Tema 1. Estadística Descriptiva (3ª parte)

Estadística

Ángel Serrano Sánchez de León

Índice

- Introducción
- Variables estadísticas
- Distribuciones de frecuencias
- Introducción a la representación gráfica de datos
- Medidas de tendencia central: media (aritmética, geométrica, armónica, cuadrática), mediana, moda, cuartiles, deciles, percentiles
- Medidas de dispersión: Recorrido (total/intercuartílico/ semiintercuartílico), desviación media, desviación típica, varianza, coeficientes de variación
- Momentos. Medidas de asimetría: Sesgo, curtosis
- Variables estadísticas bidimensionales

Introducción

- Hasta ahora hemos descrito variables de manera independiente.
- Una **muestra estadística bidimensional** consiste en 2 variables observadas de manera simultánea de un fenómeno. Ejemplos:
  - Altura y peso de los alumnos UFV.
  - Superficie y precio de la vivienda.
  - Ingresos y número de hijos en distintas familias.
  - Color de pelo y ojos de una serie de personas.
- En general, variable expresada como una pareja de valores (x, y).
- Si añadimos más variables → **Multidimensionales**.

3

Distribución de frecuencias de una variable bidimensional

- Tenemos una muestra formada por N pares de valores de una variable bidimensional:

$$(x_1, y_1), \ldots (x_N, y_N)$$

- Número de valores diferentes (variable discreta):
  - $x_1, \ldots x_k$
  - $y_1, \ldots, y_l$, donde k y l no tienen por qué ser iguales.
- En caso de que la variable sea continua, la agruparíamos en distintos intervalos.
- Dos pares de datos (x_1, y_1), (x_2, y_2) son diferentes si ambas componentes son diferentes:

$$x_1 \neq x_2, \quad y_1 \neq y_2$$

- **Frecuencia absoluta** n_{ij}: el número de veces que se repite el par (x_i, y_j).

Distribución de frecuencias de una variable bidimensional

- **Tabla de frecuencias de doble entrada (tabla de contingencia):**

$x \setminus y$	y_1	y_2	y_3	$\cdots$	y_j	$\cdots$	y_l
x_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1j}	$\cdots$	n_{1l}
x_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2j}	$\cdots$	n_{2l}
x_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3j}	$\cdots$	n_{3l}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	n_{i3}	$\cdots$	n_{ij}	$\cdots$	n_{il}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	n_{k3}	$\cdots$	n_{kj}	$\cdots$	n_{kl}

Distribución de frecuencias de una variable bidimensional

- Propiedad de la frecuencia absoluta: La suma total de todas las frecuencias absolutas es el total de datos.

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = N$$

- **Frecuencia relativa** f_{ij} del valor (x_i, y_j) respecto del total de valores:

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

- Propiedad de la frecuencia relativa: La suma total de todas las frecuencias relativas es 1.

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

Ejemplo: color de pelo y ojos

- Color de pelo y ojos (ambas son variables nominales y por tanto discretas).

pelo \ ojos	marrón	azul	avellana	verde
moreno	32	11	10	3
castaño	53	50	25	15
pelirrojo	10	10	7	7
rubio	3	30	5	8

Ejemplo: alturas y pesos

- Alturas (cm) y pesos (kg) de 20 personas (ambas son variables cuantitativas y continuas):

(164, 64), (176, 77), (179, 82), (165, 62), (168, 71),
(165, 72), (186, 85), (182, 68), (173, 72), (175, 75),
(159, 81), (187, 88), (173, 72), (157, 71), (163, 74),
(171, 69), (168, 81), (173, 67), (153, 65), (182, 73)

altura \ peso	[60, 69]	[70, 79]	[80, 89]
[150, 159]	1	1	1
[160, 169]	2	3	1
[170, 179]	2	4	1
[180, 189]	1	1	2

frontera inferior: 179,5 cm,
frontera superior: 189,5 cm

frontera inferior: 69,5 kg,
frontera superior: 79,5 kg

8

Distribuciones marginales

- Para una variable bidimensional (x, y), con k valores de x y l valores de y, se llama **frecuencia absoluta marginal** de una componente al número de parejas de valores que toma la variable independientemente de la otra componente.

$$n_{x_i} = \sum_{j=1}^{l} n_{ij} \qquad n_{y_j} = \sum_{i=1}^{k} n_{ij}$$

- **Distribución marginal de *x*:** conjunto de frecuencias marginales n_{x_i} (columna más a la derecha).
- **Distribución marginal de *y*:** conjunto de frecuencias marginales n_{y_j} (fila inferior).

Distribuciones marginales

$x \setminus y$	y_1	y_2	y_3	$\cdots$	y_j	$\cdots$	y_l	Suma
x_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{x_1}
x_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{x_2}
x_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3j}	$\cdots$	n_{3l}	n_{x_3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	n_{i3}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{x_i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	n_{k3}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{x_k}
Suma	n_{y_1}	n_{y_2}	n_{y_3}	$\cdots$	n_{y_j}	$\cdots$	n_{y_l}	N

Distribuciones marginales

- Propiedad: la suma de todas las frecuencias absolutas marginales es el total de datos.

$$\sum_{i=1}^{k} n_{x_i} = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = N \qquad \sum_{j=1}^{l} n_{y_j} = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij} = N$$

- De manera equivalente, podemos definir la frecuencia relativa marginal:

$$f_{x_i} = \frac{n_{x_i}}{N} \qquad f_{y_j} = \frac{n_{y_j}}{N}$$

- Propiedad:

$$\sum_{i=1}^{k} f_{x_i} = \sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1 \qquad \sum_{j=1}^{l} f_{y_j} = \sum_{j=1}^{l}\sum_{i=1}^{k} f_{ij} = 1$$

11

Ejemplo: color de pelo y ojos

pelo \ ojos	marrón	azul	avellana	verde	Frec. marginales pelo
moreno	32	11	10	3	56
castaño	53	50	25	15	143
pelirrojo	10	10	7	7	34
rubio	3	30	5	8	46
Frec. marginales ojos	98	101	47	33	279

Ejemplo: color de pelo y ojos

pelo \ ojos	marrón	azul	avellana	verde	Frec. marginales pelo
moreno	11,5	3,9	3,6	1,1	20,1
castaño	19,0	17,9	9,0	5,4	51,3
pelirrojo	3,6	3,6	2,5	2,5	12,2
rubio	1,1	10,8	1,8	2,9	16,6
Frec. marginales ojos	35,2	36,2	16,9	11,9	100,0

Valores en %

Ejemplo: alturas y pesos

altura \ peso	[60, 69]	[70, 79]	[80, 89]	Frec. marginales alturas
[150, 159]	1	1	1	3
[160, 169]	2	3	1	6
[170, 179]	2	4	1	7
[180, 189]	1	1	2	4
Frec. marginales pesos	6	9	5	20

