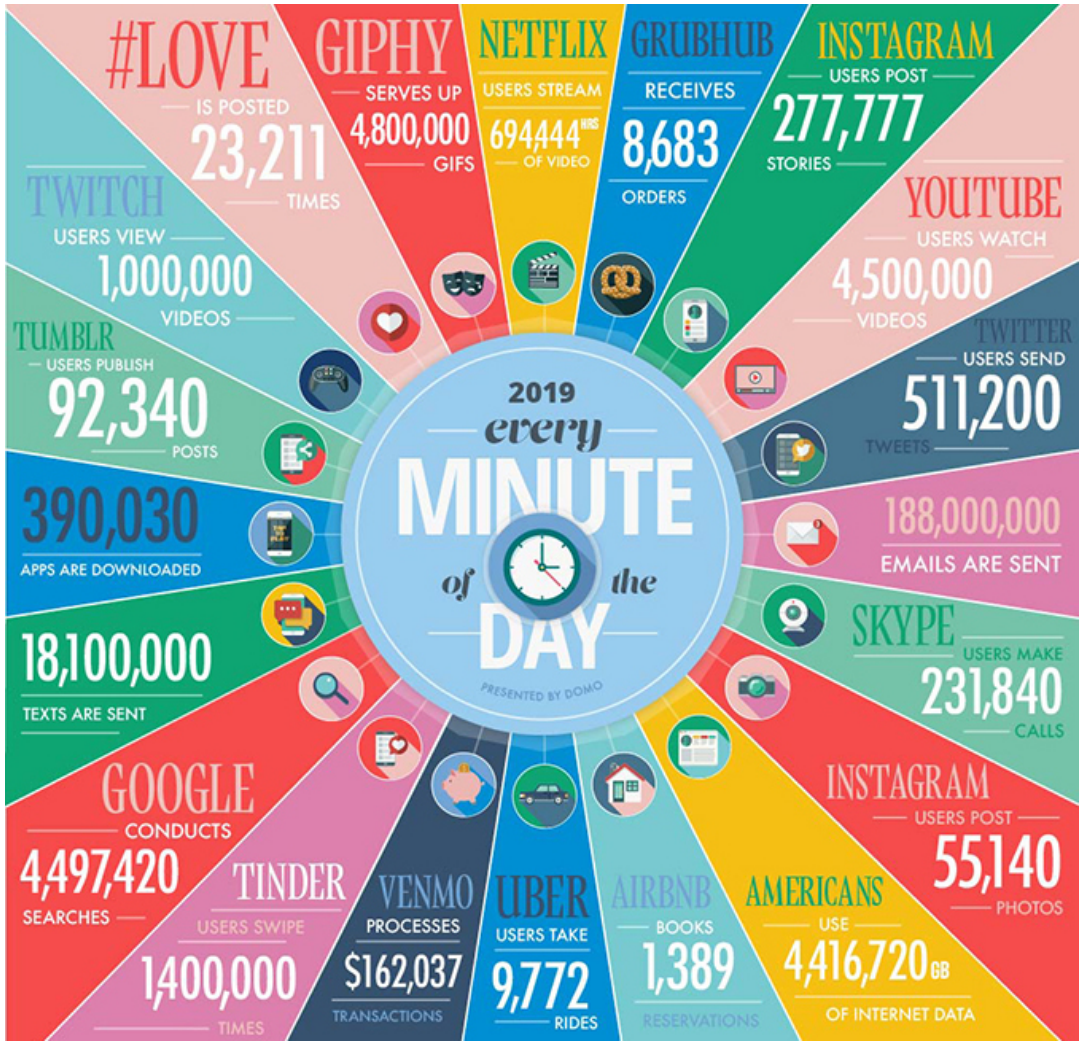
Introducción a la estadística descriptiva y análisis exploratorio de datos

Curso de análisis de datos con R
Asociación Argentina de Bioinformática y Biología Computacional
Fundación Instituto Leloir

Marzo 2021

La era de la información (¿o es de los datos?)

Se dice que "vivimos en la Era de la Información", sin embargo, una frase más adecuada debería ser que "vivimos en la Era de los Datos".



https://www.domo.com/learn/data-never-sleeps-7

Secuencias de Covid-19

Empecemos a trabajar con algunos datos. Tomemos secuencias de Covid-19 extraídas de distintos pacientes:
AAACGGTGGTTGAAGGTGGTTGATT...CATGATGTGGTTGAAAG
Y contemos cuantas A, C, G y T tiene cada una.

Cepa del virus	a	c	g	t	ubicación
2019-nCoV/USA-WA1	8913	5488	5860	9586	United States / WA
SARS-CoV-2/HZ-62	8894	5471	5850	9565	China / Zhejiang / Hangzhou
hCoV-19/USA/ID-UW260	8894	5472	5852	9572	United States / Idaho
SARS-CoV-2/Wuhan_IME-WH01	8918	5473	5852	9564	China / Wuha
SARS-CoV-2/WA-UW370	8909	5482	5859	9589	United States / WA
SARS-CoV-2/WA6-UW3	8949	5489	5866	9598	United States / Washington
SARS-CoV-2/Wuhan_YB012602	8923	5492	5860	9589	China / Hubei / Wuhan
SARS-CoV-2/WA-UW298	8956	5488	5863	9596	United States / WA
SARS-CoV-2/Hu/	8932	5492	5860	9590	Japan
2019-nCoV/Japan	8931	5490	5862	9595	Japan / Tokyo

Viendo esta tabla ¿Podemos comprender el contenido de las secuencias? ¿Podemos saber qué letra aparece más? ¿Qué cepa tiene mayor contenido de C que las otras? ¿Cuál es la cantidad más frecuente de cada una de las bases?

