Chapter 2: Probability & Statistics for ML/NN/AI

Navid Seidi

Programmers House

August 2025

Why Probability & Stats for AI/ML?

The Language of Uncertainty and Data

Machine Learning models don't live in a perfect world. They need tools to handle imperfection and draw conclusions from data.

Why Probability & Stats for AI/ML?

- **Modeling Uncertainty:** Real-world data is noisy, incomplete, and random. Probability gives us a formal way to represent and work with this uncertainty.
- **Learning from Data:** How does a model learn? By finding patterns in data. Statistics provides the core methods (like mean, variance, and regression) to describe data and identify significant patterns.
- **Model Evaluation:** Is our model actually good? We use statistical tests and metrics (like accuracy, precision, and confidence intervals) to rigorously measure performance and prove that our model's success isn't just due to luck.
- **Algorithm Design:** Many of the most powerful ML algorithms are built directly from probability. For example, a "Naive Bayes" classifier uses Bayes' theorem to classify text, and neural networks often use a "softmax" output layer, which is a probability distribution.

Let’s Start With a Question...

What is Probability?

Think about it...

How would you define it in one sentence? What is the first example that comes to your mind?

A Common Misconception: The 50/50 Trap

Is this logic correct?

"I am going to play the lottery tomorrow. There are two outcomes: I win, or I don't win. Therefore, the probability I will win is 50%!"

"There might be a lion behind that door. It's either there or it isn't. So there's a 50% chance of a lion."

A Common Misconception: The 50/50 Trap

Is this logic correct?

"I am going to play the lottery tomorrow. There are two outcomes: I win, or I don't win. Therefore, the probability I will win is 50%!"

"There might be a lion behind that door. It's either there or it isn't. So there's a 50% chance of a lion."

Why This is Wrong: Outcomes Must Be EQUALLY LIKELY

The simple formula '(Favorable / Total)' only works if all outcomes have the same chance of happening.

- **Lottery:** There is only ONE way to win, but MILLIONS of ways to lose. The outcomes are not equally likely.
- **Lion:** Based on our knowledge of the world, "no lion" is a much, much more likely outcome than "lion".

This Leads to Two Types of Problems...

Scenario A: A Game of Chance

What is the probability of rolling a "6" on a fair die?

Scenario B: A Prediction

What is the probability of rain tomorrow?

This Leads to Two Types of Problems...

Scenario A: A Game of Chance

What is the probability of rolling a "6" on a fair die? **Answer: 1/6**.

Easy! The 6 outcomes are equally likely.

Scenario B: A Prediction

What is the probability of rain tomorrow? **Problem:** We can't roll

"tomorrow" like a die. The outcomes aren't simple or equally likely. What does "probability" mean here?

Two Ways of Thinking About Probability

1. Frequentist Probability

Probability as Frequency.

- The probability of an event is its long-term frequency over many repeated trials.
- **Works for:** Repeatable experiments like coin flips, card games, dice rolls.
- This is an **objective** property of the world.

2. Bayesian Probability

Probability as Belief.

- Probability is a measure of our confidence in a statement, given evidence.
- **Works for:** Non-repeatable events like weather forecasts or medical diagnoses.
- This is a **subjective** state of knowledge.

How Do We Use These Ideas in AI?

Our Focus: Building a Foundation

We will start with the **Frequentist** approach. The tools we learn here (counting, distributions, measuring error) are the essential foundation for everything that follows.

How AI Uses Both Schools of Thought

In practice, Machine Learning is a powerful mix of both ideas:

- **We build models like a Bayesian:** We start with an initial model (a "prior belief") and update its parameters as it sees more data (the "evidence"). This is the core of what "learning" is.
- **We evaluate models like a Frequentist:** To know if our model is good, we test it on a lot of data and measure its performance. When we say a model has "95% accuracy," we are using a frequentist idea—the frequency of it being correct over many trials.

One Set of Tools, Two Philosophies

The Good News: The Math is the Same!

No matter which philosophy you follow, the mathematical language of probability is universal. The tools we learn for the Frequentist view are the **exact same tools** used by Bayesians.

The Universal Toolbox of Probability

Both schools of thought rely on:

- The Axioms of Probability (the fundamental rules).
- The formula for Conditional Probability.
- Probability Distributions (like Normal, Binomial, etc.) to model events.
- Bayes' Theorem (yes, even frequentists use it!).

The difference is not in the math, but in **what we apply it to**: repeatable events vs. degrees of belief.

Topic 1: Basic Probability

Simple Definition: A number between 0 and 1 that shows how likely an event is to happen. 0 means impossible, 1 means certain.

