Lecture notes for MA52112 (Statistics for Data Science)

Karim Anaya-Izquierdo (based on notes by Vangelis Evangelou)

2025-10-01

Table of contents

Overview of Statistics for Data Science

Synopsis

In this unit you will develop your understanding of the basic theory of probability and statistics and recognise when this theory can be applied in practice.

Learning outcomes

By the end of the unit you will be able to:

- perform elementary mathematical operations in probability and statistics
- translate real-world problems into a probabilistic or statistical framework
- solve statistical problems in abstract form
- critically interpret the outcomes of statistical analysis in a real-world context
- relate underlying theory to requirements in practical data science

Content

The laws of probability. Discrete and continuous random variables. Expectation, variance and correlation. Conditional and marginal distributions. Common distributions including the normal, binomial and Poisson. Statistical estimation including maximum likelihood. Hypothesis testing and confidence intervals.

Summative assessment

- **Exam:** 100% of unit mark.

Moodle page

Please see the Moodle page for this unit for more a more detailed overview on the organisation and expectations for Statistics for Data Science this year.

1 Probability Theory for Data Scientists

1.1 Set theory Concepts

Definition: Sample Space, Event, and Empty Set

Definition 1.1. Consider an uncertain scenario. This include a random experiment, a data-generating process or simply the future. We define the following concepts:

- **Sample Space (Ω)**: The set of all possible outcomes or results from the scenario. Sample spaces can be either countable or uncountable. If the elements of a sample space can be put into one-to-one correspondence with the set of integers, the sample space is countable. If the sample space contains only a finite number of elements, it is also countable. Otherwise is uncountable.
- **Event**: A subset of the sample space. It represents a specific outcome or a collection of outcomes of interest.
- **Empty Set ($\emptyset$)**: A set containing no elements. It represents an impossible event.

Example 1.1. If we flip a coin twice then the sample space can be written as:

$$\Omega = \{HH, HT, TH, TT\}$$

where H represents *heads* and T *tails*. This sample space is finite. An event (say A) could be *at least one head appears*, that is

$$A = \{HT, TH, HH\} \subset \Omega$$

Example 1.2. If we are analyzing customer purchase behavior for a single online transaction, the sample space could be the set of all possible combinations of items a customer might select from the store's catalog. This sample space is in principle finite and therefore countable. However, if the catalog is very large, the sample space can be considered uncontably large for practical purposes. More on this later.

An event could be "customer buys at least one item from category X", or "customer buys product Y".

Example 1.3. We measure the time (in seconds) it takes for a user to complete a task on a website. The time limit is predefined at 5 minutes. Then the sample space is $\Omega = \{0, 1, 2, 3, \dots, 300\}$ which is finite. If, however, we measure the time with arbitrary precision, then the sample space is the interval $(0, 300)$ of real numbers. This sample space is uncountable.

An event could be "user completes the task in under 2 minutes". In the former case, this corresponds to the set $A = \{1, 2 \dots, 119\}$. Int he latter case is the real interval $A = (0, 120)$.

Events can be described in many different ways. We will use set theory and notation to describe events and operations on events. This can help later in the computation of probabilities.

Basic Set Operations

Definition 1.2. Given events A, B, C in the sample space Ω:

- **Union ($A \cup B$)**: The event that A occurs, or B occurs, or both occur.
- **Intersection ($A \cap B$)**: The event that both A and B occur.
- **Complement (A^c)**: The event that A does not occur. It is the set of all outcomes in Ω that are not in A.

The following propertties hold for any events A, B, C:

- **Commutativity**:
  - Union: $A \cup B = B \cup A$
  - Intersection: $A \cap B = B \cap A$
- **Associativity**:
  - Union: $(A \cup B) \cup C = A \cup (B \cup C)$
  - Intersection: $(A \cap B) \cap C = A \cap (B \cap C)$
- **Distributive Laws**:
  - Intersection over Union: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
  - Union over Intersection: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- **De Morgan's Laws**:
  - $(A \cup B)^c = A^c \cap B^c$
  - $(A \cap B)^c = A^c \cup B^c$

Disjoint Sets and Partitions of Sample Space

Definition 1.3.

