Análisis de Datos en Python: descriptivo e inferencial 1

Dora Suárez, Juan F. Pérez

Departamento MACC
Matemáticas Aplicadas y Ciencias de la Computación
Universidad del Rosario

juanferna.perez@urosario.edu.co

2018

Contenidos

Datos: una característica de una población

Análisis descriptivo de datos

- Tomamos datos respecto a una característica de una población

Análisis descriptivo de datos

- Tomamos datos respecto a una característica de una población
- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso

Análisis descriptivo de datos

- Tomamos datos respecto a una característica de una población
- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- Población: los estudiantes de este curso

Análisis descriptivo de datos

- Tomamos datos respecto a una característica de una población
- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- Población: los estudiantes de este curso
- Característica: número de tazas de café que tomaron ayer

Análisis descriptivo de datos

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso

Análisis descriptivo de datos

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $2, 0, 1, \ldots$

Análisis descriptivo de datos

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $2, 0, 1, \ldots$
- Medidas descriptivas:
  - Mínimo, máximo

Análisis descriptivo de datos

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $2, 0, 1, \ldots$
- Medidas descriptivas:
  - Mínimo, máximo
  - Promedio

Análisis descriptivo de datos

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $2, 0, 1, \ldots$
- Medidas descriptivas:
  - Mínimo, máximo
  - Promedio
  - Variabilidad alrededor de la media

Análisis descriptivo de datos

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $2, 0, 1, \ldots$
- Medidas descriptivas:
  - Mínimo, máximo
  - Promedio
  - Variabilidad alrededor de la media
  - Frecuencia (absoluta, relativa)

Análisis descriptivo de datos (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso

Análisis descriptivo de datos (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $[0, 1, 0, 2, 3, 0, 1, 2, 0, 2, 2, 2, 1, 0, 3, 4]$

Análisis descriptivo de datos (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $[0, 1, 0, 2, 3, 0, 1, 2, 0, 2, 2, 2, 1, 0, 3, 4]$
  - Número de observaciones: 16

Análisis descriptivo de datos (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $[0, 1, 0, 2, 3, 0, 1, 2, 0, 2, 2, 2, 1, 0, 3, 4]$
  - Número de observaciones: 16
  - Mínimo: 0

Análisis descriptivo de datos (ejemplo)

