ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

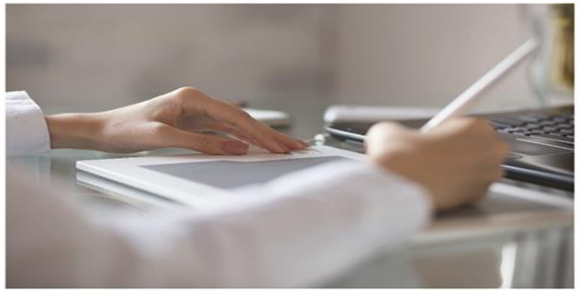
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

Preprocesamiento y Análisis de Datos

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau




Preprocesamiento

¿Por qué preprocesar los datos?

¿Por qué preprocesar los datos?

Datos incompletos

falta de valores en algunas variables

datos que vienen sólo agregados

En algunos casos es posible imputar los datos incompletos

¿Por qué preprocesar los datos?

Datos inconsistentes

diferencias en nombres y/o codificaciones de variables

diferencias de unidades

Combinar fuentes de datos distintas siempre es un desafío

¿Por qué preprocesar los datos?

Datos anómalos

datos que se alejan de la tendencia general

combinaciones de datos poco razonables

Identificar *outliers* no es sencillo

“Los datos no tienen sentido, tendremos que recurrir a la estadística”

Análisis de Datos

Definamos en qué consiste un dato

Un dato corresponde a una observación de un conjunto de variables relevantes

Si un dato es una persona, sus variables pueden ser: sexo, edad, ingreso, ocupación, orientación política, etc.

A su vez, una base de datos será un conjunto de datos observados

Definamos en qué consiste un dato

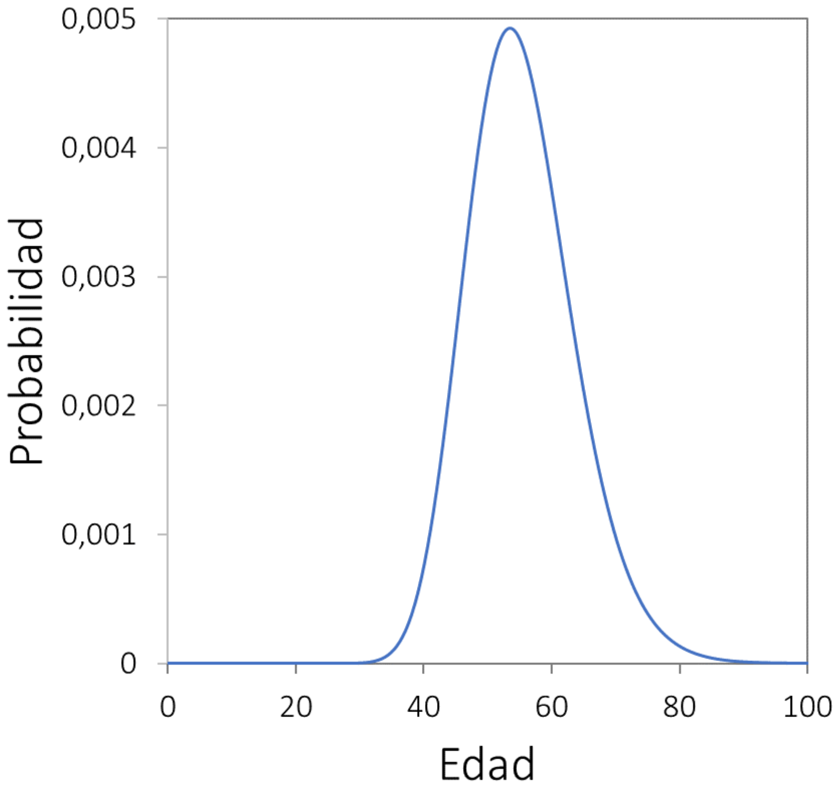
Los datos x_i corresponden a observaciones de fenómeno de interés

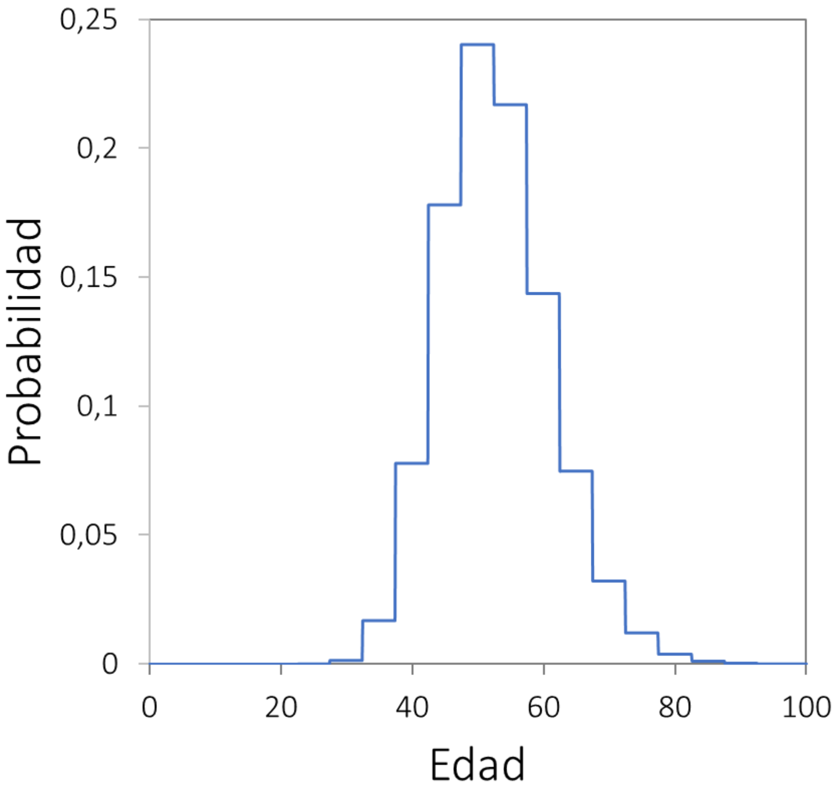
Cada dato x_i está compuesto de variables k, por lo tanto corresponde a un vector de información x_{ik}

$$x_i \rightarrow persona = \begin{pmatrix} sexo \\ edad \\ ingreso \\ \dots \end{pmatrix} \leftarrow x_{ik}$$

