

Chapter 4. Probability

The development of probability theory has been underway for hundreds of years but the formal mathematical structure proceeds from a few axioms set out by Andrey Kolmogorov in 1933.

Despite the axiomatic basis, Probability is not a purely mathematical subject. In fact, the “probability function” itself is left undefined by the axioms and requires us to supply an interpretation and a specification. This is as it should be since probability is inextricably linked with the real world of experiments, observation, and reasoning.

Therefore, we will set out the formal structure first, in an abstract form, and then inject it with meaning.

4.1 The formal structure

4.1.1 σ -algebras

First we need an algebraic structure over sets. This means we want a collection of sets that we can “add” and “multiply” together. The corresponding set operations are union, intersection, and taking the complement with respect to a universal set.

Start with a set S and a collection of subsets of S called \mathcal{F} (i.e. $\forall E \in \mathcal{F}, E \subseteq S$).

Definition 4.1.1. \mathcal{F} is called a **σ -algebra** associated with S if:

a) $S \in \mathcal{F}$

b) \mathcal{F} is closed under complements with respect to S :

$$E \in \mathcal{F} \implies \bar{E} \in \mathcal{F} \quad (4.1)$$

c) \mathcal{F} is closed under countable unions:

$$E_1, E_2, \dots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} E_i \in \mathcal{F} \quad (4.2)$$

This definition implies two quick facts.

First, \mathcal{F} must contain the empty set \emptyset :

$\emptyset \in \mathcal{F}$ since \emptyset is the complement of S , which is in \mathcal{F}

Second, \mathcal{F} must be closed under countable intersections, as well as unions. The trick to showing this is to use De Morgan's laws. Specifically, if $E_1, E_2, \dots \in \mathcal{F}$ show that $\bigcap_{i=1}^{\infty} E_i \in \mathcal{F}$.

For $E_1, E_2, \dots \in \mathcal{F}$ the complement of each of the E_i is in \mathcal{F} . But then so is the union of these complements: $\bigcup_{i=1}^{\infty} \overline{E_i} \in \mathcal{F}$. Finally, the complement of this union must be in \mathcal{F} . And we can show that this last set, $\overline{\bigcup_{i=1}^{\infty} \overline{E_i}}$, is exactly the intersection of the E_i .

Bottom line: The sets in \mathcal{F} form an algebra. We can take unions, intersections and complements of members of \mathcal{F} in any combination and the result will always be a member of \mathcal{F} .

4.1.2 Probability measure

Now we can write down Kolmogorov's three simple axioms for probability. We start with a σ -algebra \mathcal{F} associated with the set S . A **probability measure** P is a function that takes each element of \mathcal{F} (i.e. a subset of S) and returns a real number.

Definition 4.1.2 (Kolmogorov's axioms of probability). A **probability measure** P is a function from \mathcal{F} to the real numbers satisfying:

- a) $P(E) \geq 0$ for every $E \in \mathcal{F}$,
- b) $P(S) = 1$,
- c) If $E_1, E_2, \dots \in \mathcal{F}$ are disjoint (i.e. $E_i \cap E_j = \emptyset$ for all $i \neq j$) then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i). \quad (4.3)$$

The triple (S, \mathcal{F}, P) , consisting of a set S , a σ -algebra \mathcal{F} of subsets of S , and a probability measure P , is called a **probability space**.

4.2 Interpretations of the probability space

First, notice that the axioms, from which all else supposedly follows, don't define what probability actually is. On the plus side this means we can come up with several interpretations of probability and, as long as they agree with the axioms, all of our constructions and derivations based on the axioms will hold for any interpretation. That is, we will build up a single set of rules for manipulating expressions involving symbols like $P(AB|C)$ but we can interpret the results in many different contexts. The downside of the axiomatic starting point is that we have to do a little more work supplying interpretations before we can actually put probabilities into practice.

4.3 Interpretation of the σ -algebra

4.3.1 The sample space (S)

The conventional interpretation of probability starts with the idea of a statistical *experiment* — any fixed procedure with a variable or uncertain outcome (such as tossing a coin or rolling a die). The set of possible *outcomes* of an experiment is called the **sample space**. The sample space is going to correspond to the universal set we have been calling S in the probability space.