Formal Definition: The probability of an event E is the ratio of favorable outcomes to the total number of possible outcomes, assuming all outcomes are equally likely.

$$P(E) = \frac{\text{Number of Favorable Outcomes}}{\text{Total Number of Outcomes}}$$

Intuitive Example: In a bag with 3 red balls and 2 blue balls, the probability of picking a red ball is the chance you will grab a red one without looking.

Topic 2: Conditional Probability

Simple Definition: The probability of an event happening, given that another event has already happened.

Formal Definition: The probability of event A given event B is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{where } P(B) > 0$$

Intuitive Example: What is the probability a card is a "King", **given that** we know it is a "Face Card"? The new information reduces our total possible outcomes.

Topic 3: Bayes' Theorem

Simple Definition: A mathematical rule for updating our beliefs after we get new evidence. It's the engine of learning from data.

Intuitive Example: You think your friend is late because of traffic (your "prior" belief). Then you get a text saying their car broke down (the "evidence"). You update your belief to be certain it's a car problem, not traffic (your "posterior" belief).

The Formula and Its Parts:

$$\underbrace{P(A \mid B)}_{\text{Posterior}} = \frac{\overbrace{P(B \mid A)}^{\text{Likelihood}}\overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Evidence}}}$$

- **Prior:** What we believed before seeing the new data.
- **Likelihood:** How likely is our evidence, given our belief?
- **Posterior:** Our new, updated belief.

Bayes' Theorem in Python

Scenario

Your friend is late 30% of the time ($P(\text{Late}) = 0.3$). If they are late, they are stuck in traffic 80% of the time ($P(\text{Traffic} \mid \text{Late}) = 0.8$). On any given day, the chance of traffic is 40% ($P(\text{Traffic}) = 0.4$). If you see there is traffic, what is the probability your friend is late?

Bayes' Theorem in Python

```python
# What we know (our priors and likelihoods)
p_late = 0.3 # Prior probability of being late
p_traffic = 0.4 # Probability of traffic (the evidence
    )
p_traffic_given_late = 0.8 # Likelihood of traffic if
    they are late

# We want to find P(Late | Traffic)
# Apply Bayes' Theorem: P(A|B) = (P(B|A) * P(A)) / P(B
    )
p_late_given_traffic = (p_traffic_given_late * p_late)
     / p_traffic

print(f"Our updated belief that our friend is late: {
    p_late_given_traffic:.2f}")
# Output: 0.60
```

Quiz: Bayes' Theorem

What is the primary purpose of Bayes' Theorem?

1. To calculate the average of a dataset.
2. To find the probability of two independent events happening at once.
3. To formally update a probability when you receive new information.
4. To measure how spread out your data is.

Quiz: Bayes' Theorem

What is the primary purpose of Bayes' Theorem?

1. To calculate the average of a dataset.
2. To find the probability of two independent events happening at once.
3. To formally update a probability when you receive new information.
4. To measure how spread out your data is.

Answer: (3). It provides a mathematical way to change your mind based on evidence.

Topic 4: Random Variables

Simple Definition: A variable whose value is a number determined by a random event.

Two Types:

- **Discrete:** Can take a countable number of values (e.g., result of a die roll: 1, 2, 3, 4, 5, 6).
- **Continuous:** Can take any value in a range (e.g., the height of a person).

Intuitive Example: We flip a coin 3 times. Let the random variable X be "the number of heads". The possible values for X are $\{0, 1, 2, 3\}$.

Topic 5: Probability Distributions

Simple Definition: A description of how likely each possible value of a random variable is. It can be a table, graph, or formula.

Formal Definition: A function that provides the probabilities of occurrence of all different possible outcomes in an experiment.

Intuitive Example: For the "number of heads in two coin flips" random variable X, the distribution is a simple table:

Outcome (x**)** & **Probability** $P(X = x)$
0 Heads & 1/4 or 0.25
1 Head & 2/4 or 0.50
2 Heads & 1/4 or 0.25

Describing Distributions: PMF, PDF, CDF

How do we describe the probabilities for a random variable?

- **PMF (Probability Mass Function):** For **discrete** variables. Gives the probability of a specific outcome.
  - Example: $P(X = k)$. What is the probability of rolling exactly a 3?
- **PDF (Probability Density Function):** For **continuous** variables. The area under the curve over a range gives the probability.
  - The probability of any single exact value is 0! We ask about ranges.
  - Example: What is the probability a person's height is between 170cm and 180cm?
- **CDF (Cumulative Distribution Function):** For both types. Gives the probability of getting a value **less than or equal to** x.
  - Example: $P(X \leq x)$. What is the probability of rolling a 3 or less?

