Tema 1
Análisis exploratorio de datos

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

Información de contacto

José Ramón Berrendero Díaz

Correo electrónico: joser.berrendero@uam.es

Teléfono: 91 497 66 90

Despacho: Módulo 08 - Despacho 210

Página web:
http://matematicas.uam.es/~joser.berrendero


¿PROFESIÓN?
ESTADÍSTICO


LOS PRÓXIMOS DILUVIOS SERÁN DE DATOS...

Ejemplo: contaminación por mercurio en el pescado

- El agua de los ríos contiene pequeñas concentraciones de mercurio que se pueden ir acumulando en los tejidos de los peces.

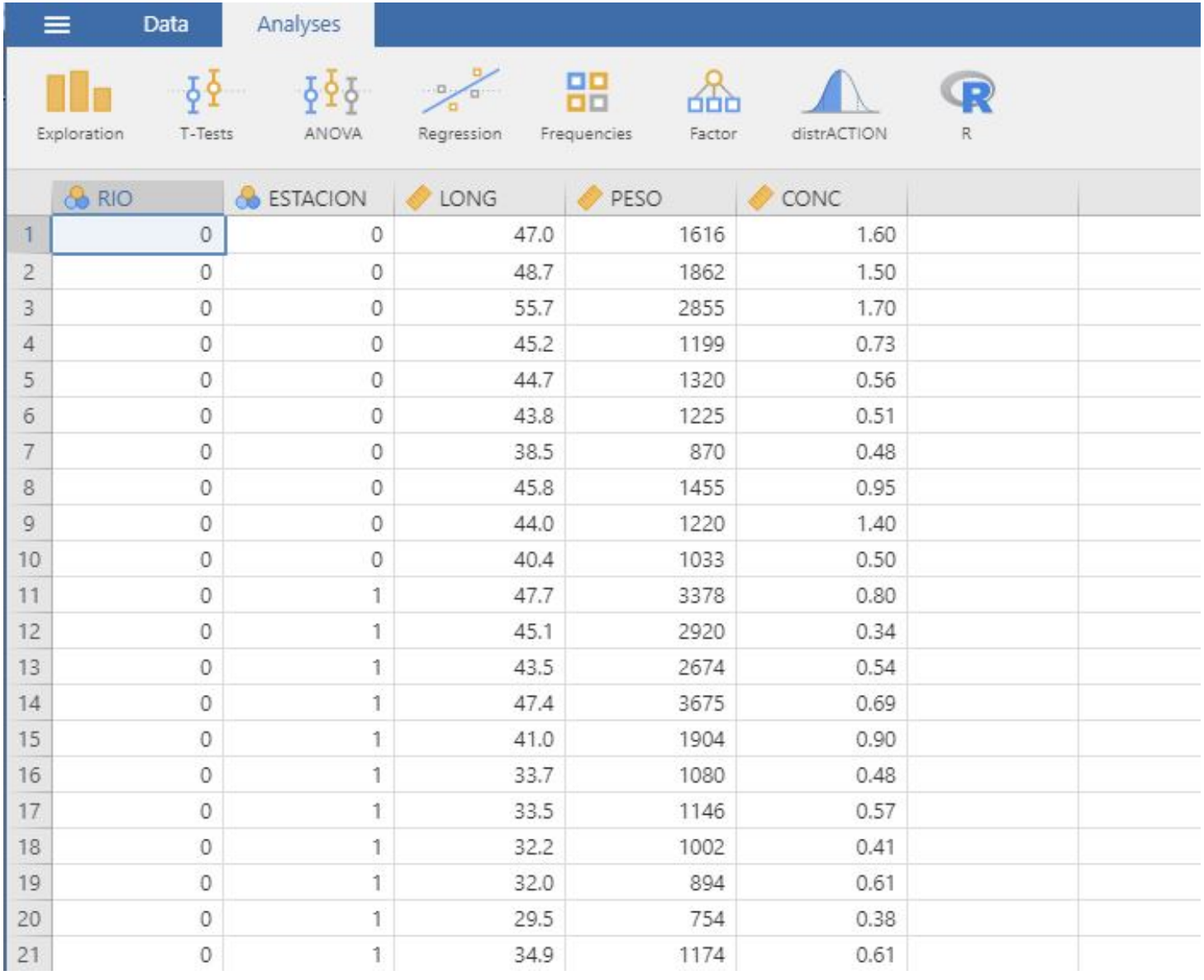
- Se ha realizado un estudio en los ríos Wacamaw y Lumber en Carolina del Norte (EE.UU.), analizando la cantidad de mercurio que contenían 171 ejemplares capturados de una cierta especie de peces.

- Los datos obtenidos se encuentran en el fichero `mercurio.omv` (la extensión `.omv` corresponde al formato de *jamovi*).

Variables

Nombre variable	Descripción
`RIO`	Código del río (0=Lumber, 1=Wacamaw)
`ESTACION`	Código de la estación (de 0 a 16)
`LONG`	Longitud (en cm) del pez
`PESO`	Peso (en g) del pez
`CONC`	Concentración (en ppm) de mercurio

Datos



	RIO	ESTACION	LONG	PESO	CONC
1	0	0	47.0	1616	1.60
2	0	0	48.7	1862	1.50
3	0	0	55.7	2855	1.70
4	0	0	45.2	1199	0.73
5	0	0	44.7	1320	0.56
6	0	0	43.8	1225	0.51
7	0	0	38.5	870	0.48
8	0	0	45.8	1455	0.95
9	0	0	44.0	1220	1.40
10	0	0	40.4	1033	0.50
11	0	1	47.7	3378	0.80
12	0	1	45.1	2920	0.34
13	0	1	43.5	2674	0.54
14	0	1	47.4	3675	0.69
15	0	1	41.0	1904	0.90
16	0	1	33.7	1080	0.48
17	0	1	33.5	1146	0.57
18	0	1	32.2	1002	0.41
19	0	1	32.0	894	0.61
20	0	1	29.5	754	0.38
21	0	1	34.9	1174	0.61

Problemas de interés relacionados con estos datos

- Resumir la información que contienen con unas pocas cifras o gráficos.
- ¿Que valores toma cada variable? ¿Cuáles son los más frecuentes? ¿Hay grandes diferencias entre ellos?
- ¿Existe algún modelo que permita saber la proporción *de la población* de peces que tiene una concentración de mercurio superior a 3 ppm?
- ¿Es significativamente más alta la concentración de mercurio en un río que en otro?
- ¿Existe relación entre la concentración de mercurio y la longitud o el peso del pez?

Temario

1. Análisis exploratorio de datos
2. Nociones elementales de inferencia estadística.
3. Contrastes de hipótesis.
4. Regresión lineal simple.
5. Análisis de la varianza.

Bibliografía

- De La Horra, J. *Estadística Aplicada* (3ª ed). Ediciones Díaz de Santos, 2003.
- Moore, D. S. *Estadística aplicada básica*. Antoni Bosch, 1999.
- Milton, S. *Estadística para Biología y Ciencias de la Salud* (3ª ed. ampliada). McGraw-Hill, 2007.
- Navarro D.J. y Foxcroft D.R. (2019). *Learning statistics with jamovi*. Libro electrónico de libre descarga (enlace en moodle).
- Samuels, M., Witmer, J. y Schaffner, A. *Statistics for the life sciences.* (4ª ed.). Pearson, 2011.
- Townend, J. *Practical Statistics for Environmental and Biological Scientists*. Wiley, 2002.

Estructura del Tema 1

- Tipos de variables.
- Distribución de una variable.
- Representación gráfica de la distribución.
- Medidas numéricas para resumir la distribución.
- Covarianza y correlación.

Introducción

La *estadística* tiene por objetivo extraer conocimiento a partir de información (principalmente) numérica.

La estadística descriptiva tiene por objetivo identificar las principales características de un conjunto de datos mediante un número reducido de gráficos y/o números.

Los conjuntos de datos que vamos a considerar proceden de medir una o más *variables* en un conjunto de *individuos*.

Para describir un conjunto de datos se comienza con un análisis individual de cada variable y posteriormente se estudian las relaciones entre variables.

Se suele comenzar con representaciones gráficas y posteriormente se calculan resúmenes numéricos.

Tipos de variables

1. **Variables cualitativas**: Describen cualidades o atributos (ej. color del pelo).
2. **Variables cuantitativas discretas**: Toman un número pequeño de valores, normalmente enteros (ej. número de hijos).
3. **Variables cuantitativas continuas**: Toman valores en un intervalo (ej. tiempo hasta que llega un autobús).

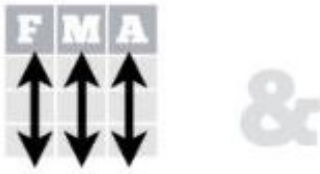
En los datos sobre contenido de mercurio, ¿de qué tipo es cada una de las variables?

En general, la técnica estadística adecuada para analizar una variable depende de su tipo.

Datos ordenados

Tidy data definition

In a tidy data set:



Each **variable** is saved in its own **column**

&

Each **observation** is saved in its own **row**

Wickham, H. (2014). Tidy Data. Journal of Statistical Software

Sample_ID	Ca	Mg	Na	Cl
P-1	234.3	12.3	4.3	33.5
P-2	432.2	22.3	2.4	12.3

Gather

Spread

Long format

Sample_ID	Key	Value
P-1	Ca	234.3
P-1	Mg	12.3
P-1	Na	4.3
P-1	Cl	33.5
P-2	Ca	432.2
P-2	Mg	22.3
P-2	Na	2.4
P-2	Cl	12.3

Distribución de una variable

La *distribución de una variable* viene determinada por los valores que toma esa variable y la frecuencia con la que los toma.

La *frecuencia absoluta* de un valor (o de un intervalo) es el número de individuos para los que la variable toma ese valor (o pertenece a ese intervalo).

La *frecuencia relativa* es igual a la frecuencia absoluta dividida por el número de datos n. Siempre es un número entre 0 y 1.

En ocasiones nos encontraremos con *datos agrupados* en intervalos o clases $A_1, \ldots, A_k$. Los valores $x_1, \ldots, x_k$ que representan cada clase (generalmente los puntos medios de los intervalos) se llaman *marcas de clase*.

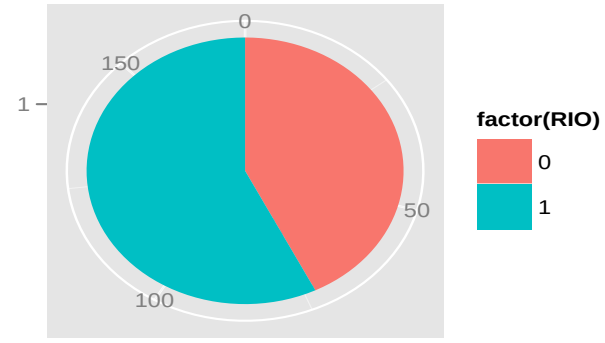
Aspectos interesantes de una distribución

- Su *posición*: en torno a qué valor central toma valores la variable.
- Su *dispersión*: el grado de concentración de los valores que toma la variable alrededor de su posición central.
- Su *forma*: por ejemplo, la simetría, es decir, si los valores se reparten de la misma forma a uno y otro lado del centro.

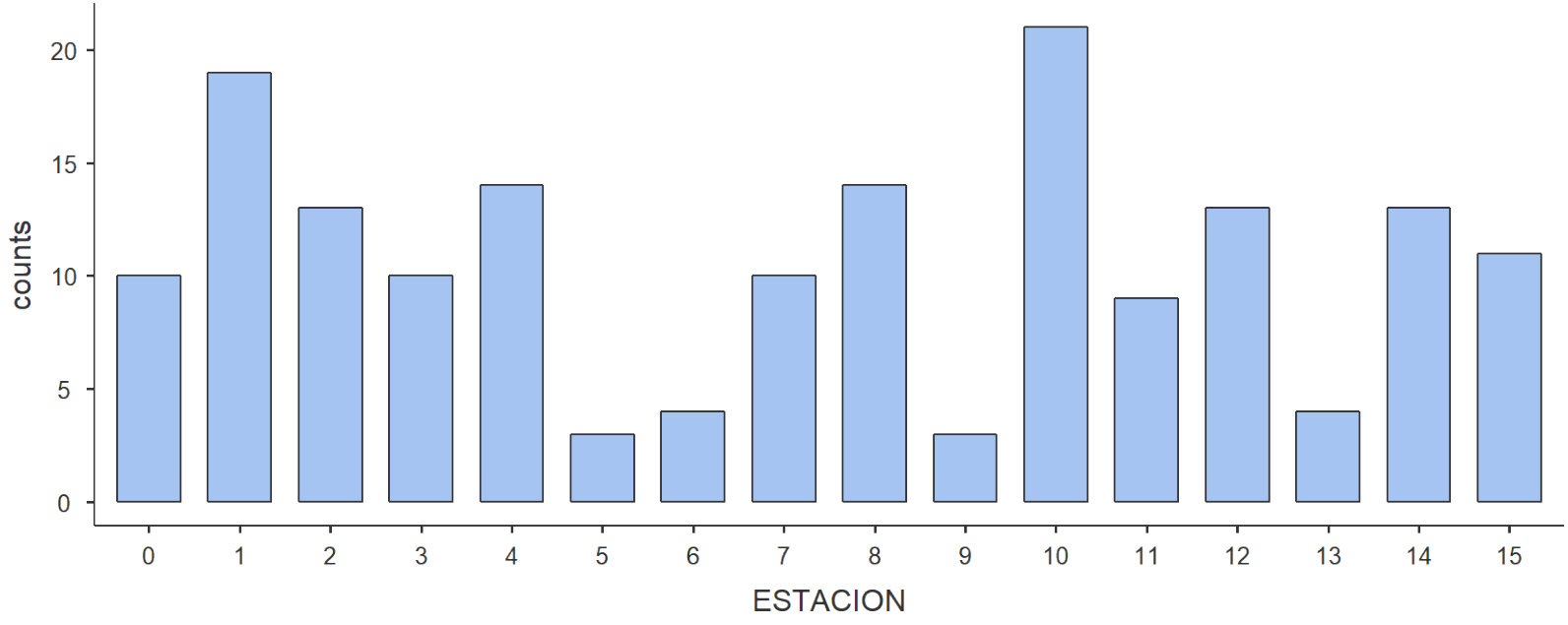
Piensa en dos conjuntos de 5 datos que tengan:

(a) La misma posición y distinta dispersión.
(b) La misma dispersión y distinta posición.

