

CPEN 455 25W1 TUTORIAL

PROBABILITY AND STATISTICS¹

Qi Yan and Felix Fu

University of British Columbia

September 8th, 2025

¹Based on materials from Goodfellow et al., 2016; Murphy, 2022; Zhang et al., 2023.

CONTENTS

1	Probability	2
2	Random Variables	4
2.1	Discrete Variables and PMFs	5
2.2	Continuous Variables and PDFs	6
3	Probability Distributions	7
3.1	Marginal Probability	7
3.2	Conditional Probability	8
3.3	The Chain Rule of Conditional Probabilities	9
3.4	Independence and Conditional Independence	10
4	Expectation, Variance and Covariance	11
4.1	Expectation and Expected Value	11
4.2	Variance and Covariance	12
5	A Bit of Information Theory	14
5.1	Entropy	14
5.2	KL Divergence and Cross-Entropy	15
6	References	18

PROBABILITY

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

- ▶ Probability: quantitative degree of belief.
 - an image classifier outputs a probability distribution given an input image.
 - a large language model (e.g., chatGPT) outputs a probability distribution over the next word when making predictions.
 - a generative model (VAEs/GANs/diffusion) starts generation from known prior distributions.

PROBABILITY

Probability theory is nothing but common sense reduced to calculation. — Pierre Laplace, 1812

- ▶ Probability: quantitative degree of belief.
 - an image classifier outputs a probability distribution given an input image.
 - a large language model (e.g., chatGPT) outputs a probability distribution over the next word when making predictions.
 - a generative model (VAEs/GANs/diffusion) starts generation from known prior distributions.
- ▶ **Frequentist** perspective: probabilities represent long run frequencies of **events** that can happen multiple times.
 - if we toss the coin many times, we expect it to land heads about half the time.
- ▶ **Bayesian** perspective: probability is used to quantify our **uncertainty** or ignorance about something; hence it is fundamentally related to information rather than repeated trials
 - we believe the coin is equally likely to land heads or tails on the next toss.

PROBABILITY

► **Probability Space:**

- A mathematical model of a random experiment.

PROBABILITY

- ▶ **Probability Space:**

- A mathematical model of a random experiment.

- ▶ **Sample Space:**

- The collection of all possible outcomes of the experiment.

PROBABILITY

► **Probability Space:**

- A mathematical model of a random experiment.

► **Sample Space:**

- The collection of all possible outcomes of the experiment.

► **Events:**

- Subsets of the sample space.
- An event *occurs* if the experiment's outcome belongs to that subset.
- Events are those subsets whose occurrence is noteworthy and decidable.

PROBABILITY

► Probability Space:

- A mathematical model of a random experiment.

► Sample Space:

- The collection of all possible outcomes of the experiment.

► Events:

- Subsets of the sample space.
- An event *occurs* if the experiment's outcome belongs to that subset.
- Events are those subsets whose occurrence is noteworthy and decidable.

► Axioms of Probability:

- **Non-negativity**: the probability of an event is a non-negative real number.
- **Unit measure**: the probability of the event representing the entire sample space occurs is 1.
- **σ -additivity**: any countable sequence of disjoint sets (mutually exclusive events) satisfies the fact that the probability of their union equals to the sum of their individual probabilities.

RANDOM VARIABLES

► **Random Variable:**

- A variable that maps events with potential various random values.
- Distinct from its possible values.

RANDOM VARIABLES

► **Random Variable:**

- A variable that maps events with potential various random values.
- Distinct from its possible values.

► **Nature of Random Variables:**

- A quantitative description of possible states.
- Requires a probability distribution to define likelihood of each state.

RANDOM VARIABLES

▶ **Random Variable:**

- A variable that maps events with potential various random values.
- Distinct from its possible values.

▶ **Nature of Random Variables:**

- A quantitative description of possible states.
- Requires a probability distribution to define likelihood of each state.

▶ **Types of Random Variables:**

- *Discrete Random Variable:*
 - ▶ Finite or countably infinite states.
 - ▶ States could be numerical or non-numerical.
- *Continuous Random Variable:*
 - ▶ Associated with real number values.