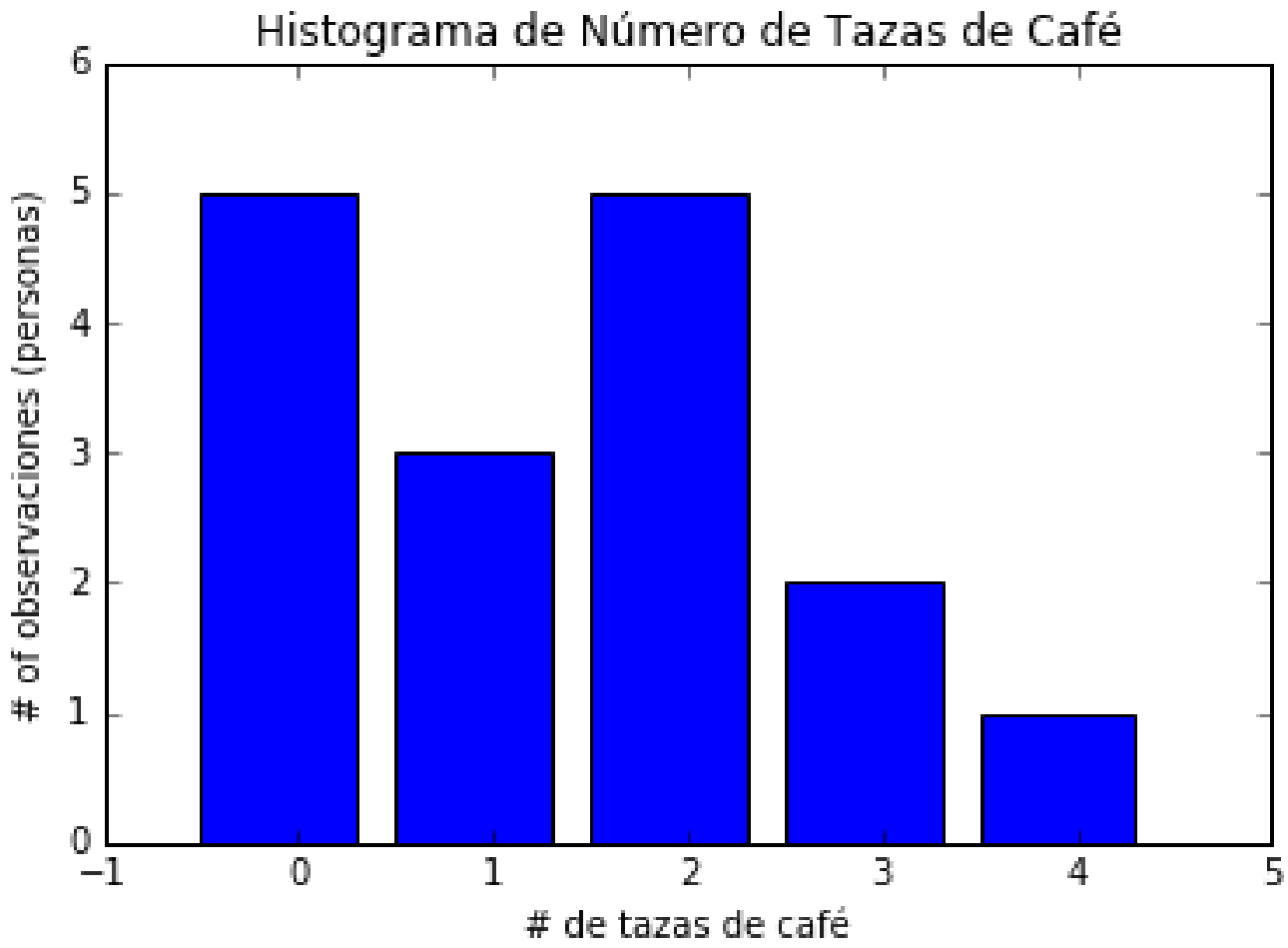
- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $[0, 1, 0, 2, 3, 0, 1, 2, 0, 2, 2, 2, 1, 0, 3, 4]$
  - Número de observaciones: 16
  - Mínimo: 0
  - Máximo: 4

Análisis descriptivo de datos (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso
- $[0, 1, 0, 2, 3, 0, 1, 2, 0, 2, 2, 2, 1, 0, 3, 4]$
  - Número de observaciones: 16
  - Mínimo: 0
  - Máximo: 4
  - Promedio: 1.4375

