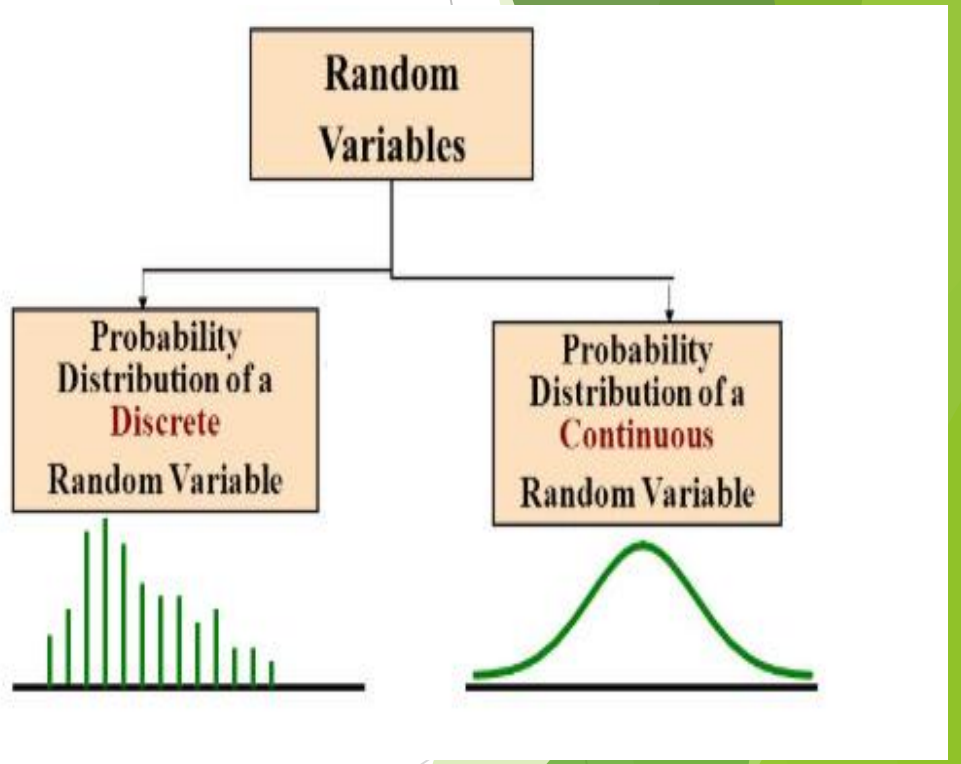
# Data Science

# Distribution

- **Definition of Distribution:**
- In statistics, a distribution is the way data points are spread or arranged over a range of values.
- It helps in understanding the underlying patterns and trends in the dataset.
- Discrete data: age = [22,64,12,89,56]
- Continuous data: weight = [26.4,38.2,11.9,76.3]



# Uniform Distribution

Example: Rolling a die.
Here probability for each digit occurance is 1/6 (uniform or same)

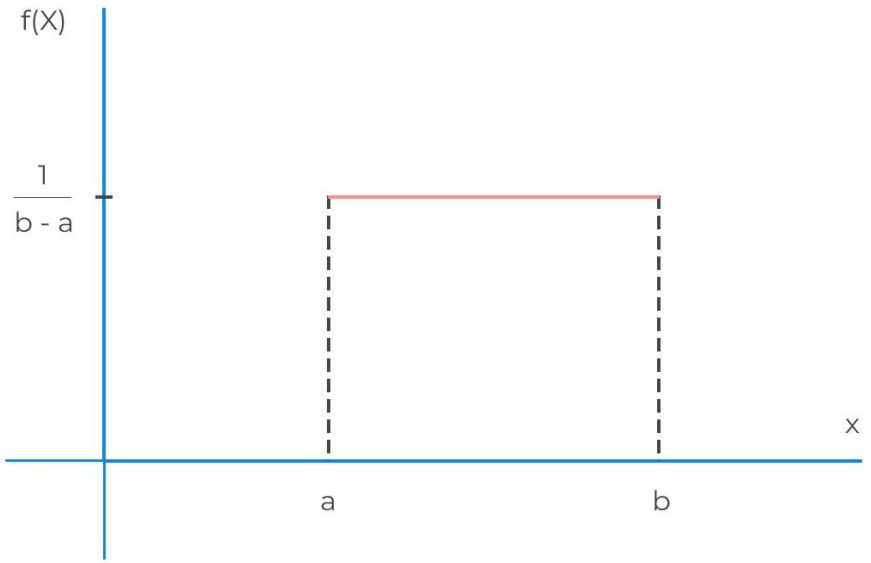
$$\text{Mean} = \frac{(a+b)}{2}$$

$$\text{Variance} = \frac{(b-a)^2}{12}$$

**P. D. F. of Uniform Probability Distribution**

$$f(x) = \begin{cases} f(x) = \dfrac{1}{b-a}, & \text{if } a < x < b. \\ 0, & \text{Otherwise.} \end{cases}$$

