Estadística Descriptiva

Datos

Número	Consumo	Cilindrada	Potencia	Peso	Aceleración	Año	País	Nº Cilindros
	l/100Km	*cc*	*CV*	*kg*	*segundos*			
1	15	4982	150	1144	12	70	EEUU	8
2	16	6391	190	1283	9	70	EEUU	8
3	24	5031	200	1458	15	70	EEUU	8
4	9	1491	70	651	21	71	EEUU	4
5	11	2294	72	802	19	71	EEUU	4
6	17	5752	153	1384	14	71	EEUU	8
7	12	2294	90	802	20	72	EEUU	4
8	17	6555	175	1461	12	72	EEUU	8
9	18	6555	190	1474	13	72	EEUU	8
10	12	1147	97	776	14	72	Japón	3
11	16	5735	145	1360	13	73	EEUU	8
12	12	1868	91	860	14	73	Europa	4
13	9	2294	75	847	17	74	EEUU	4
14	8	1295	67	666	16	74	Europa	4
15	7	1163	65	612	21	74	Japón	4
16	7	1360	61	667	19	74	Japón	4
17	12	3802	90	1070	17	75	EEUU	6
18	13	3687	95	1261	19	75	EEUU	6
19	9	1475	71	741	17	75	Europa	4
20	9	1983	115	890	14	75	Europa	4
...	...	...	...	...	...	...	...	...
391	7	1753	75	735	15	82	Japón	4

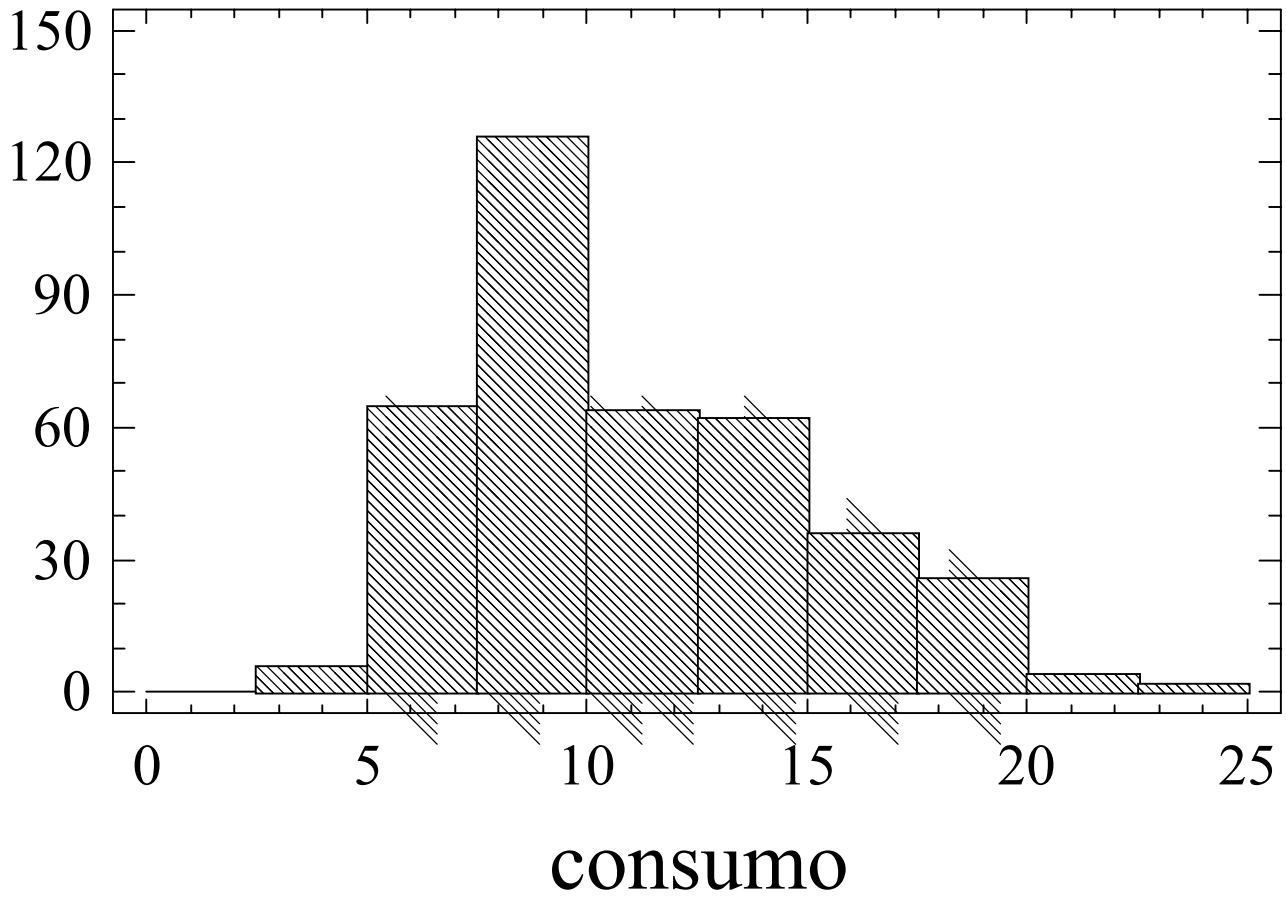
Tipos de datos

- Cuantitativos
  - Continuos: *consumo, potencia,aceleración, peso*
  - Discretos: *nº de cilindros*
- Cualitativos
  - Ordinales: categoría
  - No ordinales: país, gasolina/gasoil

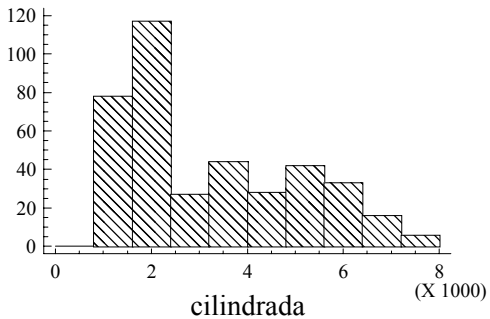
Distribución de frecuencias: consumo *l/100 km*

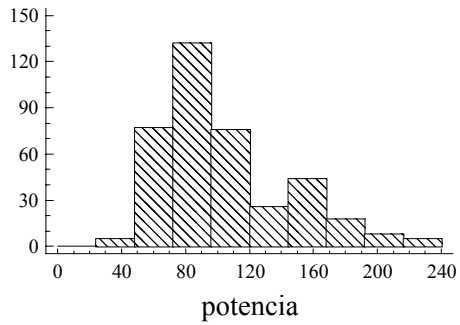
Clase	Limite Inferior	Limite Superior	Punto Medio	Frecuencia Absoluta	Frecuencia Relativa
1	0,0	2,5	1,25	0	0,0000
2	2,5	5,0	3,75	6	0,0153
3	5,0	7,5	6,25	65	0,1662
4	7,5	10,0	8,75	126	0,3223
5	10,0	12,5	11,25	64	0,1637
6	12,5	15,0	13,75	62	0,1586
7	15,0	17,5	16,25	36	0,0921
8	17,5	20,0	18,75	26	0,0665
9	20,0	22,5	21,25	4	0,0102
10	22,5	25,0	23,75	2	0,0051
Total				391	1,0000

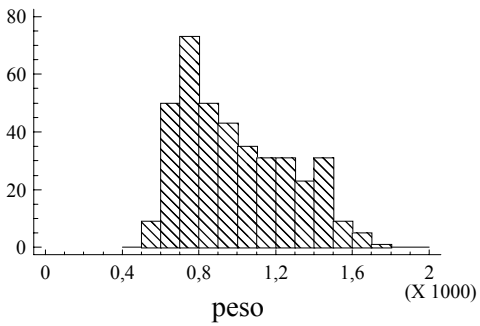
Histograma

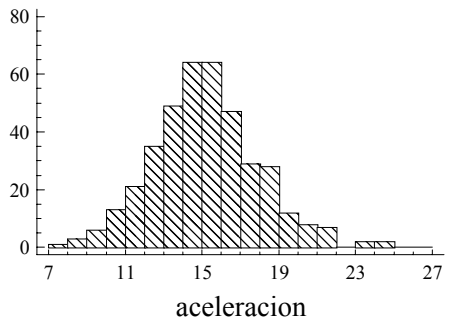


Histogramas para coches









Medidas de centro

$$x_1, x_2, \ldots, x_n$$

Media aritmética

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Media geométrica

(si $x_i > 0$ para todo i)

