Probability Distributions

What is a Probability Distribution?

Imagine you're analyzing real-world data — sales, weather, clicks, heights, etc. You'll notice:

- Some values are more **likely** than others.
- There's a **pattern** to how data is spread.

That pattern is described by a **probability distribution**.

In short:

> **A probability distribution tells you how likely different outcomes are.**

What is a Probability Distribution?

A **probability distribution** is a **mathematical function** or **table** that:

- Assigns **probabilities** to all possible outcomes of a random process.

There are two types:

1. Discrete Probability Distribution

- For outcomes you can **count** (e.g. number of heads in 3 coin tosses).
- Example: **Binomial Distribution**

> Think: "What's the probability I get exactly 2 heads in 3 coin tosses?"

---

2. Continuous Probability Distribution

- For outcomes that can be **any number in a range** (e.g. height, weight).
- Example: **Normal Distribution**

> Think: "What's the probability someone's height is between 165 cm and 170 cm?"

A Simple Analogy

Let's say you roll a die:

- Outcomes = {1, 2, 3, 4, 5, 6}
- Each has a probability of `1/6`

This is a **uniform distribution** (discrete).

Now imagine measuring people's **heights**:

- You don't get fixed values.
- Instead, you get a **curve** — most people around average height, fewer very short or very tall.

That's a **normal distribution** (continuous).

Why Are Distributions Useful in Data Science?

1. **Model real-world randomness** (user behavior, errors, arrivals, etc.)
2. **Make predictions** (how likely is a customer to buy?)
3. **Run simulations**

Summary:

Type	Example	Used For
Discrete	Binomial, Poisson	Count of events
Continuous	Normal, Uniform	Measuring quantities

Uniform Distribution

What It Is:

> A **uniform distribution** is when **every outcome is equally likely**.

Simple Example: Rolling a Fair Die

- Possible outcomes: {1, 2, 3, 4, 5, 6}
- Each number has a **1/6** chance → That's a **discrete uniform distribution**

Continuous Version:

Let's say we randomly pick a number between 0 and 1.

- Every value in that range is **equally likely**.
- That's a **continuous uniform distribution**.

Graphs to Visualize

1. Discrete Uniform (like a die roll)

```
Outcome:     1   2   3   4   5   6
Probability:|---|---|---|---|---|---|
             1/6 for each → flat bars
```

Discrete Uniform Distribution Formula:

P(X = x) = 1 / n

- Where `n` is the number of possible outcomes (e.g., for a 6-sided die, P(rolling a 4) = 1/6)

2. Continuous Uniform (0 to 1)

- It's just a **flat horizontal line** from x = 0 to x = 1
- The probability density is constant (say 1.0) across that interval

Continuous Uniform Distribution Formula:

f(x) = 1 / (b - a) for values between `a` and `b`

- Outside the range `a` to `b`, the probability is 0

Why It's Useful in Data Science

- It models **pure randomness**
- It's used to **simulate random choices**
- Used in Generating random numbers (`np.random.uniform`)

In Code (Python):

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

# Continuous uniform from 0 to 1
samples = np.random.uniform(0, 1, 10000)

plt.hist(samples, bins=50, density=True, alpha=0.6, color='skyblue')
plt.title("Continuous Uniform Distribution (0 to 1)")
plt.xlabel("Value")
```

```
plt.ylabel("Probability Density")
plt.grid(True)
plt.show()
```

You'll see a **flat histogram** showing uniform probability.

Summary:

Property	Uniform Distribution
Type	Discrete or Continuous
Shape	Flat
Real-world example	Die roll, random number gen
Python function	`np.random.uniform(a, b)`

What is the Binomial Distribution?

> The **binomial distribution** models the number of **successes** in a fixed number of **independent yes/no experiments**, where each has the **same probability** of success.

Think of this:

- Toss a coin 10 times
- What's the probability of getting **exactly 6 heads**?

That's a binomial problem.

Key Ingredients:

- `n` = number of trials (e.g., 10 tosses)
- `p` = probability of success (e.g., 0.5 for heads)
- `x` = number of successes (e.g., 6 heads)

Plain Text Formula:

P(X = x) = C(n, x) × p^x × (1 - p)^(n - x)

Where:

- `C(n, x)` is "n choose x" = combinations = number of ways to pick `x` successes out of `n`
- `p^x` is the probability of `x` successes
- `(1 - p)^(n - x)` is the probability of the remaining being failures

Example:

10 coin tosses, what's the probability of exactly 6 heads?

- n = 10
- x = 6
- p = 0.5

$P(6heads) = C(10, 6) * 0.5^6 * 0.5^4 = 210 * (0.015625) * (0.0625) \approx 0.205$

So there's a ~20.5% chance you'll get exactly 6 heads in 10 tosses.

In Python:

```
from scipy.stats import binom

# Probability of exactly 6 heads in 10 tosses (p = 0.5)
prob = binom.pmf(k=6, n=10, p=0.5)
print(prob)  # Output: ~0.205
```

When to Use:

- Email campaign: Will 40 out of 100 people click the link?
- Quality check: How many out of 10 products will be defective?
- A/B testing: Will 60 out of 200 visitors convert?

Summary:

Concept	Value
Type	Discrete

Concept	Value
Formula	P(X = x) = C(n, x) * p^x * (1 - p)^(n - x)
Python	`scipy.stats.binom.pmf(x, n, p)`
Used for	Count of successes in repeated trials

Normal Distribution (a.k.a. Gaussian Distribution)

Before we dive into the details, let’s understand two key concepts related to normal distribution, mean, and standard deviation.

1. What is Mean?

The **mean** (or average) of a dataset is the sum of all values divided by the number of values.

Formula:

mean = (x1 + x2 + x3 + ... + xn) / n

It represents the central value of the data.

---

2. What is Standard Deviation?

The **standard deviation** measures how spread out the numbers are from the mean.

Steps to calculate standard deviation:

1. Find the mean
2. Subtract the mean from each value and square the result
3. Take the average of these squared differences (this is the variance)
4. Take the square root of the variance

Formula:

standard deviation (sigma) = sqrt((1/n) * sum((xi - mean)^2))

A smaller standard deviation means the data points are close to the mean. A larger standard deviation means the data is more spread out.

What is Normal Distribution?

> A **normal distribution** is a continuous probability distribution that is **bell-shaped and symmetric** around the mean.

It describes variables where:

- Most values cluster around the **average (mean)**.
- Extreme values (very high or low) are **rare**.

Think of:

- Heights of people
- Test scores
- Measurement errors

All tend to follow a **normal curve**.

Key Properties:

- **Mean (μ):** Center of the distribution
- **Standard Deviation (σ):** Spread of the distribution

Shape of the Curve:

- Bell-shaped
- Symmetrical
- Peaks at the mean
- About **68%** of the data lies within **$\pm 1\sigma$**, **95%** within **$\pm 2\sigma$**, **99.7%** within **$\pm 3\sigma$** → This is the famous **68–95–99.7 Rule**

Formula (Plain Text):

f(x) = (1 / (σ * sqrt(2π))) * e^(-(x - μ)² / (2σ²))

Where:

- `μ` = mean
- `σ` = standard deviation
- `e` = Euler's number (≈ 2.718)
- `π` = pi (≈ 3.14159)

This gives the **probability density** for a given value `x`.

> You **don't need to memorize** this formula — but understanding its shape and behavior is essential.

In Python:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# Generate values
x = np.linspace(-4, 4, 1000)
mean = 0
std_dev = 1

# Get the probability density
y = norm.pdf(x, loc=mean, scale=std_dev)

# Plot
plt.plot(x, y)
plt.title("Standard Normal Distribution (μ=0, σ=1)")
plt.xlabel("x")
plt.ylabel("Probability Density")
```

```
plt.grid(True)
plt.show()
```

Summary

Property	Value
Type	Continuous
Shape	Bell curve
Key parameters	Mean (μ), Std. Dev (σ)
Formula	$f(x) = (1 / (\sigma\sqrt{2\pi})) * e\text{^}(-(x - \mu)^2 / 2\sigma^2)$
Python	`scipy.stats.norm.pdf(x, μ, σ)`

Central Limit Theorem Explained

1. What is the Mean?

The **mean** (or average) of a dataset is the sum of all values divided by the number of values.

Formula:

mean = (x1 + x2 + x3 + ... + xn) / n

It represents the central value of the data.

---

2. What is the Standard Deviation?

The **standard deviation** measures how spread out the numbers are from the mean.

Steps to calculate standard deviation:

1. Find the mean
2. Subtract the mean from each value and square the result
3. Take the average of these squared differences (this is the variance)
4. Take the square root of the variance

Formula:

standard deviation (sigma) = sqrt((1/n) * sum((xi - mean)^2))

A smaller standard deviation means the data points are close to the mean. A larger standard deviation means the data is more spread out.

---

3. Central Limit Theorem (CLT)

The **Central Limit Theorem** states:

> If you take many random samples of size n from any population (with finite mean and variance), then the distribution of the sample means will tend to be **approximately normal** as n becomes large — regardless of the shape of the original population.

---

4. Mathematical Expression

Let X1, X2, ..., Xn be n independent, identically distributed (i.i.d) random variables with:

- Mean = mu
- Standard deviation = sigma

Then the **sampling distribution of the sample mean** (denoted as $\bar{X}$) approaches a normal distribution with:

- Mean = mu
- Standard deviation = sigma / sqrt(n)