Exploración de datos

Mathieu Kessler

Departamento de Matemática Aplicada y Estadística
Universidad Politécnica de Cartagena

Cartagena, Enero 2010

Guión

Guión

La estadística utiliza datos para conseguir comprensión sobre un fenómeno.

- Combinación entre conocimientos previos y nuestro uso de gráficas y cálculos $\Rightarrow$ información.
- Grandes conjuntos de datos: más información disponible pero difícil de extraer
- Es fundamental un primer paso: Exploración de datos

Guión

Unos cuantos términos:

- Un conjunto de datos describe individuos. Éstos pueden ser personas o objetos.
- Asociados a un conjunto: variables. Distinguimos dos tipos de variables:
  - variable cuantitativa: asocia un número a cada individuo.
  - variable cualitativa: coloca a cada individua en una categoría

  Ejemplo de variables asociadas a la clase: peso, altura, sexo, edad, grupo sanguíneo.
- Un concepto fundamental: la distribución de una variable X en el conjunto. Establecemos la lista de los valores de X junto con su frecuencia.
  Frecuencia absoluta: número de veces que aparece
  Frecuencia relativa: proporción de veces que aparece.

Ejemplo

Distribución del grupo sanguíneo en una clase:

Grupo	Frec. absoluta	Frec. relativa
A	51	51/145=0.35
B	19	0.13
O	5	0.03
AB	70	0.49

Guión

Representaciones gráficas: una herramienta fundamental

Para variable cualitativa:

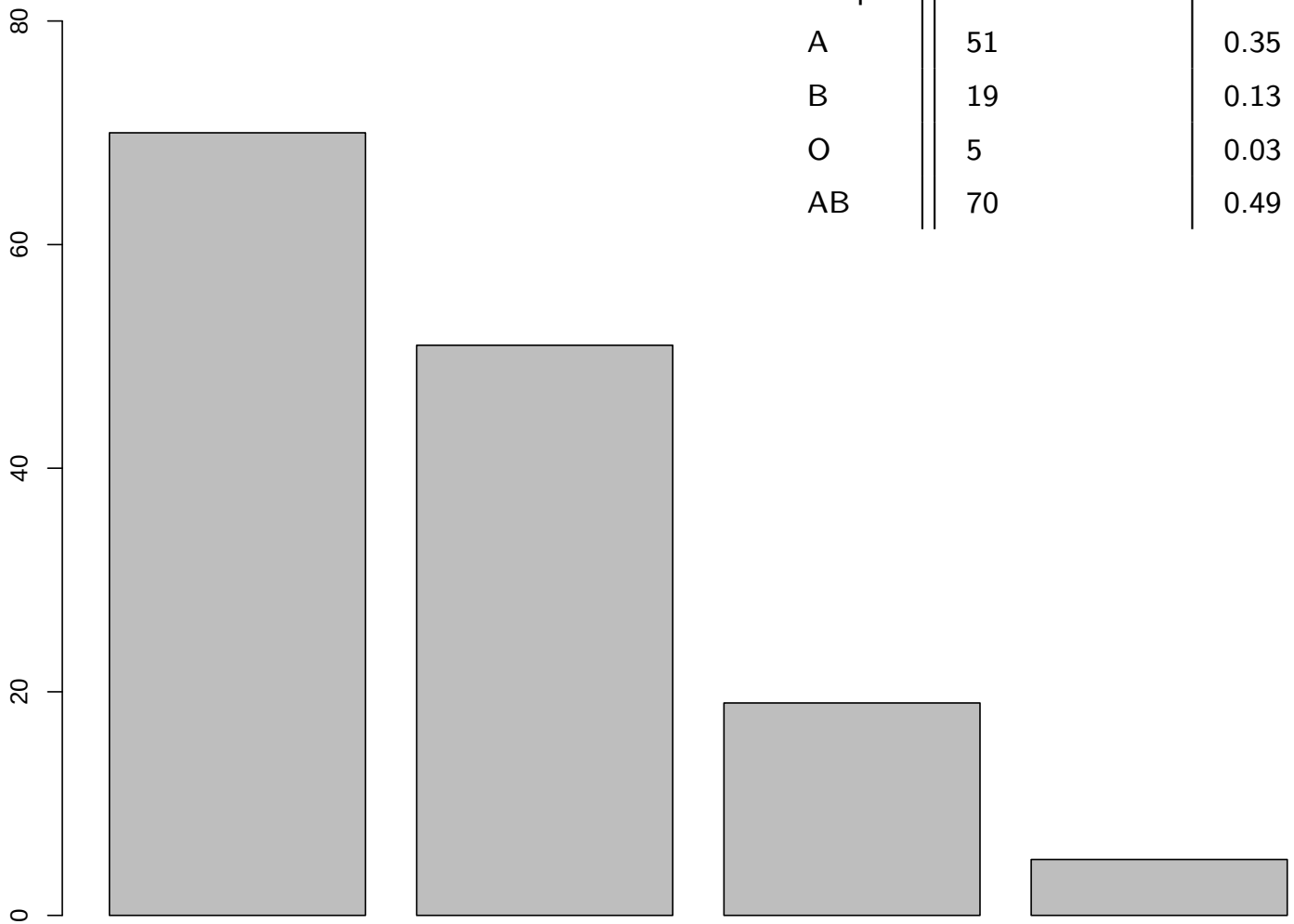
- Diagrama de barras
- Diagrama de sectores

Para variable cuantitativa:

