Distributions in Machine Learning: A Complete Guide with Visuals and Use Cases

# Why Study Distributions in ML?

Statistical distributions are foundational in machine learning. They help us:  
- Understand how data is spread  
- Model uncertainty  
- Choose the right algorithm  
- Perform probabilistic inference  
- Simulate data for experimentation  
  
Each distribution has a specific shape and purpose. Let’s explore the most used ones.

# 1. Uniform Distribution

Definition: Every value in the interval has equal probability.

## Use Cases:

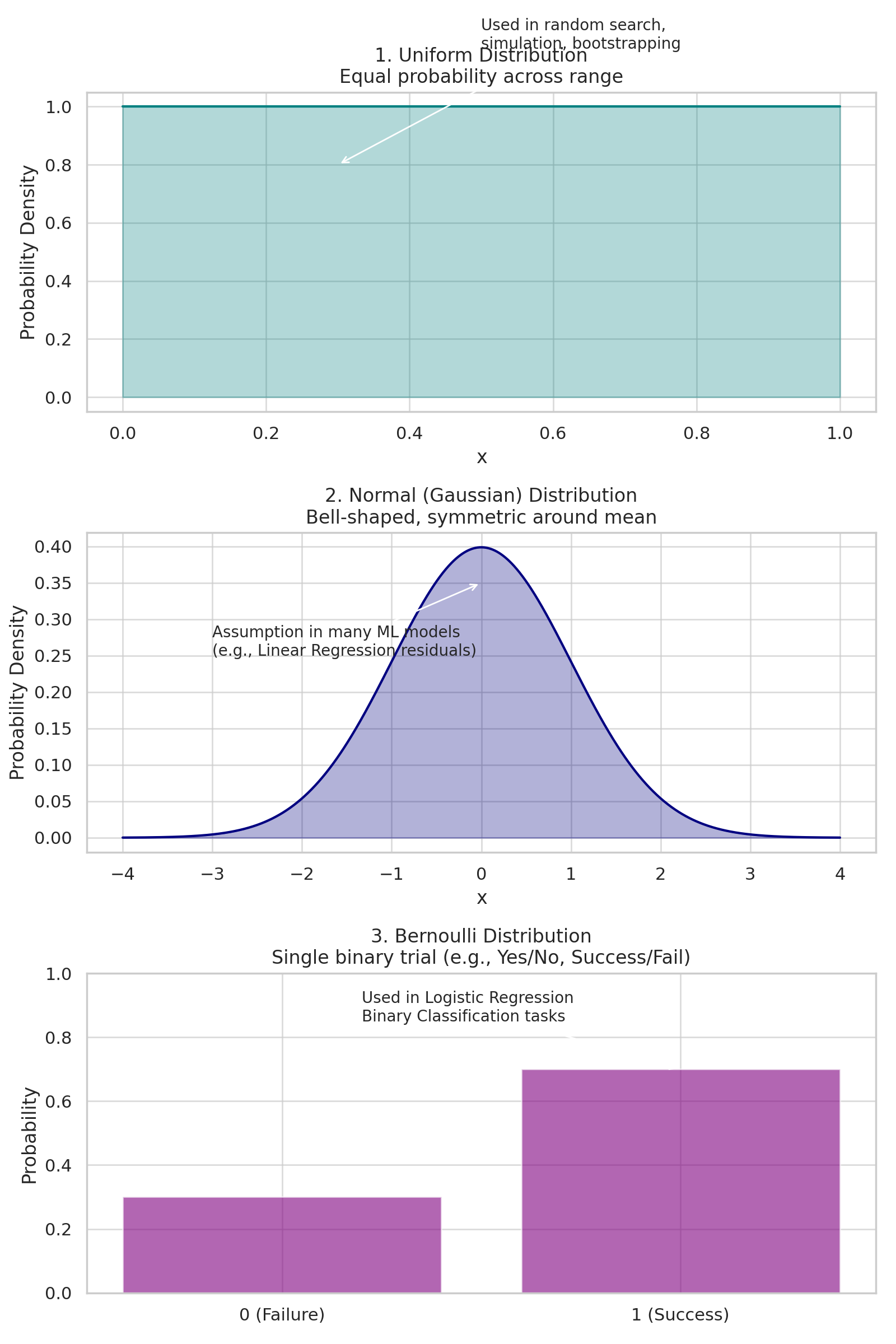
* - Random initialization of weights — Ensures each weight starts unbiased and uniformly spread
* - Bootstrapping — Sampling with replacement is uniform
* - Random search in hyperparameter tuning — All hyperparameter values are equally likely

## Formula:

f(x) = 1 / (b - a) for x in [a, b]

| **Feature** | **Description** |
| --- | --- |
| Mean | (a+b)/2 |
| Variance | ((b-a)^2)/12 |
| Shape | Flat |
| Skewness | None (perfectly symmetric) |

## Visual:



# 2. Normal (Gaussian) Distribution

Definition: A bell-shaped curve symmetric around the mean.