Entender nuestros datos para poder sacarles el jugo

Los datos del mundo real son típicamente ruidosos, vienen de fuentes heterogéneas y en enormes cantidades.
Para poder comprender nuestros datos, vamos a querer describirlos lo más exhaustivamente posible y para eso nos va a interesar responder las siguientes preguntas:

- ¿Qué tipos de atributos componen nuestros datos?
- ¿Qué valores pueden tomar esos atributos?
- ¿Cómo se distribuyen?
- ¿Es posible visualizarlos de una forma interesante?
- ¿Existen valores atípicos?

¿Qué otras preguntas formularían para comprender mejor sus datos?

Pero primero ¿qué es un dato?

Un dataset está compuesto por objetos de datos u observaciones. Un objeto de datos representa una entidad (no necesariamente del mundo real).
Por ejemplo, un dato puede ser la secuencia de COVID-19 de un paciente, un ratón en un experimento o la posición de un aminoácido en la estructura de una proteína.
Una forma habitual de representar los datos es por medio de una tabla, donde cada fila corresponde a un objeto de datos y cada columna es un atributo que compone ese dato.

Cepa del virus	a	c	g	t	ubicación
2019-nCoV/USA-WA1	8913	5488	5860	9586	United States / WA
SARS-CoV-2/HZ-62	8894	5471	5850	9565	China / Zhejiang / Hangzhou
hCoV-19/USA/ID-UW260	8894	5472	5852	9572	United States / Idaho
SARS-CoV-2/Wuhan_IME-WH01	8918	5473	5852	9564	China / Wuha
SARS-CoV-2/WA-UW370	8909	5482	5859	9589	United States / WA
SARS-CoV-2/WA6-UW3	8949	5489	5866	9598	United States / Washington
SARS-CoV-2/Wuhan_YB012602	8923	5492	5860	9589	China / Hubei / Wuhan
SARS-CoV-2/WA-UW298	8956	5488	5863	9596	United States / WA
SARS-CoV-2/Hu/	8932	5492	5860	9590	Japan
2019-nCoV/Japan	8931	5490	5862	9595	Japan / Tokyo

¿Qué es qué en esta tabla?

Tipos de atributos

Usualmente se utilizan los términos atributo, variable, dimensión o feature de forma intercambiable.
Cada atributo puede ser de tipo:

- **Nominal:** son signos o nombres de cosas. Por ejemplo China, Japón, Argentina o Negro, Blanco, Rojo. Notar que estos atributos no tienen orden.
- **Binario:** 0 o 1 o Verdadero y Falso.
- **Ordinal:** signos o nombres de cosas con orden. Por ejemplo pequeño, mediano, grande o Investigador Asistente, Investigador Adjunto, Investigador Independiente, Investigador Principal, Investigador Superior. Notar que si bien tienen orden, no es posible indicar la magnitud de la diferencia entre dos valores.
- **Numérico:** representan cantidades. Pueden ser intervalos, por ejemplo, entre 5 y 10, entre 10 y 15, entre 15 y 20, discretos, por ejemplo, la edado o continuos, por ejemplo, la altura de una persona.

Tipos de atributos

¿Qué tipo de atributo es cada uno en la tabla?

Cepa del virus	tamaño (nm)	n muestras	Gram	ubicación
nCoV/USA-WA1	125.5	5860	Si	United States / WA
nCoV/JAP	153.3	8460	Si	Japon
nCoV/CHINA	135.1	8620	No	China / Wu
nCoV/WA1	160.3	460	Si	Japon
nCoV/USA-WA1	121.7	520	No	United States / WA

Veamos cómo podemos hacer todo esto con R

Abrimos dia_3.R

Describiendo los datos desde la estadística

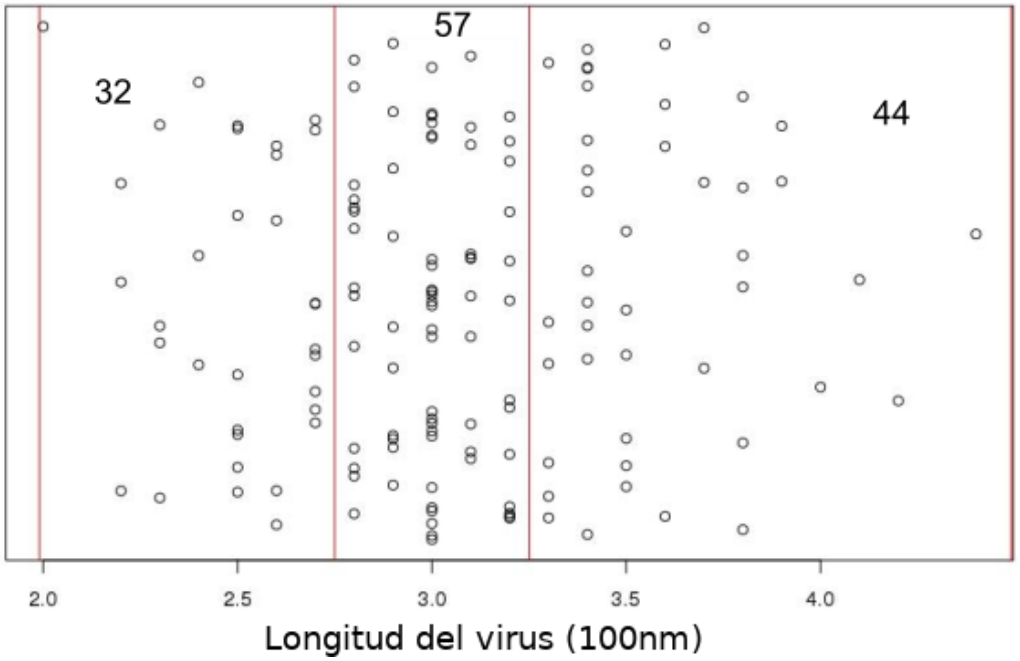
Un primer acercamiento a la descripción de los datos lo da la estadística descriptiva. La misma es útil para identificar propiedades de los datos y resaltar aquellos valores ruidosos o atípicos. Para ello utilizaremos:

