Tema 9: Estadística descriptiva

Matemáticas específicas para maestros

Grado en Educación Primaria

ÍNDICE

ÍNDICE

Introducción

Estadística: Ciencia que trata sobre la teoría y aplicación de métodos para coleccionar, representar, resumir y analizar datos, así como realizar inferencias a partir de ellos.

Tipos:

- **Estadística descriptiva:** Recogida y análisis de datos para realizar una descripción de las características de un colectivo, deduciendo conclusiones sobre su estructura y sobre las relaciones existentes con otros colectivos.
- **Inferencia estadística:** Realización de inferencias acerca de las características de una población a partir del estudio de una muestra (necesita probabilidad).
  RAE: inferir es deducir algo o sacarlo como conclusión de otra cosa.

ÍNDICE

Conceptos generales

- **Población**: conjunto total de individuos que se quiere analizar. A veces demasiado grande o inaccesible.
- **Muestra**: subconjunto de la población que *representa* al total de individuos. Hay *técnicas de muestreo* para elegir una "buena" muestra, es decir, la más representativa posible.
- **Variable o carácter estadístico**: característica o propiedad de un individuo que se quiere estudiar.
  Tipos de variables:
  - *Cualitativas*: cualidades no numéricas. Ejemplos: sexo, nacionalidad, calificación no numérica, etc.
  - *Cuantitativas*: valores numéricos.
    - *Cuantitativas discretas*: número finito de valores. Ejemplos: número de asistentes, asignaturas aprobadas.
    - *Cuantitativas continuas*: valores reales dentro de un intervalo. Ejemplos: temperatura, peso, tiempo.

ÍNDICE

Representación de datos: tablas de frecuencias y gráficos

La forma en que representamos los datos puede ayudarnos a extraer información de ellos y, por tanto, sacarles utilidad.

Ejemplo

Mala representación de datos. Tenemos las siguientes observaciones sobre el tipo de árboles/arbustos en un parque:

Abeto, Durillo, Abeto, Brezo, Ciprés,
Brezo, Brezo, Abeto, Brezo, Durillo,
Durillo, Brezo, Brezo, Ciprés, Brezo,
Abeto, Brezo, Brezo, Durillo, Brezo.

- ¿Cuántas observaciones hay?
- ¿Cuántas clases diferentes?
- ¿Repeticiones de cada clase?

Tablas de frecuencias

Se pueden utilizar para cualquier tipo de variable.
Con las $n = 20$ observaciones del ejemplo se construye una tabla:

x_i	f_i	$h_i = \frac{f_i}{n}$	$F_i = \sum_{j=1}^{i} f_j$	$H_i = \frac{F_i}{n}$
A	4	$4/20 = 0.2$	4	0.2
B	10	$10/20 = 0.5$	14	0.7
C	2	$2/20 = 0.1$	16	0.8
D	4	$4/20 = 0.2$	20=n	1
Total	$n = \sum_{i=1}^{4} f_i = 20$	$\sum_{i=1}^{4} h_i = 1$	-	-

- x_i: datos distintos observados (marcas de clase si los datos están agrupados)
- f_i: frecuencia absoluta
- h_i: frecuencia relativa
- F_i: frecuencia absoluta acumulada
- H_i: frecuencia relativa acumulada

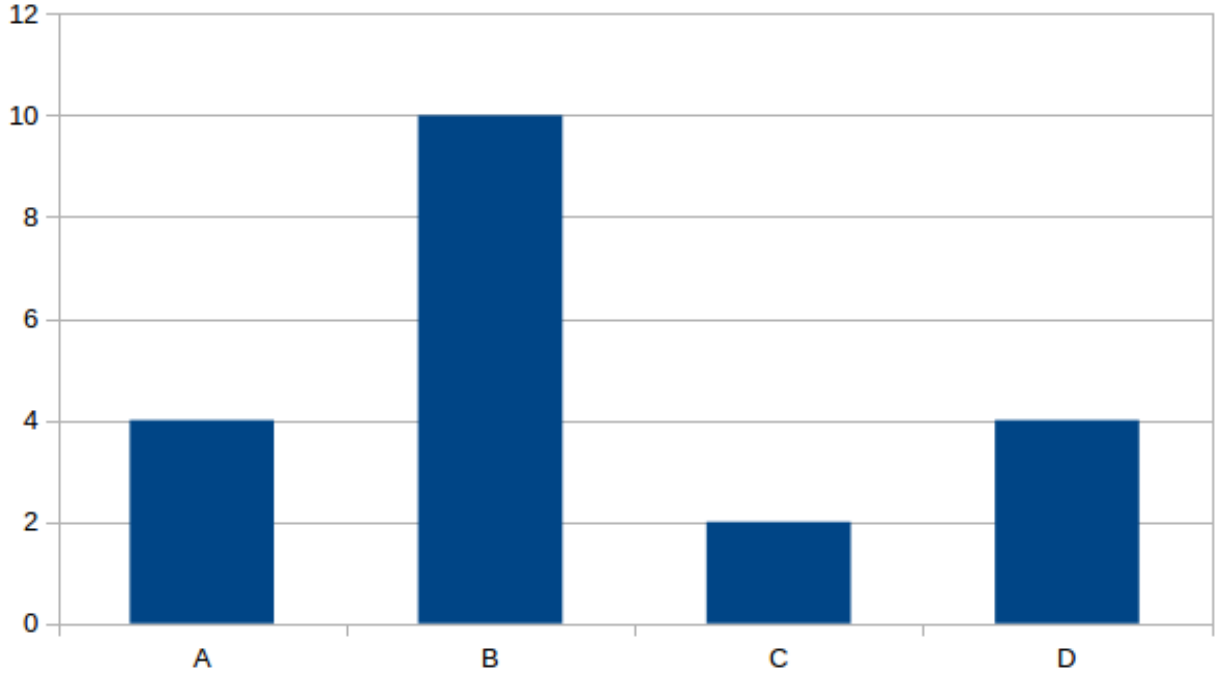
Representaciones gráficas

- Diagrama de barras (datos NO agrupados en intervalos)
- Diagrama de sectores (datos NO agrupados en intervalos)
- Histograma (datos agrupados en intervalos)
- Polígono de frecuencias (datos agrupados o no)

Representaciones gráficas

Diagrama de barras: la altura de cada barra representa la frecuencia absoluta (o relativa) de cada una de las diferentes observaciones consideradas.
Se usa con variables cualitativas y cuantitativas discretas (datos NO agrupados en intervalos). (¡Barras separadas!)

