Medidas de tendencia central

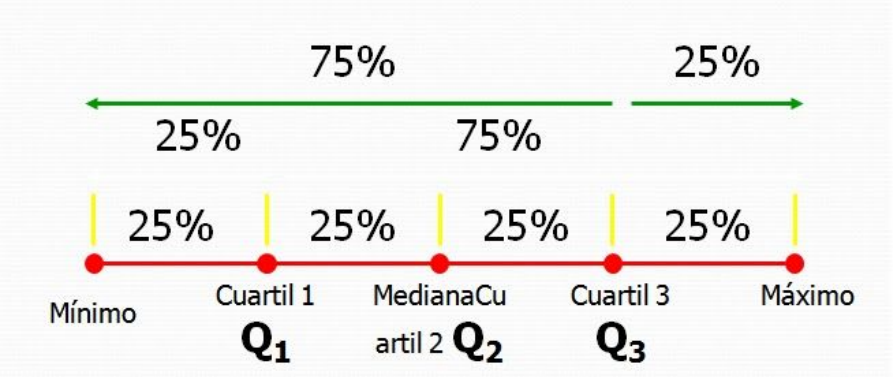
Medida de tendencia central	Definición	Fórmula
Media	Es el promedio del conjunto de datos.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$
Mediana	Es el valor que se encuentra en la posición central de los datos ordenados.	$Me = L_i + h \cdot \frac{\frac{N}{2} - F_{i-1}}{f_i}$ L_i : extremo inferior de la clase medial h : amplitud de la clase medianal N : número total de datos F_{i-1} : frecuencia absoluta acumulada del intervalo anterior a la clase medianal f_i : frecuencia absoluta de la clase medianal
Moda	Es el valor que tiene mayor frecuencia en el conjunto de datos.	$Mo = Li + \left(\frac{\Delta 1}{\Delta 1 + \Delta 2}\right) * c$

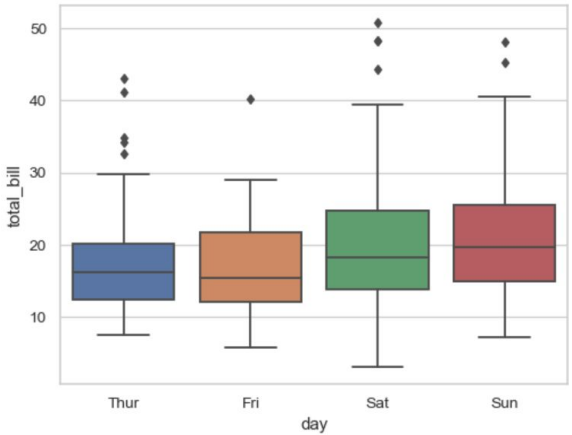
Medidas de tendencia central

Percentil: Indica la posición de un valor dado el porcentaje dentro de un conjunto de datos (el cual debe estar ordenado de menor a mayor).

Un caso específico de los percentiles son los **cuartiles** que dividen a los datos en cuatro partes iguales entre dichos porcentajes:

- 25%
- 50% (mediana)
- 75%



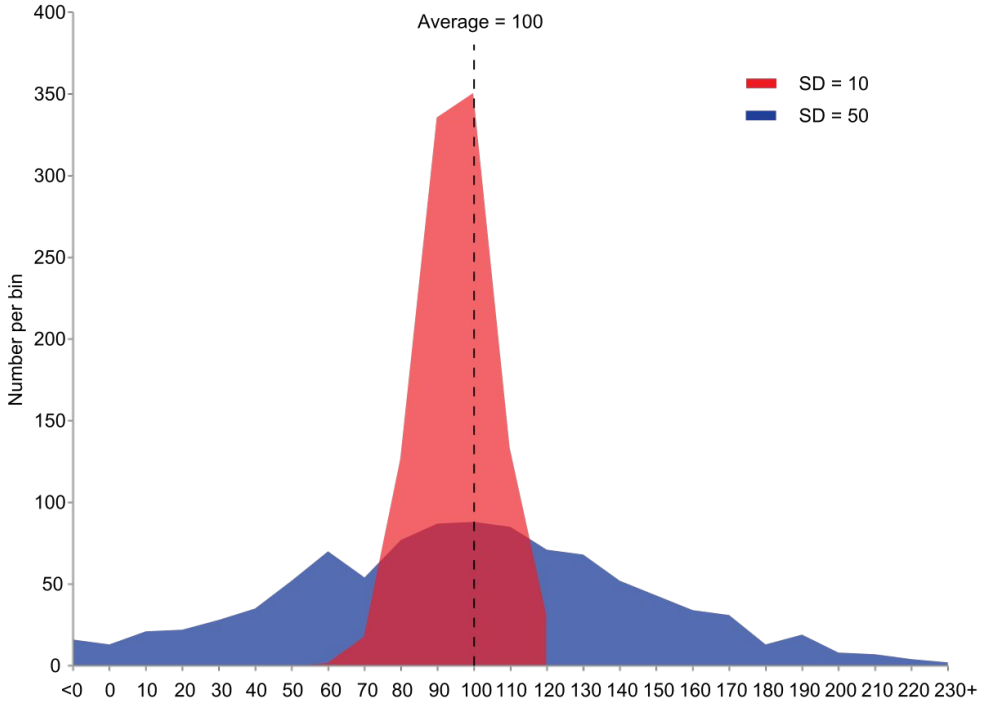


Medidas de dispersión de datos.

Varianza: Nos mide qué tan cerca están distribuidos los datos con respecto a su media.

La desviación estándar no es más que la raíz cuadrada de la varianza. Y se hace de esta manera para conservar las unidades.

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \bar{x}|^2}{n}}$$

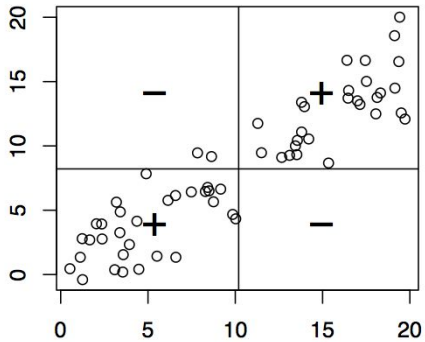


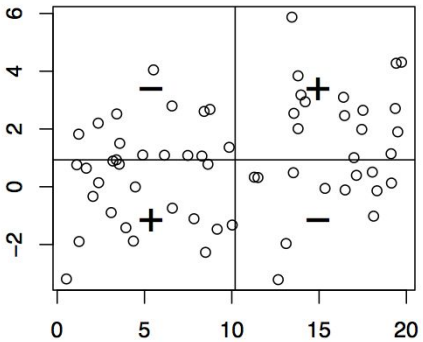
boost
accelerate your learning

Covarianza

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Es la medida de la variación en conjunto de dos variables (En el ejemplo x e y). Se puede interpretar como cuánto se mueven en conjunto estas dos variables, si es que una sube (o baja) y la otra también sube (o baja)





Uso práctico: Si tengo dos variables que sospecho que siempre se mueven juntas, puedo medir ese movimiento con la covarianza.

Pero, estas variaciones dependen en sí de las magnitudes de los datos.

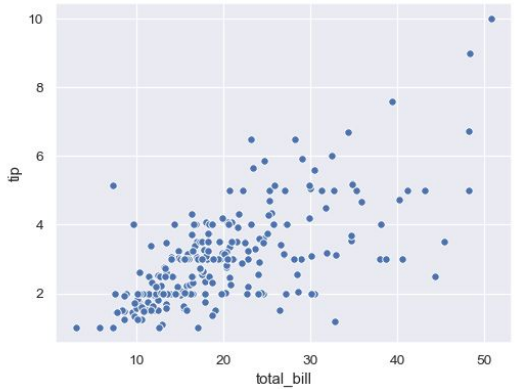
Correlación (Pearson)

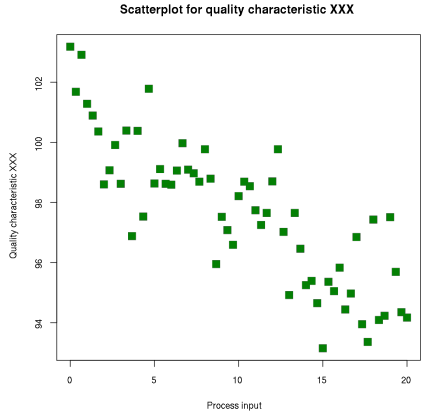
Es una forma de estandarizar los pesos de la covarianza y hacerla inmune a los cambios en los datos.

