Estadistica para negocios

Modulo 1: Estadistica descriptiva univariada y bivariada

Dr. José Ignacio Hernández

Semana del 27 de mayo de 2024

Bienvenida al curso

Estadística para Negocios

- Entregar herramientas estadísticas para la toma de decisiones.

- En otras palabras: en base a los datos, guiar el proceso de toma de decisiones en la empresa.

Bienvenida al curso

Ejemplo: Explicando las ventas de una tienda de retail

- De 473 empleados de un retail:
  - Ventas promedio $1.634.861
  - Máximo ventas: $4.999.109
  - 34,9% tuvo una capacitación
  - 43 años de edad en promedio
  - 12,14 años educación promedio

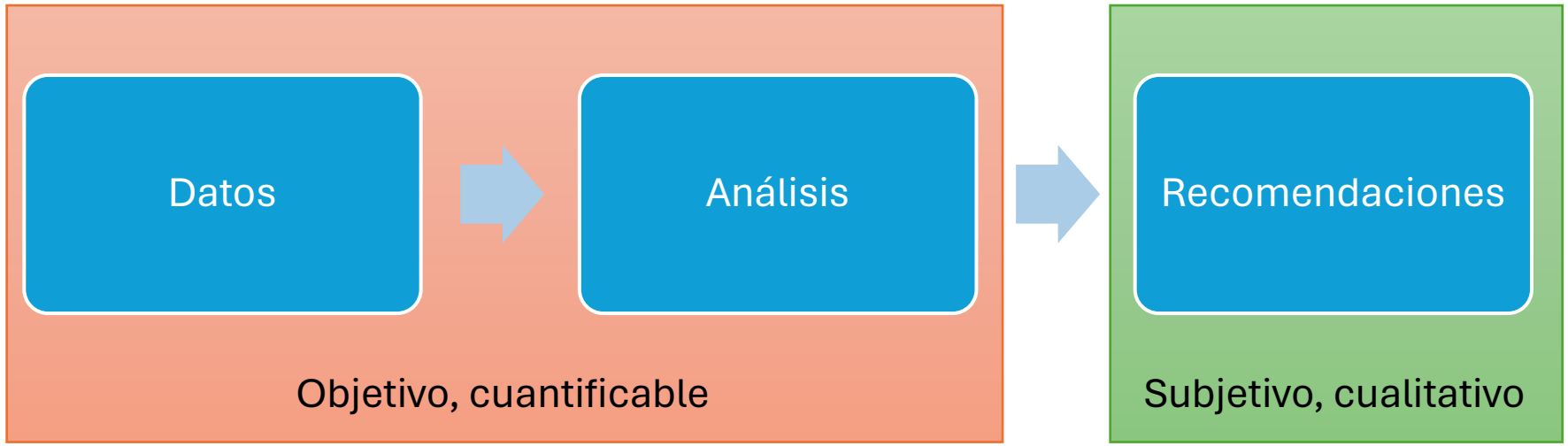
Bienvenida al curso

Ejemplo: Explicando las ventas de una tienda de retail

- ¿Qué implica esta información para la empresa?
  - Quienes venden más (menos)
  - ¿Cómo aumentar las ventas?

Bienvenida al curso

Para responder estas preguntas, durante el curso adoptaremos un enfoque sequencial:



Bienvenida al curso

Es muy importante tener desde un principio que las recomendaciones hechas son <u>subjetivas.</u>

- En otras palabras, existe un componente moral en ellas.

- <u>Por ejemplo:</u> si los trabajadores con más hijos son menos productivos debido a que deben cuidar de ellos…
  - <u>Analista 1 dice:</u> Se deben contratar trabajadores con menos hijos.
  - <u>Analista 2 dice:</u> Se debe proveer beneficio de sala cuna los trabajadores con hijos.

¿Quién está en lo correcto?

Bienvenida al curso

Resultados de aprendizaje:

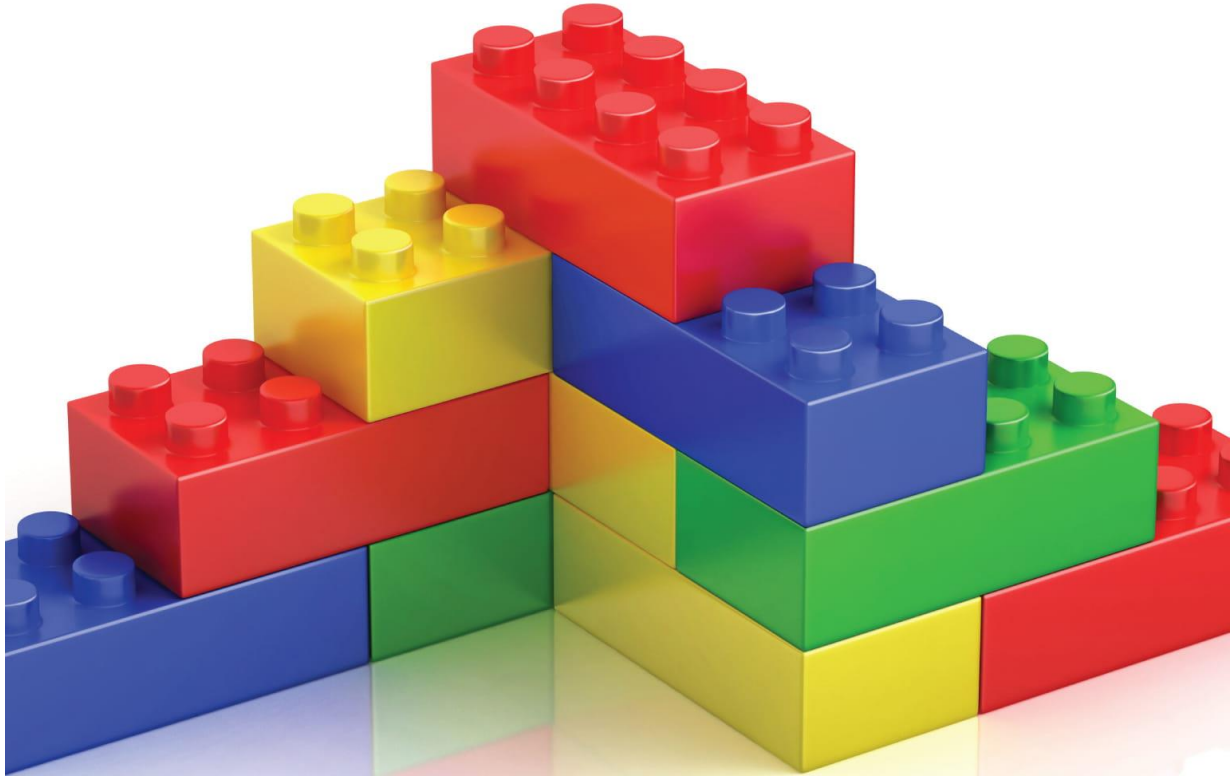
- RA1: Diseñar modelos de gestión que permitan la implementación de la estrategia en la organización.

- RA2: Analiza los datos existentes de la organización utilizando los sistemas de información adecuados.

Metodología

- Clases expositivas:
  - Dirigidas por el profesor
  - Revisión de conceptos
  - Uso de software estadistico

- Estudio de casos prácticos:
  - Trabajo individual o en grupos
  - Resolución de problemas aplicados.
  - Evaluación y feedback al final



Metodología

Modulo 1:

- Estadística descriptiva univariada y bivariada

Modulo 2:

- Fundamentos del cálculo de probabilidades de eventos

Modulo 3:

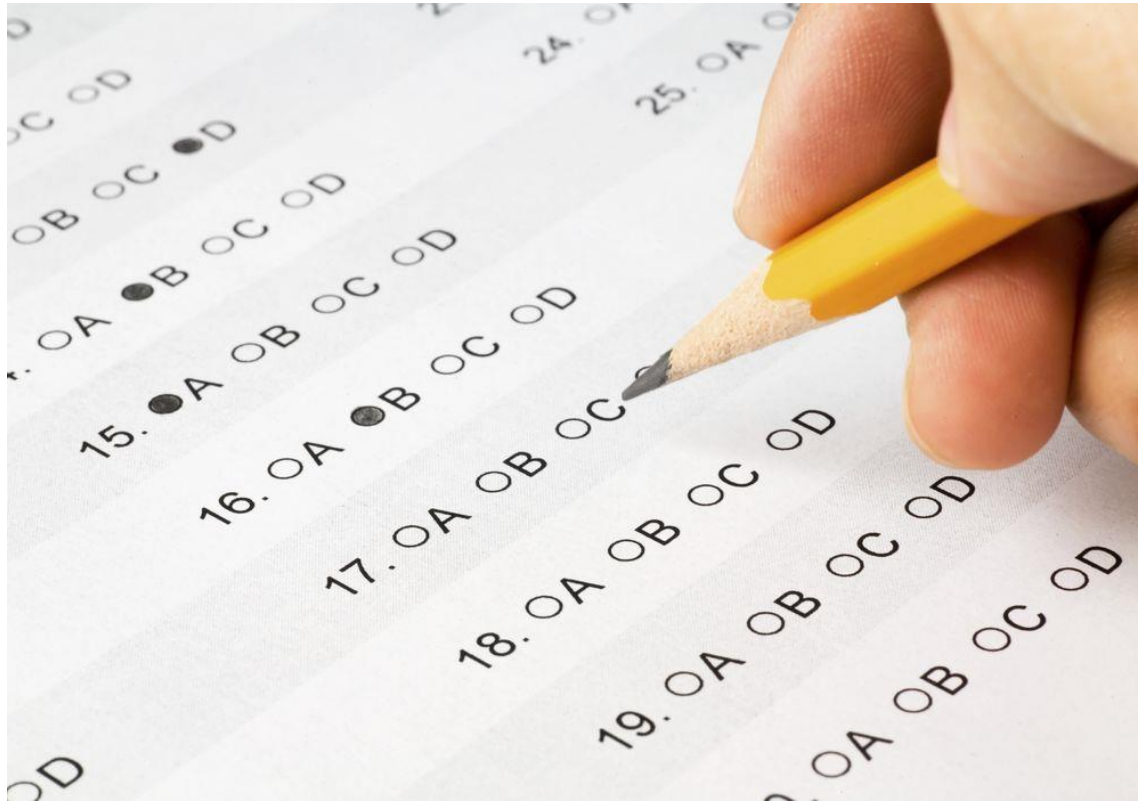
- Distribuciones muestrales

Modulo 4:

- Inferencia estadística

Evaluación

- Evaluación grupal (70%):
  - Parte 1 (30%): Se realiza y entrega durante la sesión PM del módulo 1.
  - Caso 2 (40%): Se entrega 1 semana después del modulo 2.
- Evaluación escrita (30%)
  - Caso aplicado individual
  - Se realizará en la sesión PM del módulo 4



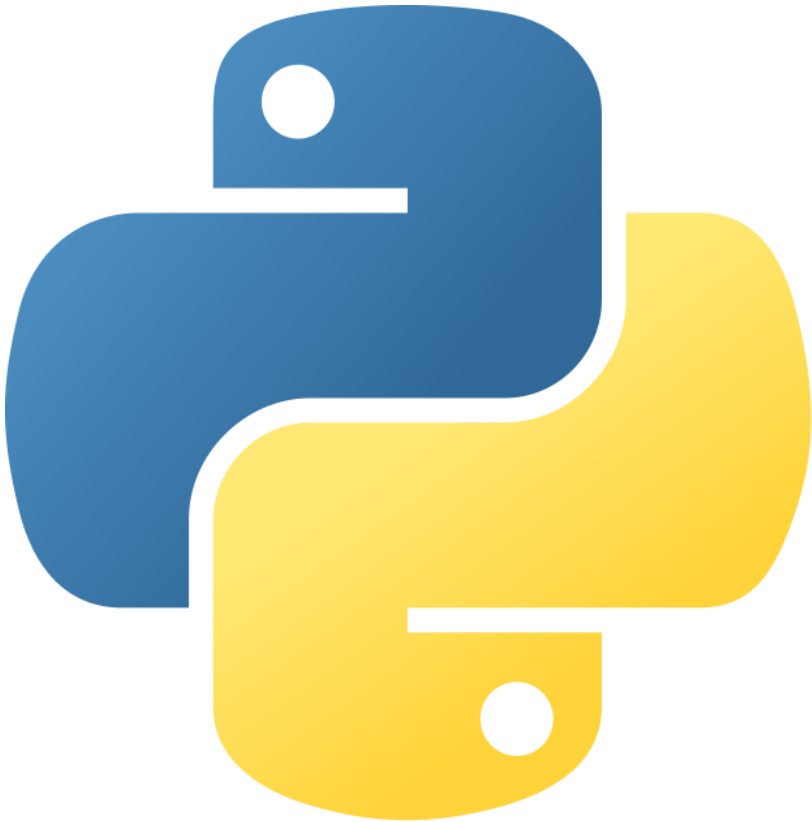
¿Me están siguiendo?

Uso de software

Durante este curso, usaremos Python

- Python es un lenguaje de programacion, con aplicaciones estadisticas.

- Hoy, es ampliamente utilizado:
  - Machine learning
  - Data science
  - Estadisticas en general



¿Por qué Python?

- **Ventajas:**
  - Software libre / sin costo alguno
  - Amplia disponibilidad de tutoriales / utilidades / documentacion
  - Respaldo de compañias que usan diariamente este software

- **Limitaciones:**
  - Complejidad: Por ser un lenguaje de programacion (puede ser resuelto)

Python es más fácil con Jupyter Notebooks!

Jupyter es un proyecto que permite el uso de Python a través de notebooks

- Un notebook es un archivo interactivo que combina texto y Código ejecutable.

- Ventajas de los notebooks:
  - Permite ejecutar codigo en tiempo real, paso a paso.
  - Permite hacer anotaciones claras para un uso más amigable.

Alternativas a Python

julia

IBM SPSS

STATA

Parte 0: Introducción a Python y Jupyter

Ejemplo: Familiarizandonos con Python

Manos a la obra: vamos a familiarizarnos con Python

- Abran el siguiente link en sus computadores:

- Luego, abran el archivo “notebook_0.ipynb” en la carpeta del Modulo 1.