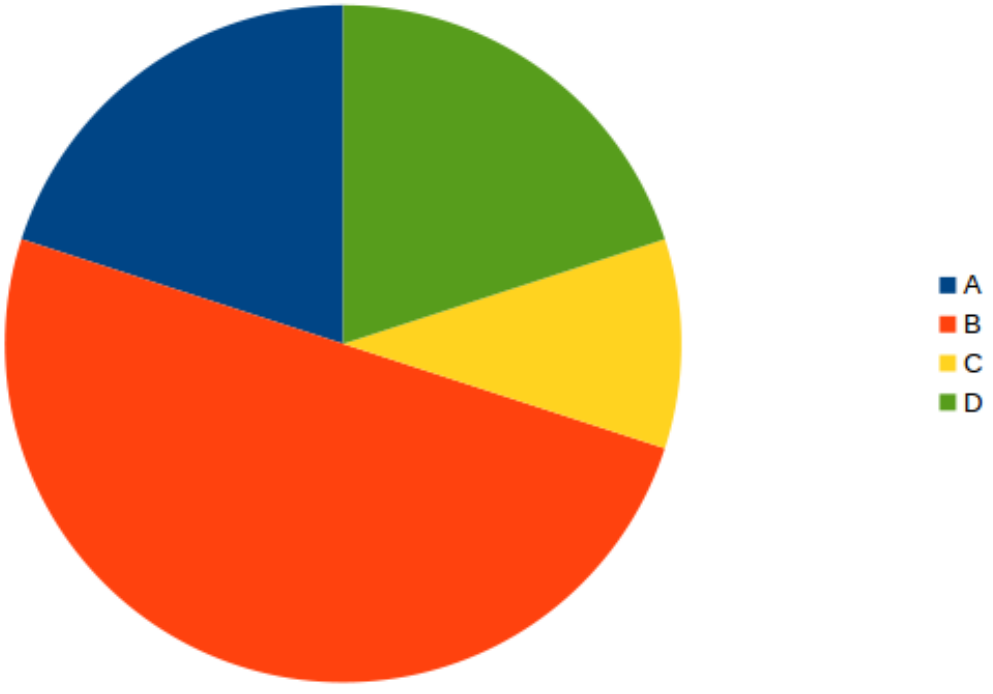
– En el ejemplo anterior:



Representaciones gráficas

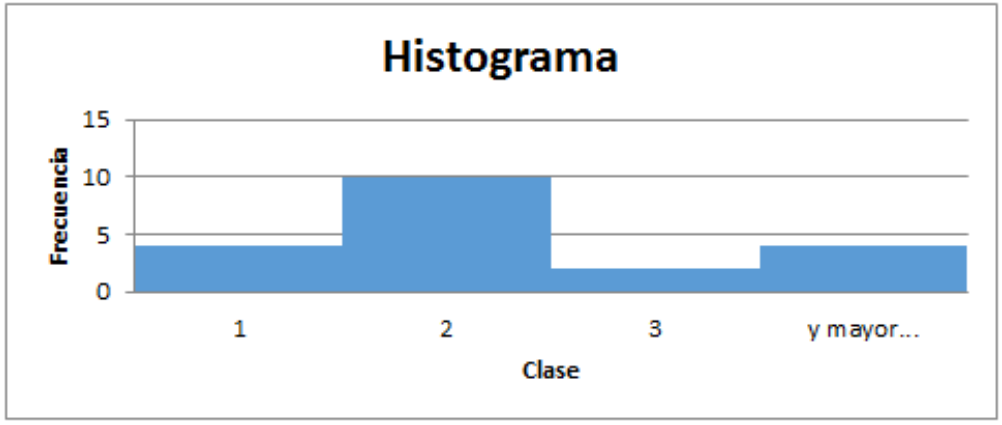
Diagrama de sectores: cada sector del círculo representa la frecuencia relativa.
Se usa con variables cualitativas y cuantitativas discretas (datos NO agrupados en intervalos).

– En el ejemplo anterior:



Representaciones gráficas

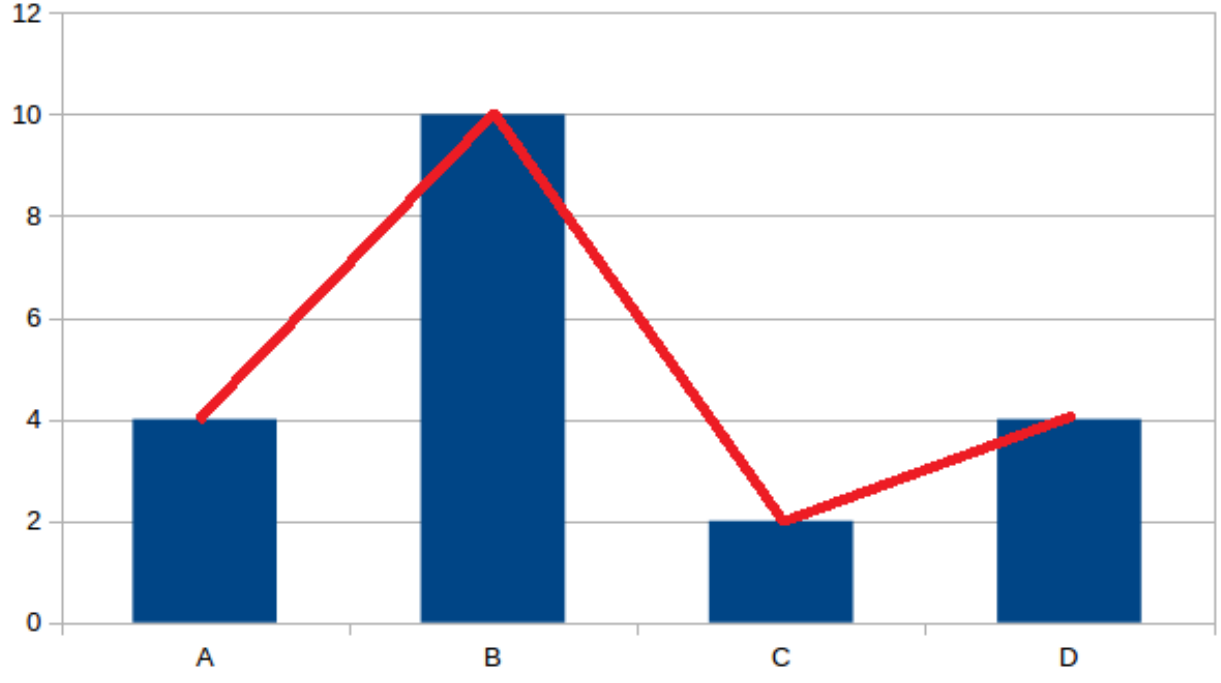
Histograma: el área de cada rectángulo es proporcional a la frecuencia absoluta (o relativa) de cada una de las clases consideradas.
Se usa con variables cuantitativas continuas (datos agrupados en intervalos). (¡Barras juntas!)



Representaciones gráficas

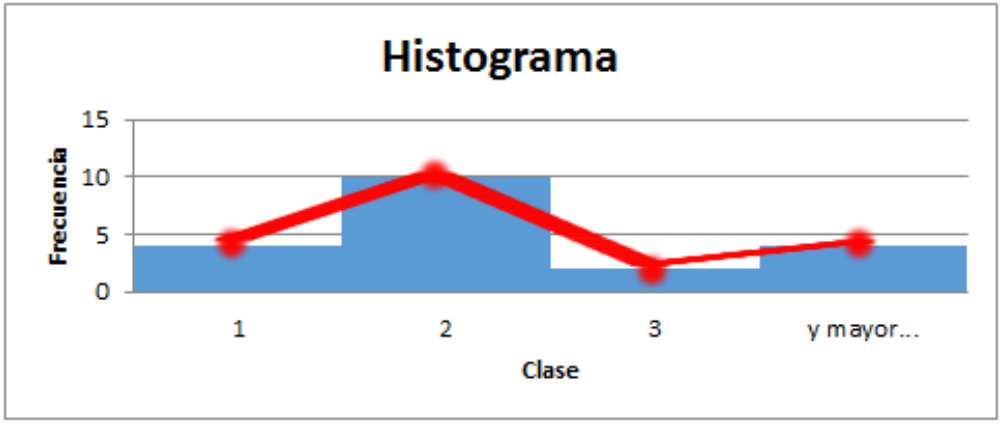
Polígono de frecuencias:

- Si la variable es discreta, el polígono de frecuencias se obtiene uniendo los extremos superiores de las barras en el diagrama de barras.



Polígono de frecuencias:

- Si la variable es continua y está agrupada en intervalos, el polígono de frecuencias se obtiene uniendo los puntos medios de las bases superiores de cada rectángulo en el histograma.



Parámetros estadísticos

Valores numéricos calculados a partir de los valores de la muestra. Proporcionan herramientas útiles para extraer información de la muestra. Varios tipos:

- **Parámetros de centralización y posición**: información del valor en torno al cual se agrupan los datos de la muestra, y valores que ocupan cierta posición.
- **Parámetros de dispersión**: información sobre la concentración o dispersión de los datos con respecto a los valores centrales.

Parámetros de centralización y posición

- Media aritmética
- Moda
- Mediana
- Percentiles y cuartiles

Parámetros de centralización y posición

Media aritmética

- Datos sin agrupar o marcas de clase de intervalos: $x_1, \ldots, x_n$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- Si algunos de los n datos aparecen repetidos, se podrían escribir como $x_1, \ldots, x_m$ con frecuencias absolutas $f_1, \ldots, f_m$ y relativas $h_1, \ldots, h_m$ y entonces

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{m} x_i f_i = \sum_{i=1}^{m} x_i h_i.$$

Propiedades de la media:

- $\overline{c + x} = c + \overline{x}$, c constante.
- $\overline{cx} = c\overline{x}$, c constante.
- $\overline{x + y} = \frac{n_x \overline{x} + n_y \overline{y}}{n_x + n_y}$

Parámetros de centralización y posición

Moda

Es el valor o valores que más se repite en la muestra, es decir, en el que se alcanza el máximo de las frecuencias relativas.
Si los datos están agrupados, se llama *clase modal*.
No tiene que ser única.

Mediana

Primer valor x_i (o intervalo) con frecuencia relativa acumulada mayor o igual que 0.5*.*
En un conjunto de datos ordenado de manera creciente, la mediana es el valor mínimo que divide la muestra en dos conjuntos con el mismo número de datos.

Parámetros de centralización y posición

Percentiles y cuartiles

Percentil P_r, primer valor tal que su frecuencia relativa acumulada es mayor o igual que el $r/100$.

Extendiendo la idea de mediana, el percentil P_{10} es el valor en el cual hemos recorrido el 10 % de los datos de la muestra.

Casos importantes:

- Primer cuartil, $P_{25} = Q_1$.
- Segundo cuartil, $P_{50} = Q_2$, ¡es la *mediana*!
- Tercer cuartil, $P_{75} = Q_3$.

