T2

Probability Theory

Salvador Ruiz Correa

August 12, 2025

Agenda:
1 Probability theory and random phenomena.
2 Probability measure key features.
3 Probability theory and random phenomena.
4 Kolmogorov axioms overview.
5 Probability space: sample space,σ-algebras.

Machine learning is a subfield of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computer systems to learn from and make predictions or decisions based on data. In essence, machine learning is the science of enabling computers to automatically learn and improve their performance on a specific task over time without being explicitly programmed for that task. Here, we briefly introduce essential machine learning theory and its relation to probability theory.

Probability Theory and Random Phenomena

Probability theory is a branch of mathematics that studies the behavior of *random phenomena* through a formal system of axioms. It is built upon and closely related to *measure theory*, a broader mathematical framework that generalizes the notions of length, area, and volume in $\mathbb{R}^n$ by assigning a measure to subsets of a given space.

Measure theory serves as the foundational framework for *integrating functions* over general spaces and plays a pivotal role in probability theory by enabling the rigorous definition of probability measures. Among the classical measures developed within this framework are the Jordan, Lebesgue, and Borel measures, which extend intuitive notions of length, area, and volume. More specialized constructs—such as complex measures, Haar measures (on locally compact groups), and probability measures—are designed to capture diverse properties of sets and functions across functional analysis, topology, and stochastic modeling.

Figure 1: Henri Léon Lebesgue (French, 1875-1941) was a French mathematician known for his theory of integration, which was a generalization of the 17th-century concept of integration—summing the area between an axis and the curve of a function defined for that axis. His theory was published originally in his dissertation Intégrale, longueur, aire ("Integral, length, area") at the University of Nancy in 1902.

Random Phenomena

Probability theory is useful for modeling *random phenomena* because it provides a mathematically rigorous way to describe uncertainty, quantify risk, and predict patterns in systems where outcomes are not deterministic.

Definition 1: Random Phenomenon

A phenomenon or procedure for generating data is *random* if
- the outcome is not predictable in advance;
- there is a predictable long-term pattern that can be described by the distribution of outcomes over very many observations.

Definition 2: Random Outcome

A *random outcome* is the result of a random phenomenon or procedure.

The outcome of an individual random experiment cannot be predicted with certainty, but the set of all *possible* outcomes is known in advance.

Box 1: Rolling a Dice

Imagine I roll a fair die privately, and I tell you if the resulting throw is odd or even.
- The possible outcomes are integers from one to six.
- The information available to you is whether the roll is odd or even.
- Probabilities are computed on the basis that each outcome is equally likely, so we have $\frac{1}{2}$ chance of obtaining odd/even.

```r
# Sample a random number in {1,2,3,4,5,6}
number = sample(1:6,1, replace=F)
if (number %% 2 > 0) {
    print("Odd number")
} else {
   print("Even number")
}
```

Figure 2: R code implementation of the random experiment described in Example 1. Outcomes have equal probability (i.e., the dice is fair).

Random Phenomena Examples

Outcomes in Health Sciences that are unpredictable can be modeled, in principle, as random phenomena. A principled modeling of these events can have a significant impact on patient care, research, and decision-making. The following are examples of random phenomena.

1. *Patient outcomes*: The response of individual patients to medical treatments or interventions can vary due to random factors. For instance, some patients may recover quickly from surgery while others with similar conditions may experience complications, and these differences can be influenced by factors that are not fully understood.
2. *Disease spread*: The spread of infectious diseases can exhibit random patterns. Factors such as the movement of infected individuals, contact patterns, and the effectiveness of preventive measures can all contribute to the randomness of disease transmission.
3. *Clinical trials*: Randomized controlled trials (RCTs) are widely used

in healthcare research to evaluate the efficacy of new treatments or interventions. The random assignment of participants to treatment and control groups helps reduce bias and accounts for random variations in outcomes.

4. *Emergency Room traffic*: The number of patients arriving at an emergency room can vary significantly from day to day or even hour to hour. Random factors like accidents, weather conditions, and disease outbreaks can influence the patient flow.
5. *Medication response*: Some patients may respond differently to the same medication due to genetic variations, environmental factors, or random fluctuations in their physiology. This can complicate medication management and dosing.
6. *Diagnostic Testing*: The results of diagnostic tests, such as blood tests or imaging scans, may show variability due to random measurement error, equipment calibration, or sample handling procedures.
7. *Disease Outbreaks*: The occurrence and severity of disease outbreaks, such as flu epidemics or COVID-19 surges, can exhibit random patterns influenced by factors like population density, vaccination rates, and individual behavior.
8. *Healthcare resource allocation*: The allocation of healthcare resources, such as hospital beds and ventilators during a public health crisis, can be influenced by random factors like the sudden surge in cases or unexpected logistical challenges.
9. *Healthcare costs*: The cost of healthcare services can vary randomly due to factors such as fluctuations in the prices of medical supplies, changes in insurance coverage, and unexpected healthcare demands.
10. *Healthcare workforce availability*: The availability of healthcare professionals, including doctors, nurses, and support staff, can be influenced by random factors like staff illnesses or scheduling conflicts.

Random Experiments

Probability theory is also useful for modeling *random experiments*.

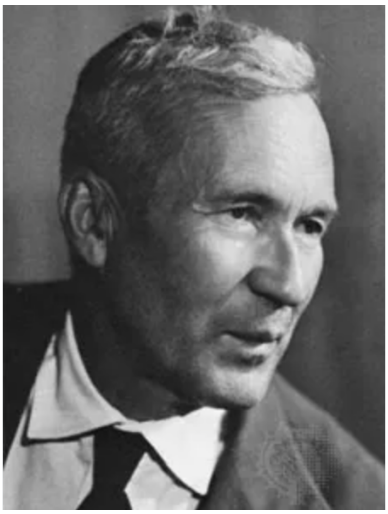


Figure 3: Andrey Nikolaevich Kolmogorov (Russian, 1903-1987) was a Soviet mathematician who contributed to the mathematics of probability theory, topology, intuitionistic logic, turbulence, classical mechanics, algorithmic information theory, and computational complexity. His contributions to probability theory provided a principled approach to the study of random phenomena.

Definition 3: Random Experiment

A *random experiment* is understood as any procedure such that when it is repeated under the same initial conditions, the result obtained is not always the same. It is characterized by three main features.

- The possible outcomes of the experiment.
- The events we can observe, i.e. the information that is revealed at the end of the experiment.
- The probabilities assigned to each event.

In practical applications, a random experiment can, in principle, be repeated numerous times under the same conditions. The outcomes of individual experiments must be independent, and must in no way be affected by any previous outcome.

Random Phenomena, Random Experiments, Measure Theory and Probability Theories

MEASURE THEORY is concerned with the problem of how to assign a size to certain sets, enabling a principle definition of a probability measure, which is the principal tool of probability theory. In daily life, assigning a size to sets is easy to do:

- count: $\{a, b, c, \ldots, x, y, x\}$ has 26 letters;
- take measurements: length (in one dimension), area (in two dimensions), volume (in three dimensions), or time;
- calculate the odds of winning the lottery.

In each case, we compare and express the result with respect to some base unit (Figure 4-1).

Probability Measure Key Features

Notice that triangles are more flexible than rectangles since we can represent every rectangle, and actually any odd-shaped quadrangle, as the 'sum' of two non-overlapping triangles Fig. 4-2. In doing so we have assumed a few things:

- In Fig. 4-3 we have chosen a *particular* baseline and the corresponding height arbitrarily. But the concept of *area* should not depend on such choice and the calculation this choice entails.
- The independence of the area from the way we calculate it is called well-definedness. Plainly, we have the choices shown in Fig 4-3. Notice that Fig 4-3 allows us to pick the most convenient method to work out the area.

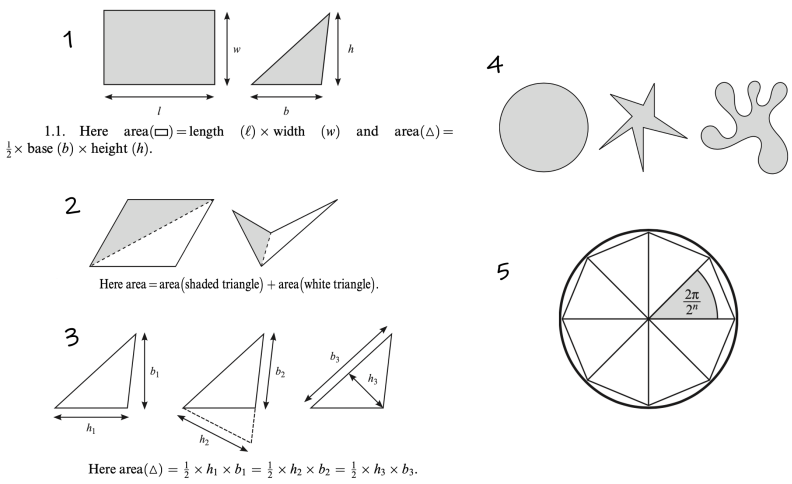


Figure 4: Measure theory is concerned with the problem of how to assign a size to certain sets. In daily life this is easy to do: we can count, take measurements, or calculate rates. In each case, we compare and express the result with respect to some base unit.

The area of the circle in 5 can be computed as follows.

$$\text{area}(\bigcirc) = lim_{n\longrightarrow\infty} 2^n \times \text{area}(\triangle \text{ at step } n).$$

Source: R. L. SchillingMeasures, Integrals, and Martingales, 2nd edition, Cambridge University Press, 2017.

- In Fig. 4-2 we actually use two facts:
  - the area of non-overlapping (disjoint) sets can be added, i.e.

$$\text{area}(A) = \alpha, \text{area}(B) = \beta, A \cap B = \emptyset \rightarrow, \text{area}(A \cup B) = \alpha + \beta.$$

This shows that the least we should expect from a reasonable measure μ is that it is

- well-defined, takes values in $[0, \infty]$, and $\mu(\emptyset) = 0$;
- additive, i.e., $\mu(A \cap B) = \mu(A) + \mu(B)$ whenever $A \cap B = \emptyset$;
- is invariant under congruences and translations, which is a characteristic property of length, area, and volume (the Lebesgue measure in $\mathbb{R}^n$).

Measures defined on the set of outcomes/events of a random phenomena/experiments enable a principled definition definition of probability.

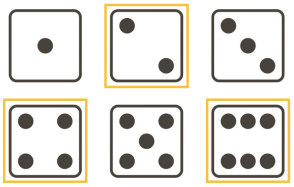


Figure 5: Measures over finite and countable sets have the properties described above. For instance, given a fair dice,

$$\mu(\{\boxdot, \boxdot\!\!\boxdot, \boxdot\!\!\boxdot\!\!\boxdot\}) = \mu(\{\boxdot\}) + \mu(\{\boxdot\!\!\boxdot\}) + \mu(\{\boxdot\!\!\boxdot\!\!\boxdot\}) = \frac{1}{2}.$$

Kolmogorov's Axioms

Kolmogorov's axioms are the foundations of probability theory. These axioms were introduced by Andrey Kolmogorov in 1933. These axioms remain central and have direct contributions to real-world probability cases.

These axioms are expressed using the fundamental concept of *probability space* $(\Omega, \mathscr{F}, P)$, where Ω is the sample space, $\mathscr{F}$ is a σ-algebra, and P is a probability measure. The elements of Ω are often called outcomes or *elementary events*, and the elements of $\mathscr{F}$, which are subsets of Ω, are called *events*.

Informally speaking a probability space is like a special playground where we play with chance and randomness. It's made up of three important parts:

- *Sample space:* This is like a list of all the possible things that can happen when we're dealing with something random. For example, if we're rolling a die, the sample space includes all the possible outcomes: 1, 2, 3, 4, 5, and 6.
- *Events*: These are like the games we play in our probability playground. Events are just groups of outcomes from the sample space. For example, the event of "getting an even number" when rolling a die includes outcomes 2, 4, and 6 (see Box 1).
- *Probability measure*: This is like a rule that tells us how likely each event is. It's like saying, "In our playground, the chance of this game happening is this much." For example, the probability of getting an even number when rolling a fair six-sided die is 1/2 because there are three even outcomes out of six possible outcomes.

Probability Space

THE PROBABILITY SPACE CONCEPT sets a principled frame to model random phenomena and random experiments. Here we provide an interpretation of each of the elements of a probability space: sample space, σ-algebra, and probability measure.

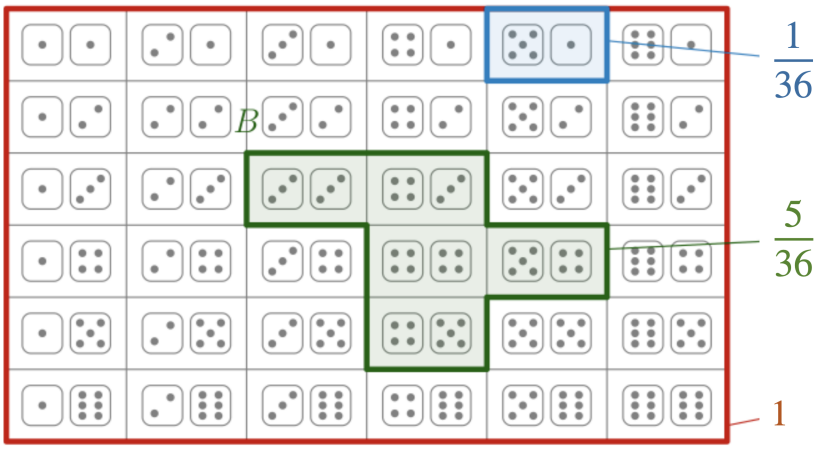


Figure 6: Probability space for throwing a die twice in succession: The sample space Ω consists of all 36 possible elementary outcomes; three different events (colored polygons) are shown with their respective probabilities (assuming that the dice are fair). Source: https://en.wikipedia.org

Definition 4: Probability Space

A probability space consists of an ordered triplet $(\Omega, \mathscr{F}, P)$, where Ω is an arbitrary *set* called *sample space*, $\mathscr{F}$ is a σ-algebra of subsets of Ω, and P is a probability measure defined on $\mathscr{F}$.

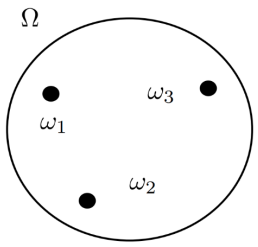


Figure 7: Sample space Ω containing three possible outcomes ω_1, ω_2, and ω_3.

Sample Space Ω

Loosely speaking, sample space in probability is like a list of all the things that could possibly happen when you're dealing with something random, like flipping a coin or rolling a die. Imagine you're flipping a coin. The sample space for this is just a list of the two things that can happen: "heads" and "tails." This is the sample space associated with flipping a coin. So, a sample space is like a simple list of all the different outcomes you might get when you do something random. It helps you see what could happen and helps you figure out the chances or probabilities of each outcome.

Definition 5: Sample Space

A *sample space* is a set Ω containing all possible outcomes of a random phenomenon or procedure. An *outcome* ω is an element in Ω, i.e., $\omega \in \Omega$ (which we may or may not observed).

Sample spaces can be classified according to their cardinality, i.e., the number of elements, $\#\Omega$, forming the set.