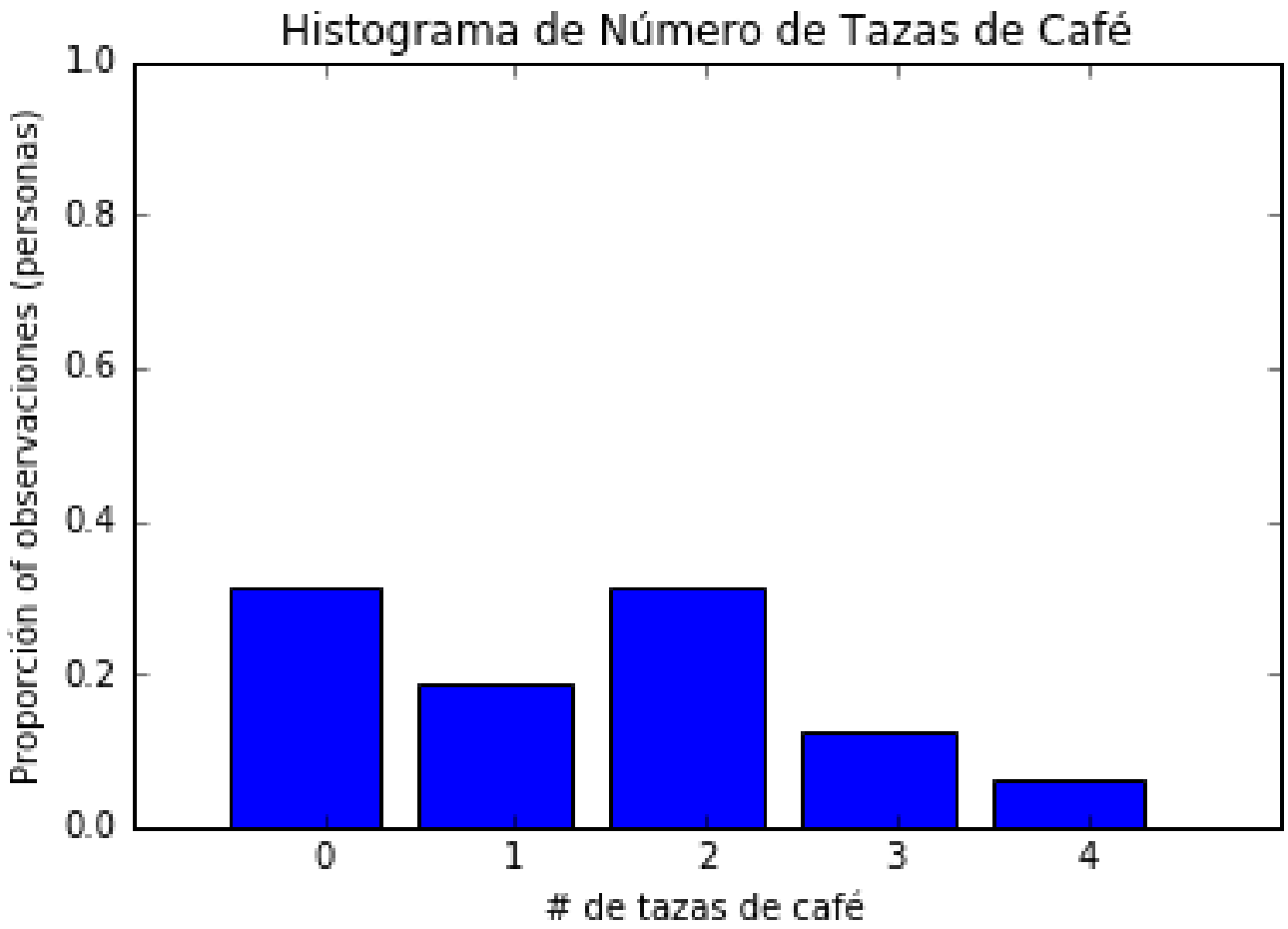
Análisis descriptivo de datos (ejemplo)

Frecuencia absoluta:



Análisis descriptivo de datos (ejemplo)

Frecuencia relativa:



Generalizando a cualquier conjunto de datos

- Datos unidimensionales: una sola característica

Generalizando a cualquier conjunto de datos

- Datos unidimensionales: una sola característica
- Número de datos: N

Generalizando a cualquier conjunto de datos

- Datos unidimensionales: una sola característica
- Número de datos: N
- **Datos**: $[x_1, x_2, \ldots, x_N]$

Generalizando a cualquier conjunto de datos

- Datos unidimensionales: una sola característica
- Número de datos: N
- **Datos**: $[x_1, x_2, \ldots, x_N]$
- **Mínimo**: $\text{mín}_i\{x_i\}$