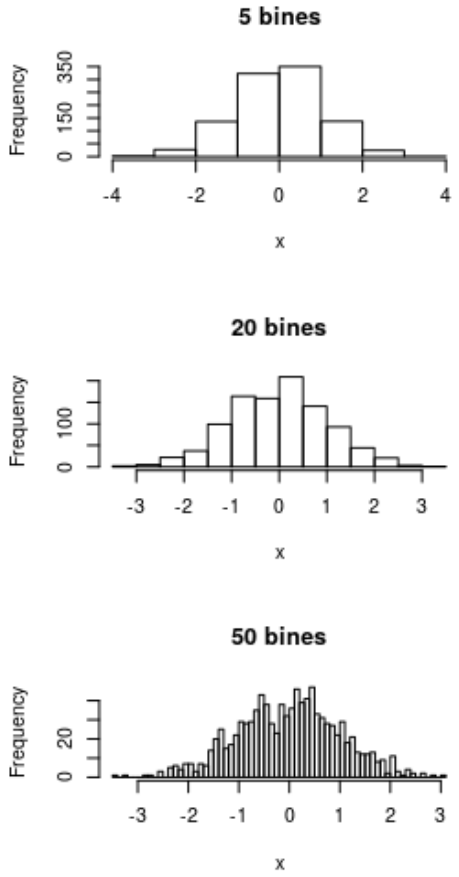
- Medidas de tendencia central que indican dónde caen los valores típicos: media, mediana y moda.
- Medidas de dispersión que indican cuán separados están los valores: varianza, desvío estándar y rango intercuantil.
- Gráficos que nos permitan evaluar visualmente nuestros datos: histogramas, gráficos de dispersión, gráficos de caja o boxplots, gráficos de coordenadas paralelas, entre otros.
- Algunas otras medidas que no sirven para variables numéricas, por ejemplo, para variables categóricas, podemos contar cuantos elementos tiene cada categoría.

¿Cómo se distribuyen nuestros datos?

Una forma de visualizar cómo son nuestros datos es usando un histograma.

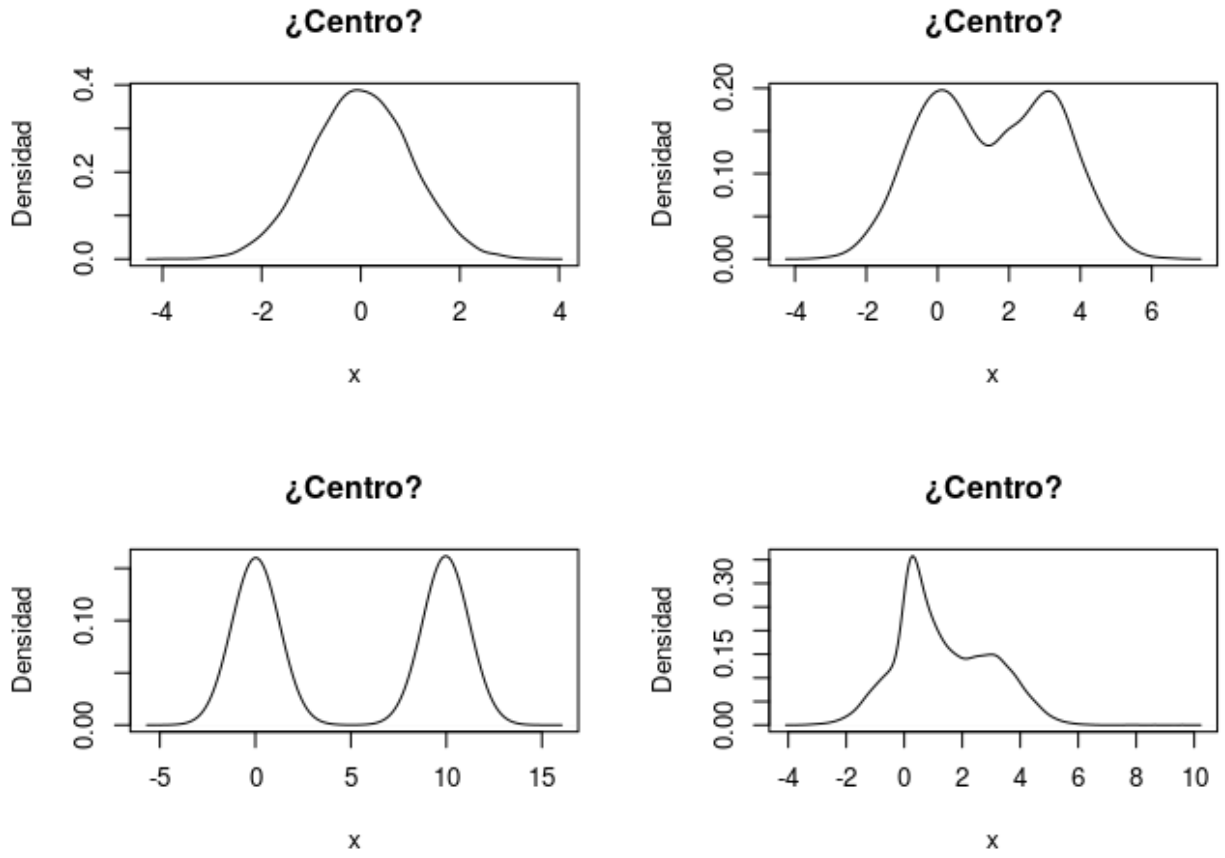
Para construir un histograma se divide el rango de las variables en intervalos y se cuenta la cantidad de datos en cada intervalo, es decir, la frecuencia.