Definition 6: Types of Sample Spaces

- A sample space consisting of a finite or a *countably infinite* number of elements is called a *discrete sample space*.
- When the sample space includes all the numbers in an interval of the real line, it is called a *continuous sample space*.

The following are examples of sample spaces.

1. *Coin toss*: When you flip a fair coin, the sample space consists of two possible outcomes: heads (H) or tails (T). $\Omega = \{\mathsf{H}, \mathsf{T}\}$.
2. *Rolling a six-sided die*: When you roll a standard six-sided die, the sample space includes the numbers 1 through 6. $\Omega = \{1, 2, 3, 4, 5, 6\}$.
3. *Flipping two coins*: If you flip two coins simultaneously, the sample space includes all possible combinations of outcomes for each coin. $\Omega = \{\{\mathsf{H},\mathsf{H}\}, \{\mathsf{H},\mathsf{T}\}, \{\mathsf{T},\mathsf{H}\}, \{\mathsf{T},\mathsf{T}\}\}$.
4. *Rolling two dice*: Rolling two six-sided dice results in a sample space that includes all possible combinations of the dice values.

$\Omega = \{$ {1,1}, {1,2}, ... {6,6}$\}$. (Figure 4).

5. *Rolling a six-sided die and flipping a coin*: If you roll a die and flip a coin, the combined sample space includes all possible pairs of outcomes.

$$\Omega = \{ \{1, H\}, \{1, T\}, \{2, H\}, \{2, T\}, ..., \{6, H\}, \{6, T\} .\}$$

6. *Diagnostic test*: In a diagnostic test for a disease, the sample space could consist of four outcomes $\omega_1 = \mathsf{TP}$, $\omega_2 = \mathsf{FP}$, $\omega_3 = \mathsf{TN}$ and $\omega_4 = \mathsf{FN}$, so $\Omega = \{\mathsf{TP}, \mathsf{FP}, \mathsf{TN}, \mathsf{FN}\}$. Key: True Positive, TP; False Positive, FP; True Negative, TN; and False Negative FN.
7. *Medication dosage*: Imagine a situation where a doctor is deciding on the dosage of a particular medication for a patient. The sample space would consist of all possible dosages that the doctor can prescribe, ranging from the lowest possible dose to the highest. $\Omega = \{\mathsf{Low}, \mathsf{Medium}, \mathsf{High}\}$.
8. *Hospital stay length*: When a patient is admitted to a hospital, the length of their stay can vary. The sample space for the length of stay could include various possibilities, such as short stays, average stays, and long stays. $\Omega = \{\mathsf{Short}, \mathsf{Average}, \mathsf{Long}\}$.
9. *Patient discharge status*: After receiving treatment, a patient may be discharged under different conditions. The sample space for discharge status might include being discharged in good health, with ongoing treatment needs, or with a referral to a specialist. $\Omega = \{\mathsf{Discharged}, \mathsf{Ongoing\ treatment}, \mathsf{Referred}\}$.
10. *Surgical outcomes*: In the case of a surgical procedure, the sample space could include different possible outcomes, such as successful surgery, or unsuccessful surgery. $\Omega = \{\mathsf{Successful}, \mathsf{Unsuccessful}\}$.
11. *Disease progression*: For patients with chronic diseases, the progression of the disease can vary. The sample space for disease progression could include different stages of the disease, such as mild, moderate, or severe. $\Omega = \{\mathsf{Mild}, \mathsf{Moderate}, \mathsf{Severe}\}$.
12. *Emergency room triage*: In an emergency room, patients are triaged based on the severity of their condition. The sample space for triage levels might include various levels of urgency, such as critical, urgent, or non-urgent. $\Omega = \{\mathsf{Critical}, \mathsf{Urgent}, \mathsf{Non\text{-}urgent}\}$.
13. *Appointment scheduling*: When scheduling patient appointments, there can be various time slots available. The sample space for appointment scheduling would consist of all the possible time slots.

$$\Omega = \{\mathsf{8{:}00\ AM}, \mathsf{10{:}00\ AM}, \mathsf{12{:}00\ PM}, \mathsf{2{:}00\ PM}, \mathsf{4{:}00\ PM}\}.$$

14. *Psychological testing:* In psychological assessments, the sample space may consist of all possible test scores or responses. For

example, when conducting an IQ test, the sample space includes all potential scores that individuals might achieve.

15. *Survey responses:* When conducting surveys or questionnaires in psychology, the sample space represents all possible responses to each question. For instance, in a survey about people's feelings, the sample space could include options like "happy," "neutral," "sad," and so on.
16. *Behavioral observations:* When observing and recording behaviors in psychological studies, the sample space could encompass various behavioral categories. For example, in a study on child behavior, the sample space might include categories like "playing," "crying," "listening," and "talking."
17. *Emotional responses:* In studies of emotional responses, the sample space can describe the full range of possible emotions that individuals might experience, such as "joy," "anger," "fear," "disgust," "surprise," and "sadness."
18. *Blood pressure measurement*: When measuring blood pressure, both systolic and diastolic pressures are continuous variables, and their sample spaces consist of real numbers within specified ranges. For example, systolic pressure might fall within the range of 90 mmHg to 180 mmHg, while diastolic pressure might range from 60 mmHg to 120 mmHg. Therefore

$$\Omega = [90\,\text{mmHg}, 180\,\text{mmHg}],$$

and

$$\Omega = [60\,\text{mmHg}, 120\,\text{mmHg}],$$

for systolic and diastolic pressures, respectively.

19. *Weather forecast*: A weather forecast might have different categories like sunny, cloudy, rainy, or snowy. $\Omega = \{\,\text{Sunny, Cloudy, Rainy, Snowy}\}$.
20. *Temperature measurement*: When measuring temperature using a thermometer, the sample space includes all possible real numbers within a specified temperature range. For instance, the temperature could be any real number within a range of -100^oC to 100^oC. $\Omega = [-100^oC, 100^oC]$.
21. *Inventory management:* The sample space could represent the different levels of inventory for a product. For example,

$$\Omega = \{\,\text{High Inventory, Moderate Inventory, Low Inventory}\}.$$

22. *Market Research:* When conducting market research, the sample space may represent customer preferences. For instance,

$$\Omega = \{\,\text{Product A Preferred, Product B Preferred, No Preference}\}.$$

23. *Financial Investments:* In the context of investment decisions, the sample space might represent investment outcomes like

$$\Omega = \{ \text{Profit, Break-even, Loss} \}.$$

24. *Project Management:* For project management, the sample space could depict project outcomes such as

$$\Omega = \{\text{On Time and On Budget, Delayed but On Budget, Delayed and Over Budget}\}.$$

25. *Product Launch Success:* When launching a new product, the sample space may represent the possible outcomes for success, like

$$\Omega = \{ \text{High Sales, Moderate Sales, Low Sales} \}.$$

26. *Employee Performance Evaluation:* In performance evaluations, the sample space might represent performance levels, such as

$$\Omega = \{ \text{High Inventory, Moderate Inventory, Low Inventory} \}.$$

27. *Customer Satisfaction:* Customer satisfaction surveys often use sample spaces like

$$\Omega = \{\text{Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied}\}.$$

28. *Risk Assessment:* In risk assessment, the sample space may include potential risks or events, like

$$\Omega = \{\text{Market Downturn, Regulatory Changes, Supplier Issues}\}.$$

29. *Marketing Campaign Effectiveness:* For assessing the effectiveness of a marketing campaign, the sample space could represent the different customer responses, such as

$$\Omega = \{\text{Conversion, Click-Through, No Response}\}.$$

30. *Production Quality Control:* In quality control, the sample space might represent the quality of products, like

$$\Omega = \{\text{Defective, Passes Quality Control, Exceptional Quality}\}.$$

31. *Sample Space of Continuous Functions $C([a,b]$*: The sample space of continuous functions often denoted as $C([a,b])$, represents the set of all functions f(x) defined on the closed interval [a, b] that are continuous over that interval. In mathematical notation:

$$\Omega = \{f(x) | f(x) \text{is continuous on } [a,b]\}.$$

32. *The sample space of texts:* In the context of natural language and textual data, represents the set of all possible text strings that can be generated within a specific language or character encoding. The sample space of texts is essentially the universe of all potential textual documents, messages, or sequences of characters. Here are some key points to consider regarding the sample space of texts:

- *Character set*: The sample space of texts depends on the character set used. For example, in English, it would be the set of all combinations of English letters (both uppercase and lowercase), digits, punctuation marks, and special characters.
- *Infinite nature*: The sample space of texts is typically considered infinite, as there is no upper limit on the length of text strings that can be generated. Texts can be of varying lengths, from a single character to entire books or larger documents.
- *Variability*: The sample space encompasses a wide variety of texts, ranging from common words and phrases to unique combinations and rare or gibberish sequences of characters. It includes valid and meaningful texts as well as invalid, meaningless, or nonsensical ones.
- *Natural language*: The sample space is most relevant in the context of natural language, where it represents the potential for human communication in written form. It includes text in languages other than English, each with its own set of characters and rules.
- *Encoding and formatting*: The sample space is influenced by text encoding standards and formatting rules. For instance, plain text documents, HTML, JSON, XML, or any other text-based data format contribute to the diversity within the sample space.
- *Applications*: The sample space of texts is fundamental in natural language processing (NLP), text analysis, information retrieval, and various data science applications that involve textual data.
- *Size and complexity*: Due to the infinite nature of the sample space, it is practically impossible to exhaustively list or describe all possible text strings. The complexity increases with the size of the character set and the length of the text.
- *Machine earning*: In machine learning and NLP tasks, understanding the sample space of texts is important for tasks like text generation, text classification, and text mining. Models and algorithms are trained on data from this sample space to perform specific tasks.

The nature of the sample space of natural language can be appreciated by understanding how deep learning architectures for Large Language Models (LLMs), conduct natural language processing tasks. Models such as transformers, often use *embeddings* and *positional encoding* to convert text sequences into number sequences.

- *Word embeddings*: Transformers use word embeddings to convert input tokens (e.g., words) into continuous vector representations. These embeddings capture semantic information, allowing the model to understand the meaning of words. Word embeddings help transformers handle a large vocabulary ef-

ficiently and generalize better because they learn to represent words with similar meanings in similar vector spaces.

- *Positional encoding*: Transformers don't have a built-in notion of word order or position, which is essential for understanding sequences (e.g., sentences or documents). Positional encoding provides this information. Positional encoding is added to the word embeddings to convey the position of words in a sequence. It allows the model to distinguish between the same word in different positions and learn the sequential relationships between words. In summary, embeddings and positional encoding in transformers play a critical role in enabling these models to handle sequences of data effectively, whether it's natural language text or other sequential data. Embeddings capture the meaning of words or tokens, while positional encoding imparts information about the order of tokens in the sequence. This combination enables transformers to excel in tasks like machine translation, text generation, and more.

Box 2: Sample Space Exercise

- Mr. Holmes now lives in Los Angeles. One morning when Homes leaves his house, he realizes that his grass is wet (H). Is it due to rain (R), or has he forgotten to turn off the sprinkler (S)? Next, he notices that the grass of his neighbor, Dr. Watson, is also wet (W).
  1. Write down the sample space for this example.
  2. Write down the outcomes related to the events:
     - A = "Holmes grass is wet."
     - B = "Holmes forgot to turn the sprinkler off."
     - C = "Holmes forgot to turn the sprinkler off and Watson's grass is wet."
- Solution.
  1. Sample space and outcome probabilities:

$\Omega's$ oucomes	$P(\omega)$
$\omega_1 = \mathsf{H^cW^cS^cR^c}$	α_1
$\omega_2 = \mathsf{H^cW^cS^cR}$	α_2
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_5 = \mathsf{H^cWS^cR^c}$	α_5
$\omega_6 = \mathsf{H^cWS^cR}$	α_6
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_9 = \mathsf{HW^cS^cR^c}$	α_9
$\omega_{10} = \mathsf{HW^cS^cR}$	α_{10}
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{13} = \mathsf{HWS^cR^c}$	α_{13}
$\omega_{14} = \mathsf{HWS^cR}$	α_{14}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

  2. Events:

A's outcomes	$P(\omega)$
$\omega_9 = \mathsf{HW^cS^cR^c}$	α_9
$\omega_{10} = \mathsf{HW^cS^cR}$	α_{10}
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{13} = \mathsf{HWS^cR^c}$	α_{13}
$\omega_{14} = \mathsf{HWS^cR}$	α_{14}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

C's outcomes	$P(\omega)$
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

σ-algebras

To gain some intuition of what a σ-algebra is, imagine you have a collection of things, like a number of different events or sets of outcomes in probability. A sigma algebra is like a special container or

group that holds these things together in an organized way (Figure 19).

Definition 7: σ-algebra

A σ-algebra $\mathscr{F}$ on a set Ω is a family of subsets of Ω such that:

- $\Omega \in \mathscr{F}$
- $A \in \mathscr{F} \implies A^c \in \mathscr{F}$
- $(A_n)_{n\in\mathbb{N}} \subset \mathscr{F} \implies \bigcup A_n \in \mathscr{F}$

The set $A \in \mathscr{F}$ is said to be *measurable* or $\mathscr{F}$*-measurable*.

To be a σ-algebra, this container has to follow a few rules:

1. It must always contain the "whole thing." In other words, it includes everything you're interested in. For example, if you're thinking about all possible outcomes when rolling a die, the σ-algebra would definitely include "rolling a 1," "rolling a 2," and so on, all the way up to "rolling a 6."
2. It should also include the "opposites" of things. Let's say you toss a coin and have "getting a head" as one of your events. The sigma σ-algebra should also have "not getting a head" (which means getting a tail) inside it.
3. When you look inside the σ-algebra, you should find all the combinations and possibilities of the things you're interested in. So, if you have "rolling an even number" and "rolling a prime number," you want the sigma-algebra to include things like "rolling an even number AND rolling a prime number."

In summary, a sigma-algebra is like a special container that holds all the possible outcomes and their opposites in an organized way, helping you study and understand different events and probabilities in a systematic manner.

The following are more examples of σ-algebras.

1. Construct a σ-algebra for the set $\Omega = \{\boxdot, \boxdot\}$.

$$\mathscr{F} = \{\Omega, \emptyset, \{\boxdot\}, \{\boxdot\}\}$$

2. Consider the set $\Omega = \{\boxdot, \boxdot, \boxdot\}$.
   - Construct the *minimal* σ-algebra for the set Ω.

$$\mathscr{F} = \{\Omega, \emptyset\}.$$

   - Construct the *maximal* σ-algebra for the set Ω (i.e., Ω's power set, $\mathscr{P}(\Omega)$).

$$\mathscr{F} = \{\emptyset, \{\boxdot\}, \{\boxdot\}, \{\boxdot\}, \{\boxdot, \boxdot\}, \{\boxdot, \boxdot\}, \{\boxdot, \boxdot\}, \Omega\}.$$

   - Construct a σ-algebra for the set Ω that is neither maximal or minimal.

$$\mathscr{F} = \{\Omega, \emptyset, B, B^c\}, B \subset \Omega$$

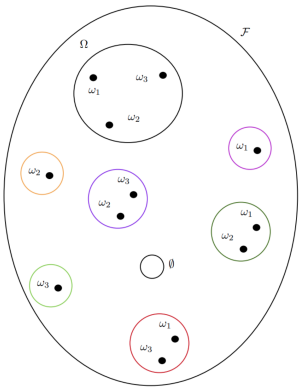


Figure 8: The event space is a σ-algebra on the sample space Ω. It is a subset of the power set, $\mathcal{P}(\Omega)$, whose elements, called events, satisfy some regularity conditions. When is finite and discrete, the power set is a valid σ-algebra that can be used as the event space. The σ-algebra satisfies $\Omega \in \mathcal{F}$; the sample space is an event called the *sure* event. $A \in F \implies A^c \in \mathcal{F}$, i.e., $\mathcal{F}$ is *closed* under complementation. $A_1, A_2, \ldots, \bigcup A_n \in \mathscr{F}$, so F is closed under *countable unions*. Source: http://bit.ly/3s11bq5.