Examples

- Coin tossing: $S = \{\bullet, \circ\}$. For typographical convenience we usually write $S = \{H, T\}$ or $S = \{0, 1\}$.
- Die rolling: $S = \{\square, \square, \square, \square, \square, \square\}$.
- Tossing a coin twice: $S = \{H, T\}^2 = \{H, T\} \times \{H, T\}$ where ' \times ' represents the cartesian product of sets. Writing out all combinations, $S = \{HH, HT, TH, TT\}$.
- Throwing two dice: $S = \{\square, \square, \square, \square, \square, \square\} \times \{\square, \square, \square, \square, \square, \square\}$. For typographical convenience we write $S = \{1, 2, 3, 4, 5, 6\}^2$.
- Lifetime of a light bulb: $S = \mathbb{R}^+$. What if we only count months?
- If we can tolerate a bit of imprecision, S can be the set of all possible worlds, or all possible states the world might be in. This gives a glimpse of how probability is relevant to essentially any situation.

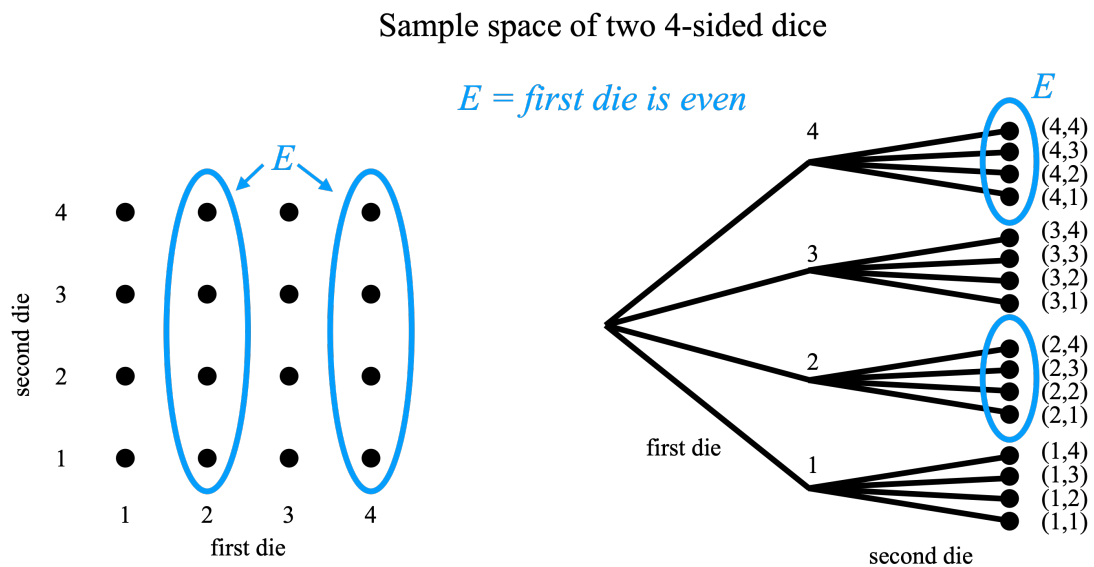
4.3.2 The event space (\mathcal{F})

An **event** E is a subset of the sample space, $E \subseteq S$. An event is a collection of some of the possible outcomes of the experiment. The σ -algebra \mathcal{F} is the set of possible events.

Examples of events

- Single coin toss: $E = \{H\}$, $E = \{T\}$, $E = \{H,T\}$, $E = \emptyset$.
- Die rolling: $E = \{\square\}$, $E = \{\text{Even numbered face}\} = \{\square, \boxtimes, \boxplus\}$.
- Toss 2 coins: $E = \{\text{First coin lands on heads}\} = \{(H,H), (H,T)\}$.