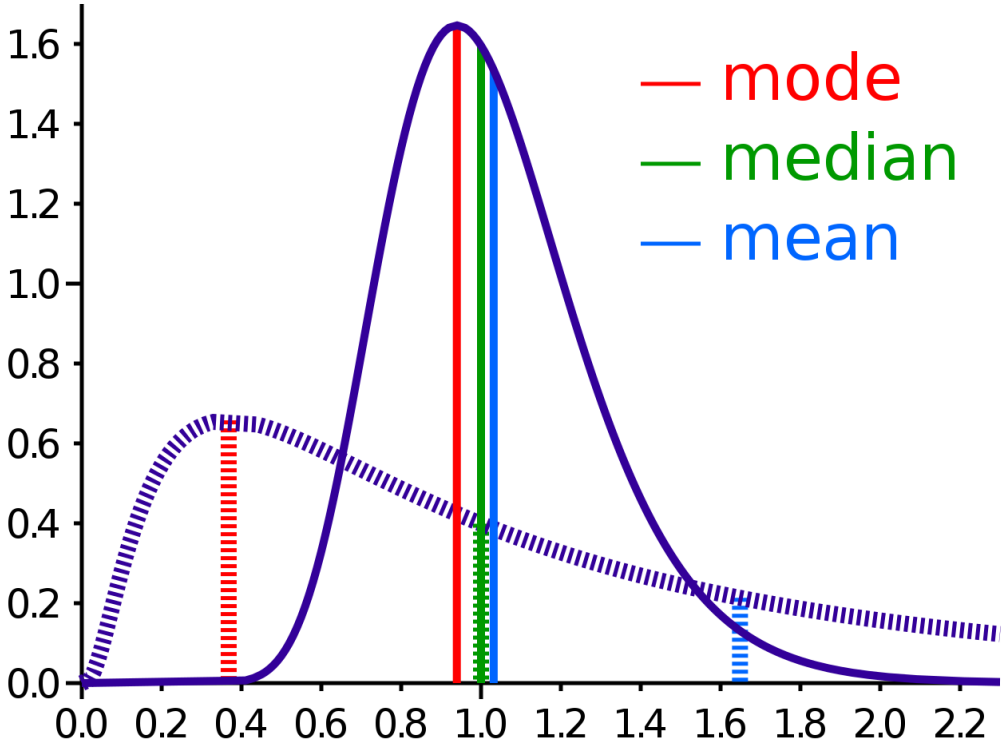
¿Un número que los represente a todos?

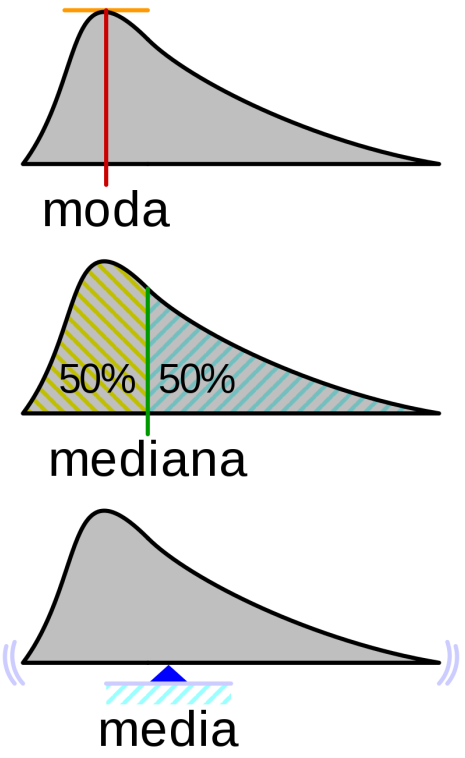
Una forma de resumir un conjunto de valores es usando un número que "represente a todos", un valor típico para ese conjunto. Nos preguntamos entonces, ¿cuál es el valor central o más representativo de esos datos?



Media, mediana y moda

Existen distintas formas de definir el centro o el número que "representa a todos".