- **Disjoint Sets (Mutually Exclusive Events)**: Two sets A and B are disjoint if they have no elements in common, i.e., $A \cap B = \emptyset$.
- **Partition**: A collection of non-empty, disjoint subsets (events) of Ω whose union is Ω. That is $A_1, A_2, ...$ is a partition if

$$\bigcup_i A_i = \Omega \quad \text{and} \quad A_i \cap A_j = \emptyset \text{ for } i \neq j$$

Representation of events using set operations

Example 1.4. When we flip a coin twice, the event A "at least one head appears" can be written in variosu ways. These include

- the union of three events $A = \{HT\} \cup \{TH\} \cup \{HH\}$. That is, A occurs if we get heads on the first flip and tails on the second flip, or tails on the first flip and heads on the second flip, or heads on both flips. Note that these three events are disjoint as they do not share any outcomes.
- the union $A = A_1 \cup A_2$ where $A_1 = \{HT, HH\}$ is the event "head on first flip" and $A_2 = \{TH, HH\}$ is the event "head on second flip". Note that A_1 and A_2 are not disjoint as they both contain the outcome HH.
- the complement $A = B^c$ where $B = \{TT\}$ is the event "no heads appear".

Three different partitions of the sample space are given by:

- The trivial partition where each event contains a single outcome:

$$\mathcal{P}_1 = \{\{HT\}, \{TH\}, \{HH\}, \{TT\}\}$$

- The partition:

$$\mathcal{P}_{equal} = \{\{HH, TT\}, \{HT, TH\}\}$$

  that is, when we throw the coin twice, either we get the same results in both throws OR different ones.

- The partition where we group the outcomes based on the number of heads:

$$\mathcal{P}_{heads} = \{\{TT\}, \{HT, TH\}, \{HH\}\}$$

that is, when we flip the coin twice, we can get no heads, one head or two heads.

Sigma Algebra

Definition 1.4. A collection $\mathcal{F}$ of subsets of Ω is a **sigma algebra** (or σ-algebra) if it satisfies the following properties:

1. $\Omega \in \mathcal{F}$ (The sample space is in the collection).
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$ (The collection is closed under complementation).
3. If $A_1, A_2, ...$ are in $\mathcal{F}$, then $A_1 \cup A_2 \cup \cdots \cup A_n \in \mathcal{F}$ (The collection is closed under arbitray number of unions).

Note the definition of sigma-algebra does not explicitly require that the intersection of two sets in $\mathcal{F}$ is also in $\mathcal{F}$. However, this property follows from the other properties and De Morgan's laws. For example, if $A, B \in \mathcal{F}$, then

$$A \cup B \in \mathcal{F} \implies (A \cup B)^c = A^c \cap B^c \in \mathcal{F} \implies (A^c \cap B^c)^c = A \cup B \in \mathcal{F}$$

Examples of sigma-algebras

Example 1.5. The trivial sigma algebra is clearly $\mathcal{F}_0 = \{\emptyset, \Omega\}$ which does not seem very useful.
The partition $\mathcal{P}_{equal} = \{\{HH, TT\}, \{HT, TH\}\}$ above, is not a sigma-algebra as it does not contain the empty set. If we add the empty set, then is still not a sigma algebra as it is not closed under union. The union of the only two elements is Ω. If we include Ω then we have the sigma algebra:

$$\mathcal{F}_{equal} = \{\emptyset, \Omega, \{HH, TT\}, \{HT, TH\}\}$$

The partition $\mathcal{P}_{heads}$ above is also not a sigma algebra but if we add all possible unions then we obtain the sigma algebra:

$$\mathcal{F}_{heads} = \{\emptyset, \Omega, \{TT\}, \{HT, TH\}, \{HH\}, \{HT, TH, HH\}, \{HT, TH, TT\}, \{HH, TT\}\}$$