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \cdots x_n}$$

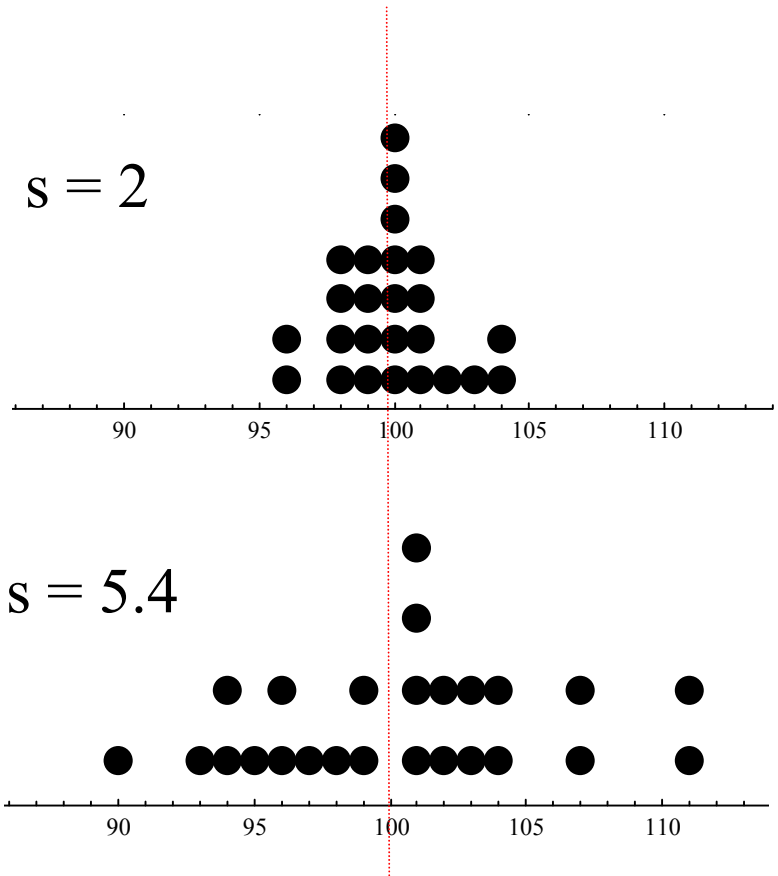
Media armónica

(si $x_i > 0$ para todo i)

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}$$

Medidas de dispersión



Media 100

$$x_1, x_2, \ldots, x_n$$

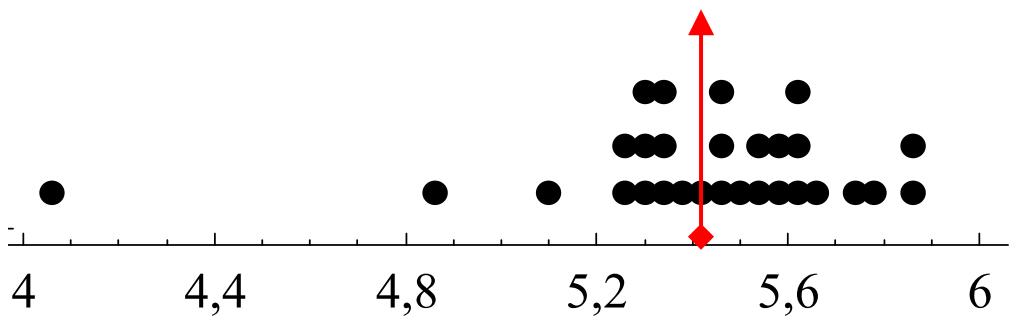
Desviación Típica

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

Varianza : s^2

Densidad de la tierra (Cavendish, 1798)

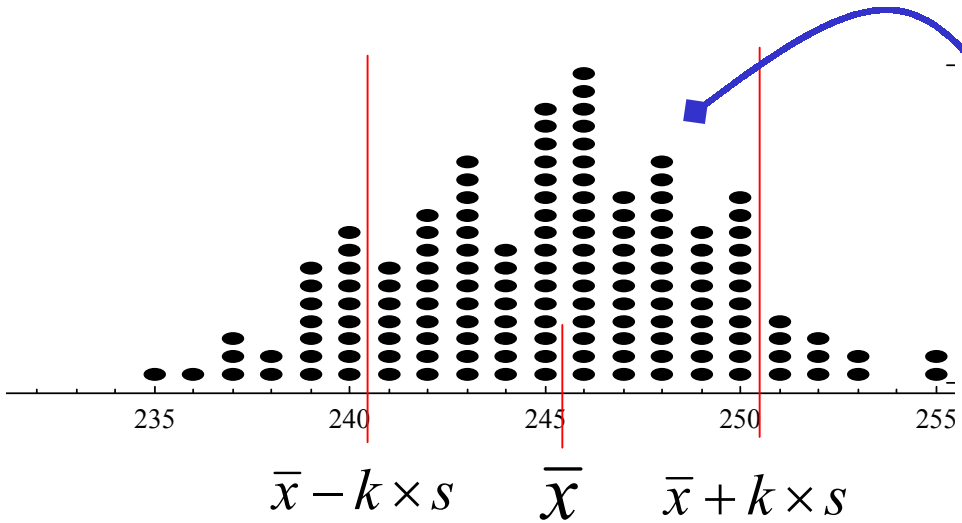
5,5	5,47	5,55	5,75	5,29	5,27
5,57	4,88	5,34	5,29	5,34	5,85
5,42	5,62	5,3	5,1	5,26	5,65
5,61	5,63	5,36	5,86	5,44	5,39
5,53	4,07	5,79	5,58	5,46	



densidad

Media = 5.42 *Desv. Típ. = 0.338*

Desigualdad de Chebychev



$$fr(|x_i-\bar{x}|\le ks)>1-\frac{1}{k^2}$$

$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}=\frac{\sum_{|x_i-\bar{x}|\le ks}(x_i-\bar{x})^2}{n}+\frac{\sum_{|x_i-\bar{x}|>ks}(x_i-\bar{x})^2}{n}$$

$$s^2\ge\frac{\sum_{|x_i-\bar{x}|>ks}(x_i-\bar{x})^2}{n}>\frac{\sum_{|x_i-\bar{x}|>ks}k^2s^2}{n}=fr(|x_i-\bar{x}|>ks)k^2s^2$$

$$fr(|x_i-\bar{x}|>ks)<\frac{1}{k^2}\quad\Leftrightarrow\quad fr(|x_i-\bar{x}|\le ks)>1-\frac{1}{k^2}$$

Mediana y Cuartiles

$x_1, x_2, \ldots, x_n$

Datos ordenados

$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$

$$\text{Mediana} \begin{cases} x_{(p)} & p = \dfrac{n+1}{2} : n \quad \text{impar} \\ \dfrac{x_{(p)} + x_{(p+1)}}{2} & p = \dfrac{n}{2} : \quad n \text{ par} \end{cases}$$

Cuartiles

$$Q_1 = x_{(r)} \qquad Q_3 = x_{(s)}$$

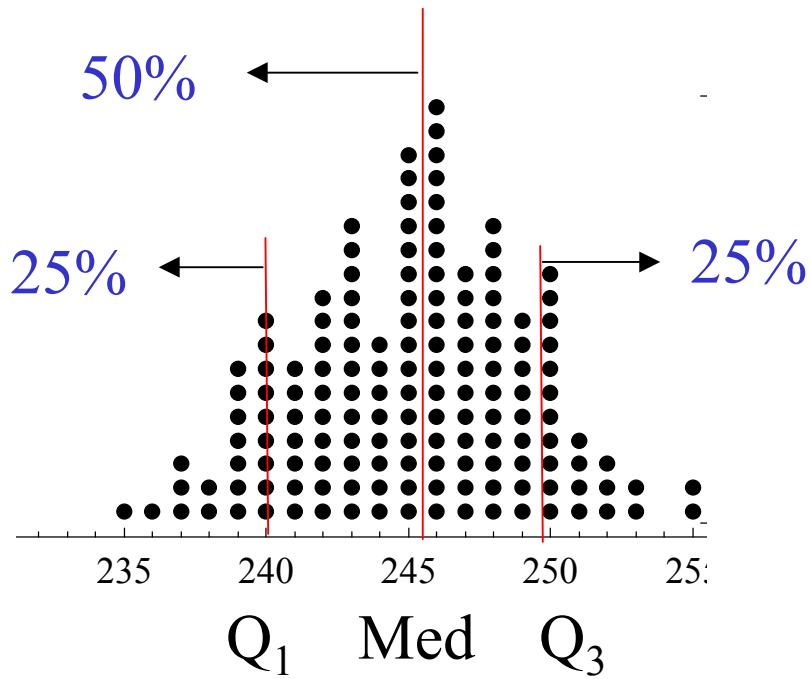
$$r = \left\lfloor \frac{p+1}{2} \right\rfloor \qquad s = n - r$$