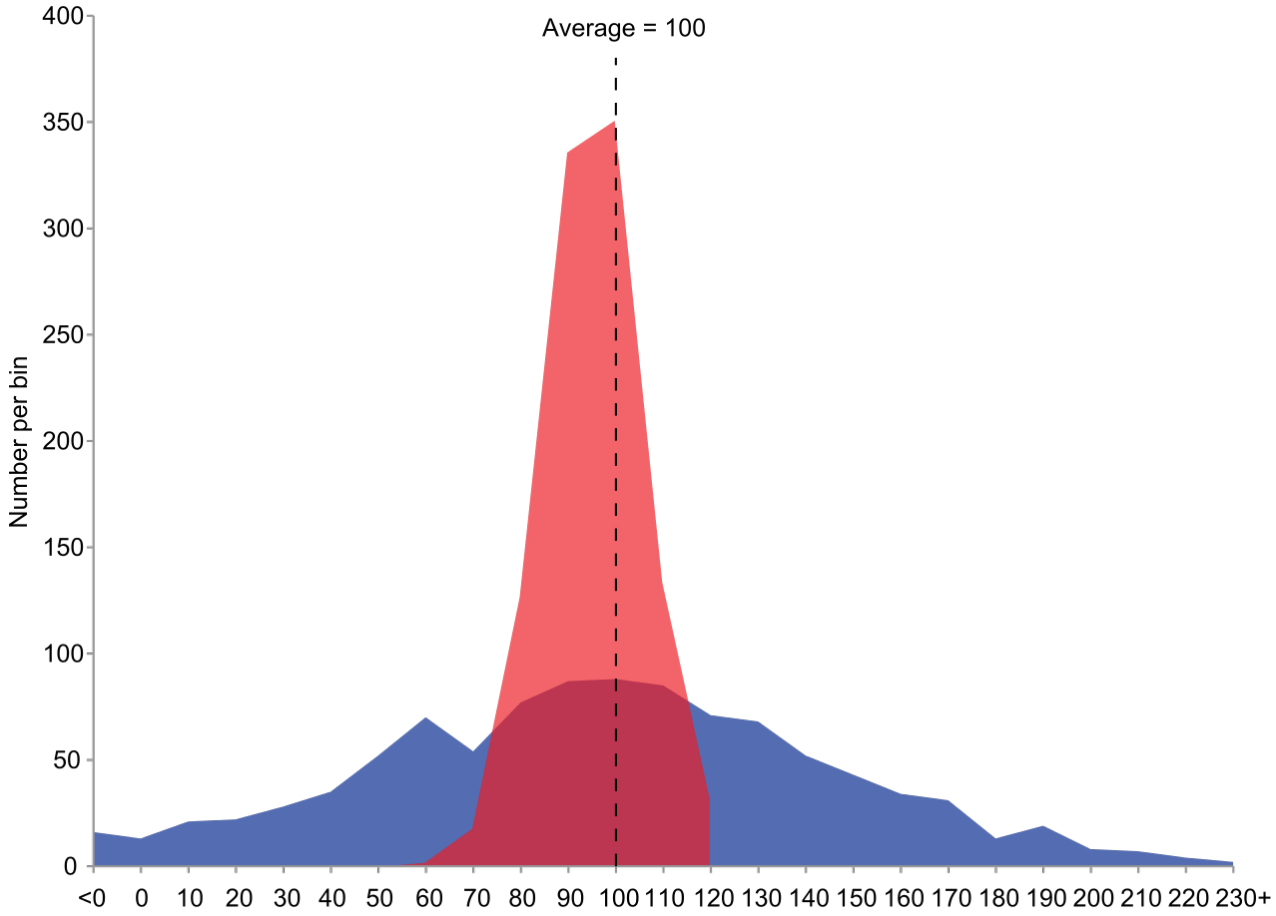




¿Son estas medidas mutuamente excluyentes o cada una nos brinda información diferente?

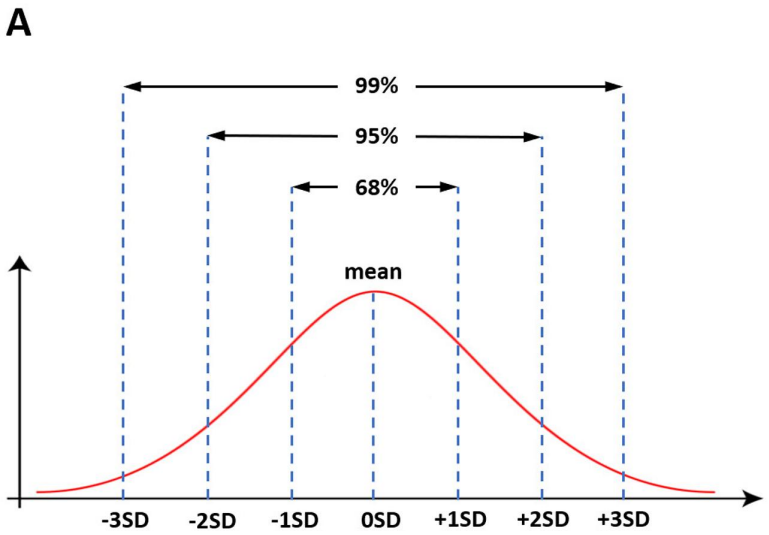
¿Alcanza con esto?

Teniendo una medida de tendencia central, ¿Ya podemos describir nuestra distribución?