Es una forma de ver relaciones lineales dentro de mis datos

$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

Sus valores van de -1 a 1.

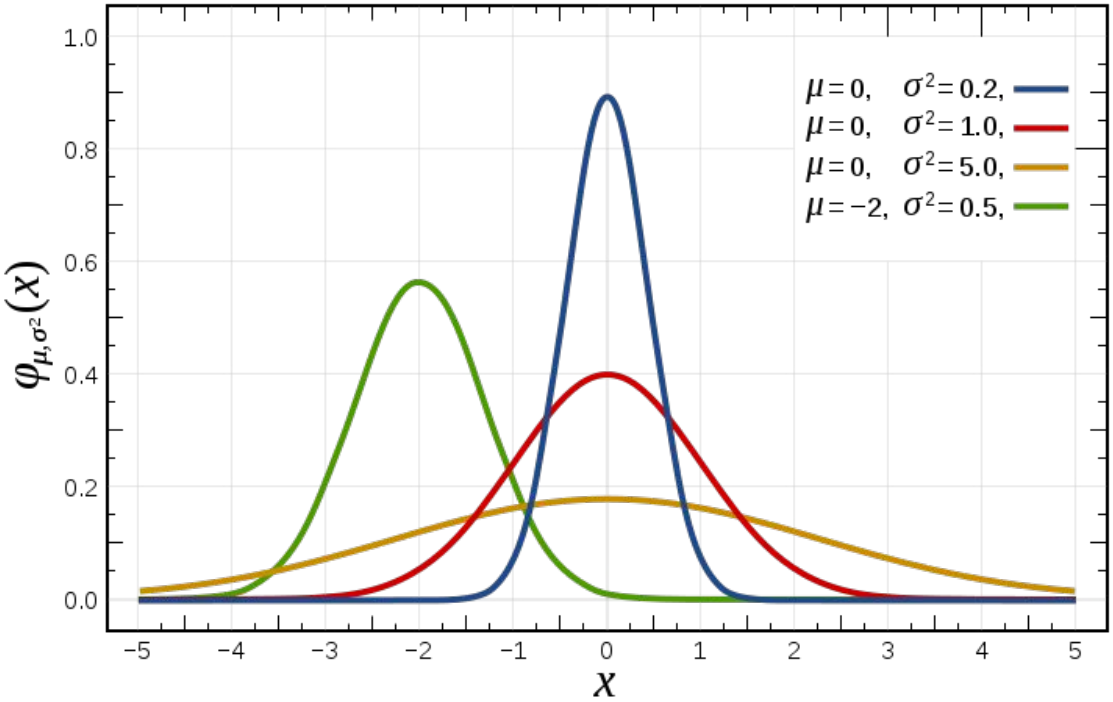




boost
accelerate your learning

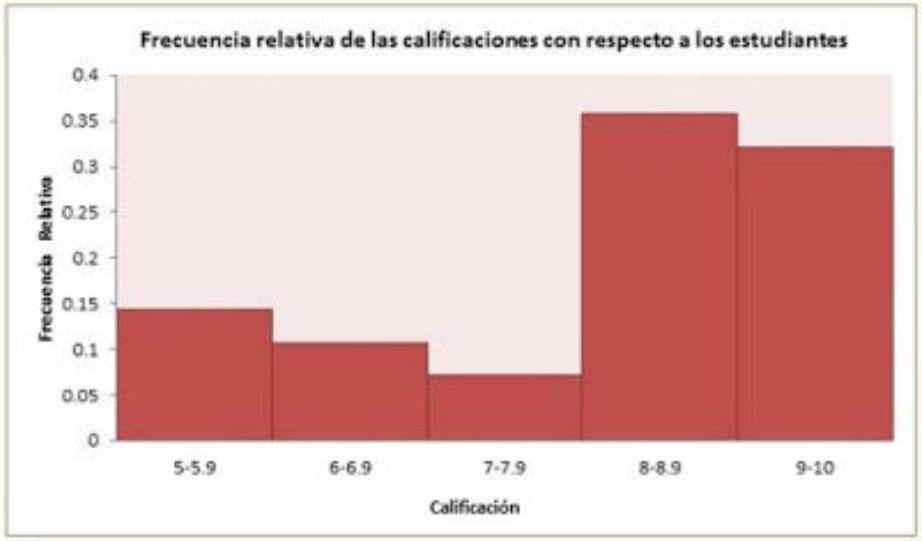
Kurtosis

Nos permite medir qué tanto influye la media a los datos aledaños a ella. Es decir, si mi distribución es muy “picuda” es porque los valores están más juntos a ella y decrecen en sus alrededores (Alta kurtosis). Si es más “cabezona” o “plana”, es porque los valores más cercanos tienen casi la misma frecuencia que la media.



Histograma de frecuencias relativas

