Empleos del Futuro

INTRODUCCIÓN A LA CIENCIA DE DATOS

Leonardo Ignacio Córdoba

Agenda del curso

Agenda del curso

- Primer encuentro:

Agenda del curso

- Primer encuentro:
  - Introducción a los datos

Agenda del curso

- Primer encuentro:
  - Introducción a los datos
  - Estadística descriptiva

Agenda del curso

- Primer encuentro:
  - Introducción a los datos
  - Estadística descriptiva
  - Introducción a Python

Agenda del curso

- Segundo encuentro:

Agenda del curso

- Segundo encuentro:
  - Visualización

Agenda del curso

- Segundo encuentro:
  - Visualización
  - Manipulación de datos

Agenda del curso

- Segundo encuentro:
  - Visualización
  - Manipulación de datos
  - Visualización

Agenda del curso

- Tercer encuentro:

Agenda del curso

- Tercer encuentro:
  - Introducción al ML

Agenda del curso

- Tercer encuentro:
  - Introducción al ML
  - Casos de uso

Agenda del curso

- Tercer encuentro:
  - Introducción al ML
  - Casos de uso
  - Regresión lineal

Agenda del curso

- Cuarto encuentro:

Agenda del curso

- Cuarto encuentro:
  - Problemas de clasificación

Agenda del curso

- Cuarto encuentro:
  - Problemas de clasificación
  - Casos de uso

Agenda del curso

- Cuarto encuentro:
  - Problemas de clasificación
  - Casos de uso
  - Regresión logística

Análisis de la información

- Un dato es una representación simbólica de una característica de la realidad.
- Los conjuntos de datos a veces son nombrados en inglés como *datasets*.


Data Science es la refinería pero
los datos son el petróleo

¿De dónde vienen los datos?

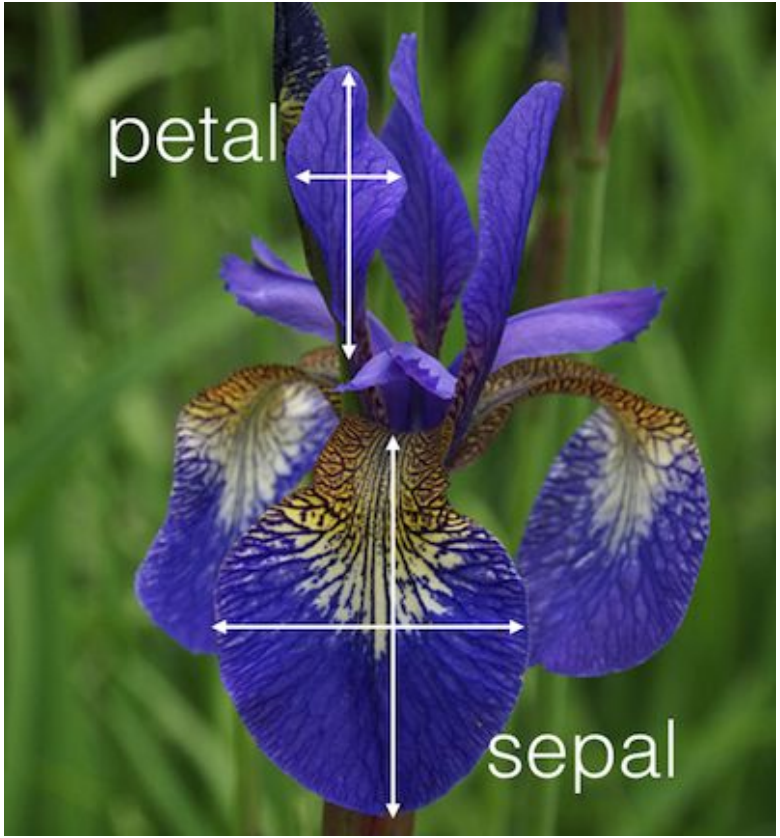
Breve recorrido histórico

- Observaciones de campo
- Experimentos científicos
- Encuestas
- Sistemas transaccionales
- CRM
- Información de sensores
- Comportamiento online
- Comportamiento en las redes sociales
- Textos
- Audios
- Imágenes

¿De dónde vienen los datos?

Observaciones de campo

- Las observaciones de campo históricamente han servido para la clasificación en, por ejemplo, las ciencias biológicas, y son la fuente de información sobre la cual se generan teorías en ciencias naturales.
- El ejemplo clásico en este caso es el Iris Dataset, generado por Fisher en 1936, cuenta con 150 observaciones de tres tipos de lirios. En CABA podemos mencionar, por ejemplo, el Relevamiento de Usos del Suelo.



¿De dónde vienen los datos?