Example Sequential events, e.g. roll two four-sided dice:

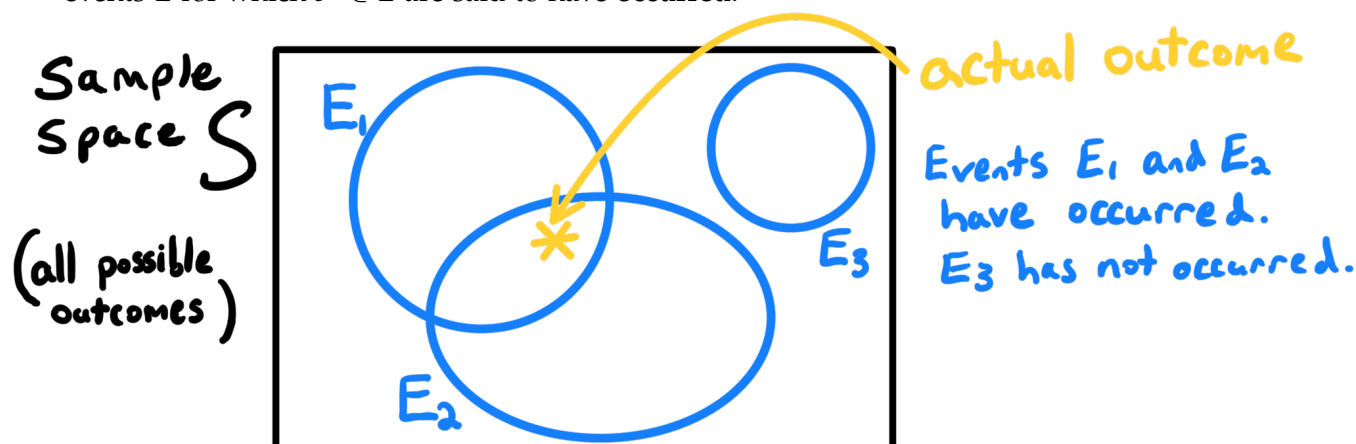


Two ways of representing a “sequential” sample space. Each dot represents a possible outcome. The blue set is the event “the first die is even”. ■

Terminology: **elementary events** are events that consist of a single element of S (e.g. $\{H\}$ or $\{\square\}$).

Why does the event space need to be a σ -algebra?

We run the experiment and we get an outcome s^* , an element of the sample space S . All the events E for which $s^* \in E$ are said to have **occurred**.



Now we see why the event space \mathcal{F} needs to be a σ -algebra: the unions, intersections, and complements of events are also valid events.

- Let's say we have two events E_1 and E_2 . We want to be able to ask "What is the probability that E_1 or E_2 occurred?" That is, what is the probability that the outcome s^* is part of event 1 or event 2? In other words, we are asking about the event $E_1 \cup E_2$. For example, "What's the probability that we rolled an odd number or we rolled a number greater than 3?"
- Similarly, we want to be able to ask about the event, $E_1 \cap E_2$, i.e. about whether both event 1 and event 2 occurred. For example, "What is the probability we rolled an odd number that is also greater than 3?"
- Being able to take a complement of an event E means being able to ask about the probability that the event *does not* occur, i.e. that the outcome $s^* \notin E$, equivalent to $s^* \in \bar{E}$.
- The event space \mathcal{F} has to contain the event S itself (part a of definition 4.1.1). In other words, S (interpreted as an element of \mathcal{F}) is the event "Some outcome occurred." (And jumping ahead we also see the connection to $P(S) = 1$ in the probability axioms: it is guaranteed that we get *some* outcome from the experiment.)
- Why countable unions and not just finite unions (part c of definition 4.1.1)?

Example A coin is tossed repeatedly until the first head comes up. We are concerned with the number of tosses it takes for this to happen. The set of all possible outcomes is the sample space $S = \{1, 2, 3, \dots\}$. The elementary events are $E_i = \{i\}$ for all $i \geq 1$. We may seek to assign a probability to the event E_{even} , that the first head occurs after an even number of tosses, that is $E_{\text{even}} = \{2, 4, 6, \dots\}$. This is an infinite countable union of events, $E_{\text{even}} = \bigcup_{i=1}^{\infty} E_{2i}$, and we require that such a set belongs to \mathcal{F} in order that we can discuss its probability.

Finally, we even get an interpretation of the fact that two events E_1 and E_2 are disjoint (i.e. $E_1 \cap E_2 = \emptyset$). The two events are **mutually exclusive**. There is no way for both events to occur simultaneously. E.g. in rolling a die E_1 is the event that we get an even number and E_2 is the event that we get an odd number. There is no outcome that makes E_1 and E_2 both occur. (Soon we will see that $P(E_1 \cap E_2) = P(\emptyset) = 0$, which reflects this situation.)