Ejemplo: alturas y pesos

altura \ peso	[60, 69]	[70, 79]	[80, 89]	Frec. marginales alturas
[150, 159]	5	5	5	15
[160, 169]	10	15	5	30
[170, 179]	10	20	5	35
[180, 189]	5	5	10	20
Frec. marginales pesos	30	45	25	100

Valores en %

Media aritmética a partir de frecuencias marginales

- Recordemos del tema anterior el cálculo de la media aritmética para datos agrupados de una variable unidimensional cuantitativa *x* según su frecuencia:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{N} = \sum_{i=1}^{k} f_i x_i$$

- Para una variable bidimensional (x, y), podemos calcular la **media aritmética de cada componente** con las frecuencias marginales:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_{x_i} x_i}{N} = \sum_{i=1}^{k} f_{x_i} x_i \qquad \bar{y} = \frac{\sum_{j=1}^{l} n_{y_j} y_j}{N} = \sum_{j=1}^{l} f_{y_j} y_j$$

Desviación típica a partir de frecuencias marginales

- De manera equivalente, la desviación típica de una variable unidimensional cuantitativa *x* con datos agrupados según su frecuencia:

$$s_x = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{N}}$$

- Para una variable bidimensional, entonces la **desviación típica de cada componente** es:

$$s_x = \sqrt{\frac{\sum_{i=1}^{k} n_{x_i} (x_i - \bar{x})^2}{N}} \qquad s_y = \sqrt{\frac{\sum_{j=1}^{l} n_{y_j} (y_j - \bar{y})^2}{N}}$$

- Desviación típica sin sesgo: $N - 1$ en el denominador.

Entre paréntesis, la marca de cada intervalo

Ejemplo: alturas y pesos

peso	[60,69] (64,5)	[70,79] (74,5)	[80,89] (84,5)
Frec. marginales pesos	6	9	5

$$\bar{p} = \frac{6 \cdot 64{,}5 + 9 \cdot 74{,}5 + 5 \cdot 84{,}5}{6+9+5} = 74 \text{ kg} \qquad s_p = \sqrt{\frac{6 \cdot (64{,}5-74)^2 + 9 \cdot (74{,}5-74)^2 + 5 \cdot (84{,}5-74)^2}{20}} = 7{,}4 \text{ kg}$$

altura	[150,159] (154,5)	[160,169] (164,5)	[170,179] (174,5)	[180,189] (184,5)
Frec. marginales alturas	3	6	7	4

$$\bar{a} = \frac{3 \cdot 154{,}5 + 6 \cdot 164{,}5 + 7 \cdot 174{,}5 + 4 \cdot 184{,}5}{3+6+7+4} = 170{,}5 \text{ cm}$$

$$s_a = \sqrt{\frac{3 \cdot (154{,}5-170{,}5)^2 + 6 \cdot (164{,}5-170{,}5)^2 + 7 \cdot (174{,}5-170{,}5)^2 + 4 \cdot (184{,}5-170{,}5)^2}{20}} = 9{,}7 \text{ cm}$$

Distribución condicionada

- Cuando fijamos el valor de una componente y estudiamos las frecuencias de los valores que toma la otra componente: **distribución condicionada**.
  - Distribución de x condicionada a $y = y_j$. Para un valor concreto x_i, esta frecuencia condicionada es:

$$n(x_i \mid y = y_j) = n_{ij}$$

  - Distribución de y condicionada a $x = x_i$. Para un valor concreto y_j, esta frecuencia condicionada es:

$$n(y_j \mid x = x_i) = n_{ij}$$

Distribución condicionada

- Distribución de x condicionada a $y = y_j$:

x	$n(x\|y = y_j)$	$f(x\|y = y_j)$
x_1	n_{1j}	f_{1j}
x_2	n_{2j}	f_{2j}
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	f_{ij}
$\vdots$	$\vdots$	$\vdots$
x_k	n_{kj}	f_{kj}
	n_{y_j}	1

Distribución condicionada

- ¡OJO! Para calcular las frecuencias relativas condicionadas se divide por la frecuencia marginal:

$$f(x_i \mid y = y_j) = \frac{n(x_i \mid y = y_j)}{n_{y_j}} = \frac{n_{ij}}{n_{y_j}}$$

$$f(y_j \mid x = x_i) = \frac{n(y_j \mid x = x_i)}{n_{x_i}} = \frac{n_{ij}}{n_{x_i}}$$

- Propiedades:

$$\sum_{i=1}^{k} n(x_i \mid y = y_j) = n_{y_j} \qquad \sum_{j=1}^{l} n(y_j \mid x = x_i) = n_{x_i}$$

$$\sum_{i=1}^{k} f(x_i \mid y = y_j) = 1 \qquad \sum_{j=1}^{l} f(y_j \mid x = x_i) = 1$$

21

Ejemplo: color de pelo y ojos

- Distribución de color del pelo condicionada a ojos azules:

pelo	n(pelo | ojos=azul)	f(pelo | ojos=azul)
moreno	11	10,9
castaño	50	49,5
pelirrojo	10	9,9
rubio	30	29,7
Suma	101	100,0
		en %

22

Ejemplo: color de pelo y ojos

- Distribución de color de ojos condicionada a pelo rubio:

ojos	n(ojos | pelo=rubio)	f(ojos | pelo=rubio)
marrón	3	6,5
azul	30	65,2
avellana	5	10,9
verde	8	17,4
Suma	46	100,0
		en %

Ejemplo: alturas y pesos

- Distribución de la altura condicionada a que el peso esté en el intervalo [70, 79]:

altura	n(altura | peso∈[70,79])	f(altura | peso∈ [70,79])
[150,159]	1	11,1
[160,169]	3	33,3
[170, 179]	4	44,4
[180, 189]	1	11,1
Suma	9	99,9 (por redondeo)