Distribuciones estadísticas

Como las variables observadas no son constantes (de lo contrario no serían útiles para entender el fenómeno de interés), la información proviene de distribuciones de probabilidad





Tipos de datos

Las variables pueden de dos tipos: cuantitativas o cualitativas

Las cuantitativas representan datos numéricos (discretos o continuos)

Ejemplos: edad de una persona, población de un país, ingreso monetario de un individuo, cantidad de ventas en un período, etc.

Tipos de datos

Las variables pueden de dos tipos: cuantitativas o cualitativas

Las cualitativas representan datos categóricos

Ejemplos: sexo de una persona, comuna de la ciudad, rango de ingreso de un individuo, mes del año, etc.

Tipos de datos

Varios de los enfoques de minería de datos que veremos en el curso requieren datos cuantitativos

Podemos obtener variables cuantitativas a partir de las cualitativas

Esto requiere transformar una variable cualitativa con N categorías en N variables binarias

Tipos de datos

Ejemplo: mes del año

Variable original X = {Enero, Febrero, Marzo,..., Diciembre}

Variables binarias

$X_{Enero} = \{0,1\}$

$X_{Febrero} = \{0,1\}$

…

$X_{Dicembre} = \{0,1\}$

Descripción de Datos

Descripción de datos

El objetivo de un análisis descriptivo de los datos es obtener una visión de algunas características generales

Es útil para el análisis inicial de la información disponible, para identificar potenciales problemas con los datos y para plantear relaciones causales preliminares que permitan entender y predecir el fenómeno de interés

Utilizaremos algunas medidas estadísticas descriptivas

Algunas medidas descriptivas

Media o promedio

$$\overline{X} = \frac{\sum_i x_i}{n}$$

$$X = \{8\,;4\,;6\,;6\,;7\,;1\,;1\,;6\,;3\,;5\}$$

$$\overline{X} = \frac{8+4+6+\ldots+5}{10} = 4{,}7$$

Algunas medidas descriptivas

Moda

Valor que más se repite dentro de los datos

$$X = \{8\,;4\,;6\,;6\,;7\,;1\,;1\,;6\,;3\,;5\}$$

Moda = 6

Algunas medidas descriptivas

Mediana

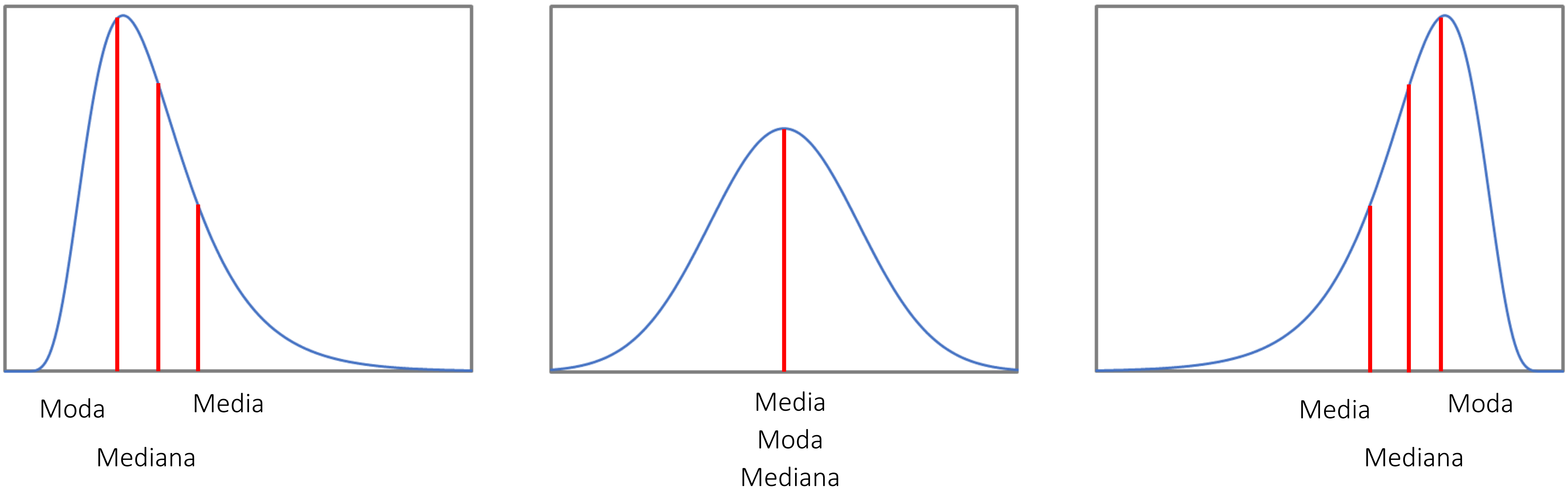
Valor donde se encuentra el 50% de los datos