Generalizando a cualquier conjunto de datos

- Datos unidimensionales: una sola característica
- Número de datos: N
- **Datos**: $[x_1, x_2, \ldots, x_N]$
- **Mínimo**: $\text{mín}_i\{x_i\}$
- **Máximo**: $\text{máx}_i\{x_i\}$

Generalizando a cualquier conjunto de datos

- Datos unidimensionales: una sola característica
- Número de datos: N
- **Datos**: $[x_1, x_2, \ldots, x_N]$
- **Mínimo**: $\text{mín}_i\{x_i\}$
- **Máximo**: $\text{máx}_i\{x_i\}$
- **Promedio**:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Generalizando a cualquier conjunto de datos

- **Datos**: $[x_1, x_2, \ldots, x_N]$

Generalizando a cualquier conjunto de datos

- **Datos**: $[x_1, x_2, \ldots, x_N]$
- **Frecuencia absoluta**:
  número de veces que aparece j en las observaciones

$$f_j = \#\{x_i = j\}$$

Generalizando a cualquier conjunto de datos

- **Datos**: $[x_1, x_2, \ldots, x_N]$
- **Frecuencia absoluta**:
  número de veces que aparece j en las observaciones

$$f_j = \#\{x_i = j\}$$

- **Frecuencia relativa**:
  proporción del veces que aparece j en las observaciones

$$p_j = \frac{f_j}{N}$$

Variable aleatoria

Variable aleatoria

- X: número de tazas de café que tomó ayer un estudiante de este curso

Variable aleatoria

- X: número de tazas de café que tomó ayer un estudiante de este curso
- Seleccionamos un estudiante al azar, ¿cuál es la probabilidad de que el estudiante haya tomado 3 tazas de café?

Variable aleatoria

- X: número de tazas de café que tomó ayer un estudiante de este curso
- Seleccionamos un estudiante al azar, ¿cuál es la probabilidad de que el estudiante haya tomado 3 tazas de café?
-

$$P(X = 3) = p_3 = \frac{2}{16} = 0{,}125$$

Variable aleatoria

- Se realiza un **experimento aleatorio**

Variable aleatoria

- Se realiza un **experimento aleatorio**
- **Espacio muestral**: conjunto de todos los posibles resultados

Variable aleatoria

- Se realiza un **experimento aleatorio**
- **Espacio muestral**: conjunto de todos los posibles resultados
- **Variable aleatoria**: función del espacio muestral en los números reales (asignamos valores numéricos al resultado del experimento)

Variable aleatoria (ejemplo)

- **Experimento aleatorio**: los estudiantes de este curso toman un número de tazas de café

Variable aleatoria (ejemplo)

- **Experimento aleatorio**: los estudiantes de este curso toman un número de tazas de café
- **Espacio muestral**: todas las posibles combinaciones de número de tazas de café consumidas por cada estudiante

Variable aleatoria (ejemplo)

- **Experimento aleatorio**: los estudiantes de este curso toman un número de tazas de café
- **Espacio muestral**: todas las posibles combinaciones de número de tazas de café consumidas por cada estudiante
- **Variable aleatoria**: el número de tazas que consumió un estudiante seleccionado al azar

Variable aleatoria discreta

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso

Variable aleatoria discreta

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso
- **Discreta**: toma valores en un conjunto finito (o contable)

Variable aleatoria discreta

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso
- **Discreta**: toma valores en un conjunto finito (o contable)
- $X \in \{0, 1, 2, 3, \dots\}$