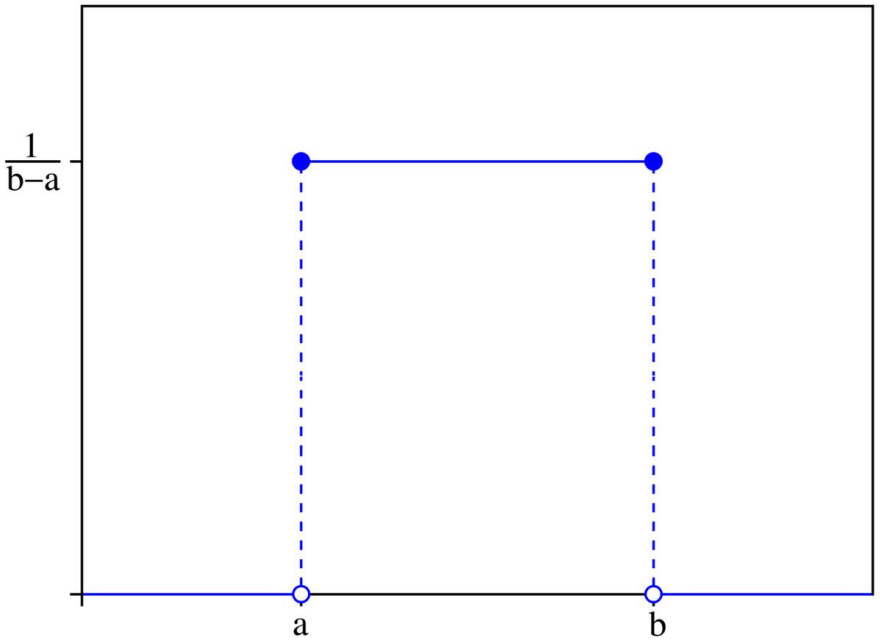
Nos agrupa los datos por intervalos de frecuencia. Es decir, cuántos valores hay en un número N de intervalos iguales. Luego, se hace un gráfico de barras con respecto a estos números para poder obtener en el eje Y la frecuencia relativa (La probabilidad de que los valores estén dentro de ese intervalo)



Distribuciones

Distribución uniforme: Es una distribución de probabilidad en la que cada valor del conjunto de datos tiene la misma probabilidad de ocurrencia.

$$f(x)= \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{en otro caso} \end{cases}$$