The set

$$\mathcal{G} = \{\emptyset, \Omega, \{HT\}, \{TH\}, \{HH\}, \{TT\}\}$$

is neither a partition nor a sigma algebra as it is not closed under union. For example, $\{HT\} \cup \{TH\} = \{HT, TH\} \notin \mathcal{G}$. However, if we add all possible unions of the elements of $\mathcal{G}$ we obtain the **power set** of Ω, that is the set of all subsets of Ω:

$$\begin{aligned}\mathcal{F}_{max} = \{&\emptyset, \{HT\}, \{TH\}, \{HH\}, \{TT\},\\ &\{HT, TH\}, \{HT, HH\}, \{HT, TT\}, \{TH, HH\}, \{TH, TT\}, \{HH, TT\}\\ &\{HT, TH, HH\}, \{HT, TH, TT\}, \{HT, HH, TT\}, \{TH, HH, TT\}, \Omega\}\end{aligned}$$

This is the largest possible sigma-algebra for this sample space. It has $2^4 = 16$ elements since the sample space has 4 elements. In general, if the sample space has n elements, then its power set has 2^n elements. Also generally, if we have a finite partition of Ω then the collection of all unions of sets in the partition (including the empty set) is a sigma-algebra.
Note that different sigma algebras serve for different purposes. For example, the sigma algebra $\mathcal{F}_{equal}$ is useful if we are only interested in whether the two coin flips are the same or different. The sigma algebra $\mathcal{F}_{heads}$ is useful if we areinterested in the number of heads. The power set $\mathcal{F}_{max}$ is a sigma algebra that me be more useful if we are interested in all possible events.

1.2 Probability

We will start by defining probability in an intuitive way. Later we will give a more formal mathematical definition .

1.2.1 Types of Probability

There are several ways to think about probability. These include

- **Classical Probability:** Assumes all possible outcomes in a finite sample space are equally likely. That is, for any event A with $n(A)$ outcomes in a sample space Ω with $n(\Omega)$ equally likely outcomes, the probability of A is:
$$P(A) = \frac{n(A)}{n(\Omega)}$$

  Example 1.6. Under this framework, the probability of rolling an even number on a die is assigned to be $P(\text{rolling an even number}) =$

$\frac{3}{6}$. More, generally this is equivalent to say the die is fair. Another example is when we assign the probability of rain tomorrow, locally at 10 AM, to be 1/2 as there are only two possible outcomes: rain or no rain.

- **Empirical (or Frequentist) Probability:** Based on observed frequencies from repeated experiments. As the number N of experiment repetitions increases, the probability of an event A approaches the true probability:

$$P(A) \approx \frac{\text{Number of times A occurred}}{N}$$

Example 1.7. If we do not what the probability of heads when flipping a coin is. We can we flip the coin 1000 times and if it lands heads 537 times, we would say the empirical probability of heads is 0.537. Furthermore we might say that the true probability of heads is ≈ 0.537 and the important aspect of thios framework is that, in theory, the more times we flip the coin the closer the empirical proportion will be to the true probability. Finally, according to historical data for our location, it has rained 33.6% of the days out of the last 10 years. The empirical probability of rain tomorrow locally at 10 AM is 0.336.

- **Subjective Probability:** Based on personal belief or judgment, often used when objective data is scarce.

Example 1.8. I had a look through the window and is a bit overcast, then I believe the probability of rain tomorrow locally at 10 AM is 0.7. On the other hand, if I am a weather expert from the point of atmospheric physics, I might believe the probability of rain tomorrow locally at 10 AM is 0.9.

1.2.2 Formal definition of probability

After we have chosen a sigma algebra $\mathcal{F}$ that contains events we are interested in, we can define probabilities for all the events in a more formal way.