$$X = \{8\,;4\,;6\,;6\,;7\,;1\,;1\,;6\,;3\,;5\}$$

$$X = \{1\,;1\,;3\,;4\,;5\,;6\,;6\,;6\,;7\,;8\}$$

Mediana = 5,5

Algunas medidas descriptivas



Algunas medidas descriptivas

Percentiles

Valor donde se encuentra el cierto % de los datos

$$X = \{8\,;4\,;6\,;6\,;7\,;1\,;1\,;6\,;3\,;5\}$$

$$X = \{1\,;1\,;3\,;4\,;5\,;6\,;6\,;6\,;7\,;8\}$$

Percentil 20% = 1 Percentil 60% = 6 Percentil 90% = 7

Algunas medidas descriptivas

Varianza y desviación estándar

Medidas de dispersión de los datos

$$s_x^2 = \frac{\sum_i \left(x_i - \overline{X}\right)^2}{n}$$

$$s_x = \sqrt{\frac{\sum_i \left(x_i - \overline{X}\right)^2}{n}}$$

Algunas medidas descriptivas

Varianza y desviación estándar

$$X = \{8\,;4\,;6\,;6\,;7\,;1\,;1\,;6\,;3\,;5\}$$

$$s_x^2 = \frac{(8-4{,}7)^2 + (4-4{,}7)^2 + \ldots + (5-4{,}7)^2}{10} = 5{,}21$$

$$s_x = \sqrt{5{,}21} = 2{,}28$$

Algunas medidas descriptivas

Covarianza y correlación

Medidas de variación conjunta de las variables

$$s_{xy} = \frac{\sum_i \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{n}$$

$$\rho_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Algunas medidas descriptivas

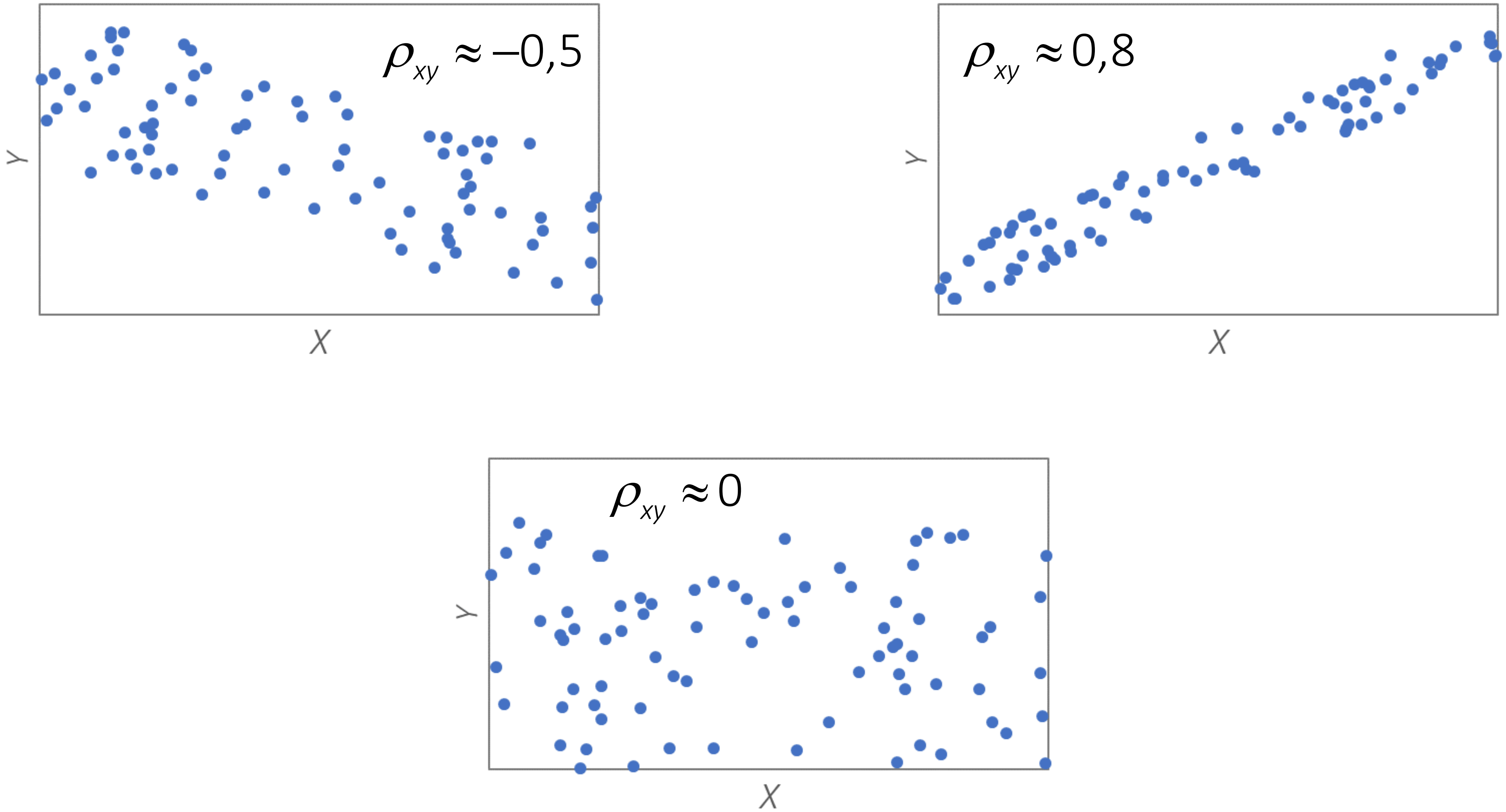
Covarianza y correlación

$$X = \{8\,;4\,;6\,;6\,;7\,;1\,;1\,;6\,;3\,;5\} \qquad Y = \{7\,;3\,;4\,;5\,;1\,;2\,;1\,;7\,;2\,;5\}$$