Finite or Countable S

If the sample space S is finite or countable we can simply set $\mathcal{F} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}$. For example, if $S = \{1, 2, 3\}$, then we can set the event space to be

$$\mathcal{F} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}.$$

Uncountable S

When S is uncountable, \mathcal{F} is chosen to contain the sets of interest. For example, suppose the outcome of an experiment is a real number between 0 and 1,

$$S = [0, 1] = \{s \mid 0 \leq s \leq 1\}.$$

Then \mathcal{F} can be chosen to contain all sets of the form

$$[a, b], [a, b), (a, b], \text{ and } (a, b), \text{ with } 0 \leq a \leq b \leq 1, \quad (4.4)$$

along with all possible complements, countable unions, and countable intersections of the above intervals.

4.4 Interpretations of the probability measure (P)

For an event $E \subseteq S$, the probability that E occurs will be written as $P(E)$. The axioms do not tell us how to construct the function P . The various *interpretations of probability* guide our understanding of what $P(E)$ means and how to calculate it based on our knowledge of the world.

We will cover three important interpretations: **classical**, **frequentist**, and **subjective**.

4.4.1 Classical interpretation

The **classical** interpretation of probability is based on the assumption that the different outcomes in the sample space S are “equally likely”.

The simplest case is when the sample space is finite. Suppose $S = \{s_1, \dots, s_n\}$, i.e. there are n possible outcomes of the experiment. If we consider the *elementary events* “equally likely” then the probability of an event E is the proportion of all outcomes in S which lie inside E ,

$$P(E) = \frac{|E|}{|S|},$$

(remember that $|A|$ is the cardinality, or size, of the set A).

Example Rolling a die: Elementary events are $\{\square\}, \{\square\}, \dots, \{\boxplus\}$.

- $P(\{\square\}) = P(\{\square\}) = \dots = P(\{\boxplus\}) = \frac{1}{6}$.
- $P(\text{Odd number}) = P(\{\square, \square, \boxplus\}) = \frac{3}{6} = \frac{1}{2}$.

■

Example If we toss a die twice, then S has 36 elements: $S = \{(i, j) \mid i, j \in \{1, \dots, 6\}\}$. If each outcome is equally likely, then $P(E) = |E|/36$ where $|E|$ denotes the number of elements in E .

What is the probability that the sum of the two rolls is 11? The event is the set of outcomes $E = \{(5, 6), (6, 5)\}$. The size of this event is $|E| = 2$ and so the probability that the sum of the two rolls is 11 is $P(E) = 2/36$.

■

It is easy to show that this interpretation of probability $\left[P(E) = \frac{|E|}{|S|}\right]$ obeys the axioms (definition 4.1.2).

1. $P(E) \geq 0$ for all events
True since both $|E|$ and $|S|$ are non-negative.
2. $P(S) = 1$
 $P(S) = |S|/|S| = 1$
3. For disjoint events, $P(E_1 \cup E_2 \cup \dots) = P(E_1) + P(E_2) + \dots$

$$P(E_1 \cup E_2 \cup \dots) = \frac{|E_1 \cup E_2 \cup \dots|}{|S|} = \frac{|E_1| + |E_2| + \dots}{|S|} = \frac{|E_1|}{|S|} + \frac{|E_2|}{|S|} + \dots = P(E_1) + P(E_2) + \dots$$

An equivalent description of the classical interpretation is that the probability of each elementary event is equal to $1/n$, where $n = |S|$ are the number of possible outcomes. Any event E can be thought of as the disjoint union of $|E|$ elementary events, yielding $P(E) = |E|/|S|$.

The “equally likely” (uniform) idea can be extended to infinite spaces, by apportioning probability to sets not by their cardinality but by other standard *measures*, like volume or mass.

Example If a meteorite were to strike Earth, the probability that it will strike land rather than sea would be given by

$$P(\text{Strike land}) = \frac{\text{Total area of land}}{\text{Total area of Earth}}.$$

■

4.4.2 Frequentist interpretation

We have a situation (an experiment) in which the event E may or may not occur. Observation shows that if we repeat the experiment many times the proportion of times in which E occurs tends to some limiting value, called the probability of E . This is the **frequentist** interpretation of probability – relative frequency of an event over many trials.

Example Proportion of heads in tosses of a coin: $H, H, T, H, T, T, H, T, T, \dots \rightarrow \frac{1}{2}$.

