Laboratorio de Datos
Visualización

Primer Cuatrimestre 2024
Turnos tarde y noche

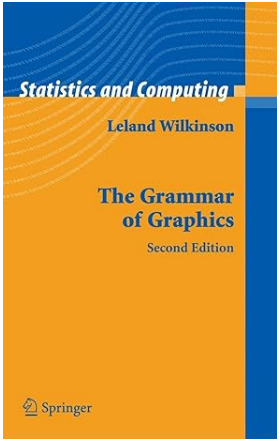
Facultad de Ciencias Exactas y Naturales, UBA

Gramática de Gráficos

La Gramática de Gráficos es un marco para pensar sobre la visualización desarrollado por Leland Wilkinson. Su idea central es que cualquier visualización se puede descomponer en varias partes constituyentes.

Referencia:

- "The Grammar of Graphics (Statistics and Computing)", Leland Wilkinson.



Elementos de un gráfico

1) Los datos

En el centro de cualquier visualización, por supuesto, están los datos que se esperaban visualizar.

2) Las marcas

Para visualizar nuestros datos, debemos representar los datos con marcas reales en nuestra figura.

Estos incluyen no sólo los ejes que dan forma a nuestra figura, sino también puntos, círculos, barras u otras formas geométricas.

Elementos de un gráfico

3) La codificación (o mapeo o mapping)

Para vincular nuestros datos a las marcas en nuestra figura, debemos decidir sobre un mapeo de datos a marcas. Esta codificación de datos en características visuales es donde ocurre la magia.

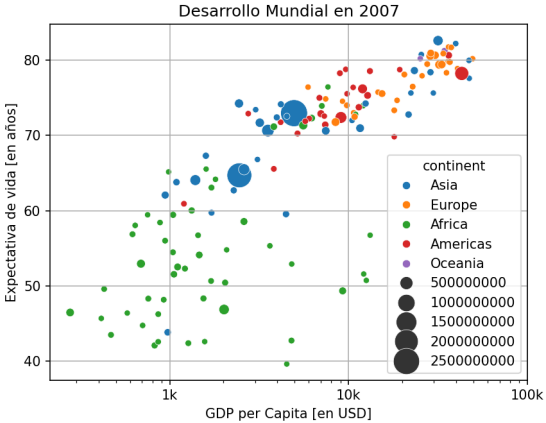
4) Canales

Las características visuales de nuestro gráfico en las que podemos codificar información se denominan "canales".

Algunos canales son obvios: un canal puede ser la ubicación de puntos a lo largo del eje x, otro canal es la ubicación de puntos a lo largo del eje y.

Hay otros canales en los que también podemos codificar datos, como el tamaño, la forma o el color de los puntos.

Gramática de Gráficos - Ejemplo



Canales: coordenada en el eje X, coordenada en el eje Y, tamaño y color.

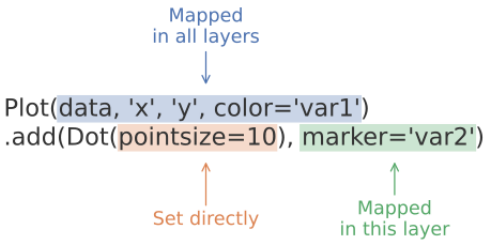
Seaborn objects en Python

Introducido a finales de 2022, el nuevo sistema está basado en el paradigma "Gramática de Gráficos" que utilizan otros paquetes como `ggplot2` de R.
No necesitamos recordar una docena de métodos para hacer gráficos, todo gráfico se hace mediante una única clase `Plot()`.

Seaborn es una biblioteca de visualización construida sobre `matplotlib`, para interactuar en forma más amigable.

Mapeo y asignación por capas

- Si asignamos una codificación (o mapeo) al definir un Plot(), el mapeo se asigna en todas las capas de marcas (objetos mark).
- Si asignamos una codificación dentro del método add() de una marca, el mapeo se realiza solo en esa capa.
- Si asignamos un parámetro de la marca, el valor se asigna directamente (no es una codificación de datos).



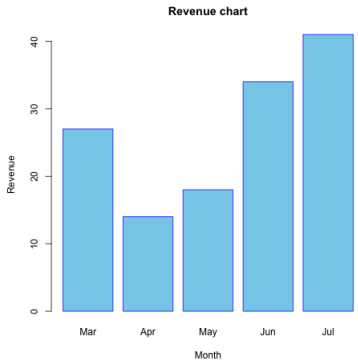
Distintos gráficos

Vimos hasta ahora:

- **Curvas.** Se utilizan para representar funciones, para cada valor de la variable en el eje x tengo un único valor de la variable en el eje y.
- **Gráficos de dispersión (scatter plot).** Se grafican puntos en el plano, para analizar la relación entre dos variables numéricas. Para cada valor de x puede haber varios valores de y.