Sectores o barras (sólo variables cualitativas o discretas)

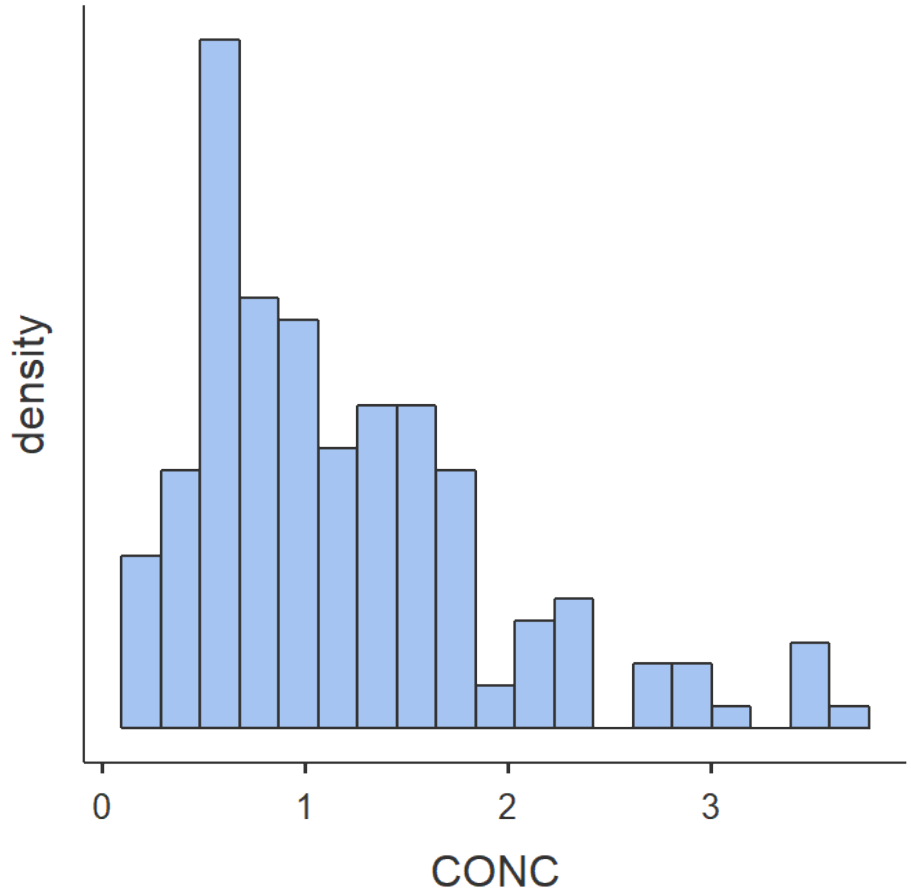


Número de observaciones en cada río



Histogramas (sólo variables continuas)

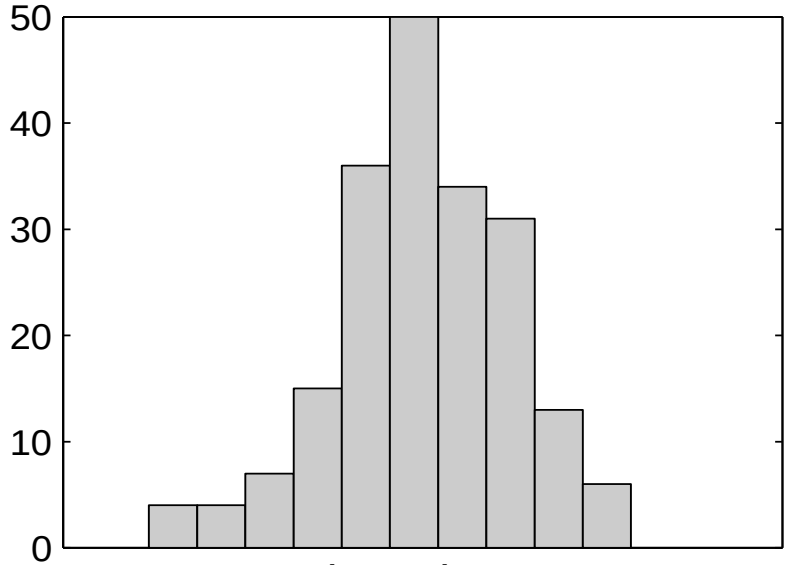
- Se divide el rango de los datos en un número adecuado de intervalos.
- Sobre cada intervalo se dibuja un rectángulo **cuya área** es proporcional a la frecuencia (relativa o absoluta) de datos en el intervalo.



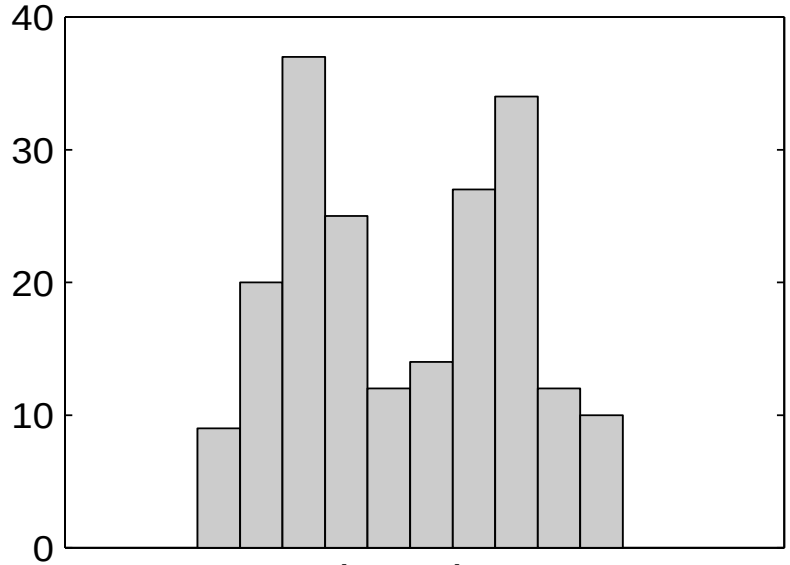
Aspectos a tener en cuenta para interpretar un histograma

- Si la base de todos los rectángulos es la misma la altura es proporcional a la frecuencia.
- ¿Cuántas modas hay?
- ¿Hay algún dato atípico en relación al resto?
- ¿Es simétrica la distribución?
- En caso de asimetría, ¿es asimétrica a la izquierda o a la derecha
- ¿En torno a qué valor aproximado están centrados los datos?
- ¿Están muy dispersos los datos en torno a este centro?

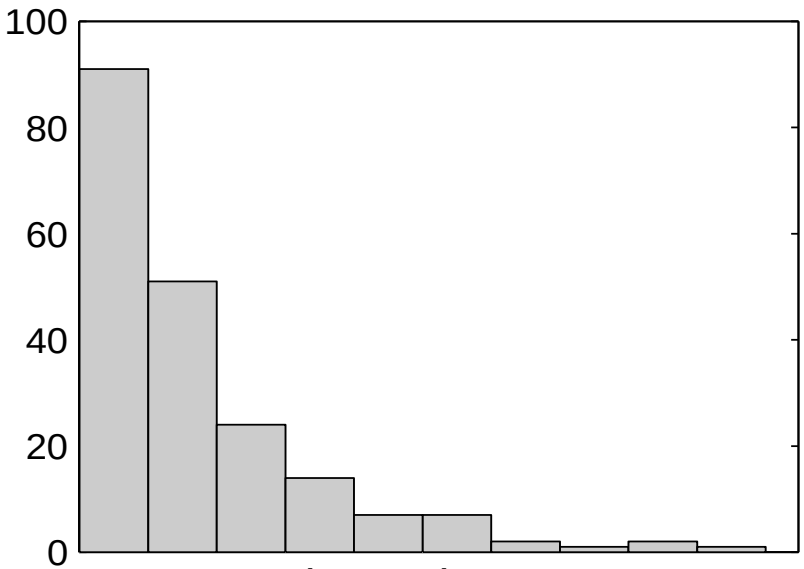
Tipos de simetría



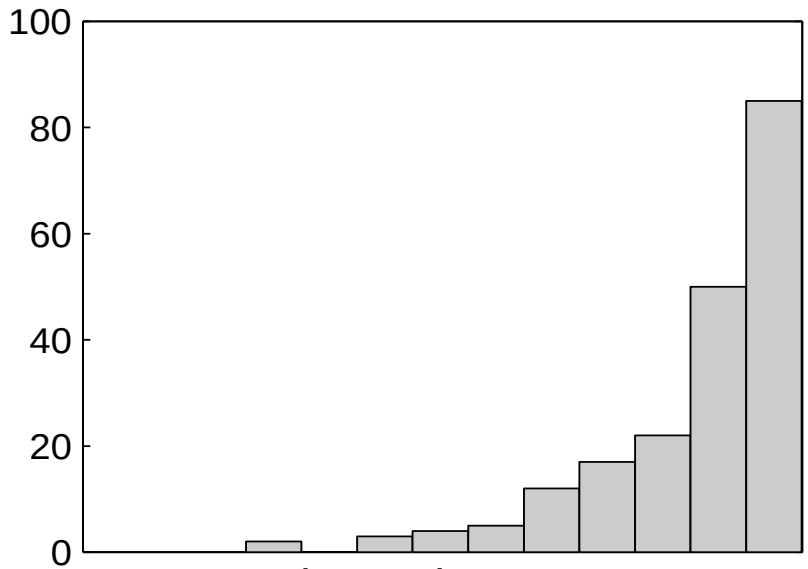
Distribución simétrica unimodal



Distribución simétrica bimodal

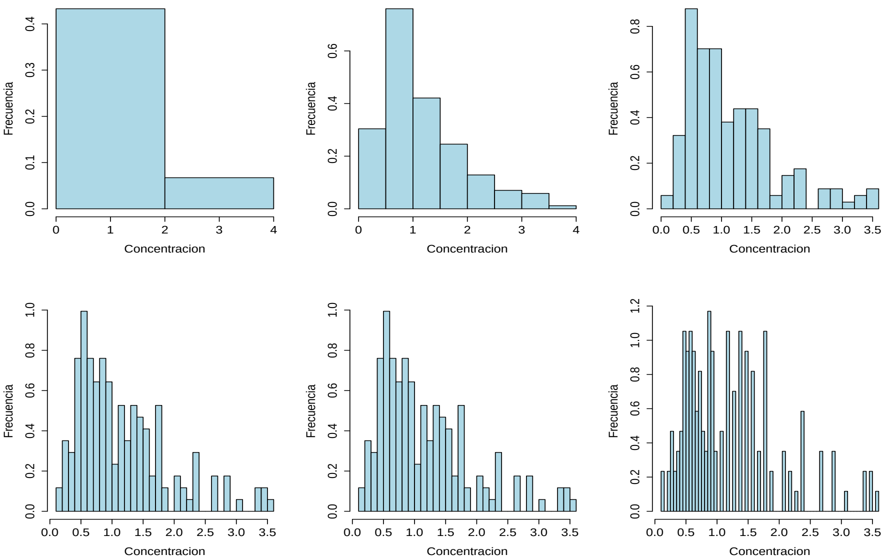


Distribución asimétrica a la derecha



Distribución asimétrica a la izquierda

La forma depende del número de intervalos



Medidas numéricas de posición: la media aritmética

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Algunas propiedades:

- La suma de las desviaciones a la media siempre es igual a cero:

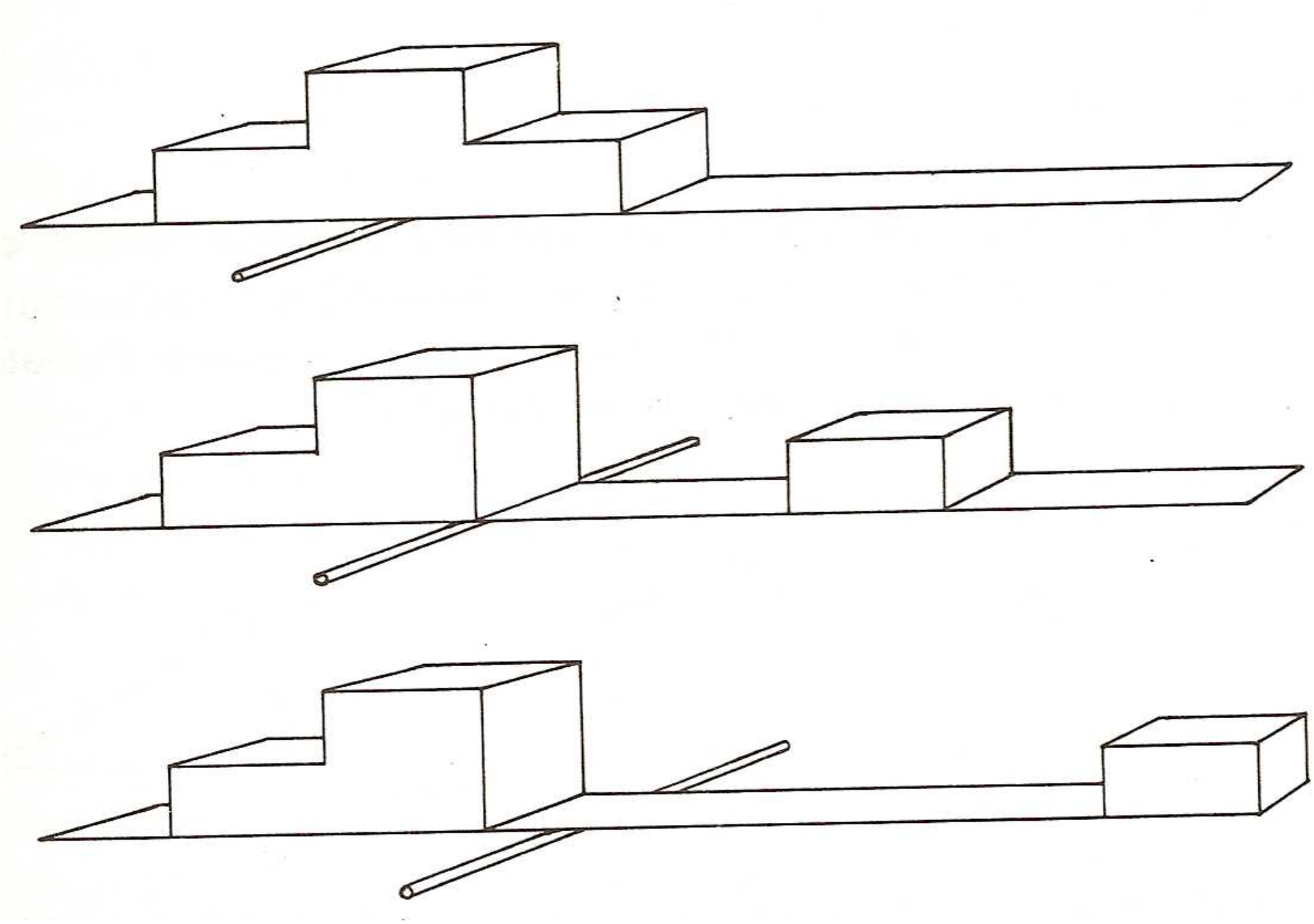
$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0.$$

- Si la distribución es muy asimétrica, la media puede distorsionar nuestra percepción de cómo son los datos.
- La media es muy sensible a la existencia de datos atípicos en los datos.

Para **datos agrupados**, si $x_1, \ldots, x_k$ son las marcas de clase y $f_1, \ldots, f_k$ son las frecuencias relativas

$$\bar{x} = x_1 f_1 + \cdots + x_k f_k.$$

Posición de la media en un histograma



Medidas numéricas de posición: la mediana

Una medida alternativa de posición es la **mediana**. Para calcular la mediana:

- Se ordenan los datos de menor a mayor.
- Si el número de datos es impar, la mediana es el dato que ocupa la posición central.
- Si el número de datos es par, la mediana es la media de los dos datos centrales.