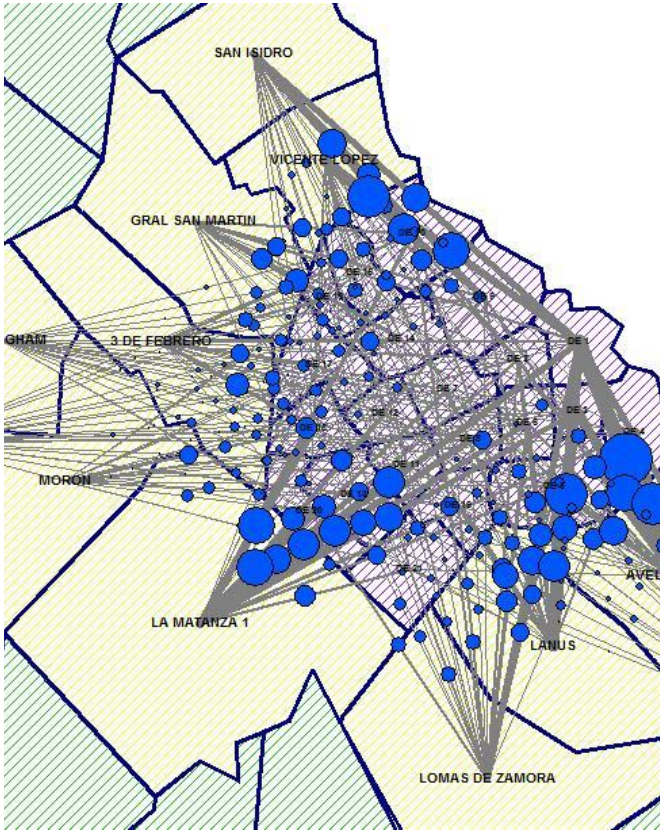
Experimentos científicos

- En ciertas ocasiones se generan datos en ambientes (más o menos) controlados para entender el comportamiento de cierto objeto en unas condiciones dadas.
- Un ejemplo de este tipo de dataset es el llamado mcycle, que se refiere a 133 observaciones de un accidente de moto simulado, en donde se mide la aceleración de la cabeza en distintos momentos del tiempo. En este caso, este dataset se empleó para analizar choques de casos

¿De dónde vienen los datos?

Encuestas

- En las encuestas se cuenta con una serie de preguntas o formulario que los relevadores preguntan a un grupo de encuestados.
- En Argentina se cuentan con distintas encuestas de distinto tamaño y periodicidad:
  - Encuesta Permanente de Hogares (EPH)
  - Censo Nacional
  - Relevamiento de Expectativas de Mercado (REM) del BCRA
  - INTRUPUBA (2007)



¿De dónde vienen los datos?

Sistemas transaccionales

- Los sistemas transacciones surgieron originalmente para el mercado de reservas de vuelos. Posteriormente su uso se extendió a otros rubros, especialmente al sistema bancario y hotelería.
- ACID:
  - Atomicidad
  - Consistencia
  - Aislamiento
  - Durabilidad

¿De dónde vienen los datos?

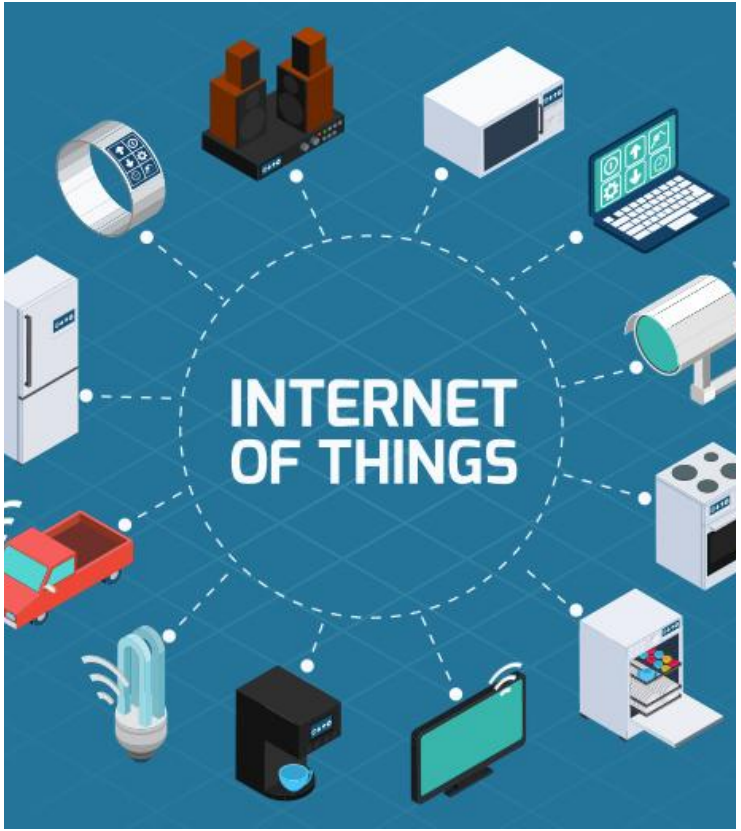
Sistemas de gestión

- La informatización de la administración de los procesos (en el sector privado o público) también genera el registro de las distintas etapas en la gestión.
- Como caso ejemplar podemos mencionar los CRM (Customer Relationship management) que son sistemas en los que se cuenta con toda la información que el cliente (o potenciales clientes) nos provee en los distintos puntos de contacto. Esto permite facilitar el trabajo para alcanzar nuevos clientes y retener mejor a los ya existentes.

¿De dónde vienen los datos?

Información de sensores - IOT

- Con la extensión de los sensores y de las redes inalámbricas (WiFi - 2G - 3G - 4G), surgió el llamado “Internet de las cosas”.
- Hoy por hoy gran parte de los objetos que consumimos cuentan con decenas de sensores que almacenan y transmiten información sobre distintos aspectos del propio objeto o del ambiente.
- Ejemplos: sensores en electrodomésticos, autos, celulares, aviones, ambientales, etc.



¿De dónde vienen los datos?

Comportamiento online

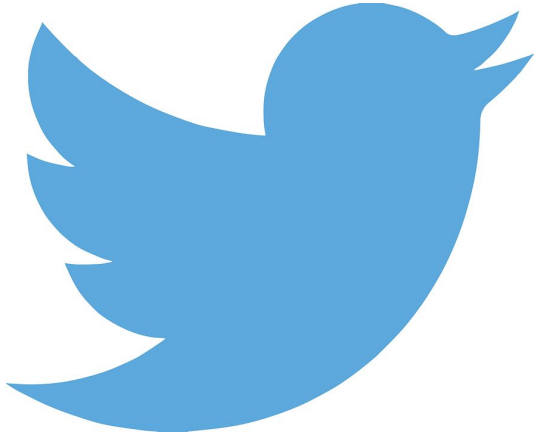
- El comportamiento que realizamos cada día a través de internet queda registrado y es, incluso, revendido.
- Esto incluye tanto cuando entramos a una aplicación web, como cuando es un app mobile o desde un smart tv.
- Existen distintas herramientas para trackear cada uno de estos aspectos, la más conocida es Google Analytics.