- *Remark:* no set is an element of itself. For example,

$$\boxdot \neq \{\boxdot\}.$$

3. *Emotional responses:* In studies of emotional responses, the sample space can describe the full range of possible emotions that individuals might experience, such as "joy," "anger," "fear," "disgust," "surprise," and "sadness." Write down the sample space Ω associated with the given outcomes. Let

$$\mathscr{F} = \{\Omega, \emptyset, B, B^c\},$$

   where $B = \{\text{joy}\}$. Is $\mathscr{F}$ is a σ-algebra?
4. *Surgical outcomes*: In the case of a surgical procedure, the sample space could include different possible outcomes, such as successful surgery, complications, or the need for additional surgeries.

$$\Omega = \{\text{Successful surgery, Complications, Additional surgery}\}$$

   Let $\mathscr{F} = \{\Omega, \emptyset, A, A^c\}$ where $A = \{\text{Successful surgery}\}$. Show that $\mathscr{F}$ is a σ-algebra?
5. *Product sales:* When launching a new product or analyzing sales data, the sample space represents all possible sales outcomes, including the range of sales volumes and revenue generated.

$$\Omega = \{\text{Low sales, Moderate sales, High sales}\}$$

   Construct a σ-algebra for Ω.
6. *Emergency room triage*: In an emergency room, patients are triaged based on the severity of their condition. The sample space for triage levels might include various levels of urgency, such as critical, urgent, or non-urgent.

$$\Omega = \{\text{Critical, Urgent, Non-urgent}\}.$$

   What is the maximal σ-algebra for Ω?

*σ-algebra Generators**

Given a set of events or outcomes, a σ-algebra generator refers to the smallest σ-algebra that contains these events. In other words, it's the collection of all possible events that can be formed by combining the original set of events through set operations like unions, intersections, and complements.

Definition 8: σ-algebra Generators*

σ-algebra generators are defined as follows:

- For every system of sets $\mathscr{G} \in \mathscr{P}(\Omega)$ there exist a smallest σ-algebra containing $\mathscr{G}$. This is the σ-algebra generated by $\mathscr{G}$, denoted by $\sigma(\mathscr{G})$, and $\mathscr{G}$ is called the generator.
- Notice that the intersection $\bigcap_{i\in I} \mathscr{A}_i$ of arbitrarily many σ-algebras in Ω is again a σ-algebra in Ω.
- Except in a few simple examples, it is hard to write down explicitly a generated σ-algebra.

Box 3 shows an example of a σ-algebra generator and the corresponding σ-algebra.

Box 3: σ-algebra Generators Example*

- Let $A, B \in \Omega$, $A \bigcap B = \emptyset$. Define $\mathscr{G} = \{A, B\}$.

$$\sigma(\mathscr{G}) = \{\emptyset, A, B, A \cup B, A^c, B^c, (A \cup B)^c, \Omega\}.$$

  Prove this statement.

Borel Algebra

Informally speaking, you can think of a Borel algebra as a special collection of sets that helps you organize everyday numbers in a neat and organized way.

Imagine you have all the real numbers, like 1, 2, 3, and so on, and also the numbers with decimals like 1.5, 2.75, and so forth. These numbers can be really messy, and it's challenging to sort them neatly. But with a Borel algebra, you create a system where you can group these numbers together in a smart way.

Here's how it works: you start with simple sets, like all the numbers that are less than 3, or all the numbers between 1.5 and 2.5. Then, you can combine these sets in various ways to create more sets. For example, you can combine the set of numbers less than 3 with the set of numbers between 1.5 and 2.5 to create a new set: all the numbers between 1.5 and 3.

So, a Borel algebra is like a toolbox that helps you neatly organize numbers into sets that make sense, making it easier to study and work with these numbers in mathematics and statistics. It's a bit like putting numbers into labeled boxes to keep them tidy and manageable.

Consider the collection of all open intervals (a, b) of $\mathbb{R}$, where $a < b$. The minimum σ-algebra generated by this collection is called

the σ-algebra of $\mathbb{R}$ and is denoted by $\mathcal{B}(\mathbb{R})$, A σ-algebra $\mathscr{F}$ on a set Ω is a family of subsets of Ω such that:

$$\mathcal{B}(\mathbb{R}) = \sigma\left(\{(a,b) \subseteq \mathbb{R} : a \leq b\}\right).$$

For any real numbers $a \leq b$, the following, are all elements of $\mathcal{B}(\mathbb{R})$.

- $(-\infty, a)$, (b, ∞), $(-\infty, a) \cup (b, \infty)$.
- $[a,b] = ((-\infty, a) \cup (b, \infty))^c$.
- $(-\infty, a] = \cup_{n=1}^{\infty}[a-n, a]$ and $[b, \infty) = \cup_{n=1}^{\infty}[b, b+n]$.
- $(a,b] = (a, \infty) \cap (-\infty, b]$
- $\{a\} = \bigcap_{n=1}^{\infty}(a - \frac{1}{n}, a + \frac{1}{n})$ and $\{a_1, a_2, \ldots, a_n\} = \cup_{k=1}^{n}\{a_k\}$.

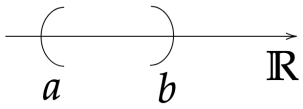


Figure 9: Open interval (a,b), $a \neq b$.

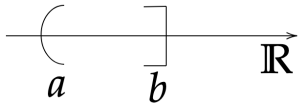


Figure 10: Semi-open interval $(a,b] = (a, \infty) \cap (-\infty, b]$, $a \neq b$.

*Borel Algebra in $\mathbb{R}^n$**

The Borel Algebra in $\mathbb{R}^n$ is defined as:

$$\mathcal{B}(\mathbb{R}^n) = \sigma(\mathcal{B}(\mathbb{R}) \times \cdots \times \mathcal{B}(\mathbb{R})).$$

The Borel σ-algebra is generated by many different systems of sets. The most interesting are:

- Open rectangles:

$$\mathscr{I}^{n,o} = \mathscr{I}^n(\mathbb{R}^n) = \{(a_1, b_1) \times \cdots \times (a_n, b_n) : a_i, b_i \in \mathbb{R}\}$$

.

- From the right half-open rectangles:

$$\mathscr{I} = \mathscr{I}(\mathbb{R}^n) = \{[a_1, b_1) \times \cdots \times [a_n, b_n) : a_i, b_i \in \mathbb{R}\}$$

.

- Specific examples
  - $\mathscr{I}^2(\mathbb{R}^2) = \sigma(\{(a,b) \times (c,d) \subseteq \mathbb{R}^2 : a \leq b, c \leq d\})$.
  - $\mathscr{I}(\mathbb{R}^2) = \sigma(\{[a,b) \times [c,d) \subseteq \mathbb{R}^2 : a \leq b, c \leq d\})$.
  - $\mathscr{I}^3(\mathbb{R}^3) = \sigma(\{(a,b) \times (c,d) \times (e,f) \subseteq \mathbb{R}^3 : a \leq b, c \leq d, e \leq f\})$.

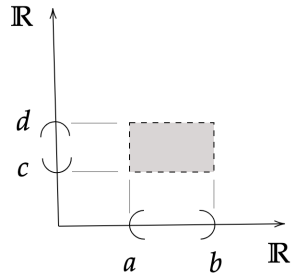


Figure 11: Rectangle $R \in \mathscr{I}(\mathbb{R}^2)$ for the real numbers a, b, c and d for which $a < b$ and $c < d$.

Measurable Space

Let $\mathscr{F}$ be a σ-algebra defined on the sample space Ω. The pair $(\Omega, \mathscr{F})$ is called *measurable space*.

Definition 9: Product Space

Let $(\Omega_1, \mathscr{F}_1)$ and $(\Omega_2, \mathscr{F}_2)$ be two measurable spaces. The measurable spaces product is defined as

$$(\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2),$$

where $\mathscr{F}_1 \otimes \mathscr{F}_2 = \sigma(\mathscr{F}_1 \times \mathscr{F}_2)$, and $\mathcal{G} = \mathscr{F}_1 \times \mathscr{F}_2$ is the σ-algebra generator. In general,

$$\mathscr{F}_1 \times \mathscr{F}_2 := \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_1 \in \mathcal{F}_2\},$$

is not a σ-algebra.

Box 4: Product Space Example

Compute $(\Omega, \mathscr{F}) = (\Omega_1 \times \Omega_2, \mathscr{F}_1 \otimes \mathscr{F}_2)$ given:

$$\Omega_1 = \{\mathsf{H}, \mathsf{T}\},$$

$$\mathscr{F}_1 = \{\emptyset, \{\mathsf{H}\}, \{\mathsf{T}\}, X_1\},$$

$$\Omega_2 = \{\mathsf{h}, \mathsf{t}\},$$

$$F_2 = \{\emptyset, \{\mathsf{h}\}, \{\mathsf{t}\}, X_2\}.$$

- The cartesian product $\Omega_1 \times \Omega_2$ is:

$$\Omega = \Omega_1 \times \Omega_2 = \{(\mathsf{H}, \mathsf{h}), (\mathsf{H}, \mathsf{t}), (\mathsf{T}, \mathsf{h}), (\mathsf{T}, \mathsf{t})\}.$$

- σ-algebra generator:

$$\begin{aligned}\mathscr{F}_1 \times \mathscr{F}_2 = \{&(\emptyset, \emptyset), (\emptyset, \mathsf{h}), (\emptyset, \mathsf{t}), (\emptyset, \Omega_2),\\ &(\mathsf{H}, \emptyset), (\mathsf{H}, \mathsf{h}), (\mathsf{H}, \mathsf{t}), (\mathsf{H}, \Omega_2),\\ &(\mathsf{T}, \emptyset), (\mathsf{T}, \mathsf{h}), (\mathsf{T}, \mathsf{t}), (\mathsf{H}, \Omega_2),\\ &(\Omega_1, \emptyset), (\Omega_1, \mathsf{h}), (\Omega_1, \mathsf{t}), (\Omega_1, \Omega_2)\}.\end{aligned}$$

- σ-algebra:

$$\begin{aligned}\mathscr{F} = \mathscr{F}_1 \otimes \mathscr{F}_2 = \{&\emptyset,\\ &\{(\mathsf{H}, \mathsf{h})\}, \{(\mathsf{H}, \mathsf{t})\}, \{(\mathsf{T}, \mathsf{h})\}, \{(\mathsf{T}, \mathsf{t})\}\\ &\{(\mathsf{H}, \mathsf{h}), (\mathsf{H}, \mathsf{t})\}, \{(\mathsf{H}, \mathsf{h}), (\mathsf{T}, \mathsf{h})\}, \{(\mathsf{H}, \mathsf{h}), (\mathsf{T}, \mathsf{t})\},\\ &\{(\mathsf{H}, \mathsf{t}), (\mathsf{T}, \mathsf{h})\}, \{(\mathsf{H}, \mathsf{t}), (\mathsf{T}, \mathsf{t})\},\\ &\{(\mathsf{T}, \mathsf{h}), (\mathsf{T}, \mathsf{t})\},\\ &\{(\mathsf{H}, \mathsf{h}), (\mathsf{H}, \mathsf{t}), (\mathsf{T}, \mathsf{h})\}, \{(\mathsf{H}, \mathsf{h}), (\mathsf{H}, \mathsf{t}), (\mathsf{T}, \mathsf{t})\},\\ &\{(\mathsf{H}, \mathsf{t}), (\mathsf{T}, \mathsf{h}), (\mathsf{T}, \mathsf{t})\},\\ &\{(\mathsf{H}, \mathsf{h}), (\mathsf{T}, \mathsf{h}), (\mathsf{T}, \mathsf{t})\},\\ &\Omega\}.\end{aligned}$$

Measures

Intuitively, a (measure-theoretic) measure is a way of giving value to things, but it's not as simple as measuring the length of a rope or the weight of an object. Instead, it's used for more complicated situations where you want to understand and compare things that aren't always straightforward to measure. Imagine you have a collection of different-sized buckets, and you want to know how much water each one can hold. Each bucket might have an irregular shape, and you can't just fill them to see how much they hold. A measure-theoretic measure helps you assign a value to each bucket's capacity, even when they're not simple shapes. This concept is used in advanced mathematics and statistics to tackle complex problems and understand things in a more abstract or generalized way.

Measure theory is indeed concerned with the problem of how to assign a size to certain sets. In daily life this is easy to do: we can count, take measurements, or calculate rates. In each case, we compare and express the result with respect to some base unit.

Definition 10: Measure Definition

A positive measure μ on Ω is a map $\mu : \mathscr{F} \to [0, \infty]$ satisfying:

- $\mathscr{F}$ is a σ-algebra in Ω.
- $\mu(\emptyset) = 0$
- If $(A_n)_{n\in\mathbb{N}} \subset \mathscr{F}$ are pair-wise disjoint, then

$$\mu\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mu(A_n).$$

Examples of Measures

- The set function on λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ that assigns every half-open rectangle $[a, b)$ the value

$$\lambda([a, b)) = b - a.$$

  is called the *one-dimensional Lebesgue measure*.
- Let $\Omega = \{\omega_1, \omega_2, \dots\}$ be a countable and $(p_1, p_2, \dots)$ a sequence of numbers $p_n \in [0, 1]$, such that $\sum_{n\in\mathbb{N}} p_n = 1$. On $(\Omega, \mathcal{P}(\omega), P)$ the set function

$$P(A) = \sum_{n:\omega_n\in A}, \quad A \in \Omega,$$

  defines a *probability measure*. The triplet $(\Omega, \mathcal{P}(\omega), P)$ is called discrete probability space.
- Box 2 shows an example of a discrete probability measure. Here we reproduce one of the tables of Box 2.

Ω	$P(\omega)$
$\omega_1 = \mathsf{H^cW^cS^cR^c}$	α_1
$\omega_2 = \mathsf{H^cW^cS^cR}$	α_2
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_5 = \mathsf{H^cWS^cR^c}$	α_5
$\omega_6 = \mathsf{H^cWS^cR}$	α_6
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_9 = \mathsf{HW^cS^cR^c}$	α_9
$\omega_{10} = \mathsf{HW^cS^cR}$	α_{10}
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{13} = \mathsf{HWS^cR^c}$	α_{13}
$\omega_{14} = \mathsf{HWS^cR}$	α_{14}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

$$P(\omega_k) = \alpha_k,$$
$$\sum_{k=1}^{16} P(\omega_k) = \sum_{k=1}^{16} \alpha_k = 1.$$

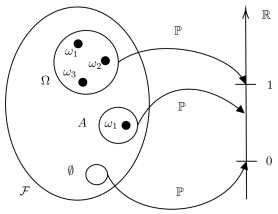


Figure 12: A probability measure is a special case of a measure. Suppose you have a measurable space. Then a function $\mu : \mathcal{F} \to [0, +\infty]$ is called a measure on $(\Omega, \mathcal{F})$ if it gives a size of zero, to the empty set and if it is countably additive. Notice how the measure takes elements of $\mathcal{F}$ and not elements of the sample space Ω; i.e., it maps events, not outcomes. A probability measure is just a measure, with the additional property that it maps to $[0, 1]$ rather than to $[0, +\infty]$. In order words, it assigns probabilities to events. We can see in the diagram above how the probability measure maps the empty set to zero, the sample space to 1, and any other event to some number in $[0, 1]$.