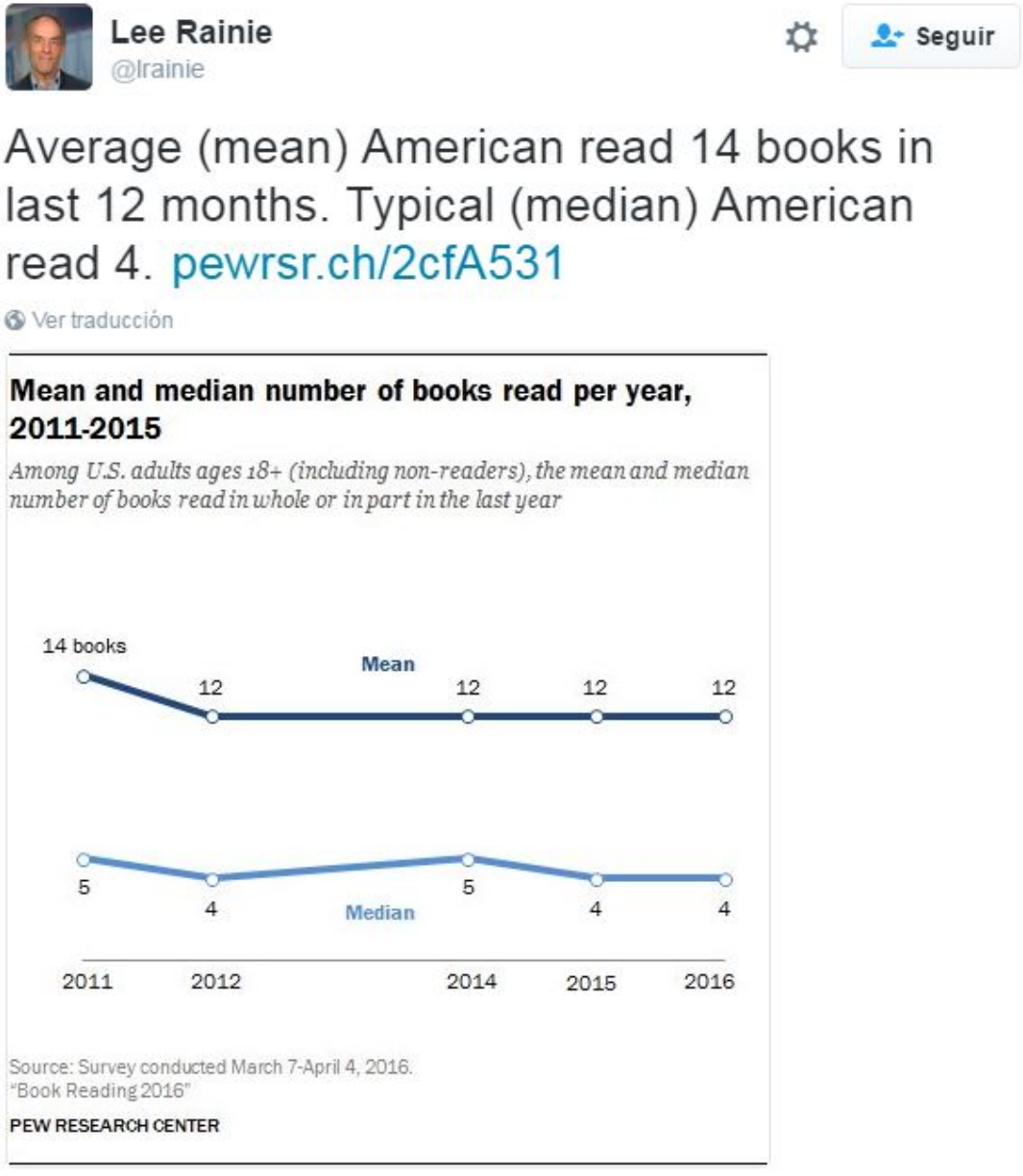
La mediana es *más robusta* que la media pero hace un uso menos eficiente de la información contenida en los datos.

Relación entre la simetría de una distribución y la posición relativa entre la media y la mediana.

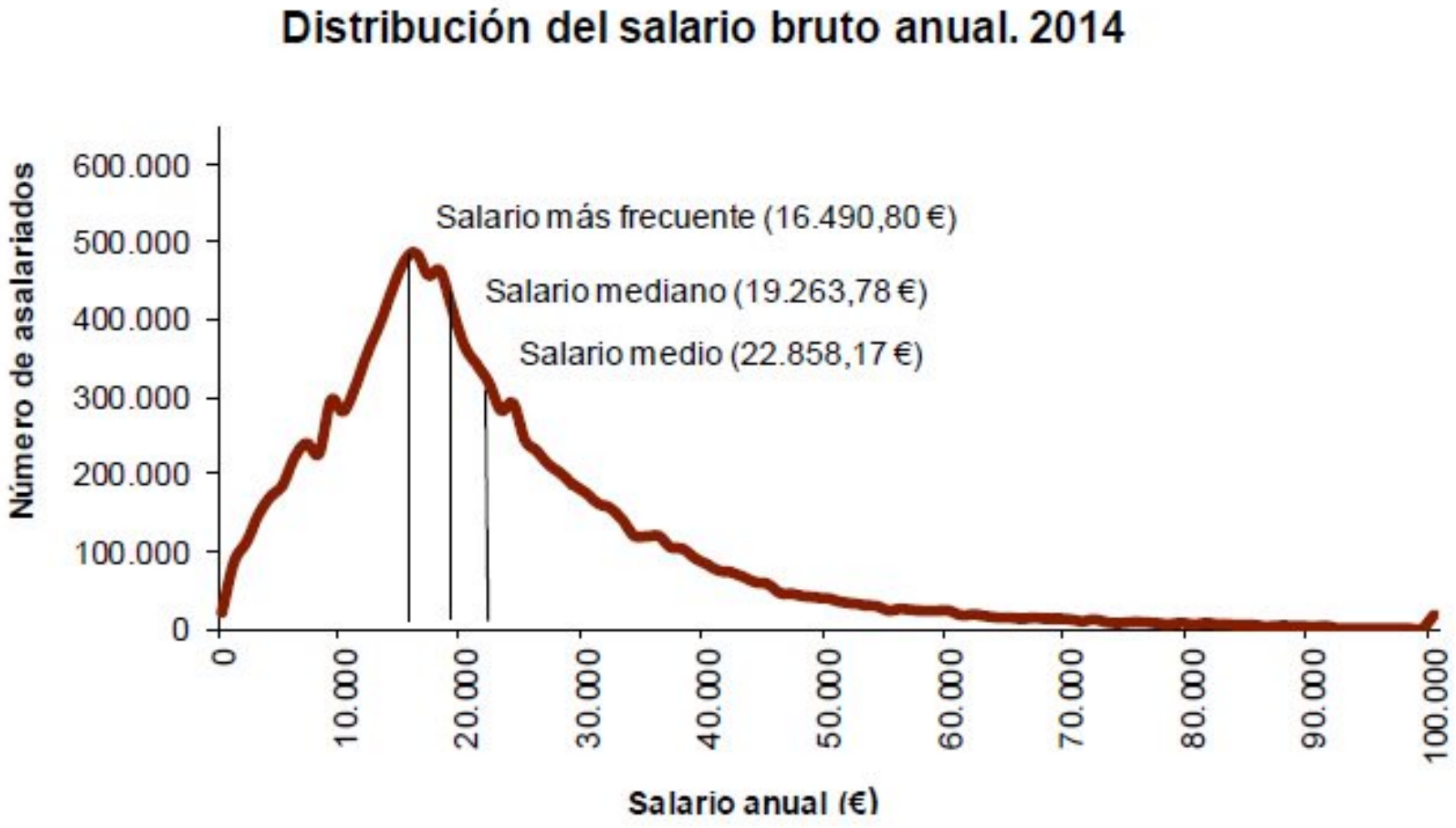
Mediana para datos agrupados.

Media y mediana

Media y mediana



Media y mediana



Fuente: Encuesta de Estructura Salarial. Año 2014. **INE**

Medidas de dispersión: rango intercuartílico

Una medida de dispersión muy sencilla es el *rango o recorrido* de los datos: el valor máximo menos el mínimo.

El rango sólo depende de los datos extremos por lo que no es muy conveniente.

Mejores propiedades tienen los cuartiles y el rango intercuartílico:

- El **primer cuartil**, Q_1, es la mediana de los datos menores que la mediana.
- El **tercer cuartil**, Q_3, es la mediana de los datos mayores que la mediana.
- El **rango, recorrido o amplitud intercuartílica** es la diferencia entre los dos cuartiles anteriores: $Q_3 - Q_1$.

De acuerdo con las anteriores definiciones, responde a las siguientes cuestiones:

¿Qué porcentaje de datos hay...

(a) ... entre Q_1 y Q_3?

(b) ... a la izquierda de Q_1?

(c) ... a la derecha de Q_3?

(d) ... entre el mínimo y Q_3?

Una descripción útil de un conjunto de datos viene dada por los cinco números siguientes:

$$\text{Mínimo, } Q_1\text{, Mediana, } Q_3\text{, Máximo}$$

El **cuantil** p (o **percentil** $100p$) es el valor que deja **a su izquierda** una proporción p (o porcentaje $100p$) de los datos.

Ejemplo: salarios en España

Encuesta de estructura salarial. Año 2006

MEDIAS Y PERCENTILES. Resultados Nacionales

Ganancia media anual por trabajador por sexo, estudios y media y percentiles.

Unidades: euros

	Media	Percentil 10	Percentil 25	Percentil 50	Percentil 75	Percentil 90
Total						
Todos los estudios	19.680,88	8.201,02	11.903,59	15.740,23	23.285,82	34.889,95
Varones						
Todos los estudios	22.051,08	10.608,18	13.483,55	17.204,27	25.671,40	38.620,68
Mujeres						
Todos los estudios	16.245,17	6.258,13	9.682,85	13.506,00	19.722,20	29.323,84

Cuestiones

- La media y la mediana de los salarios en España en 2006 fueron ... y ...
- ¿Cuál es la forma de la distribución de salarios?
- ¿Cuánto vale el rango intercuartílico?
- Un 10% de las mujeres ganaba más de ...
- Un 80% de los hombres ganaba entre ... y ...
- Un ... % de las mujeres ganaba más de 6258.13 euros.

Medidas de dispersión: la (cuasi)varianza y la (cuasi)desviación típica

Son las medidas de dispersión más utilizadas.

La **varianza** es el promedio de las desviaciones al cuadrado de los datos a su media.

Datos	$x_1, \ldots, x_n$
Desviaciones	$x_1 - \bar{x}, \ldots, x_n - \bar{x}$
Desviaciones al cuadrado	$(x_1 - \bar{x})^2, \ldots, (x_n - \bar{x})^2$

La varianza es el promedio de las desviaciones al cuadrado:

$$v_x = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$$

La cuasivarianza

Como

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0,$$

dadas $n - 1$ desviaciones, podemos despejar la restante.

En realidad sólo disponemos de $n - 1$ desviaciones independientes.

Como consecuencia, es más correcto dividir por $n - 1$ que por n.

La **cuasivarianza** muestral es

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

Fórmulas alternativas

La varianza se puede escribir como la media de los datos al cuadrado, menos el cuadrado de la media de los datos.

Por lo tanto,

$$v_x = \frac{x_1^2 + \cdots + x_n^2}{n} - \bar{x}^2$$

$$S^2 = \frac{n}{n-1}\left(\frac{x_1^2 + \cdots + x_n^2}{n} - \bar{x}^2\right)$$

Estas fórmulas suelen ser más rápidas para calcular v_x y S^2.

Cuasidesviación típica

La **cuasidesviación típica** es la raíz cuadrada de S^2:

$$S = \sqrt{\frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}$$

S se usa más que S^2 porque mide la dispersión en la misma escala que los datos originales.

Una medida adimensional relacionada es el **coeficiente de variación**:

$$\mathrm{CV} = \frac{S}{|\bar{X}|}$$

Cuestiones

1. Siempre $S^2 \geq 0$. Da un ejemplo de un conjunto de datos tal que $S^2 = 0$.
2. Dado un conjunto de observaciones medidas en kg, supongamos que cambiamos las unidades y las pasamos a gramos (es decir, multiplicamos por mil). Determina si son verdaderas o falsas las siguientes afirmaciones:
   - Tanto la media como la mediana de los nuevos datos se multiplican también por mil.
   - La varianza se multiplica también por mil.

   ¿Cómo cambiaría la desviación típica?
3. Ahora sumamos 100 a todos los datos. Determina si son verdaderas o falsas las siguientes afirmaciones:
   - Los cuartiles no cambian.
   - El rango intercuartílico no cambia.
   - La desviación típica no cambia.
4. ¿Cuál es la fórmula de la varianza para datos agrupados?

Descripción numérica

Descriptives

	LONG	PESO	CONC
N	171	171	171
Missing	0	0	0
Mean	40.0	1148	1.19
Std. error mean	0.651	67.0	0.0582
Median	39.0	873	0.930
Standard deviation	8.52	876	0.762
Variance	72.5	766556	0.580
Range	39.8	4308	3.49
Minimum	25.2	203	0.110
Maximum	65.0	4511	3.60
25th percentile	33.3	491	0.600
50th percentile	39.0	873	0.930
75th percentile	46.2	1455	1.60

Cuestiones

- Comparando los valores de la media y la mediana, ¿qué podemos decir sobre la simetría de las distribuciones?
- Verdadero o falso: Al menos para 100 peces, la concentración de mercurio es superior a 0.93 ppm.
- Verdadero o falso: La longitud de aproximadamente 42 peces es mayor que 25.20 cm y menor que 33.3 cm.
- ¿Cuál es el rango intercuartílico de la variable que mide el peso de los peces?

Estandarización o tipificación

Consiste en restarle a cada observación la media de todos los datos y dividir por la desviación típica:

$$z_i = \frac{x_i - \bar{x}}{S}$$

Representa la distancia de x_i a la media expresada en desviaciones típicas (el signo indica si el dato es mayor o menor que la media).

Utilidad de la tipificación

- Eliminar los efectos de las unidades de medida.
- Detectar posibles valores atípicos en los datos.
- Realizar comparaciones de los valores de una variable en diferentes poblaciones.

¿Cuánto vale la media y la desviación típica de los datos estandarizados?

Ejemplo

En un examen la nota media de los alumnos fue de 50 puntos y la cuasidesviación típica fue de 10.

- Estandariza las notas siguientes: 60, 45, 75.
- Si la nota estandarizada de un alumno fue -2, el alumno obtuvo ... en el examen.
- Una nota de 60 en este examen equivale después de estandarizar a otra de ... en otro examen cuya media fue 40 y cuya cuasidesviación típica fue 5.

