y superior Caja:

$$\text{Límite Inferior} = Q_1$$

$$\text{Límite Superior} = Q_3$$

(b) Límite inferior y superior Patillas:

$$\text{Límite Inferior} = \max\left\{\min(x_i), Q_1 - 1.5 \frac{Q_3 - Q_1}{2}\right\}$$

$$\text{Límite Superior} = \min\left\{\max(x_i), Q_3 + 1.5 \frac{Q_3 - Q_1}{2}\right\}$$

Los datos que quedan fuera son *outliers* o datos atípicos representados por cruces.

Los datos *extremos* son los que quedan fuera de $Q_1 - 3 \frac{Q_3 - Q_1}{2}$ o $Q_3 + 3 \frac{Q_3 - Q_1}{2}$

3. **Medidas de Dispersión Relativas.** Permiten la comparación de la variabilidad de variables diferentes.

(a) **Recorrido relativo:**

$$RR = \frac{R}{\bar{x}}$$

(b) **Coeficiente de variación de Pearson:**

$$CV_X = \frac{S_X}{|\bar{x}|}$$

4. **Ejemplo.** La siguiente tabla recoge los datos tabulados correspondientes a los ingresos anuales en millones de euros de las empresas de dos sectores industriales.

$[L_{i-1}, L_i]$	[40,80]	(80,110]	(110,150]	(150,200]
Sector I	10	25	18	7
Sector II	20	21	10	15

Se pide: a) Media, mediana, varianza, desviación típica y coeficiente de variación de Pearson de los ingresos anuales de las empresas de ambos sectores. b) En qué sector las empresas son más heterogéneas; Para el Sector I: c) el nivel de ingresos a partir del cuál se encuentra el 25% de las empresas del sector con mayores ingresos anuales, d) la proporción de empresas con ingresos anuales mayores que 160 millones de euros.

Solución.

		Sector I			Sector II		
$[L_{i-1}, L_i]$	\bar{x}_i	n_i	f_i	F_i	n_i	f_i	F_i
[40,80]	60	10	0.1667	0.1667	20	0.3030	0.3030
(80,110]	95	25	0.4167	0.5834	21	0.3182	0.6212
(110,150]	130	18	0.3000	0.8834	10	0.1515	0.7727
(150,200]	175	7	0.1166	1.0000	15	0.2273	1.0000
		60	1.0000		66	1.0000	

(a) Para el Sector I:

$$1. \bar{x} = \frac{1}{n} \sum_{i=1}^4 \bar{x}_i n_i = \frac{1}{60} (60 \cdot 10 + 95 \cdot 25 + 130 \cdot 18 + 175 \cdot 7) = 109.0$$

$$2. Me = L_{i-1} + l = 80 + \frac{110-80}{0.4167} \cdot (0.5 - 0.1667) = 99.996$$

$$3. S^2 = \frac{1}{n} \sum_{i=1}^4 \bar{x}_i^2 n_i - \bar{x}^2 = \frac{1}{60} (60^2 \cdot 10 + 95^2 \cdot 25 + 130^2 \cdot 18 + 175^2 \cdot 7) - 109.0^2 = 1122.3$$

$$4. S = \sqrt{1122.3} = 33.501$$

$$5. CV_X = \frac{S_X}{|\bar{x}|} = \frac{33.501}{109.0} = 0.3074$$

Para el Sector II:

1. $\bar{x} = \frac{1}{n} \sum_{i=1}^4 \bar{x}_i n_i = \frac{1}{66} (60 \cdot 20 + 95 \cdot 21 + 130 \cdot 10 + 175 \cdot 15) = 107.88$
 2. $Me = L_{i-1} + l = 80 + \frac{110-80}{0.3182} \cdot (0.5 - 0.3030) = 98.573$
 3. $S^2 = \frac{1}{n} \sum_{i=1}^4 \bar{x}_i^2 n_i - \bar{x}^2 = \frac{1}{66} (60^2 \cdot 20 + 95^2 \cdot 21 + 130^2 \cdot 10 + 175^2 \cdot 15) - 107.88^2 = 1845.2$
 4. $S = \sqrt{1845.2} = 42.956$
 5. $CV_X = \frac{S_X}{|\bar{x}|} = \frac{42.956}{107.88} = 0.3982$
- (b) Comparando el coeficiente de variación de Pearson, se ve que el Sector II es más heterogéneo (presenta más variabilidad relativa).
- (c) $x_{0.75} = 110 + l = 110 + \frac{40}{0.3} \cdot (0.75 - 0.5834) = 132.21$
- (d) Proporción pedida = $\frac{0.1166}{50} \cdot (200 - 160) = 0.09328$.