Variable aleatoria discreta

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso
- **Discreta**: toma valores en un conjunto finito (o contable)
- $X \in \{0, 1, 2, 3, \dots\}$
- Función de masa de probabilidad:

$$p_j = P(X = j)$$

Variable aleatoria discreta

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso
- **Discreta**: toma valores en un conjunto finito (o contable)
- $X \in \{0, 1, 2, 3, \dots\}$
- Función de masa de probabilidad:

$$p_j = P(X = j)$$

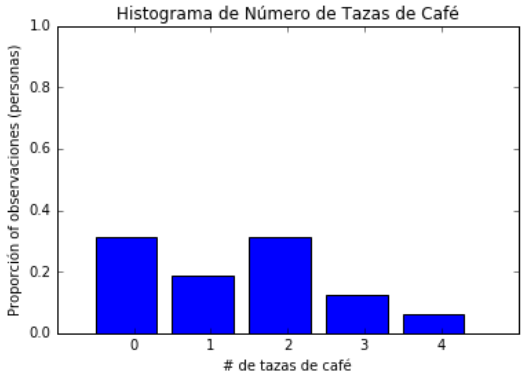
- p_j: probabilidad de que X sea igual a j

Variable aleatoria discreta (ejemplo)

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso

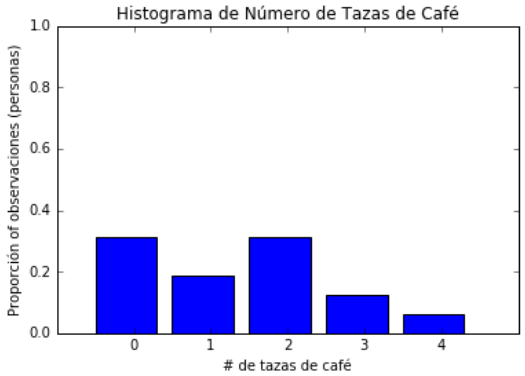
Variable aleatoria discreta (ejemplo)

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso



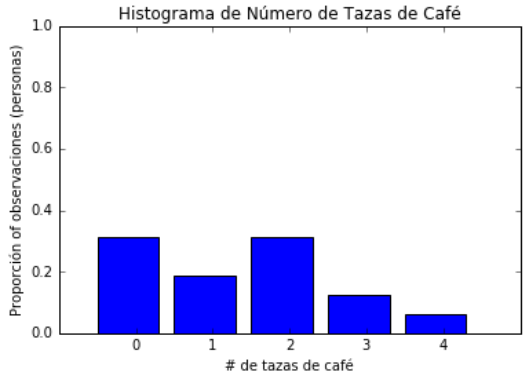
Variable aleatoria discreta (ejemplo)

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso



Variable aleatoria discreta (ejemplo)

- **Variable aleatoria** X: número de tazas de café que tomó ayer un estudiante de este curso



- Función de masa de probabilidad:

$$P(X = j) = \begin{cases} 5/16, & j = 0, \\ 3/16, & j = 1, \\ 5/16, & j = 2, \\ 2/16, & j = 3, \\ 1/16, & j = 4, \end{cases}$$

Datos: una característica de una muestra

Muestras

- Tomamos datos respecto a una característica de una población

Muestras

- Tomamos datos respecto a una característica de una población
- Ejemplo: número de tazas de café que tomaron ayer los habitantes del país

Muestras

- Tomamos datos respecto a una característica de una población
- Ejemplo: número de tazas de café que tomaron ayer los habitantes del país
- Población: los habitantes del país

Muestras

- Tomamos datos respecto a una característica de una población
- Ejemplo: número de tazas de café que tomaron ayer los habitantes del país
- Población: los habitantes del país
- Característica: número de tazas de café que tomaron ayer

Muestras

- Es infactible, muy costoso o incluso imposible medir la característica en todos los elementos de la población

Muestras

- Es infactible, muy costoso o incluso imposible medir la característica en todos los elementos de la población
- Ejemplos:

Muestras

- Es infactible, muy costoso o incluso imposible medir la característica en todos los elementos de la población
- Ejemplos:
  - Grandes poblaciones

Muestras

- Es infactible, muy costoso o incluso imposible medir la característica en todos los elementos de la población
- Ejemplos:
  - Grandes poblaciones
  - Características cambiantes