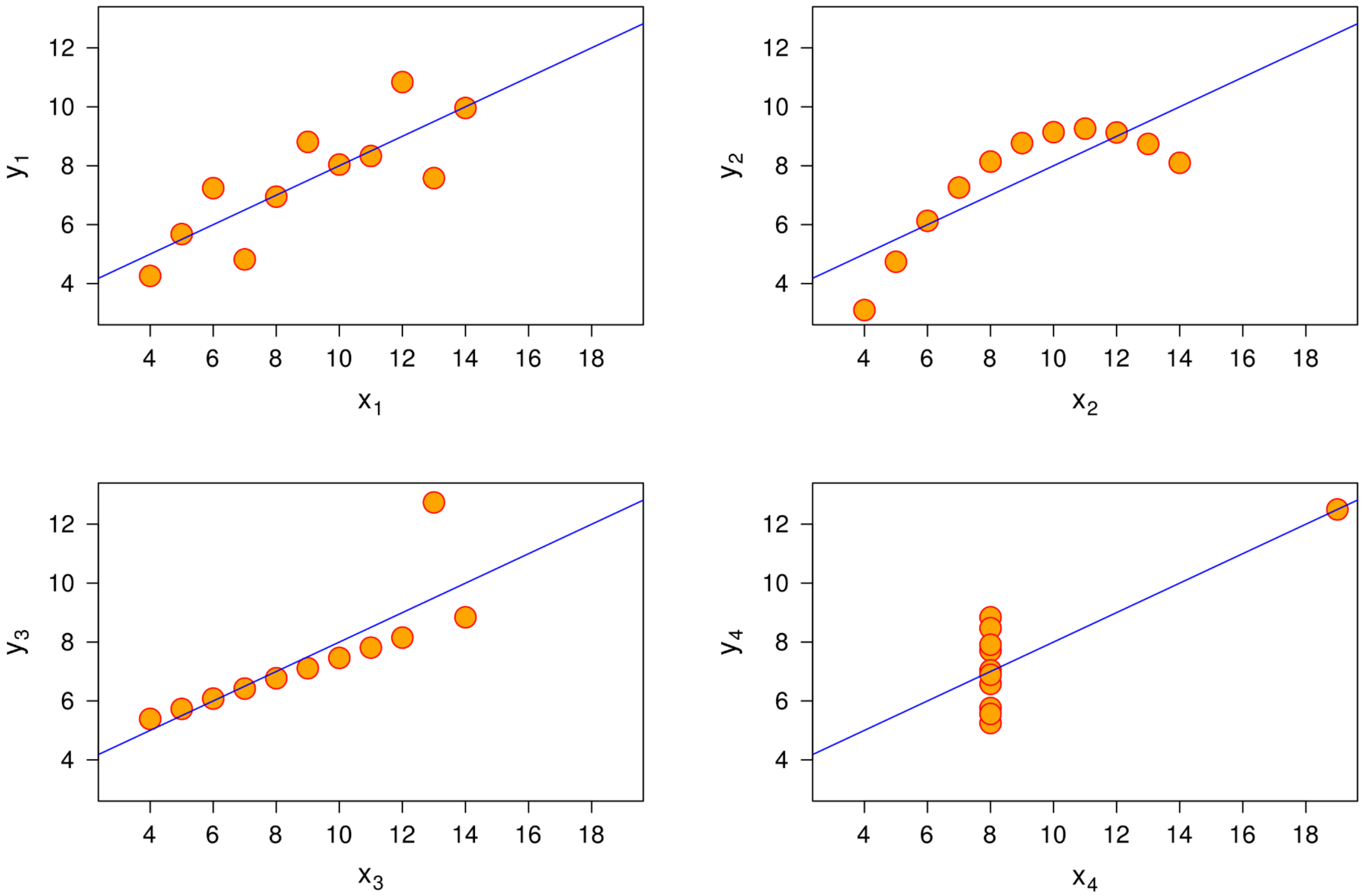
$$s_{xy} = \frac{(8-4{,}7)(7-3{,}7)+(4-4{,}7)(3-3{,}7)+\cdots+(5-4{,}7)(5-3{,}7)}{10} = 3{,}11$$