1.4.3 Momentos

1. **Momentos no centrados u ordinarios (centrados respecto al origen)** de orden r :

$$a_r = \frac{1}{n} \sum_{i=1}^v x_i^r n_i = \sum_{i=1}^v x_i^r f_i$$

2. **Momentos centrados (respecto a la media)** de orden r :

$$m_r = \frac{1}{n} \sum_{i=1}^v (x_i - \bar{x})^r n_i = \sum_{i=1}^v (x_i - \bar{x})^r f_i$$

3. Propiedades:

- (a) $a_1 = \bar{x}$ y $m_1 = 0$.
- (b) m_2 es la **varianza**.
- (c) Momentos centrados en función de momentos no centrados. Por la fórmula de Newton:

$$\begin{aligned} (x_i - \bar{x})^r &= \sum_{j=0}^r (-1)^j \binom{r}{j} x_i^{r-j} \bar{x}^j \\ m_r &= \frac{1}{n} \sum_{i=1}^v (x_i - \bar{x})^r n_i = \frac{1}{n} \sum_{i=1}^v n_i \sum_{j=0}^r (-1)^j \binom{r}{j} x_i^{r-j} \bar{x}^j \\ &= \sum_{j=0}^r (-1)^j \binom{r}{j} \bar{x}^j \frac{1}{n} \sum_{i=1}^v n_i x_i^{r-j} = \sum_{j=0}^r (-1)^j \binom{r}{j} \bar{x}^j a_{r-j} \end{aligned}$$

- (d) En particular:

$$\begin{aligned} m_2 &= a_2 - \bar{x}^2 \\ m_3 &= a_3 - 3a_2\bar{x} + 3\bar{x}^2 a_1 - \bar{x}^3 = a_3 - 3a_2\bar{x} + 3a_1^3 - a_1^3 = a_3 - 3a_2\bar{x} + 2a_1^3 \\ m_4 &= a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 3a_1^4 \end{aligned}$$

- (e) Si $Y = aX + b \Rightarrow m_{r,Y} = a^r m_{r,X}$

1.4.4 Medidas de forma

1. Asimetría:

(a) Coeficiente de Asimetría de Fisher:

$$g_1 = \frac{m_3}{S_X^3}$$

(b) Coeficiente de Asimetría de Pearson:

$$v = \frac{\bar{x} - M_o}{S_X}$$

(c) Si una distribución es simétrica, $g_1 \approx 0$ y $v \approx 0$ (no necesariamente a la inversa). Si $g_1 > 0$, $v > 0$ ($\bar{x} > M_o$) es asimétrica positiva o a la derecha y en caso contrario, a la izquierda. (g_1 vale para el caso de varias modas).

(d) Si una distribución es simétrica, $\bar{x} = M_e$ y si es unimodal, $\bar{x} = M_e = M_o$.

2. Curtosis o apuntamiento. Coeficiente de Fisher:

$$g_2 = \frac{m_4}{S_X^4} - 3$$

(a) Si $g_2 < 0$ es **platicúrtica** (colas más pesadas), $g_2 \approx 0$ **mesocúrtica** (parecida a la normal), y $g_2 > 0$ **leptocúrtica**.

(b) No confundir varianza con no apuntamiento. La curva gaussiana (campana de Gauss) siempre verifica $g_2 = 3$ independientemente de la varianza.

1.5 Distribuciones Bidimensionales de Frecuencias

- Ahora tratamos de pares de variables, (X, Y) , y todo puede extenderse a d variables (X_1, \dots, X_d) .

- Distribución bidimensional de frecuencias:**

Frecuencias	Datos No Agrupados	Datos Agrupados
Absolutas	$\{(x_i, y_j), n_{ij}\}_{i=1:u; j=1:v}$	$\{((L_{i-1}, L_i], (L_{j-1}, L_j]), n_{ij}\}_{i=1:u; j=1:v}$
Relativas	$\{(x_i, y_j), f_{ij}\}_{i=1:u; j=1:v}$	$\{((L_{i-1}, L_i], (L_{j-1}, L_j]), f_{ij}\}_{i=1:u; j=1:v}$

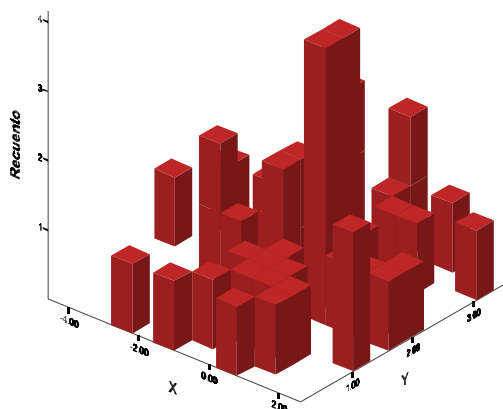
1. n_{ij} es la frecuencia absoluta asociada a (x_i, y_j) o a $((L_{i-1}, L_i], (L_{j-1}, L_j])$. Se verifica:

$$\sum_{i=1}^u \sum_{j=1}^v n_{ij} = n$$

2. $f_{ij} = n_{ij}/n$ es la frecuencia relativa asociada a lo mismo. Se verifica: $\sum_{i=1}^u \sum_{j=1}^v f_{ij} = 1$

- La distribución bidimensional de frecuencias suele disponerse en una tabla, la cuál, si las variables son cuantitativas, se denomina **tabla de correlación** y si son cualitativas o categóricas, se denomina **tabla de contingencia**. Se enfrentan las variables indicando las frecuencias (absolutas o relativas):

$X \backslash Y$	y_1	y_2	\dots	y_v	$n_{i\bullet} = \sum_{j=1}^v n_{ij}$
x_1	n_{11}	n_{12}	\dots	n_{1v}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	\dots	n_{2v}	$n_{2\bullet}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_u	n_{u1}	n_{u2}	\dots	n_{uv}	$n_{u\bullet}$
$n_{\bullet j} = \sum_{i=1}^u n_{ij}$	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet v}$	$n = \sum_{i=1}^u \sum_{j=1}^v n_{ij}$