Probability Measures

In non-technical terms, a probability measure is a way of describing how likely or unlikely something is to happen. It's a bit like looking at weather forecasts that tell you the chances of rain. When we say there's a 50% probability of rain, it means that out of every two similar situations, it's expected to rain in one of them.

So, a probability measure assigns a number between 0 and 1 to events, where 0 means the event won't happen, 1 means it will definitely happen, and values in between tell us the likelihood of something occurring. It helps us understand and work with uncertainty, make predictions, and make informed decisions based on the chances of different events (Figure 19).

Definition 11: Probability Measure Definition

A probability measure is a a map $P = \mathbb{P} : \mathscr{F} \to [0, 1]$ satisfying:

- $\mathscr{F}$ is a σ-algebra in Ω.
- $P(\emptyset) = 0$
- $P(\Omega) = 1$
- If $(A_n)_{n\in\mathbb{N}} \subset \mathscr{F}$ are pair-wise disjoint, then

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n).$$

let $(\Omega, \mathscr{F}, P)$ be a probability space, $A, B, B_n, A_n \in \Omega$, $(A_n)_{n\in\mathbb{N}}$ pairwise disjoint.

1. $A \cap B = \emptyset \implies P(A \cup B) = P(A) + P(B)$.

2. $A \subset B \implies P(A) \leq P(B)$.
3. $A \subset B, \implies P(B \setminus A) = P(B) - P(A)$.
4. $P(A \cup B) + P(A \cap B) = P(A) + P(B)$.
5. $P(A \cup B) \leq P(A) + P(B)$.
6. $P\left(\bigcup_{n\in\mathbb{N}} B_n\right) \leq \sum_{n\in\mathbb{N}} P(B_n)$.
7. $\biguplus_{n\in\mathbb{N}} A_n = \Omega \implies P(A) = \sum_{n\in\mathbb{N}} P(A \cap A_n)$.

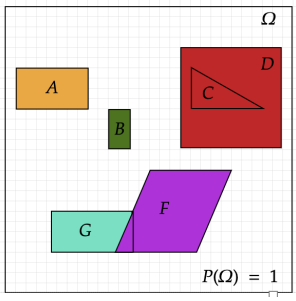


Figure 13: Events in the sample space Ω.

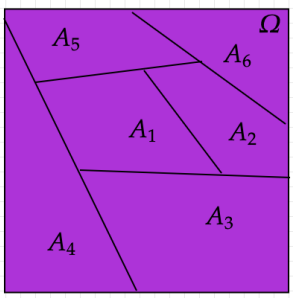


Figure 14: Property 7 example.

Probability Measure Examples

- We assume the outcome can be directly observed at the end of the experiment and thus $\mathscr{F}$ is chosen to be the largest possible σ-algebra, and we define P on it. The precise definition of P depends on the application:
  - For a finite sample space Ω where each outcome is equally likely, define P on $\mathscr{F} = \mathscr{P}(\Omega)$ via $P(A) = \frac{|A|}{|\Omega|}$ for any $A \in \mathscr{F}$.
  - To model the number of coin flip required to obtain the first head ($\Omega = \{1, 2, 3, ...\}$), define P on $\mathscr{F} = \mathscr{P}(\Omega)$ where P satisfies $P(\{\omega : \omega = k\}) = (1-p)^{k-1}p$. Here $p \in (0, 1)$ represents the chance of getting a head in a single flip.
- To represent a uniform random number draw from $\Omega = [0, 1]$, define P on $\mathscr{F} = \mathcal{B}([0, 1])$ where P satisfies $P([a, b]) = b - a$ for $0 \leq a \leq b \leq 1$. This is the Lebesgue measure on $[0, 1]$.

Independent Events

$A_m, A_n \in \mathscr{F}$ are independent $\implies P(A_m \cap A_n) = P(A_m)P(A_n)$

- $(A_n)_{n\in\mathbb{N}}$ are pair-wise independent $\implies A_m$ and A_n are independent for any $n \neq m$, $n, m \in \mathbb{N}$
- $(A_n)_{n\in I}$, are independent $\implies P(\bigcap_{n\in I} A_n) = \prod_{n\in I} P(A_n)$

Conditional probability

The conditional probability of event A conditioned to event B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

where $P(B) > 0$. A conditional probability is a *probability measure*. If A and B are independent $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

Chain rule

Consider the events $A_1, A_2, \cdots, A_n$. The chain rule is defined as follows.

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) =$$

$$P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1})$$

- Apply the chain rule to the following expression $P(A_1 \cap A_2 \cap A_3 \cap A_4)$.
- Is the following expression correct? $P(A_1 \cap A_2 \cap A_3) = P(A_1 \mid A_2 \cap A_3)P(A_2 \mid A_3)p(A_3)$.

Bayes Theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where $P(B) > 0$. Recall that $P(B) = \sum_{n\in\mathbb{N}} P(B \cap A_n)$ with $\biguplus_{n\in\mathbb{N}} A_n = \Omega$. Therefore:

$$\begin{aligned} P(A_m \mid B) &= \frac{P(B \mid A_m)P(A_m)}{P(B)} \\ &= \frac{P(B \mid A_m)P(A_m)}{\sum_{n\in\mathbb{N}} P(B \cap A_n)} = \frac{P(B \mid A_m)P(A_m)}{\sum_{n\in\mathbb{N}} P(B \mid A_n)P(A_n)} \end{aligned}$$

Applying the Bayes Theorem

- Mr. Holmes now lives in Los Angeles. One morning when Homes leaves his house, he realizes that his grass is wet (H). Is it due to rain (R), or has he forgotten to turn off the sprinkler (S)? Next, he notices that the grass of his neighbor, Dr. Watson, is also wet (W).
  - What is the probability that Holme's grass is wet (H) given that he forgot to turn the sprinkler off (S)?
  - What is the probability that Holme's grass is wet (H) given that he forgot to turn the sprinkler off (S) and Dr. Watson's grass is also wet (W)?
- Consider the following events.
  - $A =$ "Holmes' grass is wet."
  - $A^c =$"Holmes grass is dry. "
  - $B =$"Holmes forgot to turn the sprinkler off."
  - $B' =$ "Holmes forgot to turn the sprinkler off and Watson's grass is wet."

  We write down the outcomes including these events and their corresponding probabilities in the tables shown below:

Ω	$P(\omega)$
$\omega_1 = \mathsf{H^cW^cS^cR^c}$	α_1
$\omega_2 = \mathsf{H^cW^cS^cR}$	α_2
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_5 = \mathsf{H^cWS^cR^c}$	α_5
$\omega_6 = \mathsf{H^cWS^cR}$	α_6
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_9 = \mathsf{HW^cS^cR^c}$	α_9
$\omega_{10} = \mathsf{HW^cS^cR}$	α_{10}
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{13} = \mathsf{HWS^cR^c}$	α_{13}
$\omega_{14} = \mathsf{HWS^cR}$	α_{14}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

A	$P(\omega)$
$\omega_9 = \mathsf{HW^cS^cR^c}$	α_9
$\omega_{10} = \mathsf{HW^cS^cR}$	α_{10}
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{13} = \mathsf{HWS^cR^c}$	α_{13}
$\omega_{14} = \mathsf{HWS^cR}$	α_{14}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

A^c	$P(\omega)$
$\omega_1 = \mathsf{H^cW^cS^cR^c}$	α_1
$\omega_2 = \mathsf{H^cW^cS^cR}$	α_2
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_5 = \mathsf{H^cWS^cR^c}$	α_5
$\omega_6 = \mathsf{H^cWS^cR}$	α_6
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8

B	$P(\omega)$
$\omega_3 = \mathsf{H^cW^cSR^c}$	α_3
$\omega_4 = \mathsf{H^cW^cSR}$	α_4
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_{11} = \mathsf{HW^cSR^c}$	α_{11}
$\omega_{12} = \mathsf{HW^cSR}$	α_{12}
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

B'	$P(\omega)$
$\omega_7 = \mathsf{H^cWSR^c}$	α_7
$\omega_8 = \mathsf{H^cWSR}$	α_8
$\omega_{15} = \mathsf{HWSR^c}$	α_{15}
$\omega_{16} = \mathsf{HWSR}$	α_{16}

Box 5: Example solutions

We use conditional probability and the Bayes theorem to compute the requested probabilities.

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{\alpha_{11} + \alpha_{12} + \alpha_{15} + \alpha_{16}}{\sum_{k=9}^{16} \alpha_k}.$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{\alpha_{11} + \alpha_{12} + \alpha_{15} + \alpha_{16}}{\alpha_3 + \alpha_4 + \alpha_7 + \alpha_8 + \alpha_{11} + \alpha_{12} + \alpha_{15} + \alpha_{16}}.$$

$$P(B' \mid A) = \frac{P(B' \cap A)}{P(A)} = \frac{\alpha_{15} + \alpha_{16}}{\sum_{k=9}^{16} \alpha_k}.$$

$$P(A \mid B') = \frac{P(B' \mid A)P(A)}{P(B')} = \frac{\alpha_{15} + \alpha_{16}}{\alpha_7 + \alpha_8 + \alpha_{15} + \alpha_{16}}.$$

Random Variables

A random variable is also called a measurable function. Suppose you have two measurable spaces. One could be the pair of sample space and event space $(\Omega, \mathcal{F})$, and the other one could be some other arbitrary pair of a set and of its σ-algebra $(E, \mathcal{E})$, although usually, we choose $E = \mathbb{R}^n$ and $\mathcal{E} = \mathcal{B}(\mathbb{R}^n)$, $n = 1, 2, \ldots$. Then a measurable function is a function X that maps elements in Ω to elements in E with some additional properties. Notice how this function maps outcomes to elements of E, it does not map events (Figure 20).

The mapping X guarantees a correspondence between the events in our original event space and our transformed event space. For this reason, we require the random variable $X : \Omega \to E$ ($X : \Omega \to \mathbb{R}$) is such that to be such that the preimage $X^{-1}(B)$of any $\mathcal{E}$-measurable set $B \in \mathcal{B}$ is a $\mathcal{F}$ measurable set.

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} \ \ \forall B \in \mathcal{E}.$$

In the diagram below we can see how the set B, which is an element of $\mathcal{E}$ has a pre-image, $X^{-1}(B)$, which is an element of $\mathcal{F}$. From here on, we use $E = \mathbb{R}^n$, and $\mathcal{E} = \mathcal{B}(\mathbb{R}^n)$ to define a random variable.

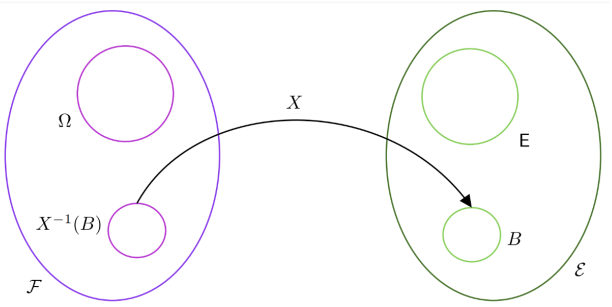


Figure 17: Random variable definition.

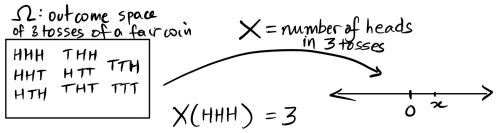


Figure 15: Random variable example defined on the real line $\mathbb{R}$. Notice that the inverse mapping X^{-1} maps real numbers into measurable events. For example $\{HHH\} = X^{-1}(3)$ and . $\{\{HHT\}, \{HTH\}, \{THH\}\} = X^{-1}(2)$ belong to the σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$.

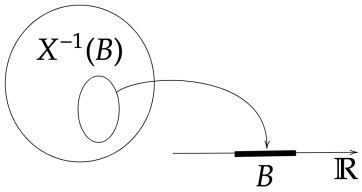


Figure 16: Random variable definition.

Definition 12: Random Variable

Let $(\Omega, \mathscr{F}, P)$ be a probability space. A *random variable* X is a $(\mathcal{F}/\mathcal{B}(\mathbb{R}^n))$ measurable map $X : \Omega \to \mathbb{R}^n$, if for every Borel set $B \in \mathcal{B}(\mathbb{R}^n)$:

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{F}.$$

Clearly,

- $X^{-1}(B) \in \mathscr{F} \quad \forall B \in \mathcal{B}(\mathbb{R}^n)$.

Types of Random Variables

In practice, we seldom bother working with the abstract concept of a probability space $(\Omega, \mathscr{F}, P)$, but rather just focus on the distributional properties of a random variable X representing the random phenomenon. We are interested in random variables that are *discrete* and *continuous*.

- A random variable X is discrete if it only takes values on a countable set $\Omega = \{\omega_1, \omega_2, \omega_3, \dots\}$, which is called the support of X.
- A random variable X is a continuous random variable if it only takes values on a non-countable set Ω.
- Random variables can also be mixed.

Examples of Discrete Random Variables

- Consider the experiment in which we roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
  - Let $\mathscr{F} = \mathscr{P}(\Omega)$. The random variable $X_1(\omega) = \omega$, is $\mathscr{F}/\mathcal{B}(\mathbb{R})$ measurable. X_1 gives the exact outcome of the roll.
  - Let $\mathscr{F}_1 = \{\emptyset, \Omega, \{1, 3, 5\}, \{2, 4, 6\}\}$. The random variable

    $$X_2(\omega) = \begin{cases} 1, & \omega \in \{1, 3, 5\} \\ 0, & \omega \in \{2, 4, 6\} \end{cases}$$

    is $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable; its value depends on whether the roll is odd or even.
  - If we only have information on whether the roll is odd or even, we can determine the value of X_2 but not the value of X_1.
  - The random variable X_1, is not $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable. For instance, $X_1^{-1}(1) = \{1\} \notin \mathscr{F}_1$.
- Consider the experiment in which we roll a three-faced dice: $\Omega = \{-1, 0, 1\}$ and $\mathscr{F}_1 = \{\emptyset, \Omega, \{-1, 1\}, \{0\}\}$.
  - $X_1(\omega) = \omega$ is not $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable. For example, $X^{-1}(1) = \{1\} \notin \mathscr{F}_1$.
  - $X_2(\omega) = \omega^2$ is $\mathscr{F}_1/\mathcal{B}(\mathbb{R})$ measurable.

