Estadística descriptiva

Nelson Enrique Vera Parra Ph.D.

Estadística descriptiva

Descripción de un conjunto de datos

- Distribución
- Tendencia central
- Dispersión
- Relación entre variables*

*Este es el límite entre la estadística descriptiva y la predictiva

Estadística descriptiva

- Distribución

Histograma

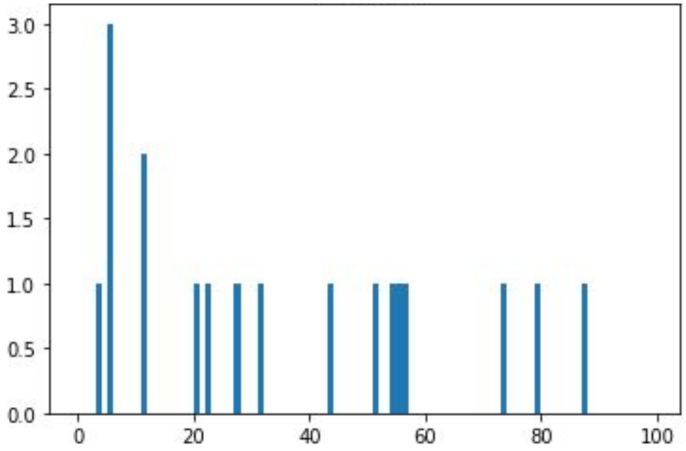
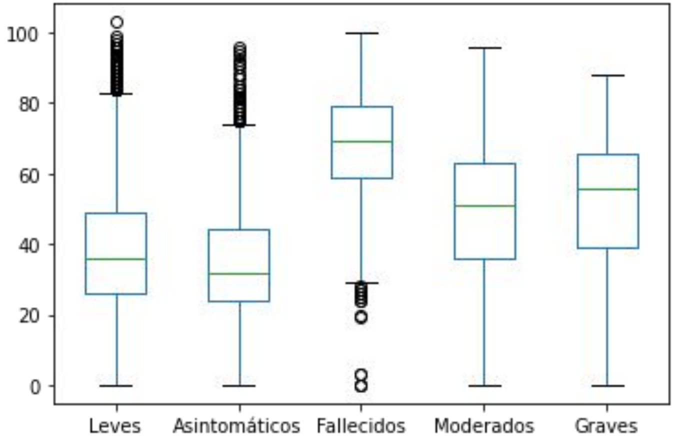