- Vamos a ver algunos elementos fundamentales de Python, además de conocer su entorno

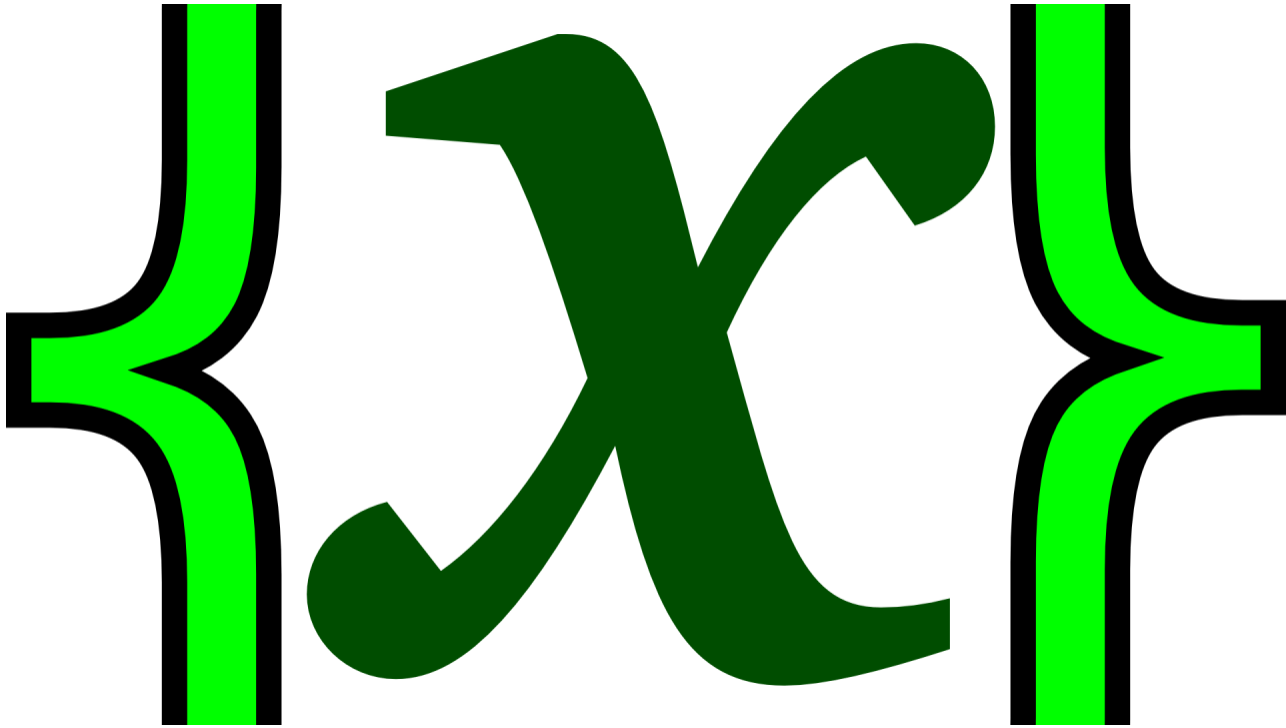
¿Preguntas?

Parte I: Estadística descriptiva univariada y bivariada

Variables

- Una **variable** es una magnitud que puede tomar distintos valores en distintos puntos de observación

- La estadística permite describir variables en base a una estructura.

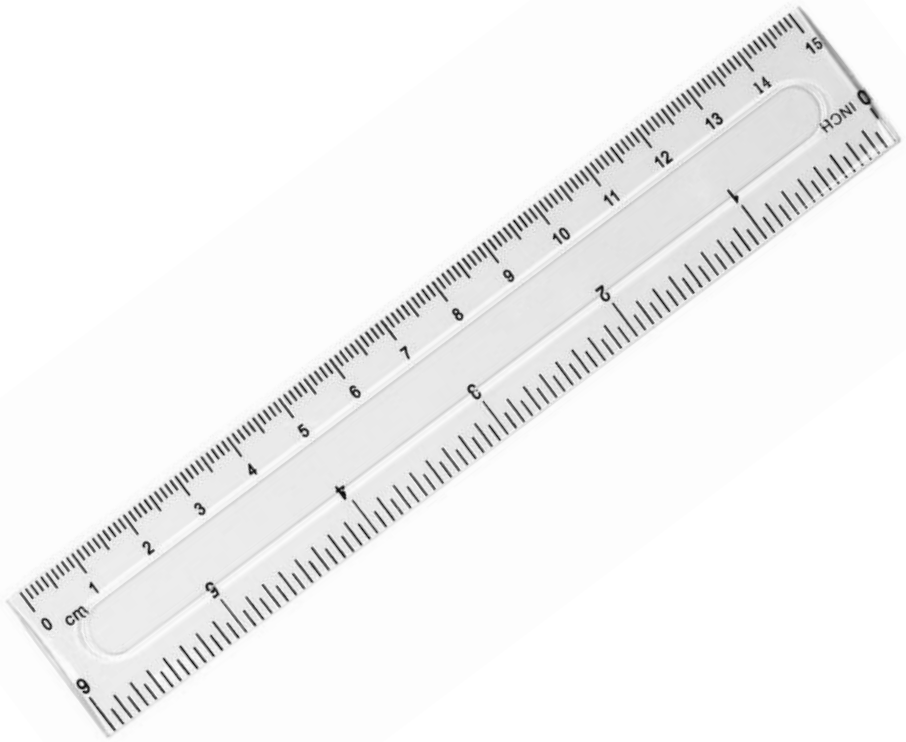


Variables discretas y continuas

Variable discreta:

- Se compone de un numero finito de numeros reales
- Ejemplos:
  - Numero de visitas al doctor
  - Genero
  - Numero de hijos

- **Variable continua:**
- Es un intervalo o la recta real completa
- Ejemplos:
  - PIB, consumo
  - Salarios
  - Temperatura del día



Variables

- **Asumiremos el rol del gerente de una tienda de retail de <u>mediano tamaño.</u>**

- En esta tienda, existe interés de aumentar las ventas totales.

- Si Ud. Fue contratado como analista de la tienda, ¿Qué variables debería sugerir observar?

Variables

Sobre las ventas, puede ser:

- Ventas por trabajador durante el periodo
- Ventas totales por periodo
- Ventas por trabajador, por periodo

Otras variables relevantes pueden ser:

- Edad de los trabajadores
- Genero de los trabajadores
- Nivel educacional
- Experiencia laboral
- Número de hijos

Práctica 1: Bases de datos en Python

Manos a la obra: Usaremos Python para explorar una base de datos, utilizando el modulo “Pandas”

- En nuestro Jupyter, abriremos el archivo “notebook_1.ipynb”

- Luego:
  - Revisaremos la estructura de la base de datos (filas y columnas)
  - Distinguiremos entre distintas variables
  - Revisaremos como crear nuevas variables en la base de datos

¿Preguntas?

Tabulación y presentación gráfica de variables unidimensionales

Frecuencia

- Frecuencia absoluta: Es el número de veces que el valor de una variable se encuentra presente en una base de datos.

- Frecuencia relativa: Es el número de veces que el valor se encuentra presente, con respecto al total de datos.

$$FR_i = \frac{\sum x_i}{N}$$

- Frecuencia porcentual: Es la frecuencia relativa, expresada en porcentaje:

$$FP_i = FR_i * 100\%$$

Frecuencia

Manos a la obra: Calcularemos la frecuencia en nuestra base de datos de ventas.

- **Problema:** La dirección de la empresa desea saber en qué edad y escolaridad se encuentran la mayoría de los trabajadores.

- **Posible estrategia:** Calcular la frecuencia de la variable de edad y escolaridad

- **Herramientas:** Base de datos + estadisticas

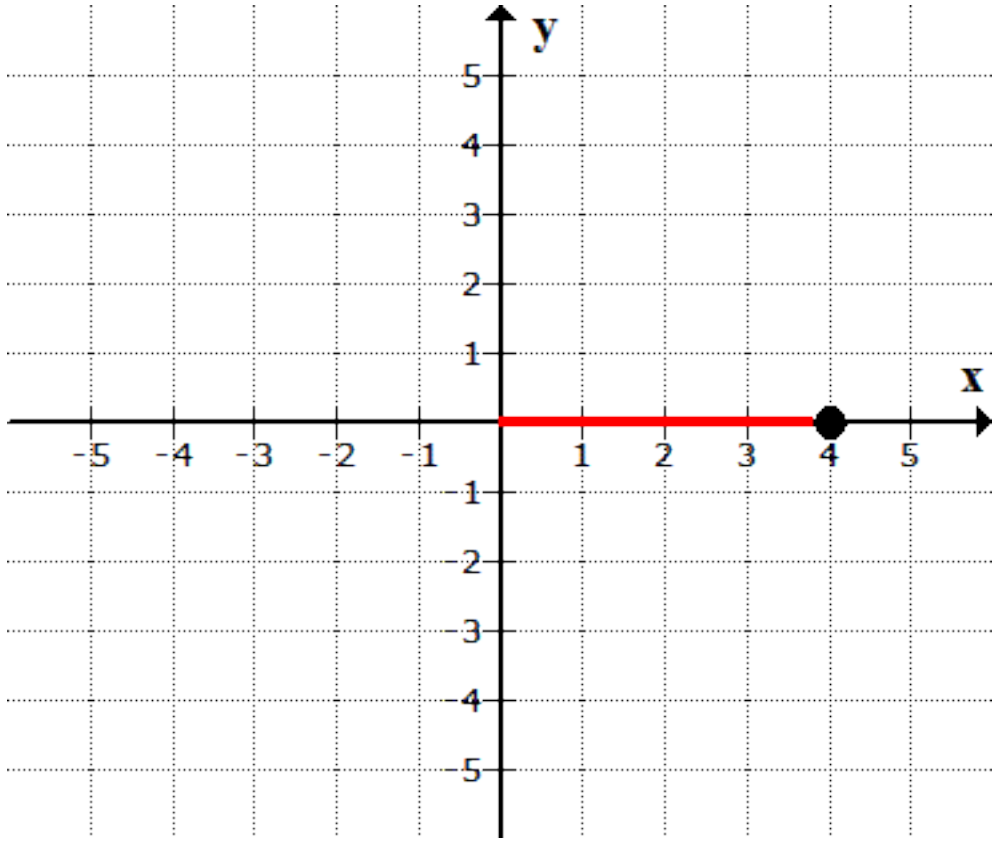
¿Preguntas?