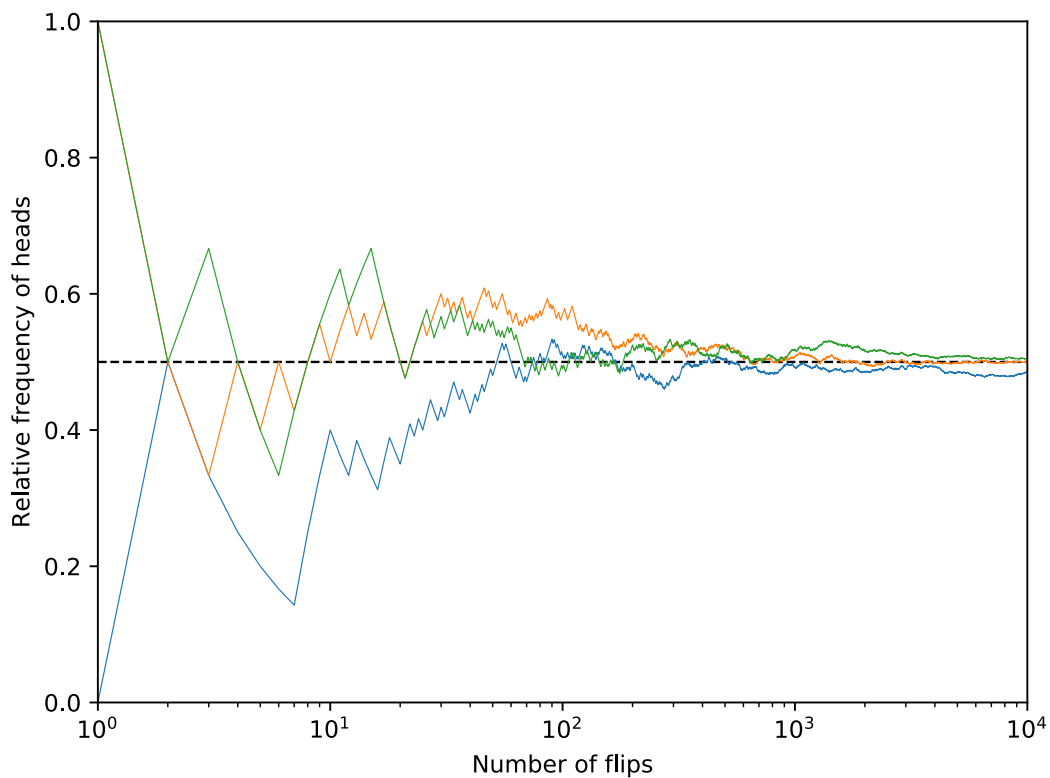


Figure 4.1: Illustration of the frequentist definition of probability. Each colored line corresponds to a simulation where a coin is flipped 10,000 times. The y -axis shows the fraction of flips that are heads as the experiment proceeds. We see that this fraction tends to the limit $1/2$ as the number of flips increases. Therefore, $P(\{H\}) = 1/2$.

■

Let's show that this notion of probability obeys the axioms.

1. $P(E) \geq 0$ for all events

True, since probability is defined as the fraction of trials in which an event occurs is a non-negative number.

2. $P(S) = 1$

The event S occurs in every single trial (since the outcome of every trial is in S). Therefore, the fraction of trials in which S occurs is exactly 1.

3. For disjoint events, $P(E_1 \cup E_2 \cup \dots) = P(E_1) + P(E_2) + \dots$

After n trials, let $k_{1,n}$ be the number of times event E_1 occurred, $k_{2,n}$ be the number of times event E_2 occurred, etc. In the frequentist interpretation, probability of event E_i means $P(E_i) = \lim_{n \rightarrow \infty} k_{i,n}/n$. Now consider the event $E = (E_1 \cup E_2 \cup \dots)$. This event occurs whenever any of the E_i occur. Since the E_i are mutually exclusive, the number of times E has occurred after the n th trial is $k_{1,n} + k_{2,n} + \dots$. Therefore, after n trials the fraction of times E occurs is

$$\frac{k_{1,n} + k_{2,n} + \dots}{n} = \frac{k_{1,n}}{n} + \frac{k_{2,n}}{n} + \dots$$

In the limit $n \rightarrow \infty$ this becomes $P(E_1) + P(E_2) + \dots$

4.4.3 Subjective interpretation

In the **subjective interpretation**, probability is a numerical measure of the **degree of belief** held by an individual.