**PDF of a Uniform Distribution**




Area = 1

Numpy: np.rand()

# Bernoulli Distribution

$$\text{PMF} = f(x, p) = \begin{cases} p & \text{if } x = 1 \\ q = 1 - p & \text{if } x = 0 \end{cases}$$

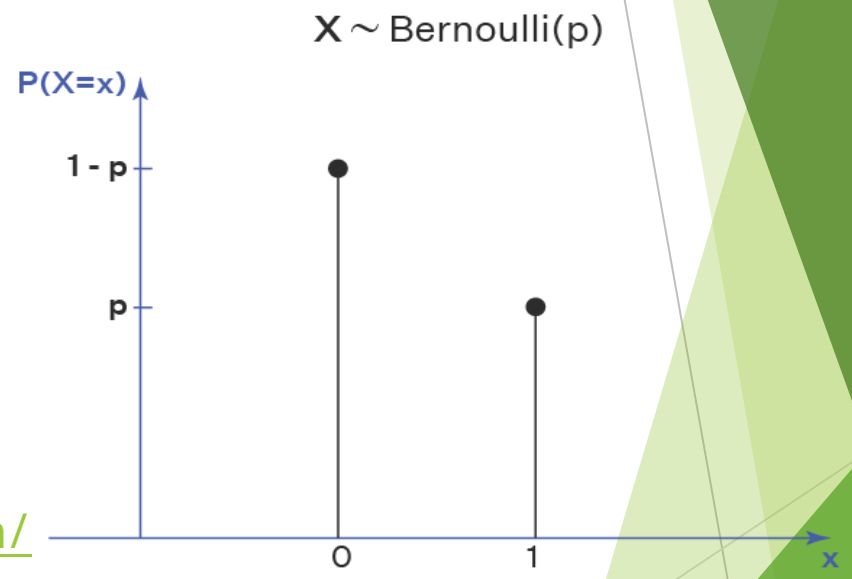
Thus, the mean or expected value of a Bernoulli distribution is given by

$E[X] = p.$

Hence, the variance of a Bernoulli distribution is $\text{Var}[X] = p(1 - p) = p \cdot q$