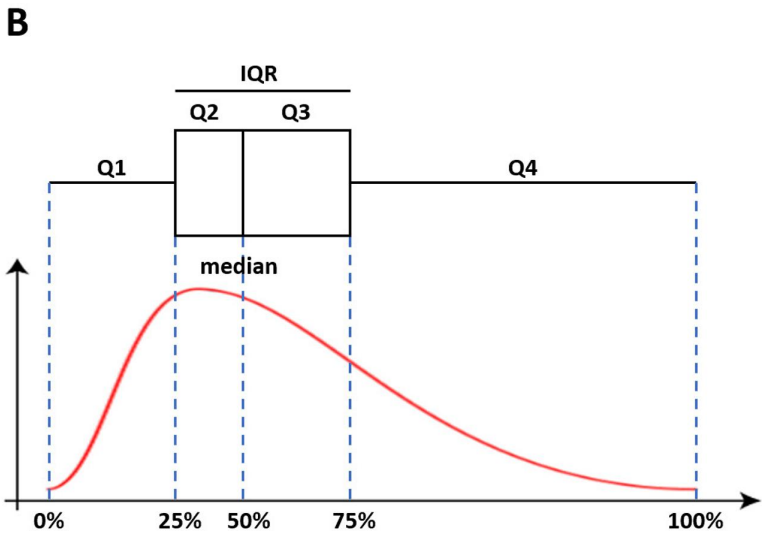


Medidas de dispersión

Otras preguntas que nos formulamos respecto a los datos es ¿cuán dispersos están los datos? ¿cuán cercanos son los datos al valor típico?



A: Media y desvío estandar.



B: Mediana y rango intercuartil

* Operating with Data - Statistics for the Cardiovascular Surgeon: Part I. Fundamentals of Biostatistics, Romero Liguori, Pinho Moreira

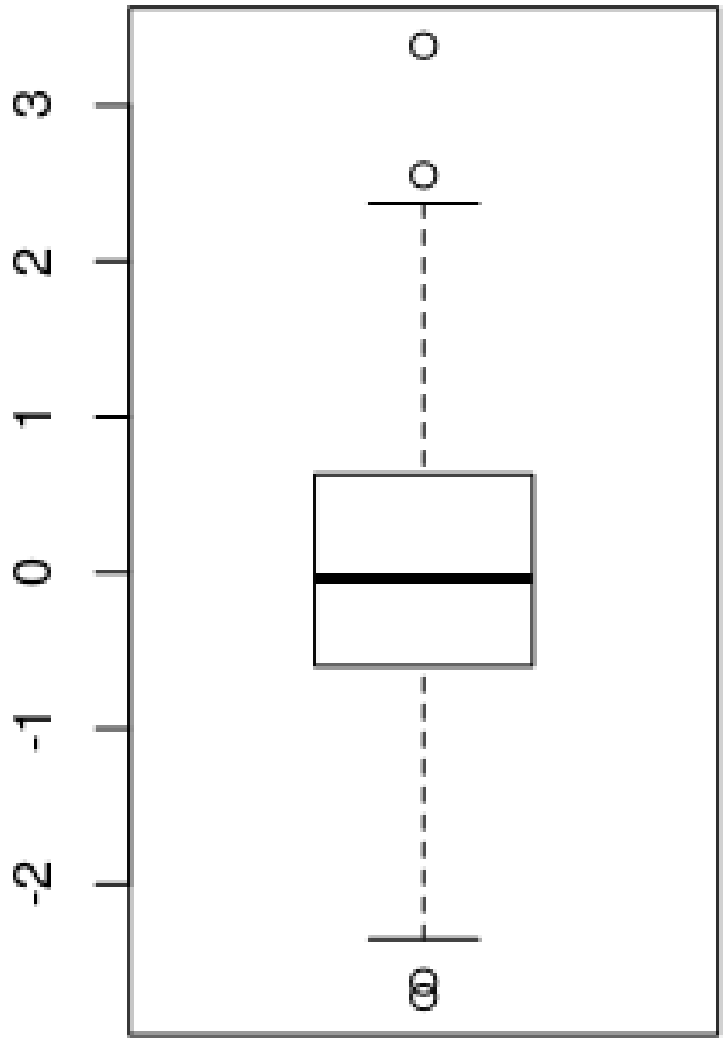
Veamos cómo hacer esto en R

Abrimos dia_3.R

Boxplot o diagrama de cajas

En un boxplot podemos visualizar varias medidas de dispersión en simultaneo y darnos una mejor idea de la distribución de nuestros datos.