- Diagrama de barras
- Histograma
- Gráfica de densidad

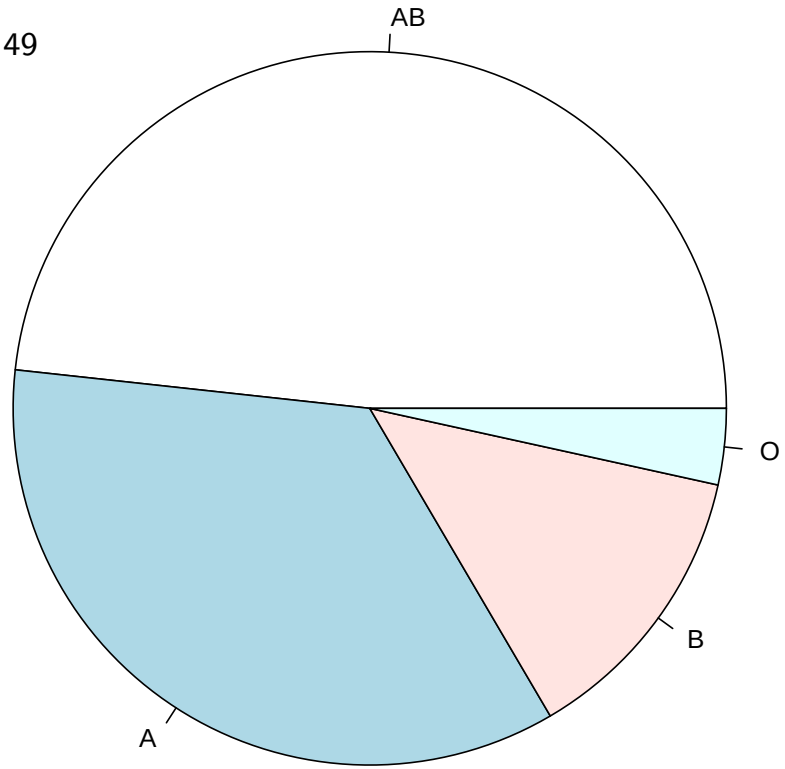
Grupo sanguíneo: diagrama de barras

Grupo	Frec. absoluta	Frec. relativa
A	51	0.35
B	19	0.13
O	5	0.03
AB	70	0.49



Grupo sanguíneo: diagrama de sectores

Grupo	Frec. absoluta	Frec. relativa
A	51	0.35
B	19	0.13
O	5	0.03
AB	70	0.49



Un primer ejemplo: mediciones de nitrato

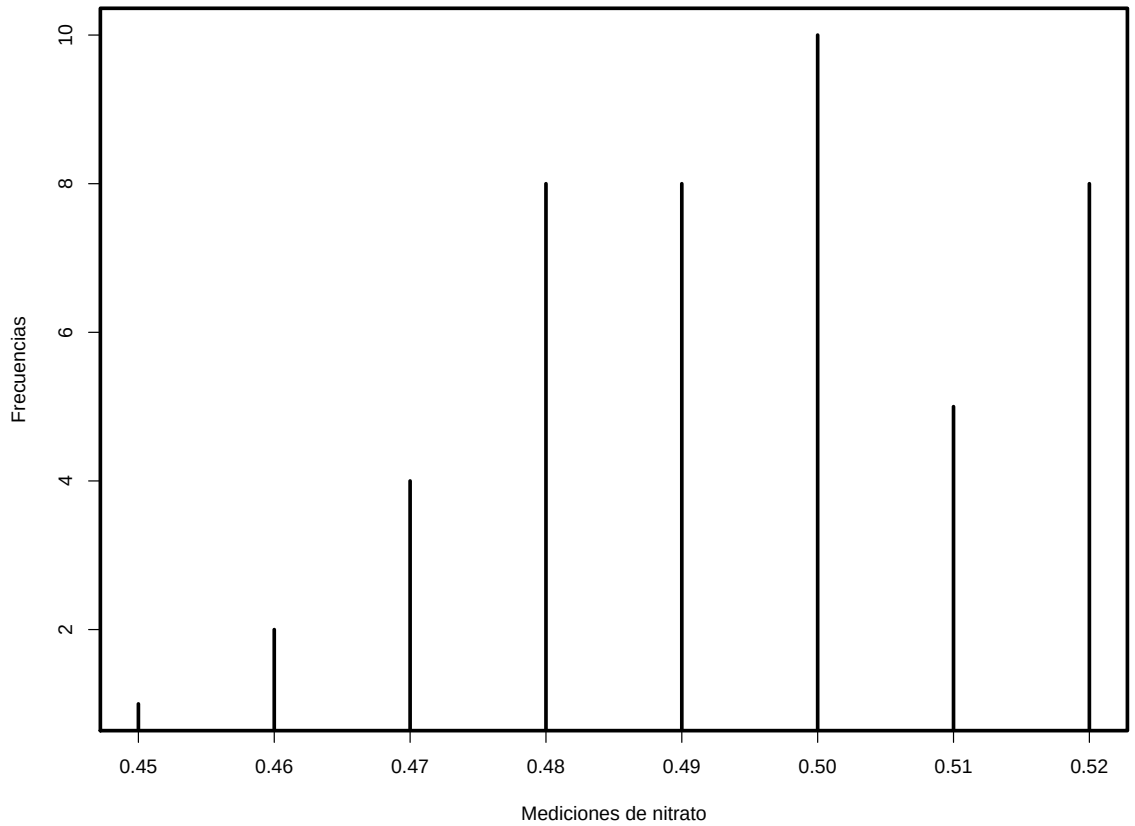
Mediciones del contenido en nitrato de una muestra de agua:

Valor	Frecuencia	Valor	Frecuencia
0.45	1	0.49	8
0.46	2	0.50	10
0.47	4	0.51	5
0.48	8	0.51	8

Valores distintos: 8, Número de datos: 46

Mediciones de nitrato: diagrama de barras

Valor	Frec.	Valor	Frec.
0.45	1	0.49	8
0.46	2	0.50	10
0.47	4	0.51	5
0.48	8	0.51	8



Ejemplo: mediciones de la velocidad de la luz

Newcomb and Michelson consiguieron una estimación bastante precisa de la velocidad de la luz en ... 1882
Midieron el tiempo que tarda la luz en recorrer una distancia de 7400m.
En nanosegundos, *tiempo* – 24800:

28, 26, 33, 24, 34, -44, 27, 16, 40, -2, 29, 22, 24, 21, 25, 30, 23, 29, 31, 19, 24, 20, 36, 32, 36, 28, 25, 21, 28, 29, 37, 25, 28, 26, 30, 32, 36, 26, 30, 22, 36, 23, 27, 27, 28, 27, 31, 27, 26, 33, 26, 32, 32, 24, 39, 28, 24, 25, 32, 25, 29, 27, 28, 29, 16, 23

Dos preguntas

- ¿Por qué repitieron tantas veces las mediciones?
- ¿Qué hacer con estos datos? ¿Cuál es el valor que damos como la velocidad de la luz?

Tabla de frecuencias e histograma

- El conjunto presenta muchos valores distintos pero próximos $\Rightarrow$ agrupamos los datos en clases.
- Ordenamos los datos, dividimos el rango en clases, colocamos cada dato en su clase.
- Realizamos el recuento de las frecuencias de cada clase.

¿Cuántas clases escoger

- Un problema sin solución perfecta
- Una regla muy usada: la regla de Sturges:

$$1 + \log_2(n)$$

Recordar: n: número de datos, $\log_2(n) = \ln(n)/\ln(2)$

Mediciones de la velocidad de la luz: datos ordenados

Pos.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Dato	-44	-2	16	16	19	20	21	21	22	22	23	23	23
Pos.	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
Dato	24	24	24	24	24	25	25	25	25	25	26	26	26
Pos.	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**
Dato	26	26	27	27	27	27	27	27	28	28	28	28	28
Pos.	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**	**52**
Dato	28	28	29	29	29	29	29	30	30	30	31	31	32
Pos.	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**	**63**	**64**	**65**
Dato	32	32	32	32	33	33	34	36	36	36	36	37	39
Pos.	**66**												
Dato	40												

Clases de amplitud 5 empezando en -45 y acabando en 40:

Clase	Frec.	Clase	Frec.	Clase	Frec.
$]-45,-40]$	1	$]-15,-10]$	0	$]15,20]$	4
$]-40,-35]$	0	$]-10,-5]$	0	$]20,25]$	17
$]-35,-30]$	0	$]-5,0]$	1	$]25,30]$	26
$]-30,-25]$	0	$]0,5]$	0	$]30,35]$	10
$]-25,-20]$	0	$]5,10]$	0	$]35,40]$	7
$]-20,-15]$	0	$]10,15]$	0		

Completamos la tabla con las frecuencias acumuladas:

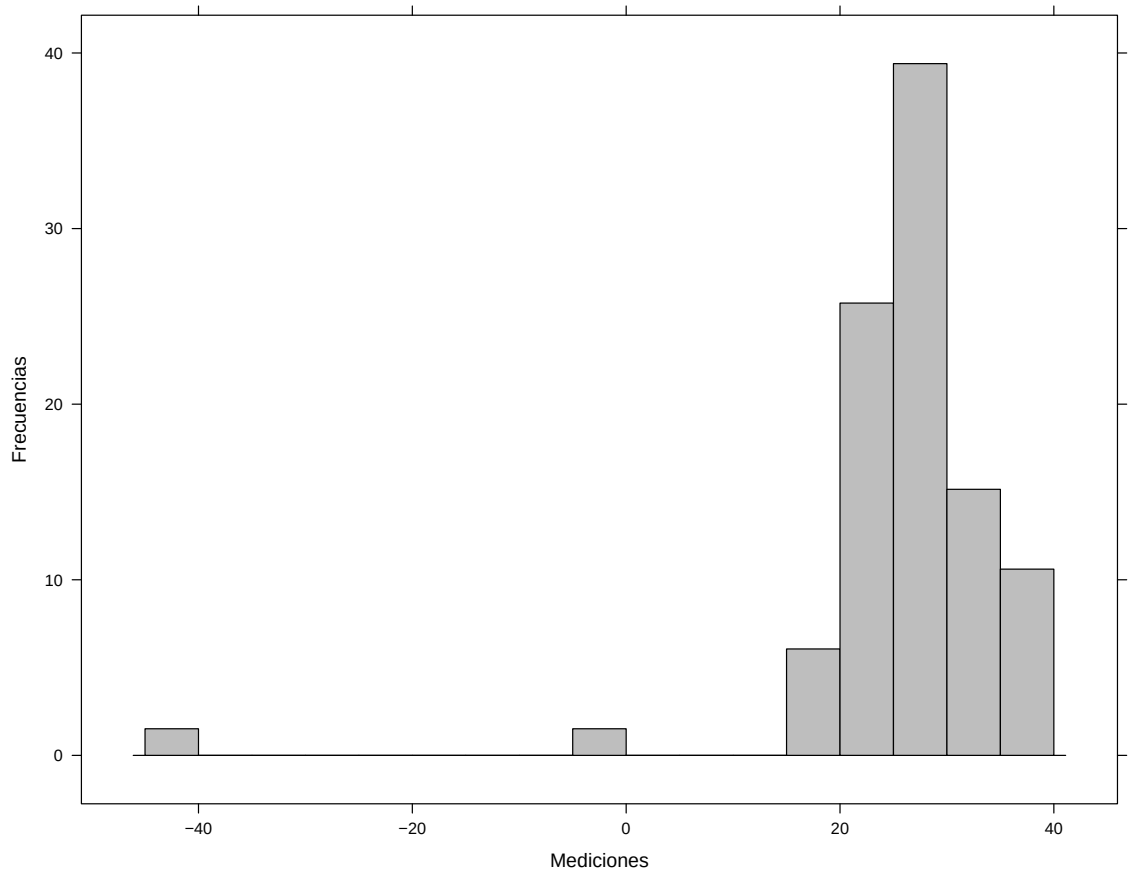
Definición de frecuencia acumulada

La frecuencia absoluta (relativa) acumulada de una clase es el número (proporción) de datos que pertenecen a esta clase o a alguna clase anterior.

Gráficas para una variable cuantitativa

Clase	Frecuencias		Frec. Acumuladas	
	Absolutas	Relativas(%)	Absolutas	Relativas(%)
$]-45,-40]$	1	1.5	1	1.5
$]-40,-35]$	0	0.0	1	1.5
$]-35,-30]$	0	0.0	1	1.5
$]-30,-25]$	0	0.0	1	1.5
$]-25,-20]$	0	0.0	1	1.5
$]-20,-15]$	0	0.0	1	1.5
$]-15,-10]$	0	0.0	1	1.5
$]-10,-5]$	0	0.0	1	1.5
$]-5,0]$	1	1.5	2	3.0
$]0,5]$	0	0.0	2	3.0
$]5,10]$	0	0.0	2	3.0
$]10,15]$	0	0.0	2	3.0
$]15,20]$	4	6	6	9
$]20,25]$	17	25.7	23	34.7
$]25,30]$	26	39.3	49	74
$]30,35]$	10	15.3	59	89.3
$]35,40]$	7	10.7	66	100
TOTAL	66	100.0		

Representación gráfica de la tabla de frecuencia: el histograma



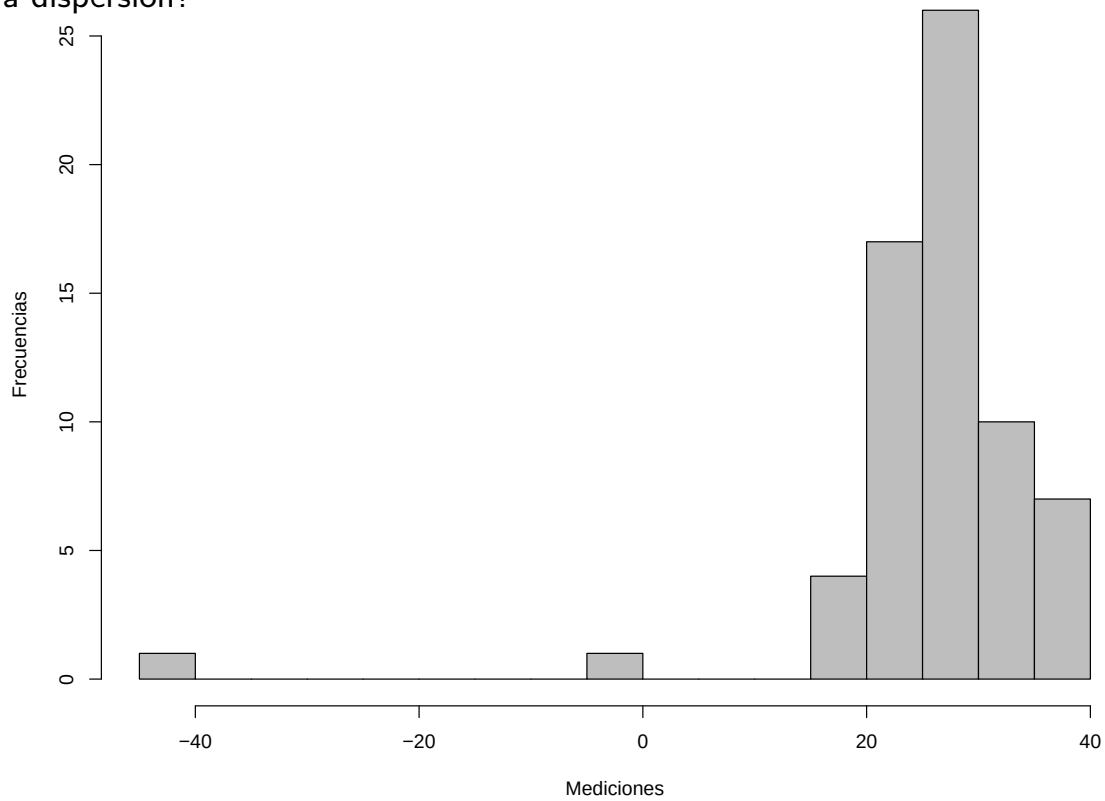
Cómo interpretar un histograma

Nos fijamos en:

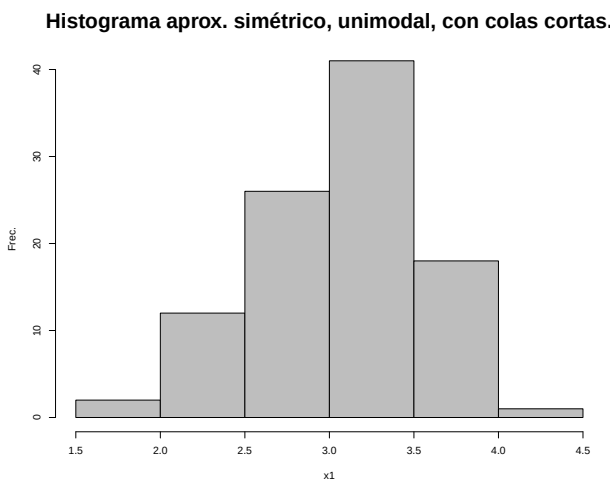
1. ¿Es la distribución simétrica?
2. ¿Tiene la distribución colas largas?
3. ¿Tiene un único máximo claro? (Histograma unimodal)
4. ¿Aparecen datos atípicos? *Un dato atípico es un dato que se aleja del patrón global del conjunto*
5. ¿Dónde está aprox. el centro de la distribución?
6. ¿Presentan los datos mucha dispersión?

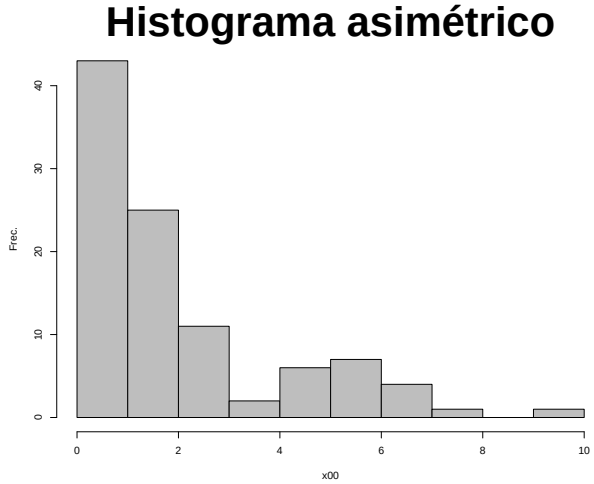
Gráficas para una variable cuantitativa

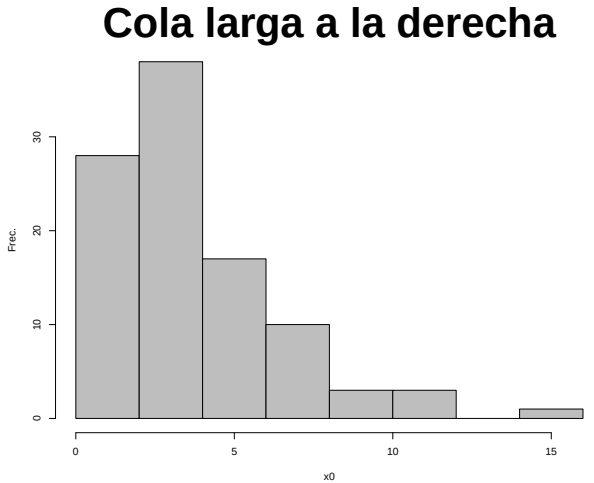
1. ¿Distribución simétrica?
2. ¿Colas largas?
3. ¿Unimodal?
4. ¿datos atípicos?
5. ¿centro aprox. de la distribución?
6. ¿Presentan los datos mucha dispersión?