Muestras

- Es infactible, muy costoso o incluso imposible medir la característica en todos los elementos de la población
- Ejemplos:
  - Grandes poblaciones
  - Características cambiantes
  - Pruebas destructivas

Muestras

- Es infactible, muy costoso o incluso imposible medir la característica en todos los elementos de la población
- Ejemplos:
  - Grandes poblaciones
  - Características cambiantes
  - Pruebas destructivas
- **Muestra**: selección de un número limitado de observaciones de la población (no todas)

Muestra aleatoria

- Muestra debe ser representativa de la población

Muestra aleatoria

- Muestra debe ser representativa de la población
- Ejemplo: los estudiantes de este curso no son una muestra representativa

Muestra aleatoria

- Muestra debe ser representativa de la población
- Ejemplo: los estudiantes de este curso no son una muestra representativa
- Selección al azar: todos los miembros del a población tienen la misma probabilidad de ser seleccionados

Muestra aleatoria

- Muestra debe ser representativa de la población
- Ejemplo: los estudiantes de este curso no son una muestra representativa
- Selección al azar: todos los miembros del a población tienen la misma probabilidad de ser seleccionados
- Si la característica poblacional es una variable aleatoria X ...

Muestra aleatoria

- Muestra debe ser representativa de la población
- Ejemplo: los estudiantes de este curso no son una muestra representativa
- Selección al azar: todos los miembros del a población tienen la misma probabilidad de ser seleccionados
- Si la característica poblacional es una variable aleatoria X ...
- ... obtenemos una muestra aleatoria de tamaño n: $X_1, \ldots, X_n$

Muestra aleatoria (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso (población de tamaño 16)

Muestra aleatoria (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso (población de tamaño 16)
- X: número de tazas de café que tomó ayer un estudiante de este curso

$$P(X = j) = \begin{cases} 5/16, & j = 0, \\ 3/16, & j = 1, \\ 5/16, & j = 2, \\ 2/16, & j = 3, \\ 1/16, & j = 4, \end{cases}$$

Muestra aleatoria (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso (población de tamaño 16)
- X: número de tazas de café que tomó ayer un estudiante de este curso

$$P(X = j) = \begin{cases} 5/16, & j = 0, \\ 3/16, & j = 1, \\ 5/16, & j = 2, \\ 2/16, & j = 3, \\ 1/16, & j = 4, \end{cases}$$

- Tomamos una muestra aleatoria de tamaño 3: X_1, X_2, X_3

Muestra aleatoria (ejemplo)

- Ejemplo: número de tazas de café que tomaron ayer los estudiantes de este curso (población de tamaño 16)
- X: número de tazas de café que tomó ayer un estudiante de este curso

$$P(X = j) = \begin{cases} 5/16, & j = 0, \\ 3/16, & j = 1, \\ 5/16, & j = 2, \\ 2/16, & j = 3, \\ 1/16, & j = 4, \end{cases}$$

- Tomamos una muestra aleatoria de tamaño 3: X_1, X_2, X_3
- Cada una de las tres observaciones tiene la misma *distribución* de X

Muestra aleatoria: inferencias

- A partir de la **muestra** buscamos **inferir** características de la **población**

Muestra aleatoria: inferencias

- A partir de la **muestra** buscamos **inferir** características de la **población**
- Muestra es una visión parcial de la población: incertidumbre

Muestra aleatoria: inferencias

- A partir de la **muestra** buscamos **inferir** características de la **población**
- Muestra es una visión parcial de la población: incertidumbre
- Al aumentar el tamaño de la muestra (aleatoria) la incertidumbre debe reducirse

Inferencias a partir de muestras aleatorias

- Característica de la **Población**: variable aleatoria X

Objetivo: a partir de estas cantidades medidas sobre la muestra inferir el comportamiento de la población

Inferencias a partir de muestras aleatorias

- Característica de la **Población**: variable aleatoria X
- **Muestra aleatoria** de tamaño n: $X_1, \ldots X_n$