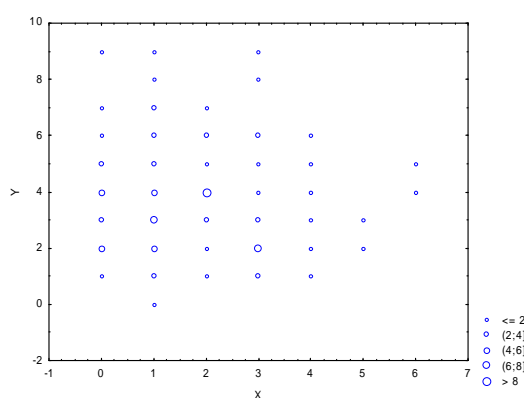
Histograma de frecuencias no acumuladas de una distribución bidimensional de frecuencias

- Otra forma de disponer los datos es la siguiente:

Datos No Agrupados		Datos Agrupados		
X	Y	X	Y	n_{ii}
x_1	y_1	$(L_0, L_1]$	$(L_0, L_1]$	n_1
\vdots	\vdots	\vdots	\vdots	\vdots
x_m	y_m	$(L_{m-1}, L_m]$	$(L_{m-1}, L_m]$	n_m

Nótese que en la tabla anterior m es el numero total de pares distintos (aunque los valores x_i o y_i pueden estar repetidos).

- Representación gráfica: gráfico de dispersión (scatterplot) o nube de puntos (con símbolos que representan la frecuencia).



- **Frecuencias marginales** (son las de las variables por separado):

– De X :

	Datos No Agrupados	Datos Agrupados	
Absolutas	$\{x_i, n_{i\bullet}\}_{i=1:u}$	$\{(L_{i-1}, L_i], n_{i\bullet}\}_{i=1:u}$	$n_{i\bullet} = \sum_{j=1}^v n_{ij}$
Relativas	$\{x_i, f_{i\bullet}\}_{i=1:u}$	$\{(L_{i-1}, L_i], f_{i\bullet}\}_{i=1:u}$	$f_{i\bullet} = \sum_{j=1}^v f_{ij}$

– De Y :

	Datos No Agrupados	Datos Agrupados	
Absolutas	$\{y_j, n_{\bullet j}\}_{j=1:v}$	$\{(L_{j-1}, L_j], n_{\bullet j}\}_{j=1:v}$	$n_{\bullet j} = \sum_{i=1}^u n_{ij}$
Relativas	$\{y_j, f_{\bullet j}\}_{j=1:v}$	$\{(L_{j-1}, L_j], f_{\bullet j}\}_{j=1:v}$	$f_{\bullet j} = \sum_{i=1}^u f_{ij}$

• **Frecuencias condicionadas.**

– De $X|Y$:

	Datos No Agrupados: $X Y = y_j$	Datos Agrupados: $X Y \in (L_{j-1}, L_j]$	
Absolutas	$\{x_i y_j, n_{ij}\}_{i=1:u}$	$\{(L_{i-1}, L_i] (L_{j-1}, L_j], n_{ij}\}_{i=1:u}$	$\sum_{i=1}^u n_{ij} = n_{\bullet j}$
Relativas	$\{x_i y_j, f_{i j}\}_{i=1:u}$	$\{(L_{i-1}, L_i] (L_{j-1}, L_j], f_{i j}\}_{i=1:u}$	$f_{i j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{f_{ij}}{f_{\bullet j}}$

– De $Y|X$:

	Datos No Agrupados: $Y X = x_i$	Datos Agrupados: $Y X \in (L_{i-1}, L_i]$	
Absolutas	$\{y_j x_i, n_{ij}\}_{j=1:v}$	$\{(L_{j-1}, L_j] (L_{i-1}, L_i], n_{ij}\}_{j=1:v}$	$\sum_{j=1}^v n_{ij} = n_{i\bullet}$
Relativas	$\{y_j, f_{j i}\}_{j=1:v}$	$\{(L_{j-1}, L_j] (L_{i-1}, L_i], f_{j i}\}_{j=1:v}$	$f_{j i} = \frac{n_{ij}}{n_{i\bullet}} = \frac{f_{ij}}{f_{i\bullet}}$

Estas distribuciones son auténticas distribuciones unidimensionales y admiten todos los tratamientos vistos en la sección anterior: representación gráfica, momentos, forma, etc.

- **Ejemplo:** La siguiente tabla expresa la distribución conjunta de las variables X : "Sexo" e Y : "Provincia" de los alumnos de una clase.

X/Y	C	L	O	P	$n_{i\bullet}$
H	3	2	7	13	25
M	1	0	6	22	29
$n_{\bullet j}$	4	2	13	35	54

Se pide: a) obtener utilizando frecuencias relativas: la distribución conjunta, la distribución marginal de la Y , y de la X , b) la distribución (condicionada) de la variable $X|Y = "O"$ y de la variable $Y|X = "M"$, c) la frecuencia relativa de $X = "H"|Y = "P"$ y de $Y = "O"|X = "H"$.