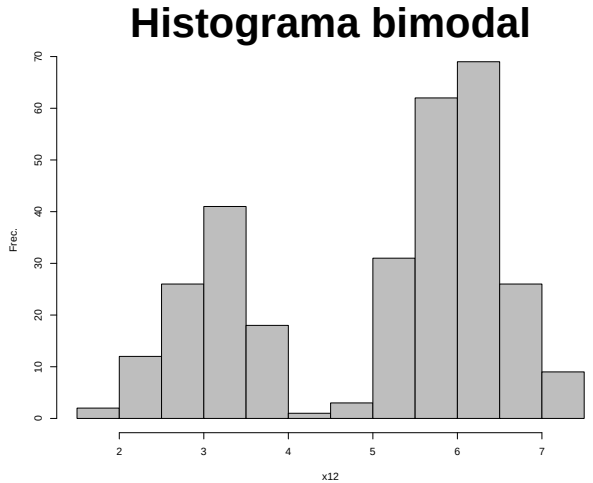


Ejemplos de histogramas



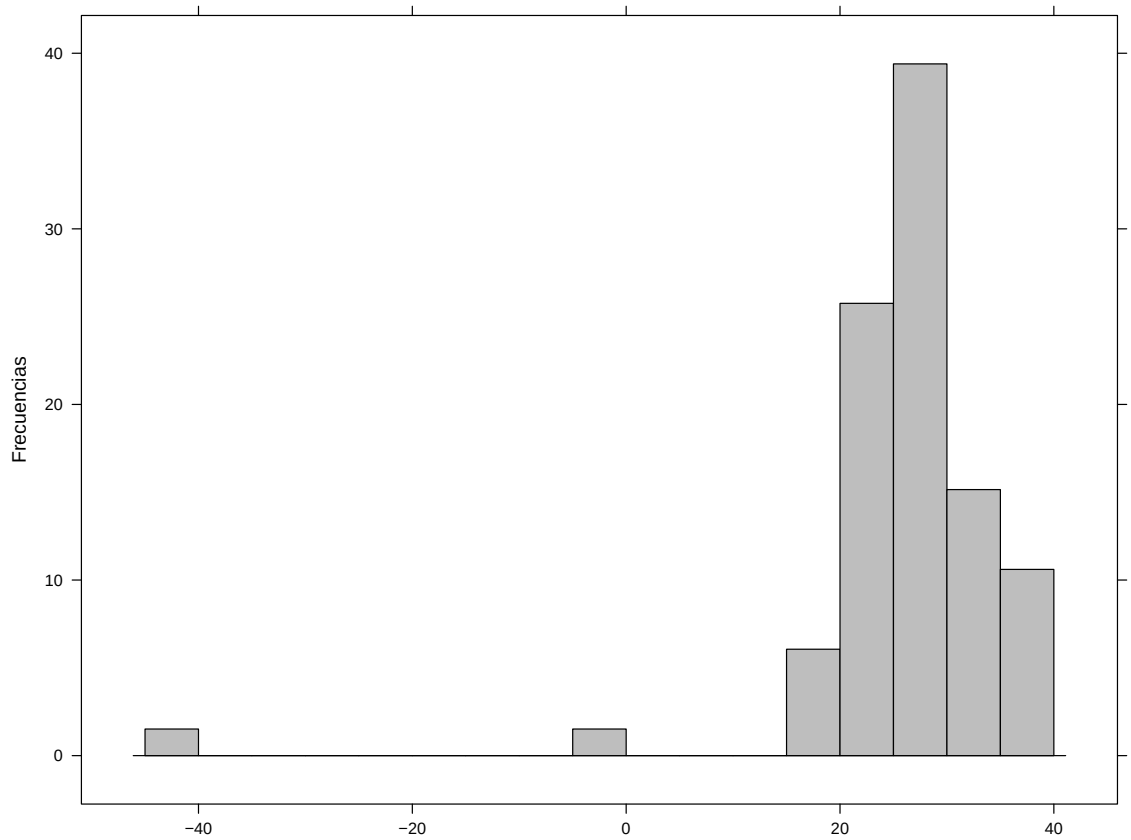






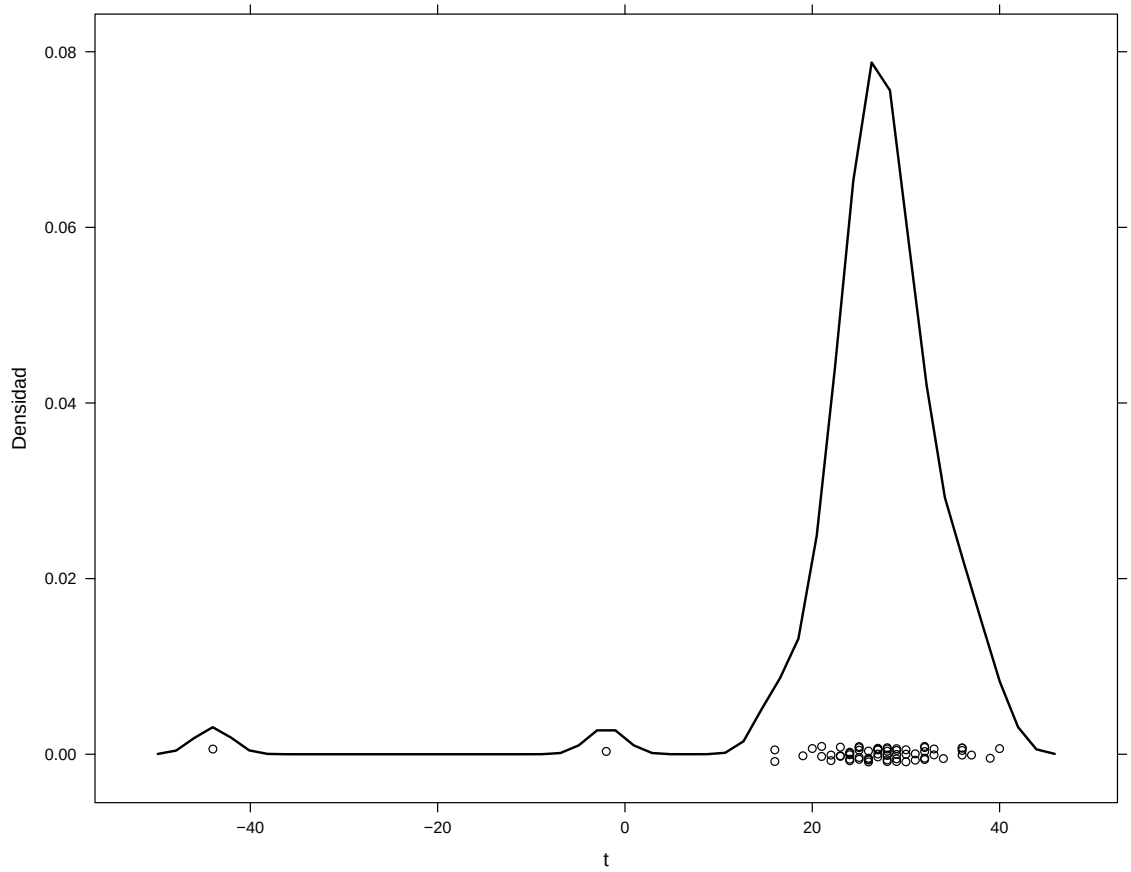
Gráfica de densidad: ejemplo de las mediciones de la luz:

Buscamos visualizar la densidad de los datos, según las regiones, partiendo del histograma:



Gráfica de densidad: ejemplo de las mediciones de la luz:

Obtenemos la gráfica de densidad (en este caso con R):



Guión

Buscamos resúmenes de las características de la distribución

Para variable cuantitativa, calculamos medidas numéricas que buscan contestar a las preguntas planteadas ante el histograma. Veremos:

- Medidas de centro
- Medidas de dispersión
- Un resumen visual: el diagrama de cajas-bigotes.