Mediana y Cuartiles

$x_1, x_2, \ldots, x_n$

Mediana : (Med)

$fr(x_i \leq Med) = 0.50$



Cuartiles

Q_1

$fr(x_i \leq Q_1) = 0.25$

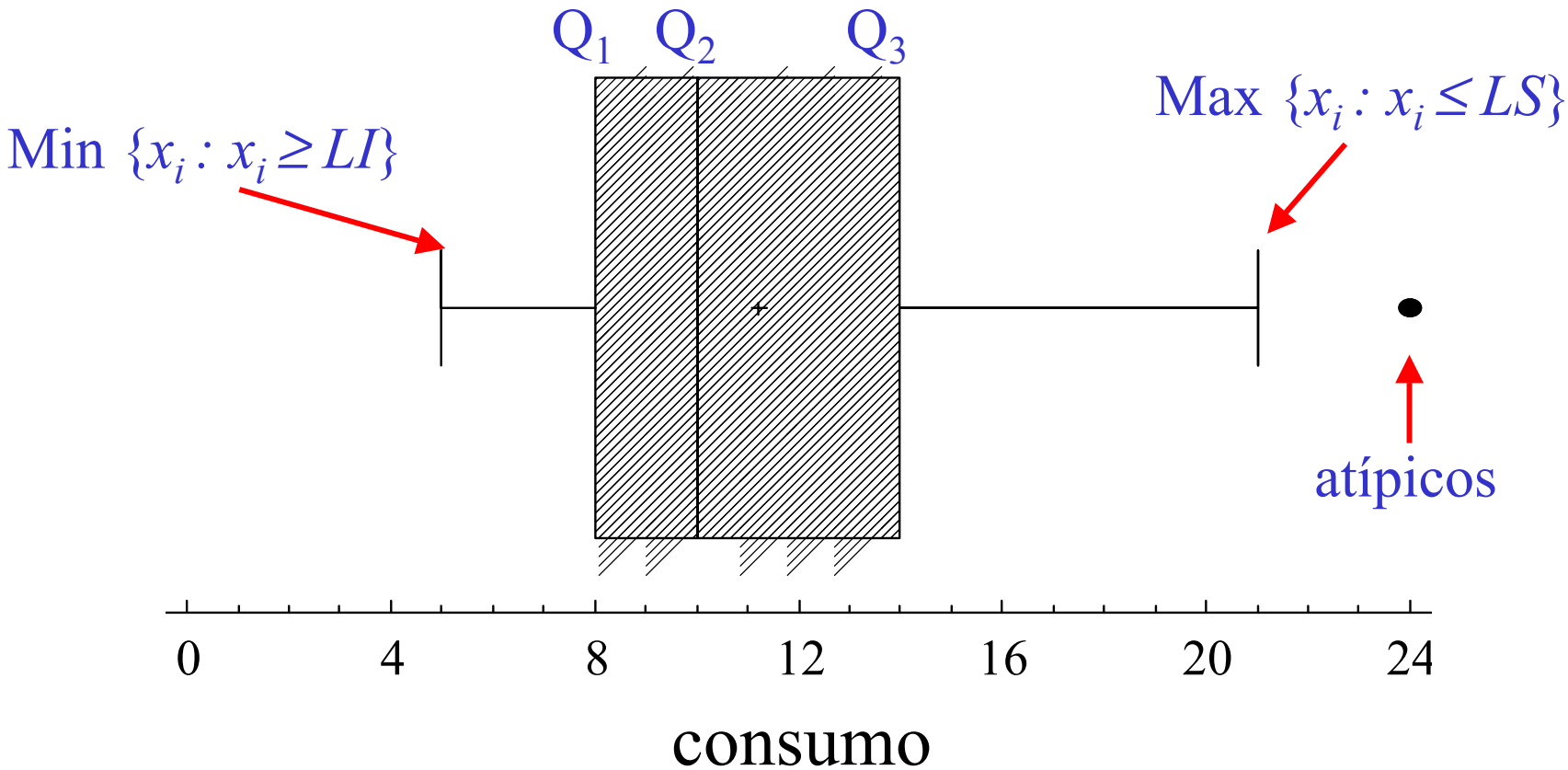
Q_3

$fr(x_i \leq Q_3) = 0.75$

Medidas características

	Consumo	Cilindrada	Potencia	Peso	Aceleración
Media	11.2	3181.2	104.2	990.7	15.7
Desv. Típica	3.9	1714.6	38.3	281.9	2.8
Primer Cuartil	8	1721	75	741.5	14
Mediana	10	2474	93	933	16
Tercer Cuartil	13.5	4334	125	1203.5	17
Rango Intercuartílico	5.5	2613	50	462	3

Diagrama de caja



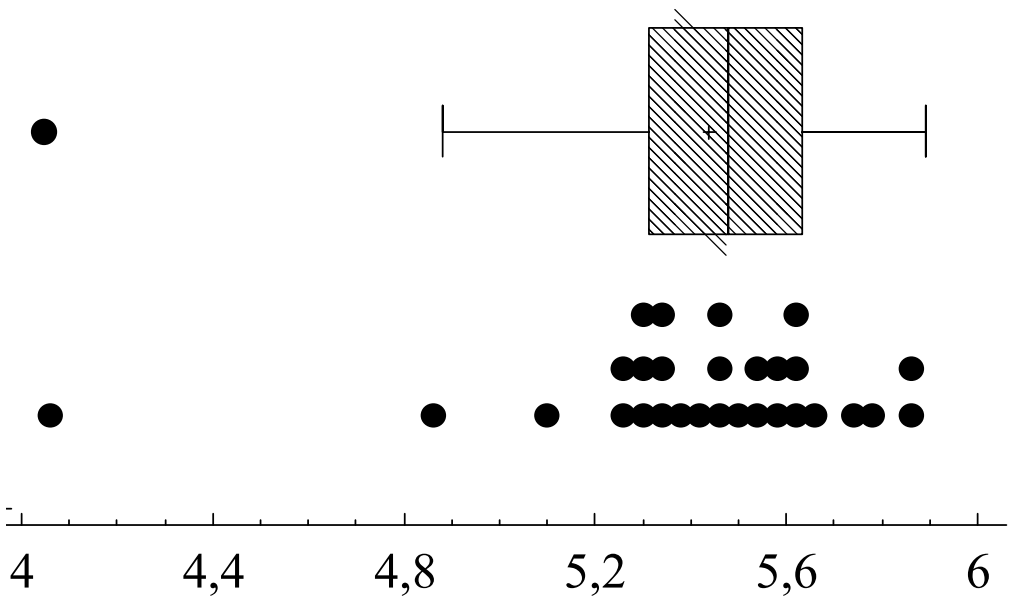
$LI = Q_1 - 1.5\ RI$

$LS = Q_3 + 1.5\ RI$

$RI = Q_3 - Q_1$

Densidad de la tierra (Cavendish, 1798)

5,5	5,47	5,55	5,75	5,29	5,27
5,57	4,88	5,34	5,29	5,34	5,85
5,42	5,62	5,3	5,1	5,26	5,65
5,61	5,63	5,36	5,86	5,44	5,39
5,53	4,07	5,79	5,58	5,46	