Sampling From a Distribution

Simple Definition: ”Sampling” or ”drawing a sample” means generating a random value according to the rules of a distribution.

Notation in Literature: The tilde symbol ‘ ‘ means ”is distributed as”.

Examples:

- $x \sim \mathcal{U}(1,6)$: We are drawing a sample x from a Uniform distribution with integer outcomes from 1 to 6 (like a die roll).
- $h \sim \mathcal{N}(175, 6^2)$: A person’s height h is sampled from a Normal distribution with a mean of 175cm and a standard deviation of 6cm.

This notation is a very common and compact way to describe where your data comes from.

Topic 6: Mean (Expected Value)

Simple Definition: The long-term average value of a random variable.

Formal Definition: For a discrete random variable X, the expected value, denoted μ or $E[X]$, is:

$$\mu = E[X] = \sum_x x \cdot P(X = x)$$

Intuitive Example: The expected value of a six-sided die roll is 3.5.

$$E[X] = (1 \cdot \frac{1}{6}) + (2 \cdot \frac{1}{6}) + \cdots + (6 \cdot \frac{1}{6}) = 3.5$$

Topic 7: Variance and Standard Deviation

Simple Definition: Measures of how "spread out" the data is from the mean.

- **Variance:** The average of the squared differences from the Mean. High variance = very spread out.
- **Standard Deviation:** The square root of the variance. Easier to interpret because it is in the same units as the mean.

Formal Definitions:

- Variance: $\sigma^2 = \text{Var}(X) = E[(X - \mu)^2]$
- Standard Deviation: $\sigma = \sqrt{\text{Var}(X)}$

Mean, Var, Std Dev in Python

```python
import numpy as np
# Scores for two different classes
class_A_scores = [85, 86, 84, 85, 85] # Low spread
class_B_scores = [60, 100, 70, 95, 80] # High spread
# --- Class A ---
mean_A = np.mean(class_A_scores)
var_A = np.var(class_A_scores)
std_A = np.std(class_A_scores)
print(f"Class A: Mean={mean_A:.1f}, Var={var_A:.2f},
   Std Dev={std_A:.2f}")
# --- Class B ---
mean_B = np.mean(class_B_scores)
var_B = np.var(class_B_scores)
std_B = np.std(class_B_scores)
print(f"Class B: Mean={mean_B:.1f}, Var={var_B:.2f},
   Std Dev={std_B:.2f}")
```

Quiz: Spread

A weather forecast says the average temperature tomorrow will be 25°C. Which forecast implies the temperature will be more stable and predictable?

1. Mean: 25°C, Standard Deviation: 5°C
2. Mean: 25°C, Standard Deviation: 1°C
3. Mean: 30°C, Standard Deviation: 1°C
4. Mean: 20°C, Standard Deviation: 5°C

Quiz: Spread

A weather forecast says the average temperature tomorrow will be 25°C. Which forecast implies the temperature will be more stable and predictable?

1. Mean: 25°C, Standard Deviation: 5°C
2. Mean: 25°C, Standard Deviation: 1°C
3. Mean: 30°C, Standard Deviation: 1°C
4. Mean: 20°C, Standard Deviation: 5°C

Answer: (2). A smaller standard deviation means the values are less spread out and closer to the mean.

Binomial Distribution

What it is: Describes the number of successes in a fixed number of independent trials.

Intuitive Example: Flip a coin 10 times. What is the probability of getting exactly 7 heads?

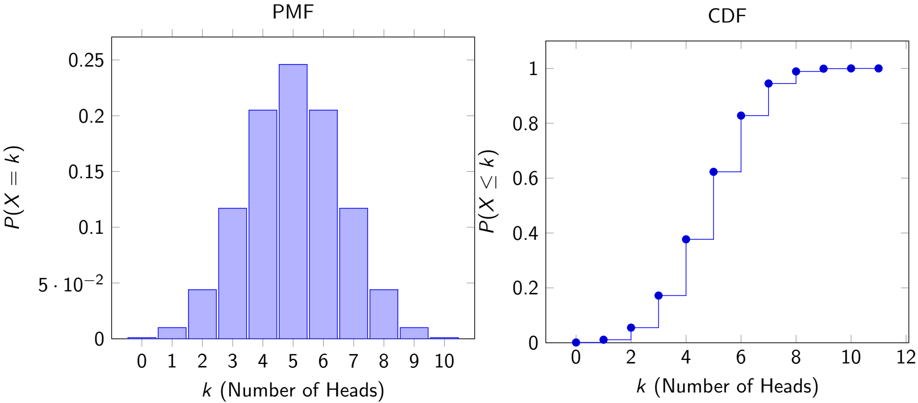
Parameters:

- n: number of trials (e.g., 10 coin flips).
- p: probability of success on a single trial (e.g., 0.5 for heads).