https://www.cuemath.com/data/bernoulli-distribution/

Bernoulli Distribution Graph

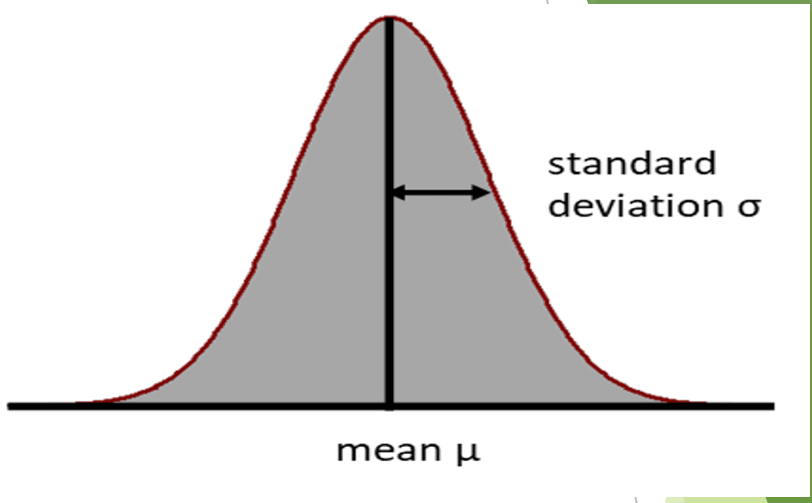


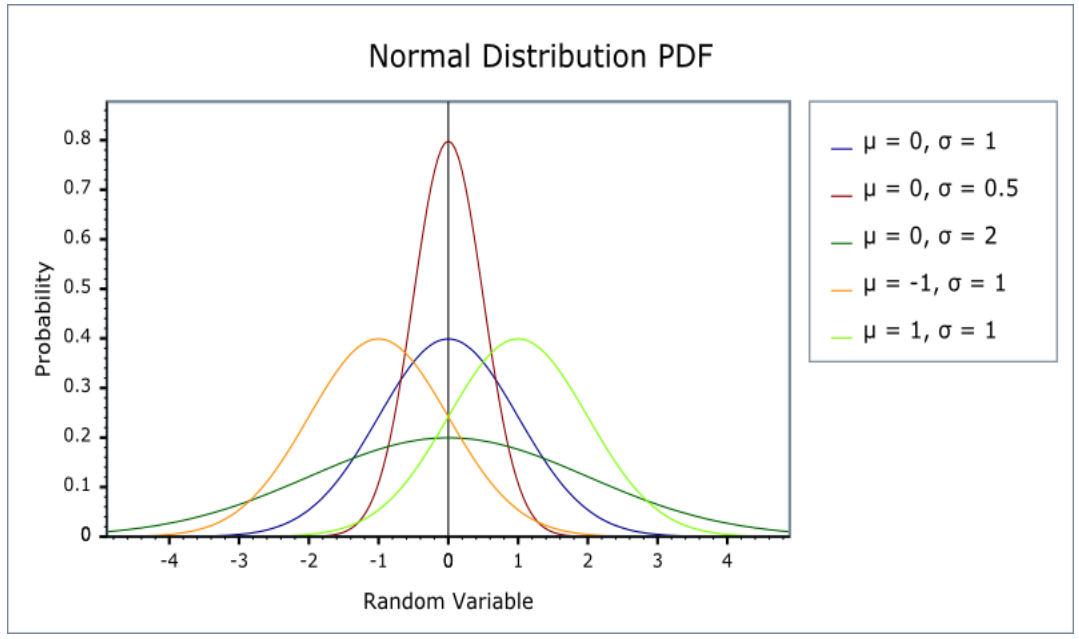

Example = Tossing a coin (heads or tails)
If p(X=head) = 0.5
Then p(X=tail) = 1- p(X=head) = 1-0.5 = 0.5

# Normal (Gaussian) Distribution

The normal distribution is an arrangement of data points in which most values form a cluster in the middle of the range and the rest taper off symmetrically toward either extreme ends