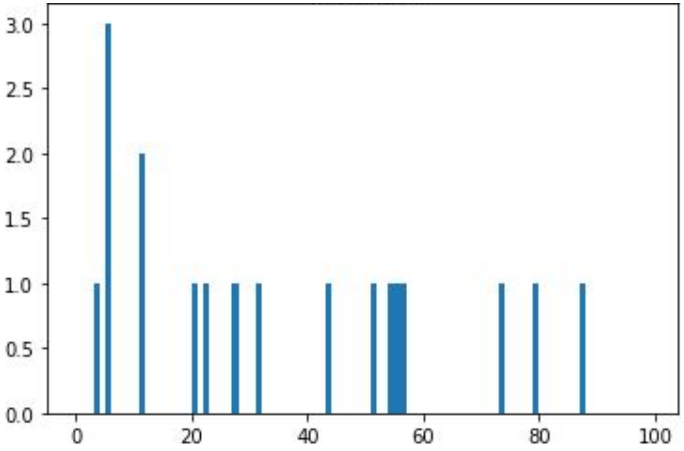
Diagrama de cajas



Estadística descriptiva

- Distribución

Histograma

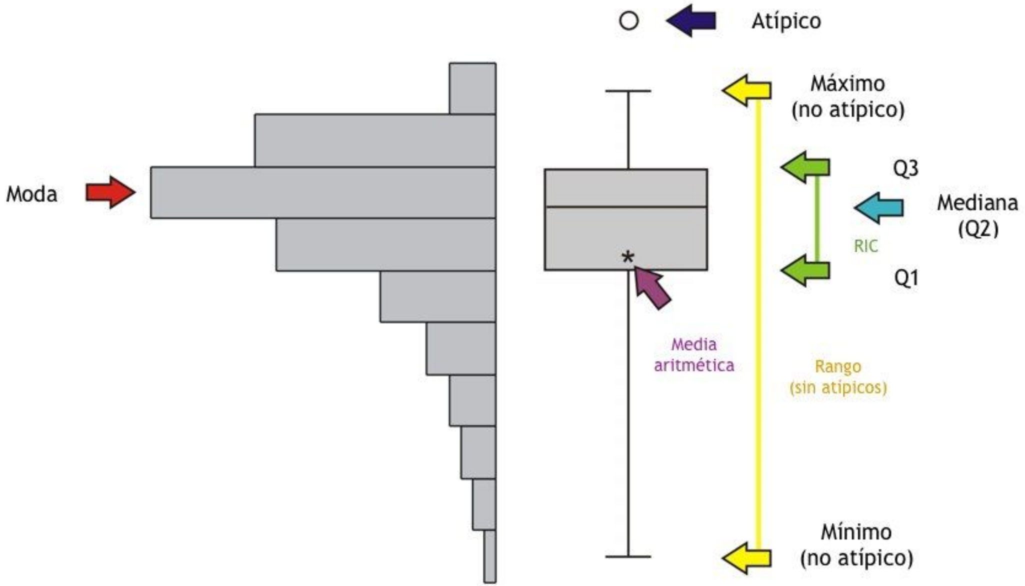



```
from matplotlib import pyplot as plt
import numpy as np

edades =
np.array([22,87,5,43,56,73,55,54,11,20,51,5,79,31,27,11,3,5])
plt.hist(edades, bins = range(100))
plt.title("histogram")
plt.show()
```

Estadística descriptiva

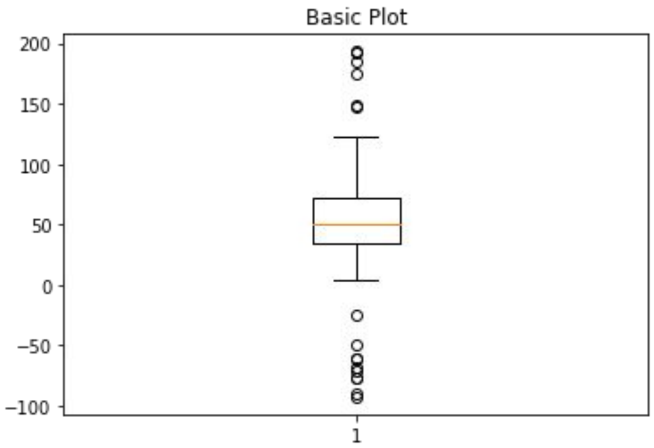
- Distribución

Diagrama de cajas



Estadística descriptiva

- Distribución




```
import numpy as np
import matplotlib.pyplot as plt

# Fixing random state for reproducibility
np.random.seed(19680801)

# fake up some data
spread = np.random.rand(50) * 100
center = np.ones(25) * 50
flier_high = np.random.rand(10) * 100 + 100
flier_low = np.random.rand(10) * -100
data = np.concatenate((spread, center,flier_high, flier_low))
fig1, ax1 = plt.subplots()
ax1.set_title('Basic Plot')
ax1.boxplot(data)
```

Estadística descriptiva

- Distribución

  https://docs.scipy.org/doc/numpy-1.14.0/reference/routines.random.html

Distributions

`beta`(a, b[, size])	Draw samples from a Beta distribution.
`binomial`(n, p[, size])	Draw samples from a binomial distribution.
`chisquare`(df[, size])	Draw samples from a chi-square distribution.
`dirichlet`(alpha[, size])	Draw samples from the Dirichlet distribution.
`exponential`([scale, size])	Draw samples from an exponential distribution.
`f`(dfnum, dfden[, size])	Draw samples from an F distribution.
`gamma`(shape[, scale, size])	Draw samples from a Gamma distribution.
`geometric`(p[, size])	Draw samples from the geometric distribution.
`gumbel`([loc, scale, size])	Draw samples from a Gumbel distribution.
`hypergeometric`(ngood, nbad, nsample[, size])	Draw samples from a Hypergeometric distribution.
`laplace`([loc, scale, size])	Draw samples from the Laplace or double exponential distribution with specified location (or mean) and scale (decay).
`logistic`([loc, scale, size])	Draw samples from a logistic distribution.
`lognormal`([mean, sigma, size])	Draw samples from a log-normal distribution.
`logseries`(p[, size])	Draw samples from a logarithmic series distribution.

Estadística descriptiva

- Tendencia central: Media, Mediana, Moda

Estadística descriptiva

- Tendencia central: Media

Es el promedio aritmético

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

```
import statistics as st
st.mean([2,3,5,7,9,12])
Out[ ]: 6.333333333333333
```

Estadística descriptiva

- Tendencia central: Mediana

Es el valor ubicado en la posición central de un conjunto de datos ordenados. Existen dos tipos de mediana: para datos no agrupados y para datos agrupados.