RANDOM VARIABLES

DISCRETE VARIABLES AND PMFS

Discrete Random Variables

Described using a **Probability Mass Function (PMF)**, typically denoted by P . Associates each possible state with a probability.

PMF Characteristics

- ▶ Maps a state of a random variable to its probability of occurrence.
- ▶ PMFs can describe joint probabilities for multiple variables, e.g., $P(x, y)$.
- ▶ Must satisfy two conditions:
 1. $0 \leq P(x) \leq 1$ for all states x .
 2. $\sum_x P(x) = 1$ (Normalization).

RANDOM VARIABLES

DISCRETE VARIABLES AND PMFS

Discrete Random Variables

Described using a **Probability Mass Function (PMF)**, typically denoted by P . Associates each possible state with a probability.

PMF Characteristics

- ▶ Maps a state of a random variable to its probability of occurrence.
- ▶ PMFs can describe joint probabilities for multiple variables, e.g., $P(x, y)$.
- ▶ Must satisfy two conditions:
 1. $0 \leq P(x) \leq 1$ for all states x .
 2. $\sum_x P(x) = 1$ (Normalization).

Example 2.1 (Uniform Distribution PMF)

Given a discrete random variable x with k states, a uniform distribution assigns:

$$P(x = x_i) = \frac{1}{k}$$

This satisfies the normalization condition since $\sum_i \frac{1}{k} = 1$.

RANDOM VARIABLES

CONTINUOUS VARIABLES AND PDFs

Continuous Random Variables

Probability Density Functions (PDFs) p describe probability distributions for continuous variables:

- ▶ Domain is all possible states of x .
- ▶ $p(x) \geq 0$ for all x in the domain.
- ▶ $\int p(x) dx = 1$ (Total probability is 1).

Understanding PDFs

- ▶ PDFs give the probability density, not the probability of specific states.
- ▶ Probability for an infinitesimal region is $p(x) dx$.
- ▶ The probability that x is in a set S is the integral of $p(x)$ over S .

RANDOM VARIABLES

CONTINUOUS VARIABLES AND PDFs

Continuous Random Variables

Probability Density Functions (PDFs) p describe probability distributions for continuous variables:

- ▶ Domain is all possible states of x .
- ▶ $p(x) \geq 0$ for all x in the domain.
- ▶ $\int p(x) dx = 1$ (Total probability is 1).

Understanding PDFs

- ▶ PDFs give the probability density, not the probability of specific states.
- ▶ Probability for an infinitesimal region is $p(x) dx$.
- ▶ The probability that x is in a set S is the integral of $p(x)$ over S .

Example 2.2 (Uniform Distribution PDF)

Consider $u(x; a, b)$ for a uniform distribution over $[a, b]$, where $a < b$:

$$u(x; a, b) = \begin{cases} 0 & \text{for } x \notin [a, b] \\ \frac{1}{b-a} & \text{for } x \in [a, b] \end{cases}$$

This function is always nonnegative and integrates to 1, representing the uniform distribution $x \sim U(a, b)$

PROBABILITY DISTRIBUTIONS

MARGINAL PROBABILITY

Concept of Marginal Probability

The probability distribution over a subset of variables, derived from a joint distribution of multiple variables.

Discrete Random Variables

Given discrete random variables x and y , and the joint distribution $P(x, y)$, the marginal probability $P(x)$ is calculated as:

$$\forall x \in X, P(x = x) = \sum_{y \in Y} P(x = x, y = y)$$

Continuous Random Variables

For continuous variables, marginal probability is found using integration:

$$p(x) = \int p(x, y) dy$$

Origin of the Term

The term "marginal" refers to the practice of summing probabilities in a table and writing the totals in the margins.

PROBABILITY DISTRIBUTIONS

CONDITIONAL PROBABILITY

Definition

The probability of an event given that another event has occurred, denoted as $P(y = y \mid x = x)$.

Computation Formula

$$P(y = y \mid x = x) = \frac{P(y = y, x = x)}{P(x = x)}$$

Note: Defined only if $P(x = x) > 0$.