Objetivo: a partir de estas cantidades medidas sobre la muestra inferir el comportamiento de la población

Inferencias a partir de muestras aleatorias

- Característica de la **Población**: variable aleatoria X
- **Muestra aleatoria** de tamaño n: $X_1, \ldots X_n$
- **Datos**: $[x_1, x_2, \ldots, x_n]$

Objetivo: a partir de estas cantidades medidas sobre la muestra inferir el comportamiento de la población

Inferencias a partir de muestras aleatorias

- Característica de la **Población**: variable aleatoria X
- **Muestra aleatoria** de tamaño n: $X_1, \ldots X_n$
- **Datos**: $[x_1, x_2, \ldots, x_n]$
- Mínimo, Máximo

Objetivo: a partir de estas cantidades medidas sobre la muestra inferir el comportamiento de la población

Inferencias a partir de muestras aleatorias

- Característica de la **Población**: variable aleatoria X
- **Muestra aleatoria** de tamaño n: $X_1, \ldots X_n$
- **Datos**: $[x_1, x_2, \ldots, x_n]$
- Mínimo, Máximo
- Promedio

Objetivo: a partir de estas cantidades medidas sobre la muestra inferir el comportamiento de la población

Inferencias a partir de muestras aleatorias

- Característica de la **Población**: variable aleatoria X
- **Muestra aleatoria** de tamaño n: $X_1, \ldots X_n$
- **Datos**: $[x_1, x_2, \ldots, x_n]$
- Mínimo, Máximo
- Promedio
- Frecuencia absoluta

Objetivo: a partir de estas cantidades medidas sobre la muestra inferir el comportamiento de la población

Inferencias a partir de muestras aleatorias

- Característica de la **Población**: variable aleatoria X
- **Muestra aleatoria** de tamaño n: $X_1, \ldots X_n$
- **Datos**: $[x_1, x_2, \ldots, x_n]$
- Mínimo, Máximo
- Promedio
- Frecuencia absoluta
- Frecuencia relativa

Objetivo: a partir de estas cantidades medidas sobre la muestra inferir el comportamiento de la población

Un primer código en Python

¿Qué es Python?

Respuestas...

¿Qué es Python?

Respuestas...

- Un lenguaje de programación...

¿Qué es Python?

Respuestas...

- Un lenguaje de programación...
- ... de alto nivel.

¿Qué es Python?

Respuestas...

- Un lenguaje de programación...
- ... de alto nivel.
- ... imperativo.

Algunas características de Python

Algunas características de Python

- Requiere pocas líneas de código comparado con otros lenguajes (Java o C++)

Algunas características de Python

- Requiere pocas líneas de código comparado con otros lenguajes (Java o C++)
- Usa indentación (sangrado) para separar bloques de código

Algunas características de Python

- Requiere pocas líneas de código comparado con otros lenguajes (Java o C++)
- Usa indentación (sangrado) para separar bloques de código
- Simple pero poderoso, usado en proyectos de gran escala

Algunas características de Python

- Requiere pocas líneas de código comparado con otros lenguajes (Java o C++)
- Usa indentación (sangrado) para separar bloques de código
- Simple pero poderoso, usado en proyectos de gran escala
- Usado en muchas áreas científicas (e.g., aprendizaje de máquina)

Algo más de historia

- Primer desarrollo en 1989 (Guido Van Rossum, Centrum Wiskunde en Informatica, Amsterdam)

Algo más de historia

- Primer desarrollo en 1989 (Guido Van Rossum, Centrum Wiskunde en Informatica, Amsterdam)
- Python 2.0: lanzado en 2000

Algo más de historia

- Primer desarrollo en 1989 (Guido Van Rossum, Centrum Wiskunde en Informatica, Amsterdam)
- Python 2.0: lanzado en 2000
- Python 3.0: lanzado en 2008

