Uniform & Log Normal Distribution

Types of Non-Gaussian Probability Distributions

1. Continuous: **Uniform, Log Normal, Pareto**
2. Discrete

Uniform Distribution

- A probability distribution
- Every value in a given range is equally likely
- Often called a **rectangular distribution** because its probability density function (PDF) is a flat, constant line.

Notation → $X \sim U(a, b)$

✅ Key Idea:

- **All values have the same probability** within a given range.
- There are **no peaks** (unlike a normal distribution).
- Used when **all outcomes are equally likely**, like rolling a fair die.

Types of Uniform Distributions

Type	Definition	Example
Discrete Uniform	A **finite** set of equally likely values	Rolling a **fair die (1,2,3,4,5,6)**
Continuous Uniform	A range of values with equal probability	Random number between **0 and 1**

Probability Density Function (PDF) for Continuous Uniform Distribution:

$$f(x) = \frac{1}{b-a}, \quad \text{for } a \leq x \leq b$$

- **a**: The lower bound of the distribution (minimum value).
- **b**: The upper bound of the distribution (maximum value).
- The PDF is flat between a and b, meaning each value within this range has the same probability

Cumulative Distribution Function (CDF):

- The CDF for a continuous uniform distribution is the integral of the PDF and gives the probability that a random variable X will take a value less than or equal to a given point x.

$$F(x) = \frac{x-a}{b-a} \quad \text{for} \quad a \leq x \leq b$$

🔷 **Example:**

For $a = 2$, $b = 6$, and $x = 4$:

$$F(4) = \frac{4-2}{6-2} = \frac{2}{4} = 0.5$$

This means **50% of the values** are ≤ **4**.

Properties of the Uniform Distribution:

1. **Mean** (Expected Value):

$$\mu = \frac{a+b}{2}$$

2. **Variance:**

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

3. **Standard Deviation:**

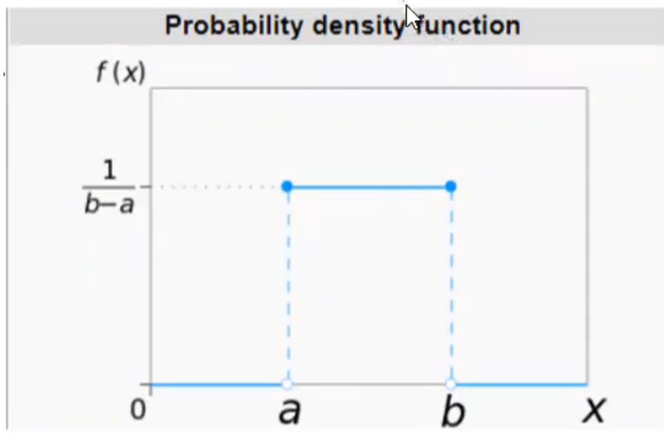
$$\sigma = \frac{b-a}{\sqrt{12}}$$

Continuous Uniform Distribution

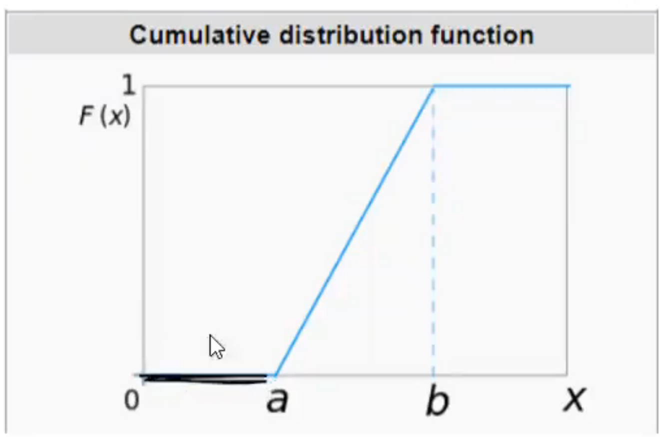
Examples:

- The height of a person randomly selected from a group of individuals whose height range from 5'6" to 6'0" would follow a continuous uniform distribution.
- The time it takes for a machine to produce a product, where the production time ranges from 5 to 10 minutes, would follow a continuous uniform distribution.
- The distance that a randomly selected car travels on a tank of gas, where the distance ranges from 300 to 400 miles, would follow a continuous uniform distribution.
- The weight of a randomly selected apple from a basket of apples that weighs between 100 and 200 grams, would follow a continuous uniform distribution.