Gráficos

- Además de las tablas de frecuencias, es posible presentar información de forma gráfica.

- <u>Ventajas:</u>
  - Facilidad de interpretación
  - Amigables con público en general

- <u>Desventajas:</u>
  - Riesgo de menor nivel de detalle
  - Posible pérdida de información



Gráficos

Algunos tipos de gráficos:

- Para frecuencias absolutas:
  - Gráficos de barra
  - Histogramas
- Para frecuencias relativas:
  - Gráficos de pastel

- Para comparar dos variables (los veremos más adelante)
  - Gráficos de puntos
  - Gráficos de lineas

Frecuencia

Manos a la obra: Calcularemos la frecuencia en nuestra base de datos de ventas.

- **Problema:** La dirección de la empresa tiene dos dudas:
  - ¿Cuál es la distribución de edad en la empresa?
  - ¿Qué proporción de personas capacitadas hay en la empresa?

- **Posible estrategia:** Mostrar gráficamente ambas variables

- **Herramientas:** Base de datos + Python

¿Preguntas?

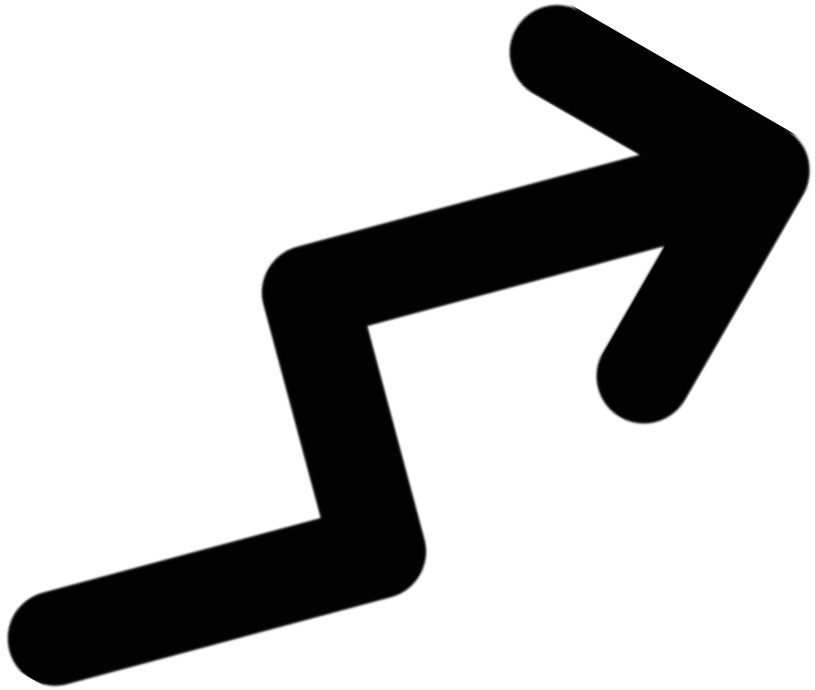
Estadística descriptiva

Medidas de tendencia central

Propósito: Describir variables en razón de una medida de tendencia.

- Permite identificar patrones o repeticiones de los datos en una variable especifica.

- **En particular, veremos dos medidas de tendencia central:**
  - Media
  - Mediana



Medidas de tendencia central

Media: Corresponde al promedio de los datos en una variable. Su formula es:

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

Suma de los valores

Número de observaciones

- **Uso:** Conocer la tendencia promedio de una variable especifica.
  - Ejemplo: ¿Cuál es la edad promedio en la empresa? ¿Cuál es la venta promedio del periodo?

- Ventajas:
  - Simple interpretación

- Desventajas:
  - Sensible a datos atípicos (*outliers*)

Medidas de tendencia central

Mediana: Corresponde al dato del medio de una variable previamente ordenada.

- La mediana puede variar si el número de observaciones es:
  - Impar: La mediana es el dato ubicado en el medio
  - Par: La mediana es el promedio entre los dos datos en el medio.

- **Uso:** Conocer dónde está el 50% de los individuos en una muestra
  - **Ejemplo:** El 50% de los trabajadores vendieron más de $XX.-

Medidas de tendencia central

- Ventajas de la mediana:
  - Menor sensibilidad a datos atípicos que la media

- Desventajas de la mediana:
  - Información limitada.