PMF Formula: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

Binomial Distribution: Graphs

PMF and CDF for Binomial(n=10, p=0.5)



Normal (Gaussian) Distribution

What it is: The most famous continuous distribution. It is a symmetric "bell curve".

Intuitive Example: The distribution of human heights, blood pressure, or measurement errors. Many natural things follow this pattern.

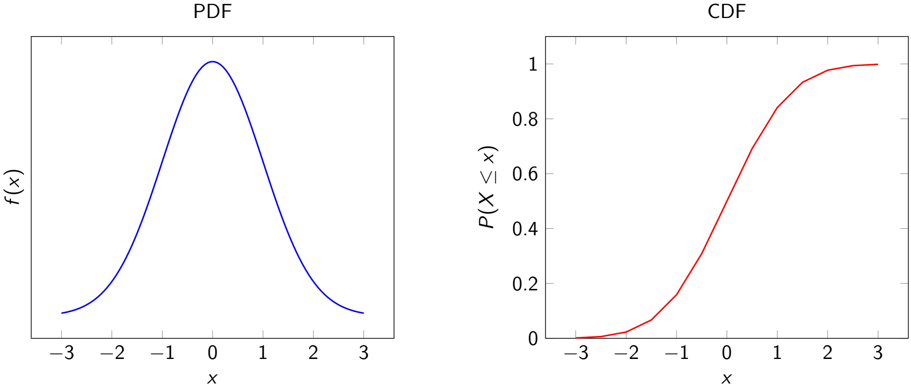
Parameters:

- μ: the mean, which is the center of the bell.
- σ^2: the variance, which controls how wide or narrow the bell is.

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

Normal Distribution: Graphs

PDF and CDF for Normal($\mu = 0, \sigma = 1$)



PDF

CDF

Uniform Distribution

What it is: A distribution where all outcomes in a range are equally likely.

Intuitive Example: Rolling a fair six-sided die. Each number from 1 to 6 has an equal probability (1/6).

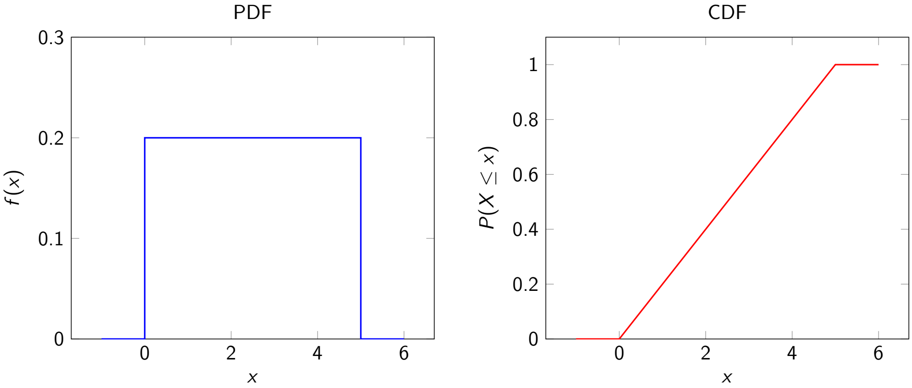
Parameters:

- a: the minimum value.
- b: the maximum value.

PDF Formula: $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$, and 0 otherwise.

Uniform Distribution: Graphs

PDF and CDF for Uniform(a=0, b=5)



What You Now Know

- The two main philosophies of probability (Frequentist and Bayesian).
- Key concepts like Conditional Probability, Random Variables, PDF, and CDF.
- How to describe data with Mean (μ) and Standard Deviation (σ).
- The meaning of sampling ($x \sim \mathcal{D}$).
- The properties and uses of three major distributions:
  - **Binomial:** For counting successes in trials.
  - **Normal:** For continuous data that clusters around an average.
  - **Uniform:** When all outcomes are equally likely.

Example 1: A Frequentist Problem

The Scenario: Drawing Cards

We have a standard, well-shuffled deck of 52 cards. This is a perfect frequentist problem because the total number of outcomes is fixed and known.

The Question

What is the probability of drawing two Kings in a row from the deck, without replacing the first card?

Solution 1: A Frequentist Problem

The Logic: We calculate the probability of the first event, and then the probability of the second event given the first has occurred.