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



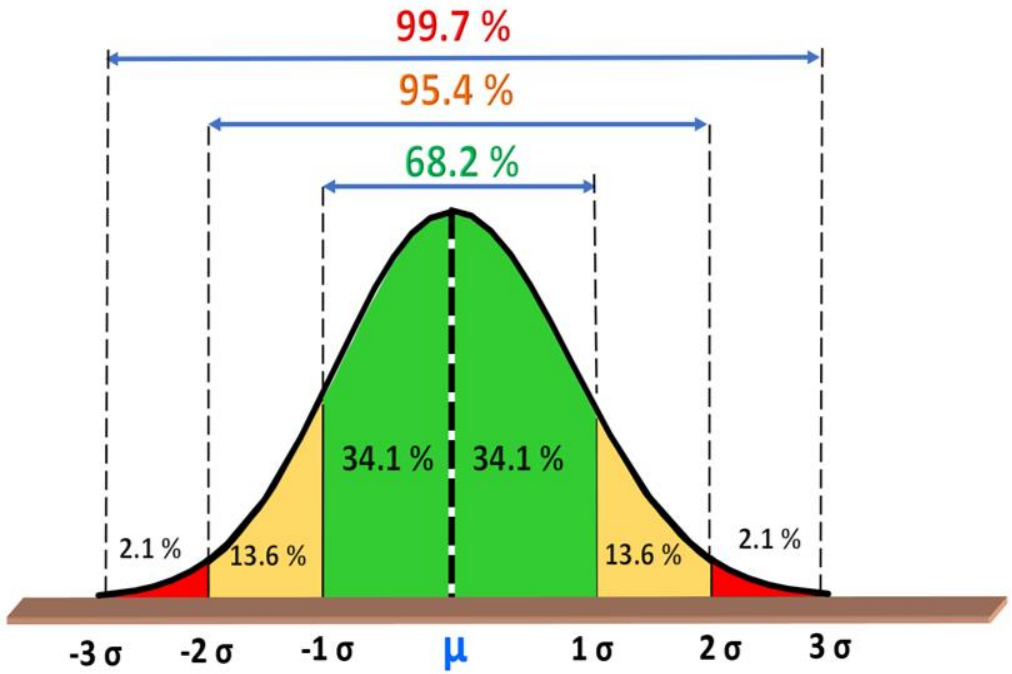



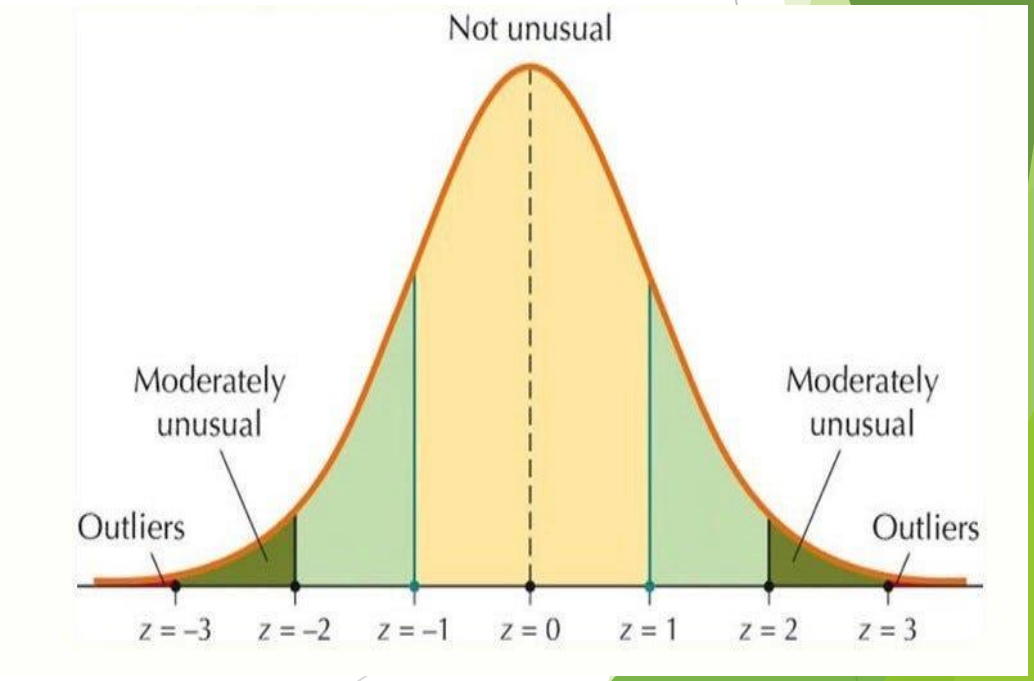

Example: Height or weight of people in country

# Normal (Gaussian) Outlier detection

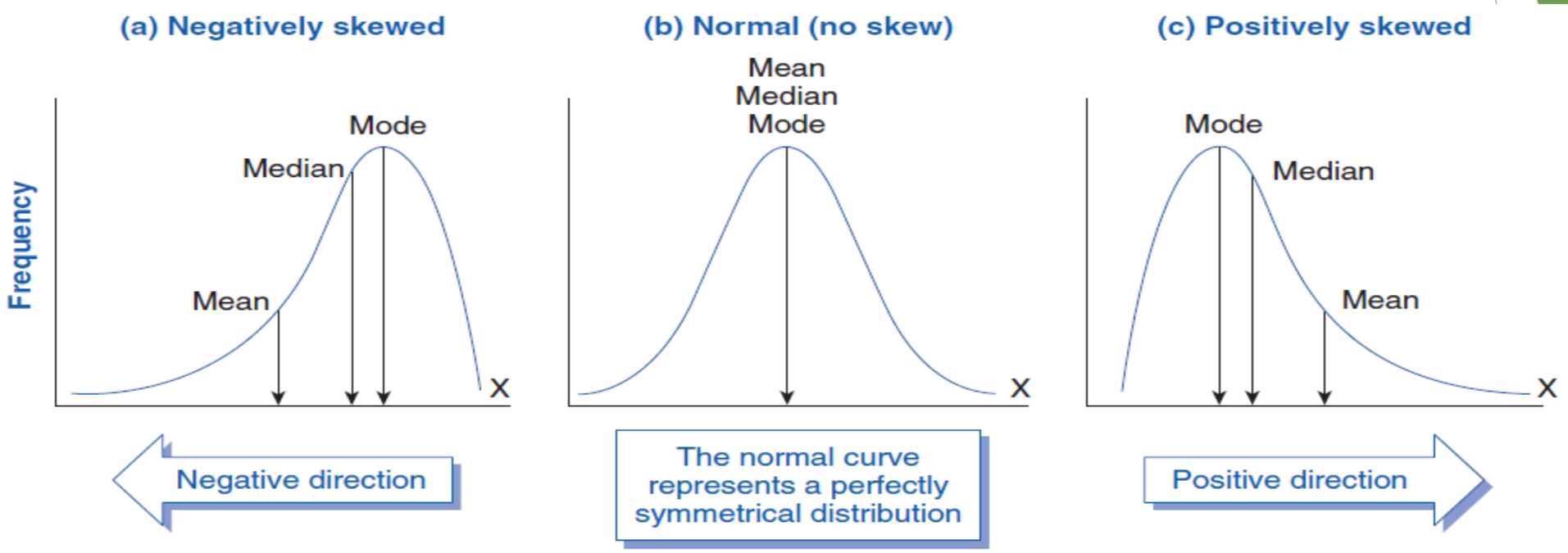
→The 68-95-99 rule:
→68% of the data is within 1 standard deviation of the mean.
→95% of the data is within 2 standard deviation of the mean.
→99.7% of the data is within 3 standard deviation of the mean.

Numpy:
np.randn()



# Skewed distribution




Left skewed:
Mean < Median < Mode

Example: Exam scores where most students scored high but a few scored very low.

Mode == Median == Mean

Example: Height of individuals in a country

Right skewed:
Mode < Median < Mean

Example: Income distribution where most earn less, and a few earn much more.

https://statisticsbyjim.com/basics/skewed-distribution/