Media = 5.42 *Desv. Típ. = 0.338*

Diagrama de caja múltiple

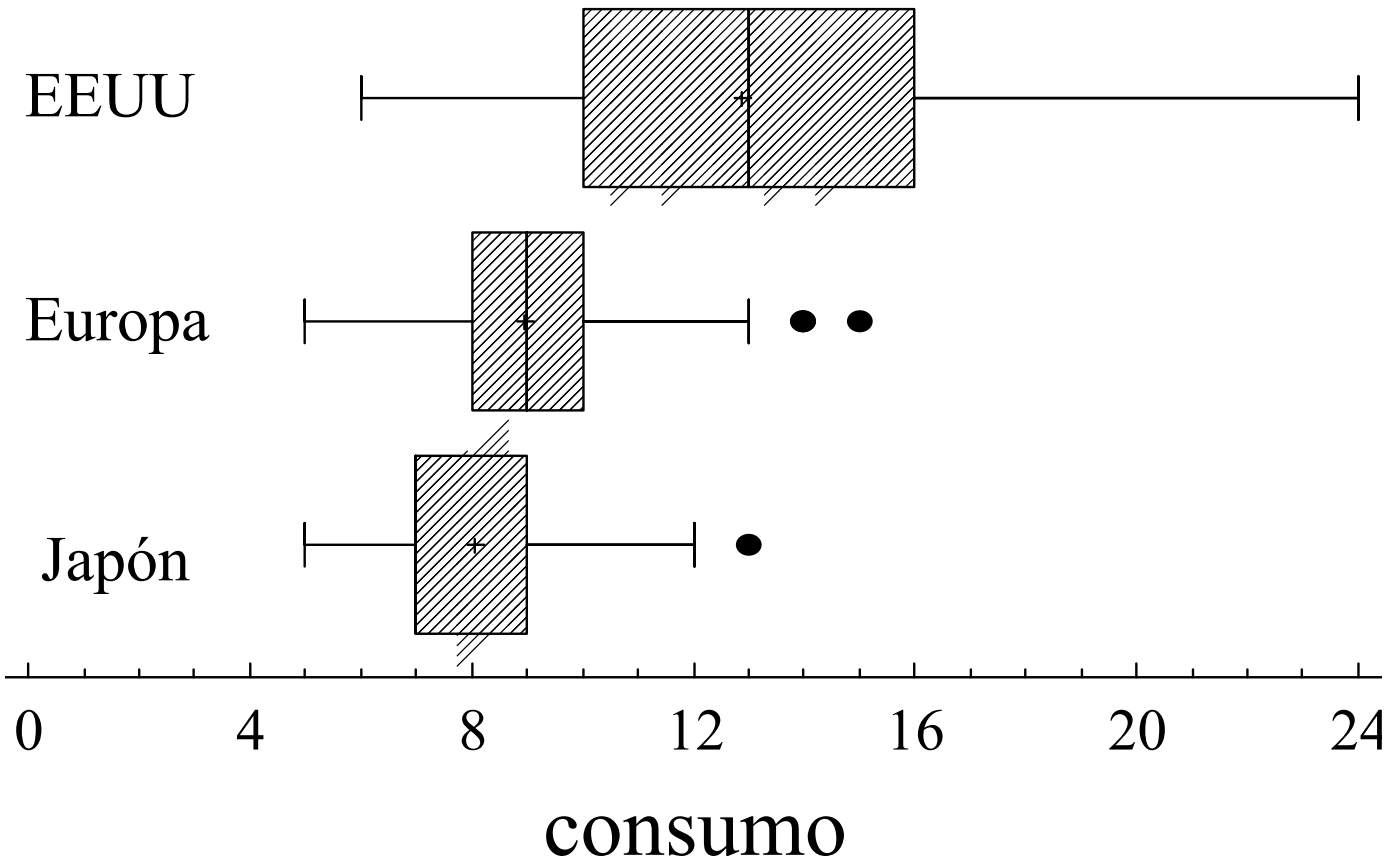
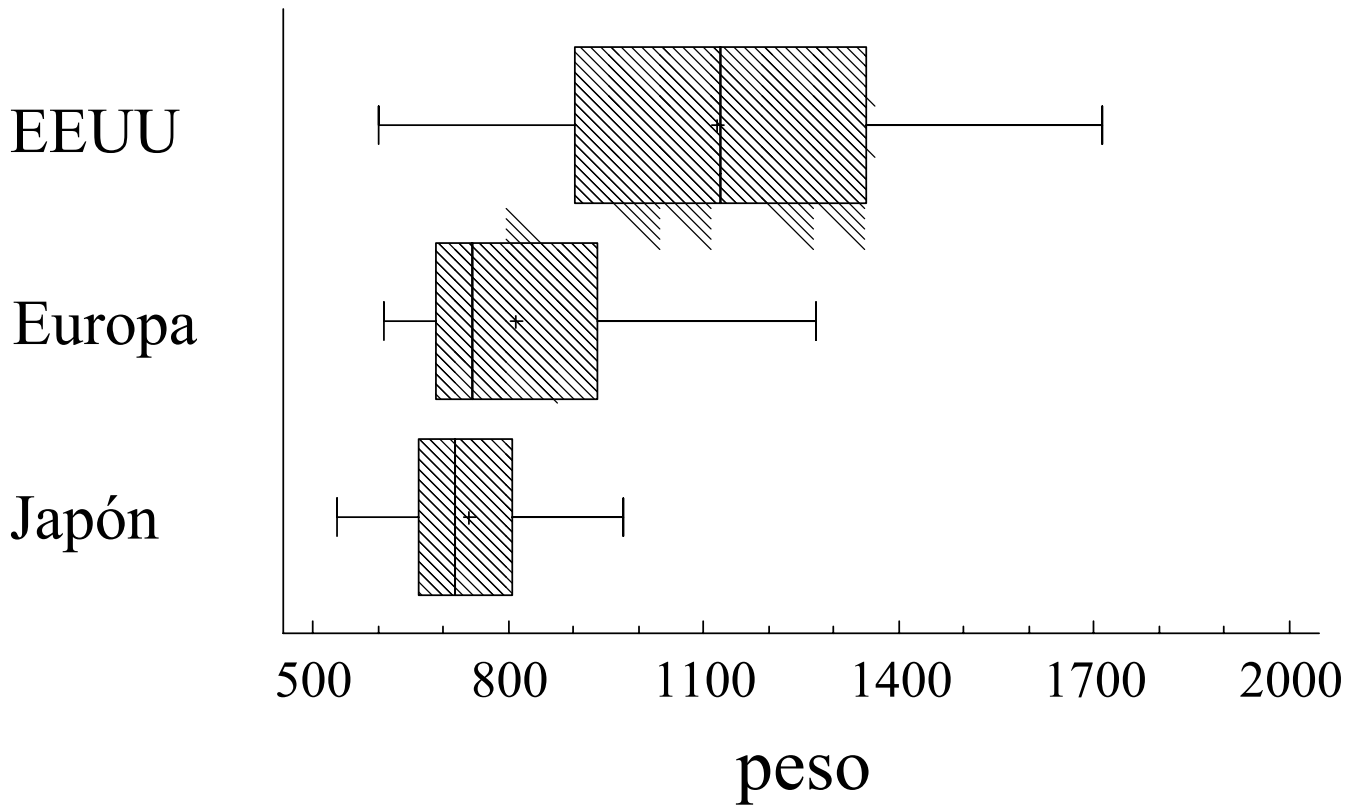