$$\rho_{xy} = \frac{3{,}11}{2{,}28 \cdot 2{,}15} = 0{,}63$$

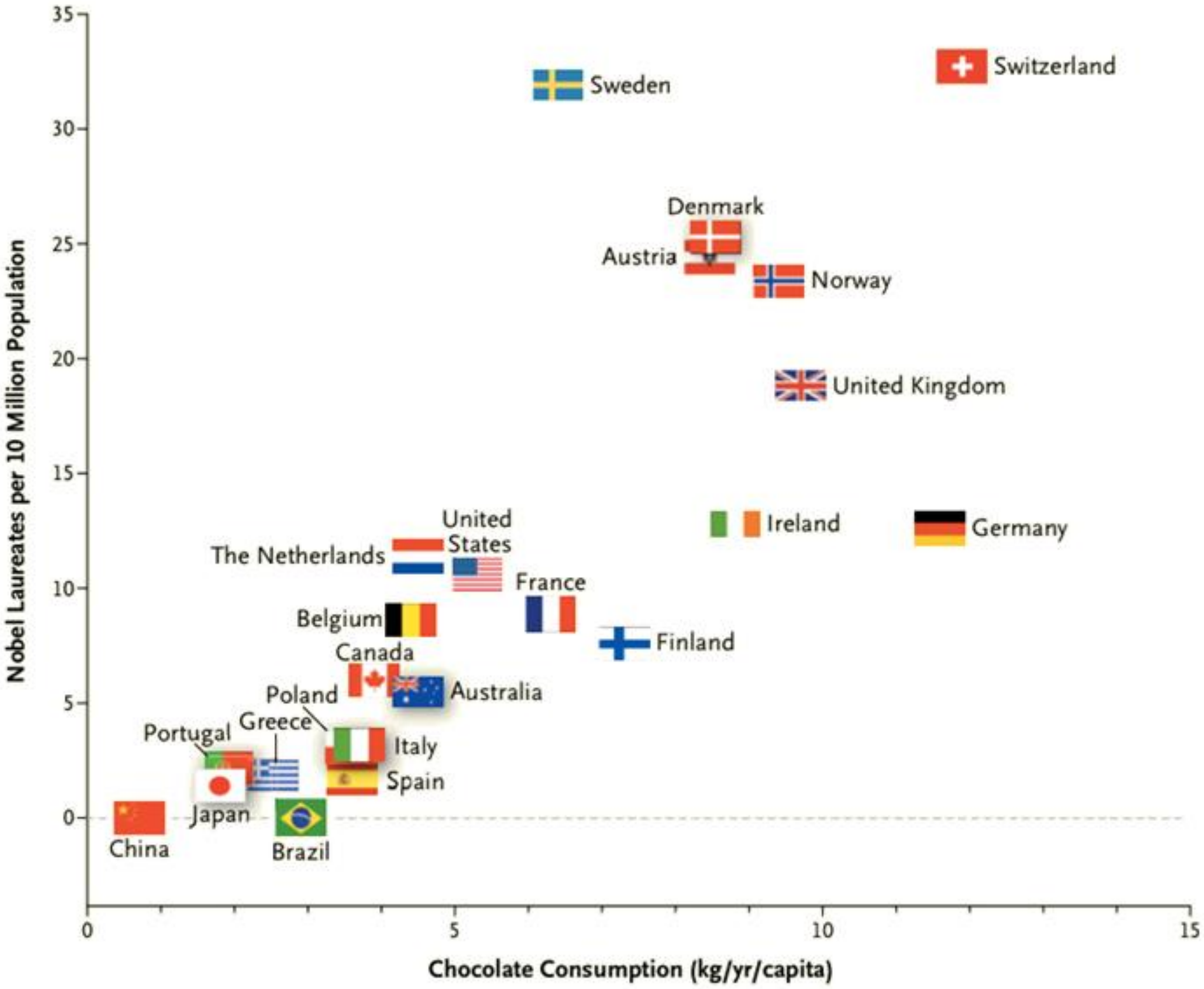
Gráficos de dispersión



Cuarteto de Anscombe



Relaciones espurias



http://tylervigen.com/spurious-correlations


SOLÍA CREER QUE LA CORRELACIÓN IMPLICABA CAUSALIDAD.
LUEGO DI UNA ASIGNATURA DE ESTADÍSTICA Y DEJÉ DE CREERLO.
PARECE QUE ESA CLASE TE AYUDÓ.
BUENO, QUIZÁ.

Detección de *outliers*

¿Qué son los *outliers*?

Los *outliers* son datos atípicos que al parecer han sido generados de manera distinta al resto de los datos

Pueden ser causados por ejemplo por errores de medición o digitación de los datos, cambios en los instrumentos de medición

También simplemente pueden representan una heterogeneidad intrínseca del fenómeno estudiado

¿Qué son los *outliers*?

La caracterización de un dato *outlier* en muchos casos puede ser sencilla, ya que por definición debe estar alejado del resto de los datos

Podemos calcular la "distancia" de cada dato a la media:

$$d\left(x_i,\overline{X}\right)=\sqrt{\left(x_i-\overline{X}\right)^T\left(x_i-\overline{X}\right)}$$

¿Qué son los *outliers*?

Ejemplo, un dato con:	Ingreso	= 1.800.000	(media = 860.000)
	Sexo	= 1	(media = 0,51)
	Edad	= 45	(media = 54,6)

$$d = \sqrt{(1.8000.000 - 860.000)^2 + (1 - 0,51)^2 + (45 - 54,6)^2}$$

$$d = 940,000$$

¿Qué son los *outliers*?

Como las variables pueden tener escalas distintas, podemos estandarizarlas:

$$z_{ik} = \frac{x_{ik} - \bar{X}_k}{s_k}$$

Todas las variables z_{ik} tendrán media 0 y desviación estándar 1

¿Qué son los *outliers*?

Otra alternativa es normalizarlas:

$$z_{ik} = \frac{x_{ik} - \min_{\ddot{i}}\left\{x_{ik}\right\}}{\max_{\ddot{i}}\left\{x_{ik}\right\} - \min_{\ddot{i}}\left\{x_{ik}\right\}}$$

Todas las variables z_{ik} estarán en el rango entre 0 y 1

¿Qué son los *outliers*?

Ejemplo, un dato con:

Ingreso estandarizado = 0,28

Sexo estandarizado = 0,98

Edad estandarizada = -0,48

$$d = \sqrt{0{,}28^2 + 0{,}98^2 + \left(-0{,}48\right)^2}$$

$$d = 1{,}13$$

Efecto de *outliers*

En algunos casos, incluir datos *outlier* puede tener consecuencias graves:

distorsionar las medias y desviaciones estándar

sesgar los resultados del análisis

destruir relaciones existentes entre variables

Lamentablemente, no es fácil distinguir cuándo un dato “distinto” es un *outlier* y cuando simplemente se debe a la variabilidad del fenómeno

Detección de *outliers*

Podemos utilizar métodos estadísticos para detectar potenciales *outliers*

Método 1 - Percentiles

Método 2 - Intervalos de variabilidad

Método 3 - Valor-z robusto

Método 1 - Percentiles

Una de las prácticas más comunes es detectar potenciales *outliers* usando percentiles

Por ejemplo, podemos considerar el 1% superior y 1% inferior como potenciales *outliers*, y quedarnos con el 98% restante de los datos

El percentil a utilizar es arbitrario y no tiene que ser necesariamente simétrico

La cantidad de *outliers* dependerá del tamaño de la base de datos

Método 2 - Intervalos de variabilidad

Los *outliers* serán aquellos que se alejen de la tendencia general de los datos

Para esto podemos construir un intervalo centrado en la media, en función de la desviación estándar de los datos

$$\bar{X} \pm \delta \cdot s$$

Todo dato que no pertenezca a este intervalo será un potencial *outlier*

Método 2 - Intervalos de variabilidad

El valor de δ determinará cuántos datos pertenecen al intervalo

Es posible demostrar que el intervalo contendrá al menos una proporción $1 - \frac{1}{\delta^2}$ de los datos ("desigualdad de Chebyshev")