1. Distribución conjunta: es la misma tabla 2×3 dividiendo sus valores por 54, es decir $f_{ij} = n_{ij}/54$, $i = 1, 2$ y $j = 1, 2, 3$.

La distribución (marginal) de Y sale de la última fila:

Y	C	L	O	P
$f_{\cdot j}$	$\frac{4}{54}$	$\frac{2}{54}$	$\frac{13}{54}$	$\frac{35}{54}$

y la de X sale de la última columna:

X	$n_{i\cdot}$
H	$\frac{25}{54}$
M	$\frac{29}{54}$

2. La distribución (condicionada) de $X|Y = "O"$ sale de la tercera columna:

$X Y = "O"$	$n_{i 3}$
H	$\frac{7}{13}$
M	$\frac{6}{13}$

3. La distribución (condicionada) de $Y|X = "M"$ sale de la última fila:

$Y X = "M"$	C	L	O	P
$f_{j 2}$	$\frac{1}{29}$	$\frac{0}{29}$	$\frac{6}{29}$	$\frac{22}{29}$

4. Las frecuencias relativas pedidas son $f_{i=1|j=4} = 13/35 = 0.37143$ y $f_{j=3|i=1} = 7/25 = 0.28$, respectivamente.

1.5.1 Características de una variable bidimensional cuantitativa

- **Vector de medias:** (\bar{x}, \bar{y})

- **Covarianza:**

- Datos tabulados:

$$\text{Cov}(X, Y) = S_{XY} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v (x_i - \bar{x})(y_j - \bar{y})n_{ij} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v x_i y_j n_{ij} - \bar{x}\bar{y}$$

- Datos no tabulados (n observaciones de una variable bidimensional (X, Y) en los que no importan las repeticiones): $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$:

$$\text{Cov}(X, Y) = S_{XY} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x}\bar{y}$$

- Propiedades:

$$* S_{XX} = S_X^2, S_{YY} = S_Y^2.$$

$$* \left\{ \begin{array}{l} U = a_1 X + b_1 \\ V = a_2 Y + b_2 \end{array} \right\} \Rightarrow S_{UV} = a_1 a_2 S_{XY}$$

- La covarianza expresa el grado de relación lineal entre las variables, pero su valor no permite determinar la magnitud de dicha relación pues aquél depende de la escala utilizada en las variables. Para evitar esto se define:

- **Coefficiente de correlación lineal:**

$$r_{X,Y} = \frac{S_{XY}}{S_X S_Y} \in [-1, 1]$$

1. Cuanto mayor es el valor absoluto del coeficiente de correlación lineal mayor relación lineal existe entre las variables.
2. El signo indica el signo de dicha relación.
3. Si $r_{X,Y} = 0$ no hay relación lineal entre las variables, pero puede haber relación de tipo no lineal.

- **Matriz de varianzas-covarianzas:**

$$\text{Var}[(X, Y)] = \mathbf{S} = \begin{bmatrix} S_X^2 & S_{XY} \\ S_{XY} & S_Y^2 \end{bmatrix} = \frac{1}{n} \sum_{i=1}^v (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

donde $\mathbf{x}_i = (x_i, y_i)^T$ y $\bar{\mathbf{x}} = (\bar{x}, \bar{y})^T$. Al determinante $|\mathbf{S}|$ se le denomina **varianza generalizada de (X, Y)** .

1.5.2 Momentos Bidimensionales

Momentos Respecto al origen	Momentos Centrados respecto a la Media
$a_{r,s} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v x_i^r y_j^s n_{ij}$	$m_{r,s} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v (x_i - a_{1,0})^r (y_j - a_{0,1})^s n_{ij}$

- Ejemplos:

Momentos Respecto al origen	Momentos Centrados respecto a la Media
$a_{1,0} = \frac{1}{n} \sum_{i=1}^u x_i n_{i\bullet} = \bar{x}$ $a_{0,1} = \frac{1}{n} \sum_{j=1}^v y_j n_{\bullet j} = \bar{y}$	$m_{1,0} = \frac{1}{n} \sum_{i=1}^u (x_i - a_{1,0}) n_{i\bullet} = 0$ $m_{0,1} = \frac{1}{n} \sum_{j=1}^v (y_j - a_{0,1}) n_{\bullet j} = 0$
$a_{2,0} = \frac{1}{n} \sum_{i=1}^u x_i^2 n_{i\bullet}$	$m_{2,0} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v (x_i - a_{1,0})^2 n_{ij} = S_X^2 = a_{2,0} - a_{1,0}^2$
$a_{0,2} = \frac{1}{n} \sum_{j=1}^v y_j^2 n_{\bullet j}$	$m_{0,2} = \frac{1}{n} \sum_{j=1}^v \sum_{i=1}^u (y_j - a_{0,1})^2 n_{ij} = S_Y^2 = a_{0,2} - a_{0,1}^2$
$a_{1,1} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v x_i y_j n_{ij}$	$m_{1,1} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v (x_i - a_{1,0})(y_j - a_{0,1}) n_{ij} = S_{XY}$ $= a_{1,1} - a_{1,0} a_{0,1}$

1.6 Dependencia o Asociación Estadística entre Dos Variables

- Por la definición de frecuencia relativa condicionada:

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}} = \frac{n_{ij}/n}{n_{\bullet j}/n} = \frac{f_{ij}}{f_{\bullet j}} \Leftrightarrow f_{ij} = f_{i|j} f_{\bullet j}$$

Si $f_{i|j} = f_{i\bullet} \forall j \in \{1 : v\}$ significa que el valor y_j que tome Y no influye en la distribución de $X|Y = y_j$ que es igual que la distribución de la marginal de X :

$$\{x_i|y_j, f_{i|j}\} \equiv \{x_i, f_{i\bullet}\}$$

Entonces, $\forall i \in \{1 : u\}, \forall j \in \{1 : v\}$:

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

(Nótese que al ser $f_{ij} = f_{j|i} f_{i\bullet}$ se obtiene igualmente $f_{j|i} = f_{\bullet j}$ para la variable $Y|X = x_i$)

- En términos de las frecuencias absolutas, la condición anterior equivale a:

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n} \Leftrightarrow n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$$

Definición 1 Dos variables X, Y son **independientes** si y sólo si $\forall i \in \{1 : u\}, \forall j \in \{1 : v\}$:

$$f_{ij} = f_{i\bullet} f_{\bullet j} \quad (1.1)$$

o bien:

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n} \quad (1.2)$$

Definición 2 Si dos variables no son independientes, se dice que están asociadas estadísticamente (presentan **asociación** o **dependencia** estadística).

1.6.1 Asociación entre dos Variables Cuantitativas

Proposición 1 Si X, Y son variables cuantitativas, se verifica que:

1. $r_{XY} \neq 0 \iff$ existe algún grado de relación lineal entre X e Y (tanto mayor cuanto mayor sea $|r_{XY}|$).
2. X, Y con relación lineal exacta ($Y = a + bX$) $\Leftrightarrow |r_{X,Y}| = 1$.
3. X, Y independientes $\Rightarrow \text{Cov}(X, Y) = r_{X,Y} = 0$. **No** se verifica la inversa.

Como consecuencia, se tiene que:

1. Dos variables cuantitativas X, Y presentan una relación exacta o funcional si dado un valor de X , se obtiene el valor de Y de manera exacta como $Y = f(X)$ para alguna función f .
2. Si dos variables cuantitativas presentan relación funcional (de tipo lineal o no lineal), entonces, lógicamente, están asociadas estadísticamente (no son independientes).
3. Si dos variables cuantitativas están asociadas, no tiene por qué ser de manera exacta o funcional, sino que puede ser que estén asociadas sólo de manera estadística.
4. Si dos variables cuantitativas presentan asociación estadística, puede ser que dicha asociación no sea de tipo lineal.

5. La covarianza (la correlación) refleja una asociación de tipo lineal entre las variables, pero puede haber asociación de tipo no lineal que queda escondida para la covarianza.

Ejemplo 3 Para la tabla de correlación siguiente, identificar las distribuciones marginales y condicionadas, y evaluar la dependencia estadística entre X e Y .

$X \backslash Y$	1	2	3	$n_{i\bullet}$
2	1	4	1	6
3	2	4	2	8
4	1	2	1	4
$n_{\bullet j}$	4	10	4	18

Ejemplo 4 Determinar la independencia de X e Y en las siguientes tablas, calculando también la covarianza.

1)	$Y = X^2$; 2)	$Y = 2X$; 3)	$Y = X^2$				
	X/Y	0	1	$n_{i\bullet}$		X/Y	-2	0	2	$n_{i\bullet}$		X/Y	1	4	9	$n_{i\bullet}$
	-1	0	1	1		-1	1	0	0	1		1	1	0	0	1
	0	1	0	1		0	0	1	0	1		2	0	1	0	1
	1	0	1	1		1	0	0	1	1		3	0	0	1	1
$n_{\bullet j}$				1	2	3	$n_{\bullet j}$	1	1	1		3	$n_{\bullet j}$	1	1	1

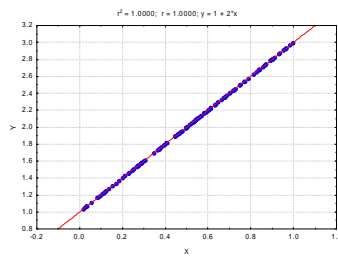
- $\bar{x} = 0$, $\bar{y} = 2/3$, $\text{Cov}(X, Y) = S_{XY} = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} = 0 - 0 = 0$, y $r_{X,Y} = 0$.
- $\bar{x} = 0$, $\bar{y} = 0$, $\text{Cov}(X, Y) = 3/4 - 0 = 3/4$ y

$$r_{X,Y} = \frac{4/3}{\sqrt{2/3} \sqrt{8/3}} = 1$$

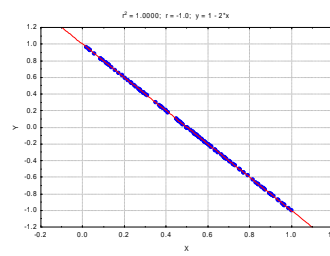
- $\bar{x} = 2$, $\bar{y} = 14/3$, $\text{Cov}(X, Y) = 36/3 - 2 \cdot 14/3 = 8/3$.

$$r_{X,Y} = \frac{8/3}{\sqrt{(1+4+9)/3 - 4} \sqrt{(1+16+81)/3 - (14/3)^2}} = 0.9897$$