Google Analytics

¿De dónde vienen los datos?

Comportamiento en las redes sociales

- Nuestra actividad en las distintas redes sociales puede ser consumida y analizada.
- Las principales redes sociales son Facebook, Instagram, Twitter y Youtube.
- El reciente escándalo de Cambridge Analytica tiene que ver con ésto.



¿De dónde vienen los datos?

Textos

- Los textos se pueden extraer desde cualquier lugar en que estén, especialmente online.
- Libros públicos, blogs, diarios, revistas, etc.
- Scraping de páginas web.

¿De dónde vienen los datos?

Audios

- El audio hoy en día es muy utilizado para entrenar los sistemas que procesan el habla, como los asistentes de voz de los celulares.
- La información puede venir de canciones, personas hablando en distintos idiomas, sonidos de ambiente, etc.

¿De dónde vienen los datos?

Imágenes

- Las imágenes son muy estudiadas para los sistemas capaces de procesar o interpretar imágenes, desde un OCR hasta un sistema de IA moderno.

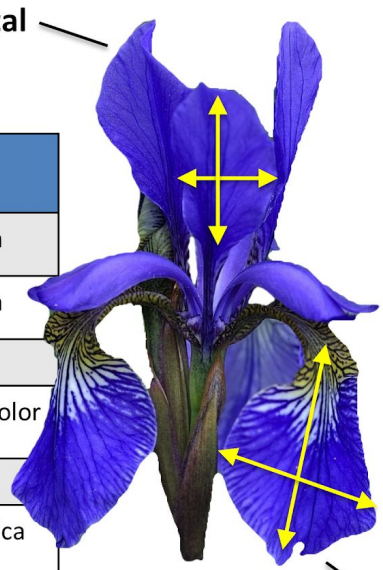

ESTADÍSTICA DESCRIPTIVA

Introducción a la estadística descriptiva

Observaciones y atributos

Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

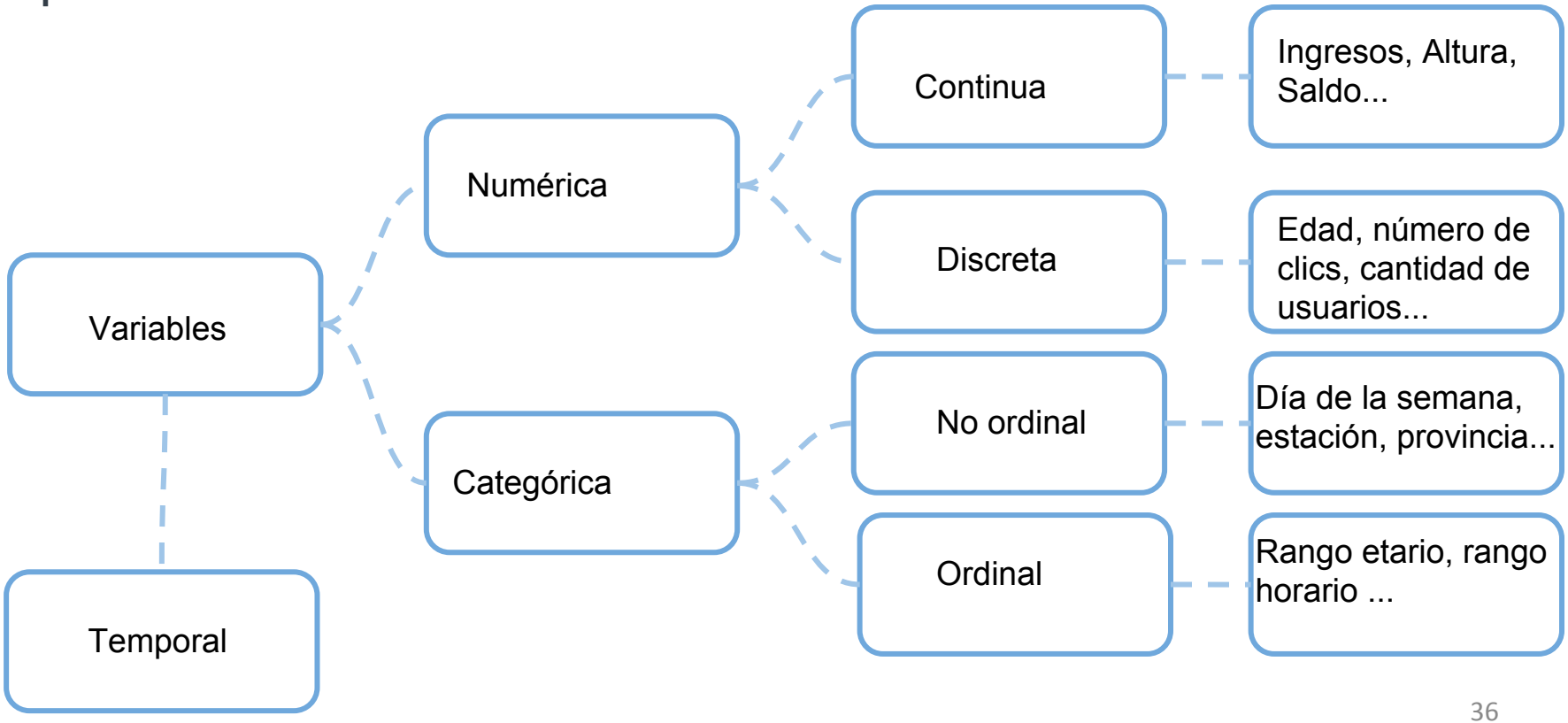


Features
(attributes, measurements, dimensions)