Efecto de estandarizar un conjunto de datos

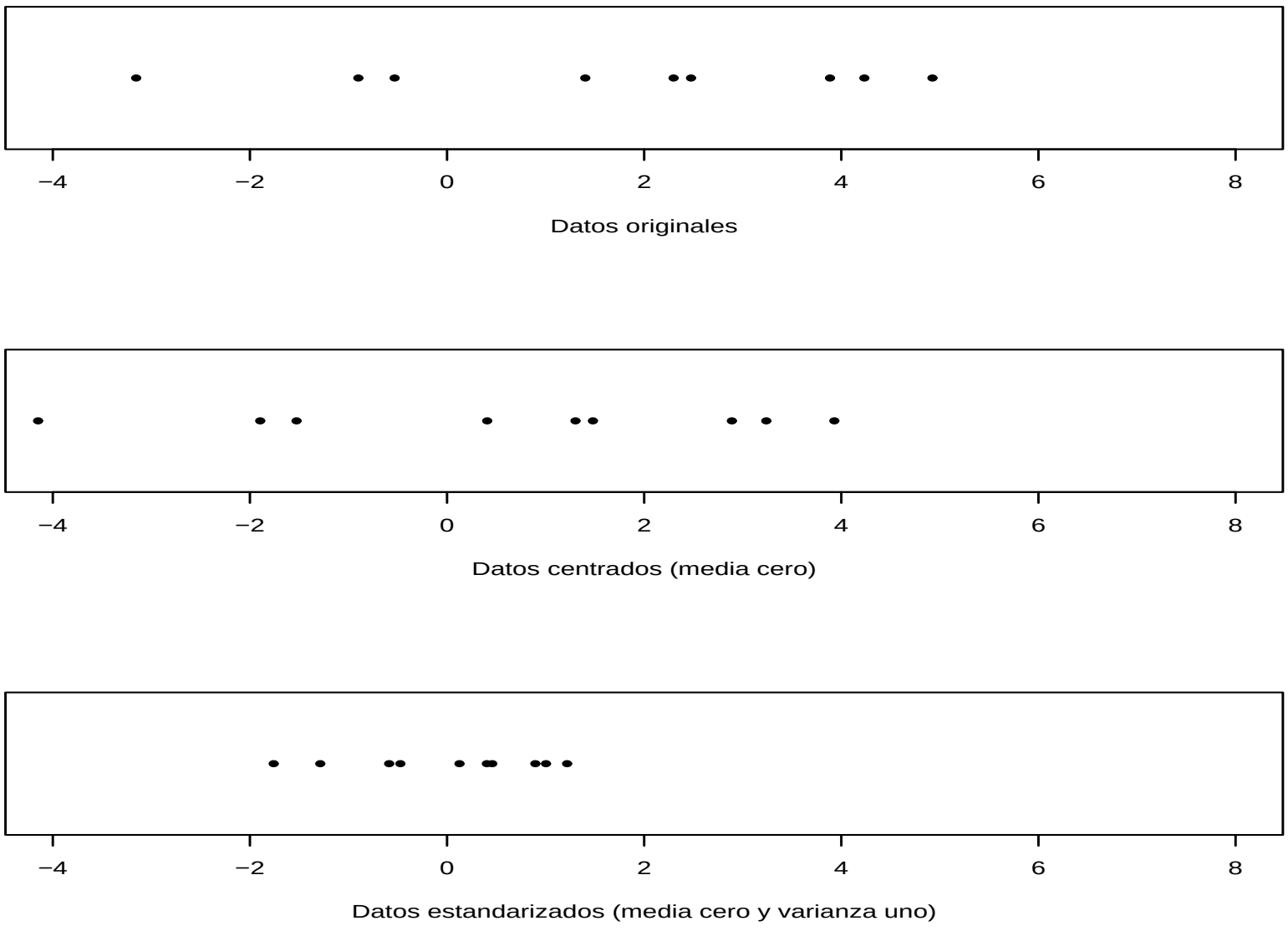
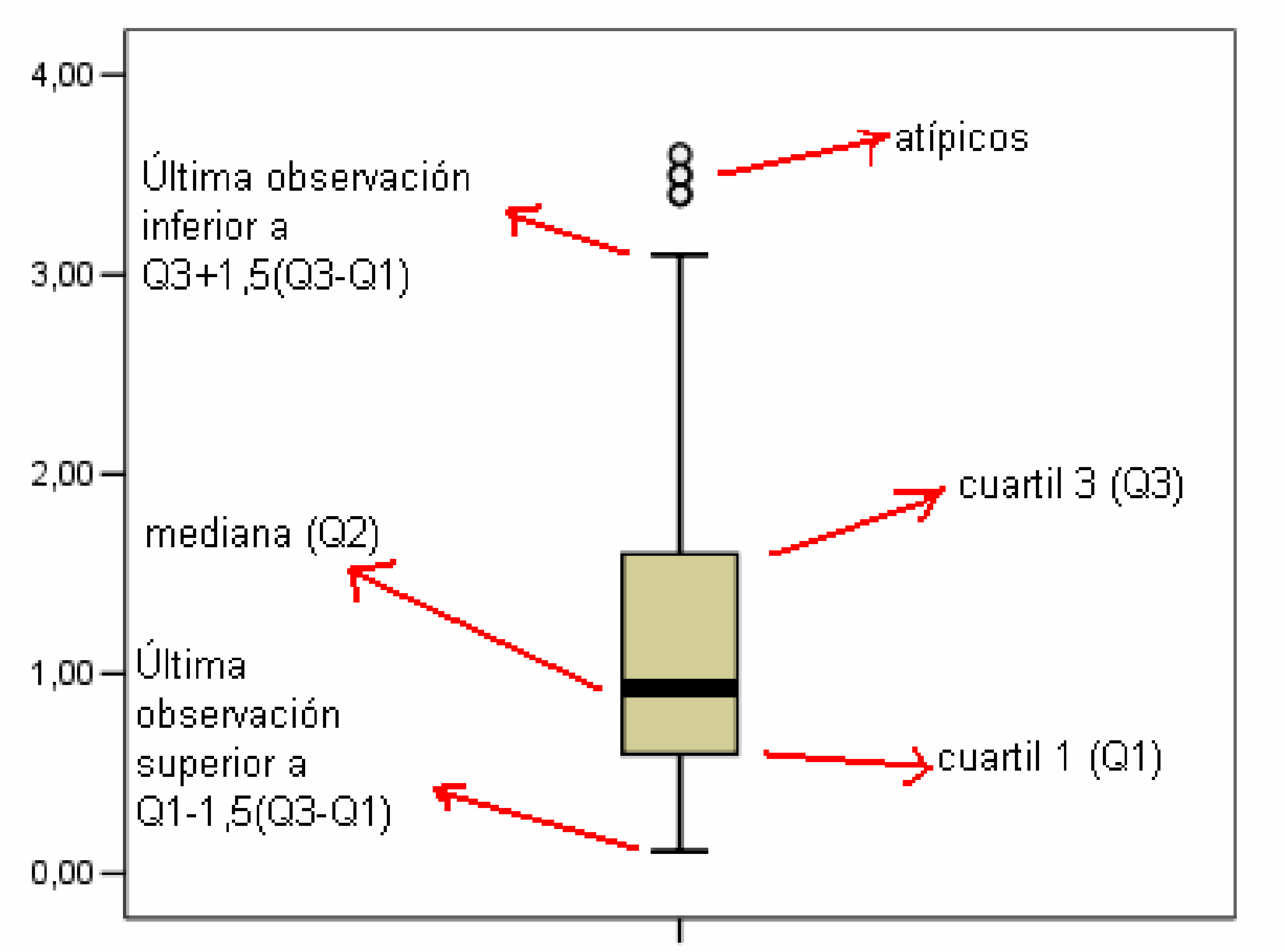


Diagrama de cajas



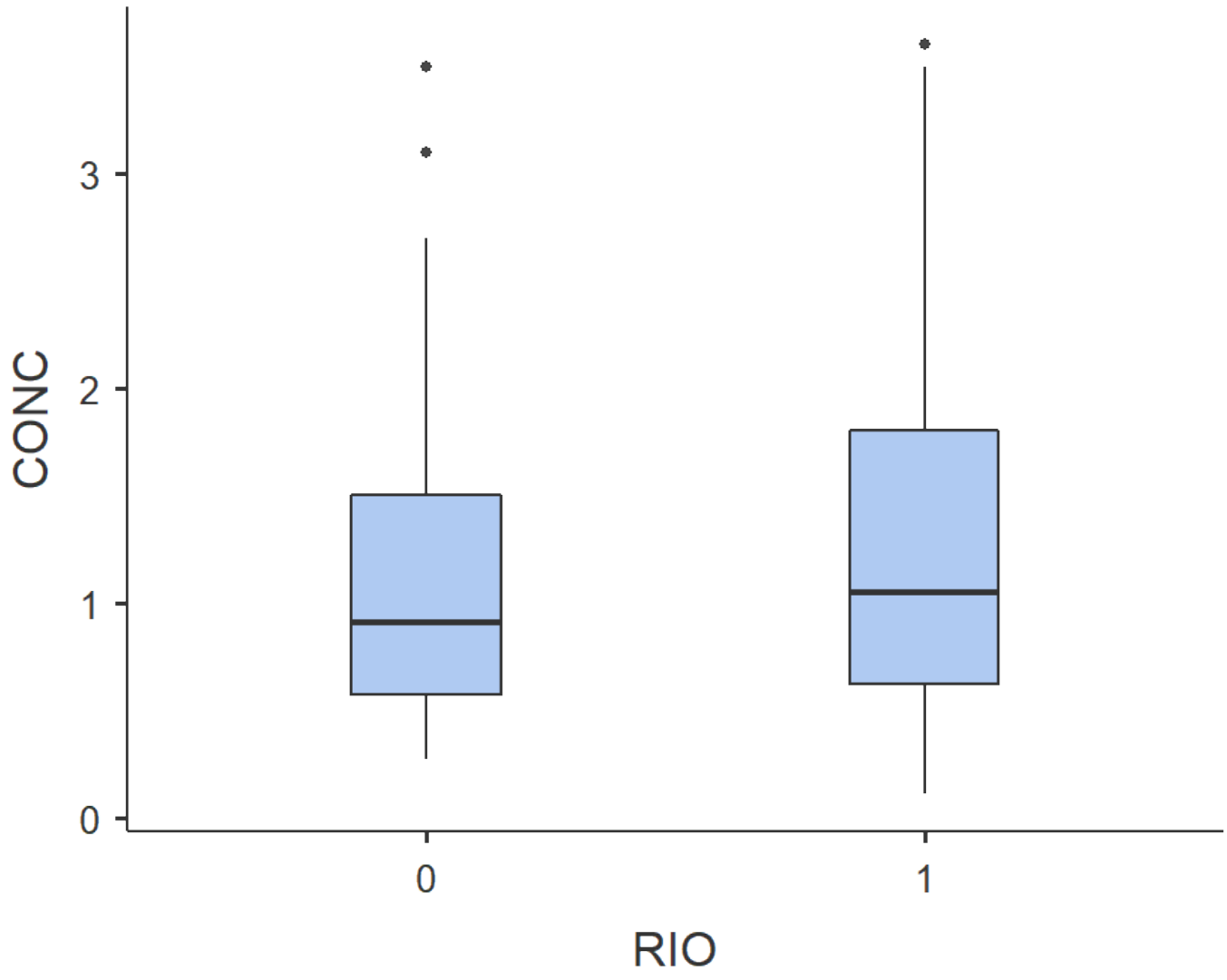
¿Para qué sirven?

Los diagramas de cajas son especialmente útiles para comparar varios conjuntos de datos.

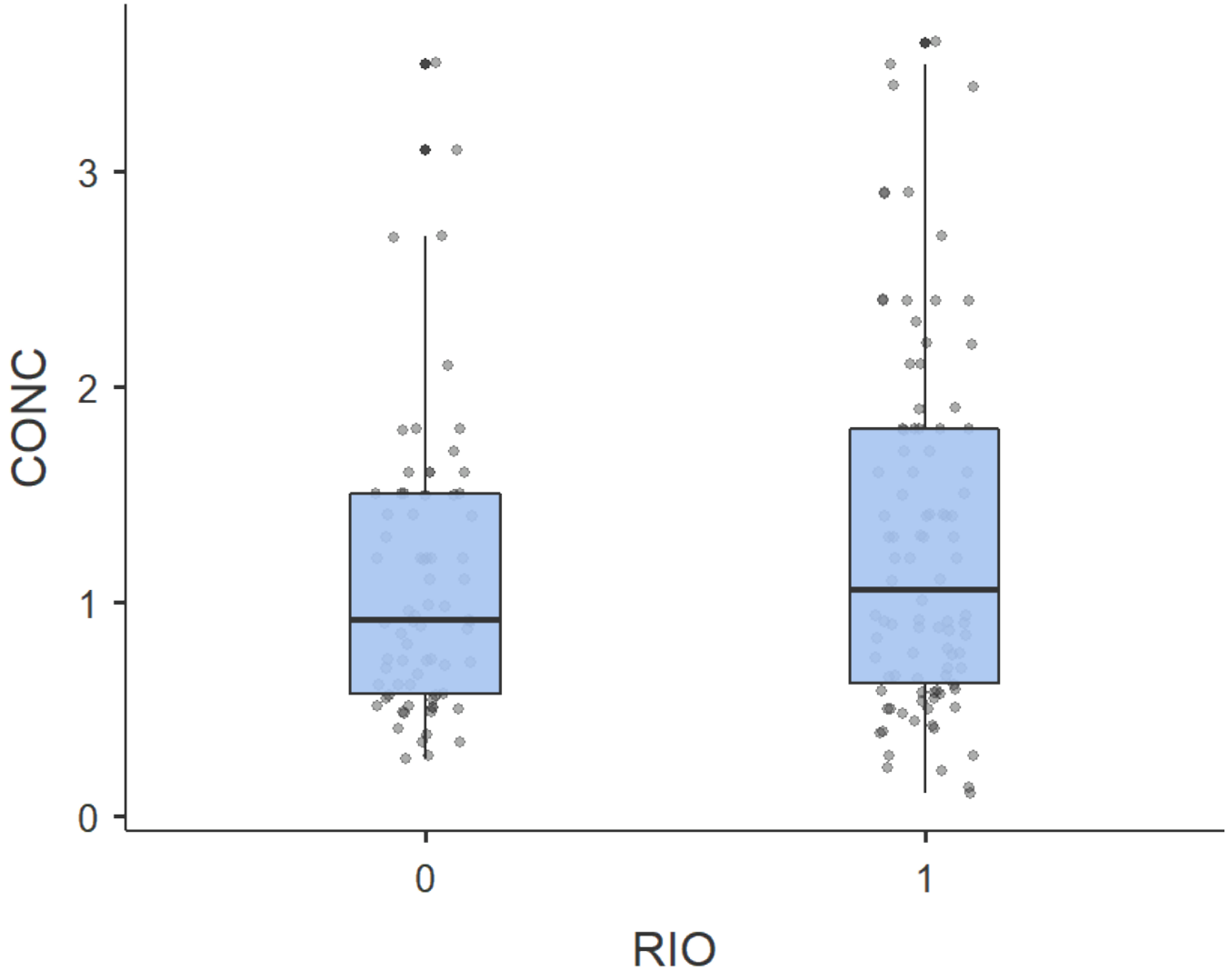
Además, proporcionan información sobre:

- La posición (mediana) y la dispersión (rango intercuartílico) de los datos.
- La simetría de la distribución (comparamos el tamaño de las cajas).
- La existencia de datos que se desvían del patrón general (datos atípicos).