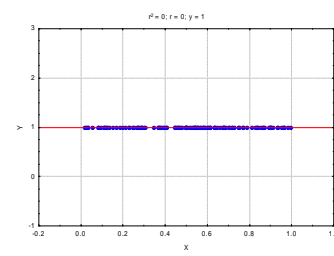
- O sea, dependiendo de los datos, el coeficiente de correlación lineal (la covarianza normalizada) a veces captura la relación no lineal. Pero, si hay relación lineal, entonces $|r_{X,Y}| = 1$ siempre.
- Las tablas anteriores son ejemplo de una relación de **dependencia funcional** entre X e Y . Si se verifica que $f_{ij} \neq f_{i\bullet} f_{\bullet j}$ entonces hay asociación (**dependencia estadística**) aunque pueda no haber asociación de tipo exacto o funcional.



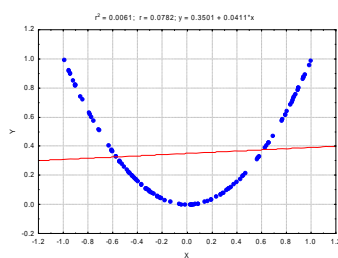
Relación funcional lineal
 $r = 1$



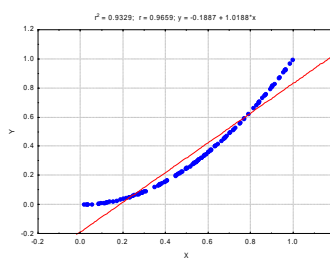
Relación funcional lineal
 $r = -1$



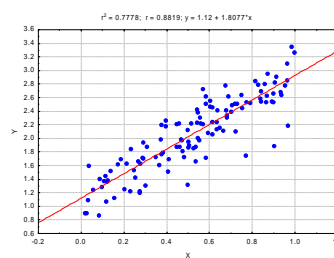
Ausencia de relación
 $r = 0$



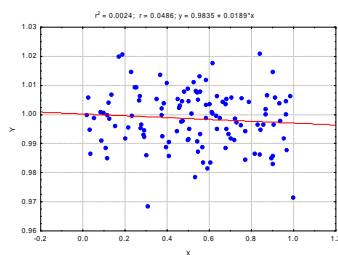
Relación funcional no lineal
 $r = 0.078$



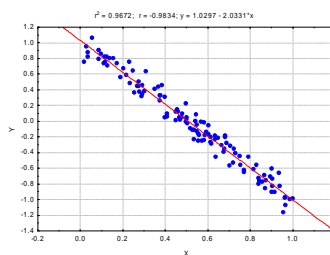
Relación funcional no lineal
 $r = 0.966$



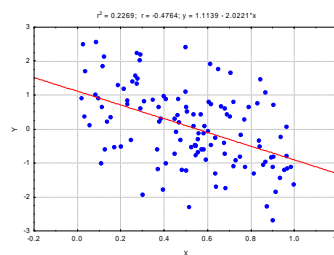
Asociación estadística lineal
 $r = 0.882$



Sin asociación estadística
 $r = -0.0486$



Asociación estadística lineal
 $r = -0.983$



Poca asociación lineal
 $r = -0.476$

1.6.2 Regresión Lineal

1. Dados n datos $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ de una variable bidimensional (X, Y) , se trata de determinar si existe una relación estadística de tipo lineal entre X e Y y, en su caso, estimar dicha relación.

Concretamente, se trata de ver si es cierto el siguiente modelo de regresión de Y sobre X :

$$y_i = ax_i + b + \varepsilon_i, i = 1, \dots, n$$

donde ε_i son errores aleatorios de pequeña magnitud (ya veremos que significa "pequeña magnitud").

2. A la variable Y se le denomina variable **dependiente o respuesta** (o output). A la variable X se le denomina variable **independiente o regresora** (o input).
3. El término b es el **término independiente**. El término a es la **pendiente de la recta** y representa el incremento de y por cada unidad adicional en x .

1.6.2.1 Método de los Mínimos Cuadrados

- El método de los mínimos cuadrados obtiene las estimaciones \hat{a}, \hat{b} como resultado de minimizar la suma de los errores al cuadrado:

$$\{\hat{a}, \hat{b}\} = \arg \min_{\alpha, \beta} \sum_{i=1}^n [y_i - (\alpha x_i + \beta)]^2$$

- La solución es:

$$\begin{aligned} \hat{b} &= \bar{y} - \hat{a}\bar{x} \\ \hat{a} &= \frac{S_{XY}}{S_X^2} = r_{XY} \frac{S_Y}{S_X} \end{aligned} \quad (1.3)$$

- Por tanto, la recta estimada es:

$$\hat{y} = \frac{S_{XY}}{S_X^2} x + \bar{y} - \frac{S_{XY}}{S_X^2} \bar{x}$$

Nótese que la recta estimada pasa por (\bar{x}, \bar{y}) ya que $\bar{y} = \hat{a}\bar{x} + b$ (debido a 1.3)