Medidas de tendencia central

Manos a la obra: Calcularemos la media y la mediana de cada variable en nuestra base de datos de retail

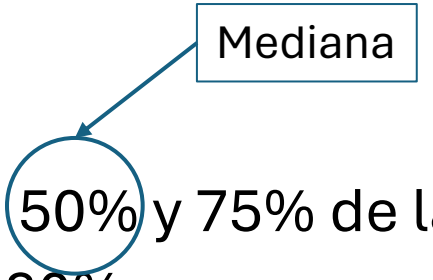
- **Problema:** La dirección de la empresa tiene dos dudas:
  - ¿Cuál es la media y la mediana de las ventas?
  - Con esa información ¿qué implica que la mediana de las ventas esté por debajo/sobre la media?

- **Posible estrategia:** Calcular e interpretar

- **Herramientas:** Base de datos + Python

Medidas de posición

Cuantiles: permiten determinar la distribución de los datos, una vez ordenados:



- Los cuantiles pueden clasificarse en:
  - Cuartiles: Determinar dónde está el 25%, 50% y 75% de la población
  - Quintiles: Dónde está el 20%, 40%, 60%, 80%...
  - Deciles: 10%, 20%, 30%...

- **Uso:** Conocer dónde se ubican porcentajes específicos de la población

Medidas de posición

Manos a la obra: Calcularemos la distribución de los datos

- **Problema:** La dirección de la empresa quiere saber:
  - ¿Cuáles son los cuartiles de ventas?
  - ¿Cómo se ve la distribución de las ventas de forma gráfica?

- **Posible estrategia:** Calcular cuantiles y un histograma.

- **Herramientas:** Base de datos + Python

Medidas de dispersión

Varianza (muestral): Es una medida de la variabilidad de los datos con respecto a su media:

$$\sigma^2 = \frac{\sum_i (x_i - \bar{X})^2}{N - 1}$$

Notar que el término cuadrático hace que la varianza no pueda ser interpretada en relación a los datos

Desviación estándar: Es la raíz cuadrada de la varianza, lo cual lo hace interpretable:

$$\sigma = \sqrt{\frac{\sum_i (x_i - \bar{X})^2}{N - 1}}$$

Medidas de dispersión

Manos a la obra: Calcularemos la desviación estandar de los datos

- **Problema:** La dirección de la empresa quiere saber la desviación promedio de los datos.

- **Posible estrategia:** Calcular desviación estándar

- **Herramientas:** Base de datos + Python

¿Preguntas?

Tabulación de variables bidimensionales

Análisis de variables bidimensionales

Analizar variables en dos dimensiones permiten identificar patrones o comportamientos entre ellas.

- Ejemplo: La empresa de retail desea saber si:
  - Trabajadores con mayor edad tienen mayor o menor escolaridad.
  - Trabajadores capacitados venden más o menos.

Tabla de frecuencias en dos dimensiones

Propósito: Conocer la frecuencia (absoluta o relativa) en dos variables:

- Las filas representan los elementos de variable 1
- Las columnas representan los elementos de variable 2
- Las celdas representan la frecuencia entre dos variables.

	A	B	C
X			
Y			

Frecuencia en dos dimensiones

Manos a la obra: Calcularemos una tabla de frecuencias absolutas en dos dimensiones

- **Problema:** La empresa se pregunta:
  - ¿Qué edad y escolaridad tiene la mayor cantidad de trabajadores?
  - ¿En qué tramo de edad está la mayor cantidad de trabajadores capacitados?

- **Posible estrategia:** Calcular tablas de frecuencias

- **Herramientas:** Base de datos + Python

¿Preguntas?

Medidas de correlación

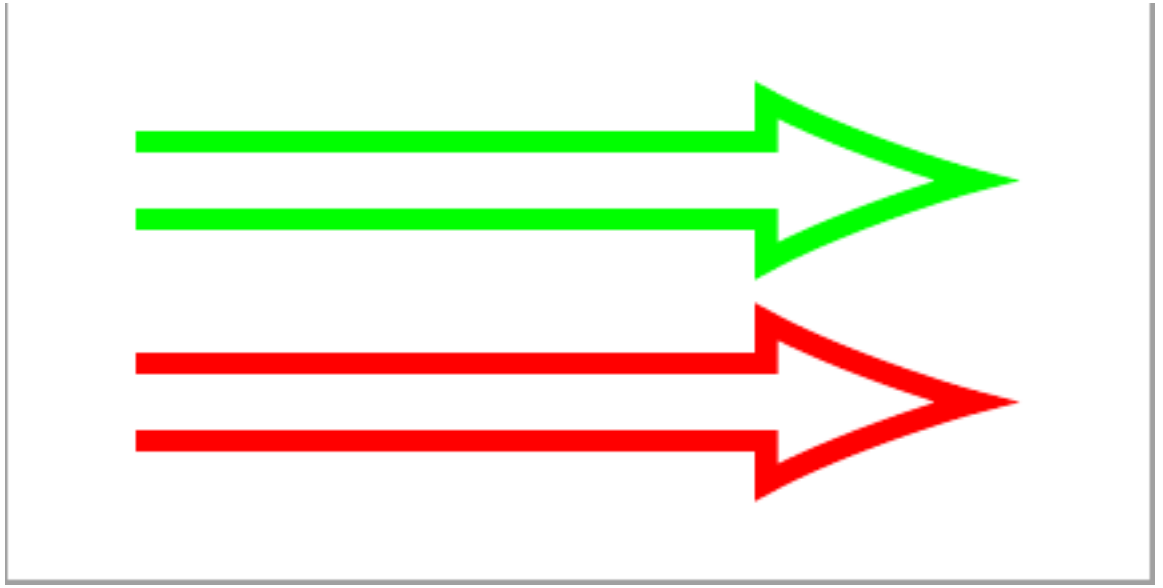
Covarianza: Permite conocer el grado de variabilidad y comportamiento entre dos variables.