Understanding Conditional Probability

- ▶ Not to be confused with the consequences of actions or interventions.
- ▶ The conditional probability of an event y given x is different from the probability that y would happen if x were to be caused by some action (called intervention query for causal modeling).

PROBABILITY DISTRIBUTIONS

THE CHAIN RULE OF CONDITIONAL PROBABILITIES

Chain Rule

A joint probability distribution can be decomposed into conditional probabilities:

$$P(x^{(1)}, \dots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^n P(x^{(i)} | x^{(1)}, \dots, x^{(i-1)})$$

Application of the Chain Rule

- ▶ Simplifies the computation of joint distributions.
- ▶ Derived from the definition of conditional probability.

PROBABILITY DISTRIBUTIONS

THE CHAIN RULE OF CONDITIONAL PROBABILITIES

Chain Rule

A joint probability distribution can be decomposed into conditional probabilities:

$$P(x^{(1)}, \dots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^n P(x^{(i)} | x^{(1)}, \dots, x^{(i-1)})$$

Application of the Chain Rule

- ▶ Simplifies the computation of joint distributions.
- ▶ Derived from the definition of conditional probability.

Example 3.1

Applying the rule to three variables a , b , and c :

$$P(a, b, c) = P(a|b, c)P(b, c)$$

$$P(b, c) = P(b|c)P(c)$$

$$P(a, b, c) = P(a|b, c)P(b|c)P(c)$$

PROBABILITY DISTRIBUTIONS

INDEPENDENCE AND CONDITIONAL INDEPENDENCE

Independence of Random Variables

Two random variables x and y are **independent** if:

$$\forall x \in X, y \in Y, \quad p(x = x, y = y) = p(x = x)p(y = y)$$

Conditional Independence

x and y are **conditionally independent** given z if:

$$\forall x \in X, y \in Y, z \in Z, \quad p(x = x, y = y | z = z) = p(x = x | z = z)p(y = y | z = z)$$

Notation

Independence is denoted by $x \perp y$, while conditional independence given z is denoted by $x \perp y | z$.

EXPECTATION, VARIANCE AND COVARIANCE

EXPECTATION AND EXPECTED VALUE

Definition

The expectation or expected value of a function $f(x)$ with respect to a distribution $P(x)$ is the mean value f takes when x is drawn from P .

Computation

For discrete variables:

$$\mathbb{E}_{x \sim P}[f(x)] = \sum_x P(x)f(x),$$

For continuous variables:

$$\mathbb{E}_{x \sim P}[f(x)] = \int p(x)f(x)dx.$$

Properties of Expectation

- Expectations are linear:

$$\mathbb{E}_x[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_x[f(x)] + \beta \mathbb{E}_x[g(x)],$$

where α and β are constants.

- Notation can be simplified to $\mathbb{E}[f(x)]$ when the context is clear.

EXPECTATION, VARIANCE AND COVARIANCE

VARIANCE AND COVARIANCE

Variance

The measure of the spread of a function $f(x)$ of a random variable:

$$\text{Var}(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

Standard deviation is the square root of the variance.

Covariance

Indicates how two variables $f(x)$ and $g(y)$ linearly relate to each other:

$$\text{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$$

Positive (negative) covariance implies a positive (negative) linear relationship.

EXPECTATION, VARIANCE AND COVARIANCE

VARIANCE AND COVARIANCE

Variance

The measure of the spread of a function $f(x)$ of a random variable:

$$\text{Var}(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

Standard deviation is the square root of the variance.

Covariance

Indicates how two variables $f(x)$ and $g(y)$ linearly relate to each other:

$$\text{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$$

Positive (negative) covariance implies a positive (negative) linear relationship.

Correlation

Correlation normalizes covariance to measure the strength of the linear relationship (scale-invariant).

Independence v.s. Covariance

Independence implies zero covariance but not vice versa. Zero covariance does not imply independence.