Parámetros de dispersión

- **Rango o recorrido**
  Diferencia entre el valor más grande x_{max} y el valor más pequeño x_{min} en un conjunto ordenado de datos.

$$R = x_{max} - x_{min}.$$

- **Recorrido intercuartílico**
  Indica dónde se encuentra el 50 % de los datos centrales. Se calcula como la diferencia entre el tercer cuartil y el primer cuartil.

$$RI = Q_3 - Q_1.$$

Parámetros de dispersión

Varianza es la media aritmética de las diferencias (al cuadrado) entre las observaciones y la media:

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

También se puede expresar en función de las frecuencias relativas y absolutas.

$$s_X^2 = \frac{1}{n}\sum_{j=1}^{m} f_j(x_j - \bar{x})^2 = \sum_{j=1}^{m} h_j(x_j - \bar{x})^2$$

Y con esta fórmula equivalente:

$$s_X^2 = \overline{(x^2)} - (\overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.$$

Desviación típica es la raíz cuadrada (positiva) de la varianza.

$$s_X = +\sqrt{s_X^2}.$$

Análisis descriptivo de una variable estadística

Tipo de variable	**Tablas**	**Gráficos**	**Medidas numéricas**
Cualitativa	Frecuencias absolutas y relativas/porcentajes (no acumuladas)	Diagrama de sectores o barras	Moda
Cuantitativa discreta	Frecuencias de valores aislados	Diagrama de barras o sectores	Todas las medidas características
Cuantitativa continua	Frecuencias de valores agrupados	Histograma	Todas las medidas características

ÍNDICE

Estadística descriptiva bidimensional

Hasta ahora hemos considerado una única variable.
A continuación vamos a estudiar la relación entre dos variables de una misma muestra, es decir, dos características de cada uno de los individuos de la muestra.
Se puede describir esta relación mediante una tabla de doble entrada o mediante un gráfico.

Ejemplos:

1. Edad y peso de una persona.
2. Edad y nivel de inglés (bajo-medio-alto).
3. Color de ojos y pelo.

Las dos variables X e Y pueden ser:

- Las dos cuantitativas: ejemplo 1.
- Una cuantitativa y la otra cualitativa: ejemplo 2.
- Las dos cualitativas: ejemplo 3.

Tablas de doble entrada

Ejemplo

X : “color de ojos”
Y : “color de pelo”

X \ Y	Claro	Oscuro
Claro	7	2
Oscuro	1	10

Posibles preguntas:

- ¿ Cuántas personas tienen ojos de color claro?
- ¿ Cuántas personas tienen ojos de color oscuro?
- ¿ Cuántas personas tienen pelo de color claro?
- ¿ Cuántas personas tienen pelo de color oscuro?

Tablas de doble entrada

Ejemplo

X : "calificación en la asignatura de Matemáticas"
Y : "calificación en la asignatura de Didáctica de las Matemáticas"

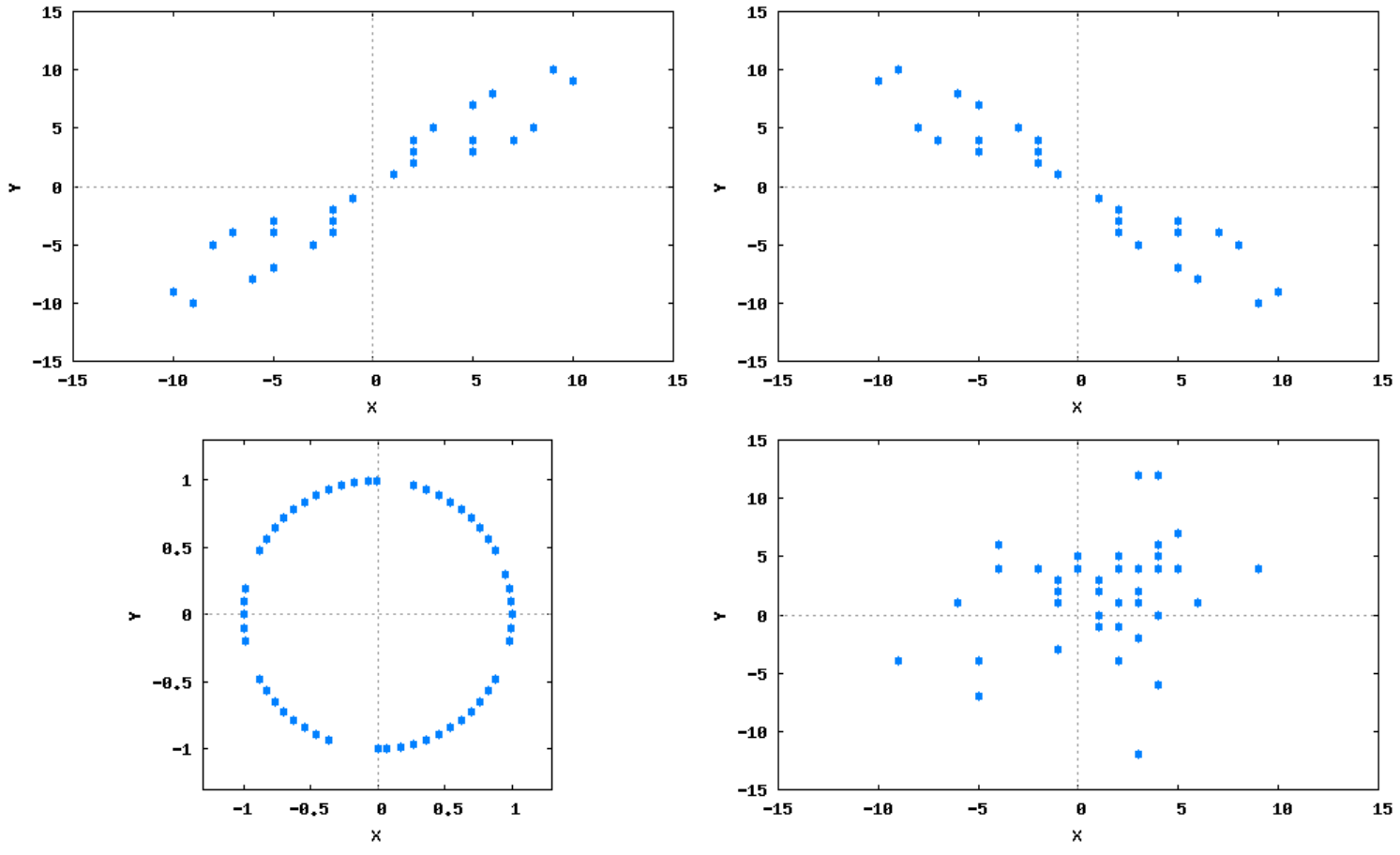
X	5	3	5	5	5	6	3	9	9
Y	6	1	8	8	8	7	1	10	10