- Let $(\Omega, \mathscr{F})$ describe throwing two fair dice, i.e. $\Omega := \{(i,k) : 1 \leq i,k \leq 6\}$, $\mathscr{F} = \mathscr{P}(\Omega)$. The total number of points thrown $X : \Omega \to \{2,3,\ldots,12\}$, $U((i,j)) = i+j$ is a measurable map.
- *A σ-algebra generated by a random variable U, denoted by $\sigma(U)$, is the smallest σ-algebra for which U is measurable. For example, for $\Omega = \{\mathsf{HH}, \mathsf{HT}, \mathsf{TH}, \mathsf{TT}\}$.

$$X_1(\omega) = \begin{cases} 1, & \omega \in \{\mathsf{HH}, \mathsf{HT}\}, \\ 0, & \omega \in \{\mathsf{TH}, \mathsf{TT}\}. \end{cases} \quad X_2(\omega) = \begin{cases} 2, & \omega \in \{\mathsf{HH}\}, \\ 1, & \omega \in \{\mathsf{HT}\}, \\ -1, & \omega \in \{\mathsf{TH}\}, \\ -2, & \omega \in \{\mathsf{TT}\}. \end{cases}$$

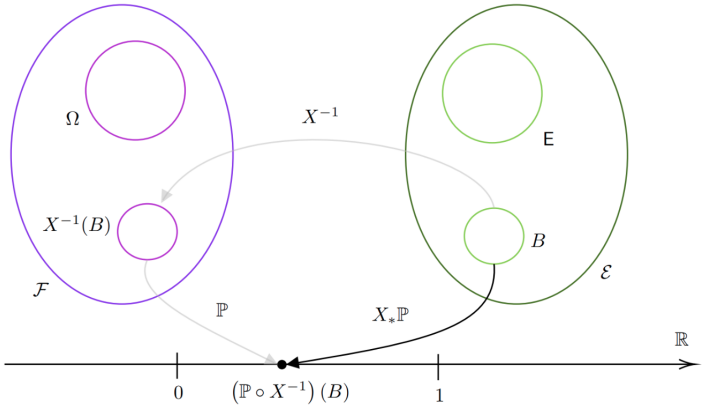
  $\sigma(X_1) = \{\emptyset, \{\mathsf{HH}, \mathsf{HT}\}, \{\mathsf{TH}, \mathsf{TT}\}, \Omega\}$ and $\sigma(X_2) = \mathscr{P}(\Omega)$. In particular, $\sigma(X_1) \subset \sigma(X_2) = \mathscr{P}(\Omega)$.

Probability Distributions

A probability distribution is also called a *push-forward* ($P_X = P_* X = P \circ X^{-1}$) measure of the *probability measure* P, via the random variable X. Suppose we have a probability space $(\Omega, \mathcal{F}, P)$ This means we can assign probabilities to events in $\mathcal{F}$. Now suppose we have a measurable space $(E, \mathcal{E})$ but we don't yet have a probability measure to measure events in it. How can we go about measuring sets in $\mathcal{E}$?

The key idea is that, given a set B in $\mathcal{E}$, we can use a random variable X to find the pre-image of such set in the event space $\mathcal{F}$ and then measure this set via the probability measure P. This will then be our probability measurement for B, as shown in the figure below.

The *probability distribution*, is defined as $P_X = P_* X = P \circ X^{-1} : \mathcal{E} \to [0,1]$. The probability distribution is therefore a function mapping sets in $\mathcal{E}$ into $[0,1]$. Here we use $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, that is, $P \circ X^{-1} : \mathcal{B}(\mathbb{R}^n) \to [0,1]$.



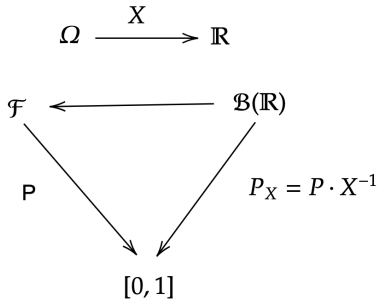


Figure 18: Probability distribution defined on the real line $\mathbb{R}$.

Figure 19: Probability distribution definition.

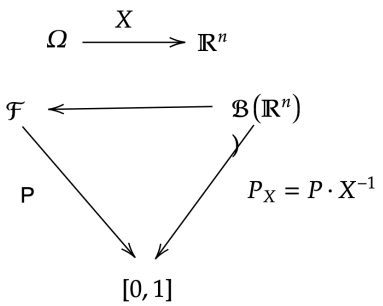


Figure 20: Probability distribution defined on the n-dimensional real space $\mathbb{R}$.

Definition 13: Probability Distribution

Let $(\Omega, P, \mathscr{F})$ be a measure space, and X a random variable $X : \Omega \to \mathbb{R}^n$, i.e. a measurable map. Then

$$P(X^{-1}(B)) = P(\{\omega : X(\omega) \in B\}) = P(X \in B).$$

is a probability measure called the *law* or *distribution* of the random variable X. Note: Here we use $(E, \mathcal{E}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, that is, $P \circ X^{-1} : \mathcal{B}(\mathbb{R}^n) \to [0, 1]$.

Examples of Probability Distributions for Discrete Random Variables

- Let $(\Omega, \mathscr{F})$ describe throwing two fair dice, i.e. $\Omega := \{(i,k) : 1 \leq i,k \leq 6\}$, $\mathscr{F} = \mathscr{P}(\Omega)$, and $P(\{i,j\}) = \frac{1}{36}$. The total number of points thrown $X : \Omega \to \{2,3,\ldots,12\}$, $X((i,j)) = i + j$ is a measurable map (Table 1 and Figure 21).

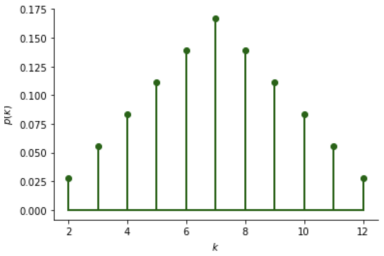


Figure 21: Probability distribution $P(K = k)$ in Table 1.

k	2	3	4	5	6	7	8	9	10	11	12
$P(X = k)$	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Table 1: The distribution of the random variable X.

- The *Bernoulli random* variable $X \in \{0,1\}$ with parameter p, $0 < \mathsf{p} \leq 1$ has the following probability distribution:

$$P(X = x \mid \mathsf{p}) := \mathsf{Bernoulli}(X = x \mid \mathsf{p}) = \mathsf{p}^x(1-\mathsf{p})^{1-x}.$$

  Hint: $P(X = 1 \mid \mathsf{p}) = \mathsf{p}$.

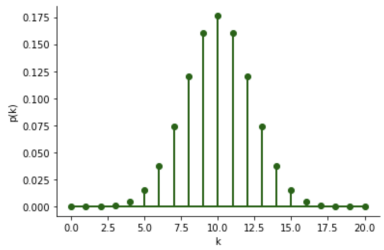


Figure 22: Binomial probability distribution function for $N = 20$ and $p = 0.5$.

- The *Binomial distribution* with parameters N and p is the discrete probability distribution of the number K of successes in a sequence of N independent Bernoulli trials (with parameter p). The probability distribution is

$$P(K = k \mid \mathsf{N}, \mathsf{p}) := \mathsf{Binomial}(K = k \mid \mathsf{p}, \mathsf{N}) = \binom{\mathsf{N}}{k} \mathsf{p}^k (1-\mathsf{p})^{N-k}$$

  for $k = 0, 1, 2, \ldots$ where

$$\binom{\mathsf{N}}{k} = \frac{\mathsf{N}!}{(\mathsf{N}-k)!k!}$$

  is the number of ways of choosing $K = k$ objects out of a total of N identical objects.

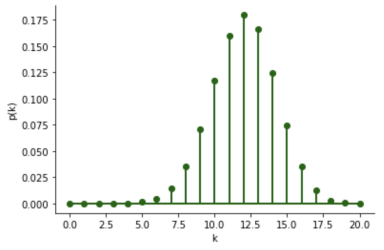


Figure 23: Binomial probability distribution function for $N = 20$ and $p = 0.6$.

Continuous Random Variables

A random variable X is a continuous random variable if there exists a non-negative function $f_X(\cdot)$ such that:

$$P(X \leq \omega) = \int_{-\infty}^{\omega} f_X(\alpha) d\alpha$$

for any $\omega \in \mathbb{R}$. The function f_X is called the probability density function of X. To simplify notation in practice we use $p(x) = f(x) = f_X(x)$. We remark on the abuse of notation.

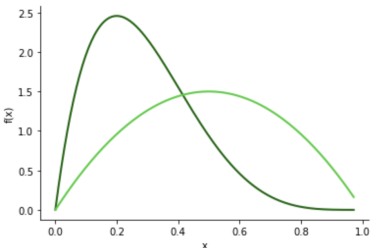


Figure 24: Beta probability density function. $\alpha = 2, \beta = 5$ (dark green), $\alpha = 2, \beta = 2$ (lime green).

Examples of Continuous Random Variables

- A random variable $M \in [0,1]$ has a Beta distribution of variable with parameters α and β if the density function has the form

$$f_M(m \mid \alpha, \beta) := \text{Beta}(m \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} m^{\alpha-1}(1-m)^{\beta-1},$$

  where $\Gamma(x)$ is the Gamma function $\Gamma(x) = \int_0^x u^{x-1}e^{-u}du$.
- A random variable $X \in \mathbb{R}$ has a Gaussian or Normal distribution of variable with parameters μ and σ^2 if the density function has the form

$$f_X(x \mid \mu, \sigma^2) := \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\},$$

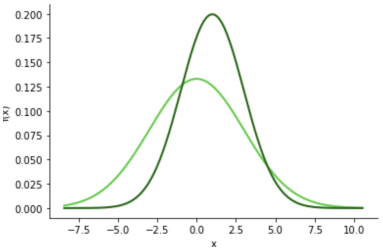


Figure 25: Gaussian probability density function. $\mu = 0, \sigma^2 = 3$ (dark green), $\mu = 2, \sigma^2 = 2$ (lime green).

Cummulative Distribution Function

For a random variable X, its cumulative distribution function (CDF) is defined as:

$$F_X(x) := P(X \leq x), \quad -\infty \leq x \leq \infty$$

Note that $P(X \leq x) = P \circ X^{-1}((-\infty, x])$, and:

- $F_X(x)$ is non-decreasing and right-continuous.
- $\lim_{x\to\infty} F_X(x) = 1$ and
- $\lim_{x\to-\infty} F_X(x) = 0$

Conversely, if a given function F_X satisfies the above properties, then it is a CDF of some random variable.

As an example, we show below the cumulative distribution of the random variable X in Table 1 (see Figures 21 and 22).

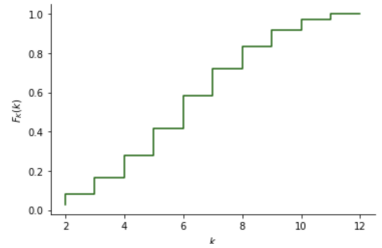


Figure 26: Cumulative distribution of the random variable X in Table 1.

k	2	3	4	5	6	7	8	9	10	11	12
$F_K(k)$	0.02	0.08	0.16	0.27	0.41	0.58	0.72	0.83	0.91	0.97	1

Drawing Samples from Random Variable

Drawing samples from a random variable refers to the process of generating individual realizations or instances of that random variable according to its probability distribution. In probability theory and statistics, a random variable is a variable whose possible values are outcomes of a random phenomenon. The probability distribution of a random variable describes the likelihood of different values it can take.

When you draw samples from a random variable, you are essentially simulating or generating data points that follow the probability distribution of that variable. This process is often used in various fields such as statistics, machine learning, and simulations to understand the behavior of random phenomena or to make predictions.

For example, if you have a random variable representing the outcome of a fair six-sided die roll, drawing samples from this random variable would involve simulating the roll of the die and obtaining values like 1, 2, 3, 4, 5, or 6 with equal probability.

The concept of drawing samples is fundamental to Monte Carlo simulations, where random sampling is used to estimate numerical results and analyze complex systems that involve randomness. In statistical terms, the more samples you draw, the better your approximation of the true underlying distribution or properties of the random variable.

Expected Value and Variance of a Discrete Random Variable

For a discrete random variable X, its expected value is defined as:

$$\mu_X = \mathrm{E}[X] = \sum_{\omega \in \Omega} X(\omega) p_X(\omega) = \sum_x x p(x).$$

For example, for the random variable K of Table 1, the expected value is

$$\begin{aligned}\mu_K &= 2(\frac{1}{36}) + 3(\frac{1}{18}) + 4(\frac{1}{12}) + 5(\frac{1}{9}) + 6(\frac{5}{36}) + 7(\frac{1}{6}) + 8(\frac{5}{36}) + 9(\frac{1}{9}) + 10(\frac{1}{12}) + 11(\frac{1}{18}) + 12(\frac{1}{36}) \\ &= 7.\end{aligned}$$

More generally, for a given function $g(\cdot)$ we define:

$$\mu_{g(X)} = \mathrm{E}[g(X)] = \sum_{\omega \in \Omega} g(X(\omega)) p_X(\omega) = \sum_x g(x) p(x).$$

The variance of X is defined as:

$$\mathrm{var}[X] = \sum_{\omega \in \Omega} (X(\omega) - E[X])^2 p_X(\omega) = \sum_x (x - \mu_X)^2 p(x) = E\left[X^2\right] - (E[X])^2$$

For example, for the random variable K of Table 1, the variance is computed as follows

$$\begin{aligned}
\mathrm{E}[K^2] &= 2^2(\frac{1}{36}) + 3^2(\frac{1}{18}) + 4^2(\frac{1}{12}) + 5^2(\frac{1}{9}) + 6^2(\frac{5}{36}) + 7^2(\frac{1}{6}) + 8^2(\frac{5}{36}) + 9^2(\frac{1}{9}) + 10^2(\frac{1}{12}) + 11^2(\frac{1}{18}) + 12^2(\frac{1}{36}) \\
&= 54.83 \\
(E[K])^2 &= 49 \\
\mathrm{var}[K] &= 5.83.
\end{aligned}$$

Expected Value and Variance of a Continuous Random Variable

For a continuous random variable X with probability density $f_X(x)$, its expected value is defined as:

$$\mu_x = \mathrm{E}[X] = \int_{-\infty}^{\infty} \omega f_X(\omega) d\omega.$$

More generally, for a given function $g(\cdot)$ we define:

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} f_X(\omega) f_X(\omega) d\omega.$$

The variance of X is defined as:

$$\mathrm{var}[X] = \int_{-\infty}^{\infty} (\omega - E[X])^2 f_X(\omega) d(\omega) = E\left[X^2\right] - (E[X])^2.$$

Joint Probability Distribution

When dealing with multiple random variables, it is sometimes useful to use vector and matrix notations. Let $X_1, X_2, \ldots, X_N$ be N discrete random variables.