Probability Measure (Kolmogorov's Axioms)

Definition 1.5. A **probability measure** P on a sample space Ω with a σ-algebra $\mathcal{F}$ is a function $P : \mathcal{F} \to [0, 1]$ that assigns a probability to each event in $\mathcal{F}$ and satisfies the following three axioms:

1. **Non-negativity**: For any event $A \in \mathcal{F}$, $P(A) \geq 0$. The probability of any event is non-negative.
2. **Normalization**: $P(\Omega) = 1$. The probability of the entire sample space (the certain event) is 1.
3. **Additivity (for disjoint events)**: If $A_1, A_2, \ldots, A_n$ are disjoint events in $\mathcal{F}$ (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

For a countably infinite sequence of disjoint events, this extends to:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

The probability of the union of disjoint events is the sum of their individual probabilities.

Probability measure for the equality of two coin flips

Example 1.9. For the sigma algebra $\mathcal{F}_{equal} = \{\emptyset, \Omega, \{HH, TT\}, \{HT, TH\}\}$ we can define a probability measure simply by specifying:

- $P(\emptyset) = 0$
- $P(\{HH, TT\}) = 0.4$

Note we can compute the probability of the other two events in $\mathcal{F}_{equal}$ using the axioms:

- $P(\Omega) = 1$ (by axiom 2)
- $P(\{HT, TH\}) = P(\Omega) - P(\{HH, TT\}) = 1 - 0.4 = 0.6$ (by axiom 3)

The assignment of probability of $\{HH, TT\}$ to be 0.4 maybe frequentist or subjective but regardless of this, it generates is a valid probability measure as it satisfies all three axioms.

Probability measure for the number of heads in two coin flips

Example 1.10. For the sigma algebra $\mathcal{F}_{heads}$ we can define a probability measure simply by specifying:

- $P(\{HT, TH\}) = 0.5$
- $P(\{TT\}) = 0.1$

The probabilities for the rest of the event in $\mathcal{F}_{heads}$ can be computed using axiom 3 as follows:

- $P(\{HH\}) = 1 - 0.1 - 0.5 = 0.4$
- $P(\{HT, TH, HH\}) = P(\{HT, TH\}) + P(\{HH\}) = 0.5 + 0.4 = 0.9$

- $P(\{HT, TH, TT\}) = P(\{HT, TH\}) + P(\{TT\}) = 0.5 + 0.1 = 0.6$
- $P(\{HH, TT\}) = 0.1 + 0.4 = 0.5$
- $P(\Omega) = 1$ (Trivial but good to double check in practice)
- $P_{heads}(\emptyset) = 1 - 1 = 0$ (Trivial, always true)

As before the probability assignment maybe frequentist or subjective but regardless of this, it generates is a valid probability measure as it satisfies all three axioms.

Probability measure for power set

Example 1.11. For the largest sigma algebra $\mathcal{F}_{max}$ we can define a probability measure simply by specifying probabilities for the four singletons or atoms:

- $P(\{HH\}) = 0.3$
- $P(\{HT\}) = 0.2$
- $P(\{TH\}) = 0.4$

The probabilities for the rest of the events in $\mathcal{F}_{max}$ can be computed using the axioms as follows:

- $P(\{TT\}) = 1 - 0.3 - 0.2 - 0.4 = 0.1$
- $P(\{HT, TH\}) = 0.2 + 0.4 = 0.6$
- $P(\{HT, HH\}) = 0.2 + 0.3 = 0.5$
- $P(\{HT, TT\}) = 0.2 + 0.1 = 0.3$
- $P(\{TH, HH\}) = 0.4 + 0.3 = 0.7$
- $P(\{TH, TT\}) = 0.4 + 0.1 = 0.5$
- $P(\{HH, TT\}) = 0.3 + 0.1 = 0.4$
- $P(\{HT, TH, HH\}) = 0.2 + 0.4 + 0.3 = 0.9$
- $P(\{HT, TH, TT\}) = 0.2 + 0.4 + 0.1 = 0.7$
- $P(\{HT, HH, TT\}) = 0.2 + 0.3 + 0.1 = 0.6$

- $P(\{TH, HH, TT\}) = 0.4 + 0.3 + 0.1 = 0.8$

As before the probability assignment maybe frequentist or subjective but regardless of this, it generates is a valid probability measure as it satisfies all three axioms.

