

Análisis de Datos en Python: descriptivo e inferencial 2

Dora Suárez, Juan F. Pérez

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

juanferna.perez@urosario.edu.co

2018

Contenidos

- 1 Análisis descriptivo de datos
- 2 Media, varianza y desviación estándar
- 3 Cuantiles, percentiles, cuartiles y mediana
- 4 Diagramas de caja y bigotes

Análisis descriptivo de datos

Cargue y descripción de datos

```
# -*- coding: utf-8 -*-  
import pandas as pd  
  
filename = "data_blood.txt"  
datos = pd.read_csv(filename, header=None, sep="\s+",  
                    names = [u'Indice', u'Uno',  
                             u'Edad', u'Presión Sangre'])  
  
print(datos)  
  
print(type(datos))  
print(datos.head())  
print(datos.tail())
```

Cargue y descripción de datos

```
print(datos.index)
print(datos.columns)
print(datos.dtypes)
print(datos.shape)
print(datos.values)
print(datos.info())

print(datos.describe())
```

Cargue y descripción de datos

	Indice	Uno	Edad	Presion Sangre
count	30.000000	30.0	30.000000	30.000000
mean	15.500000	1.0	45.133333	142.533333
std	8.803408	0.0	15.294203	22.581245
min	1.000000	1.0	17.000000	110.000000
25 %	8.250000	1.0	36.750000	125.750000
50 %	15.500000	1.0	45.500000	141.000000
75 %	22.750000	1.0	56.000000	157.000000
max	30.000000	1.0	69.000000	220.000000

Descripción de datos

- **count**: número de observaciones

Descripción de datos

- **count**: número de observaciones
- **mean**: promedio

Descripción de datos

- **count**: número de observaciones
- **mean**: promedio
- **std**: desviación estándar

Descripción de datos

- **count**: número de observaciones
- **mean**: promedio
- **std**: desviación estándar
- **min**: mínimo

Descripción de datos

- **count**: número de observaciones
- **mean**: promedio
- **std**: desviación estándar
- **min**: mínimo
- **max**: máximo

Descripción de datos

- **count**: número de observaciones
- **mean**: promedio
- **std**: desviación estándar
- **min**: mínimo
- **max**: máximo
- **25 %**: percentil 25 (primer cuartil)

Descripción de datos

- **count**: número de observaciones
- **mean**: promedio
- **std**: desviación estándar
- **min**: mínimo
- **max**: máximo
- **25 %**: percentil 25 (primer cuartil)
- **50 %**: percentil 50 (segundo cuartil, mediana)

Descripción de datos

- **count**: número de observaciones
- **mean**: promedio
- **std**: desviación estándar
- **min**: mínimo
- **max**: máximo
- **25 %**: percentil 25 (primer cuartil)
- **50 %**: percentil 50 (segundo cuartil, mediana)
- **75 %**: percentil 75 (tercer cuartil)

Media, varianza y desviación estándar

Análisis descriptivo de datos: valor esperado

- X : característica de la población (variable aleatoria)

Análisis descriptivo de datos: valor esperado

- X : característica de la población (variable aleatoria)
- **Valor esperado** de X :

$$E[X] = \sum_j j \times p_j$$

Análisis descriptivo de datos: valor esperado

- X : característica de la población (variable aleatoria)
- **Valor esperado** de X :

$$E[X] = \sum_j j \times p_j$$

- Promedio ponderado de los valores que toma X

Análisis descriptivo de datos: valor esperado

- X : característica de la población (variable aleatoria)
- **Valor esperado** de X :

$$E[X] = \sum_j j \times p_j$$

- Promedio ponderado de los valores que toma X
- Medida de localización de X

Análisis descriptivo de datos: varianza

- X : característica de la población (variable aleatoria)

Análisis descriptivo de datos: varianza

- X : característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

Análisis descriptivo de datos: varianza

- X : característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

- Promedio ponderado de las diferencias de los valores que toma X respecto a su valor esperado

Análisis descriptivo de datos: varianza

- X : característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

- Promedio ponderado de las diferencias de los valores que toma X respecto a su valor esperado
- Medida de la variabilidad de X respecto a $E[X]$

Análisis descriptivo de datos: varianza

- X : característica de la población (variable aleatoria)
- Varianza de X

$$V[X] = E[(X - E[X])^2] = \sum_j (j - E[X])^2 \times p_j$$

- Promedio ponderado de las diferencias de los valores que toma X respecto a su valor esperado
- Medida de la variabilidad de X respecto a $E[X]$
- Desviación estándar:

$$\sigma_X = \sqrt{V[X]}$$

Análisis descriptivo de datos (ejemplo)

■

$$P(X = x) = \begin{cases} 1/3, & x = 1, \\ 1/3, & x = 2, \\ 1/3, & x = 3. \end{cases}$$