- The *joint probability function* of $X_1, X_2, \ldots, X_N$ is denoted as

  $$p(x_1, x_2, \ldots, x_N) := p_{X_1, X_2, \ldots, X_N}(x_1, x_2, \ldots, x_N) = P(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N).$$

- Using vector notation we can write this distribution as

  $$p(\mathbf{x}) := p_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x})$$

  where $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$ is a column (random) vector having $X_1, X_2, \ldots, X_N$ as its components. Similarly $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$. Note the abuse of notation in defining $p(\mathbf{x})$ and $p(x_1, x_2, \ldots, x_N)$ above.
- Let $X_1, X_2, \ldots X_N$ be a set of discrete random variables with joint distribution $p(x_1, x_2, \ldots, x_N)$. Variable X_n can be *marginalized* from the joint distribution function as follows.

  $$p(\mathbf{x}_{\neg n}) = p(x_1, x_2, \ldots, x_{n-1}, x_{n+1}, \ldots, x_N) \quad = \sum_{x_n} p(x_1, x_2, \ldots, x_N).$$

x_1	x_2	x_3	$p(x_1, x_2, x_3)$
0	0	0	θ_1
0	0	1	θ_2
0	1	0	θ_3
0	1	1	θ_4
1	0	0	θ_5
1	0	1	θ_6
1	1	0	θ_7
1	1	1	θ_8

x_1	x_2	$p(x_1, x_2) = \sum_{x_3} p(x_1, x_2, x_3)$
0	0	$\theta_1 + \theta_2$
0	1	$\theta_3 + \theta_4$
1	0	$\theta_5 + \theta_6$
1	1	$\theta_7 + \theta_8$

- Let X_1, X_2 be two discrete random variables. The *conditional distribution function* of X_1 given X_2 is defined as

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)},$$
$$p(x_1, x_2) = p(x_1 \mid x_2)p(x_2) = p(x_2 \mid x_1)p(x_1),$$

with $p(x_2) > 0$.

x_1	x_2	$p(x_1, x_2)$
0	0	$\theta_1 + \theta_2$
0	1	$\theta_3 + \theta_4$
1	0	$\theta_5 + \theta_6$
1	1	$\theta_7 + \theta_8$

x_2	$p(x_2) = \sum_{x_1} p(x_1, x_2)$
0	$\theta_1 + \theta_2 + \theta_5 + \theta_6$
1	$\theta_3 + \theta_4 + \theta_7 + \theta_8$

x_1	x_2	$p(x_1 \mid x_2)$
0	0	$(\theta_1 + \theta_2)/(\theta_1 + \theta_2 + \theta_5 + \theta_6)$
0	1	$(\theta_3 + \theta_4)/(\theta_3 + \theta_4 + \theta_7 + \theta_8)$
1	0	$(\theta_5 + \theta_6)/(\theta_1 + \theta_2 + \theta_5 + \theta_6)$
1	1	$(\theta_7 + \theta_8)/(\theta_3 + \theta_4 + \theta_7 + \theta_8)$

$\sum_{x_1} p(x_1 \mid x_2) = 1$

- Let X_1, X_2, and X_3 be three discrete random variables. Compute $p(x_2 \mid x_3)$ from the table given below. Compute $\sum_{x_2} p(x_2 \mid x_3, x_1)$.

x_2	x_1	x_3	$p(x_2 \mid x_3, x_1)$
0	0	0	θ_1
0	0	1	θ_2
0	1	0	θ_3
0	1	1	θ_4
1	0	0	$1 - \theta_1$
1	0	1	$1 - \theta_2$
1	1	0	$1 - \theta_3$
1	1	1	$1 - \theta_4$

x_2	x_3	$p(x_2 \mid x_3)$
0	0	$\theta_1 + \theta_3$
0	1	$\theta_2 + \theta_4$
1	0	$1 - (\theta_1 + \theta_3)$
1	1	$1 - (\theta_2 + \theta_4)$

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ be two sets of discrete random variables. The *conditional distribution function* of $\mathbf{X}$

given $\mathbf{Y}$ is defined as

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

with $p(\mathbf{y}) > 0$.

- Let X_1, X_2 be two discrete random variables. The Bayes rule establishes that

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} = \frac{p(x_2 \mid x_1)p(x_1)}{p(x_2)} = \frac{p(x_2 \mid x_1)p(x_1)}{\sum_{x_1} p(x_2 \mid x_1)p(x_1)}$$

with $p(x_2) > 0$.

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ be two sets of discrete random variables. The *Bayes rule* establishes that

$$p(\mathbf{x} \mid \mathbf{y}) = \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})} = \frac{p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \mathbf{x})p(\mathbf{y})}{\sum_{\mathbf{x}} p(\mathbf{y} \mid \mathbf{x})p(\mathbf{x})}$$

with $p(\mathbf{y}) > 0$.

- Let X_1, X_2 be two discrete random variables. These variables are independent if

$$p(x_1, x_2) = p(x_1)p(x_2),$$

or alternatively,

$$p(x_1 \mid x_2) = p(x_1).$$

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ be two sets of discrete random variables. These sets of variables are *independent* if

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{x})p(\mathbf{y}),$$

or alternatively,

$$p(\mathbf{x} \mid \mathbf{y}) = p(\mathbf{x}).$$

- Let X_1, X_2, X_3 be three discrete random variables. We say that X_1 is *conditionally independent* of X_2 given X_3 (denoted by $X_1 \perp\!\!\!\perp X_2 \mid X_3$) if

$$p(x_1, x_2 \mid x_3) = p(x_1 \mid x_3)p(x_2 \mid x_3)$$

or alternatively,

$$p(x_1 \mid x_2, x_3) = p(x_1 \mid x_3).$$

- Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_M\}$ and $\mathbf{Z} = \{Z_1, Z_2, \ldots, Z_L\}$ be three sets of discrete random variables. We say that $\mathbf{X}$ is *conditionally independent* of $\mathbf{Y}$ given $\mathbf{Z}$ (denoted by $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}$) if

$$p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})$$

or alternatively,

$$p(\mathbf{x} \mid \mathbf{y}, \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z}).$$

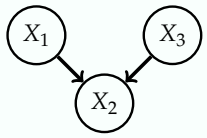
- *Bayesian networks in brief* A Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG). Bayesian networks are directed acyclic graphs (DAGs) whose nodes represent variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters, or hypotheses. Each edge represents a direct conditional dependency. Any pair of nodes that are not connected (i.e. no path connects one node to the other) represent variables that are conditionally independent of each other. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node. Bayesian networks having $X_1, X_2, \ldots, X_N$ variables factorize as

$$p(x_1, x_2, \ldots, x_N) = \prod_{n=1}^{N} p(x_n | \mathsf{pa}(X_n)),$$

where $\mathsf{pa}(\mathsf{x_n})$ represent the parents of variable X_a in the associated DAG.

Box 6: Bayesian Network Example

Consider the example of a Bayesian network having three random variables X_1, X_2, X_3.



$$p(x_1, x_2, x_3) = p(x_2 \mid x_1, x_3) p(x_3) p(x_1)$$

x_2	x_1	x_3	$p(x_2 \mid x_3, x_1)$
0	0	0	θ_1
0	0	1	θ_2
0	1	0	θ_3
0	1	1	θ_4
1	0	0	$1 - \theta_1$
1	0	1	$1 - \theta_2$
1	1	0	$1 - \theta_3$
1	1	1	$1 - \theta_4$

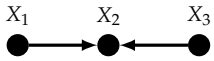
x_1	$p(x_1)$
0	θ_5
1	$1 - \theta_5$

x_3	$p(x_3)$
0	θ_6
1	$1 - \theta_6$

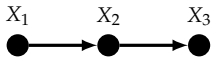
- *Bayesian Network Examples*
Solution.

Draw the DAG associated with the following probability distributions:

1. $p(x_1, x_2, x_3) = p(x_2)p(x_3)p(x_2 \mid x_1, x_3)$. Show that $X_1 \perp\!\!\!\perp X_3$.

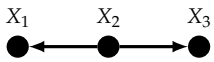


2. $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_2)$. Show that $X_1 \perp\!\!\!\perp X_3|X_2$.



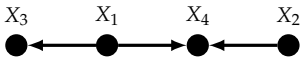
$$
\begin{aligned}
p(x_1, x_3 \mid x_2) &= \frac{p(x_1, x_2, x_3)}{p(x_2)} \\
&= \frac{p(x_1)p(x_2|x_1)p(x_3|x_2)}{p(x_2)} \\
&= \frac{p(x_1, x_2)p(x_3|x_2)}{p(x_2)} \\
&= \frac{p(x_1 \mid x_2)p(x_2)p(x_3|x_2)}{p(x_2)} \\
&= p(x_1|x_2)p(x_3|x_2) \\
&\implies X_1 \perp\!\!\!\perp X_3|X_2
\end{aligned}
$$

3. $p(x_1, x_2, x_3) = p(x_2)p(x_1|x_2)p(x_3|x_2)$. Show that $X_1 \perp\!\!\!\perp X_3|X_2$.

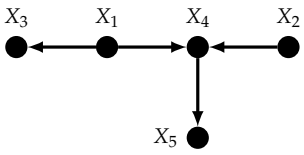


$$
\begin{aligned}
p(x_1, x_3 \mid x_2) &= \frac{p(x_1, x_2, x_3)}{p(x_2)} \\
&= \frac{p(x_2)p(x_1|x_2)p(x_3|x_2)}{p(x_2)} \\
&= p(x_1|x_2)p(x_3|x_2) \\
&\implies X_1 \perp\!\!\!\perp X_3|X_2
\end{aligned}
$$

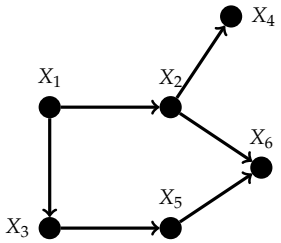
4. $p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_2)$. Show that $X_3 \perp\!\!\!\perp X_4|X_1$.



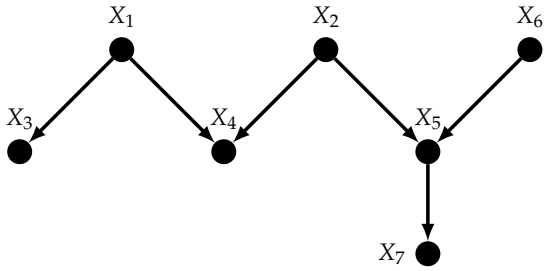
5. $p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1)p(x_4|x_1, x_2)p(x_5|x_4)$. Show that $X_1 \perp\!\!\!\perp X_5|X_4$.



6. $p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_2, x_5)$. Show that $X_2 \perp\!\!\!\perp X_3|X_1$.



7. Write down the factorization of $p(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ encoded in the following DAG.



Box 7: Application Example (Was it the burglar?)

Mary lives in San Francisco City. One afternoon, she is driving back home and receives a phone call from her neighbor Jane. She told her that her house alarm was set off (A). While driving, she also heard on the radio (R) that a small earthquake (E) hit the city. Small earthquakes sometimes activate the alarm, and perhaps this is the reason why the alarm was sounding.

- What is the probability that a burglar (B) broke into the house?
- What is the probability that a burglar broke into the house given that the alarm was set off?
- What is the probability that a burglar broke into the house given that the alarm was set off and a small earthquake hit the city?

To solve this problem we make several considerations.

1. Define the random variables $R(\mathsf{R}) = 1$, $R(\mathsf{R}^c) = 0$, $A(\mathsf{A}) = 1$, $A(\mathsf{A}^c) = 0$, $E(\mathsf{E}) = 1$, $E(\mathsf{E}^c) = 0$, and $B(\mathsf{B}) = 1$, $B(\mathsf{B}^c) = 0$.
2. Assume that the joint distribution of random variables R, A, E, and B factorizes as

$$p(A, R, E, B) = p(E = e)p(B = b)p(A = a \mid B = b, E = e)p(R = r \mid E = e)$$

our using our simplified notation

$$p(a, r, e, b) = p(e)p(b)p(a \mid b, e)p(r \mid e).$$

The factors of the probability distribution correspond to the following tables.

	b	$p(b)$
B^c	0	$\mathsf{b}_1 = .9$
B	1	$1 - \mathsf{b}_1$

	e	$p(e)$
E^c	0	$\mathsf{e}_1 = 0.95$
E	1	$\mathsf{e}_2 = 1 - \mathsf{e}_1$

		b	e	$p(b,e)$
B^c	E^c	0	0	$\mathsf{b}_1\mathsf{e}_1$
B^c	E	0	1	$\mathsf{b}_1\mathsf{e}_2$
B	E^c	1	0	$\mathsf{b}_2\mathsf{e}_1$
B.	E	1	1	$\mathsf{b}_2\mathsf{e}_2$

		r	e	$p(r \mid e)$
R^c	E^c	0	0	$\mathsf{f}_1 = 0.99$
R^c	E	0	1	$\mathsf{f}_2 = 0.01$
R	E^c	1	0	$\mathsf{f}_3 = 1 - \mathsf{f}_1$
R.	E	1	1	$\mathsf{f}_4 = 1 - \mathsf{f}_2$

			a	b	e	$p(a \mid b,e)$
A^c	B^c	E^c	0	0	0	q_1
A^c	B^c	E	0	0	1	q_2
A^c	B	E^c	0	1	0	q_3
A^c	B	E	0	1	1	q_4
A	B^c	E^c	1	0	0	$\mathsf{q}_5 = 1 - \mathsf{q}_1$
A	B^c	E	1	0	1	$\mathsf{q}_6 = 1 - \mathsf{q}_2$
A	B	E^c	1	1	0	$\mathsf{q}_7 = 1 - \mathsf{q}_3$
A	B	E	1	1	1	$\mathsf{q}_8 = 1 - \mathsf{q}_4$

The table representing the joint probability function is

				a	r	b	e	$p(r,a,e,b)$
A^c	R^c	B^c	E^c	0	0	0	0	$\mathsf{p}_1 = \mathsf{b}_1\mathsf{e}_1\mathsf{f}_1\mathsf{q}_1$
A^c	R^c	B^c	E	0	0	0	1	$\mathsf{p}_2 = \mathsf{b}_2\mathsf{e}_1\mathsf{f}_1\mathsf{q}_1$
A^c	R^c	B	E^c	0	0	1	0	$\mathsf{p}_3 = \mathsf{b}_1\mathsf{e}_2\mathsf{f}_2\mathsf{q}_2$
A^c	R^c	B	E	0	0	1	1	$\mathsf{p}_4 = \mathsf{b}_2\mathsf{e}_2\mathsf{f}_2\mathsf{q}_2$
A^c	R	B^c	E^c	0	1	0	0	$\mathsf{p}_5 = \mathsf{b}_1\mathsf{e}_1\mathsf{f}_3\mathsf{q}_3$
A^c	R	B^c	E	0	1	0	1	$\mathsf{p}_6 = \mathsf{b}_2\mathsf{e}_1\mathsf{f}_3\mathsf{q}_3$
A^c	R	B	E^c	0	1	1	0	$\mathsf{p}_7 = \mathsf{b}_1\mathsf{e}_2\mathsf{f}_4\mathsf{q}_4$
A^c	R	B	E	0	1	1	1	$\mathsf{p}_8 = \mathsf{b}_2\mathsf{e}_2\mathsf{f}_4\mathsf{q}_4$
A	R^c	B^c	E^c	1	0	0	0	$\mathsf{p}_9 = \mathsf{b}_1\mathsf{e}_1\mathsf{f}_1\mathsf{q}_5$
A	R^c	B^c	E	1	0	0	1	$\mathsf{p}_{10} = \mathsf{b}_2\mathsf{e}_1\mathsf{f}_1\mathsf{q}_5$
A	R^c	B	E^c	1	0	1	0	$\mathsf{p}_{11} = \mathsf{b}_1\mathsf{e}_2\mathsf{f}_2\mathsf{q}_6$
A	R^c	B	E	1	0	1	1	$\mathsf{p}_{12} = \mathsf{b}_2\mathsf{e}_2\mathsf{f}_2\mathsf{q}_6$
A	R	B^c	E^c	1	1	0	0	$\mathsf{p}_{13} = \mathsf{b}_1\mathsf{e}_1\mathsf{f}_3\mathsf{q}_7$
A	R	B^c	E	1	1	0	1	$\mathsf{p}_{14} = \mathsf{b}_2\mathsf{e}_1\mathsf{f}_3\mathsf{q}_7$
A	R	B	E^c	1	1	1	0	$\mathsf{p}_{15} = \mathsf{b}_1\mathsf{e}_2\mathsf{f}_4\mathsf{q}_8$
A	R	B	E	1	1	1	1	$\mathsf{p}_{16} = \mathsf{b}_2\mathsf{e}_2\mathsf{f}_4\mathsf{q}_8$

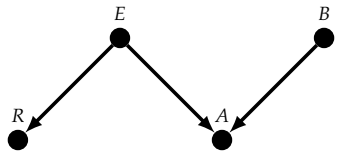


Figure 27: Directed Acyclic Graph encoding the probability distribution $p(a,r,e,b) = p(e)p(b)p(a \mid b,e)p(r \mid e)$. Dark circles represent random variables R, A, E, and B.