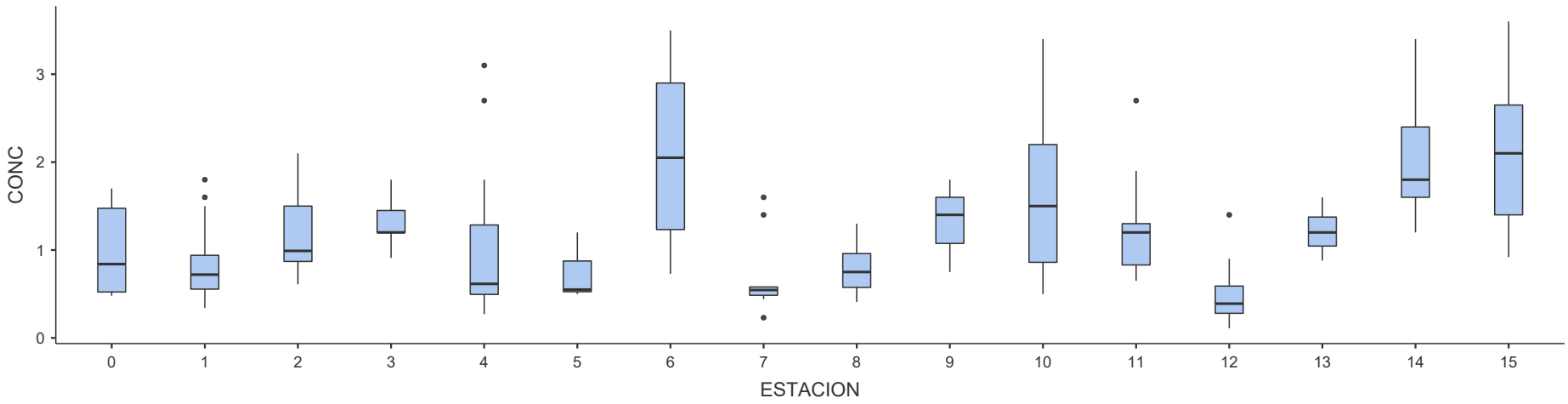
Concentración de mercurio y río



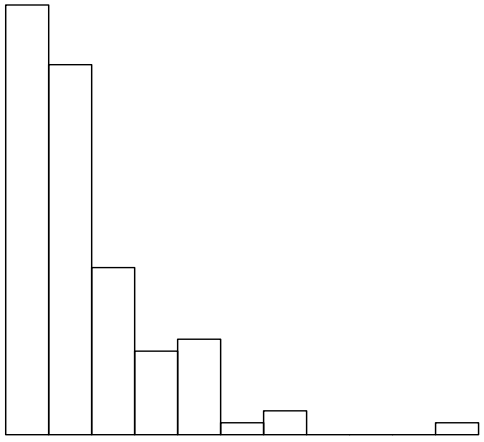
Concentración de mercurio y río

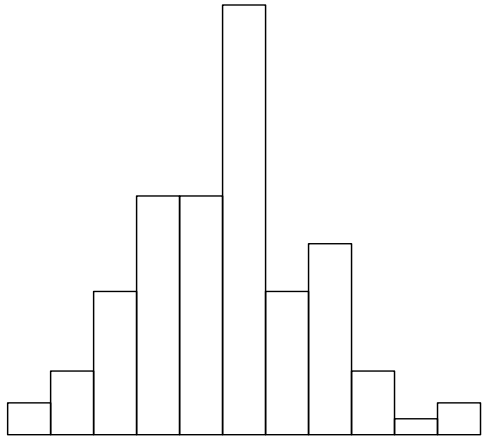


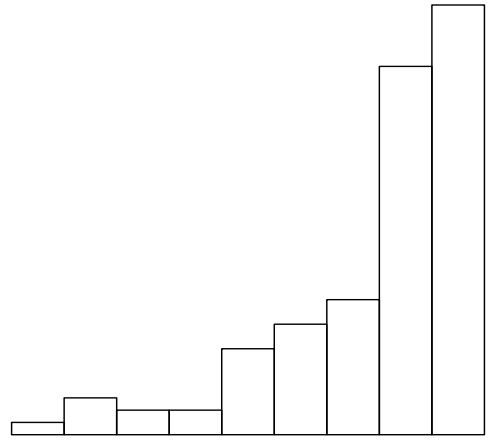
Concentración de mercurio y estación

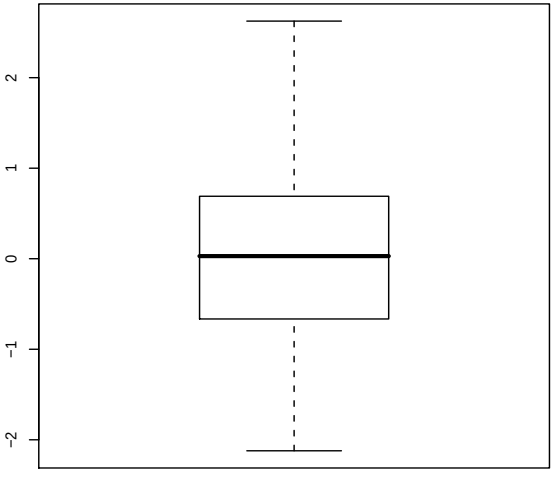


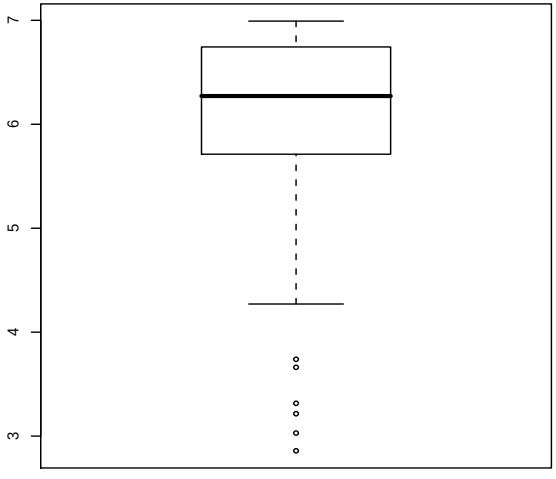
Relaciona cada histograma con su diagrama de cajas











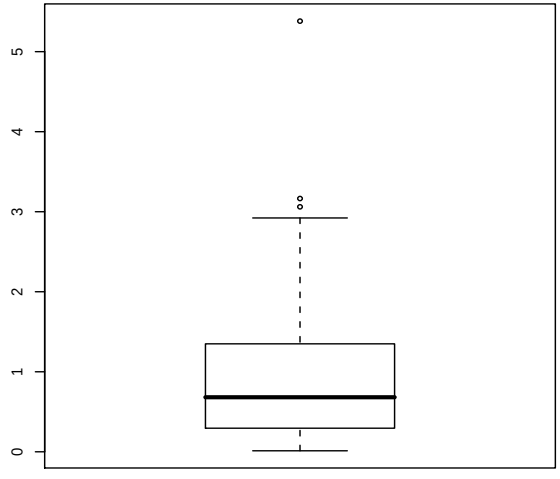
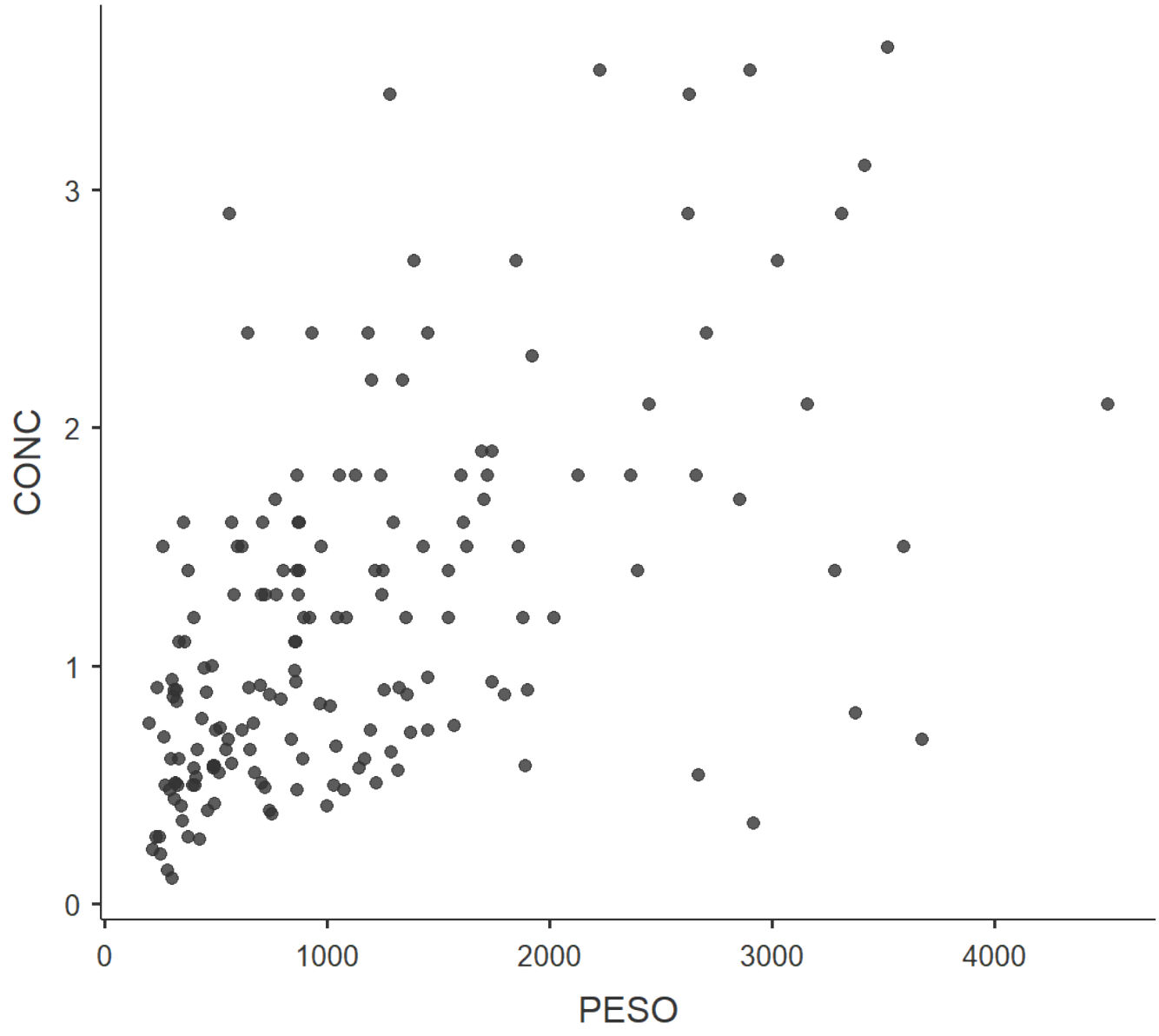


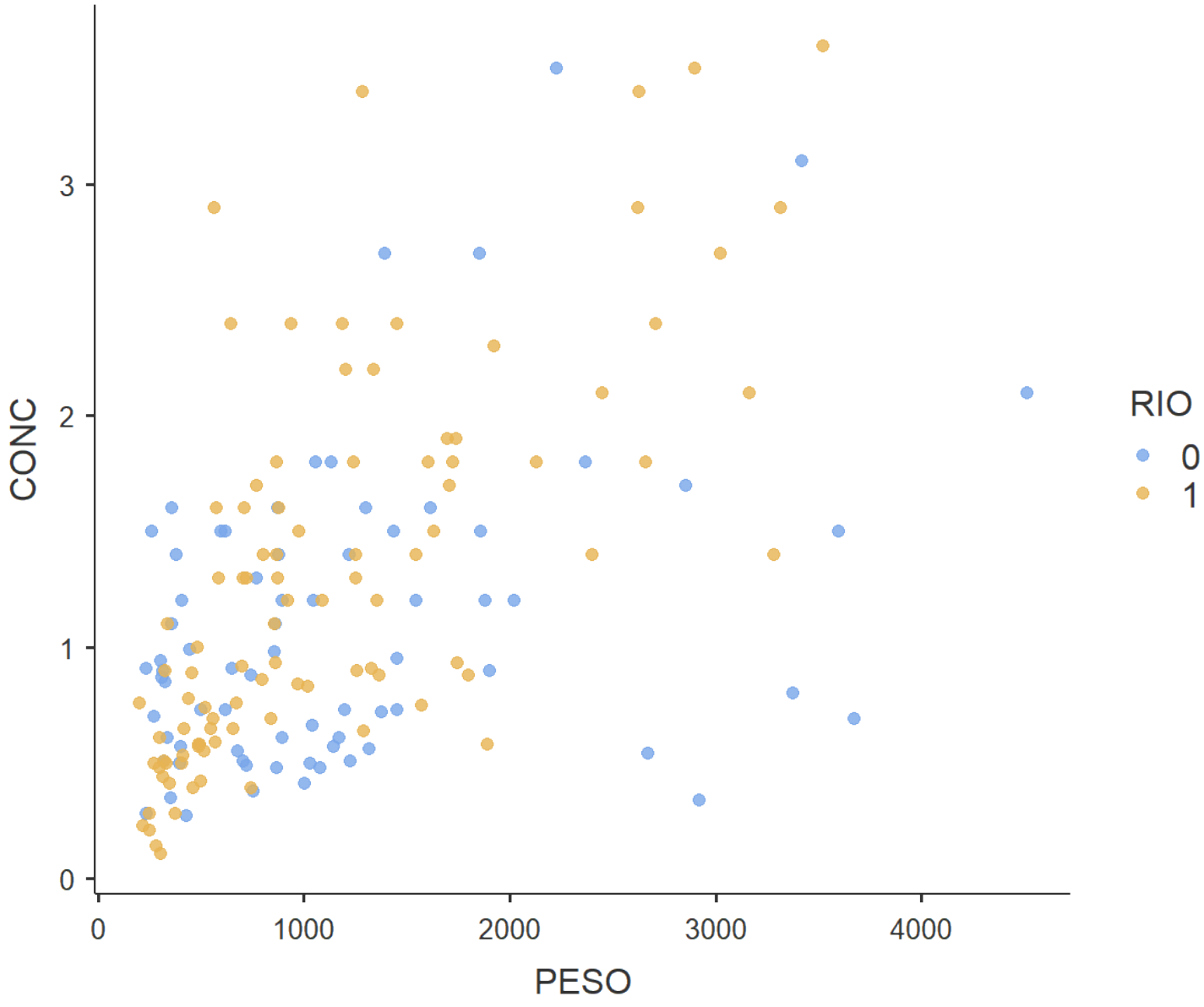
Diagrama de dispersión: Concentración frente a peso