La primera medida de centro: la media

Cálculo

- Datos: $x_1, \ldots, x_n$, la media es

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

- Datos ya agrupados: tenemos los valores distintos $x_1, \ldots, x_m$ junto con sus frecuencias $n_1, \ldots, n_m$, la media es

$$\bar{x} = \frac{n_1 \cdot x_1 + \cdots + n_m \cdot x_m}{(n_1 + \ldots + n_m)}.$$

Aspectos de la media

IMPORTANTE

- La media se interpreta como el centro de gravedad de los datos.
- $\Rightarrow$ es muy sensible a datos atípicos.

Otra medida de centro: la mediana

¿Qué es la mediana?

La mediana es el punto que deja el 50% de los datos a su izquierda y el otro 50% a su derecha.

¿Cómo se calcula?

Si tenemos n datos, $x_1, x_2, \ldots, x_n$, ordenamos los datos por orden creciente. La mediana es el dato ordenado n^o $(n+1)/2$.

Ejemplos

- 125, 129, 134, 185, 200 Me es el dato ordenado número 3, $\Rightarrow Me = 134$.
- 11, 15, 20, 23: Me es el dato ordenado n^o 2.5, (¿?)$\Rightarrow$, por convención, punto intermedio entre el dato n^o 2 y el dato n^{o}3.$\Rightarrow$ $Me = 17.5$.

La mediana no es sensible a datos atípicos:

- 125, 129, 134, 185, 200 Me es el dato ordenado número 3, $\Rightarrow Me = 134$
- 125, 129, 134, 185, 2000 Me es el dato ordenado número 3, $\Rightarrow Me = 134$
- 125, 129, 134, 185, 20000000 Me es el dato ordenado número 3, $\Rightarrow Me = 134$

La desviación típica

- Mide lo "lejos" que están situados los datos respecto a su centro de gravedad (la media)
- Se denota por s, es la raíz cuadrada de la varianza s^2,

$$s = \sqrt{s^2}.$$

Cálculo de la varianza

- Definición:

$$s^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}.$$

- Fórmula alternativa:

$$s^2 = \frac{n}{n-1}(\overline{x^2} - (\bar{x})^2)$$

Ejemplo de cálculo

La fórmula alternativa:

$$s^2 = \frac{n}{n-1}(\overline{x^2} - (\bar{x})^2)$$

- $\overline{x^2}$: elevamos todos los datos al cuadrado y después calculamos su media.
- $(\bar{x})^2$: calculamos la media de los datos y después la elevamos al cuadrado.

Ejemplo

Datos: 4, 5.5, 6.5, 8.
$\overline{x^2} = (4^2 + 5.5^2 + 6.5^2 + 8^2)/4 = 38.125.$
$\bar{x} = (4 + 5.5 + 6.5 + 8)/4 = 6 \Rightarrow (\bar{x})^2 = 36.$
Deducimos $s^2 = 2.8333$ y $s = 1.683251$

Algunos aspectos de la desviación típica

- La desviación típica es representativa de la dispersión del conjunto de datos solo si la media es representativa de su centro.
- Unidades de la varianza y de la desviación típica.

El rango intercuartílico

Cuartiles y percentiles

- La mediana separa el conjunto en dos partes de mismo tamaño.
- Los cuartiles separan el conjunto en 4 partes de mismo tamaño.
- Los percentiles separan el conjunto en 100 partes de mismo tamaño.

Cuartiles

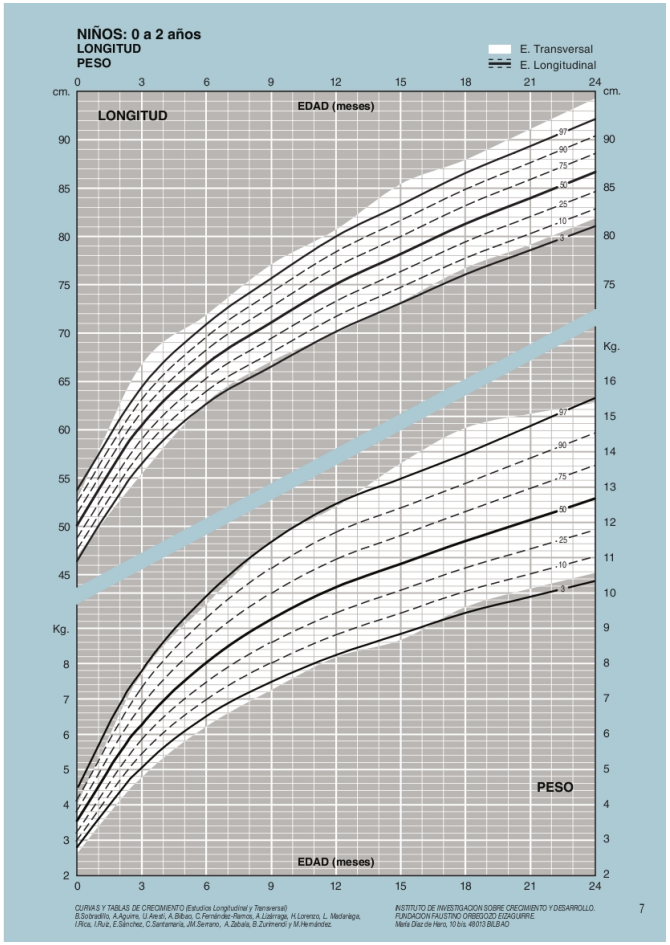
- Q_1: primer cuartil. Deja el 25% de los datos ordenados a su izquierda.
- Q_3: tercer cuartil. Deja el 75% de los datos ordenados a su izquierda.
- ¿y Q_2? segundo cuartil. $Q_2 = Me$.