- The *graphical model* (a Directed Acyclic Graph or DAG) encoding the probability distribution factorization is shown in Figure 21.
- Assume that $q_1 = 0.98$, $q_2 = 0.9$, $q_3 = 0.1$, $q_4 = 0.01$.
- What is the probability that a burglar broke into the house?

$$\begin{aligned} P(B = 1) = P(\mathsf{B}) = \\ = \mathsf{p}_3 + \mathsf{p}_4 + \mathsf{p}_7 + \mathsf{p}_8 + \mathsf{p}_{11} + \mathsf{p}_{12} + \mathsf{p}_{15} + \mathsf{p}_{16}, \end{aligned}$$

$$P(B = 1) = \sum_a \sum_r \sum_e p(r, a, B = 1, e) = 0.1.$$

- What is the probability that a burglar broke into the house given that the alarm was set off?

$$\begin{aligned} P(B = 1 \mid A = 1) &= \frac{P(B = 1, A = 1)}{P(A = 1)} \\ &= \frac{\mathsf{p}_{11} + \mathsf{p}_{12} + \mathsf{p}_{15} + \mathsf{p}_{16}}{\mathsf{p}_9 + \mathsf{p}_{10} + \mathsf{p}_{11} + \mathsf{p}_{12} + \mathsf{p}_{13} + \mathsf{p}_{14} + \mathsf{p}_{15} + \mathsf{p}_{16}}, \end{aligned}$$

$$P(B = 1 \mid A = 1) = \frac{\sum_r \sum_e p(r, A = 1, B = 1, e)}{\sum_r \sum_b \sum_e p(r, A = 1, b, e)} \qquad = 0.81.$$

- What is the probability that a burglar broke into the house given that the alarm was set off and a small earthquake hit the city?

$$\begin{aligned} P(B = 1 \mid A = 1, E = 1) &= \frac{P(B = 1, A = 1, E = 1)}{P(A = 1, E = 1)} \\ &= \frac{\mathsf{p}_{12} + \mathsf{p}_{16}}{\mathsf{p}_{10} + \mathsf{p}_{12} + \mathsf{p}_{14} + \mathsf{p}_{16}}, \end{aligned}$$

$$\begin{aligned} P(B = 1 \mid A = 1, E = 1) &= \frac{\sum_r p(r, A = 1, B = 1, E = 1)}{\sum_r \sum_b p(r, A = 1, b, E = 1)} \\ &= 0.51. \end{aligned}$$

Box 8: Solution Summary

- What is the probability that a burglar broke into the house?

$$P(B = 1) = 0.1.$$

- What is the probability that a burglar broke into the house given that the alarm was set off?

$$P(B = 1 \mid A = 1) = 0.81.$$

- What is the probability that a burglar broke into the house given that the alarm was set off and a small earthquake hit the city?

$$P(B = 1 \mid A = 1, E = 1) = 0.51.$$

- The *a priori* probability that the burglar broke into the house is 0.1. The probability that a burglar broke into the house given that the alarm was set off is much higher (0.81). However, knowing that a small earthquake hit the city *explains away* the observation that the alarm was set off, diminishing the probability that a burglar broke (0.51).

- Note that the factorization features of the probability distribution and the sum-product distributive property help us reduce the computational complexity.

$$\begin{aligned}
p(b \mid a) = \frac{p(a,b)}{p(a)} &= \frac{\sum_r \sum_e p(r,a,e,b)}{\sum_b \sum_r \sum_e p(r,a,e,b)} \\
&= \frac{\sum_r \sum_e p(e)p(b)p(a \mid e,b)p(r \mid e)}{\sum_b \sum_r \sum_e p(e)p(b)p(a \mid e,b)p(r \mid e)} \\
&= \frac{p(b)\sum_e p(e)p(a \mid e,b)\sum_r p(r \mid e)}{\sum_b p(b)\sum_e p(e)p(a \mid e,b)\sum_r p(r \mid e)} \\
&= \frac{p(b)\sum_e p(e)p(a \mid e,b)\phi_1(e)}{\sum_b p(b)\sum_e p(e)p(a \mid e,b)\phi_1(e)} \\
&= \frac{p(b)\phi_2(a,b)}{\sum_b p(b)\phi_2(a,b)} \\
&= \frac{p(b)\phi_2(a,b)}{\phi_3(a)}
\end{aligned}$$

Box 9: Application Example (Pairs of dice)

You are told that there are two pairs of coins. The first pair is fair,

$$\theta_1(\mathsf{H}) = \theta_1(\mathsf{T}) = \frac{1}{2}.$$

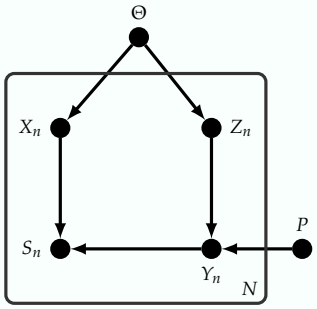
The second pair is biased as both coins have probability

$$\theta_2(\mathsf{H}) = \frac{2}{3},\ \theta_2(\mathsf{T}) = \frac{1}{3}$$

of producing heads and tails, respectively. One of the two pairs is chosen at random with probability 0.5 and thrown 10 times. The sum of the coins is recorded for each throw.

1. If the throw results in the sequence "1010101010"? Which pair of dice was picked for the throw?
2. Repeat (1) for the sequence "2222200000".
3. Assume that the result of the second coin is flipped (F) with probability $\mathsf{p} = 0.5$ before recording the result of the throw with probability. Repeat 1 and 2 under this assumption. Consider the case for which $\mathsf{p} = 0.2$. Repeat exercises 1 and 2.

The graphical model associated with this problem is as follows.



- X_n represents the n-th result from the throw of the first coin; $X_n(\mathsf{H}) = 1$ and $X_n(\mathsf{T}) = 0$, and

$$p(x_n \mid \theta) = \theta^{x_n}(1-\theta)^{(1-x_n)}.$$

- Z_n epresents the unobserved result from the n-th throw of the second coin; $Z_n(\mathsf{F}) = 1$ and $Z_n(\mathsf{F}^c) = 0$, and

$$p(z_n \mid \theta) = \theta^{z_n}(1-\theta)^{(1-z_n)}.$$

- Y_n represents the recorded result from the n-th throw of the second coin; $Y_n(\mathsf{H}) = 1$ and $Y_n(\mathsf{T}) = 0$, and

$$p(y_n \mid z_n, p) = p^{I(y_n \neq z_n)}(1-p)^{I(y_n = z_n)}.$$

- S_n is a deterministic function of X_n and Y_n,

$$S_n = X_n + Y_n,$$

$$p(s_n \mid x_n, y_n) = \mathrm{I}(s_n == x_n + y_n),$$

and

$$I(e) = \begin{cases} 1, & \text{if } e = \mathsf{true}; \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)$, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)$, $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_N)$, $\mathbf{S} = (S_1, S_2, \ldots, S_N)$, $f_n = I(y_n \neq z_n) = 1 - I(y_n = z_n)$, and $\mathbf{f} = (f_1, f_2, \ldots, f_N)$. ($f_n$ indicates whether the second was flipped or not.) The joint probability distribution for $\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s}$ given θ and p is

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s}, \theta, p) &= p(\theta)p(p) \prod_{n=1}^{N} p(x_n \mid \theta)p(z_n \mid \theta)p(y_n \mid z_n, p)p(s_n \mid x_n, y_n) \\
p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s} \mid \theta, p) &= \frac{p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s}, \theta, p)}{p(\theta)p(p)} \\
p(\mathbf{x}, \mathbf{z}, \mathbf{y}, \mathbf{s} \mid \theta, p) &= \prod_{n=1}^{N} p(x_n \mid \theta)p(z_n \mid \theta)p(y_n \mid z_n, p)p(s_n \mid x_n, y_n) \\
&= \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{(1-x_n)}\theta^{z_n}(1-\theta)^{(1-z_n)}p^{I(y_n \neq z_n)}(1-p)^{I(y_n = z_n)}I(s_n = x_n + y_n). \\
&= \theta^{\sum_{n=1}^{N} x_n}(1-\theta)^{\sum_{n=1}^{N}(1-x_n)}\theta^{\sum_{n=1}^{N} z_n}(1-\theta)^{\sum_{n=1}^{N}(1-z_n)}p^{\sum_{n=1}^{N} I(y_n \neq z_n)}(1-p)^{\sum_{n=1}^{N} I(y_n = z_n)}
\end{aligned}
$$

Notice that

$$
\begin{aligned}
x_n &= s_n - y_n, \\
y_n &= z_n(1 - f_n) + (1 - z_n)f_n,
\end{aligned}
$$

and

$$
x_n = s_n - z_n + 2z_n f_n - f_n.
$$

We define N_x, and N_z as follows,

$$
N_x = \sum_{n=1}^{N} (s_n - z_n + 2z_n f_n - f_n) = N_s - N_z + 2N_{zf} - N_f \geq 0,
$$

$$
\begin{aligned}
N_s &= \sum_{n=1}^{N} s_n, \\
N_z &= \sum_{n=1}^{N} z_n, \\
N_{zf} &= \sum_{n=1}^{N} z_n f_n,
\end{aligned}
$$

and

$$
N_f = \sum_{n=1}^{N} f_n.
$$

Therefore we can write the probability distribution function as follows,

$$\begin{aligned} p(\mathbf{x},\mathbf{z},\mathbf{y},\mathbf{s} \mid \theta, p) &= \theta^{\sum_{n=1}^{N} x_n}(1-\theta)^{\sum_{n=1}^{N}(1-x_n)}\theta^{\sum_{n=1}^{N} z_n}(1-\theta)^{\sum_{n=1}^{N}(1-z_n)}p^{\sum_{n=1}^{N} f_n}(1-p)^{\sum_{n=1}^{N}(1-f_n)} \\ &= \theta^{N_x}(1-\theta)^{(N-N_x)}\theta^{N_z}(1-\theta)^{(N-N_z)}p^{N_f}(1-p)^{N-N_f} \\ &= \theta^{N_x+N_z}(1-\theta)^{(2N-(N_x+N_z))}p^{N_f}(1-p)^{N-N_f} \\ &= p(\mathbf{s},\mathbf{z},\mathbf{f} \mid \theta, p) \end{aligned}$$

Observe that $N_x = N_x(\mathbf{s},\mathbf{z},\mathbf{f})$, $N_z = N_z(\mathbf{z})$, and therefore

$$p(\mathbf{x},\mathbf{z},\mathbf{y},\mathbf{s} \mid \theta, p) = p(\mathbf{s},\mathbf{z},\mathbf{f} \mid \theta, p).$$

Also,

$$p(\mathbf{s} \mid \theta, p) = \sum_{\mathbf{x}}\sum_{\mathbf{z}}\sum_{\mathbf{y}} p(\mathbf{x},\mathbf{z},\mathbf{y},\mathbf{s} \mid \theta, p) = \sum_{\mathbf{f}}\sum_{\mathbf{z}} p(\mathbf{s},\mathbf{z},\mathbf{f} \mid \theta, p),$$

and

$$\begin{aligned} p(\theta \mid \mathbf{s}, p) &= \frac{p(\mathbf{s},\theta,p)}{p(\mathbf{s},p)} = \frac{p(\mathbf{s} \mid \theta, p)p(\theta,p)}{p(\mathbf{s},p)} \\ &= \frac{p(\mathbf{s} \mid \theta, p)p(\theta \mid p)p(p)}{p(\mathbf{s} \mid p)p(p)} \\ &= \frac{p(\mathbf{s} \mid \theta, p)p(\theta \mid p)}{p(\mathbf{s} \mid p)} \\ &= \frac{p(\mathbf{s} \mid \theta, p)p(\theta)}{p(\mathbf{s} \mid p)}. \end{aligned}$$

since $\theta \perp\!\!\!\perp p$.

Box 10: Decision Rule

$$\mathrm{I}\left(\frac{p(\theta_1 \mid, \mathbf{s}, p)}{p(\theta_2 \mid, \mathbf{s}, p)} > 1\right) \Rightarrow \text{choose } \theta_1,$$

which is equivalent to

$$\mathrm{I}\left(\frac{p(\mathbf{s} \mid \theta_1, p)p(\theta_1)}{p(\mathbf{s} \mid \theta_2, p)p(\theta_2)} > 1\right) \Rightarrow \text{choose } \theta_1,$$

since

$$p(\theta \mid, \mathbf{s}, p) \propto p(\mathbf{s} \mid \theta, p)p(\theta) = \sum_{\mathbf{f}}\sum_{\mathbf{z}} p(\mathbf{s},\mathbf{z},\mathbf{f} \mid \theta, p)p(\theta).$$

Notice that

$$p(\mathbf{s}|p) = \sum_{\theta}\sum_{\mathbf{f}}\sum_{\mathbf{z}} p(\mathbf{s},\mathbf{z},\mathbf{f} \mid \theta, p)p(\theta),$$

with $\theta \in \{\theta_1, \theta_2\}$.

These equations lead to the following results regarding question 3.

$$\frac{p([2,2,2,2,2,0,0,0,0,0] \mid \frac{2}{3},\frac{1}{2})p(\frac{2}{3})}{p([2,2,2,2,2,0,0,0,0,0] \mid \frac{1}{2},\frac{1}{2})p(\frac{1}{2})} = 1.80 \implies \theta_2 = \frac{2}{3}.$$
$$\frac{p([1,0,1,0,1,0,1,0,1,0] \mid \frac{2}{3},\frac{1}{2})p(\frac{2}{3})}{p([1,0,1,0,1,0,1,0,1,0] \mid \frac{1}{2},\frac{1}{2})p(\frac{1}{2})} = 0.13 \implies \theta_1 = \frac{1}{2}.$$

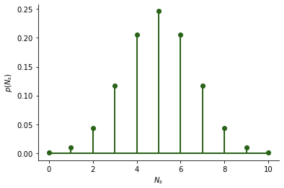


Figure 28: Probability distribution $p(N_s)$ in Box for $\theta = \frac{1}{2}$ for problem 3 in Box 9. The expected value $\mu_{N_s} = 5$ and the entropy is $E(N_s) = 2.71$.