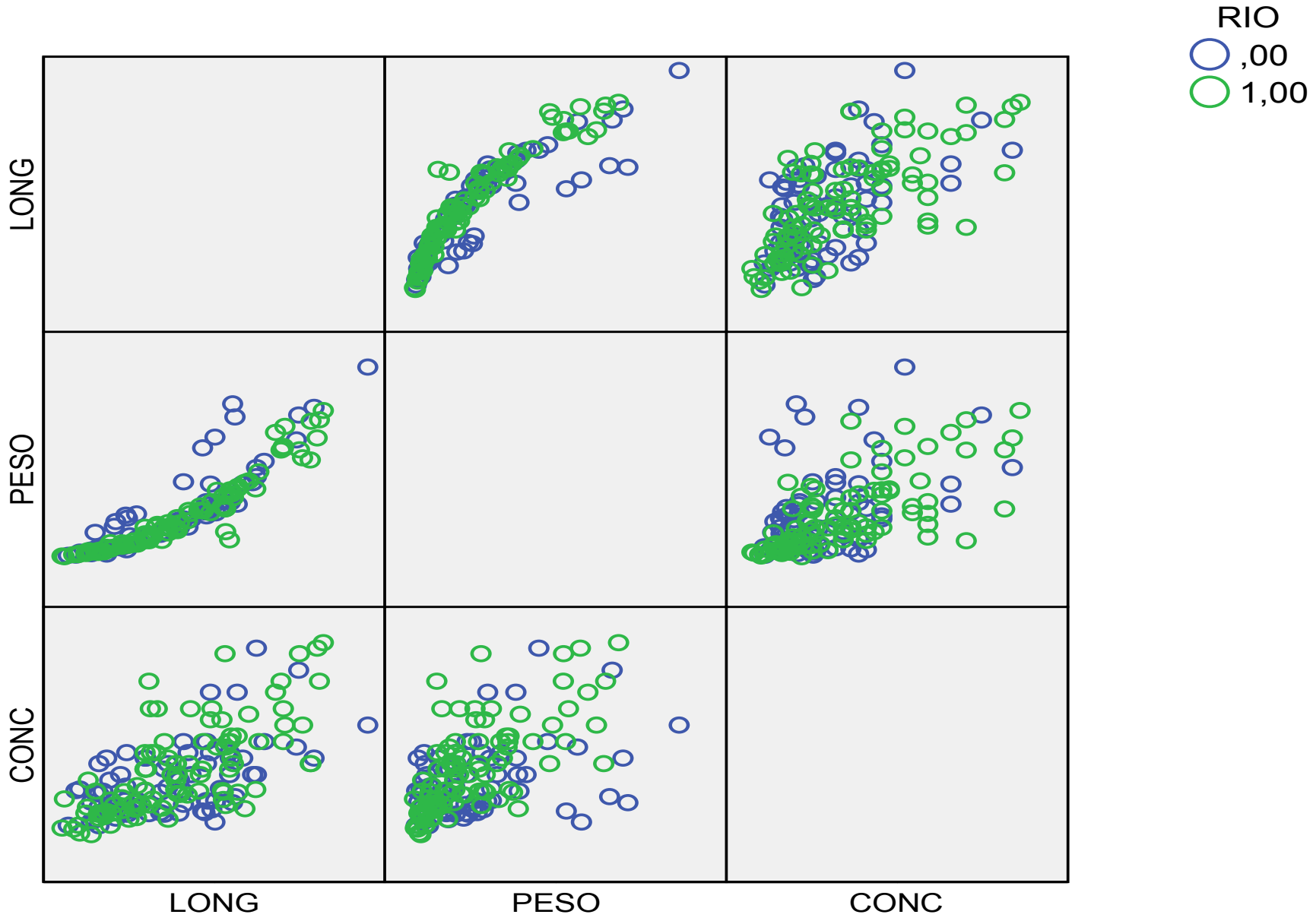
Interpretación de un diagrama de dispersión

- Es importante fijarse en las unidades de cada eje
- ¿Se observa alguna asociación entre las variables?
- ¿Cómo es de estrecha la asociación entre las variables?
- ¿Cuál es la "dirección" de la asociación entre las variables?
- ¿Hay algún punto o colección de puntos que no siga el patrón general del resto?
- Si hay una tercera variable cualitativa, resulta conveniente utilizar símbolos o colores diferentes para cada valor de esta tercera variable.

Concentración frente a peso (color según río)



Matriz de diagramas de dispersión



Covarianza

Se dispone de un conjunto de n pares de observaciones

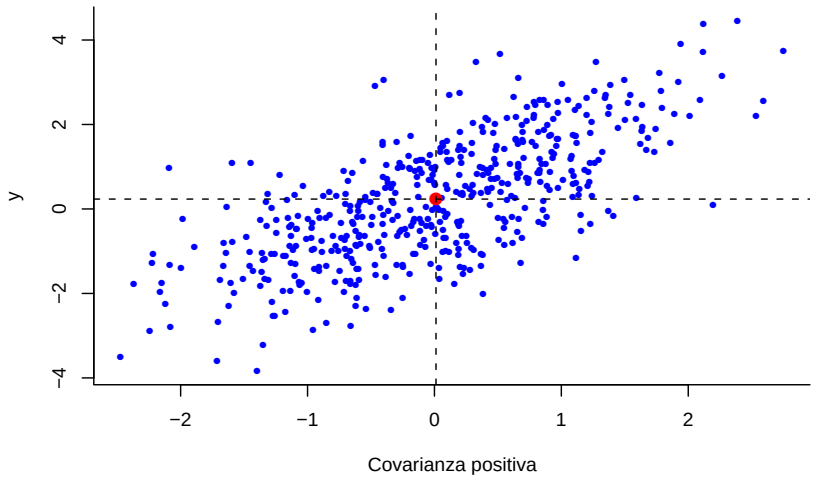
$$(x_1, y_1), \ldots, (x_n, y_n).$$

La covarianza entre x e y es una medida numérica para cuantificar el grado de asociación lineal entre x e y:

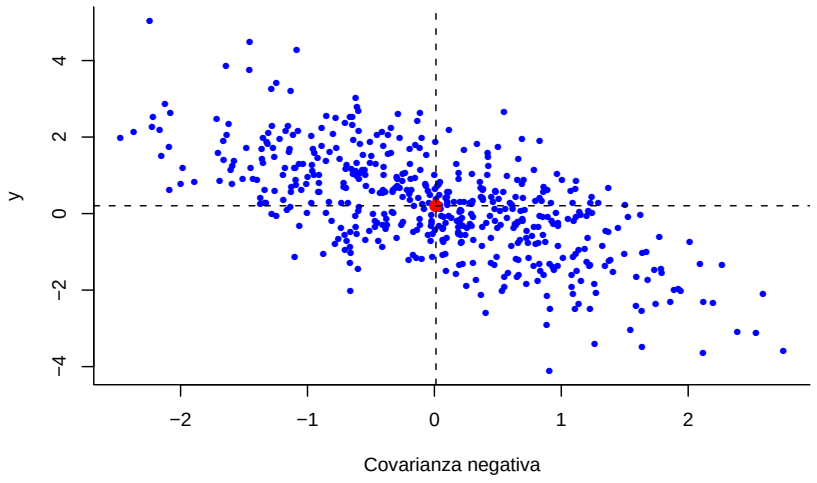
x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
x_1	y_1	$x_1 - \bar{x}$	$y_1 - \bar{y}$	$(x_1 - \bar{x})(y_1 - \bar{y})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	y_n	$x_n - \bar{x}$	$y_n - \bar{y}$	$(x_n - \bar{x})(y_n - \bar{y})$

$$S_{xy} = \frac{1}{n-1}[(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})]$$

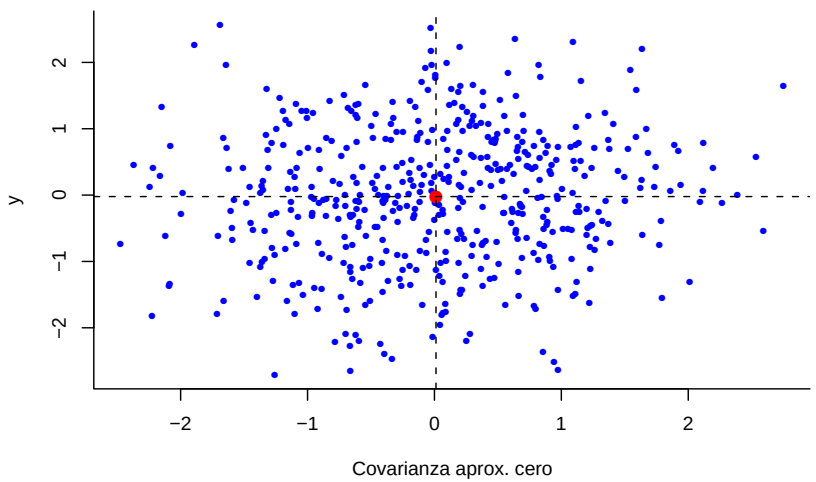
Interpretación de la covarianza



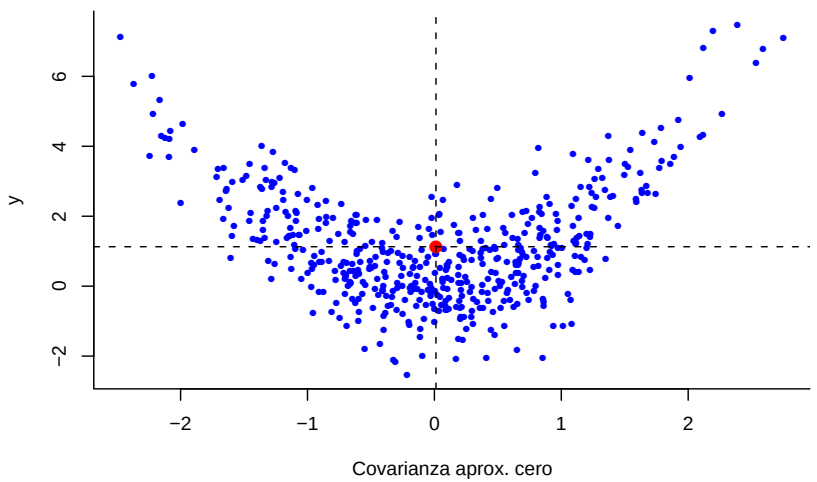
Covarianza positiva



Covarianza negativa



Covarianza aprox. cero



Covarianza aprox. cero

Covarianza

Fórmula alternativa:

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

Propiedades:

- $S_{xy} = S_{yx}$.
- S_{xx} es la cuasivarianza de x.
- S_{xy} depende de las unidades en que se midan x e y.
- También a veces se define la covarianza dividiendo por n en lugar de $n-1$. En este caso, $S_{xx} = v_x$.

Coeficiente de correlación

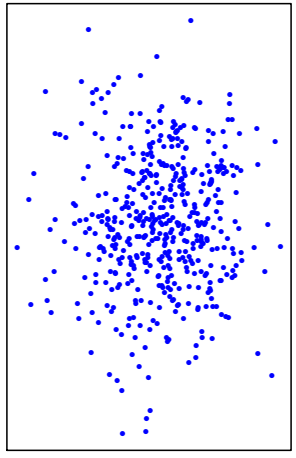
Resulta conveniente disponer de una medida de relación lineal que no dependa de las unidades. Para ello, se normaliza S_{xy} dividiendo por el producto de desviaciones típicas, lo que lleva al **coeficiente de correlación**:

$$r_{xy} = \frac{S_{xy}}{S_x S_y}.$$

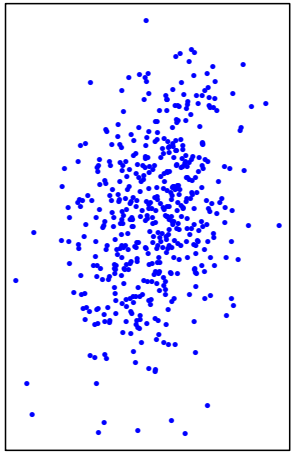
Propiedades del coeficiente de correlación:

- No depende de las unidades.
- ¿Cuánto vale r_{xx}?
- Siempre toma valores entre -1 y 1. Sólo vale 1 ó -1 cuando los puntos están perfectamente alineados.
- Su signo se interpreta igual que el de la covarianza.

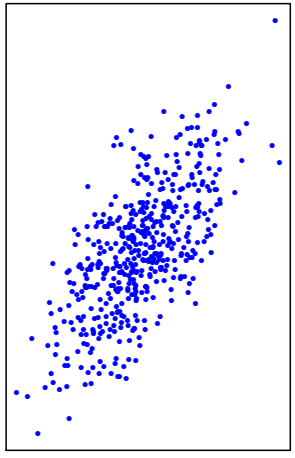
Ejemplos de correlaciones



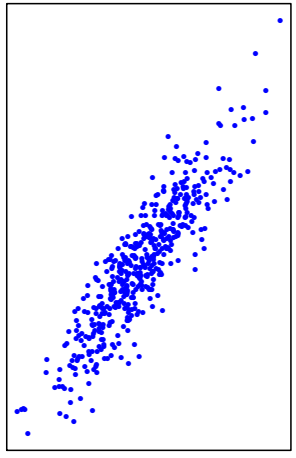
r = 0.1



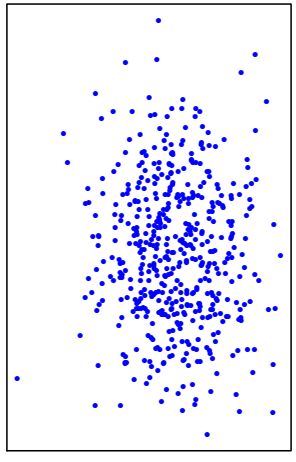
r = 0.3



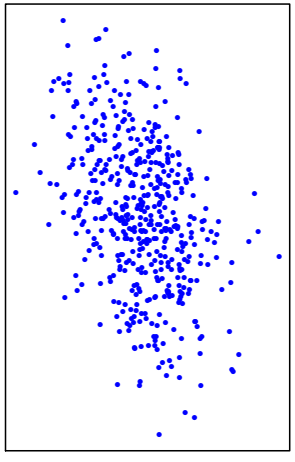
r = 0.7



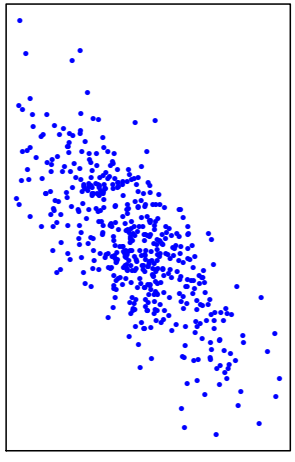
r = 0.9



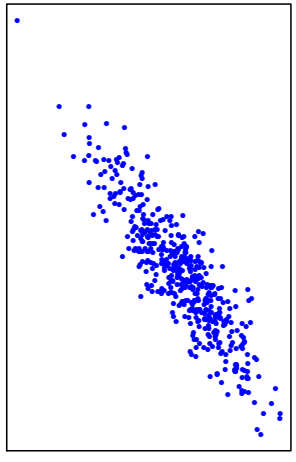
r = −0.1



r = −0.3



r = −0.7



r = −0.9

Covarianzas y correlaciones de los datos

Correlaciones

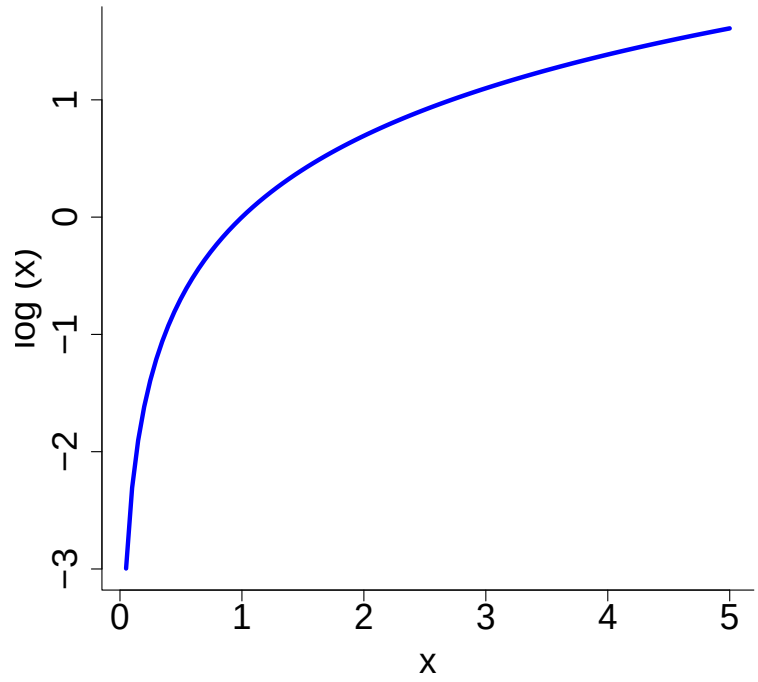
		LONG	PESO	CONC
LONG	Correlación de Pearson	1	,900	,650
	Sig. (bilateral)		,000	,000
	Suma de cuadrados y productos cruzados	12332,114	1141004	716,835
	Covarianza	72,542	6711,790	4,217
	N	171	171	171
PESO	Correlación de Pearson	,900	1	,554
	Sig. (bilateral)	,000		,000
	Suma de cuadrados y productos cruzados	1141004	1E+008	62786,546
	Covarianza	6711,790	766555,9	369,333
	N	171	171	171
CONC	Correlación de Pearson	,650	,554	1
	Sig. (bilateral)	,000	,000	
	Suma de cuadrados y productos cruzados	716,835	62786,546	98,622
	Covarianza	4,217	369,333	,580
	N	171	171	171