Percentiles

Percentil k

Si k es un entero, entre 0 y 100, P_k deja el $k\%$ de los datos ordenados a su izquierda.

Ejemplo de percentiles: curvas de crecimiento



Fuente: Fundación Faustino Orbegozo Eizaguirre

El rango intercuartílico: RIC

- $RIC = Q_3 - Q_1$.
- Mide la dispersión de los datos

También sirve para detectar atípicos:

Se considera posible atípico un data menor de $Q_1 - 1.5 \times RIC$, o mayor de $Q_3 + 1.5 \times RIC$.

El diagrama de caja-bigotes (boxplot)

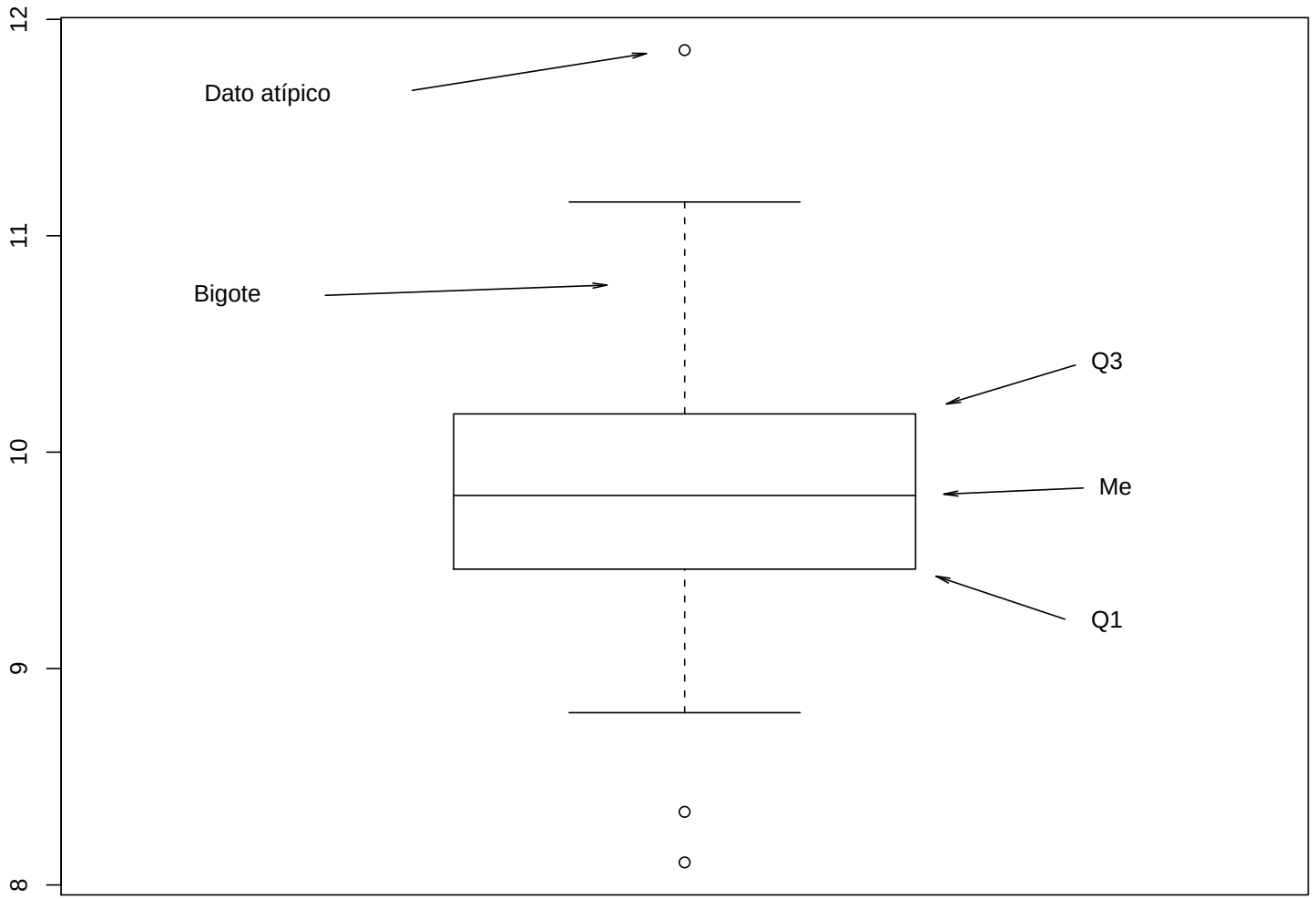
Permite visualizar la tendencia central, la dispersión, los datos atípicos.

Los componentes del diagrama de caja-bigotes

En un eje vertical:

- Tres segmentos horizontales y paralelos a la altura de Q_1, Me y Q_3. Se cierra la caja resultante.
- Dos segmentos verticales (bigotes) de una longitud máxima de $1.5 \times RIC$. Se recortan hasta los últimos datos del conjunto que no sean atípicos.

Un resumen gráfico: el diagrama de caja-bigotes



Muy útil para comparar subconjuntos

Calificaciones de los aprobados en la prueba de acceso, Distrito Único de la Región de Murcia