en %

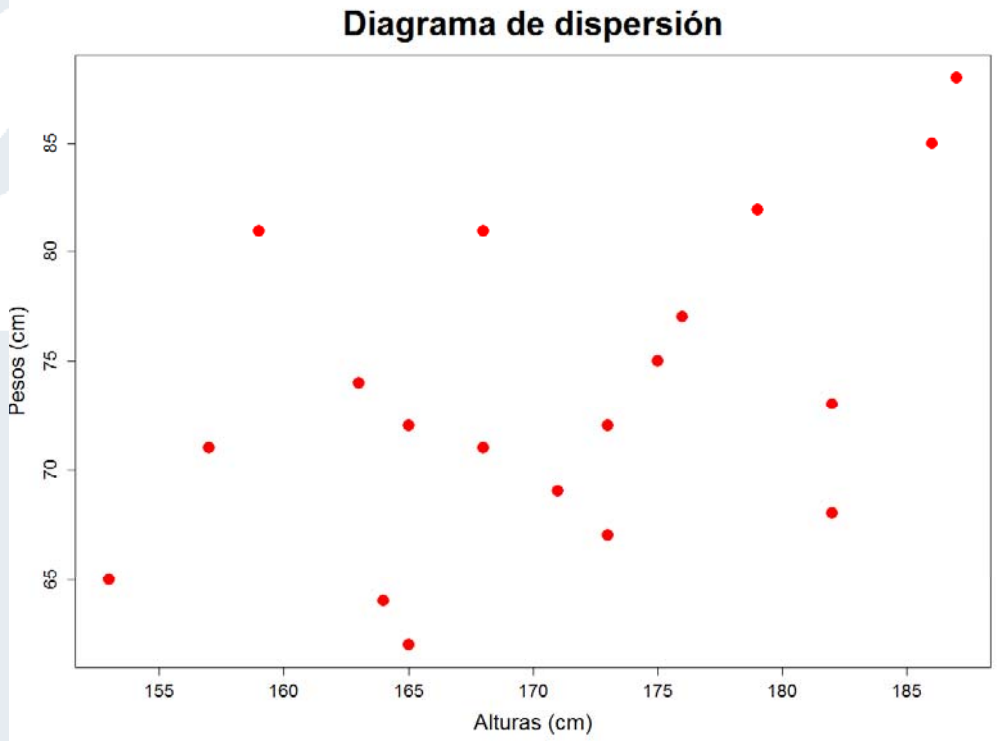
Ejemplo: alturas y pesos

- Distribución del peso condicionada a que la altura esté en el intervalo [170, 179]:

peso	n(peso | altura∈[170,179])	f(peso | altura∈[170,179])
[60,69]	2	28,6
[70,79]	4	57,1
[80,89]	1	14,3
Suma	7	100,0
		en %

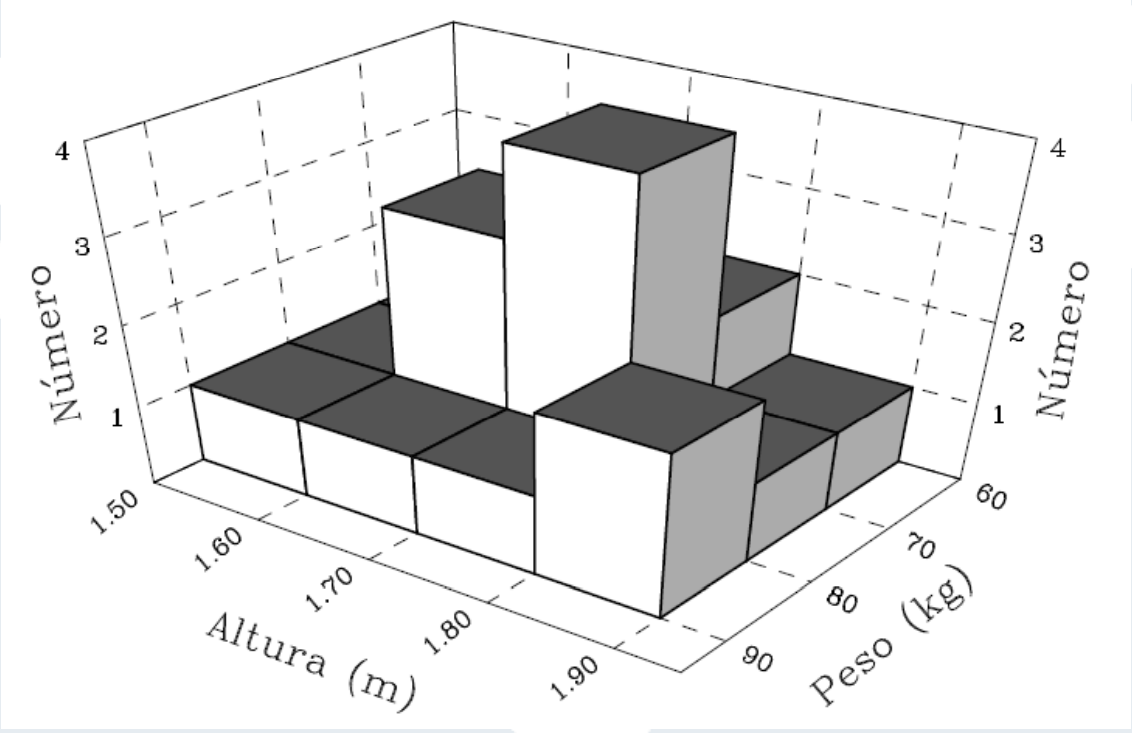
Representaciones gráficas

- Diagrama de dispersión (y frente a x):



Representaciones gráficas

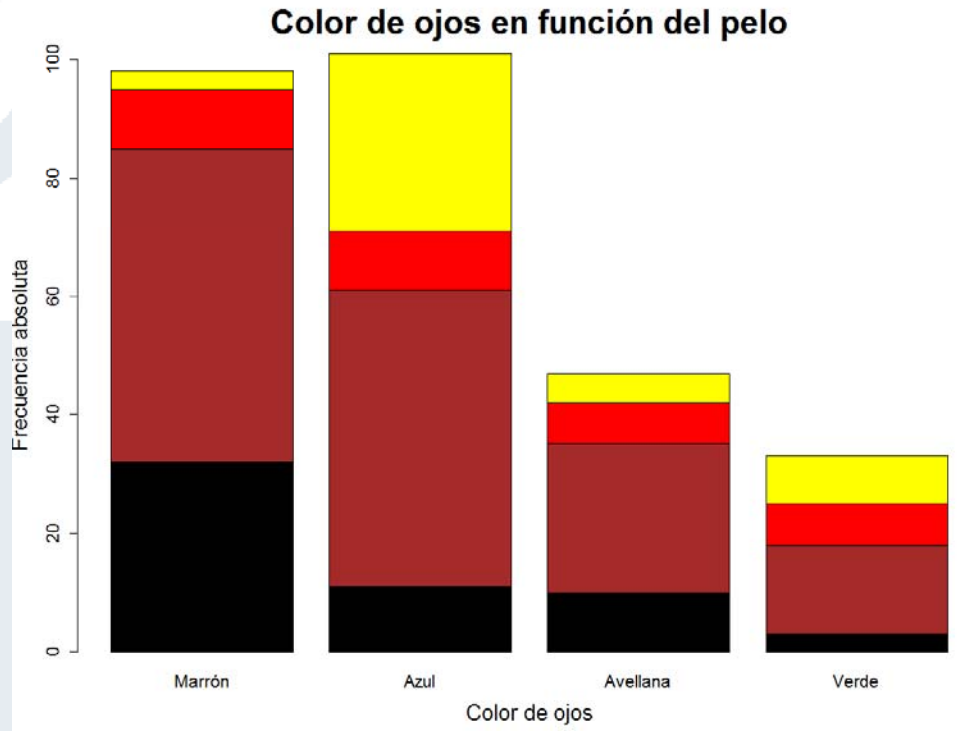
- Histograma 3D:



27

Representaciones gráficas

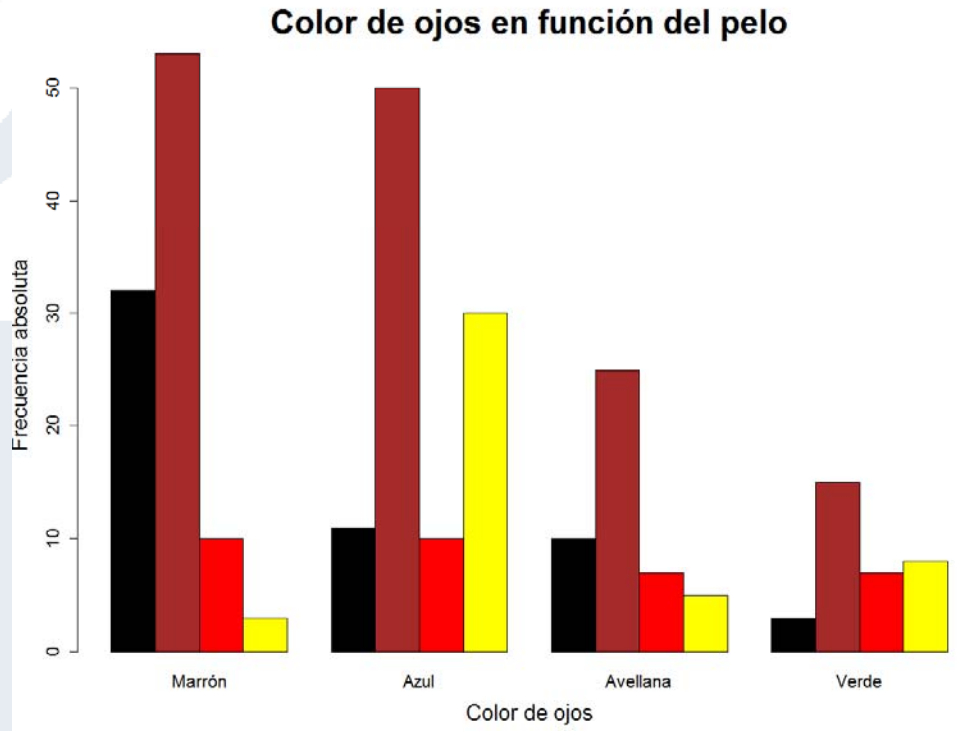
- Diagrama de barras apiladas:



28

Representaciones gráficas

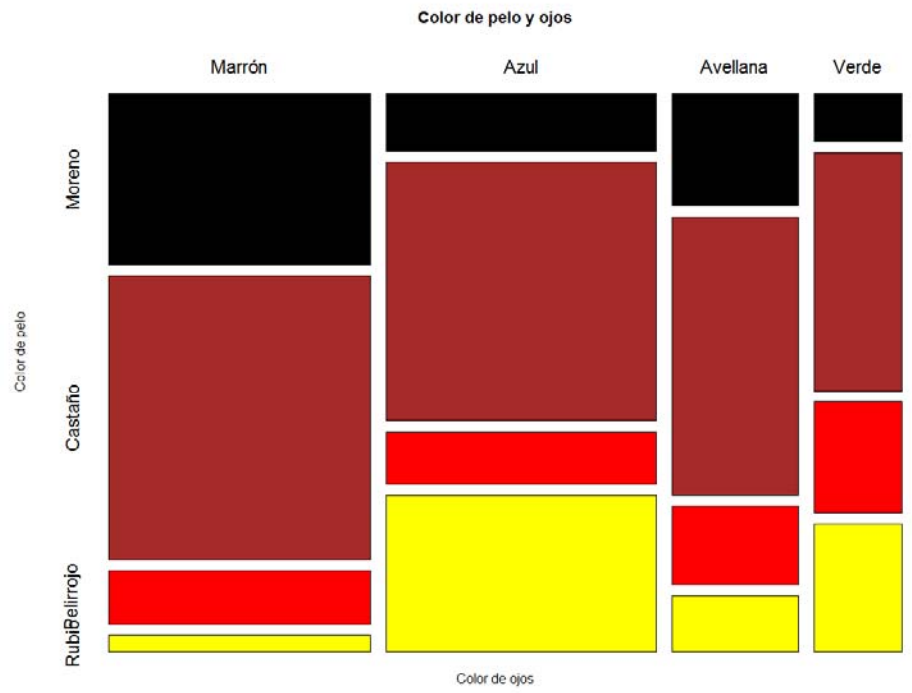
- Diagrama de barras agrupadas:



29

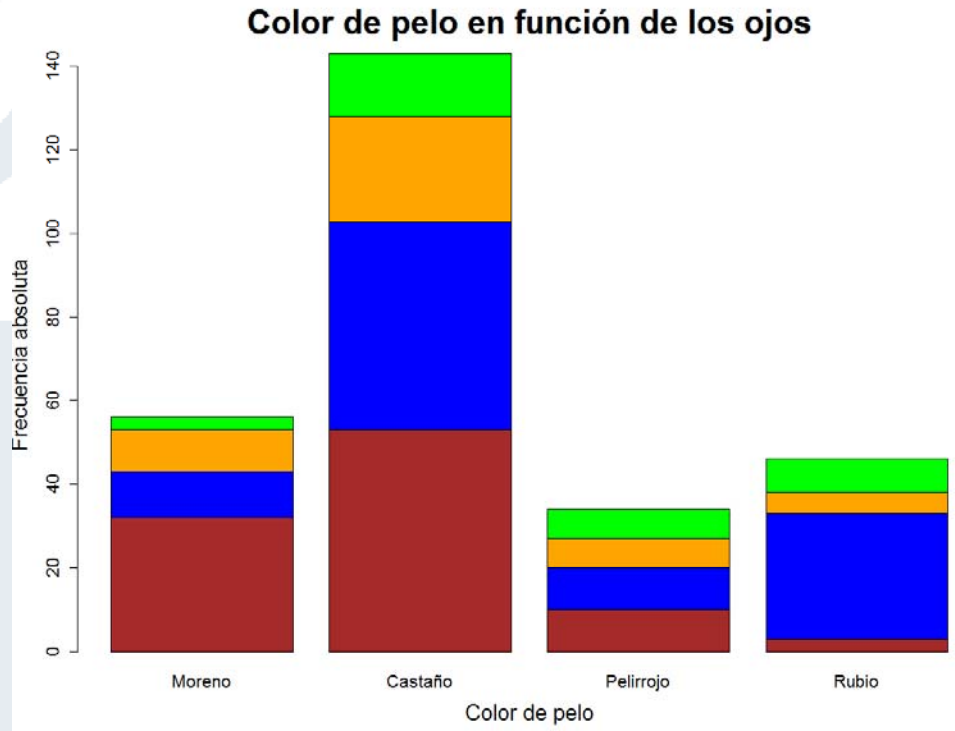
Representaciones gráficas

- Diagrama de mosaico:



Representaciones gráficas

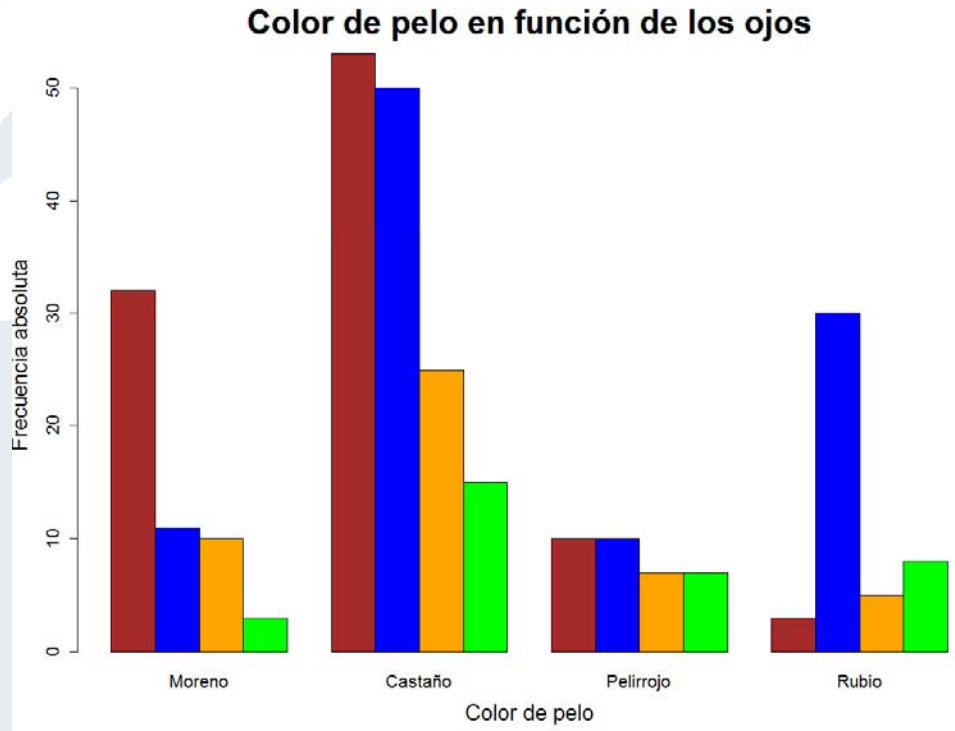
- Diagrama de barras apiladas:



31

Representaciones gráficas

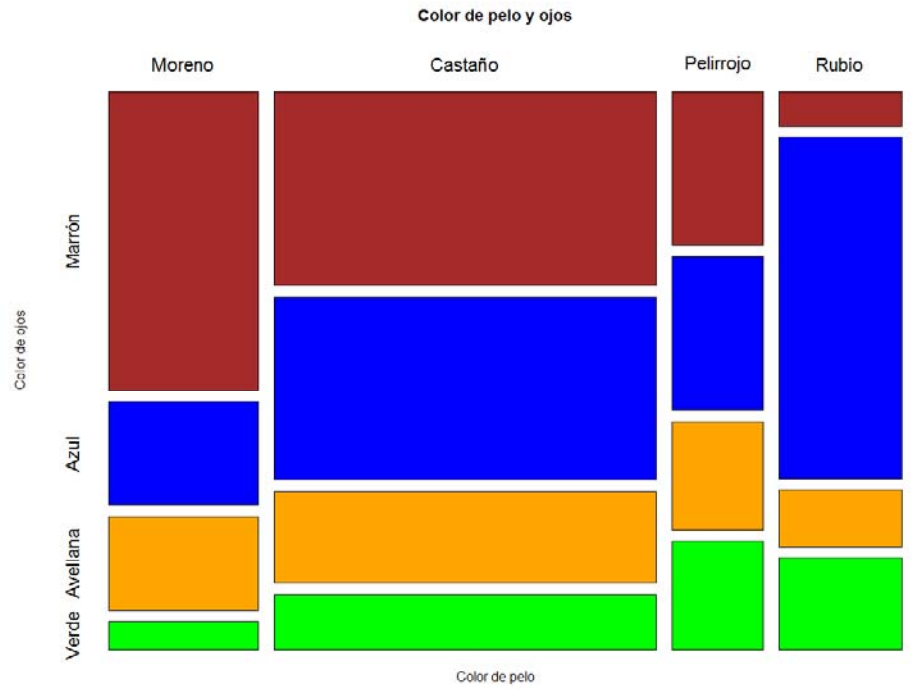
- Diagrama de barras agrupadas:



32

Representaciones gráficas

- Diagrama de mosaico:

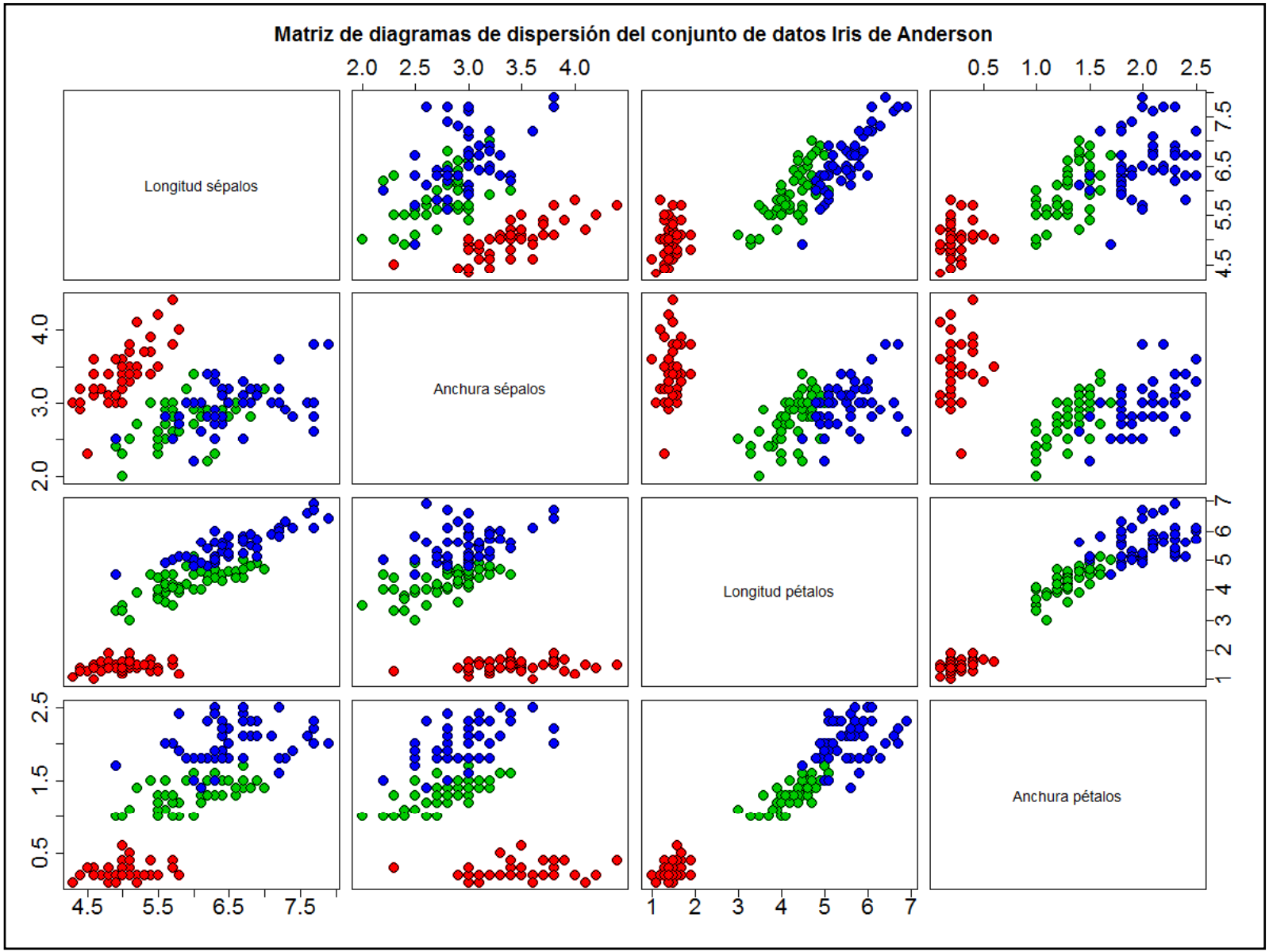


33

Representaciones gráficas

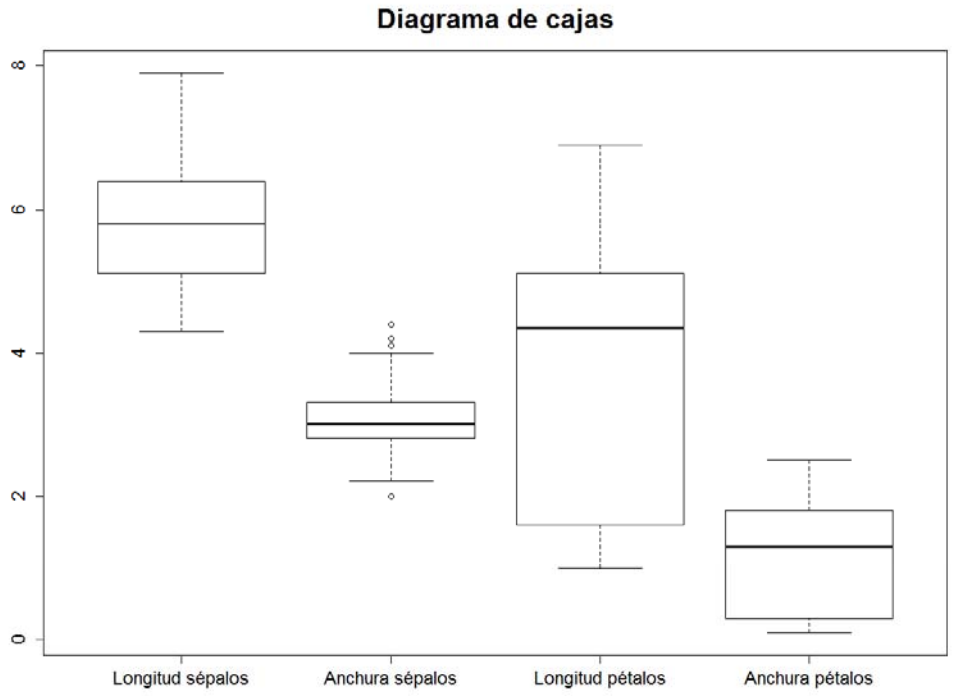
- **Matriz de diagramas de dispersión:** Para variables multidimensionales (x_1, x_2, ..., x_n), es una representación en formato matricial de diagramas de dispersión entre pares de componentes (x_i, x_j).
- Ej.: Conjunto de datos de la flor "Iris" de Anderson.
  - Longitud de sépalos (cm).
  - Anchura de sépalos (cm).
  - Longitud de pétalos (cm).
  - Anchura de pétalos (cm).
  - Especie (setosa, versicolor, virginica).

http://bit.ly/1hTyn4O


Matriz de diagramas de dispersión del conjunto de datos Iris de Anderson
Longitud sépalos
Anchura sépalos
Longitud pétalos
Anchura pétalos
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
2.0 3.0 4.0
1 2 3 4 5 6 7
0.5 1.5 2.5

Representaciones gráficas

- **Diagrama de caja y bigotes múltiple.**



36

Covarianza

- Recordemos la definición de varianza de una variable x unidimensional:

$$s_x^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \qquad s_x^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x})^2}{N}$$

- Para una variable bidimensional (x, y), se define la **covarianza** como:

$$s_{xy}^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N} \qquad s_{xy}^2 = \frac{\sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij} (x_i - \bar{x})(y_j - \bar{y})}{N}$$

- A veces el denominador se pone como $N - 1$.

Covarianza

- La covarianza mide el grado de dispersión de los valores de *x* respecto de su media de manera coordinada a la dispersión de los valores de *y* respecto de su media.
- Es una generalización de la varianza.
- La covarianza de una variable consigo misma es precisamente su varianza.

Covarianza

- Por otro lado, se puede demostrar (¡ejercicio!) que la covarianza también se puede calcular como la diferencia entre la media de los productos de los valores de *x* e *y* y el producto de las medias de *x* e *y*:

$$s_{xy}^2 = \overline{xy} - \overline{x}\cdot\overline{y} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \left(\frac{\sum_{i=1}^{N} x_i}{N}\right)\left(\frac{\sum_{i=1}^{N} y_i}{N}\right)$$

- Para ello se ha de tomar la definición de covarianza con denominador *N*.