Cuestiones

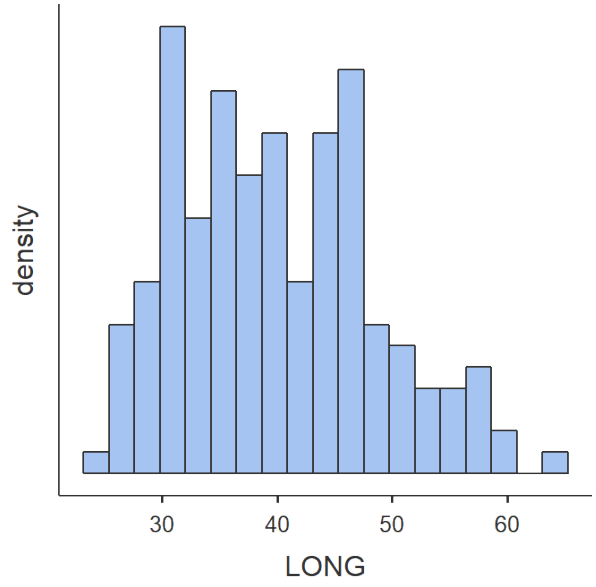
- ¿Cuánto vale la cuasivarianza de la longitud?
- Verdadero o falso: si multiplicamos por -1 una de las variables, la correlación no cambia.
- Verdadero o falso: Si $r = 0$, no hay relación entre las dos variables.
- En los siguientes conjuntos de puntos determina, si es posible, el valor c de forma que $r = 1$. Si no es posible explica por qué:
  - Conjunto 1: $(1, 1),\ (2, 3),\ (2, 3),\ (4, c)$.
  - Conjunto 2: $(1, 1),\ (2, 3),\ (3, 4),\ (4, c)$

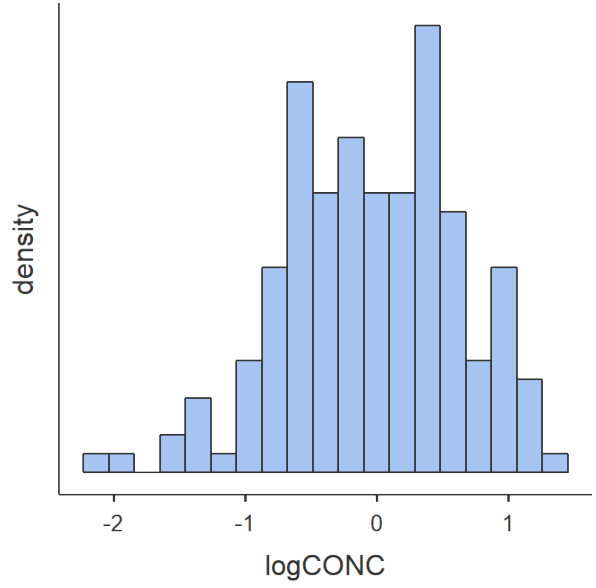
Tomar logaritmos

Si las observaciones x_i son positivas, a veces es conveniente trabajar con sus logaritmos $\log x_i$ en lugar de con las variables originales.

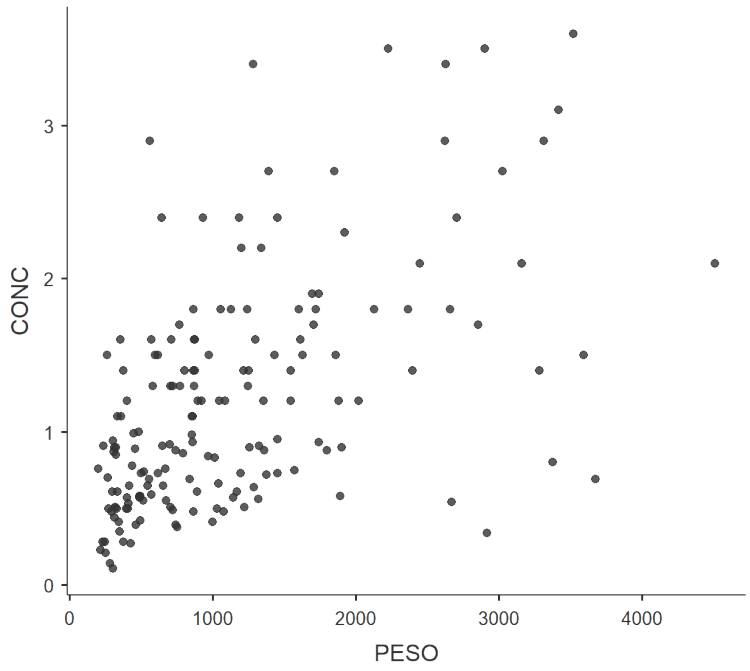


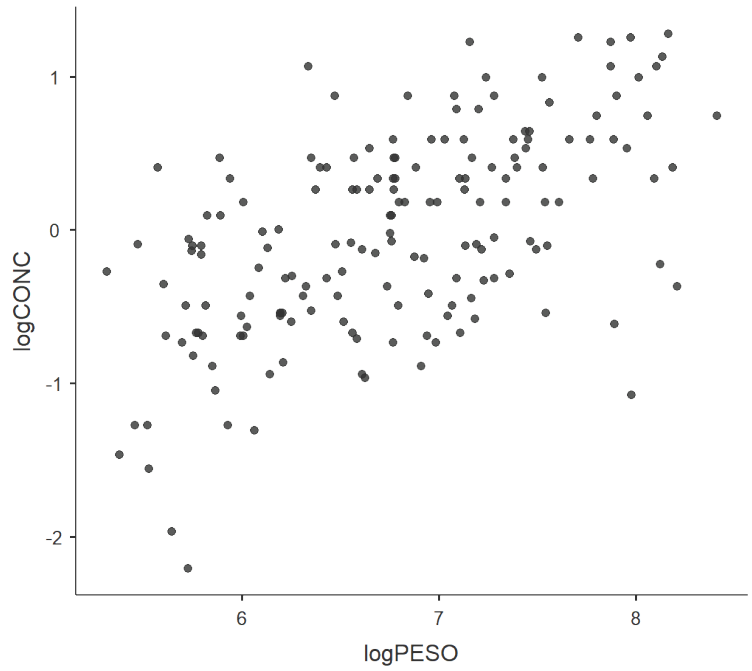
Efecto del logaritmo en la forma de la distribución



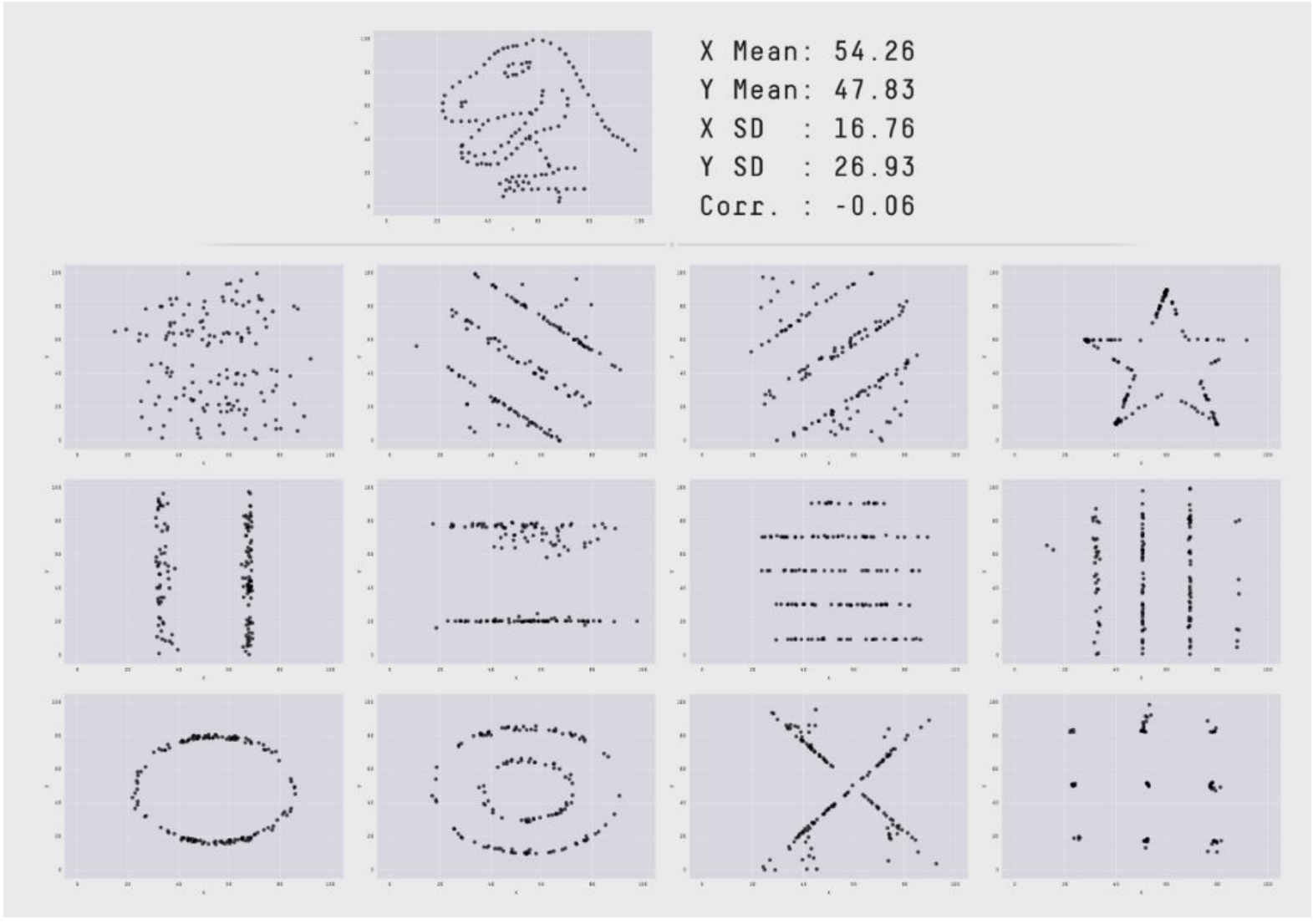


Efecto del logaritmo en el diagrama de dispersión





Siempre hay que representar gráficamente los datos



https://www.autodeskresearch.com/publications/samestats

Calorías y contenido en sodio en salchichas

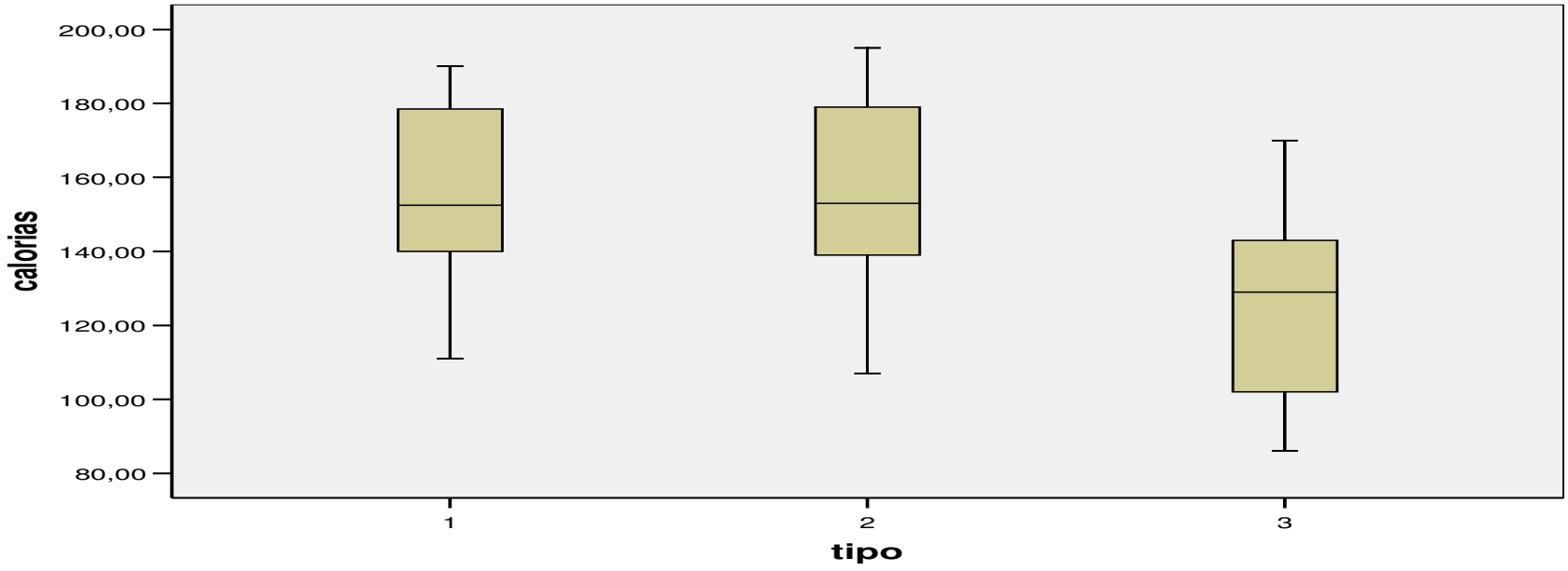
- Se ha considerado la cantidad de calorías y de sodio en salchichas de varias marcas de cada uno de los tipos siguientes:
  - Carne de ternera
  - Mezcla (hasta 15% de carne de pavo)
  - Carne de pavo

