Probabilidad y estadística

Media:

Es la medida más utilizada para encontrar el promedio. De hecho, la gente siempre utiliza la palabra “promedio” para referirse a la “media”. Encontrarla es simple; sólo se suman todas los mediciones y se dividen por la cantidad de números.

Mediana:

Es el número del medio en un grupo de datos. Los datos deben estar ordenados numéricamente antes de encontrar esta medida. Si el número del medio está entre dos números, entonces encuentra la media entre esos dos. (súmalos y divídelos entre 2)

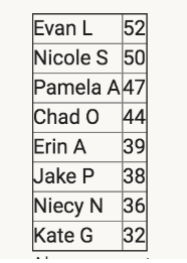
Moda:

La moda es probablemente la forma menos común de encontrar el promedio, y en la mayoría de los casos es la menos útil. Para encontrar la moda, solo encuentra el número que más se repite. Puede haber más de una moda, o ninguna.

Rango:

El rango mide la extensión de los datos, qué tan alejados se encuentran el menos del mayor. Para encontrar el rango se resta el valor más pequeño del más grande.

Ejemplo 1:



Evan L	52
Nicole S	50
Pamela A	47
Chad O	44
Erin A	39
Jake P	38
Niecy N	36
Kate G	32

Estadística	Cómo encontrarla	Explicación
Media	$\frac{52+50+47+44+39+38+36+32}{8} = \frac{338}{8} = 42.25$	Suma las puntuaciones y divídelas entre 8, el número total de concursantes. La media es 42.25.
Mediana	52 50 47 44 39 38 36 32 $\frac{44+39}{2} = \frac{83}{2} = 41.5$	Primero ordena las puntuaciones, y después encuentra el valor del medio. En este grupo, el valor medio está entre 44 y 39, así que sumamos estos dos y los dividimos entre 2.
Moda	No tiene moda.	No hay ninguna puntuación que ocurra varias veces, así que no hay moda para este grupo de datos.
Rango	52 – 32 = 20	Resta el más pequeño del más grande. El rango es 20 puntos.

Ejemplo 2:
Este sería el contenido de un archivo llamado “notas.csv”

```
alumno,nota
Araceli,9
Manuel,5
Pablo,7
Íñigo,4
Mario,3
Raúl,4
Verónica,6
Darío,10
Laura,4
Silvia,6
Eduardo,2
Susana,8
María,5
```

Código:

```
import pandas as pd

readfile = pd.read_csv(“notas.csv”)

media = readfile[“nota”].mean()
mediana = readfile[“nota”].median()
moda = readfile[“nota”].mode()

rango = readfile[“nota”].max() - readfile[“nota”].min()

print(media, mediana, moda, rango)
```

Desviación estándar
La desviación estándar (σ) mide cuánto se separan los datos.

La fórmula es fácil: es la raíz cuadrada de la varianza.

Varianza
la varianza (que es el cuadrado de la desviación estándar: σ2) se define así:

Es la media de las diferencias con la media elevadas al cuadrado.

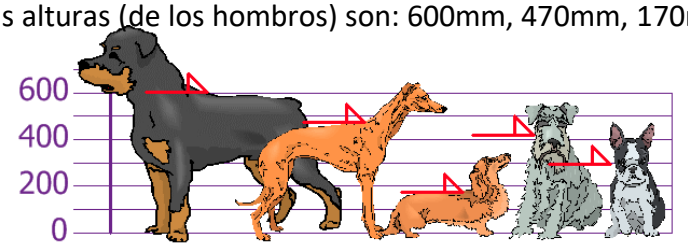
En otras palabras, sigue estos pasos:

1. Calcula la media (el promedio de los números)
2. Ahora, por cada número resta la media y eleva el resultado al cuadrado (la diferencia elevada al cuadrado).
3. Ahora calcula la media de esas diferencias al cuadrado.

Ejemplo

Tú y tus amigos miden las alturas de sus perros (en milímetros):

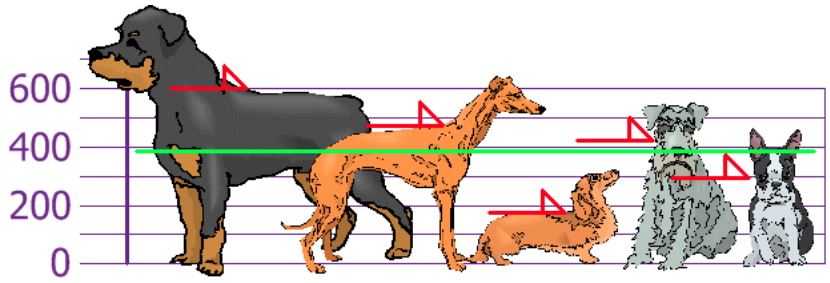
Las alturas (de los hombros) son: 600mm, 470mm, 170mm, 430mm y 300mm.