1.6.2.2 Bondad del Ajuste

- Los errores producidos por la recta de regresión son: $e_i = y_i - \hat{y}_i, i = 1, \dots, n$. Si estos errores son pequeños en magnitud el ajuste será bueno y la relación entre X e Y puede considerarse lineal.
- La **variabilidad (cantidad de información) de Y no explicada** por la regresión lineal se define como:

$$\text{VNE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

- La **variabilidad total** de la variable Y es una magnitud similar a la varianza S_Y^2 :

$$\text{VT} = \sum_{i=1}^n (y_i - \bar{y})^2 = nS_Y^2$$

- La **variabilidad de Y explicada por la regresión** se define como la diferencia y se puede demostrar que vale:

$$\text{VE} = \text{VT} - \text{VNE} = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

- Se define como **medida de la bondad del ajuste** el denominado **coeficiente de determinación R^2** :

$$\begin{aligned} R^2 &= \frac{\text{VE}}{\text{VT}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{S_Y^2}{S_Y^2} \\ &= 1 - \frac{\text{VNE}}{\text{VT}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{S_e^2}{S_Y^2} \end{aligned}$$

- Propiedades de R^2 .

- (a) $R^2 \in [0, 1]$.
- (b) Si $R^2 \approx 1$, la recta ajusta muy bien los datos. Puede decirse que hay una clara relación lineal entre X e Y .
- (c) Si R^2 pequeño, no existe relación lineal entre X e Y .
- (d) Puede que R^2 sea pequeño y exista relación estadística de otro tipo (no lineal).
- (e) Se verifica que:

$$R^2 = r_{XY}^2$$

Nota 1 Un mal ajuste (indicado por un R^2 pequeño) puede ser debido a:

1. Que la relación entre las variables no es esencialmente lineal (o es no lineal o no hay asociación estadística)
2. La relación entre las variables puede considerarse esencialmente lineal pero existe mucha dispersión (o ruido) en los datos (para cada valor de X , la variable Y posee mucha varianza).
3. Una combinación de las dos razones anteriores.

1.6.3 Regresión Lineal Múltiple

- Si interesa estudiar la relación lineal que puede existir entre una variable dependiente Y y varias independientes, regresoras o covariables X_1, \dots, X_d , se plantea el modelo de regresión lineal múltiple:

$$Y_i = a_0 + a_1X_1 + \dots + a_dX_d + \varepsilon$$

- Las estimaciones por mínimos cuadrados de a_0, a_1, \dots, a_d se obtienen minimizando:

$$\{\hat{a}_0, \hat{a}_1, \dots, \hat{a}_d\} = \arg \min_{\{\alpha_0, \dots, \alpha_d\}} \sum_{i=1}^n [y_i - (\alpha_0 + \alpha_1x_{1i} + \dots + \alpha_dx_{di})]^2$$

1.6.4 Regresión No Lineal

- Pueden estimarse modelos no lineales de regresión transformando convenientemente las variables. Supongamos que queremos estimar el modelo:

$$y = be^{ax}$$

Tomando logaritmos:

$$\ln y = \ln b + ax$$

transformando los datos mediante $t = \ln y$, y estimando la recta de regresión:

$$t = b' + ax$$

la solución al problema original será: \hat{a} y $\hat{b} = e^{b'}$ (ya que $\ln b = b' \Leftrightarrow b = e^{b'}$).

- Modelos y transformaciones asociadas son:

Modelo		Transformación	Modelo Lineal
Exponencial	$y = be^{ax}$	$t = \ln y$	$t = ax + \ln b$
Logarítmico	$y = a \ln x + b$	$z = \ln x$	$y = az + b$
Potencia	$y = bx^a$	$t = \ln y, z = \ln x$	$t = az + \ln b$
Inverso	$y = a\frac{1}{x} + b$	$z = 1/x$	$y = az + b$

1.6.5 Asociación entre una Variable Continua y una Variable Categórica

- **Ratio de correlación.**

Sea (X, Y) con $X \in \{c_1, \dots, c_m\}$ categorías e Y variable continua y el conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$. El ratio de correlación se define como:

$$\eta_{XY}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{j=1}^m n_j (\bar{y}_j - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \in [0, 1]$$

donde: si $x_i = c_j$, $\hat{y}_i = \bar{y}_j$ la media en la categoría j , n_j es el número de datos en la categoría $j = 1, \dots, m$ y $n = \sum_{j=1}^m n_j$ es el número total de datos.

- **Propiedades:**

- El ratio de correlación η^2 se interpreta como el tanto por uno de variabilidad de la variable Y explicado por X . En terminología de análisis de la varianza, es el ratio de varianza de Y explicada por el factor X .
- $\eta^2 = 0$ si las medias \bar{y}_j son iguales (conocer X no aporta información sobre Y).
- $\eta^2 = 1$ si todas las observaciones de cada grupo son iguales (conocer X determina completamente Y).
- Si la variable X es cuantitativa pero discreta, si sólo hay dos categorías, $\eta^2 = r^2$. En caso contrario, $\eta^2 > r^2$ y η^2 captura la posible relación no lineal entre X e Y .

- Ejercicio: comprobar que $\eta^2 = 1$ para las tres tablas anteriores (sólo hay una observación de Y por cada categoría de X).