Matemáticamente:

$$Cov(X,Y) = \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{N}$$

Limitación:

- La varianza no tiene una interpretación en sí misma

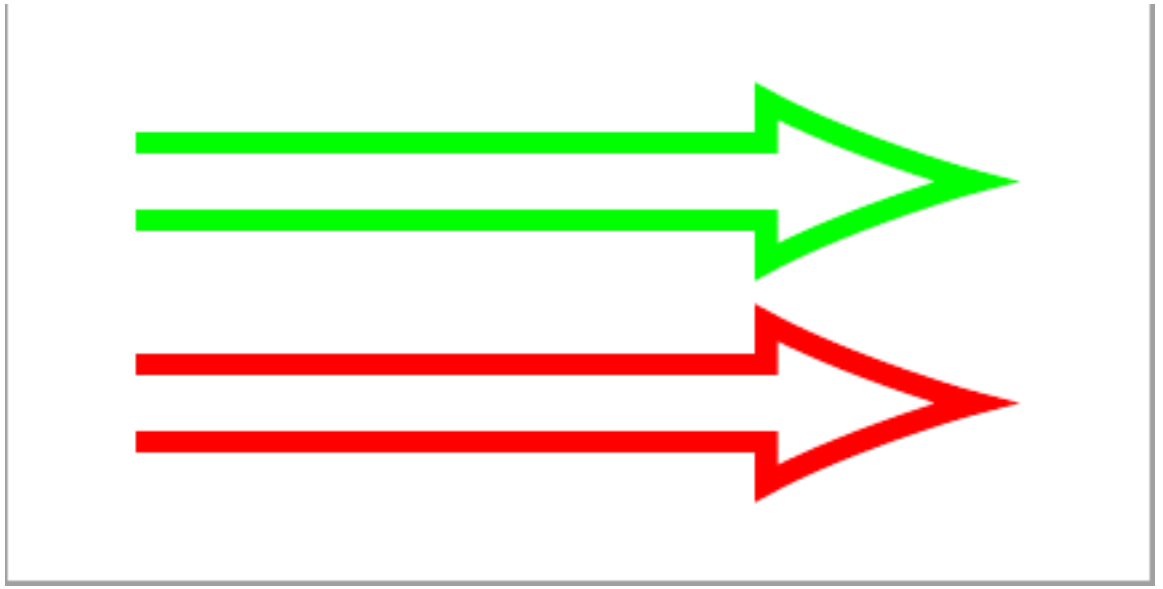


Medidas de correlación

Coeficiente de correlación de Pearson: Permite conocer el grado de asociación lineal entre dos variables

Matemáticamente:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} \in [-1,+1]$$



Medidas de correlación

Interpretación:

$-1 < \rho_{XY} < 0$

X e Y tienen una relación lineal inversa.

$\rho_{XY} = 0 < \rho_{XY} < 1$

X e Y tienen una relación lineal directa.

$\rho_{XY} = -1$

X e Y tienen una relación lineal inversa y perfecta.

$\rho_{XY} = 0$

X e Y no tienen relación lineal

$\rho_{XY} = 1$

X e Y tienen una relación lineal directa y perfecta.

Frecuencia en dos dimensiones

Manos a la obra: Calcularemos la correlación entre las variables de la base de datos

- **Problema:** La empresa se pregunta:
  - ¿Cuál es la asociación entre tener capacitación y las ventas?
  - ¿Cuál es la asociación entre la edad de los trabajadores y las ventas?

- **Posible estrategia:** Calcular coeficientes de correlación

- **Herramientas:** Base de datos + Python

¿Preguntas?

Resumen

- Hoy vimos distintas formas de analizar datos de forma descriptiva
  - Frecuencias
  - Medidas de tendencia central, posición, dispersión
  - Correlación entre variables.

- Además, estudiamos cómo estas medidas pueden proveer de información a la empresa a través de un caso aplicado.

- **Sin embargo, lo que hemos visto solo describe a una parte de la población: la muestra con que contamos.**

- **Es posible hacer inferencia de la población usando la muestra. Para ello, requerimos el uso de la teoría de la probabilidad, lo cual veremos en el próximo módulo**

¡Gracias!