Construimos la tabla de doble entrada (de frecuencias absolutas):

Y \ X	3	5	6	9
1	2	0	0	0
6	0	1	0	0
7	0	0	1	0
8	0	3	0	0
10	0	0	0	2

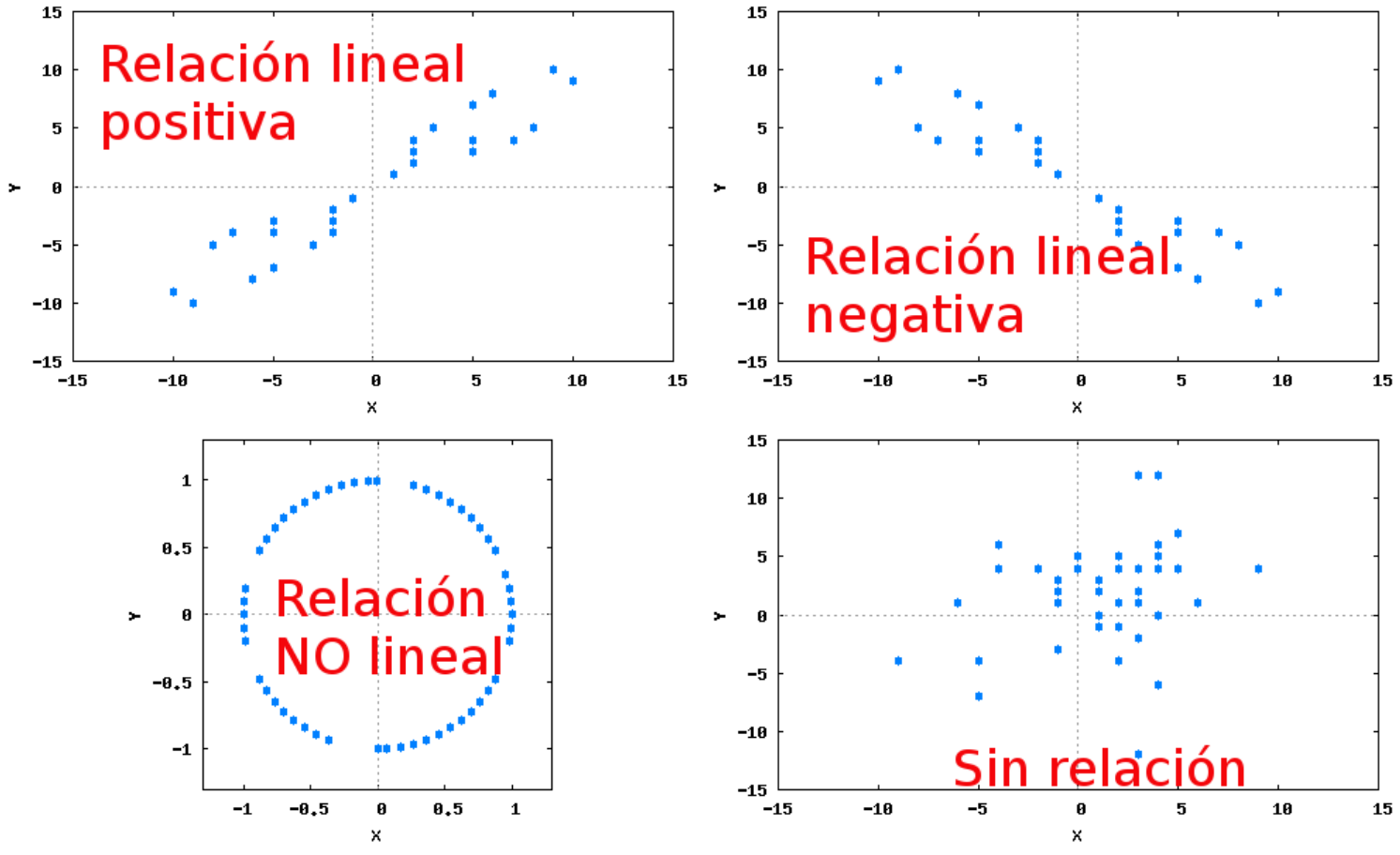
Diagramas de dispersión o nube de puntos

Dadas dos variables cuantitativas X, Y, la forma más intuitiva de ver si están relacionadas es representarlas mediante un *diagrama de dispersión*, también llamado *nube de puntos*.



Diagramas de dispersión o nube de puntos

Dadas dos variables cuantitativas X, Y, la forma más intuitiva de ver si están relacionadas es representarlas mediante un *diagrama de dispersión*, también llamado *nube de puntos*.



Correlación lineal

Nos centraremos en estudiar si hay dependencia lineal entre X, Y. Para ello, definiremos la **Covarianza muestral**:

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

que también puede calcularse como

$$s_{XY} = \overline{(xy)} - \bar{x}\cdot\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Sabemos que

$$s_{XY}\begin{cases} > 0, & \text{relación lineal directa} \\ < 0, & \text{relación lineal inversa} \\ = 0, & \text{no hay relación lineal (posible otro tipo de relación)} \end{cases}$$

Pregunta: Si existe relación lineal ($s_{XY} \neq 0$), ¿cómo de fuerte es la relación? La covarianza depende de la escala de los datos, luego necesitamos calcular el Coeficiente de correlación de Pearson.