PDF:



CDF:



https://en.wikipedia.org/wiki/Continuous_uniform_distribution

- **Skewness=0**
  - Because it's symmetric like Normal Distribution

Application in Machine learning and Data Science

- **Random initialization**: Uniform distribution is used in machine learning (e.g., neural networks, k-means clustering) to randomly initialize parameters, ensuring equal probability for all values in a specified range.

- **Sampling**: Uniform distribution helps randomly select a subset of data that is representative of all classes, especially when each class has an equal number of samples.
- **Data augmentation**: Uniform distribution can generate new data points within a specified range to artificially increase the size of a dataset.
- **Hyperparameter tuning**: Uniform distribution is used in hyperparameter tuning to sample from a defined range of hyperparameters, allowing exploration of the hyperparameter space.

Log Normal Distribution

1. What is a "Log"?

Imagine you have a magic machine that squishes big numbers into smaller ones.

- **Example**:
    - If you put `100` into the machine, it gives you `2` because 102=100102=100.
    - If you put `1000` into the machine, it gives you `3` because 103=1000103=1000.

This machine is called a **logarithm (log)**. It answers the question:

"What power do we need to raise 10 to, to get this number?"

For example, in base 10:

$$\log_{10}(1000) = 3$$

Because:

$$10^3 = 1000$$

Natural Log (`ln` or `log_e`) (We use this):

- Uses base **e** (Euler's number ≈ 2.718).
- Written as: `ln(x) = log_e(x)`
- Example: **ln(7.389) = 2**, because **e^2 ≈ 7.389**.
  - **ln(10) →** $e^{?} = 10$

---

2. What is "Exp"?

"Exp" (exponential) is the **opposite of log**. It takes a small number and makes it big!

- **Example**:
  - Exp(2)=10^2=100
  - Exp(3)=10^3=1000

Fun Fact: In math, "exp" often refers to *ex*, where ***e* (~2.718)** is a special number, but the idea is the same!

- `exp(1) = e^1 = 2.718`
- `exp(2) = e^2 = 7.389`
- `exp(0) = e^0 = 1`

- Most common in building internet applications

Notation	$\text{Lognormal}(\mu, \sigma^2)$

What is a Log-Normal Distribution?

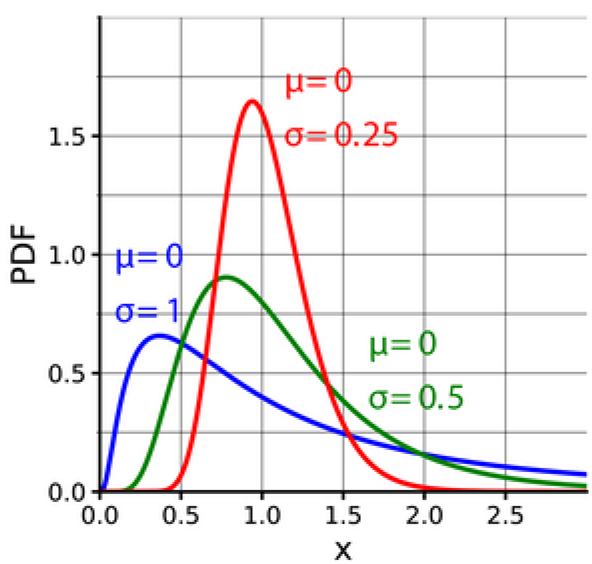
Imagine you plant a bunch of seeds. Every day, each plant grows by a **percentage** of its current height (like 10% taller each day).

- **Most plants** will stay small or medium.
- **A few plants** will grow SUPER tall, like magic beanstalks!