One way to look at the subjective interpretation is that probability becomes an extension of logic. A logical statement A is just a sentence and the statement A is either true or it's false. In classical logic these are the only two options. But we want to be able to capture a person's sense of uncertainty about the statement. So let's put degree of belief on a sliding scale from 0 to 1. If we say $P(A) = 0$ it means we are absolutely certain that the statement A is false and $P(A) = 1$ means we are absolutely certain it's true. In this view $P(A)$ is a measure of how *plausible* someone thinks A is.

For instance, $A = \text{"After rolling a die it will land on 3."}$ Coming up with a suitable value for $P(A)$ in this case is easy in the classical and frequentist interpretations. But what about a statement like $B = \text{"The mass of the proton is between 0.9 and 1.0 GeV."}$ It is hard to imagine how to think of $P(B)$ in a classical or frequentist sense (what "experiment" is being run over and over with different mass of the proton each time?).

If we are willing to consider subjective degrees of belief then we can use the machinery of probability to compute the values of statements like B above. This use of probability is called **Bayesian** after Thomas Bayes, who first put forward the concept in the 1700s. As we will see, conventional, non-Bayesian statistics also allows us to tackle the problem of quantifying the uncertainty in the proton's mass but not directly in terms of a statement like $P(B) = 0.95$.

Now, degree of belief is an inherently subjective concept and any person might assign any numerical values as their personal degree of belief in any statement. However, in order to use the machinery of probability we will require that our set of subjective probabilities conforms to the Kolmogorov axioms in the following sense¹:

1. $P(A) \geq 0$ for any statement A
2. If the statement A must be true (i.e. we are absolutely certain of it) then $P(A) = 1$
3. For two statements A and B which are mutually exclusive (i.e. we are absolutely certain that they cannot both be true at the same time) our degree of belief in the statement " A or B " obeys $P(A \text{ or } B) = P(A) + P(B)$.

¹In the Bayesian framework the event space \mathcal{F} would be the set of all possible statements. It shares the properties of a σ -algebra in that you can take analogs of unions, intersections, and complements. For instance, the "union" of statements A and B is the logical OR operator (e.g. "The die lands on 3 or it lands on an even number"). The "intersection" is the logical AND ("The first coin flip lands H and the second flip lands T") and the "complement" is the logical NOT (e.g. "The die does not land on an even number."). One way to think about it is that the sample space S is the set of "all possible worlds". For every logical statement A , the event E_A is the subset of possible worlds where the statement A is true. Then the OR, AND, and NOT perfectly correspond to union, intersection, and complement on the collection of E_A 's (e.g. $E_A \cap E_B$ is the subset of worlds in which both A and B hold true, in other words the set of worlds $E_{A \text{ and } B}$). A statement that is always true corresponds to S since it is true in all possible worlds. Probability can be thought of then as a measure over the space of possible worlds: $P(A)$ is our degree of belief that we live in E_A , i.e. in one of the worlds where the statement A is true.

4.5 A few derivations from the axioms

The following facts flow directly from the axiomatic definition of probability. We can prove them completely abstractly using the axioms and set theory. But we can also *interpret* each of these equations in terms of the classical, frequentist, and subjective frameworks.

1. $P(\emptyset) = 0$.

Start with result 3 below for the event $E = S$. Then $\bar{E} = \emptyset$. Use the second probability axiom ($P(S) = 1$) to get $P(\emptyset) = 1 - P(S) = 1 - 1 = 0$.

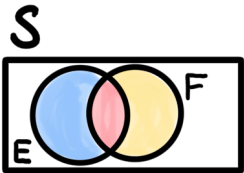
2. $P(E) \leq 1$.

Result 3 says $P(E) + P(\bar{E}) = 1$ for any event E . But axiom 1 tells us the probability of any event is non-negative. Two non-negative numbers that add to 1 implies that both numbers are less than or equal to 1.

3. $P(\bar{E}) = 1 - P(E)$.