Simple Probability Operations

Proposition 1.1. *From the axioms, we can derive several useful properties:*

- ***Probability of the Complement**: For any event $A \in \mathcal{F}$,*

$$P(A^c) = 1 - P(A)$$

- ***Probability of the empty set**: $P(\emptyset) = 0$.*
- ***Probability of the Union of Two Events (General)**: For any two events $A, B \in \mathcal{F}$:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

This is known as the addition rule. It accounts for the overlap between events.

Probability of the union

Example 1.12.

$$\begin{aligned} P(\{HT, TH, TT\} \cup \{HH, TT\}) &= P(\{HT, TH, TT\}) + P(\{HH, TT\}) - P(\{TT\}) \\ &= 0.6 + 0.5 - 0.1 \\ &= 1 \end{aligned}$$

clearly this is correct as the union of these two events is Ω.

Boole and Bonferroni inequalities

Theorem 1.1.

- ***Boole's inequality*** *For any events $A_1, A_2, \ldots, A_n$ in $\mathcal{F}$:*

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

This inequality provides an upper bound for the probability of the union of events.

Bonferroni Inequality*: For any events $A_1, A_2, \ldots, A_n$ in $\mathcal{F}$:*

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(A_i^c)$$

This inequality provides a lower bound for the probability of the intersection of events.

These inequalities, specially Bonferroni's will be useful later. Booles inequality can be proved by induction and Bonferroni's inequality follows from Booles inequality and the properties of complements. These facts can be verified by the reader.

1.3 Conditional Probability

Conditional Probability

Definition 1.6. The **conditional probability** of event A given that event B has occurred, denoted $P(A|B)$, is defined as:

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

Example of Conditional Probability

Example 1.13. What is the probability of getting heads on the first coin flip GIVEN that at least one head appears in two flips? This can be expressed as $P(A|B)$ where $A = \{HT, HH\}$ is "head on first flip" and $B = \{HT, TH, HH\}$ is "at least one head appears". We have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{P(\{HT, HH\})}{P(\{HT, TH, HH\})}$$

since $A \subset B$ in this case. We notice a subtlety here. The event $A = \{HT, HH\}$ (head on the first flip) is not a member of the sigma-algebra $\mathcal{F}_{heads}$. So cannot use the probability measure P_{heads} to compute this conditional probability. However, it is a member of the sigma algebra (the power set) $\mathcal{F}_{max}$ so we might need to define probabilities using such sigma algebra as follows:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{HT, HH\})}{P(\{HT, TH, HH\})} = \frac{0.5}{0.9} \approx 0.556$$

On a more practical situation, if A is "a user makes a purchase" and B is "a user clicks on an advertisement", then $P(A|B)$ is the probability that a user makes a purchase GIVEN that they clicked on the advertisement. This is a key metric for evaluating ad campaign effectiveness.

Two very useful consequences of the above are: the law of total probability that combines the notion of partition with that of conditional probability and Bayes rule that allows us to reverse conditional probabilities.

Law of Total Probability

Proposition 1.2. *Let $B_1, B_2, \ldots, B_n$ be a partition of the sample space Ω (i.e., they are disjoint and their union is Ω). Then for any event $A \in \mathcal{F}$:*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Bayes' Rule

Proposition 1.3. *For events A and B where $P(B) > 0$:*

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

If $B_1, \ldots, B_n$ form a partition of Ω, and $P(B_i) > 0$ for all i, then Bayes' Rule can be written using the Law of Total Probability for the denominator:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{i=1}^{n} P(B|B_i)P(B_i)}$$

The proof of this result is staightforward and left to the reader.