Estadística descriptiva

- Tendencia central: Mediana – datos no agrupados

- Si el número de elementos es impar, la mediana es el valor del elemento central

```
import statistics as st
st.median([2,3,5,7,9,12,30])
Out[ ]: 7
```

- Si el número de elementos es par, la mediana es el promedio aritmético de los valores de los dos elementos centrales

```
import statistics as st
st.median([2,2,3,5,7,9,12,30])
Out[61]: 6.0
```

Estadística descriptiva

- Tendencia central: Mediana – datos agrupados

$$Mediana = x_{i1} + \left(\frac{(N_M/2) - N_{i-1}}{f_i} \right) . (x_{i2} - x_{i1})$$

donde:

x_{i1} = es el límite inferior de la clase de la mediana.

$(\frac{N_M}{2})$ = es la posición de la mediana.

N_{i-1} = es la frecuencia acumulada de la clase premediana.

f_i = es la frecuencia absoluta de la clase de la mediana.

$(x_{i2} - x_{i1})$ = A_i = Amplitud del intervalo de la clase de la mediana.

Estadística descriptiva

- Tendencia central: Mediana – datos agrupados

```
import statistics as st
st.median_grouped([1, 2, 2, 3, 4, 4, 4, 4, 4, 5])
Out[ ]: 3.7
```

Rangos	f	F	h	H
0.5 - 1.5	1	1	0.1	0.1
1.5 - 2.5	2	3	0.2	0.3
2.5 - 3.5	1	4	0.1	0.4
3.5 - 4.5	5	9	0.5	0.9
4.5 - 5-5	1	10	0.1	1.0

$$Mediana = x_{i1} + \left(\frac{(N_M/2) - N_{i-1}}{f_i}\right) . (x_{i2} - x_{i1})$$

Mediana = 3.5 + ((5-4)/5)*1 = 3.7

Estadística descriptiva

- Tendencia central: Moda

Es el valor que presenta mayor frecuencia

```
import statistics as st
st.mode([1, 2, 2, 3, 4, 4, 4, 4, 4, 5])
Out[ ]: 4
```

Estadística descriptiva

- Dispersión: Rango, Varianza, Desviación estándar

Estadística descriptiva

- Dispersión: Rango

```
import numpy as np

def data_range(x):
    return max(x) - min(x)

edades = np.array([22,87,5,43,56,73,55,54,11,20,51,5,79,31,27,11,3,5])
data_range(edades)

Out[ ]: 84
```

Estadística descriptiva

- Dispersión: Varianza

La varianza es la esperanza del cuadrado de la desviación de dicha variable respecto a su media.
La varianza mide qué tan dispersos están los datos alrededor de la media
La varianza es el cuadrado de la desviación estándar

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

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

Varianza de una población

Varianza de una muestra

Estadística descriptiva

- Dispersión: Varianza

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

```
import statistics as st
st.pvariance([20, 22, 22, 23, 24])
Out[ ]: 1.76
st.variance([20, 22, 22, 23, 24])
Out[ ]: 2.2
```

$$\sigma^2 = \frac{(20-22.2)^2 + (22-22.2)^2 + (22-22.2)^2 + (23-22.2)^2 + (24-22.2)^2}{5}$$

$$\sigma^2 = 1.76\ años^2$$

Estadística descriptiva

- Dispersión: Desviación estándar

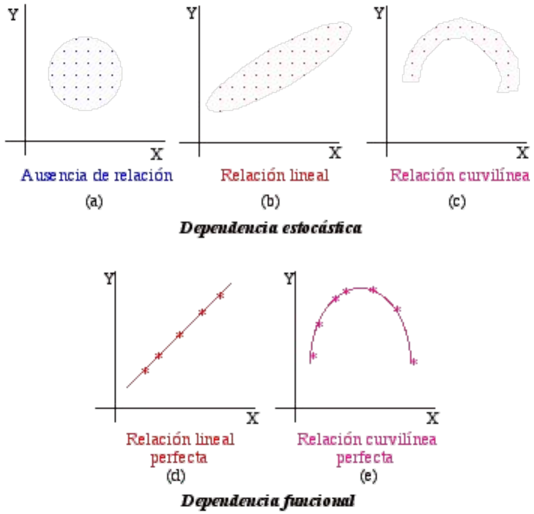
```
import numpy as np
np.std([20, 22, 22, 23, 24])
Out[ ]: 1.32664991614216
```