■

$$P(Y = y) = \begin{cases} 1/4, & y = 1, \\ 1/2, & y = 2, \\ 1/4, & y = 3. \end{cases}$$

■

$$P(Z = z) = \begin{cases} 1, & z = 2. \end{cases}$$

■

$$P(W = x) = \begin{cases} 1/2, & x = 1, \\ 1/2, & x = 3. \end{cases}$$

Análisis descriptivo de datos (ejemplo)

- $E[X] = E[Y] = E[Z] = E[W] = 2$
- $V(X) = (1 - 2)^2(1/3) + (2 - 2)^2(1/3) + (3 - 2)^2(1/3) = 2/3$
- $V(Y) = (1 - 2)^2(1/4) + (2 - 2)^2(1/2) + (3 - 2)^2(1/4) = 1/2$
- $V(Z) = (2 - 2)^2(1) = 0$
- $V(W) = (1 - 2)^2(1/2) + (3 - 2)^2(1/2) = 1$

Análisis descriptivo de datos (ejemplo)

- $\sigma_X = \sqrt{2/3}$
- $\sigma_Y = \sqrt{1/2}$
- $\sigma_Z = \sqrt{0} = 0$
- $\sigma_W = \sqrt{1} = 1$

Análisis descriptivo de datos: muestra aleatoria

Para estimar el valor esperado (μ):

- A partir de una muestra $\{X_1, \dots, X_n\}$
- **Media muestral:** \bar{X}

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

Análisis descriptivo de datos: muestra aleatoria

Para estimar el valor esperado (μ):

- A partir de una muestra $\{X_1, \dots, X_n\}$
- **Media muestral:** \bar{X}

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Promedio de los datos de la muestra

Análisis descriptivo de datos: muestra aleatoria

Para estimar el valor esperado (μ):

- A partir de una muestra $\{X_1, \dots, X_n\}$
- **Media muestral:** \bar{X}

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Promedio de los datos de la muestra
- Mean, media, promedio

Análisis descriptivo de datos: muestra aleatoria

Para estimar la varianza (σ^2):

- A partir de una muestra $\{X_1, \dots, X_n\}$
- **Varianza muestral:** S^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Análisis descriptivo de datos: muestra aleatoria

Para estimar la varianza (σ^2):

- A partir de una muestra $\{X_1, \dots, X_n\}$
- **Varianza muestral:** S^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Promedio de la diferencia (al cuadrado) entre los datos de la muestra y el promedio muestral

Análisis descriptivo de datos: muestra aleatoria

Para estimar la varianza (σ^2):

- A partir de una muestra $\{X_1, \dots, X_n\}$
- **Varianza muestral:** S^2

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Promedio de la diferencia (al cuadrado) entre los datos de la muestra y el promedio muestral
- **Desviación estándar muestral (std):** $S = \sqrt{S^2}$

Volviendo a la descripción de datos

	Indice	Uno	Edad	Presión Sangre
count	30.000000	30.0	30.000000	30.000000
mean	15.500000	1.0	45.133333	142.533333
std	8.803408	0.0	15.294203	22.581245
min	1.000000	1.0	17.000000	110.000000
25 %	8.250000	1.0	36.750000	125.750000
50 %	15.500000	1.0	45.500000	141.000000
75 %	22.750000	1.0	56.000000	157.000000
max	30.000000	1.0	69.000000	220.000000

Cuantiles, percentiles, cuartiles y mediana

Cuantiles

- X : característica de la población (variable aleatoria)

Cuantiles

- X : característica de la población (variable aleatoria)
- p : número entre 0 y 1

Cuantiles

- X : característica de la población (variable aleatoria)
- p : número entre 0 y 1
- El **cuantil** p es el número c más pequeño tal que la probabilidad de que X sea menor que c es al menos p , i.e.,:

$$P(X \leq c) \geq p$$

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & \text{d/c.} \end{cases}$$

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & \text{d/c.} \end{cases}$$

- El cuantil 0.25 es 1

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & \text{d/c.} \end{cases}$$

- El cuantil 0.25 es 1
- El cuantil 0.5 es 2.

Cuantiles (ejemplo)

$$P(X = x) = \begin{cases} 1/4, & x = 1, \\ 1/2, & x = 2, \\ 1/4, & x = 3, \\ 0, & \text{d/c.} \end{cases}$$

- El cuantil 0.25 es 1
- El cuantil 0.5 es 2.
- Los cuantiles 0.3 y 0.75 también son 2.