$\delta = 2$ → El intervalo contendrá al menos el 75,0% de los datos

$\delta = 3$ → El intervalo contendrá al menos el 88,9% de los datos

$\delta = 4$ → El intervalo contendrá al menos el 93,8% de los datos

Método 3 - Valor-*z* robusto

A diferencia de los métodos anteriores, el resultado de este método no depende de:

- La cantidad de datos
- Medidas que se ven afectadas por los *outliers*

Analizamos la desviación de los datos con respecto a la mediana

Método 3 - Valor-*z* robusto

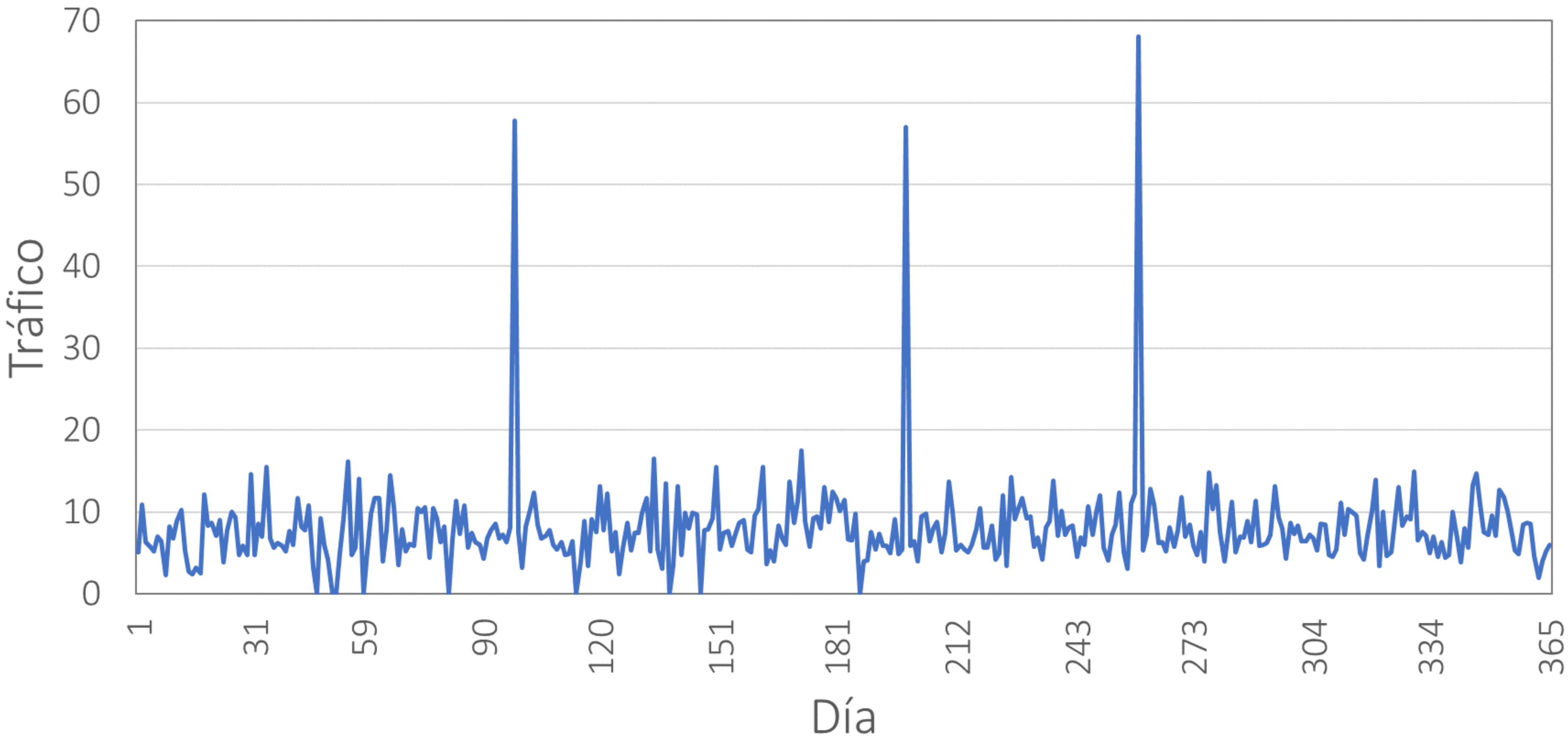
Una regla simple para detectar *outliers* es:

$$\frac{\left|x_i - mediana\left(x_i\right)\right|}{MEDA\left(x_i\right)} > \Delta$$