Example of Law of Total Probability and Bayes' Rule

Example 1.14 (Medical Testing)**.** Suppose a rare disease affects 1 in 10,000 people. A test for this disease is 99% accurate:

- If a person has the disease, the test correctly identifies it 99% of the time (True Positive).
- If a person does not have the disease, the test correctly identifies it 99% of the time (True Negative).

Let D be the event that a person has the disease, and T^+ be the event that the test is positive. The probabilities we know are:

- $P(D) = \frac{1}{10000} = 0.0001$ (Prevalence)
- $P(T^+|D) = 0.99$ (Sensitivity - True Positive Rate)
- $P(T^-|D^c) = 0.99$ (Specificity - True Negative Rate)

Before we proceed we note the probability specifications above are emprical.
Suppose we want to find $P(D|T^+)$, the probability that a person actually has the disease given a positive test result.
First, we need $P(T^+)$. A positive test can occur in two ways:

- $(D \cap T^+)$ or
- $(D^c \cap T^+)$

e.g. a partition of A. We also have:

- $P(T^+|D^c) = 1 - P(T^-|D^c) = 1 - 0.99 = 0.01$ (False Positive Rate)
- $P(D^c) = 1 - P(D) = 1 - 0.0001 = 0.9999$

Using the law of total probability:

$$\begin{aligned} P(T^+) &= P(T^+ \cap D) + P(T^+ \cap D^c) \\ &= P(T^+|D)P(D) + P(T^+|D^c)P(D^c) \\ &= (0.99)(0.0001) + (0.01)(0.9999) \\ &= 0.000099 + 0.009999 = 0.010098 \end{aligned}$$

Now, using Bayes' Theorem:

$$\begin{aligned} P(D|T^+) &= \frac{P(T^+|D)P(D)}{P(T^+)} \\ &= \frac{(0.99)(0.0001)}{0.010098} \approx 0.0098 \end{aligned}$$

This may look counter-intuitive. Even with a positive test, there's only about a 0.98% (less than 1%) chance the person actually has the disease! In particular it is a rare disease. This highlights the importance of understanding base rates and conditional probabilities in interpreting results.

The Law of Total probability allows us to calculate the probability of an event A by considering the different ways it can occur through the events in a partition.

Customer churn

Example 1.15. Suppose we have three models, M1, M2, and M3, that are used to predict customer churn. Let $P(M1) = 0.5$, $P(M2) = 0.3$, $P(M3) = 0.2$ be the probabilities that each model is the "best" for a given customer. Let A be the event "customer churns". If we know the probability of churn given each best model (e.g., $P(A|M1) = 0.1$, $P(A|M2) = 0.2$, $P(A|M3) = 0.15$), the Law of Total Probability allows us to find the overall probability of churn:

$$\begin{aligned} P(A) &= P(A|M1)P(M1) + P(A|M2)P(M2) + P(A|M3)P(M3) \\ &= (0.1)(0.5) + (0.2)(0.3) + (0.15)(0.2) \\ &= 0.05 + 0.06 + 0.03 = 0.14 \end{aligned}$$

1.4 Independence

1.4.1 Independence of events

First intuitively, two events, A and B, are considered **independent** if the occurrence of one event does not affect the probability of the other event occurring. Formally,

Independent Events

Definition 1.7. Tvents A and B are independent if:

$$P(A \cap B) = P(A) \times P(B)$$

or equivalently, if either of the following conditions hold:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$

This means that knowing that event B has occurred gives us no new information about the probability of event A occurring, and vice versa.

Independent event when flipping a coin twice

Consider the following events when flipping a coin twice:

- $A = \{HT, HH\}$ the first flip is heads
- $B = \{TT, HH\}$ the two flips are the same

Then using the probabilities in Example 1.11 we have:

$$P(A \cap B) = P(\{HH\}) = 0.3 \neq P(\{HT, HH\})P(\{TT, HH\}) = 0.5 \times 0.4 = 0.2$$

Therefore these two events are not independent. Of course, the assignment of probabilities here play a role. In this way, if we had assigned $P(\{HH\}) = 0.2$ then the events would have been independent.