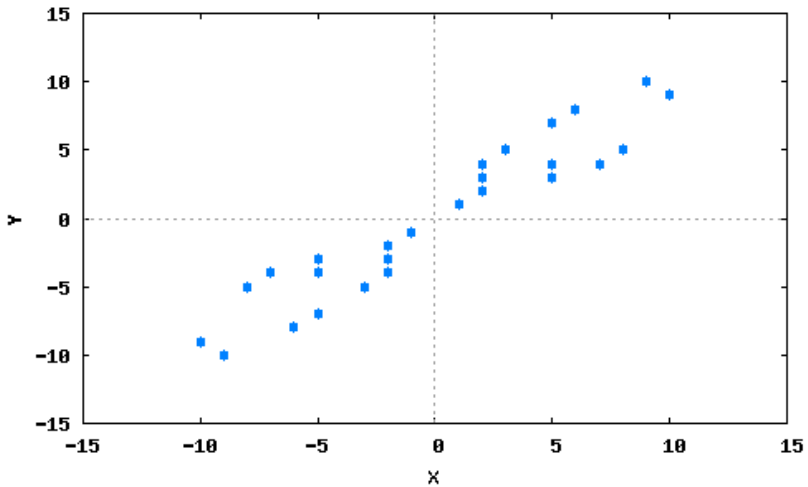
Correlación lineal

Correlación muestral o **coeficiente de correlación de Pearson**

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \in [-1, 1], \quad s_X \text{: desviación típica.}$$

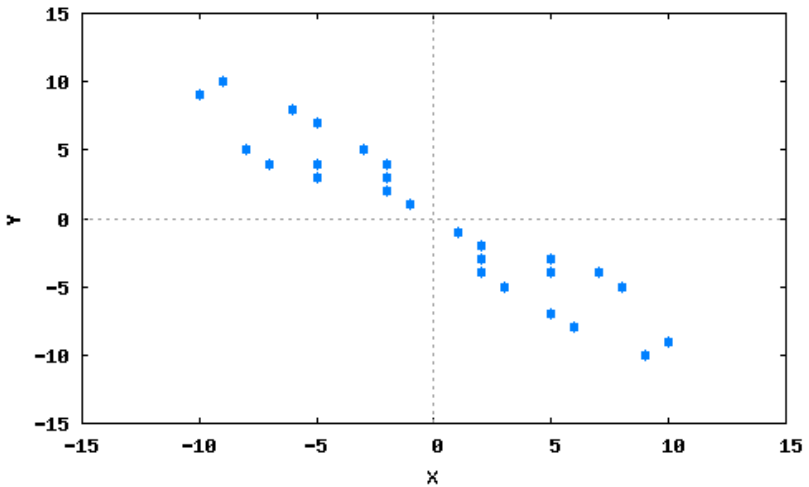
$$r_{XY} \begin{cases} > 0 & \text{, relación lineal directa} \\ < 0 & \text{, relación lineal inversa} \\ \approx \pm 1 & \text{, relación fuerte} \\ \approx 0 & \text{, no relación lineal (posible otro tipo de relación)} \end{cases}$$

Algunos ejemplos



Relación lineal directa:

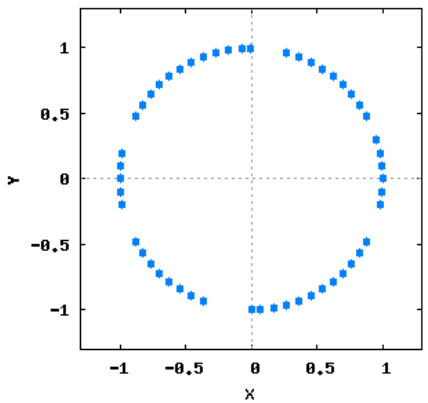
$s_{XY} > 0, \quad r_{XY} \approx 1.$



Relación lineal inversa:

$s_{XY} < 0, \quad r_{XY} \approx -1.$

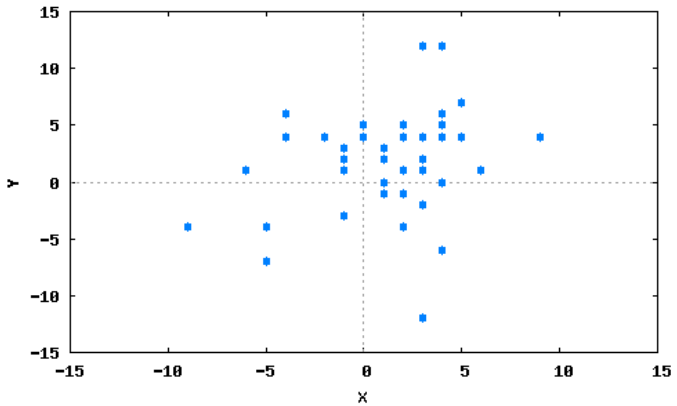
Algunos ejemplos



Relación **no** lineal:

$s_{XY} = 0, \quad r_{XY} = 0.$

(¡Pero **sí** otro tipo de relación!)



Sin Relación:

$s_{XY} \approx 0, \quad r_{XY} \approx 0.$

Regresión lineal

Si $|r_{XY}| \approx 1$, es decir X e Y tienen una *fuerte relación lineal*, podemos preguntarnos cuál es la recta que mejor se ajusta a los datos.
Esta recta, llamada de regresión lineal, es muy útil para hacer predicciones.