Matriz de covarianza

- Para variables multidimensionales (x_1, x_2, ... x_n), la matriz de covarianza es una matriz cuyo elemento (i, j) es la covarianza entre la componente x_i y la componente x_j.
  - Es simétrica respecto de la diagonal principal.
  - La diagonal principal contiene las varianzas.

$$\begin{bmatrix} s^2_{x_1} & s^2_{x_1x_2} & \cdots & s^2_{x_1x_n} \\ s^2_{x_2x_1} & s^2_{x_2} & & \\ \vdots & \cdots & s^2_{x_ix_j} \cdots s^2_{x_i} \cdots & \vdots \\ s^2_{x_nx_1} & & \cdots & s^2_{x_n} \end{bmatrix}$$

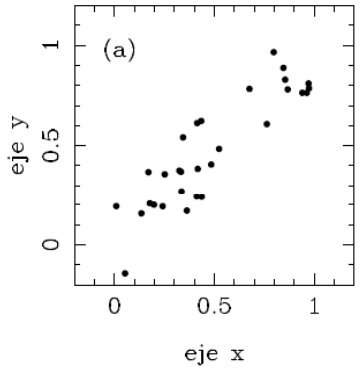
Correlación lineal

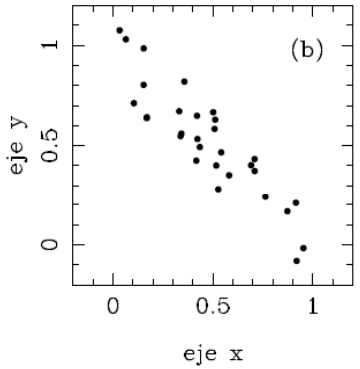
- La **correlación** es una medida de la dependencia de una variable respecto de otra:
  - Cuando una variable crece, ¿crece también la otra? ¿Decrece? ¿Permanece constante o indiferente?
- **Correlación perfecta:** existe una dependencia funcional entre ambas variables. Ej: longitud y área de un círculo.
- **Sin correlación**. Ej.: altura y apellido.
- **Correlación lineal:** estudiamos si existe alguna dependencia de tipo lineal entre ambas variables.

41

Correlación lineal

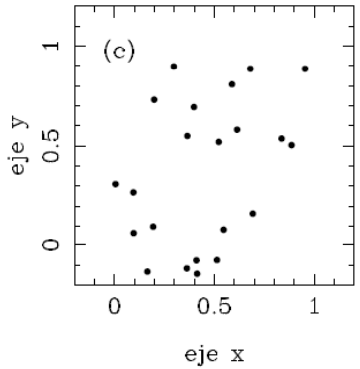
correlación positiva

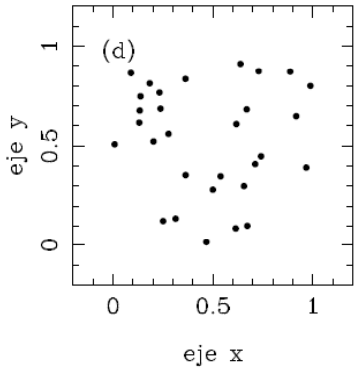




correlación negativa

correlación débilmente positiva





sin correlación

42

Correlación lineal

- **Coeficiente de correlación lineal de Pearson** entre dos variables *x* e *y*: cociente entre la covarianza y las desviaciones típicas de *x* y de *y*.

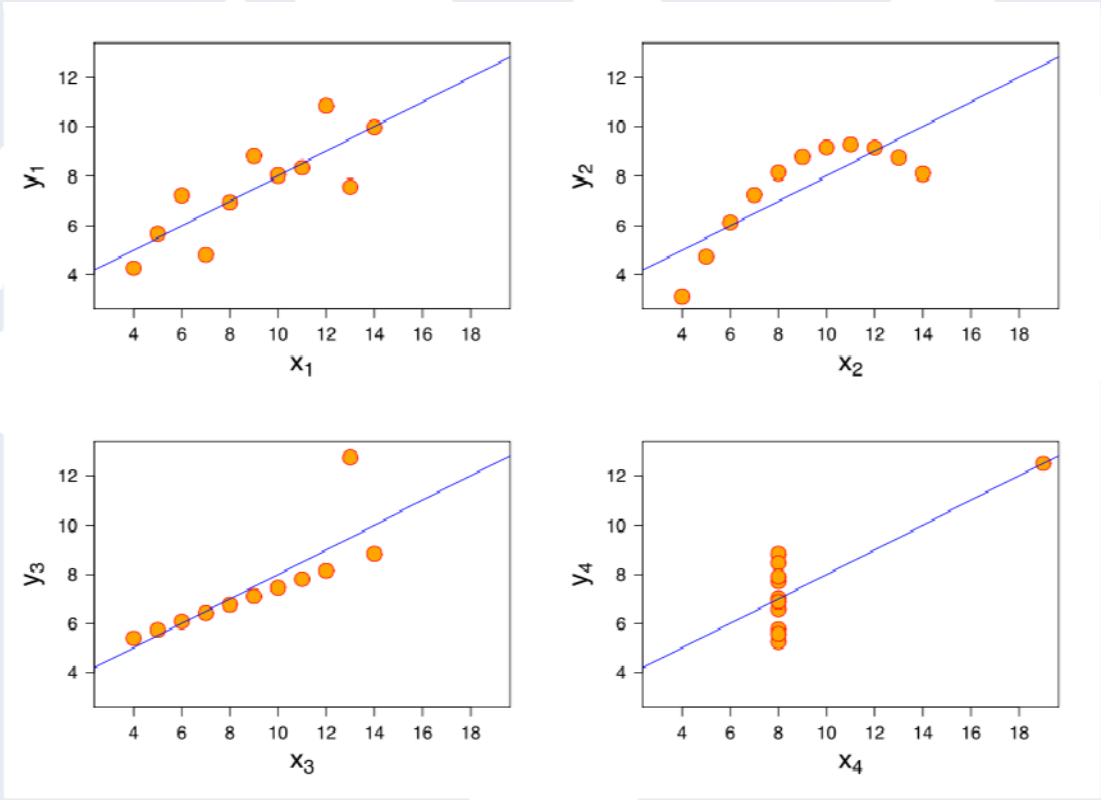
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

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

Correlación lineal

- Los valores de r están comprendidos entre -1 y 1:
  - $0 < r \leq 1$ ➔ Correlación positiva (ambas crecen).
  - $r = 0$ ➔ Sin correlación (el valor de una variable no afecta a la otra variable.
  - $-1 \leq r < 0$ ➔ Correlación negativa (una crece, la otra decrece).
- Precauciones con la correlación:
  - Correlación no implica causalidad.
  - Correlación no implica tampoco necesariamente una dependencia funcional lineal entre las variables (ver siguiente transparencia).