An obvious consequence of Definition 1.6 of conditional probability is the so-called multiplication rule.

Multiplication Rule

Proposition 1.4. *For any two events A and B*

$$P(A \cap B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

Drawing Cards without Replacement

Example 1.16. Imagine drawing two cards from a standard deck without replacement. Let A be the event that the first card is a Heart. $P(A) = \frac{13}{52}$. Let B be the event that the second card is a Heart. Since the first card is not replaced, these events are dependent. The probability of the second card being a Heart *depends* on the first card drawn. $P(B|A)$ (the probability the second card is a Heart, given the first was a Heart) is $\frac{12}{51}$ (as there are 12 Hearts left and 51 total cards). So, the probability of drawing two Hearts in a row is $P(A \cap B) = P(A) \times P(B|A) = \frac{13}{52} \times \frac{12}{51}$.

Note

The outcome of flipping a coin maybe independent of the outcome of any previous coin flips. If you flip a coin and get heads, the probability of getting heads on the next flip should remain as before. Of course, this a simplifying assumption that may not hold in practice. In this course we will make these kind of assumption specially when it involves sequences of events. Not assuming independence for sequences of events may things more complicated for what we want to achieve in this course.

1.5 Random Variables

So far we have talked about events, which are subsets of the sample space. In many applications, especially in data science, we are interested in quantifying outcomes numerically. This is where random variables come into play.

Random Variable

Definition 1.8. A **random variable** X is a function that maps outcomes from the sample space Ω to real numbers. That is, $X : \Omega \to \mathbb{R}$. It quantifies the outcomes of a random phenomenon numerically.

Random variable examples

Example 1.17. If Ω is the set of all possible customer orders, a random variable X could be "the total dollar amount spent in an order". For each order (an outcome in Ω), X assigns a specific monetary value. As another example: for a user's session on a website, X could be "the number of pages visited" or the "overall time spent in the website".

Note a random variable evaluates from the outcomes in the sample space rather than the events defined in the sigma algebra. However, events can be defined in terms of random variables. For example, the event "the total amount spent in an order is greater than 50 dollars" can be expressed as $\{X > 50\}$.

Random variable: Number of equal coin flips

Example 1.18. When we flip a coin twice, the sample space is $\Omega = \{TT, HT, TH, HH\}$. We can define a very simple random variable X as the "number of times the flips are the same". The

Random variable: Number of heads in two coin flips

Example 1.19. When we flip a coin twice, the sample space is $\Omega = \{TT, TH, HT, HH\}$. We can define a random variable X as the "number of heads" in the two flips. The mapping would be:

- $X(\{TT\}) = 0$ (both flips are the same)
- $X(\{TH\}) = 0$ (flips are different)
- $X(\{HT\}) = 1$ (one head)
- $X(\{HH\}) = 2$ (two heads)

The possible values of X are $\{0, 1, 2\}$. This random variable quantifies the outcome of the coin flips in terms of the number of heads observed. Also note the order in which the heads appear does not matter for this random variable.

The above two random variables are discrete random variables as they take on a finite or countable number of values, that is $X(\Omega)$ is finite or countable. There are also continuous random variables that can take on any value in a continuous range.

Continuous vs Discrete Random Variables

Example 1.20. Going back to Example 1.3 we have already defined a random variable X as the time to complete a task in a website with a limit of 5 minutes. If we round to the nearest second, then the possible values of X are $\{0, 1, 2, 3, \dots, 300\}$ and X is a discrete random variable. However, if we do not round then X can take any value in the interval $(0, 300)$ and X is a continuous random variable.

The definition of continuous random variables requires a bit more than simply having an uncountably infinite image set $X(\Omega)$. The definition is a bit thechnical as it involves the notion of probability density function.