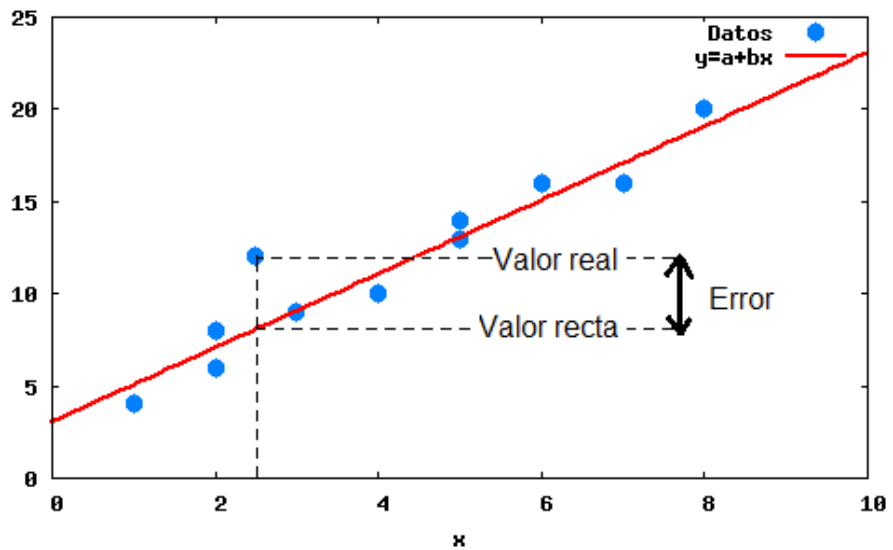
Pero... ¿cómo la calculamos?

Recta de regresión lineal

Recta de regresión de Y sobre X:

- Conocido un valor x_i, queremos estimar y_i.
- Buscamos una recta $y = a + bx$ que minimice las distancias (*errores*) entre los puntos de la muestra (x_i, y_i) y la predicción de la recta, obteniendo:

$$y - \bar{y} = \frac{s_{XY}}{s_X^2}(x - \bar{x}).$$



Recta de regresión lineal

Recta de regresión de Y sobre X: (conocido x_i, queremos estimar y_i)

$$y - \bar{y} = \frac{s_{XY}}{s_X^2}(x - \bar{x}).$$

Recta de regresión de X sobre Y: (conocido y_i, queremos estimar x_i)

$$x - \bar{x} = \frac{s_{XY}}{s_Y^2}(y - \bar{y}).$$

Importante:

- ¡LAS DOS RECTAS NO SON IGUALES EN GENERAL!
- Las rectas de regresión pasan por el punto $(\bar{x}, \bar{y})$.
- El producto de sus pendientes es r_{XY}^2.

Problema 6 del Boletín (Regresión lineal simple)

Observando las edades y los pesos de 5 niños se obtuvieron los siguientes resultados:

Edad (en años)	2	4.5	6	7.2	8
Peso (en kg)	15	19	25	33	34

a) Halla las medias y las desviaciones típicas de cada una de las variables.

b) Calcula el coeficiente de correlación lineal y la recta de regresión del peso sobre la edad.

c) ¿Qué indica el coeficiente de correlación lineal?

a) Medias y desviaciones típicas

Edad (en años)	2	4.5	6	7.2	8
Peso (en kg)	15	19	25	33	34

Llamemos X a la edad e Y al peso. Sus medias serán:

$$\bar{x} = \frac{1}{5}(2 + 4.5 + 6 + 7.2 + 8) = \frac{27.7}{5} = 5.54 \text{ años}$$

$$\bar{y} = \frac{1}{5}(15 + 19 + 25 + 33 + 34) = \frac{126}{5} = 25.2 \text{ kg}$$

Sus varianzas y desviaciones típicas:

$$s_x^2 = \frac{1}{5}(2^2 + 4.5^2 + 6^2 + 7.2^2 + 8^2) - 5.54^2 = 4.5264$$

$$\Rightarrow s_x = \sqrt{4.5264} = 2.13 \text{ años}$$

$$s_y^2 = \frac{1}{5}(15^2 + 19^2 + 25^2 + 33^2 + 34^2) - 25.2^2 = 56.16$$

$$\Rightarrow s_y = \sqrt{56.16} = 7.49 \text{ kg}$$

b) Coeficiente de correlación lineal

Edad (en años)	2	4.5	6	7.2	8
Peso (en kg)	15	19	25	33	34

Covarianza:

$$s_{XY} = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{5}\sum_{i=1}^{5} x_i y_i - \bar{x}\bar{y}$$

$$= \frac{1}{5}(2 \cdot 15 + 4.5 \cdot 19 + 6 \cdot 25 + 7.2 \cdot 33 + 8 \cdot 34) - 5.54 \cdot 25.2$$

$$= \frac{775.1}{5} - 5.54 \cdot 25.2 = 15.412$$

Luego el coeficiente de correlación lineal es:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{15.412}{2.13 \cdot 7.49} = 0.94\,.$$

b) Recta de regresión del peso sobre la edad

Recta de regresión de Y (peso) sobre X (edad)

$$y - \bar{y} = \frac{s_{XY}}{s_X^2}(x - \bar{x})\,.$$

En este caso:

$$y - 25.2 = \frac{15.412}{4.5264}(x - 5.54)\,.$$

$$y - 25.2 = 3.4(x - 5.54)\,.$$

c) Significado del coeficiente de correlación

Como $r_{xy} = 0.97$ es un valor positivo, existe una relación lineal directa o positiva. Por ser próximo a 1 sabemos que esta relación es muy fuerte.