If you use the **magic ruler (log)**, all heights get squished into a bell curve:

- **10 cm → 1 (because log(10) ≈ 1).**
- **100 cm → 2 (because log(100) ≈ 2).**

This pattern of growth creates a **Log-Normal Distribution**. It's called "log-normal" because if you use a magic ruler (called a **logarithm**) to measure the heights, the magic ruler squishes the tall plants down, making the whole thing look like a **bell curve** (normal distribution)!



Characteristics

- **Right-skewed** (long tail on the right).
- **Defined for positive values only** (X>0).
- **Not symmetric**, unlike the normal distribution.

- **Mean ≠ Median ≠ Mode** (due to skewness).

Real-World Applications

The **Log-Normal Distribution** is widely used in:

- **Finance**: Stock prices, income distributions.
- **Engineering**: Failure rates of materials.
- **Biology**: Growth of bacteria, reaction times.
- **Machine Learning**: Modeling positive-skewed features.

> 💡 🛑 **Note:** Unlike normal distribution, Log-Normal does not have mean and std directly, because they change after taking the exponent.

Notation for log Normal:

- If a variable Y follows a Normal Distribution, then the exponential of Y, written as $X = e^Y$, follows a Log-Normal Distribution.

$$Y \sim N(\mu, \sigma^2) \quad \Rightarrow \quad X = e^Y \sim \text{Log-Normal}(\mu, \sigma)$$

Mathematical Relationship:

If $Y \sim N(\mu, \sigma^2)$, then

$X = e^Y$ follows a **Log-Normal Distribution.**

In simpler terms:

- If we take **log of a Log-Normal variable**, we get a **Normal Distribution**.
- If we **exponentiate a Normal variable**, we get a **Log-Normal Distribution**.

1. **Probability Density Function (PDF):**

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

   - μ: Mean of $\ln(x)$.
   - σ: Standard deviation of $\ln(x)$.

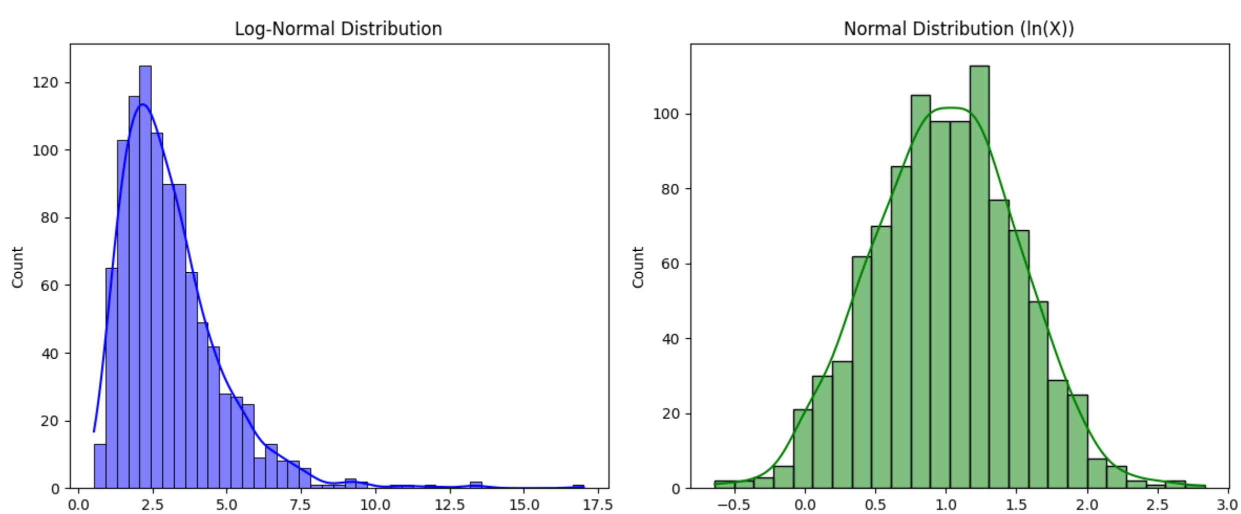
2. **Relationship to Normal Distribution:**

   If $Y = \ln(X)$ is Normal, then X is Log-Normal.