Diagrama de caja múltiple



Consumo según año de fabricación

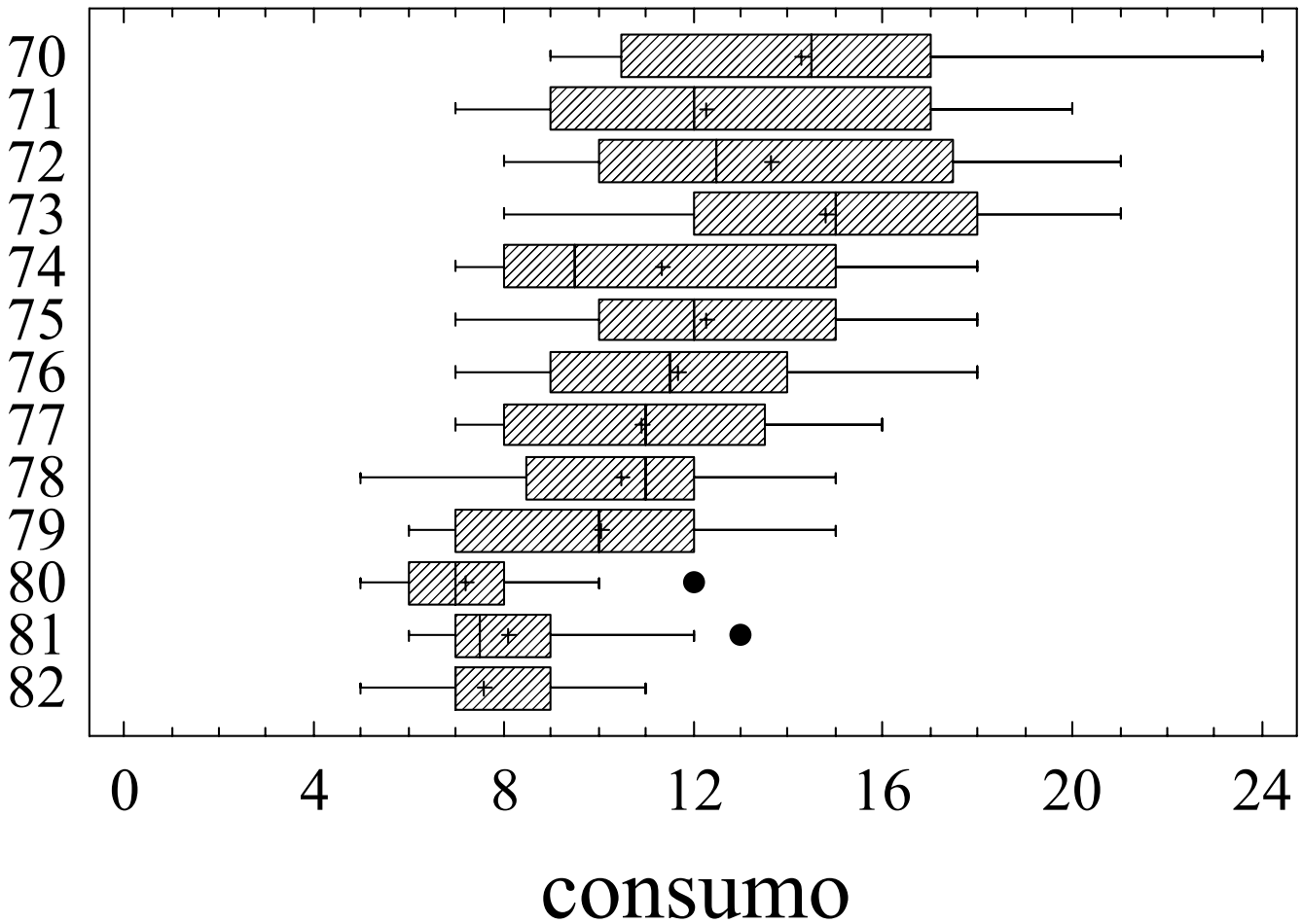


Diagrama de Caja Múltiple

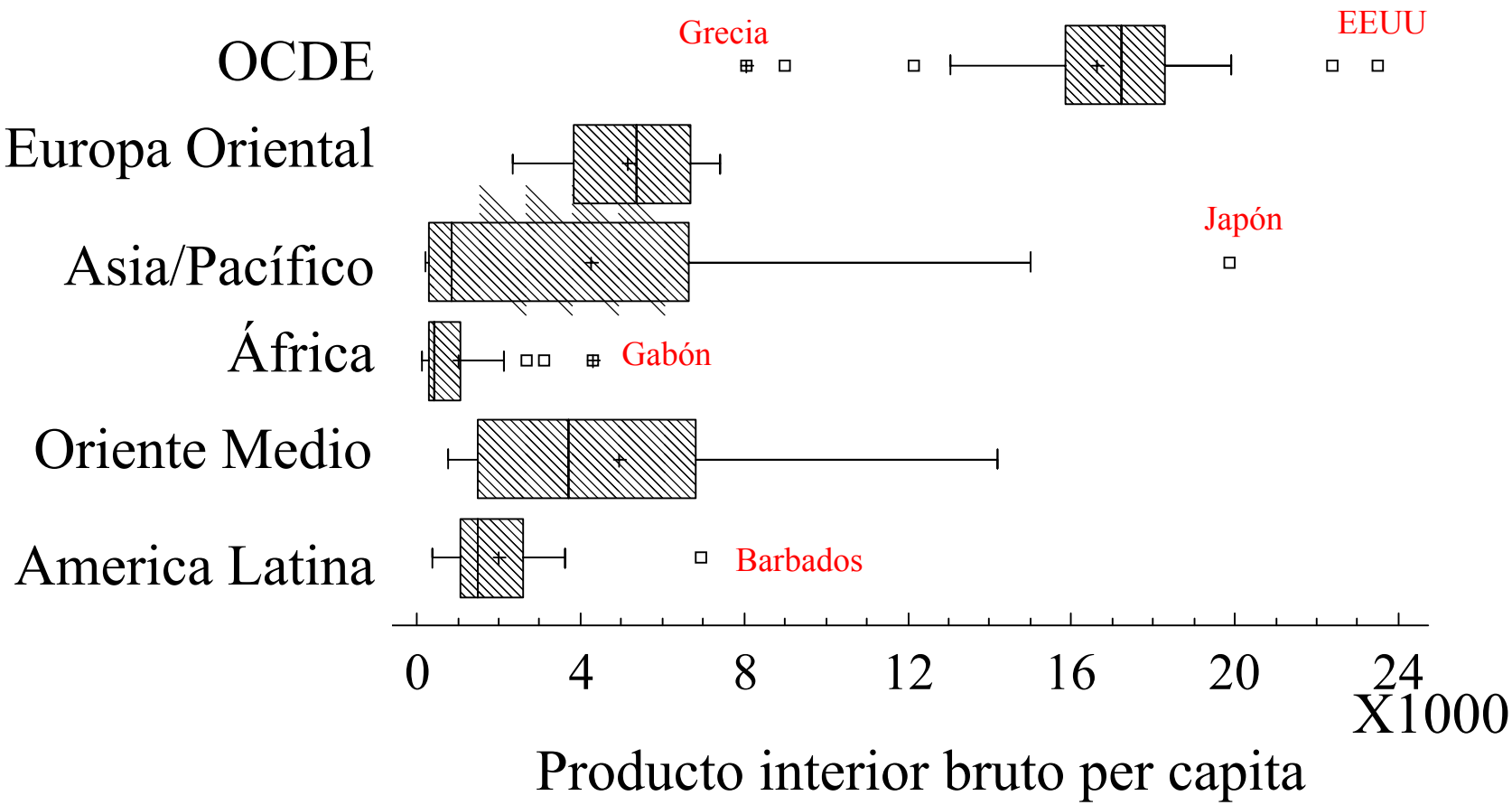


Diagrama de tallos y hojas

```
      LO|4,07

 1     4|
 1     4|
 1     4|
 1     4|
 2     4|8
 3     5|1
12     5|222233333
(9)    5|444455555
 8     5|666677
 2     5|88
```

- Media 5,419
- Des. Típica 0,339
- Mínimo 4,07
- Máximo 5,86
- Cuartil 1 5.3
- Mediana 5.46
- Cuartil 3 5.61

Medidas características de forma (asimetría y curtosis)

Coeficiente de asimetría

$$C_{AS} = \frac{m_3}{s^3} = \frac{\sum (x_i - \bar{x})^3}{n s^3}$$

Coeficiente de curtosis o apuntamiento

$$C_{AP} = \frac{m_4}{s^4} = \frac{\sum (x_i - \bar{x})^4}{n s^4}$$

Momento respecto al origen

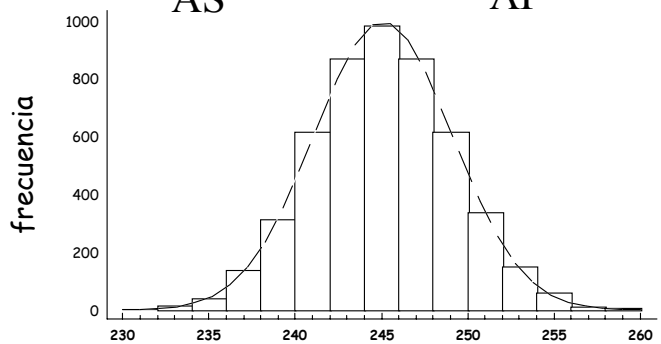
$$a_k = \frac{\sum_{i=1}^{n} x_i^k}{n}$$