1.6.6 Asociación entre Dos Variables Categóricas

Tabla de Contingencia

Variables nominales, cualitativas o categóricas o **atributos**. Sus valores se denominan **modalidades**. La **tabla de correlación** ahora se denomina **tabla de contingencia** (que también puede expresarse en frecuencias relativas):

$X \backslash Y$	y_1	y_2	\dots	y_v	$n_{i\bullet}$
x_1	n_{11}	n_{12}	\dots	n_{1v}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	\dots	n_{2v}	$n_{2\bullet}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_u	n_{u1}	n_{u2}	\dots	n_{uv}	$n_{u\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet v}$	n

Medidas de Asociación

Basándose en la condición de independencia 1.2, se definen las siguientes **Medidas de Asociación entre variables categóricas**:

1. Coeficiente de Contingencia Chi cuadrado χ^2 (o cuadrado de la contingencia):

$$\chi^2 = \sum_{i=1}^u \sum_{j=1}^v \frac{(e_{ij} - n_{ij})^2}{e_{ij}} = \sum_{i=1}^u \sum_{j=1}^v \frac{n_{ij}^2}{e_{ij}} - n$$

donde: $e_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$ frecuencias esperadas de ser los atributos independientes

- (a) Siempre $\chi^2 \geq 0$. Si $\chi^2 = 0$, hay independencia total entre los atributos pues coinciden las observadas con las esperadas.
- (b) Si las variables fuesen independientes, el estadístico χ^2 sigue una distribución χ_m^2 con $m = (u - 1) \times (v - 1)$ *grados de libertad*.
- (c) Bajo esta hipótesis de independencia, la probabilidad de que la magnitud χ^2 alcance o supere el valor χ_0^2 que se obtiene con los datos puede obtenerse en Excel mediante la función

$$DISTR.CHI(\chi_0^2, m)$$

Si esta probabilidad es muy pequeña (p. ej. < 0.05), hay que concluir que las variables están asociadas estadísticamente. En caso contrario, no puede afirmarse la existencia de tal asociación.

2. Cuadrado medio de la Contingencia (Phi):

$$\varphi^2 = \frac{\chi^2}{n} = \frac{1}{n} \sum_{i=1}^u \sum_{j=1}^v \frac{n_{ij}^2}{e_{ij}} - 1$$

Siempre $\varphi \geq 0$. Si $\varphi = 0$, independencia total entre los atributos. Por ejemplo, el **SPSS** utiliza $\varphi' = \sqrt{\chi^2/n}$ que está en $[0, 1]$ para tablas 2×2 y $r_{XY} = \varphi'$.

3. Coeficiente de Contingencia de Pearson:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{\varphi^2}{1 + \varphi^2}} \in [0, 1)$$

Si $C \rightarrow 1$ gran asociación y si $C = 0$ asociación nula. Si $u = v$ (tabla cuadrada) el valor máximo es $C_{\max} = \sqrt{(v-1)/v}$.

4. Coeficiente V de Cramer:

$$V = \sqrt{\frac{\chi^2}{n[\min(u, v) - 1]}} \in [0, 1]$$

En tablas 2×2 , $V = \varphi' = r_{XY}$.

Ejemplo 5 La siguiente tabla de contingencia recoge los resultados de una encuesta pública sobre la eficacia de las mujeres policías en asuntos de tráfico comparada con la de sus compañeros hombres, en términos del nivel de estudios de los encuestados:

<i>Más Eficaz \ Nivel de Estudios</i>	<i>Primarios</i>	<i>Medios</i>	<i>Superiores</i>	$n_{i\bullet}$
<i>Hombre</i>	10	28	32	70
<i>Mujer</i>	13	37	38	88
<i>Por Igual</i>	29	169	311	509
$n_{\bullet j}$	52	234	381	667

Determinar el grado de asociación entre el nivel de estudios de los encuestados y su opinión sobre la efectividad de los agentes según su sexo.

Solución. Se obtiene:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^u \sum_{j=1}^v \frac{(e_{ij} - n_{ij})^2}{e_{ij}} = \sum_{i=1}^u \sum_{j=1}^v \frac{n_{ij}^2}{e_{ij}} - n = 20.361 \\ \varphi^2 &= \frac{\chi^2}{n} = \frac{20.361}{667} = 0.03053 \\ \varphi &= \sqrt{\frac{\chi^2}{n}} = \sqrt{0.03053} = 0.1747 \\ V &= \sqrt{\frac{\chi^2}{n[\min(u, v) - 1]}} = \sqrt{\frac{20.361}{667 \cdot 2}} = 0.1236\end{aligned}$$

Veamos si estos valores son significativos utilizando Excel. Tenemos que $m = (u - 1) \times (v - 1) = (3 - 1) \times (3 - 1) = 4$, y que:

$$\text{distr.chi}(20.361; 4) = 0.0004$$

que es una probabilidad tan pequeña, que hay que concluir que las variables no son independientes.

Al calcular las frecuencias relativas de la distribución marginal del Nivel de Estudios, se obtiene:

Más Eficaz \ Nivel de Estudios	Primarios	Medios	Superiores	$f_{i\bullet}$
Hombre	19.23%	11.97%	8.40%	10.49%
Mujer	25.00%	15.81%	9.97%	13.19%
Por Igual	55.77%	72.22%	81.63%	76.32%
Total	100.00%	100.00%	100.00%	100.00%

donde puede verse que la asociación existente se manifiesta sobre todo en los mayores porcentajes en favor de la igualdad entre los encuestados con estudios superiores.