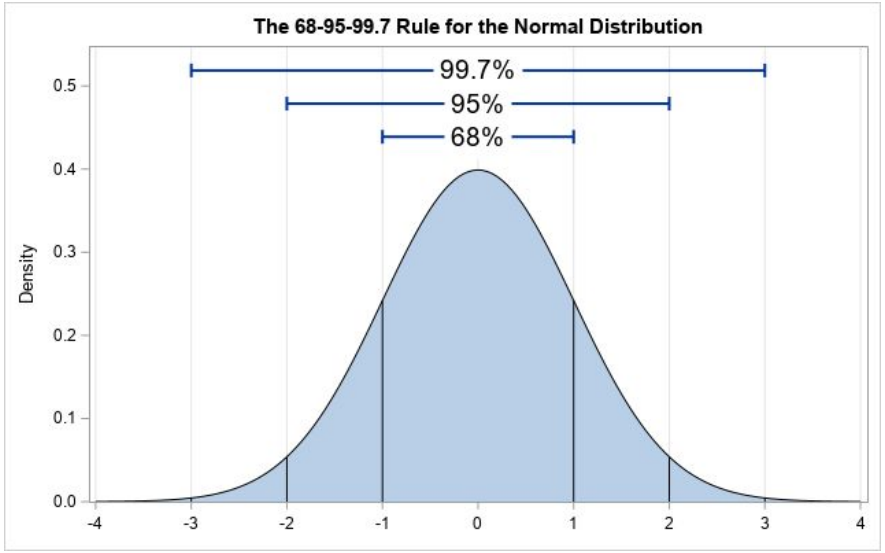


Distribuciones

Distribución normal:
O también llamada distribución Gaussiana. Es la distribución más usada y conocida en el mundo, nos permite modelar y trabajar bajo supuestos de que nuestros datos siempre van a tender a la media.

Usualmente se dice que debe de tener media 0 y desv. estándar de 1.

Varios modelos asumen que los datos tienen una distribución normal, y es nuestro deber poder transformar los datos para que podamos llegar a ello. Por ejemplo:

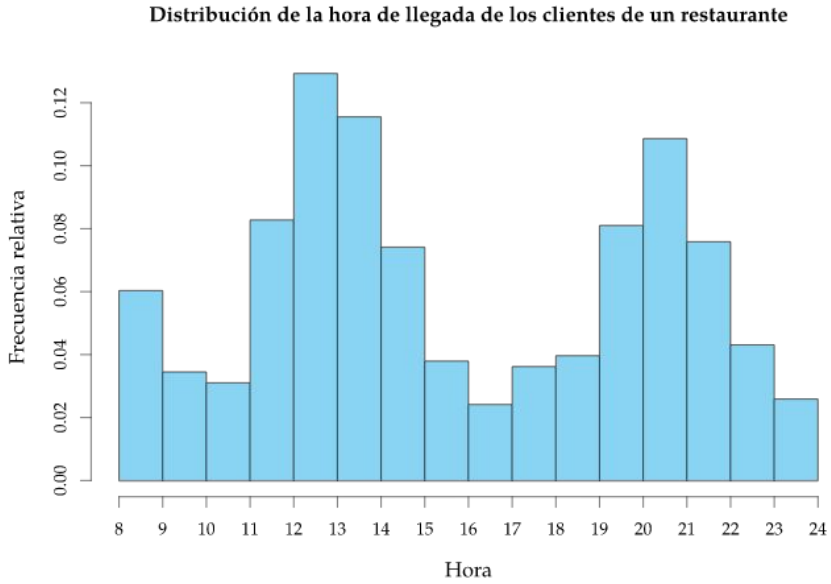


Distribuciones

Distribución multimodales: Es una distribución de probabilidad con múltiples modas (máximos locales).

Un caso específico son las **distribuciones bimodales** que presentan dos máximos locales

Existen formas de poder explicar estas bimodalidades, y es ahí donde viene nuestro trabajo. Tenemos que ver y explicar estas dos tendencias con factores que encontremos dentro de nuestras variables.



boost
accelerate your learning

Asimetría

$$\frac{\sum_{i}^{N}(X_i - \bar{X})^3}{(N-1)*\sigma^3}$$

Es un coeficiente que nos va a medir qué tan oblicua está nuestra distribución. Si la media mediana y moda son las mismas o de qué forma una difiere de la otra.