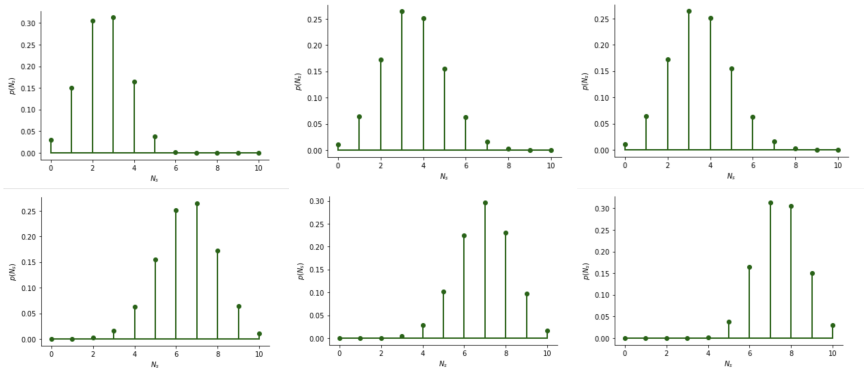


Figure 29: From left to right and from top to bottom, probability distributions $p(N_s)$ in Box for $\theta = \frac{1}{100}$, θ are $\frac{1}{9}$, $\frac{1}{5}$, $\frac{99}{100}$, $\frac{8}{9}$, $\frac{4}{5}$ for problem 3 in Box 9. The corresponding mean values, μ_{N_s}, are 2.55, 3.06, 3.5, 6.5, 6.9, respectively. and $\mu_{N_s} = 7.45$, respectively. The corrresponding entropies, $E(N_s)$, are 2.44, 2.56, 2.56, 2.44, and 2.22, respectively.

Also, notice that

$$\mathrm{E}[\mathbf{s} \mid \theta, p] = \sum_{\mathbf{s}} \sum_{\mathbf{f}} \sum_{\mathbf{z}} \mathbf{s} p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p),$$
$$\mathrm{E}[N_s \mid \theta, p] = \sum_{N_s} N_s p(N_s \mid \theta, p).$$

The *effect* of θ on N_s is

$$e(\theta_1, \theta_2 \mid p) = \mathrm{E}[N_s \mid \theta_1, p] - E[N_s \mid \theta_2, p]$$

Definition 14: Entropy

The average amount of information of a discrtete random variable X is the expectation of $\mathrm{I}(x) = -\log_2(p(x))$ with respect to the distribution $p(x)$ and is given by

$$H(X) = \mathrm{E}[\mathrm{I}(x))] = \mathrm{E}[-\log_2(p(x))] = -\sum_x p(x)\log_2(p(x)).$$

The entropy of a Bernoulli random variable $X \sim \mathsf{Bernoulli}(\mathsf{x} \mid \mathsf{p})$ is

$$H(X \mid p) = -p\log_2(p) - (1-p)\log_2(1-p).$$

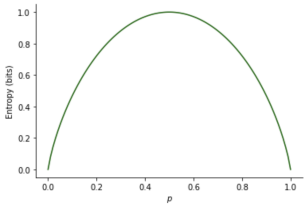


Figure 30: Entropy of the random variable $X \sim \mathsf{Bernoulli}(\mathsf{x} \mid \mathsf{p})$.

Kullback–Leibler Divergence

The Kullback–Leibler divergence (also called KL-divergence, relative entropy, and I-divergence) is a measure of how one probability distribution P_1 is different from a second, reference probability distribution P_2. For discrete probability distributions, P_1 and P_2 defined on the same sample space, the relative entropy from P_2 to P_1 is defined to be

$$D_{KL}(P_1 \| P_2) = -\sum_x p_1(x) \log_2(\frac{p_2}{p_1}).$$

In other words, it is the expectation of the logarithmic difference between the probabilities P_1 and P_2, where the expectation is taken using the probabilities P_1.

Conjugate Priors

Let X and Θ be two random variables. In Bayesian probability theory, if the posterior distribution $p(\theta|x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a *conjugate prior* for the likelihood function $p(x|\theta)$. Recall that:

$$\underbrace{p(\theta \mid x)}_{posterior} = \frac{\overbrace{p(x \mid \theta)}^{likelihood} \overbrace{p(\theta)}^{prior}}{\underbrace{p(x)}_{evidence}},$$

where

$$p(x) = \underbrace{\int p(x \mid \theta) p(\theta) d\theta}_{marginalization}.$$

For example, if $p(\theta)$ has a Beta distribution, and $p(x \mid \theta)$ has a binomial distribution, then $p(\theta \mid x)$ also has a Beta distribution. More specifically,

$$\begin{aligned}
p(x \mid \theta) &\sim \mathsf{Binomial}(x, \mid N, p) \\
p(\theta \mid \alpha, \beta) &\sim \mathsf{Beta}(\theta \mid \alpha, \beta) \\
p(\theta \mid x) &\sim \mathsf{Beta}(x \mid \alpha + \sum_{n=1}^{N} x_i, \beta + N - \sum_{n=1}^{N} x_i)
\end{aligned}$$

Consider the probability distribution from Box 9:

$$p(\mathbf{s}, \mathbf{z}, \mathbf{f} \mid \theta, p) = \theta^{N_x + N_z} (1-\theta)^{(2N - (N_x + N_z))} p^{N_f} (1-p)^{N - N_f}$$

Clearly,

$$p(\theta, p \mid \mathbf{s}, \mathbf{z}, \mathbf{f}) = \mathsf{Beta}(\theta \mid \alpha + \alpha_1, \beta + \beta_1)\mathsf{Beta}(p \mid \gamma + \gamma_1, \delta + \delta_1)$$

where $\alpha_1 = N_x + N_z$, $\beta_1 = N - (N_x + N_z)$, $\delta_1 = N_f$, $\gamma_2 = N - N_f$, $p(\theta \mid \alpha, \beta) = \mathsf{Beta}(\theta \mid \alpha, \beta)$ and $p(s \mid \gamma, \delta) = \mathsf{Beta}(s \mid \gamma, \delta)$.

Marginalizing with respect to p, $\mathbf{z}$, and $\mathbf{f}$ we obtain:

$$\begin{aligned} p(\theta \mid \mathbf{s}) &= \sum_{\mathbf{z}} \sum_{\mathbf{f}} \mathsf{Beta}(\theta \mid \alpha + \alpha_1, \beta + \beta_1) \int_p \mathsf{Beta}(p \mid \gamma + \gamma_1, \delta + \delta_1) dp, \\ &= \sum_{\mathbf{z}} \sum_{\mathbf{f}} \mathsf{Beta}(\theta \mid \alpha + \alpha_1, \beta + \beta_1). \end{aligned}$$

Expected Value and Covariance Matrix of Random Vectors

Let $X_1, X_2, \ldots, X_N$ be N discrete random variables with joint probability distribution $p(x_1, x_2, \ldots, x_N)$. Let $\mathbf{X} = [X_1, X_2, \ldots, X_N]^T$

- The expected value of $\mathbf{X}$ is given by

$$\mathrm{E}[\mathbf{X}] = \sum_{\mathbf{x}} \mathbf{x} p(\mathbf{x}) = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_N]]^T.$$

- The *covariance* of two discrete random variables X and Y is given by

$$\mathrm{cov}(X, Y) = \mathrm{E}[(X - \mu_x)(Y - \mu_y)] = \mathrm{E}[XY] - \mu_x \mu_y,$$

  where

$$\mathrm{E}[XY] = \sum_{\omega_x \in \Omega_x} \sum_{\omega_y \in \Omega_y} X(\omega_x) Y(\omega_y) P(X = X(\omega_x), Y = Y(\omega_y)).$$

  Using the notation defined in the previous section we can write

$$\mathrm{E}[XY] = \sum_x \sum_y xy p(x, y).$$

- The covariance of vector $\mathbf{X}$ is defined as

$$\mathrm{cov}(\mathbf{X}) = \begin{bmatrix} \mathrm{cov}(X_1) & \mathrm{cov}(X_1, X_2) & \ldots & \mathrm{cov}(X_1, X_N) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2, X_N) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(X_1, X_N) & \mathrm{cov}(X_2, X_N) & \ldots & \mathrm{var}(X_N) \end{bmatrix}$$

- The covariance of the random vectors $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\mathrm{cov}(\mathbf{X}, \mathbf{Y}) = \begin{bmatrix} \mathrm{cov}(X_1, Y_1) & \mathrm{cov}(X_1, Y_2) & \ldots & \mathrm{cov}(X_1, Y_N) \\ \mathrm{cov}(X_2, Y_1) & \mathrm{cov}(X_2, Y_2) & \ldots & \mathrm{cov}(X_2, Y_N) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(X_N, Y_1) & \mathrm{cov}(X_N, Y_2) & \ldots & \mathrm{cov}(X_N, Y_N) \end{bmatrix}$$

The Multivariate Gaussian Distribution

The Gaussian, also known as the normal distribution, is a widely used model for the distribution of continuous variables. In the case of a single variable x, the Gaussian distribution (probability density function) can be written in the form

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\},$$

where μ is the mean and σ^2 is the variance. For a N-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{\|\boldsymbol{\Sigma}\|^{\frac{1}{2}}} \exp\{-(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\},$$

where $\boldsymbol{\mu}$ is a N-dimensional mean vector, $\boldsymbol{\Sigma}$ is a $N \times N$ covariance matrix, and $\|\boldsymbol{\Sigma}\|$ denotes the determinant of Σ.

The Gaussian distribution arises in many different contexts and can be motivated from a variety of different perspectives. For example, the Gaussian distribution arises when we consider the sum of multiple random variables. The central limit theorem (due to Laplace) tells us that, subject to certain mild conditions, the sum of a set of random variables, which is of course itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increases. We can illustrate this by considering N variables $N_1, \ldots, X_N$ each of which has a uniform distribution over the interval $[0, 1]$, and then considering the density of the mean $(X_1 + \cdots + X_N)/N$. For large N, this density tends to be a Gaussian density (Figures 31 and 32).

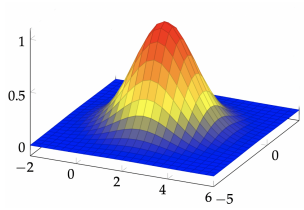


Figure 31: Surface plot of a two-dimensional Gaussian density.

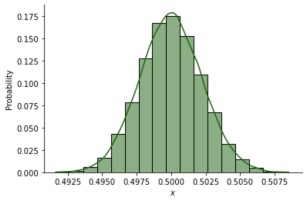


Figure 32: Histogram and KDE plots from 20,000 samples of $(X_1 + \cdots + X_N)/N$ for $N = 50{,}000$, with $X_n \sim \text{Bernoulli}(x \mid p)$. The estimated mean and variance are 0.49, and $5.04e-06$, respectively.

The Central Limit Theorem

In practice, the convergence to a Gaussian as N increases can be rapid. One consequence of this result is that the binomial distribution, which is the sum of N observations of the Bernoulli random variable, will tend to a Gaussian as $N \to \infty$ (Figure 31).

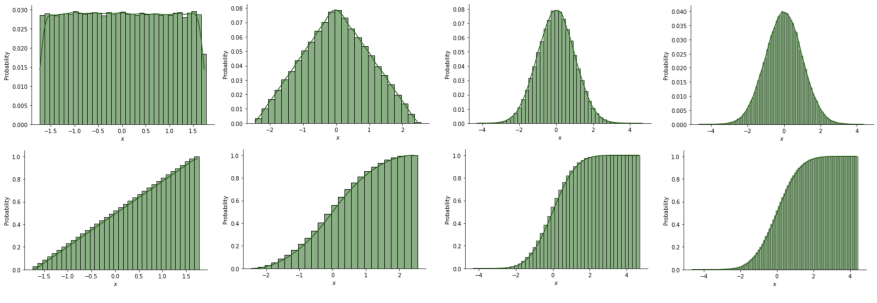


Figure 33: Top: Histogram and KDE plots of Z_N for various values of N: $1, 2, 10$, and 100 (see Definition 15). We observe that as N increases, the density tends towards a Gaussian density. Bottom: Corresponding cumulative distribution functions. Here $X_n \sim \text{Uniform}(x_n \mid 0, 1)$ for wich $\mu = 0.5$ and $\sigma^2 = \frac{1}{12}$.

Theorem 1: The Central Limit Theorem (CLT)

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with expected value $\mathrm{E}[X_i] = \mu < \infty$ and variance $0 < \mathrm{var}(X_i) = \sigma^2 < \infty$. Then, the random variable

$$Z_N = \frac{\left[\sum_{n=1}^{N} X_n\right] - \mu}{\sqrt{N}\sigma}$$

converges to the standard Gaussian (Normal) random variable as $N \to \infty$,

$$\lim_{N\to\infty} F_{Z_N}(z) = \lim_{N\to\infty} P(Z_N \leq z) = \Phi(z) \quad , \forall z \in \mathbb{R},$$

where $\Phi(z)$ is the standard Gaussian cummulative distribution function

$$\Phi(z) = \int_{-\infty}^{z} \mathcal{N}(\alpha \mid, 0, 1) d\alpha.$$

That is,

$$Z_\infty \sim \mathcal{N}(z \mid, 0, 1).$$

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables with expected value $\mathrm{E}[X_i] = \mu < \infty$ and variance $0 < \mathrm{var}(X_i) = \sigma^2 < \infty$. Then, the random variable

$$\bar{X}_N = \frac{1}{N}\sum_{n=1}^{N} X_n,$$

i.e. the average of $X_1, X_2, ..., X_N$, has a Gaussian probability density with mean μ and variance $\frac{\sigma^2}{N}$:

$$\bar{X}_N \sim \mathcal{N}(x \mid \mu, \frac{\sigma^2}{N}) \text{ as } N \to \infty.$$

Clearly, $\lim_{N\to\infty} \bar{X}_N = \mu$, so the average is an *unbiased estimator* of the mean.

Chebyshev's Inequality

Let X be a random variable with a finite expected value μ and finite non-zero variance σ^2. Then for any real number $k > 0$,

$$P(\mid X - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}.$$

Only the case $k > 1$ is useful. When $k \leq 1$, right-hand side $\frac{1}{k^2} \geq 1$ and the inequality is trivial as all probabilities are ≤ 1.

Box 11: Example

Suppose that an unbiased coin is thrown 100 times. What is the bound that the number of heads will be greater than 70 or less than 30?

- Let K be the number of heads. Because K has a binomial distribution with $\mu = 0.5$:
- $\mathrm{E}[K] = N\mu = 50$.
- $\mathrm{var}[K] = N\mu(1-\mu) = 25$.
- The standard deviation is $\mathrm{std}[K] = \sqrt{\mathrm{var}[K]} = \sqrt{25} = 5$.
- The values 70 and 30 are 20 units from the average, which is 4 standard deviations (i.e., 20/5).

$$P(\mid K - \mathrm{E}[K] \mid \geq 4\mathrm{var}[K]) \leq \frac{1}{4^2} = 0.0625.$$

- Repeat the previous exercise for a) $\mu = 0.6$ and b) $\mu = 0.4$.