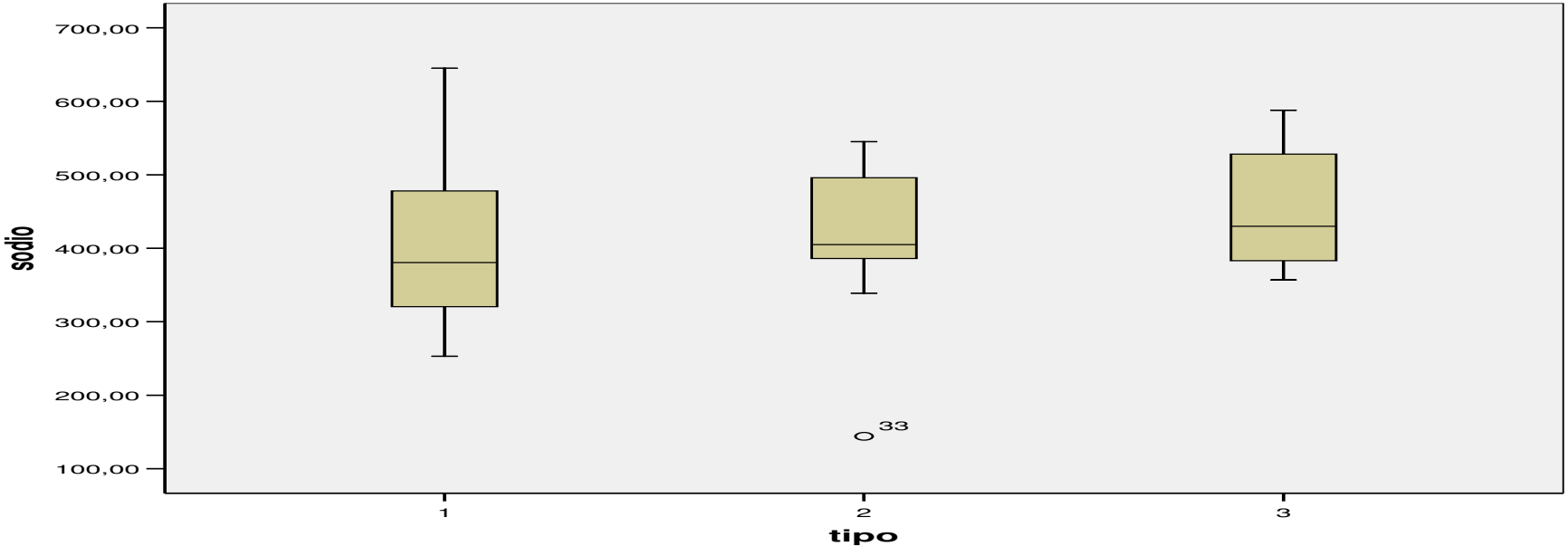
Nombre variable	Descripción
`tipo`	Tipo de carne (1=ternera, 2=mezcla, 3=pavo)
`calorias`	Cantidad de calorías
`sodio`	Cantidad de sodio

Medidas descriptivas numéricas

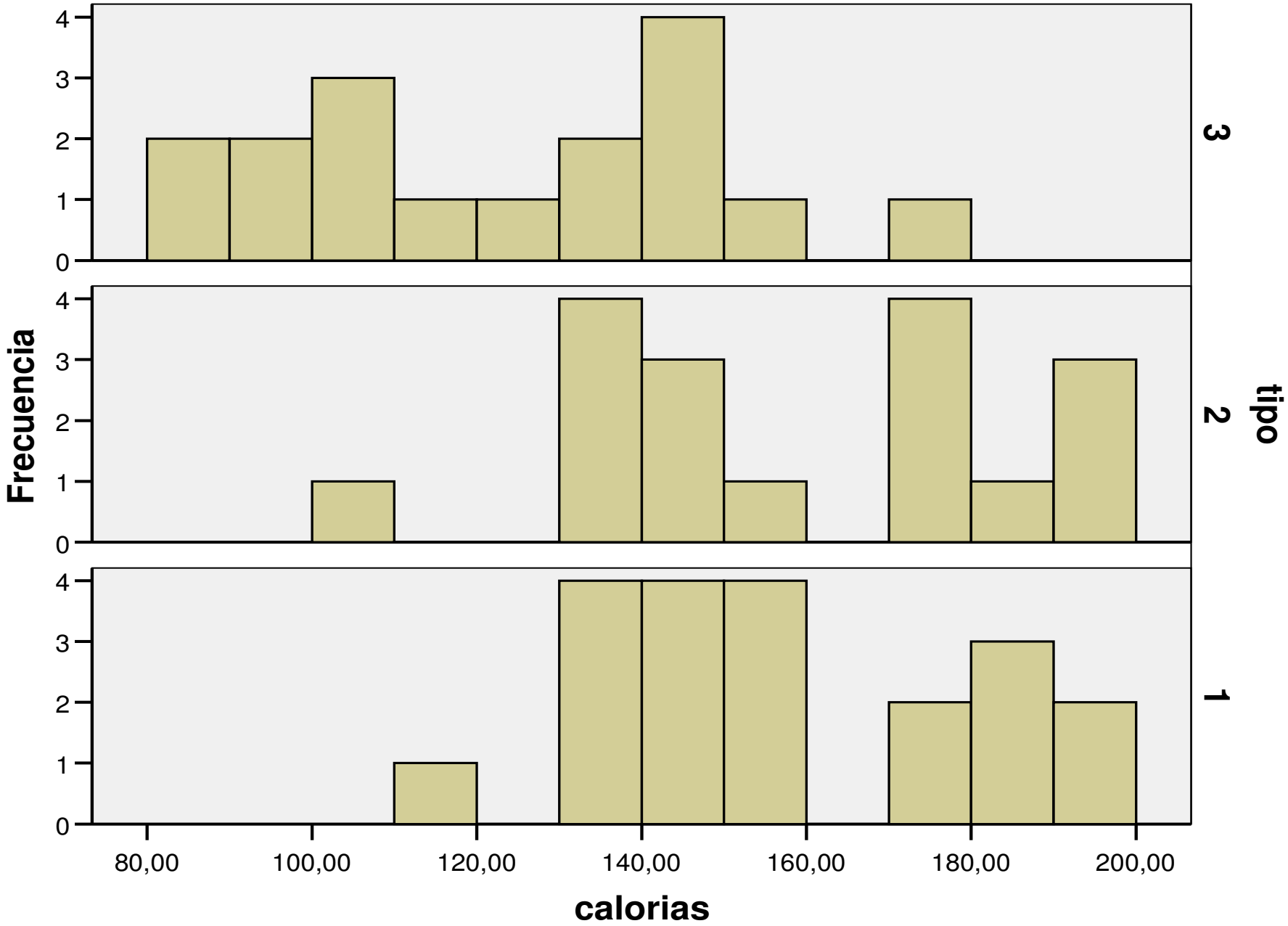
		calorias	sodio
N	Válidos	54	54
	Perdidos	0	0
Media		146,6111	424,8333
Error típ. de la media		3,95691	13,04440
Mediana		146,0000	405,0000
Desv. típ.		29,07727	95,85637
Varianza		845,487	9188,443
Mínimo		86,00	144,00
Máximo		195,00	645,00
Percentiles	25	132,0000	359,7500
	50	146,0000	405,0000
	75	173,5000	506,2500

Diagramas de cajas





Histogramas: cantidad de calorías



Histogramas: cantidad de sodio

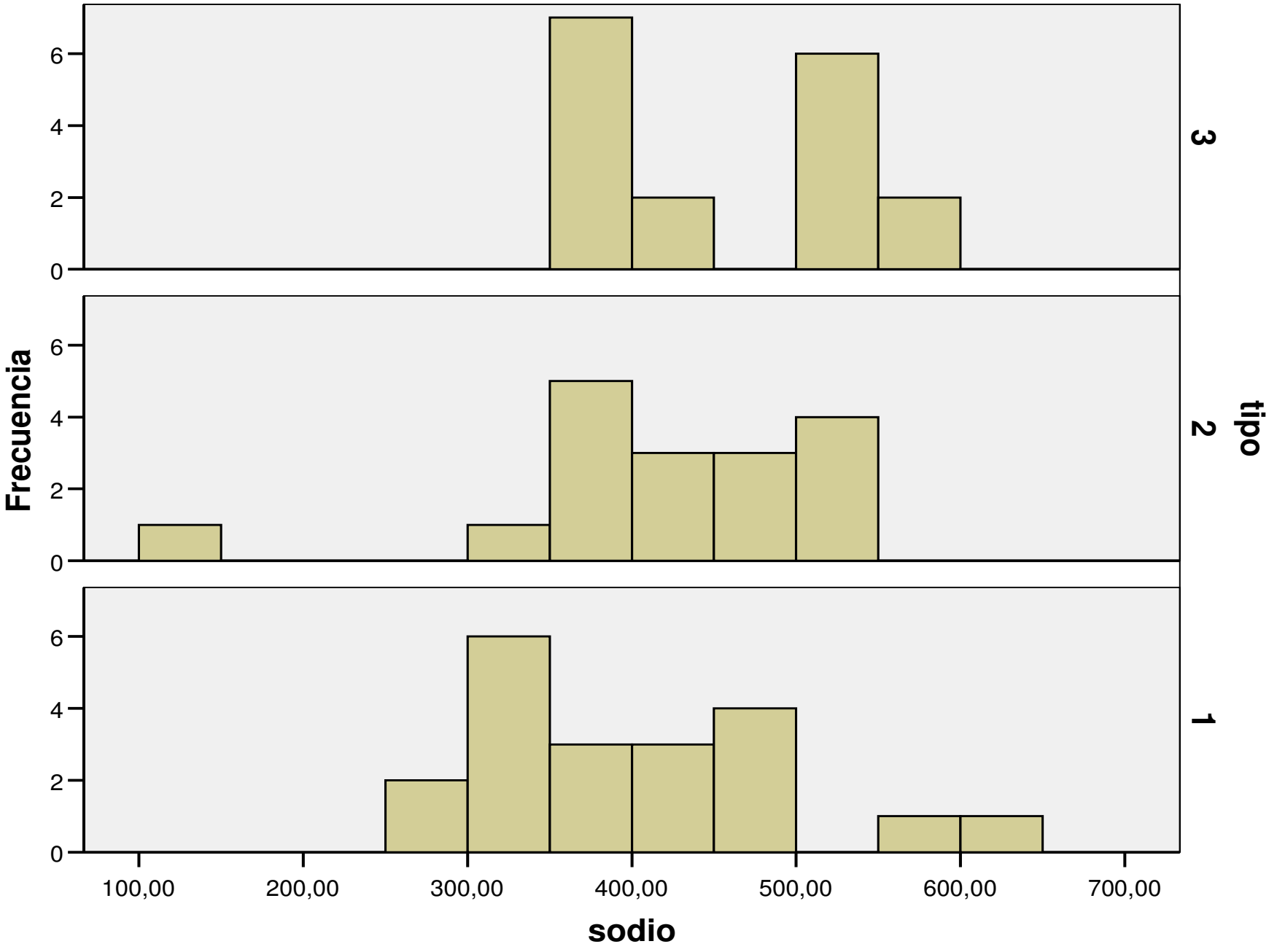
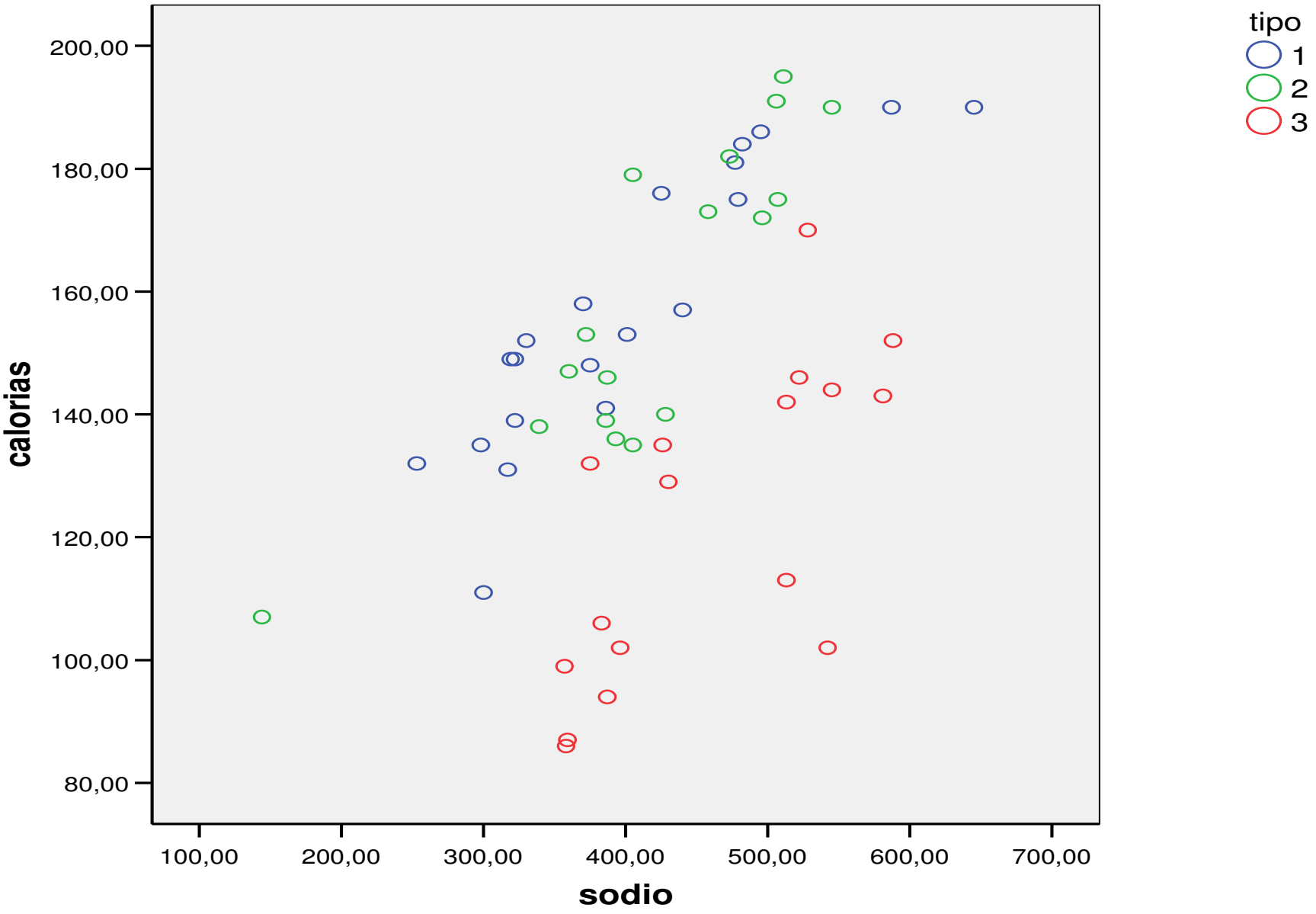


Diagrama de dispersión



Covarianzas y correlaciones

Correlaciones

		calorias	sodio
calorias	Correlación de Pearson	1	,516
	Sig. (bilateral)		,000
	Suma de cuadrados y productos cruzados	44810,833	76233,500
	Covarianza	845,487	1438,368
	N	54	54
sodio	Correlación de Pearson	,516	1
	Sig. (bilateral)	,000	
	Suma de cuadrados y productos cruzados	76233,500	486987,50
	Covarianza	1438,368	9188,443
	N	54	54

Cuestiones

- (V ó F) Aproximadamente 27 marcas de salchichas tienen entre 132 y 173 calorías.
- ¿Cuál es el rango intercuartílico de la cantidad de sodio?
- Calcula el coeficiente de variación de ambas variables.
- (V ó F) Aproximadamente 13 marcas de salchichas tienen un contenido de sodio entre 506.25 y 645.
- (V ó F) Con la información disponible en la tabla de medidas descriptivas numéricas es posible calcular la correlación entre ambas variables.
- (V ó F) Al menos el 75% de las marcas de salchichas de mezcla tienen más sodio que la mediana de las marcas de ternera.
- Identifica en el diagrama de dispersión el dato atípico que se observa en los diagramas de cajas.