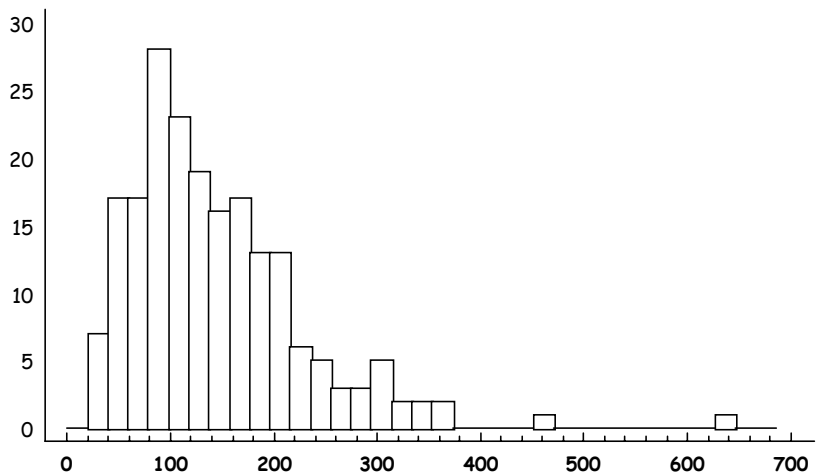
Momentos respecto a la media

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

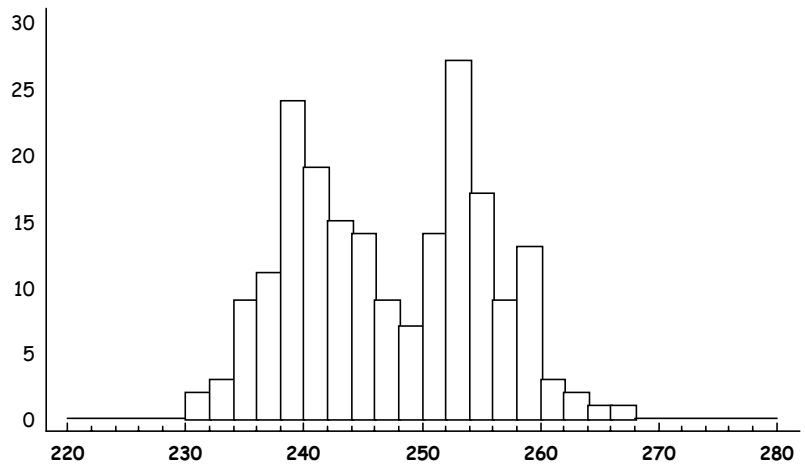
Modelo ideal

$C_{AS} = 0$ $C_{AP} = 3$





$C_{AS} > 0$



$C_{AP} < 3$

Transformaciones de datos

- Transformaciones Lineales

$$y_i = a + bx_i$$

$$\bar{y} = a + b\bar{x}$$

$$s_y = |b| s_x$$

La "*forma*" de la distribución no cambia

(Asimetría y curtosis no cambia)

- Transformaciones no-lineales

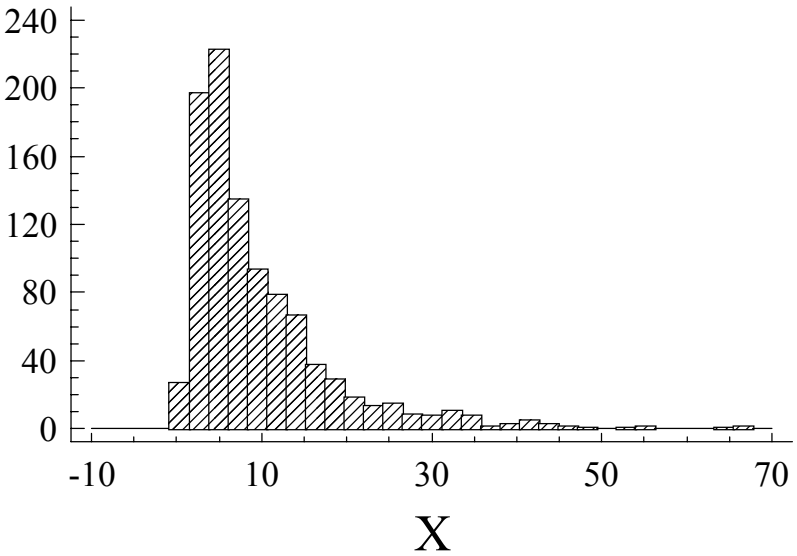
$$y_i = h(x_i)$$

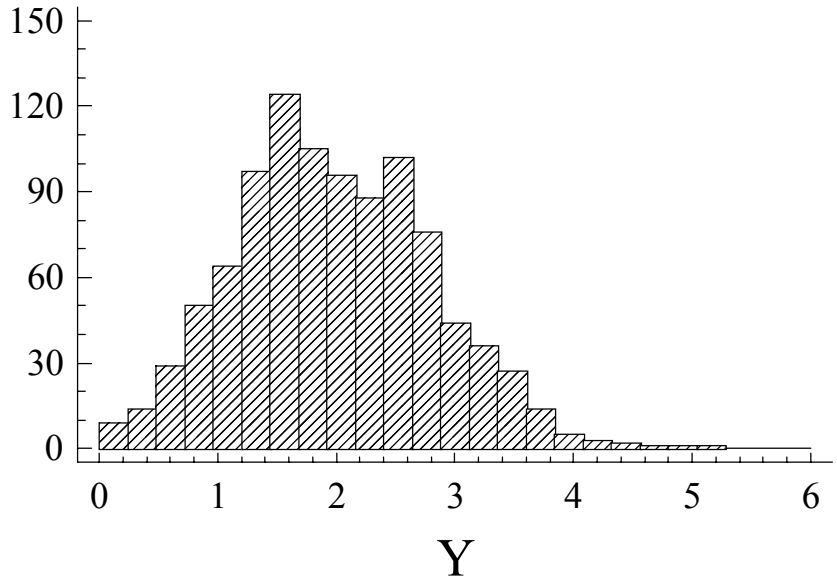
$$\bar{y} \neq h(\bar{x})$$

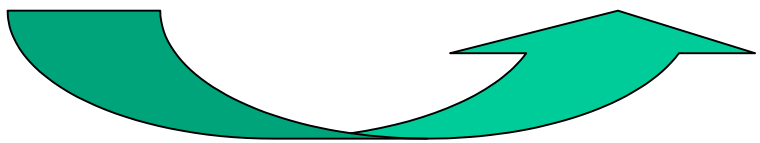
Cambia la "*forma*" de la distribución

(coeficientes de asimetría y curtosis cambian)