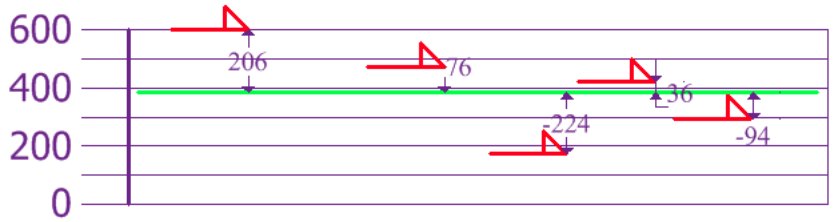
Calcula la media, la varianza y la desviación estándar.

Media = 600 + 470 + 170 + 430 + 300/5 = 1970/5 = 394

Así que la altura media es 394 mm. Vamos a dibujar esto en el gráfico:



Ahora calculamos la diferencia de cada altura con la media



Para calcular la varianza, toma cada diferencia, elévala al cuadrado, y haz la media:

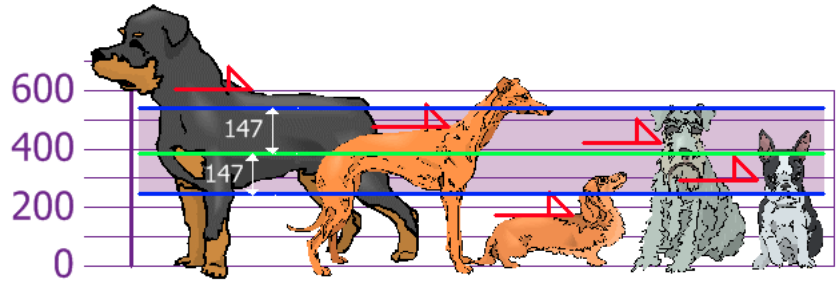
Varianza: σ2 = 206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2/5 = 108,520/5 = 21,704

Así que la varianza es 21,704.

Y la desviación estándar es la raíz de la varianza, así que:

Desviación estándar: σ = √21,704 = 147

y lo bueno de la desviación estándar es que es útil: ahora veremos qué alturas están a distancia menos de la desviación estándar (147mm) de la media:



Así que usando la desviación estándar tenemos una manera "estándar" de saber qué es normal, o extra grande o extra pequeño.

¿Por qué al cuadrado?
Elevar cada diferencia al cuadrado hace que todos los números sean positivos (para evitar que los números negativos reduzcan la varianza)

Y también hacen que las diferencias grandes se destaquen. Por ejemplo 1002=10,000 es mucho más grande que 502=2,500.

Pero elevarlas al cuadrado hace que la respuesta sea muy grande, así que lo deshacemos (con la raíz cuadrada) y así la desviación estándar es mucho más útil.

```
import numpy as np
from scipy import stats

arr = np.array([1,2,3,4,5,6])

np.mean(arr) – media
np.median(arr)- mediana
stats.mode(arr) – moda
np.var(arr) – varianza
np.std(arr) – desviación estándar
```

Los cuantiles son aquellos valores de la variable, que ordenados de menor a mayor, dividen a la distribución en partes, de tal manera que cada una de ellas contiene el mismo número de frecuencias.

Los cuantiles más conocidos son:

a) Cuartiles (Qi)
Son valores de la variable que dividen a la distribución en 4 partes, cada una de las cuales engloba el 25 % de las mismas. Se denotan de la siguiente forma: Q1 es el primer cuartil que deja a su izquierda el 25 % de los datos; Q2 es el segundo cuartil que deja a su izquierda el 50% de los datos, y Q3 es el tercer cuartil que deja a su izquierda el 75% de los datos. (Q2 = Me)

b) Deciles (Di)
Son los valores de la variable que dividen a la distribución en las partes iguales, cada una de las cuales engloba el 10 % de los datos. En total habrá 9 deciles. (Q2 = D5 = Me)

c) Centiles o Percentiles (Pi)
Son los valores que dividen a la distribución en 100 partes iguales, cada una de las cuales engloba el 1 % de las observaciones. En total habrá 99 percentiles. (Q2 = D5 = Me P50)