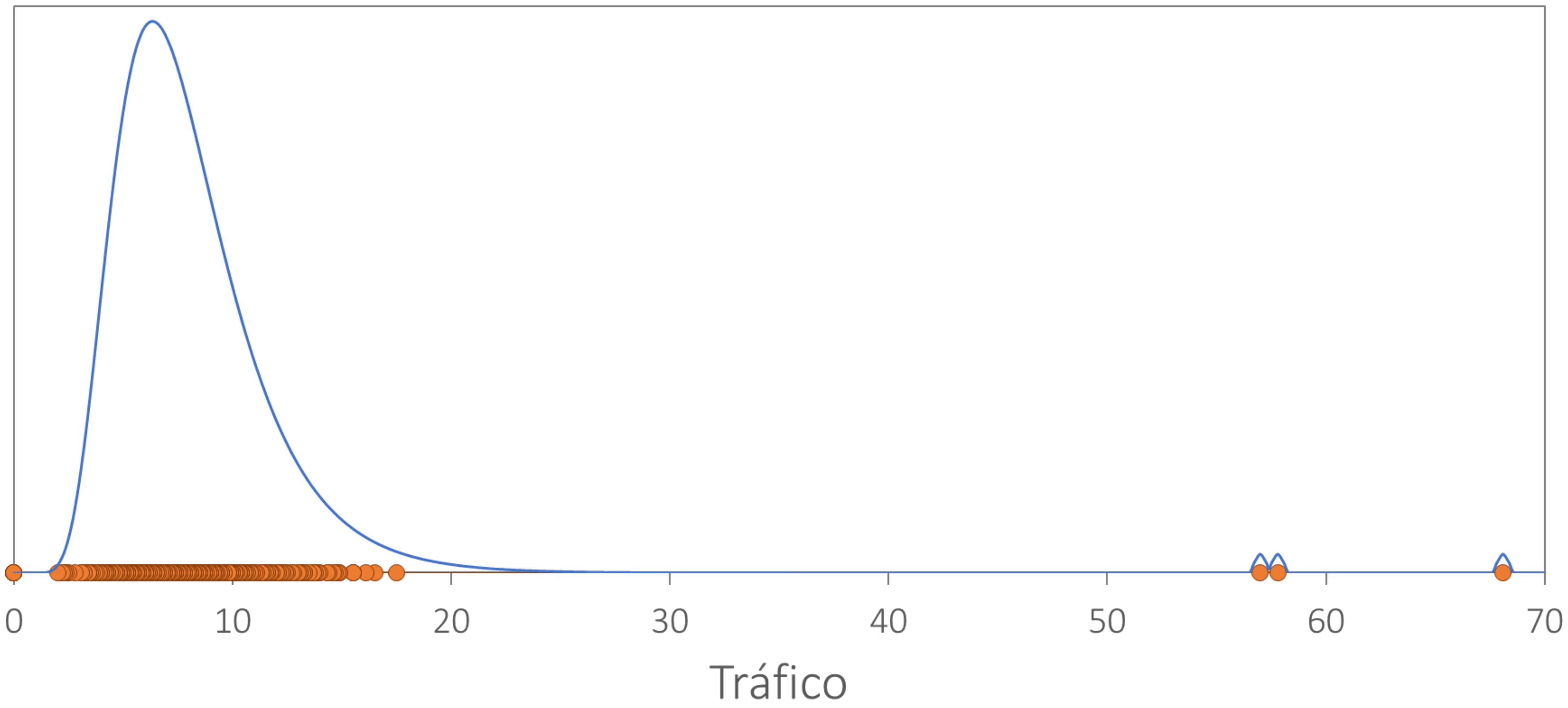
MEDA es la mediana de las desviaciones absolutas, es decir la mediana de los valores $\left|x_i - mediana\left(x_i\right)\right|$