Efecto de la transformación de datos

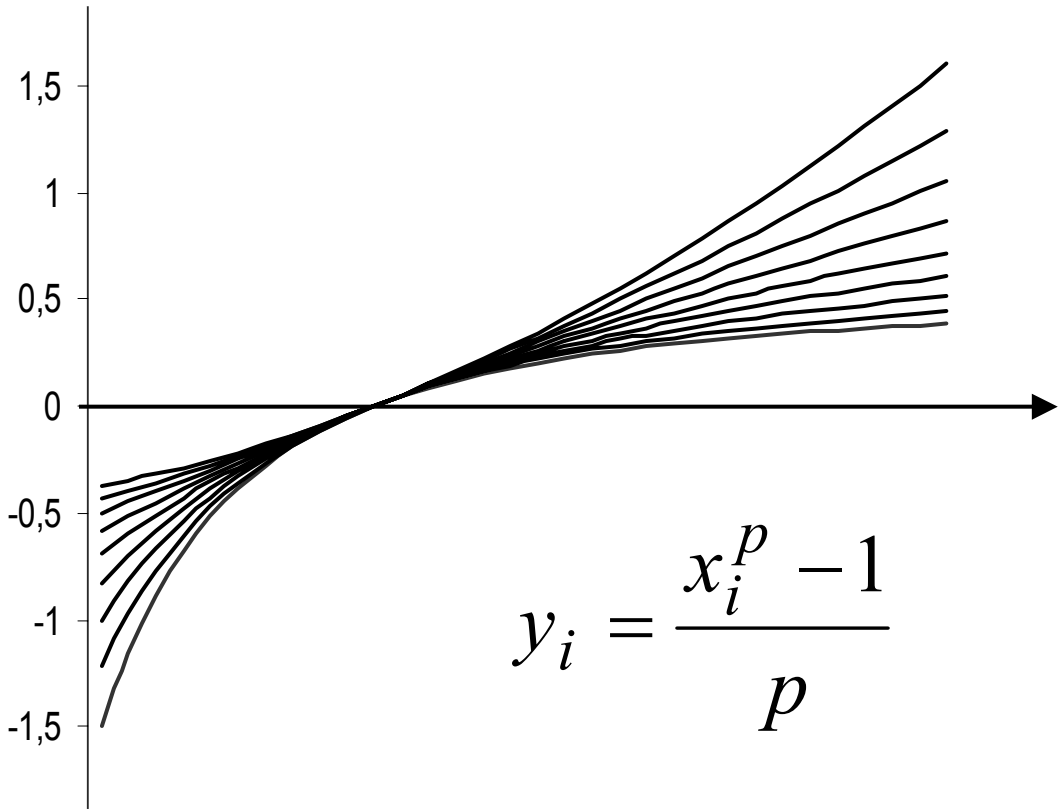






$$y_i = \log x_i$$

Transformaciones Box-Cox



$$y_i = \frac{x_i^p - 1}{p}$$

$$p = 0 \Rightarrow y_i = \log x_i$$

Datos

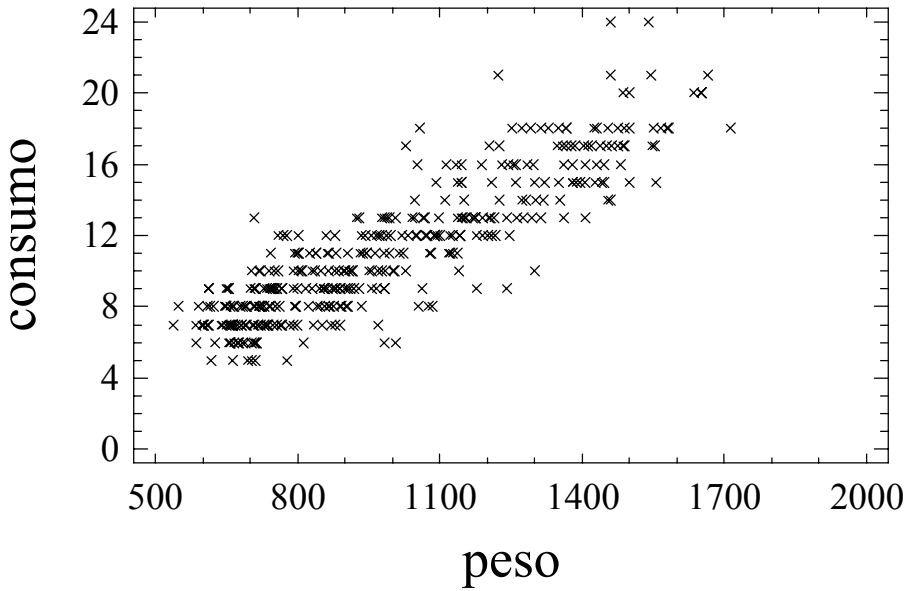
Variables

Observaciones	Y_1	Y_2	$\cdots$	Y_k		
1	x_{11}	x_{21}	$\cdots$	x_{k1}	$\longrightarrow$	$\mathbf{x_1}$
2	x_{12}	x_{22}	$\cdots$	x_{k2}	$\longrightarrow$	$\mathbf{x_2}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\longrightarrow$	$\vdots$
n	x_{1n}	x_{2k}	$\cdots$	x_{kn}	$\longrightarrow$	$\mathbf{x}_n$

Vector de Medias

$$\mathbf{x_i} = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}; \qquad \overline{\mathbf{x}} = \frac{\sum_{i=1}^{n} \mathbf{x_i}}{n} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_n \end{pmatrix}$$

Covarianza



Coche	*Peso*	*Consumo*
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

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

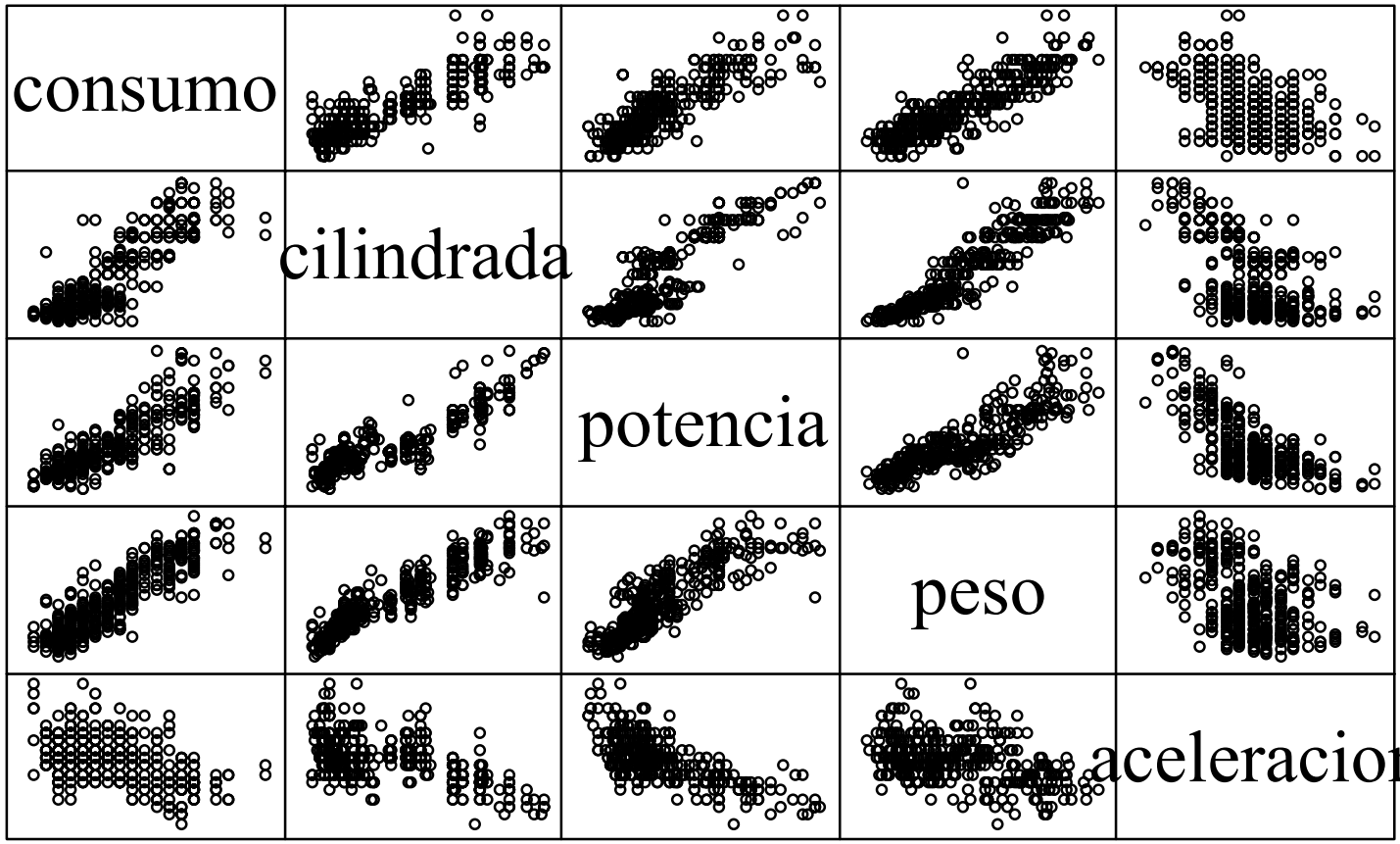
Matriz de Varianzas

$$\mathbf{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} x_{1i}-\bar{x}_1 \\ x_{2i}-\bar{x}_2 \\ \vdots \\ x_{ki}-\bar{x}_k \end{pmatrix}\begin{pmatrix} x_{1i}-\bar{x}_1 & x_{2i}-\bar{x}_2 & \cdots & x_{ki}-\bar{x}_k \end{pmatrix}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} (x_{1i}-\bar{x}_1)^2 & (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2) & \cdots & (x_{1i}-\bar{x}_1)(x_{ki}-\bar{x}_k) \\ (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2) & (x_{2i}-\bar{x}_2)^2 & \cdots & (x_{2i}-\bar{x}_2)(x_{ki}-\bar{x}_k) \\ \vdots & \vdots & \ddots & \vdots \\ (x_{1i}-\bar{x}_1)(x_{ki}-\bar{x}_k) & (x_{2i}-\bar{x}_2)(x_{ki}-\bar{x}_k) & \cdots & (x_{ki}-\bar{x}_k)^2 \end{pmatrix}$$

$$= \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{12} & s_2^2 & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1k} & s_{2k} & \cdots & s_k^2 \end{pmatrix}$$