Class labels
(targets)

Introducción a la estadística descriptiva

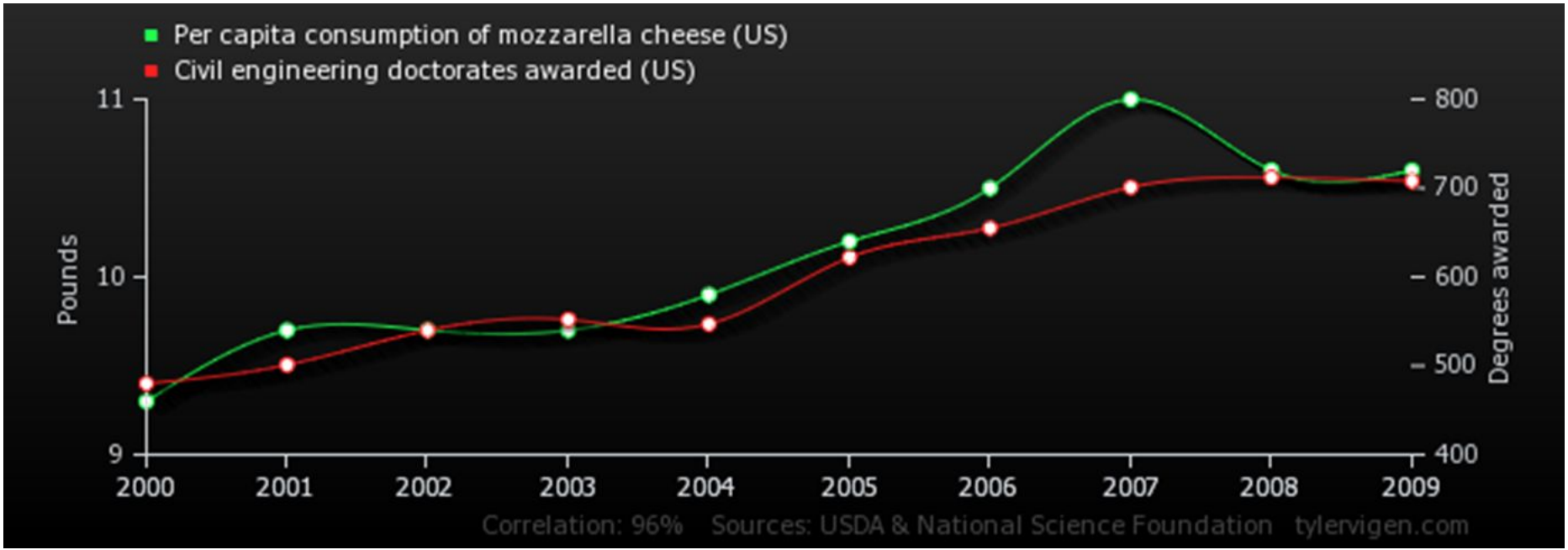
Tipos de variables



Correlación y causalidad

Introducción a la estadística descriptiva

Correlación y causalidad



Medidas de centralidad

Las medidas de centralidad sirven para caracterizar un conjunto de datos a partir de una medida que nos indique qué valor es usual o, de alguna manera, “representativo”.

Vamos a destacar tres medidas:

- Media
- Mediana
- Moda

Media

La **media** se define de la siguiente manera:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Por ejemplo, para la muestra 8, 5 y -1, su **media** es:

$$\bar{x} = \frac{8 + 5 + (-1)}{3} = 4$$

Mediana

La **mediana** puede pensarse de manera simple como el valor del ”medio” de una lista ordenada de datos (o el valor que separa la primera mitad y la segunda mitad de una distribución).
Para una lista ordenada la mediana es calculada de diferente manera dependiendo de la cantidad de elementos de la misma:

- **Impar:**

  [1, 2, 3, 5, **7**, 8, 9, 10, 15]
  #elementos: 9
  La mediana es el valor de la posición 5 (la posición del “medio”)
  Mediana = 7

- **Par:**

  [-5, -1, 0, **1**, **2**, 3, 8, 20]
  #elementos: 8
  La mediana es la media de los valores en las dos posiciones centrales
  Mediana = (1+2)/2 = 1.5

Moda

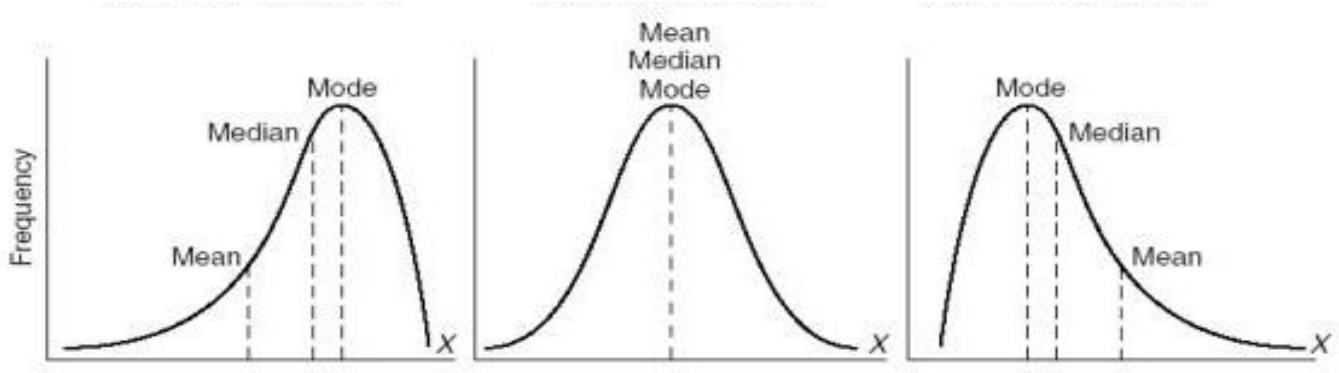
La **moda** es el valor que aparece con mayor frecuencia o más veces en la distribución.