EXPECTATION, VARIANCE AND COVARIANCE

VARIANCE AND COVARIANCE

Covariance Matrix

For a random vector $\mathbf{x} \in \mathbb{R}^n$, the covariance matrix is $n \times n$ with:

$$\text{Cov}(x_i, x_j) = \text{Var}(x_i, x_j)$$

Diagonal elements represent the variance.

A BIT OF INFORMATION THEORY

ENTROPY

Definition

Entropy, denoted as $H(X)$, is a measure of the uncertainty or unpredictability in a random variable X .

Shannon Entropy

For a discrete random variable X with possible values $\{x_1, x_2, \dots, x_n\}$ and probability mass function $P(X)$, entropy is defined as:

$$H(X) = -\mathbb{E}_{x \sim P}[\log P(x)] = -\sum_{i=1}^n P(x_i) \log P(x_i)$$

where the logarithm is base 2 for bits.

A BIT OF INFORMATION THEORY

ENTROPY

Definition

Entropy, denoted as $H(X)$, is a measure of the uncertainty or unpredictability in a random variable X .

Shannon Entropy

For a discrete random variable X with possible values $\{x_1, x_2, \dots, x_n\}$ and probability mass function $P(X)$, entropy is defined as:

$$H(X) = -\mathbb{E}_{x \sim P}[\log P(x)] = -\sum_{i=1}^n P(x_i) \log P(x_i)$$

where the logarithm is base 2 for bits.

Interpretation

Entropy quantifies the expected amount of information conveyed by identifying the outcome of X . Higher entropy implies a more uncertain outcome, while lower entropy implies a more predictable outcome.

Properties of Entropy

- ▶ $H(X) \geq 0$ for all X .
- ▶ $H(X) = 0$ if and only if one outcome has a probability of 1 (no uncertainty).

A BIT OF INFORMATION THEORY

KL DIVERGENCE AND CROSS-ENTROPY

Kullback-Leibler (KL) Divergence

Measures how one probability distribution P diverges from a second, reference probability distribution Q :

$$D_{\text{KL}}(P\|Q) = \mathbb{E}_{x \sim P} \left[\log \frac{P(x)}{Q(x)} \right]$$

It represents the extra amount of information needed to code samples from P using a code optimized for Q .

Properties of KL Divergence

- ▶ Non-negative: $D_{\text{KL}}(P\|Q) \geq 0$
- ▶ Zero if and only if P and Q are the same distribution.
- ▶ Non-symmetric: $D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$

A BIT OF INFORMATION THEORY

KL DIVERGENCE AND CROSS-ENTROPY

Cross-Entropy

Related to KL divergence, cross-entropy combines the entropy of P with the KL divergence between P and Q :

$$H(P, Q) = H(P) + D_{\text{KL}}(P \| Q)$$

Interpretation

Minimizing cross-entropy with respect to Q equates to minimizing the KL divergence, often used in optimization problems such as training machine learning models.

A BIT OF INFORMATION THEORY

KL DIVERGENCE AND CROSS-ENTROPY

Example 5.1 (Cross-Entropy and KL Divergence in Image Classification)

In image classification, a model predicts a probability distribution Q over classes for a given image, while the true distribution P is typically one-hot encoded (1 for the correct class, 0 for others).

Cross-entropy in this context measures the difference between the distributions P and Q :

$$H(P, Q) = - \sum_c P(c) \log Q(c)$$

where c indexes over the classes.

Connection to KL Divergence

Cross-entropy decomposes into the sum of the true distribution's entropy and the KL divergence:

$$H(P, Q) = H(P) + D_{\text{KL}}(P \| Q)$$

Since $H(P)$ is constant (the true label is fixed), minimizing cross-entropy $H(P, Q)$ with respect to Q is equivalent to minimizing $D_{\text{KL}}(P \| Q)$.

During training, optimizing cross-entropy encourages the model to make predictions Q that match the true distribution P , effectively reducing the divergence from the true label distribution.

REFERENCES

-  Goodfellow, I., Bengio, Y., & Courville, A. (2016). ***Deep learning***. MIT press.
-  Murphy, K. P. (2022). ***Probabilistic machine learning: An introduction***. MIT Press.
-  Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). ***Dive into deep learning***. Cambridge University Press.