Correlación lineal



http://en.wikipedia.org/wiki/File:Anscombe%27s_quartet_3.svg

Matriz de correlación lineal

- Para una variable multidimensional construimos la **matriz de correlación lineal** como aquella cuyo elemento (i, j) contiene la correlación lineal entre la componente x_i y la componente x_j.
  - Matriz simétrica.
  - En la diagonal principal contiene 1 (correlación perfecta de una variable consigo misma).

$$\begin{bmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_n} \\ r_{x_2x_1} & 1 & & \\ & & \ddots & \\ \vdots & \cdots & r_{x_ix_j} \cdots 1 & \cdots \\ & & & \ddots \\ r_{x_nx_1} & & \cdots & 1 \end{bmatrix}$$

Ejemplo

- Conjunto de datos de la especie Iris de Anderson.
- Matriz de covarianza:

	Longitud sépalos	Anchura sépalos	Longitud pétalos	Anchura pétalos
Longitud sépalos	0.6856935	-0.0424340	1.2743154	0.5162707
Anchura sépalos	-0.0424340	0.1899794	-0.3296564	-0.1216394
Longitud pétalos	1.2743154	-0.3296564	3.1162779	1.2956094
Anchura pétalos	0.5162707	-0.1216394	1.2956094	0.5810063

- Matriz de correlación:

	Longitud sépalos	Anchura sépalos	Longitud pétalos	Anchura pétalos
Longitud sépalos	1.0000000	-0.1175698	0.8717538	0.8179411
Anchura sépalos	-0.1175698	1.0000000	-0.4284401	-0.3661259
Longitud pétalos	0.8717538	-0.4284401	1.0000000	0.9628654
Anchura pétalos	0.8179411	-0.3661259	0.9628654	1.0000000

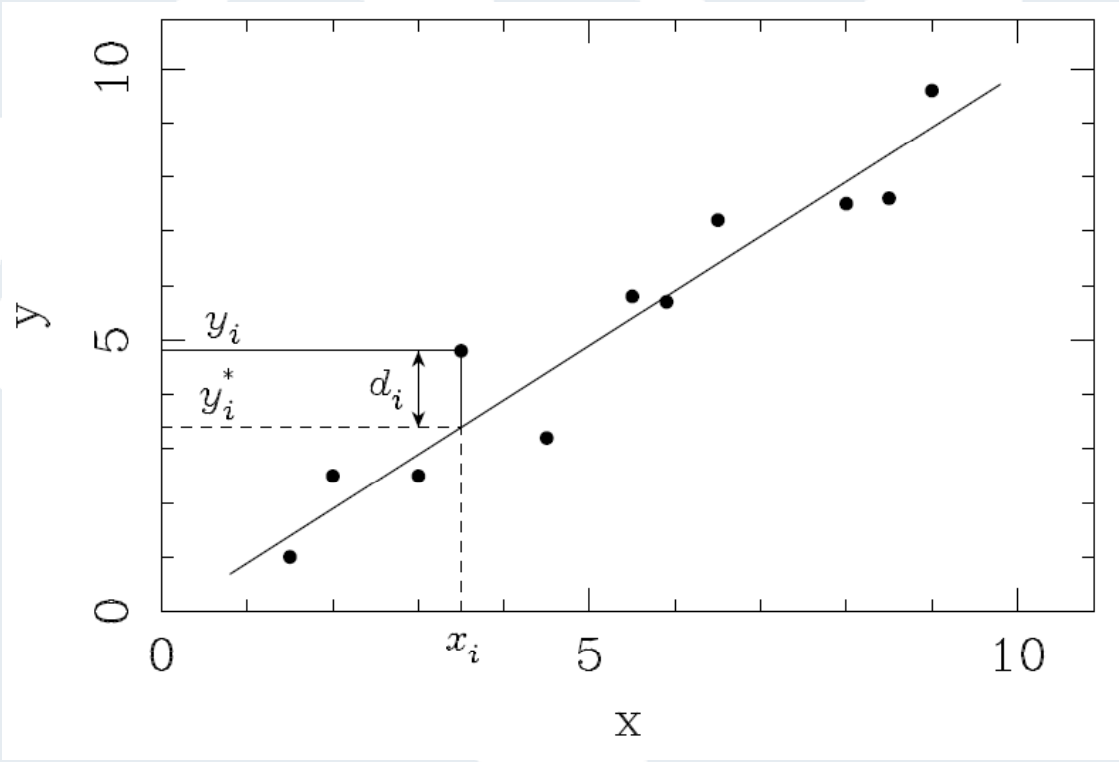
Regresión (o ajuste) lineal

- Dada una variable bidimensional (x,y), a veces interesa estudiar la posible relación funcional (matemática) entre x e y.
- En general si la función es del tipo $y = f(x)$, hablaremos de **regresión de *y* sobre *x***.
- El caso más sencillo es cuando la función es lineal, es decir, es del tipo $y = f(x) = a + b\,x$ (ecuación de una recta), siendo a y b dos parámetros que hay que calcular.
  - a es la ordenada en el origen.
  - b es la pendiente.

Método de Mínimos Cuadrados

- Método para calcular los parámetros del ajuste, en este caso, a y b.
- Datos: $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$.
- Cada x_i tiene un valor observado y_i, pero le corresponde un valor ajustado y_i^* según el ajuste.
- La distancia entre ambos valores es $d_i = y_i^* - y_i$.
- Estas distancias pueden ser positivas (cuando el dato observado queda por debajo de la recta) o negativa (cuando queda por encima).

Método de Mínimos Cuadrados



Método de Mínimos Cuadrados

- Para no tener en cuenta el signo, el método minimiza la suma de los cuadrados de las distancias, que llamaremos *M*.

$$M = \sum_{i=1}^{N} d_i^2 = \sum_{i=1}^{N} \left(y_i^* - y_i\right)^2 = \sum_{i=1}^{N} \left(a + bx_i - y_i\right)^2 = M(a,b)$$

- Para minimizar la función *M*, que depende de dos variables (*a*, *b*), hay que tomar las derivadas parciales e igualar a 0.

$$\begin{cases} \dfrac{\partial M}{\partial a} = \displaystyle\sum_{i=1}^{N} 2\left(a + bx_i - y_i\right) = 0 \\ \dfrac{\partial M}{\partial b} = \displaystyle\sum_{i=1}^{N} 2x_i\left(a + bx_i - y_i\right) = 0 \end{cases}$$

51

Método de Mínimos Cuadrados

- Simplificando la notación (todos los sumatorios van desde $i = 1$ hasta N), y desarrollando las ecuaciones, obtenemos:

$$\begin{cases} \sum (a + bx_i - y_i) = 0 \\ \sum (ax_i + bx_i^2 - x_i y_i) = 0 \end{cases} \qquad \begin{cases} \sum a + \sum bx_i - \sum y_i = 0 \\ \sum ax_i + \sum bx_i^2 - \sum x_i y_i = 0 \end{cases}$$

- Surge un sistema de ecuaciones lineales con dos incógnitas: a y b.

$$\begin{cases} aN + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{cases}$$

52

Método de Mínimos Cuadrados

- La solución del sistema de ecuaciones es:

$$\begin{cases} a = \dfrac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \\ b = \dfrac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \end{cases}$$

- Se puede reescribir como:

$$\bar{x} = \frac{\sum x_i}{N},\ \bar{y} = \frac{\sum y_i}{N} \qquad \begin{cases} a = \bar{y} - b\bar{x} \\ b = \dfrac{\frac{1}{N}\sum x_i y_i - \bar{x}\,\bar{y}}{\frac{1}{N}\sum x_i^2 - \bar{x}^2} = \dfrac{s_{xy}^2}{s_x^2} \end{cases}$$

53