Por ejemplo, la moda de [0,1,1,2,2,2,2,3,3,4,4,4,5] es 2.

La moda no es necesariamente única. Puede ocurrir que haya dos valores diferentes que sean los más frecuentes. Por ejemplo, para [10, 13, 13, 20, 20], tanto 13 como 20 son la moda.

Introducción a la estadística descriptiva

Asimetría



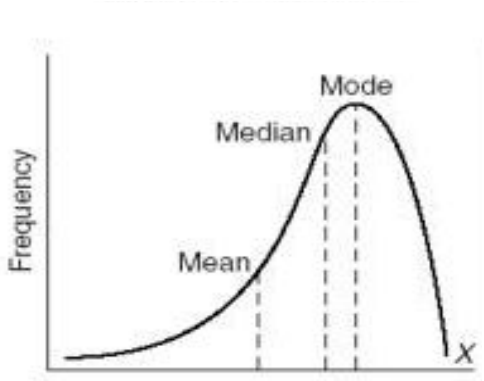
Una distribución con **asimetría a derecha** significa que la cola del lado derecho es más larga que la de la izquierda (gráfico a la derecha)

De la misma manera, una distribución con **asimetría a izquierda**, significa que la cola de la izquierda es más larga que la de la derecha (gráfico a izquierda).

Por último, una **distribución simétrica** no presenta este fenómeno dado que sus colas son de igual longitud al ser simétrica.

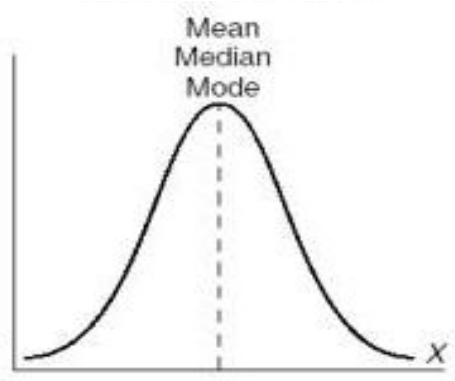
Asimetría

La media, mediana y moda son afectadas por la asimetría:



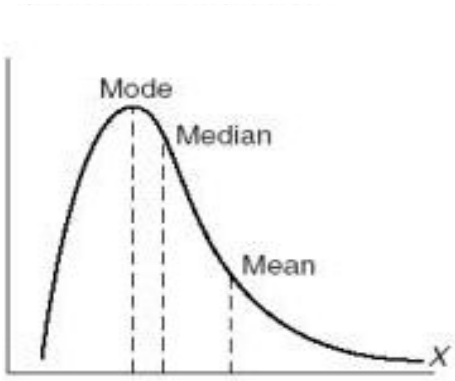
Asimetría a izquierda

Media < Mediana < Moda



Simetría

Media = Mediana = Moda

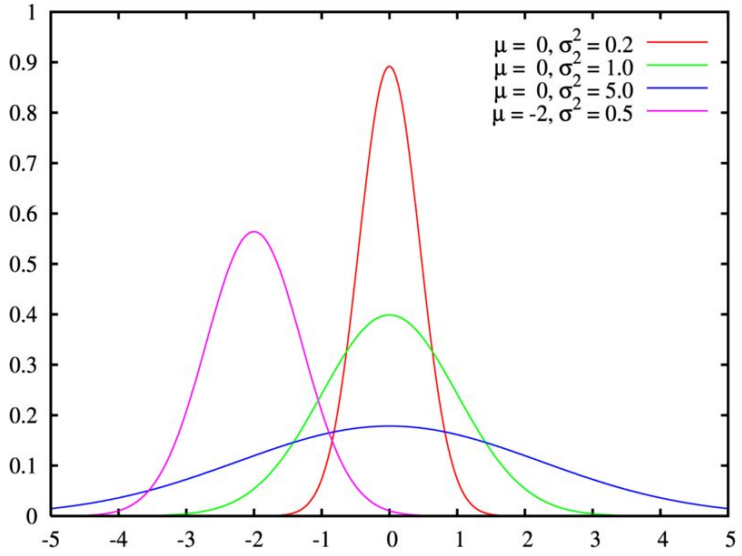


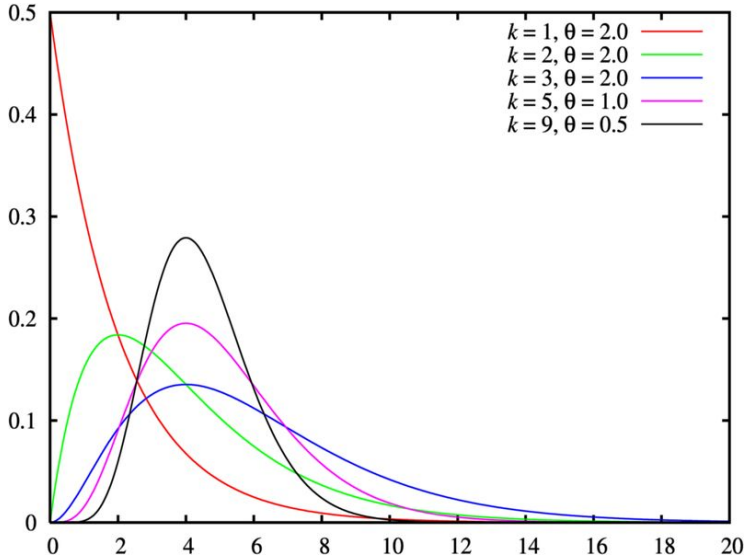
Asimetría a derecha

Moda < Mediana < Media

Introducción a la estadística descriptiva

Normal y Gamma





La distribución normal es un ejemplo de distribución **simétrica**

La distribución gamma es un ejemplo de distribución **asimétrica**

Medidas de dispersión

Las medidas de dispersión o variabilidad nos permiten resumir qué tanto se alejan usualmente las observaciones del centro de la distribución.