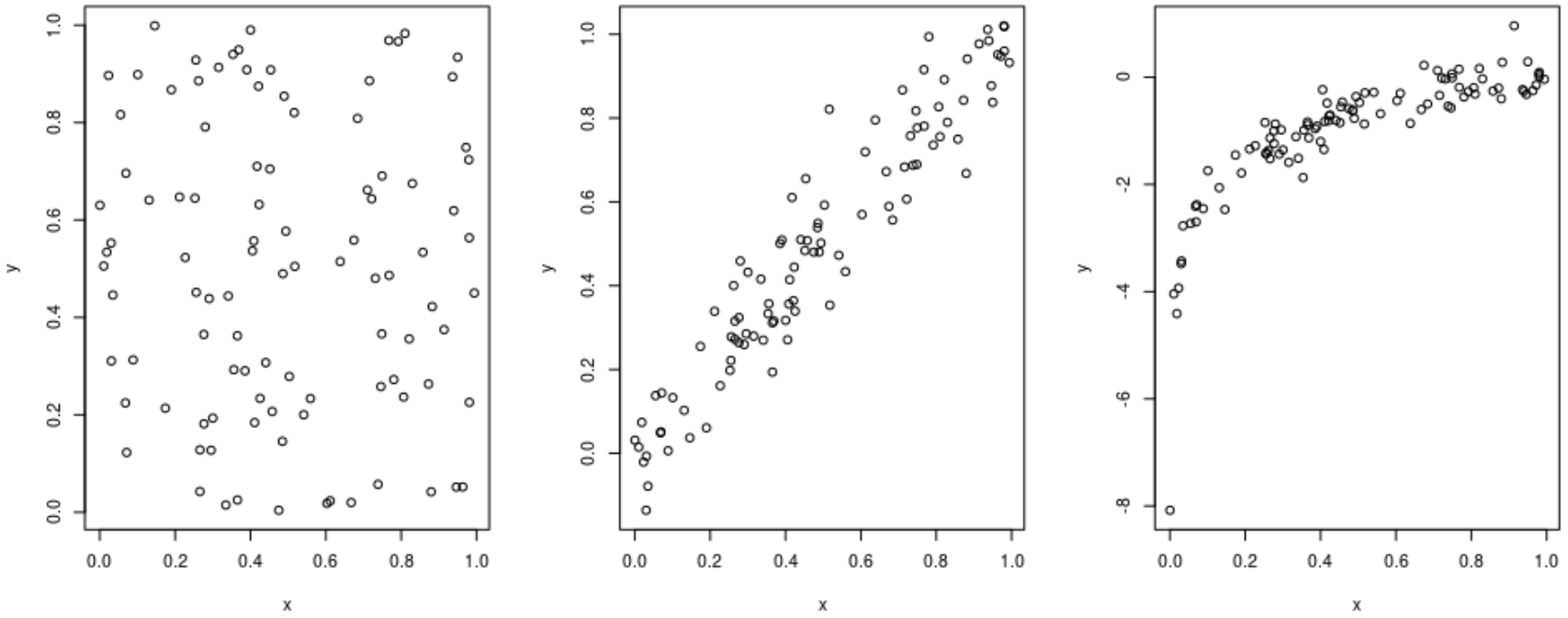
- La caja contiene el rango desde el primer cuartil hasta el tercer cuartil.
- La barra horizontal dentro de la caja indica la mediana.
- El bigote sale desde el 1er cuartil (3er cuartil) hasta la mínima (máxima) observación siempre y cuando esta observación esté a menos de 1.5 veces la distancia inter cuartil.
- Si hay elementos a mayor distancia aparecen como “o”.



Scatterplot o gráfico de dispersión

Un gráfico de dispersión es una forma muy efectiva de observar si existe una relación, patrón o tendencia entre dos variables numéricas.

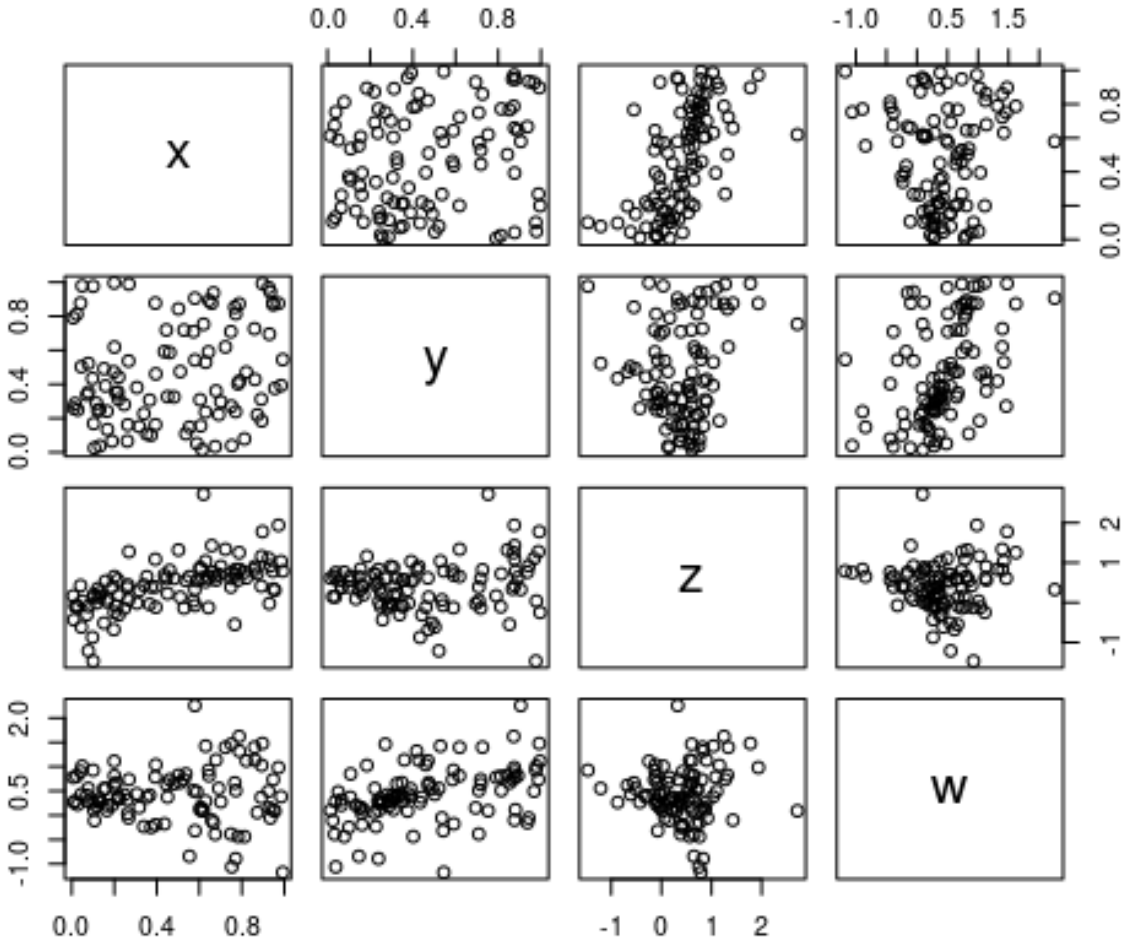
Para construir un scatterplot, cada par de valores se considera una coordenada en el plano.