Cuantiles especiales

- **Percentiles:** el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99

Cuantiles especiales

- **Percentiles:** el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99
- Ejemplo: el percentil 75 es el cuantil 0,75

Cuantiles especiales

- **Percentiles:** el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99
- Ejemplo: el percentil 75 es el cuantil 0,75
- La **mediana** es el percentil 50 (cuantil 0.5)

Cuantiles especiales

- **Percentiles:** el percentil p es el cuantil $p/100$, donde p es un número entero entre 1 y 99
- Ejemplo: el percentil 75 es el cuantil 0,75
- La **mediana** es el percentil 50 (cuantil 0.5)
- Los tres **cuartiles** son los percentiles 25, 50 y 75

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \dots, X_n\}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \dots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \dots, X_{(n)}\}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \dots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \dots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \dots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \dots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$
- Antes de $X_{(1)}$ se han observado 0 muestras menores o iguales a $X_{(1)}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \dots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \dots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$
- Antes de $X_{(1)}$ se han observado 0 muestras menores o iguales a $X_{(1)}$
- Justo en $X_{(k)}$ se han observado k muestras menores o iguales a $X_{(k)}$

Cuantiles muestrales

- A partir de una muestra aleatoria $\{X_1, \dots, X_n\}$
- Se **ordena** la muestra de menor a mayor $\{X_{(1)}, \dots, X_{(n)}\}$
- Cada muestra tiene un peso de $\frac{1}{n}$
- Antes de $X_{(1)}$ se han observado 0 muestras menores o iguales a $X_{(1)}$
- Justo en $X_{(k)}$ se han observado k muestras menores o iguales a $X_{(k)}$
- A partir de $X_{(n)}$ se han observado todas las n muestras

Cuantiles muestrales

- Cuantil c es la primera muestra (en la lista ordenada) en la que se han observado $(c \times n)$ muestras menores o iguales.

Cuantiles muestrales

- Cuantil c es la primera muestra (en la lista ordenada) en la que se han observado $(c \times n)$ muestras menores o iguales.
- Cuantil c es la muestra $(c \times n)$ -ésima en la lista ordenada

Cuantiles muestrales

- Cuantil c es la primera muestra (en la lista ordenada) en la que se han observado $(c \times n)$ muestras menores o iguales.
- Cuantil c es la muestra $(c \times n)$ -ésima en la lista ordenada
- Ejemplo: si tenemos una muestra de tamaño $n = 1000$, el cuantil 0,2 es la muestra 200 de la lista ordenada

Cuantiles muestrales (ejemplo)

```
# -*- coding: utf-8 -*-
import pandas as pd
import matplotlib.pyplot as plt

filename = "data/data_blood.txt"
datos = pd.read_csv(filename, header=None, sep="\s+",
                    names = [u'Indice', u'Uno',

ser = pd.Series(datos[u'Presión Sangre'])
print(ser)
print(ser.describe())
```


Cuantiles muestrales (ejemplo)

```
ser.hist(cumulative=True, bins=100)  
plt.show()
```

Cuantiles muestrales (ejemplo)

```
ser.hist(bins=10)  
plt.show()
```

Cuantiles muestrales (ejemplo)

count	30.000000
mean	142.533333
std	22.581245
min	110.000000
25 %	125.750000
50 %	141.000000
75 %	157.000000
max	220.000000

Cuantiles muestrales (ejemplo)

count	30.000000
mean	142.533333
std	22.581245
min	110.000000
25 %	125.750000
50 %	141.000000
75 %	157.000000
max	220.000000

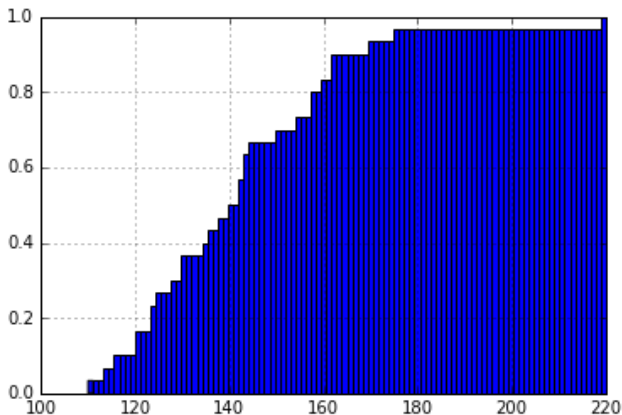
- Primer cuartil (Q1): 125.75
- Segundo cuartil (Q2): 141

Cuantiles muestrales (ejemplo)

count	30.000000
mean	142.533333
std	22.581245
min	110.000000
25 %	125.750000
50 %	141.000000
75 %	157.000000
max	220.000000

- Primer cuartil (Q1): 125.75
- Segundo cuartil (Q2): 141
- Tercer cuartil (Q3): 157

Cuantiles muestrales (ejemplo)