Podemos mencionar:

- Rango
- Varianza
- Desvío Estándar

Varianza

La **varianza** es un valor numérico utilizado para describir cuánto varían los números de una distribución respecto a su media.

La varianza puede ser calculada como:

$$s^2 = \frac{\sum (X - \overline{X})^2}{N}$$

Esto es el **promedio de la diferencia elevada al cuadrado entre cada valor y la media**.

Desvío estándar

El **desvío estándar** es la raíz cuadrada de la varianza.

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

No es la desviación promedio con respecto de la media. Como los desvíos están elevados al cuadrado los desvíos muy grandes cuentan más que proporcionalmente.

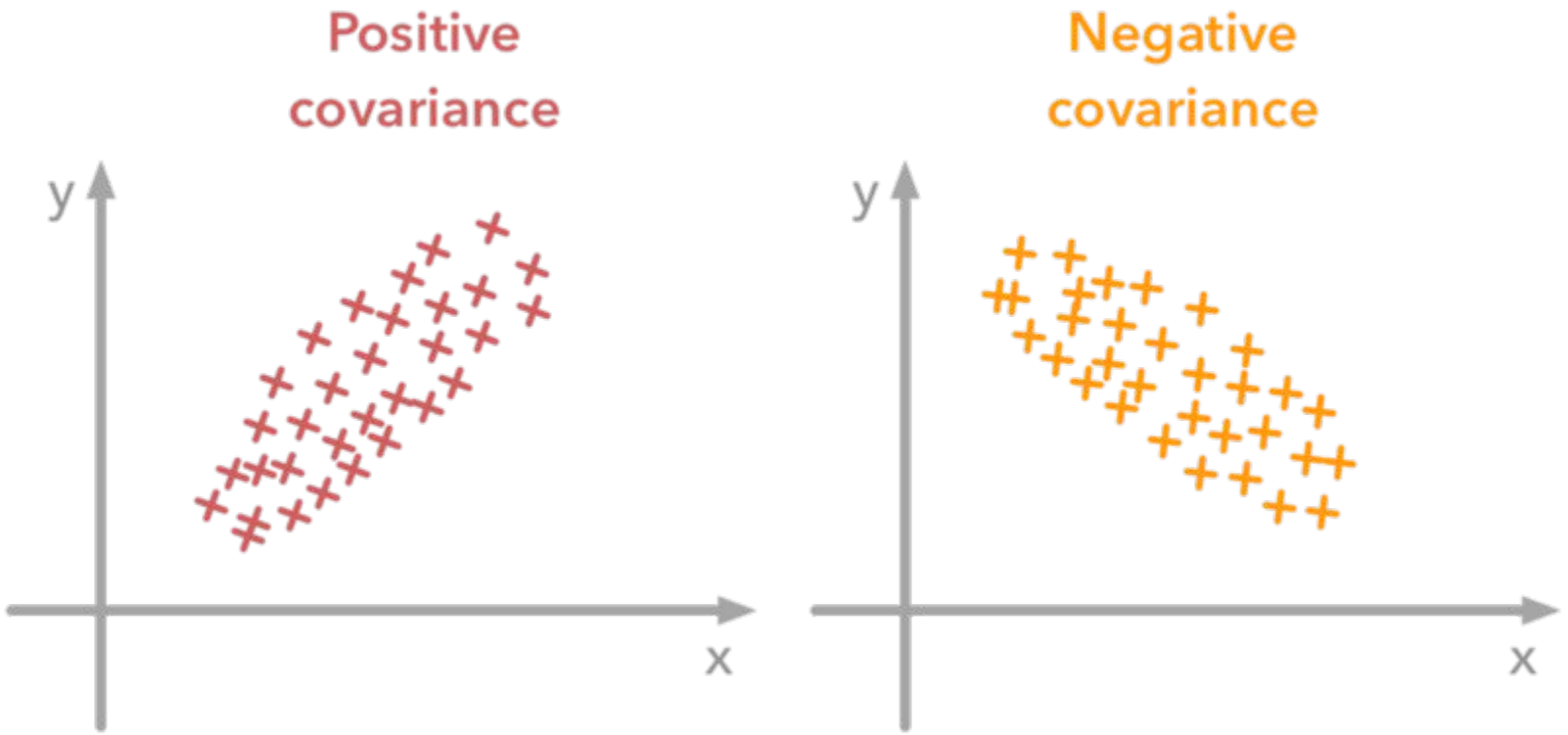
Coeficiente de variación

El **coeficiente de variación** es el desvío estándar dividido por la media

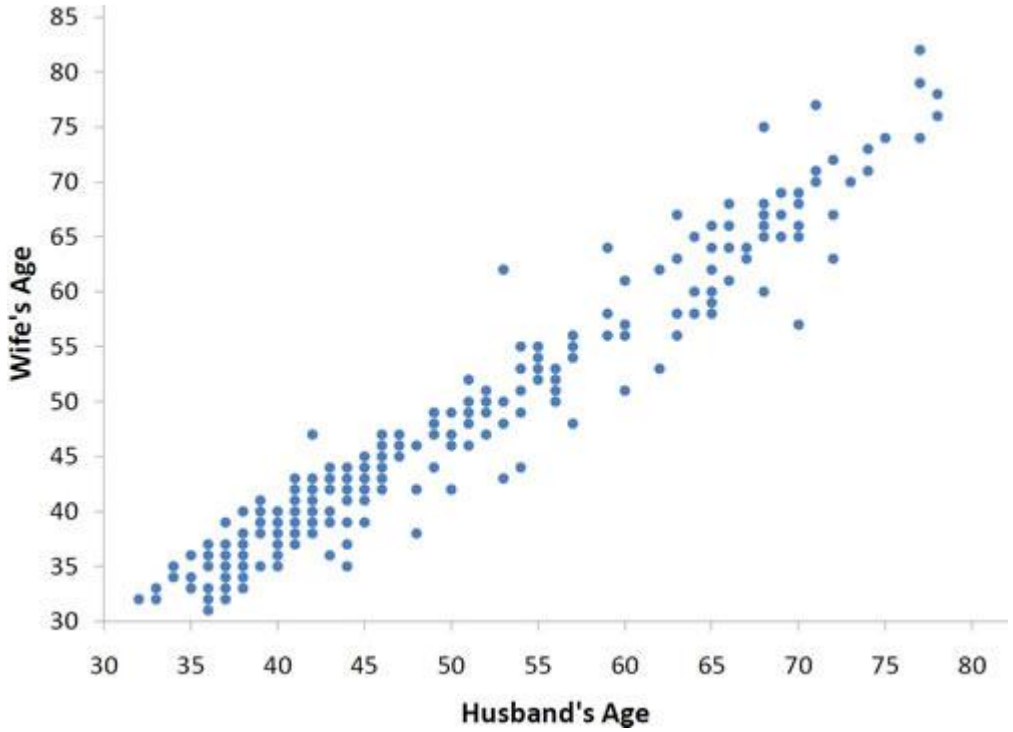
$$CV = \left(\frac{S}{\bar{X}} \right) \cdot 100\%$$

- El coeficiente de variación permite **comparar la dispersión de variables diferentes**.
  - Sirve si las variables tienen medias distintas.
  - También si las variables están expresadas en unidades distintas.
- El coeficiente de variación **no tiene unidades**.

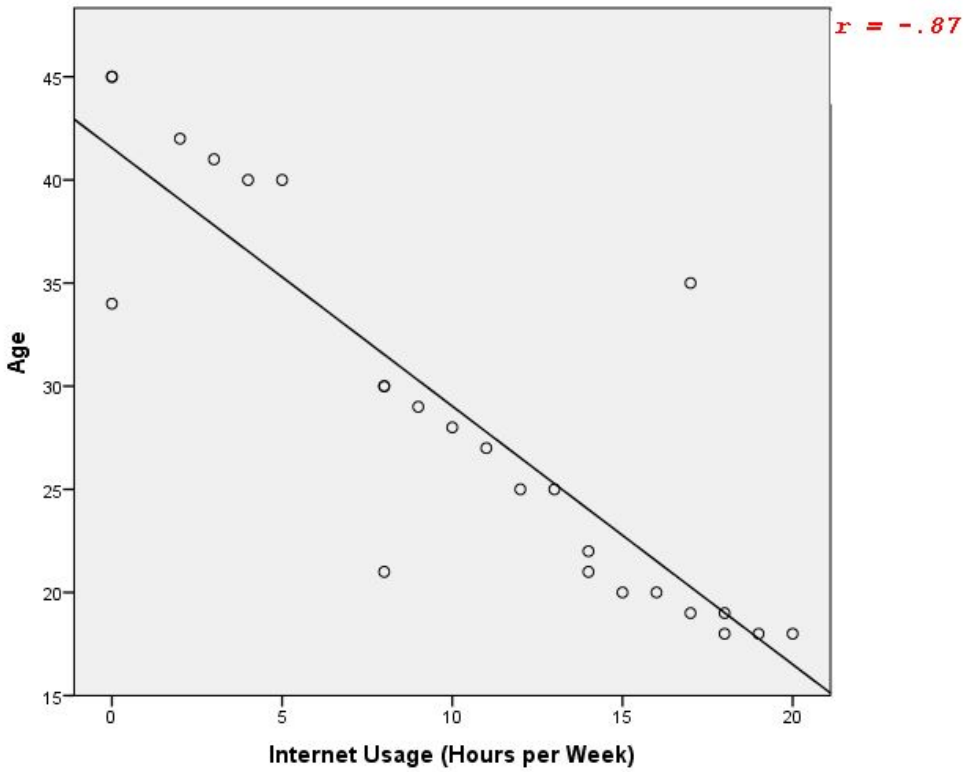
Medidas de asociación lineal entre variables: covarianza