Discrete and continuous random variables

Definition 1.9. We say a random variable X is

- **discrete** if it takes on a finite or countably infinite number of distinct values. That is if the image set $X(\Omega)$ is either finite or countably infinite.

The function:

$$f_X(x) = P(X = x) := P(\{\omega : X(\omega) = x\}) \quad \text{for } x \in X(\Omega)$$

is called the **probability mass function (PMF)** of the discrete random variable X. The PMF satisfies:

- $f_X(x) \geq 0$ for all $x \in X(\Omega)$.
- $\sum_{x \in X(\Omega)} f_X(x) = 1$
- **continuous** if there exists a function $f_X(x)$ such that for any two numbers a and b with $a < b$:

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\}) = \int_a^b f_X(x)dx$$

where

- $f_X(x) \geq 0$ for all x and
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

The function $f_X(x)$ is called the **probability density function (PDF)** of the random variable X.

The idea is that there are no "gaps", which would correspond to real numbers which have a finite probability of occurring. Instead, continuous random variables never take an exact prescribed value, that is $P(X = x) = 0$ for all x but there is a positive probability that its value will lie in particular intervals which can be arbitrarily small.

Cumulative Distribution Function (CDF)

Definition 1.10. The **cumulative distribution function (CDF)** of a random variable X, denoted by $F_X(x)$, is the function $F : \mathcal{R} \to [0, 1]$ defined by

$$F_X(x) = P(X \leq x) := P(\{\omega : X(\omega) \leq x\})$$

for any real number x. The CDF gives the probability that the random variable X takes on a value less than or equal to x.

We note that

$$\begin{aligned} P(a \leq X < b) &= F_X(b) - F_X(a) \\ &= \int_a^b f_X(x)dx \end{aligned}$$

so that

$$f_X(x) = \frac{d}{dx} F_X(x) \qquad \forall x \in \mathbb{R}$$

CDF of a discrete random variable

Example 1.21. Consider a discrete random variable X with possible values in the set $\{0, 1, 2, 3\}$. Assume we probability mass function (pmf) is given by:

$$f_X(x) = P(X = x) = \begin{cases} 0.1 & \text{if } x = 0 \\ 0.3 & \text{if } x = 1 \\ 0.4 & \text{if } x = 2 \\ 0.2 & \text{if } x = 3 \\ 0 & \text{otherwise} \end{cases}$$

then the CDF is given by

$$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 0.1 & \text{if } 0 \leq x < 1 \\ 0.4 & \text{if } 1 \leq x < 2 \\ 0.8 & \text{if } 2 \leq x < 3 \\ 1.0 & \text{if } x \geq 3 \end{cases}$$

Properties of Cumulative Distribution Functions

For any random variable X, its CDF $F_X(x)$ has the following properties:

1. **Monotonicity**: $F_X(x)$ is non-decreasing. For any $x_1 < x_2$, $F_X(x_1) \leq F_X(x_2)$.
2. **Limits**: $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$.

3. **Right-continuity**: $F_X(x)$ is right-continuous, meaning $\lim_{h\to 0^+} F_X(x+h) = F_X(x)$ for all x.

The properties above hold for both discrete and continuous random variables. For discrete random variables, the CDF is a step function (continuous from the right), while for continuous random variables, the CDF is a continuous function.

CDF of a continuous random variable

Example 1.22. Consider a random variable X representing the time (in hours) a server remains operational before crashing. X can take any non-negative real value. Assume the probability density function (pdf) is given by:

$$f_X(x) = \begin{cases} \frac{1}{100}e^{-x/100} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

This probability distribution is called an **exponential distribution with a mean of 100 hours**. We will define and talk about mean later. The CDF is computed as follows:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-x/100} & \text{if } x \geq 0 \end{cases}$$

The CDF $F_Y(y) = P(Y \leq y)$ would give the probability that the server operates for at most y hours. For instance, $F_Y(10)$ would be the probability the server fails within the first 10 hours.