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

```
# Parameters for the underlying Normal Distribution
mu = 1.0    # Mean of ln(X)
sigma = 0.5 # Std dev of ln(X)

# Generate Log-Normal data
data = np.random.lognormal(mean=mu, sigma=sigma, size=1000)
```

```
plt.figure(figsize=(12, 5))

# Plot Log-Normal Data
plt.subplot(1, 2, 1)
sns.histplot(data, kde=True, color='blue')
plt.title('Log-Normal Distribution')

# Plot Normal Data (Log-Transformed)
plt.subplot(1, 2, 2)
sns.histplot(np.log(data), kde=True, color='green')
plt.title('Normal Distribution (ln(X))')

plt.tight_layout()
plt.show()
```

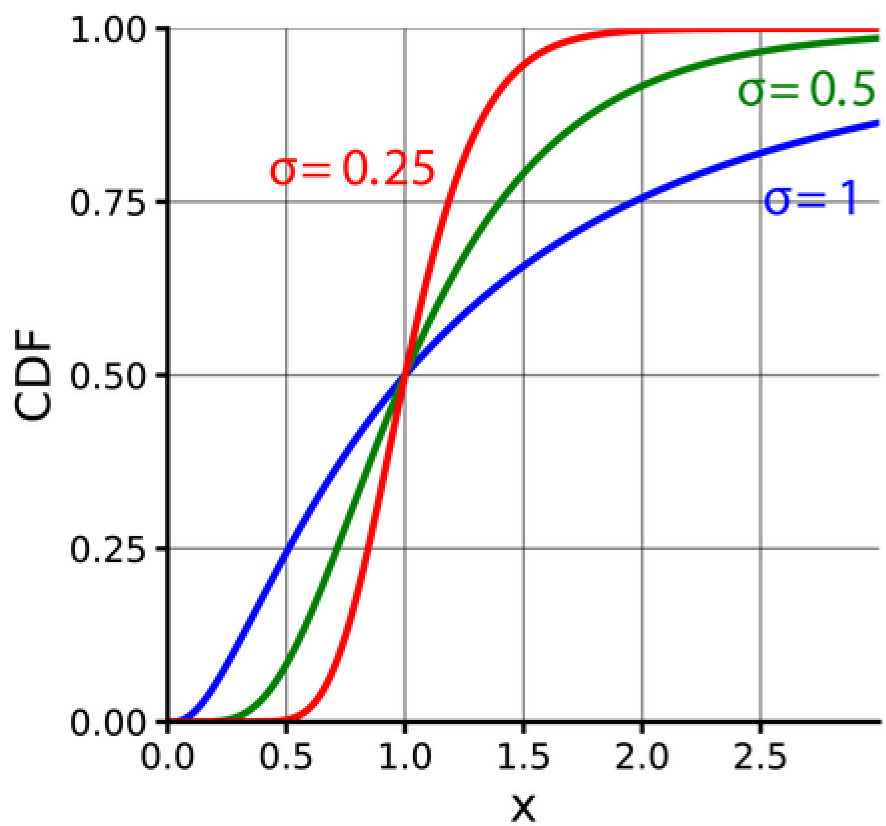


`plt.subplot(1, 2, 1)` : Divides the figure into a **1×2 grid** and selects the **1st cell** (left side).

`plt.subplot(1, 2, 2)` : Selects the **2nd cell** (right side).

`plt.tight_layout()` : Fixes spacing between plots to avoid overlapping.

CDF of Log Normal Distribution



Q. How to check if a random variable is log normally distributed?

Step 1: Take the Natural Log (`ln`) of the Data

- Convert the dataset **X** into **log(X)** using `np.log()`.
- If **log(X) is normally distributed**, then **X is log-normal**.

Step 2: Check if `log(X)` is Normally Distributed

We can check normality using multiple methods:

🔶 2.1 Visual Methods

1️⃣ Histogram & KDE Plot (Check Shape)

- Plot the histogram and Kernel Density Estimation (KDE) of **log(X)**.
- If it looks like a **bell curve**, then it is approximately normal.

2️⃣ Q-Q Plot (Quantile-Quantile Plot)

- Compare `log(X)` against a theoretical normal distribution.
- If points lie on a straight line, `log(X)` is normally distributed.