¿Cómo se relacionan las variables x e y en cada caso?

Scatterplot o gráfico de dispersión de a pares

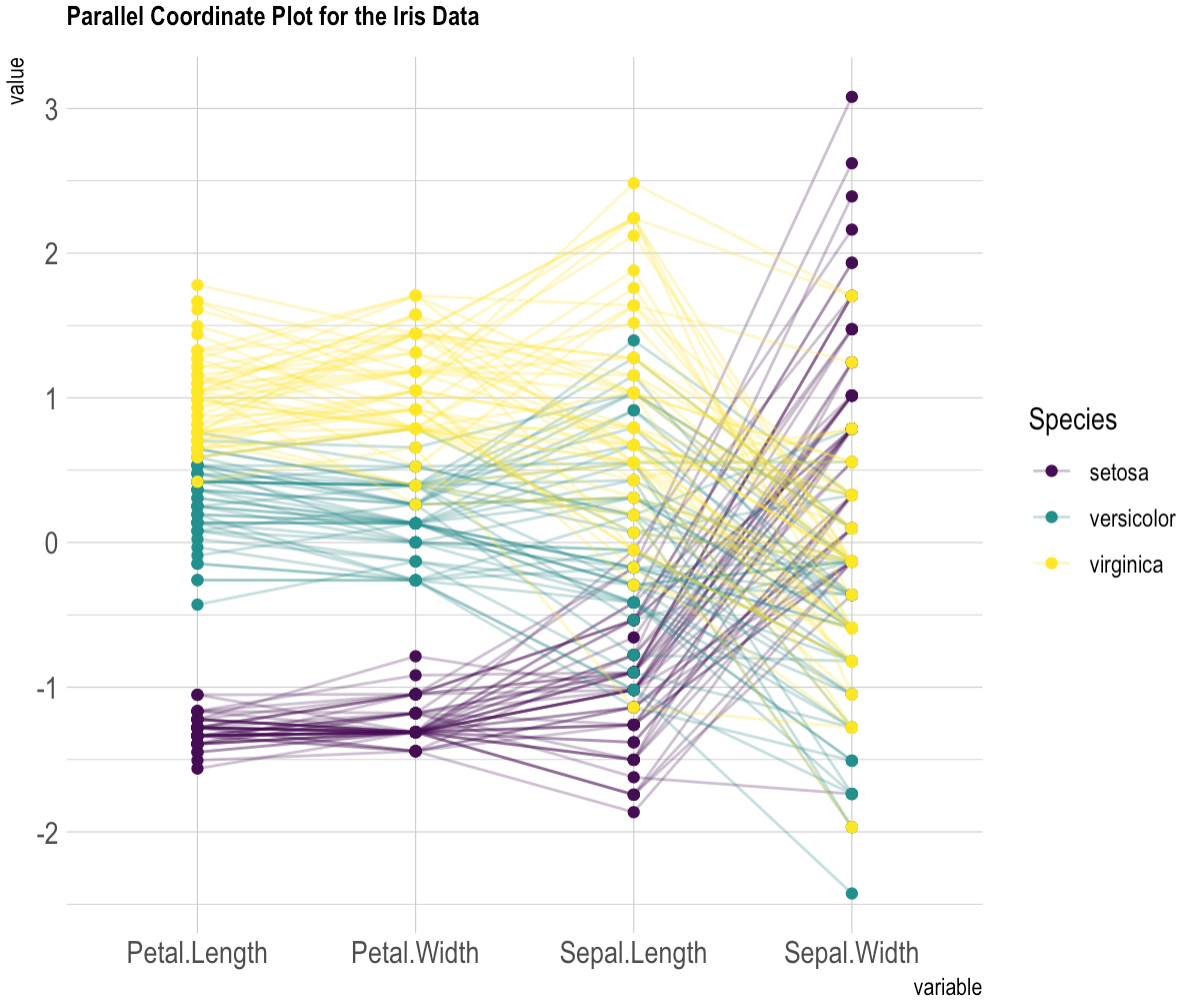
¿Y si tenemos muchas variables? Graficamos todas contra todas.



Matplot o gráfico de coordenadas paralelas

En caso de tener más de tres variables, ya no es posible graficarlas en un único scatterplot. Una opción, si todas tienen la misma escala, es realizar un gráfico de coordenadas paralelas.

En este tipo de gráficos, el eje "x" representa cada una de las variables que queremos graficar y el eje "y" su correspondiente valor.



https://www.r-graph-gallery.com/parallel-plot-ggally.html

Veamos cómo hacer esto en R

Abrimos dia_3.R