- Probability of the 1st card being a King: $P(\text{1st is King}) = \frac{4}{52}$
- Given the 1st was a King, there are now 51 cards left, and only 3 Kings.
- Probability of the 2nd card being a King: $P(\text{2nd is King} \mid \text{1st was King}) = \frac{3}{51}$
- The total probability is the product of these two: $P(\text{Both Kings}) = \frac{4}{52} \times \frac{3}{51}$

Solution 1: A Frequentist Problem

```
# Initial conditions
total_cards = 52
kings_in_deck = 4
# Probability of the first card being a King
p_first_king = kings_in_deck / total_cards
# After drawing one king, the deck changes
cards_left = total_cards - 1
kings_left = kings_in_deck - 1
# Probability of the second card being a King
p_second_king_given_first = kings_left / cards_left
# Total probability is the product of the two events
p_both_kings = p_first_king *
    p_second_king_given_first
print(f"The probability is: {p_both_kings:.4f}") #
    Output: 0.0045
```

Example 2: A Bayesian Problem

The Scenario: Medical Testing

A specific disease is present in 1 out of every 1000 people. A test for the disease is very good, but not perfect. This is a classic Bayesian problem where we need to update our belief based on new evidence (the test result).

The Question

The test correctly identifies 99% of people who have the disease, and has a 2% false positive rate (it incorrectly says people have the disease 2% of the time). If a random person tests positive, what is the actual probability they have the disease?

Solution 2: A Bayesian Problem

The Logic: We use Bayes' Theorem to update our prior belief $P(\text{Disease})$ with the evidence $P(\text{Positive})$.

- $P(D) = 0.001$ (Prior probability of having the disease)
- $P(\text{Pos} \mid D) = 0.99$ (Test is positive, given you have the disease)
- $P(\text{Pos} \mid \text{Not } D) = 0.02$ (Test is positive, given you don't have it)
- We want to find $P(D \mid \text{Pos})$.

Solution 2: A Bayesian Problem

```
disease_prevalence = 1 / 1000
test_accuracy = 0.99 # P(Pos | Disease)
false_positive_rate = 0.02 # P(Pos | No Disease)
p_disease = disease_prevalence
p_no_disease = 1 - p_disease
p_pos_given_disease = test_accuracy
p_pos_given_no_disease = false_positive_rate
# Calculate the total probability of a positive test (
    the evidence)
# P(Pos) = P(Pos|D)P(D) + P(Pos|Not D)P(Not D)
p_positive = (p_pos_given_disease * p_disease) + \
             (p_pos_given_no_disease * p_no_disease)
# Apply Bayes' Theorem => P(D|Pos) = (P(Pos|D) * P(D))
     / P(Pos)
p_disease_given_pos = (p_pos_given_disease * p_disease
   ) / p_positive
print(f"The probability is: {p_disease_given_pos:.3f}"
   ) # Output: 0.047
```

Example 3: The AI Combination

The Scenario: Spam Filtering

This is a real-world AI problem. We use frequentist statistics from our data to build a Bayesian model that makes predictions.

The Question

We have a dataset of 1000 emails. We observe that 300 are Spam. The word ”lottery” appears in 100 of the spam emails, but only 5 of the non-spam emails.

If a new email arrives containing the word ”lottery”, what is the probability that it is spam?

Solution 3: The AI Combination

Step 1 (Frequentist): Calculate probabilities from our data.

- $P(\text{Spam}) = 300/1000 = 0.3$
- $P(\text{"lottery"} \mid \text{Spam}) = 100/300 \approx 0.333$
- $P(\text{"lottery"} \mid \text{Not Spam}) = 5/700 \approx 0.007$

Step 2 (Bayesian): Use these stats in Bayes' Theorem to predict.

Solution 3: The AI Combination

```
total_emails = 1000
spam_emails = 300
not_spam_emails = total_emails - spam_emails
lottery_in_spam = 100
lottery_in_not_spam = 5
p_spam = spam_emails / total_emails
p_not_spam = not_spam_emails / total_emails
p_lottery_given_spam = lottery_in_spam / spam_emails
p_lottery_given_not_spam = lottery_in_not_spam /
   not_spam_emails
# First, find the total probability of seeing the word
    "lottery"
p_lottery = (p_lottery_given_spam * p_spam) + \
            (p_lottery_given_not_spam * p_not_spam)
# Now, apply Bayes’ Theorem to make the prediction
p_spam_given_lottery = (p_lottery_given_spam * p_spam)
    / p_lottery
print(f"Probability it’s spam if it contains ’lottery
   ’: {p_spam_given_lottery:.2f}") # Output: 0.95
```