Medidas de asociación lineal entre variables: covarianza

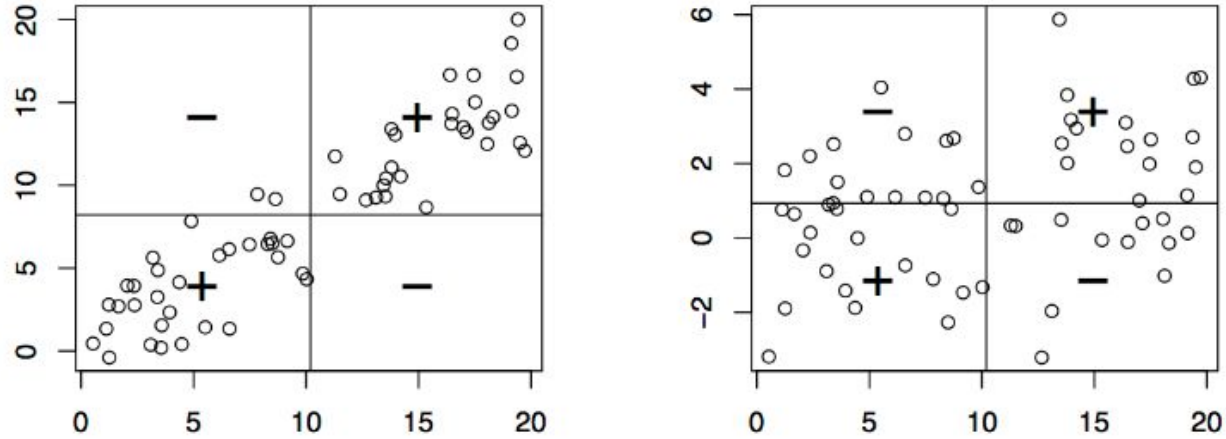


Medidas de asociación lineal entre variables: covarianza



Introducción a la estadística descriptiva

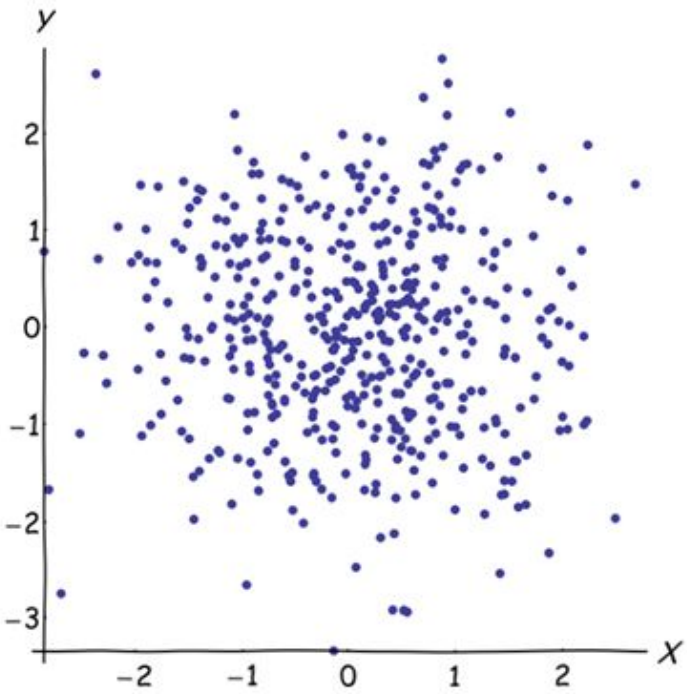
Medidas de asociación lineal entre variables: covarianza



Decimos que dos variables X e Y, tienen covarianza positiva cuando tienden a encontrarse por encima de su media al mismo tiempo y tienen covarianza negativa cuando al mismo tiempo, tienden a encontrarse una por debajo y otra por encima.
En cambio X e Y tienen covarianza cercana a cero cuando las variables pueden encontrarse por encima o por debajo de su media independientemente de lo que haga la otra.

Introducción a la estadística descriptiva

Medidas de asociación lineal entre variables: covarianza



Medidas de asociación lineal entre variables: covarianza y correlación

La covarianza se mide como: $$cov_{XY} = \frac{1}{N} \sum_{i=1}^{N} (X - \overline{X})(Y - \overline{Y})$$

La correlación es una versión estandarizada (dividida por los desvíos estándar) de la covarianza:

$$r_{xy} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_i - \overline{X}}{S_x}\right)\left(\frac{Y_i - \overline{Y}}{S_y}\right)$$

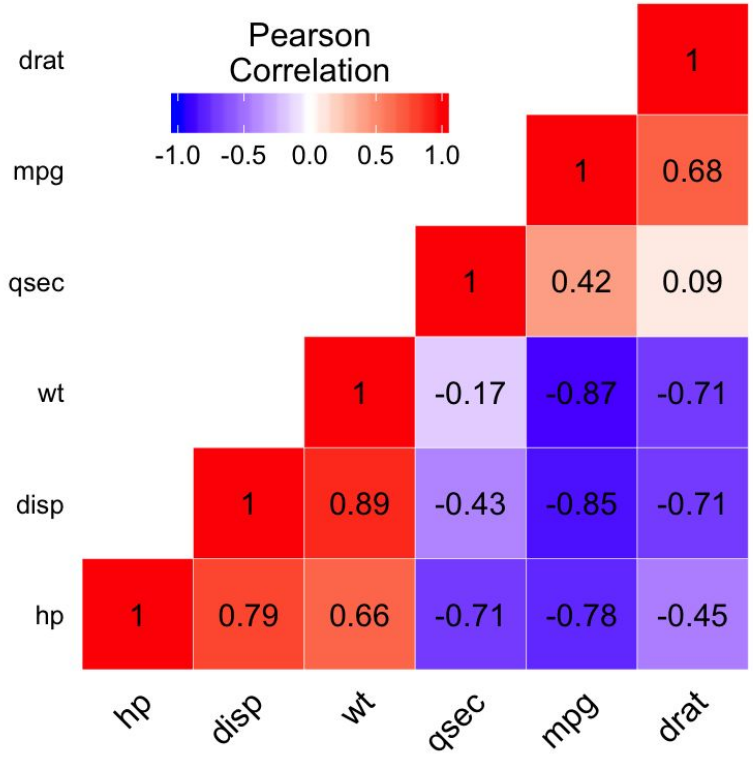
La correlación está acotada entre 1 y -1.

Siempre que la covarianza es positiva, la correlación es positiva y viceversa.

Mientras que la correlación no tiene unidades físicas, la covarianza sí.

Introducción a la estadística descriptiva

Medidas de asociación lineal entre variables: matriz de correlación



Tecnologías

Sistemas propietarios vs Open source

Fuentes de datos abiertas

Links

https://data.buenosaires.gob.ar/

http://datos.gob.ar/

https://www.gba.gob.ar/provinciaabierta

Para data science:

https://www.kaggle.com/datasets

https://archive.ics.uci.edu/ml/datasets.html

GRACIAS

Contacto
Docente: Leonardo Ignacio Córdoba

E-mail: cordoba.leonardoignacio@gmail.com

LinkedIn Leonardo Ignacio Córdoba