Gráficos de dispersión: ejemplo coches



Matriz de varianzas: ejemplo coches

$$\mathbf{S^2} = \begin{array}{c} \begin{array}{ccccc} \text{consumo} & \text{c.c.} & \text{pot.} & \text{peso} & \text{acel.} \end{array} \\ \begin{pmatrix} 15{,}2 & 5.824{,}4 & 127{,}3 & 971{,}5 & -5{,}0 \\ 5.824{,}4 & 2{,}94E6 & 58.965{,}4 & 451.461{,}0 & -2.597{,}4 \\ 127{,}3 & 58.965{,}4 & 1.465{,}2 & 9.312{,}8 & -73{,}5 \\ 971{,}5 & 451.461{,}0 & 9.312{,}8 & 7.949{,}5 & -328{,}0 \\ -5{,}0 & -2.597{,}4 & -73{,}5 & -328{,}0 & 7{,}6 \end{pmatrix} \end{array}$$

Propiedades de S^2

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & \cdots & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{k2} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & \cdots & x_{kn} - \bar{x}_k \end{pmatrix}$$

$$\mathbf{S^2} = \frac{1}{n}\tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}}$$

- Cuadrada k x k
- Simétrica
- Semidef. positiva

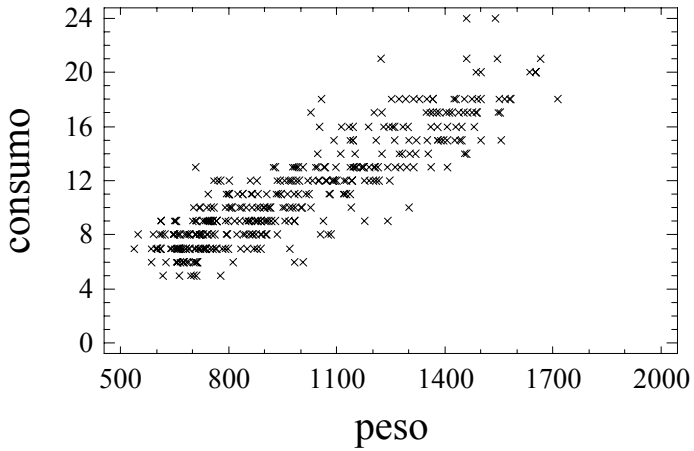
$\mathbf{S^2}$ es *semidefinida positiva* :

$$\forall \mathbf{w} \in \Re^k, \quad \mathbf{w}^{\mathbf{T}}\mathbf{S^2}\mathbf{w} \geq 0$$

$$\mathbf{w}^{\mathbf{T}}\mathbf{S^2}\mathbf{w} = \mathbf{w}^{\mathbf{T}}\left(\frac{1}{n}\tilde{\mathbf{X}}^{\mathbf{T}}\tilde{\mathbf{X}}\right)\mathbf{w} = \frac{1}{n}(\tilde{\mathbf{X}}\mathbf{w})^{\mathbf{T}}(\tilde{\mathbf{X}}\mathbf{w})$$

$$\mathbf{v} = \tilde{\mathbf{X}}\mathbf{w}, \quad \Rightarrow \mathbf{w}^{\mathbf{T}}\mathbf{S^2}\mathbf{w} = \frac{1}{n}\mathbf{v}^{\mathbf{T}}\mathbf{v} = \frac{\sum_{i=1}^{n} v_i^2}{n} \geq 0$$

Correlación



Obs.	*Var* – 1	*Var* – 2
1	x_1	y_1
2	x_2	y_2
⋮	⋮	⋮
n	x_n	y_n

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

- Adimensional
- $-1 \le r_{xy} \le +1$
- $|r_{xy}| = 1 \Leftrightarrow y_i = a + b\,x_i$

Matriz de correlaciones ejemplo coches

consumo	c.c.	pot.	peso	Acel.

$$\mathbf{R} = \begin{pmatrix} 1 & 0{,}873 & 0{,}854 & 0{,}885 & -0{,}466 \\ 0{,}873 & 1 & 0{,}898 & 0{,}934 & -0{,}549 \\ 0{,}854 & 0{,}898 & 1 & 0{,}863 & -0{,}696 \\ 0{,}885 & 0{,}934 & 0{,}863 & 1 & -0{,}422 \\ -0{,}466 & -0{,}549 & -0{,}696 & -0{,}422 & 1 \end{pmatrix}$$

Las variables están muy correlacionadas

Transformaciones Lineales

$$y_i = a_1 x_{1i} + a_2 x_{2i} + \cdots + a_k x_{ki} = (a_1 \quad a_2 \quad \cdots \quad a_k) \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix} = \mathbf{a}^{\mathbf{T}} \mathbf{x_i}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} \mathbf{a}^{\mathbf{T}} \mathbf{x_i}}{n} = \frac{\mathbf{a}^{\mathbf{T}} (\sum_{i=1}^{n} \mathbf{x_i})}{n} = \underline{\mathbf{a}^{\mathbf{T}} \bar{\mathbf{x}}}$$

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^{n} (\mathbf{a}^{\mathbf{T}} \mathbf{x_i} - \mathbf{a}^{\mathbf{T}} \bar{\mathbf{x}})(\mathbf{x_i}^{\mathbf{T}} \mathbf{a} - \bar{\mathbf{x}}^{\mathbf{T}} \mathbf{a})}{n}$$

$$= \mathbf{a}^{\mathbf{T}} \left(\frac{\sum_{i=1}^{n} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^{\mathbf{T}}}{n} \right) \mathbf{a} = \underline{\mathbf{a}^{\mathbf{T}} \mathbf{S}^{\mathbf{2}} \mathbf{a}}$$

Transformaciones lineales II

$$\left.\begin{aligned} y_{1i} &= a_{11}x_{1i} + a_{12}x_{2i} + \cdots + a_{1k}x_{ki} \\ y_{2i} &= a_{21}x_{1i} + a_{22}x_{2i} + \cdots + a_{mk}x_{ki} \\ &\vdots \\ y_{mi} &= a_{m1}x_{1i} + a_{m2}x_{2i} + \cdots + a_{mk}x_{ki} \end{aligned}\right\} \quad \begin{pmatrix} y_{1i} \\ y_{2i} \\ \vdots \\ y_{mi} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{pmatrix} \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ki} \end{pmatrix}$$

$$\mathbf{y_i} = \mathbf{A}\mathbf{x_i}$$

$$\bar{\mathbf{y}} = \frac{\sum_{i=1}^{n} \mathbf{y_i}}{n} = \frac{\sum_{i=1}^{n} \mathbf{A}\mathbf{x_i}}{n} = \frac{\mathbf{A}(\sum_{i=1}^{n} \mathbf{x_i})}{n} = \mathbf{A}\bar{\mathbf{x}}$$

$$\mathbf{S_Y^2} = \frac{\sum_{i=1}^{n} (\mathbf{y_i} - \bar{\mathbf{y}})(\mathbf{y_i} - \bar{\mathbf{y}})^T}{n}$$

$$= \frac{\sum_{i=1}^{n} (\mathbf{A}\mathbf{x_i} - \mathbf{A}\bar{\mathbf{x}})(\mathbf{x_i^T}\mathbf{A^T} - \bar{\mathbf{x}}^{\mathbf{T}}\mathbf{A^T})}{n} = \mathbf{A}\underbrace{\left(\frac{\sum_{i=1}^{n} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T}{n}\right)}_{\mathbf{S_X^2}}\mathbf{A^T}$$

$$= \mathbf{A}\mathbf{S_X^2}\mathbf{A^T}$$

Efecto de las transformaciones

(no lineales)