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

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Estadística descriptiva

- Relación entre variables: Covarianza, Correlación



Estadística descriptiva

- Relación entre variables: Covarianza

La covarianza es el grado de variación conjunta de dos variables aleatorias respecto a sus medias. Si la covarianza es positiva indica una relación lineal directa y si es negativa indica una relación lineal indirecta. Si la covarianza es cero las variables son independientes.
Desventaja -> el rango depende de la magnitud de las variables

$$\sigma(x,y) = \mathrm{E}\left[(x - \mathrm{E}[x])(y - \mathrm{E}[y])\right]$$

$$\sigma(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Estadística descriptiva

- Relación entre variables: Covarianza

$$\sigma(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

x	y
-2.1	3
-1	1.1
4.3	0.12
3	0.9

$$\sigma = \frac{(-2.1 - 1.05)(3 - 1.28) + (-1 - 1.05)(1.1 - 1.28) + (4.3 - 1.05)(0.12 - 1.28) + (3 - 1.05)(0.9 - 1.28)}{4}$$

$$\sigma = -2.39$$

$$S = \frac{(-2.1 - 1.05)(3 - 1.28) + (-1 - 1.05)(1.1 - 1.28) + (4.3 - 1.05)(0.12 - 1.28) + (3 - 1.05)(0.9 - 1.28)}{3}$$

$$S = -3.1867$$

Estadística descriptiva

- Relación entre variables: Covarianza

x	y
-2.1	3
-1	1.1
4.3	0.12
3	0.9

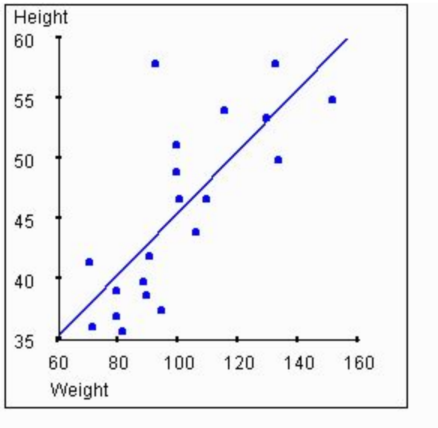
$$S = -3.1867$$

$$\sigma = -2.39$$

```
import statistics as st
from matplotlib import pyplot as plt
import numpy as np

x = [-2.1, -1,  4.3, 3]
y = [3,  1.1,  0.12, 0.9]

st.variance(x)
Out[ ]: 9.496666666666666
st.variance(y)
Out[ ]: 1.4936
np.cov(x,y)
Out[ ]:
array([[ 9.49666667, -3.18666667],
       [-3.18666667, 1.4936 ]])

st.pvariance(x)
Out[ ]: 7.1225
st.pvariance(y)
Out[ ]: 1.1202
np.cov(x, y, bias=True)
Out[ ]:
array([[ 7.1225, -2.39 ],
       [-2.39 , 1.1202]])
```

Estadística descriptiva

- Relación entre variables: Correlación

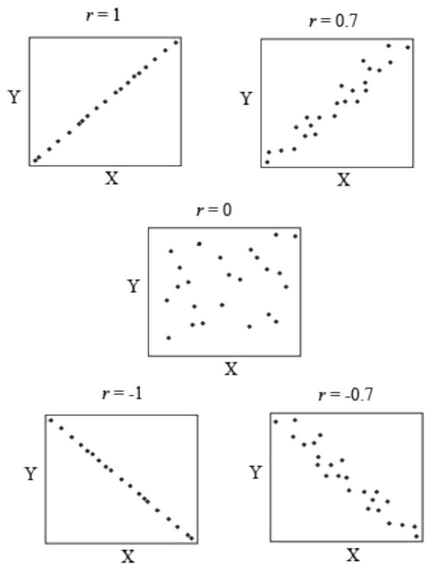


La recta que mejor modele la relación entre Y y X

Coeficiente de correlación -> Fuerza y sentido

Estadística descriptiva

- Relación entre variables: Correlación



r - Coeficiente de correlación de Pearson

Estadística descriptiva

- Relación entre variables: Correlación

r - Coeficiente de correlación de Pearson

Coeficiente	Interpretación
r = 1	Correlación perfecta
0.80 < r < 1	Muy alta
0.60 < r < 0.80	Alta
0.40 < r < 0.60	Moderada
0.20 < r < 0.40	Baja
0 < r < 0.20	Muy baja
r = 0	Nula

Estadística descriptiva

- Relación entre variables: Correlación

r - Coeficiente de correlación de Pearson

Se define como el cociente entre la covarianza y el producto de las desviaciones típicas (raíz cuadrada de las varianzas)

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}$$

Estadística descriptiva

- Relación entre variables: Correlación

r - Coeficiente de correlación de Pearson

```
import numpy as np
np.corrcoef(x, y)
```

```
from scipy.stats.stats import pearsonr
scipy.stats.pearsonr(x, y)
```