## Use Cases:

* - Assumption in Linear Regression (residuals) — Helps estimate parameters accurately
* - Feature scaling and normalization — Standard normal improves convergence in many algorithms
* - PCA (Principal Component Analysis) — Projects data along directions of maximal variance assuming Gaussian spread

## Formula:

f(x) = (1 / sqrt(2πσ²)) \* exp(-(x - μ)² / (2σ²))

# 3. Bernoulli Distribution

Definition: Models a binary outcome (success/failure) and only one trial.

## Use Cases:

* - Binary classification — Each prediction is either success (1) or failure (0)
* - Logistic regression outputs — Output is probability of success (between 0 and 1)

# 4. Binomial Distribution

Definition: Number of successes in a fixed number of independent Bernoulli trials.

The properties of a Binomial Distribution are:

1. Each trial is independent.
2. There are only two possible outcomes in a trial – success or failure.
3. A total number of n identical trials are conducted.
4. The probability of success and failure is the same for all trials. (Trials are identical.)

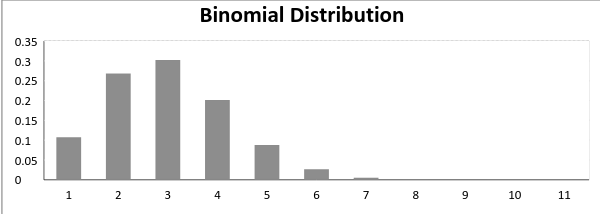
## Use Cases:

* - Classification metrics like precision/recall — These are based on multiple binary outcomes
* - Modeling user behavior outcomes — Like clicks, purchases out of total sessions

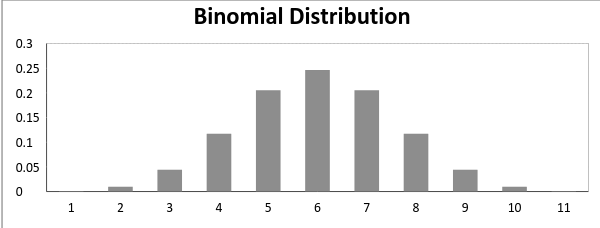
## Formula:

P(X = k) = (n choose k) \* p^k \* (1 - p)^(n - k)

A binomial distribution graph where the probability of success does not equal the probability of failure looks like this.



Now, when the probability of success = probability of failure, in such a situation, the graph of binomial distribution looks like

5. Poisson Distribution

Suppose you work at a call center; approximately how many calls do you get in a day? It can be any number. Now, the entire number of calls at a call center in a day is modeled by Poisson distribution. Some more examples are:

## The number of emergency calls recorded at a hospital in a day.

## The number of thefts reported in an area in a day.

## The number of customers arriving at a salon in an hour.

## The number of suicides reported in a particular city.

## The number of printing errors on each page of the book.

## Poisson Distribution is applicable in situations where events occur at random points of time and space wherein our interest lies only in the number of occurrences of the event.

A distribution is called a Poisson distribution when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal its probability over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Some notations used in Poisson distribution are:

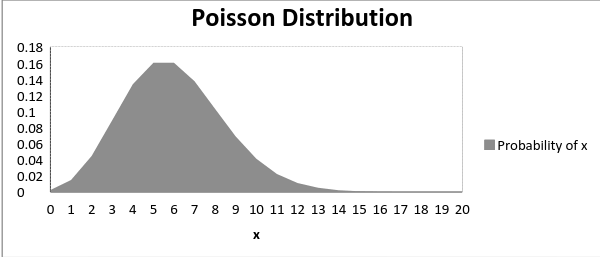
* λ is the rate at which an event occurs,
* t is the length of a time interval,
* And X is the number of events in that time interval.

Here, X is called a Poisson Random Variable, and the probability distribution of X is called Poisson distribution.

Let µ denote the mean number of events in an interval of length t. Then, µ = λ\*t.