Algo más de historia

- Primer desarrollo en 1989 (Guido Van Rossum, Centrum Wiskunde en Informatica, Amsterdam)
- Python 2.0: lanzado en 2000
- Python 3.0: lanzado en 2008
  - No es compatible con la versión 2.X
  - Muchos paquetes aún funcionan con la versión 2.X

Algo más de historia

- Primer desarrollo en 1989 (Guido Van Rossum, Centrum Wiskunde en Informatica, Amsterdam)
- Python 2.0: lanzado en 2000
- Python 3.0: lanzado en 2008
  - No es compatible con la versión 2.X
  - Muchos paquetes aún funcionan con la versión 2.X
- Nosotros trabajaremos con la versión 2.X (2.7) pero las diferencias con las 3.X son mínimas en nuestro código.

Iniciar en Python

- Python website: `https://www.python.org/`

Iniciar en Python

- Python website: `https://www.python.org/`
  - Distribución estándar de Python 2 y 3

Iniciar en Python

- Python website: `https://www.python.org/`
  - Distribución estándar de Python 2 y 3
  - Editor básico: IDLE

Iniciar en Python

- Python website: `https://www.python.org/`
  - Distribución estándar de Python 2 y 3
  - Editor básico: IDLE
- Distribución Anaconda: `https://www.anaconda.com/download/`

Iniciar en Python

- Python website: `https://www.python.org/`
  - Distribución estándar de Python 2 y 3
  - Editor básico: IDLE
- Distribución Anaconda: `https://www.anaconda.com/download/`
  - Distribución que incluye muchas librerías

Iniciar en Python

- Python website: `https://www.python.org/`
  - Distribución estándar de Python 2 y 3
  - Editor básico: IDLE
- Distribución Anaconda: `https://www.anaconda.com/download/`
  - Distribución que incluye muchas librerías
  - Administra dependencias

Iniciar en Python

- Python website: `https://www.python.org/`
  - Distribución estándar de Python 2 y 3
  - Editor básico: IDLE
- Distribución Anaconda: `https://www.anaconda.com/download/`
  - Distribución que incluye muchas librerías
  - Administra dependencias
  - Editores: Spyder, Jupyter

Un primer código en Python

```
# -*- coding: utf-8 -*-

num_cafes = [0, 1, 0, 2, 3, 0, 1, 2, 0, 2, 2, 2,
             1, 0, 3, 4]

n_obs = len(num_cafes)
sum_cafes = 0.0
for x in num_cafes:
    sum_cafes = sum_cafes + x
prom_cafes = sum_cafes/len(num_cafes)
max_cafes = max(num_cafes)
min_cafes = min(num_cafes)
```

Un primer código en Python

```
print("Número de observaciones: ", n_obs)
print("Número promedio de cafés: ", prom_cafes)
print("Número máximo de cafés: ", max_cafes)
print("Número mínimo de cafés: ", min_cafes)
```

Un primer código en Python: frecuencias absolutas

```
# -*- coding: utf-8 -*-
from collections import Counter
import matplotlib.pyplot as plt

num_cafes = [0, 1, 0, 2, 3, 0, 1, 2, 0, 2, 2, 2,
             1, 0, 3, 4]

num_cafes_cuenta = Counter(num_cafes)

xs = range(max(num_cafes)+1)
ys = [num_cafes_cuenta[x] for x in xs]
```

Un primer código en Python: frecuencias absolutas

```
print("Frecuencias absolutas: ", ys)

plt.bar(xs, ys)
plt.axis([-1, max(num_cafes)+1, 0, max(ys)+1])
plt.title(u"Histograma de Número de Tazas de Café")
plt.xlabel(u"# de tazas de café")
plt.ylabel(u"# of observaciones (personas)")
plt.show()
```

Un primer código en Python: frecuencias relativas

```
ys = [num_cafes_cuenta[x]/len(num_cafes) for x in xs]
```