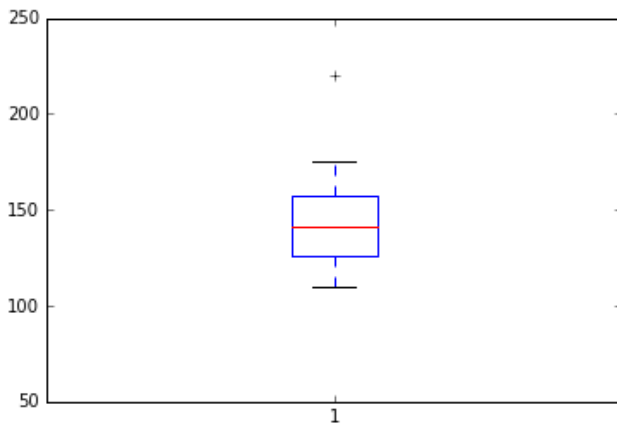
Diagramas de caja y bigotes

Diagramas de caja y bigotes

```
print(ser)
print(ser.describe())

plt.boxplot(ser)
plt.ylim(0, 300)
plt.show()
```


Diagramas de caja y bigotes(ejemplo)



Diagramas de caja y bigotes

- **Caja:**

Diagramas de caja y bigotes

- **Caja:**

- Punto medio: mediana (cuartil 2, Q_2)

Diagramas de caja y bigotes

■ Caja:

- Punto medio: mediana (cuartil 2, Q_2)
- Punto inferior: primer cuartil (Q_1)
- Punto superior: tercer cuartil (Q_3)

Diagramas de caja y bigotes

■ Caja:

- Punto medio: mediana (cuartil 2, Q2)
- Punto inferior: primer cuartil (Q1)
- Punto superior: tercer cuartil (Q3)

■ Recorrido intercuartílico: $RQ = Q3 - Q1$

Diagramas de caja y bigotes

- **Caja:**
 - Punto medio: mediana (cuartil 2, Q2)
 - Punto inferior: primer cuartil (Q1)
 - Punto superior: tercer cuartil (Q3)
- **Recorrido intercuartílico:** $RQ = Q3 - Q1$
- **Bigotes:**

Diagramas de caja y bigotes

■ Caja:

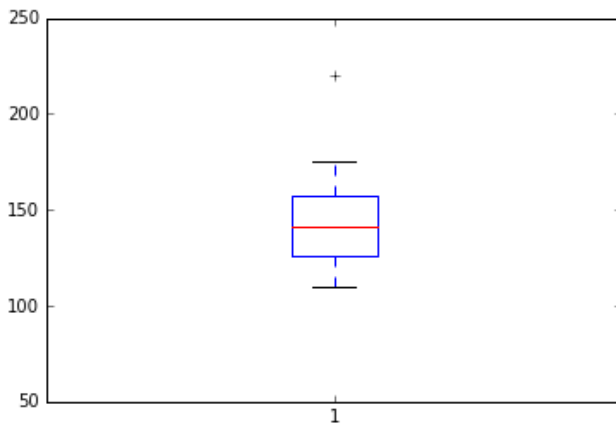
- Punto medio: mediana (cuartil 2, Q2)
- Punto inferior: primer cuartil (Q1)
- Punto superior: tercer cuartil (Q3)

■ Recorrido intercuartílico: $RQ = Q3 - Q1$

■ Bigotes:

- Punto inferior: observación más pequeña a menos de 1.5 RQ del primer cuartil (Q1)
- Punto superior: observación más grande a menos de 1.5 RQ del tercer cuartil (Q3)

Diagramas de caja y bigotes



Datos atípicos

- Datos que son:

Datos atípicos

- Datos que son:
 - Mayores al tercer cuartil más 1.5 RQ ($Q3 + 1.5RQ$)

Datos atípicos

- Datos que son:
 - Mayores al tercer cuartil más 1.5 RQ ($Q3+1.5RQ$)
 - Menores al primer cuartil menos 1.5 RQ ($Q1-1.5RQ$)

Datos atípicos

- Datos que son:
 - Mayores al tercer cuartil más 1.5 RQ ($Q3 + 1.5RQ$)
 - Menores al primer cuartil menos 1.5 RQ ($Q1 - 1.5RQ$)
- En el diagrama de caja y bigotes:

Datos atípicos

- Datos que son:
 - Mayores al tercer cuartil más 1.5 RQ ($Q3+1.5RQ$)
 - Menores al primer cuartil menos 1.5 RQ ($Q1-1.5RQ$)
- En el diagrama de caja y bigotes:
 - Puntos por fuera de los bigotes

Datos atípicos

- Datos que son:
 - Mayores al tercer cuartil más 1.5 RQ ($Q3+1.5RQ$)
 - Menores al primer cuartil menos 1.5 RQ ($Q1-1.5RQ$)
- En el diagrama de caja y bigotes:
 - Puntos por fuera de los bigotes
 - Se marcan gráficamente

Datos atípicos (ejemplo)

```
print(ser.sort_values())
```

```
ser.hist()
```

```
plt.show()
```

Datos atípicos (ejemplo)