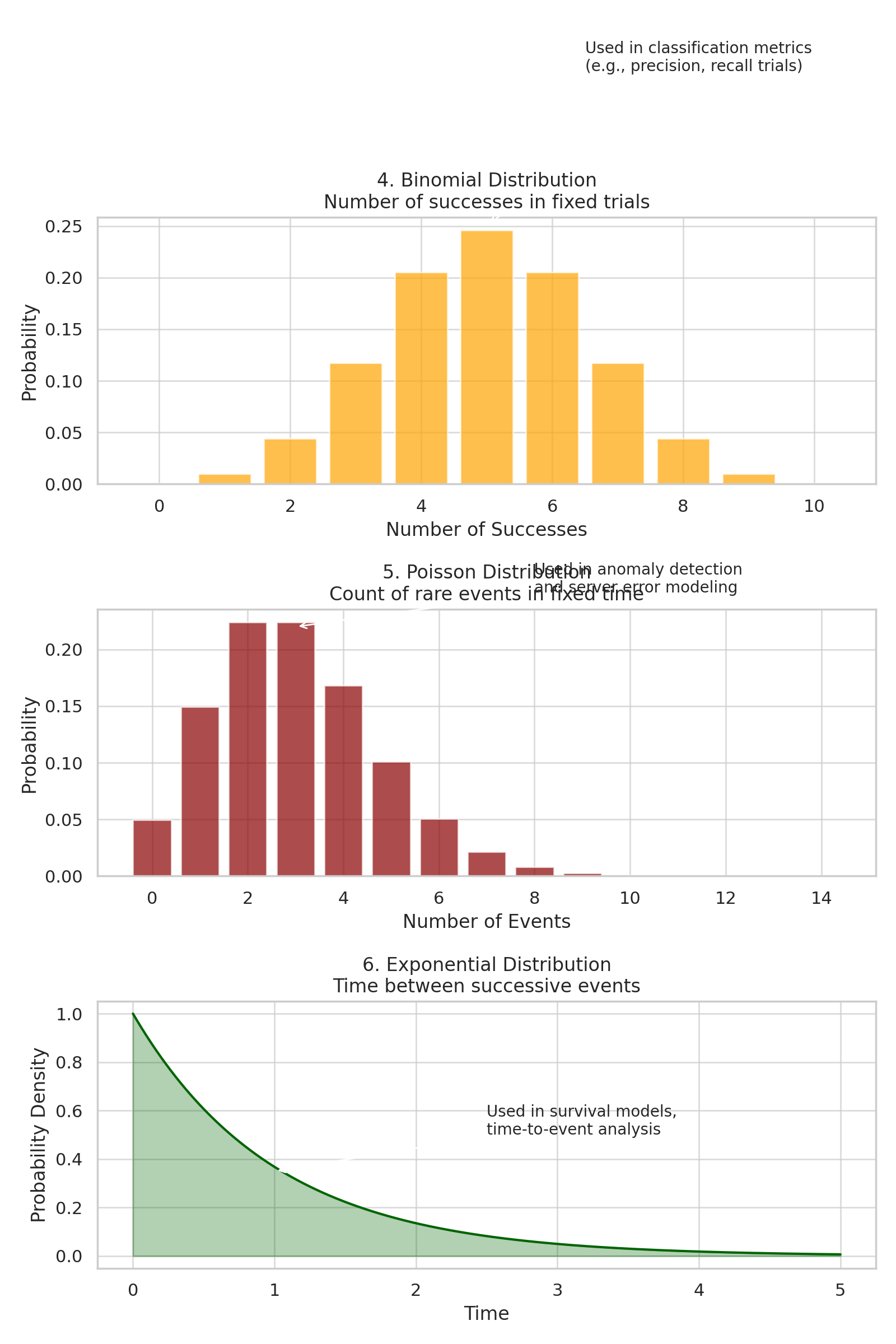
The PMF of X following a Poisson distribution is given by:





## Use Cases:

* - Modeling rare events (errors, traffic) — Predicts frequency over time
* - Anomaly detection — Sudden spike in rare event count can signal anomaly



# 6. Exponential Distribution

## Let’s consider the call center example one more time. What about the interval of time between the calls? Here, the exponential distribution comes to our rescue. Exponential distribution models the interval of time between the calls.

## Other examples are:

## Length of time between metro arrivals

## Length of time between arrivals at a gas station

## The life of an air conditioner

## The exponential distribution is widely used for survival analysis. From the expected life of a machine to the expected life of a human, exponential distribution successfully delivers the result.

## Use Cases:

* - Time-to-failure modeling — Determines when the next failure might occur
* - Survival analysis — Measures duration until event (like churn, death)

## Formula:

f(x) = λ \* e^(-λx), x ≥ 0

# 7. Multinomial Distribution

Definition: Generalization of binomial for multi-class outcomes.

## Use Cases:

* - Multiclass text classification — Predicts probabilities across multiple classes (e.g., spam, promotion, inbox)
* - Bag-of-words models — Count words occurring in multiple categories

## Visual:

# 8. Categorical Distribution

Definition: Generalization of Bernoulli for more than two categories.

## Use Cases:

* - Output of softmax layer — Converts logits into probabilities for each class
* - Label generation in multi-class problems — Useful when only one label out of many is true

# 9. Multivariate Normal Distribution

Definition: Joint distribution of multiple normally distributed variables.

## Use Cases:

* - Gaussian Mixture Models (GMMs) — Clustering with complex shaped data
* - PCA assumptions — Data spread along correlated dimensions
* - Anomaly detection — Low joint probability may imply an outlier

# 10. Student’s t-Distribution

Definition: Similar to normal but with heavier tails; useful when data has outliers or is based on small sample sizes.

## Use Cases:

* - Hypothesis testing — t-tests use this distribution to evaluate significance
* - Confidence intervals — Better estimates with small or noisy samples