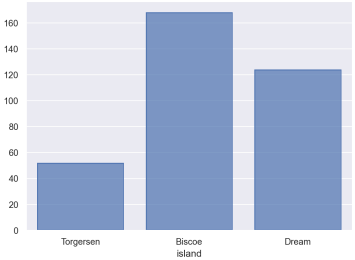
Gráfico de barras

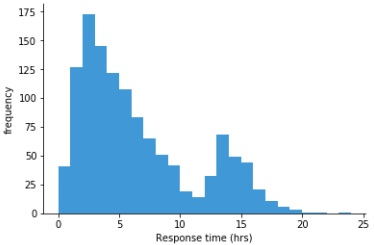
Un **gráfico de barras** muestra la relación entre una variable categórica y una variable numérica. A cada valor de la variable categórica corresponde una barra. El tamaño de la barra representa el valor númerico correspondiente al dato categórico.



Histograma

Un **histograma** puede pensarse como un caso particular de gráfico de barras, para una serie de datos de una variable numérica o categórica, donde en el eje y codificamos cantidades o frecuencias.



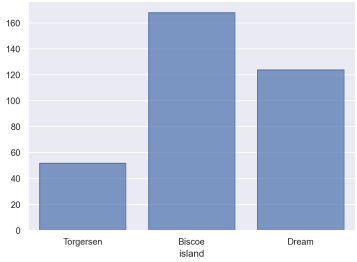


Histograma categórico

Si la variable es categórica:

- en el eje X representamos los distintos valores de la variable.
- en el eje Y representamos la cantidad de veces que aparece cada valor en la serie.

El tamaño de la barra representa la cantidad de veces que se repite cada valor de la variable categórica en la serie.

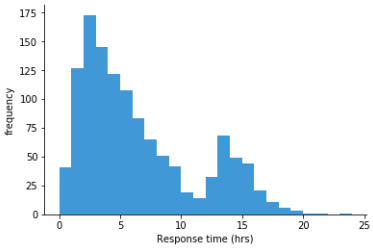


Histograma de una variable numérica

Si la variable es numérica:

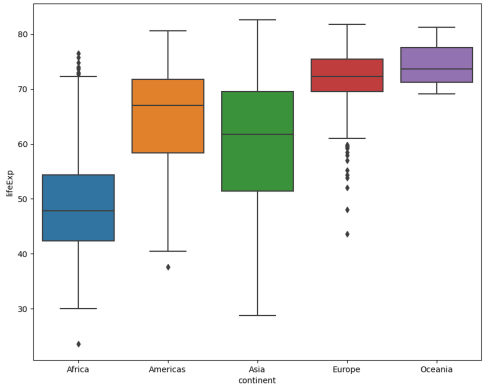
- en el eje X representamos intervalos contiguos de los valores que toma la variable.
- en el eje Y representamos la cantidad de veces que los valores de la serie caen en cada intevalo.

El tamaño de la barra representa la cantidad de veces que el valor de la serie cae en ese intervalo.



Box plot

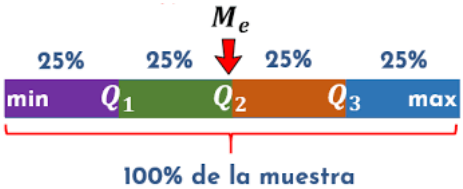
Un gráfico **box plot** usa cajas y líneas para mostrar información de la distribución de uno o más grupos de datos numéricos.



Elementos de un Box plot: cuartiles

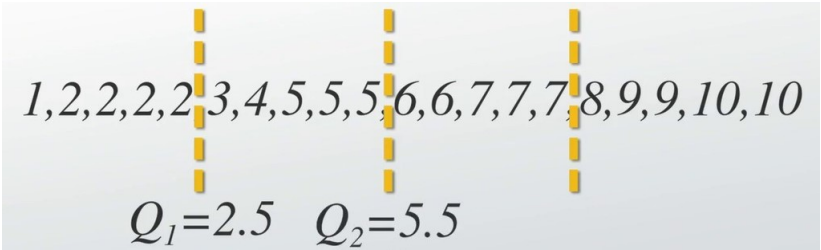
Dada una variable numérica, ordenamos los valores y los partimos en 4 grupos de igual tamaño.

- Primer cuartil (Q1): es un valor mayor que el 25% de los datos y menor que el otro 75%.
- Segundo cuartil (Q2): es un valor mayor que el 50% de los datos y menor que el otro 50% (corresponde a la mediana).
- Tercer cuartil (Q3): es un valor mayor que el 75% de los datos y menor que el otro 25%.



Cuartiles: ejemplo

Ordenamos los datos de menor a mayor y tomamos valores que dividen a los datos en 4 partes iguales.



Cálculo de cuartiles en Wikipedia...