El valor de *Δ* lo determina el analista; en general ronda en torno a 4,5

Veamos un ejemplo



Distribución de los datos



Aplicando percentiles

Seleccionamos el 1% inferior y 1% superior de los datos

Tenemos 365 datos, eso implica seleccionar los 3 ó 4 menores y mayores valores

Aplicando percentiles

Aplicando intervalos de variabilidad

El promedio de los datos es 8,0

La desviación estándar de los datos es 5,8

Identificamos los datos en el intervalo $8{,}0 \pm 3 \cdot 5{,}8$

Estos datos están en el intervalo $-9{,}4 < x_i < 25{,}4$

Aplicando intervalos de variabilidad

Aplicando intervalos de variabilidad

El promedio sin estos tres datos es 7,6

La desviación estándar sin estos tres datos es 3,1

Identificamos los datos en el intervalo $7{,}6 \pm 3 \cdot 3{,}1$

Estos datos están en el intervalo $-1{,}7 < x_i < 16{,}9$

Aplicando intervalos de variabilidad

Aplicando el valor-*z* robusto

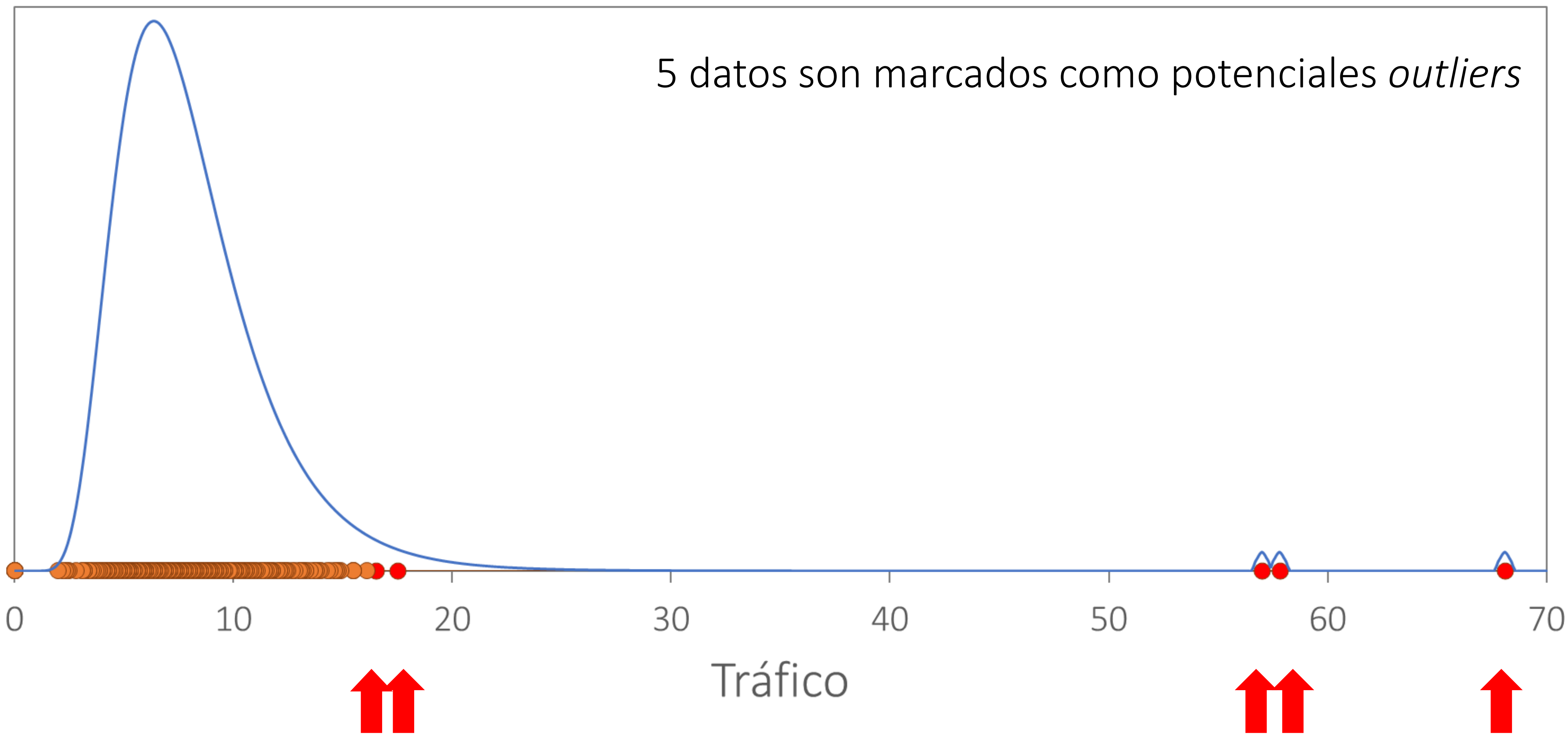
La mediana de los datos es 7,3

La mediana de las desviaciones absolutas (MEDA) es 2,0

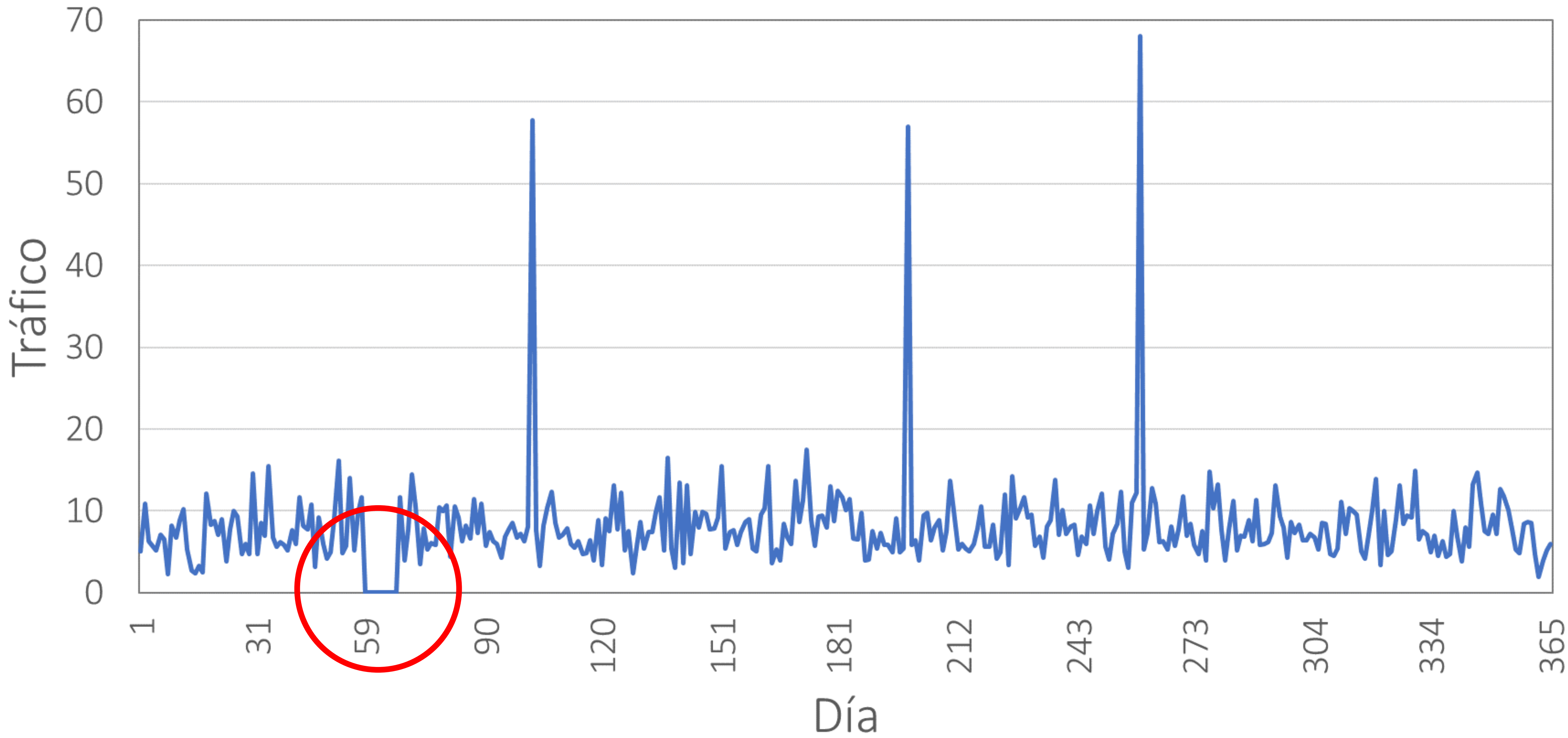
Identificamos los datos donde $\dfrac{|x_i - 7{,}3|}{2{,}0} > 4{,}5$

Estos datos están en el intervalo $-1{,}7 < x_i < 16{,}3$

Aplicando el valor-z robusto



Analizar la serie de datos también puede entregar información



Detección de *outliers*

Los métodos de detección de outliers vistos en esta sesión se basan en propiedades estadísticas de los datos

percentil, media, mediana, desviación estándar, etc.

Más adelante en el curso veremos otros enfoques de minería de datos que pueden ser utilizados para detectar *outliers*

k-NN, *clustering*, etc.

Preprocesamiento y análisis de datos

"Nunca hay que usar la estadística de la misma forma que los borrachos usan los postes de luz: no para iluminarse, sino que para disimular su inestabilidad"

A. E. Housman (1903)