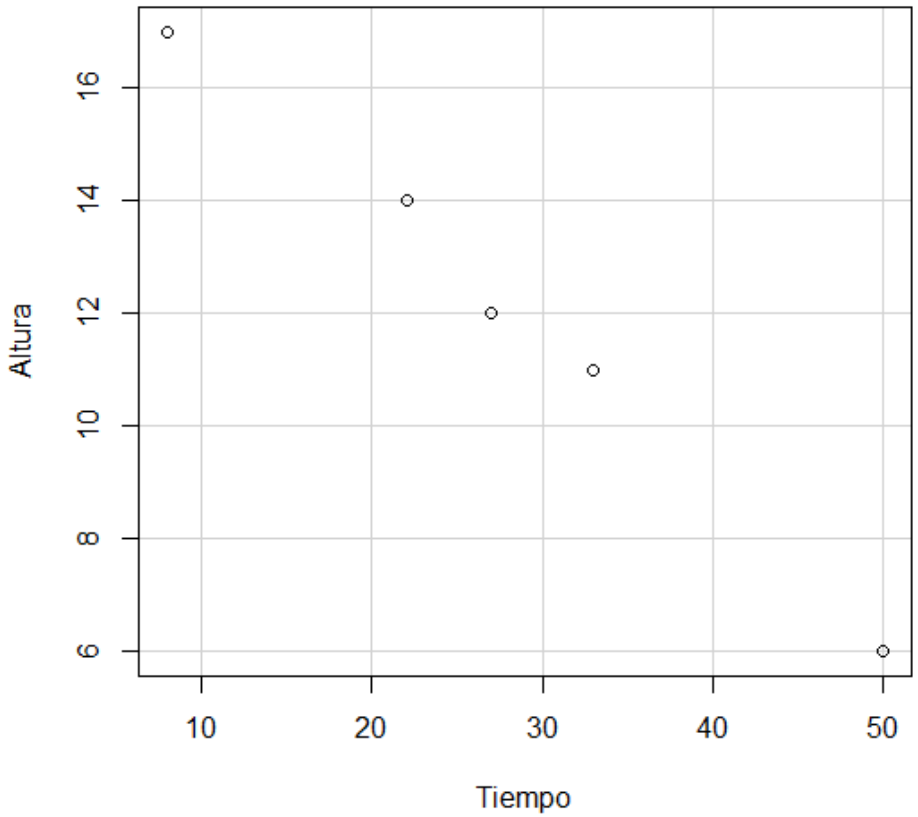
Ejercicio extra de Regresión lineal simple

En un depósito cilíndrico, la altura del agua que contiene varía conforme pasa el tiempo, según la siguiente tabla.

X (tiempo en horas)	8	22	27	33	50
Y (altura en metros)	17	14	12	11	6

1. Realiza un gráfico que represente la altura en función del tiempo. ¿Qué conclusiones se pueden extraer de este gráfico?
2. Halla e interpreta la covarianza.
3. Halla e interpreta el coeficiente de correlación lineal entre el tiempo y la altura.
4. ¿Cuál será la altura del agua cuando hayan transcurrido 60 horas?
5. Cuando la altura del agua es de 2 metros, suena una alarma. ¿Qué tiempo ha de transcurrir para que avise la alarma?

Solución de 1: Diagrama de dispersión.



X	8	22	27	33	50
Y	17	14	12	11	6

En el gráfico de dispersión se muestra que a medida que aumenta el tiempo (variable X), la altura del agua (variable Y) disminuye, es decir, hay una relación inversa entre las dos variables, lo que se traduce en una covarianza y un coeficiente de correlación negativo.
Además, como los puntos forman aproximadamente una recta deducimos que el coeficiente de correlación debe ser próximo a -1.

Solución de 2: Covarianza

$$s_{XY} = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{5}\sum_{i=1}^{5} x_i y_i - \bar{x}\bar{y}$$

Calculamos

$$\frac{1}{5}\sum_{i=1}^{5} x_i y_i = \frac{1}{5}(8 \cdot 17 + 22 \cdot 14 + 27 \cdot 12 + 33 \cdot 11 + 50 \cdot 6) = 286.2,$$

$$\bar{x} = \frac{1}{5}(8 + 22 + 27 + 33 + 50) = 28,$$

$$\bar{y} = \frac{1}{5}(17 + 14 + 12 + 11 + 6) = 12.$$

Luego

$$s_{XY} = 286.2 - 28 \cdot 12 = -49.8\,.$$

La covarianza negativa implica que existe una relación lineal inversa entre el tiempo y la altura del agua en el depósito.

Solución de 3: Coeficiente de correlación lineal

Calculamos primero las desviaciones típicas:

$$s_x^2 = \frac{1}{5}\sum_{i=1}^{5}(x_i - \bar{x})^2 = \frac{1}{5}\sum_{i=1}^{5} x_i^2 - \bar{x}^2 = 189.2 \Rightarrow s_x = 13.755,$$

$$s_y^2 = \frac{1}{5}\sum_{i=1}^{5}(y_i - \bar{y})^2 = \frac{1}{5}\sum_{i=1}^{5} y_i^2 - \bar{y}^2 = 13.2 \Rightarrow s_y = 3.6332.$$

Luego el coeficiente de correlación lineal es:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{-49.5}{13.755 \cdot 3.6332} = -0.9965\,.$$

Por ser un valor negativo existe una relación lineal inversa y por ser próximo a -1 sabemos que esta relación es muy fuerte.

Solución de 4: altura del agua cuando hayan transcurrido 60 horas

Recta de regresión de Y (Altura) sobre X (Tiempo)

$$y - \bar{y} = \frac{s_{XY}}{s_X^2}(x - \bar{x})\,.$$

Por tanto, la altura estimada cuando hayan transcurrido 60 horas es:

$$y - 12 = \frac{-49.8}{189.2}(60 - 28)$$

$$\Rightarrow y = 12 - \frac{49.8}{189.2}(60 - 28) = 3.58 \text{ metros}.$$

Solución de 5: tiempo transcurrido para que la altura del agua sea 2 metros

Recta de regresión de X (tiempo) sobre Y (altura).

$$x - \bar{x} = \frac{s_{XY}}{s_Y^2}(y - \bar{y})\,.$$

En este caso, el tiempo transcurrido cuando suene la alarma es

$$x - 28 = \frac{-49.8}{13.2}(2 - 12)$$

$$\Rightarrow x = 28 - \frac{49.8}{13.2}(2 - 12) = 65.73 \text{ horas}.$$