Métodos de computación [editar]

Distribuciones discretas [editar]

Para distribuciones discretas, no existe un acuerdo universal sobre la selección de los valores de los cuartiles. [2]

Método 1 [editar]

1. Utilice la mediana para dividir el conjunto de datos ordenados en dos mitades. La mediana se convierte en los segundos cuartiles.
   - Si hay un número impar de puntos de datos en el conjunto de datos ordenados original, **incluya** la mediana (el valor central en la lista ordenada) en ninguna de las mitades.
   - Si hay un número par de puntos de datos en el conjunto de datos ordenado original, divida este conjunto de datos exactamente por la mitad.
2. El valor del cuartil inferior es la mediana de la mitad inferior de los datos. El valor del cuartil superior es la mediana de la mitad superior de los datos.

Esta regla la emplean el diagrama de caja de la calculadora TI-83 y las funciones "Estadísticas de 1 var".

Método 2 [editar]

1. Utilice la mediana para dividir el conjunto de datos ordenados en dos mitades. La mediana se convierte en los segundos cuartiles.
   - Si hay un número impar de puntos de datos en el conjunto de datos ordenados original, **incluya** la mediana (el valor central en la lista ordenada) en ambas mitades.
   - Si hay un número par de puntos de datos en el conjunto de datos ordenado original, divida este conjunto de datos exactamente por la mitad.
2. El valor del cuartil inferior es la mediana de la mitad inferior de los datos. El valor del cuartil superior es la mediana de la mitad superior de los datos.

Los valores encontrados por este método también se conocen como " bisagras de Tukey ", [4] ver también bisagra media .

Método 3 [editar]

1. Utilice la mediana para dividir el conjunto de datos ordenados en dos mitades. La mediana se convierte en los segundos cuartiles.
   1. Si hay un número impar de puntos de datos, vaya al siguiente paso.
   2. Si hay números pares de puntos de datos, entonces el Método 3 comienza igual que el Método 1 o el Método 2 anteriores y puede optar por incluir o no la mediana como un nuevo punto de datos. Si elige incluir la mediana como el nuevo punto de datos, continúe con el paso 2 o 3 a continuación porque ahora tiene un número impar de puntos de datos. Si no elige la mediana como nuevo punto de datos, continúe con el Método 1 o 2 donde comenzó.
2. Si hay ($4\,n$ +1) puntos de datos, entonces el cuartil inferior es el 25% del *enésimo* valor de datos más el 75% del (n +1)ésimo valor de datos; el cuartil superior es el 75 % del punto de datos ($3\,n$ +1) más el 25 % del punto de datos ($3\,n$ +2).
3. Si hay ($4\,n$ +3) puntos de datos, entonces el cuartil inferior es el 75% del (n +1)ésimo valor de datos más el 25% del (n +2)ésimo valor de datos; el cuartil superior es el 25 % del punto de datos ($3\,n$ +2) más el 75 % del punto de datos ($3\,n$ +3).

Método 4 [editar]

Si tenemos un conjunto de datos ordenados$x_1, x_2, \ldots, x_n$, entonces podemos interpolar entre puntos de datos para encontrar elp^{th} cuantil empírico six_i está en el$i/(n+1)$cuantil. Si denotamos la parte entera de un númeroαpor$\lfloor \alpha \rfloor$, entonces la función cuantil empírica viene dada por,

$$q(p/4) = x_k + \alpha(x_{k+1} - x_k),$$

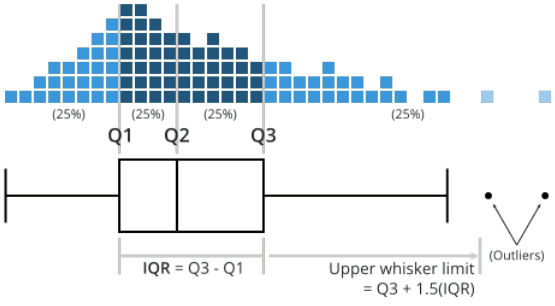
dónde$k = \lfloor p(n+1)/4 \rfloor$y$\alpha = p(n+1)/4 - \lfloor p(n+1)/4 \rfloor$. [1]

Para encontrar el primer, segundo y tercer cuartil del conjunto de datos, evaluaríamos$q(0.25)$.$q(0.5)$.$yq(0.75)$respectivamente.

Elementos de un Box plot

Dada una variable numérica, ordenamos los valores y los partimos en 4 grupos de igual tamaño.

En un box plot, dibujamos una caja, con límites en Q1 y Q3 y una línea central marcando el valor de Q2.



Elementos de un Box plot

La distancia entre Q3 y Q1 se conoce como *rango intercuartil* (IQR) y se utilizan para trazar los "bigotes".

Cada bigote se extiende hasta el valor más lejano de los datos a una distancia menor a 1.5 veces el valor IQR.

Cualquier valor de los datos más allá de esa distancia se considera un dato atípico (outlier) y se marca con un punto.