For *any* subset E of S , $\{E, \bar{E}\}$ is a partition of S (i.e. E and \bar{E} are disjoint and $S = E \cup \bar{E}$). Then the second and third probability axioms tell us,
 $1 \stackrel{(\text{ax. } 2)}{=} P(S) = P(E \cup \bar{E}) \stackrel{(\text{ax. } 3)}{=} P(E) + P(\bar{E})$.

4. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

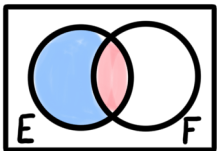


Partition the event E into the disjoint union $E = (E \cap \bar{F}) \cup (E \cap F)$. The third axiom then says $P(E) = P(E \cap \bar{F}) + P(E \cap F)$. Similarly, $P(F) = P(F \cap \bar{E}) + P(E \cap F)$.

Partition $E \cup F$ into the disjoint union $E \cup F = (E \cap \bar{F}) \cup (E \cap F) \cup (F \cap \bar{E})$. The third axiom says $P(E \cup F) = P(E \cap \bar{F}) + P(E \cap F) + P(F \cap \bar{E})$. Add and subtract $P(E \cap F)$ on the righthand side and recognize the righthand side as $P(E) + P(F) - P(E \cap F)$.

S

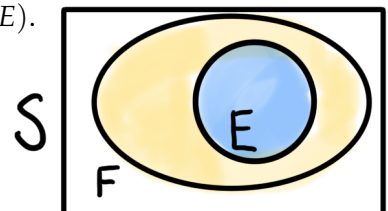
5. $P(E \cap \bar{F}) = P(E) - P(E \cap F)$



Partition E as the disjoint union $E = (E \cap \bar{F}) \cup (E \cap F)$ and use the third axiom to get $P(E) = P(E \cap \bar{F}) + P(E \cap F)$. Rearrange.

6. If $E \subset F$ then $P(E) \leq P(F)$.

Partition F as the disjoint union $F = E \cup (F \cap \bar{E})$. The third axiom says $P(F) = P(E) + P(F \cap \bar{E})$ and the first axiom says the last term is non-negative. Therefore, $P(F)$ must be greater than or equal to $P(E)$.



4.6 Conditional Probability

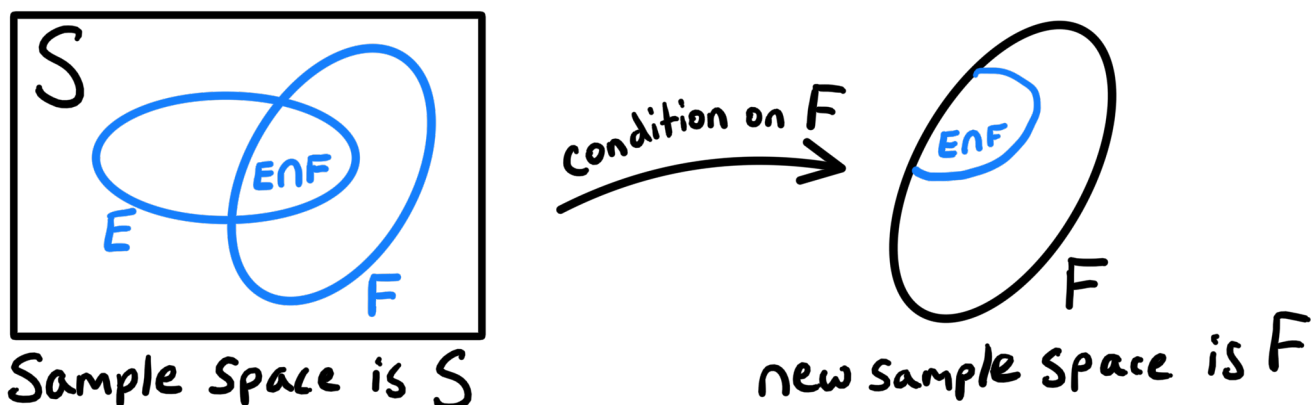
If $P(F) > 0$, we define the conditional probability of E given that F has occurred as follows:

Definition 4.6.1. If $P(F) > 0$ then the **conditional probability** of E given F is

$$P(E | F) = \frac{P(E \cap F)}{P(F)}.$$

$$P(E | F) = \text{“The probability of } E \text{ given } F\text{”}$$

Conditioning means shrinking the sample space from S to F . I.e. the universe of possible outcomes is now the set F rather than S .



Example Suppose a normal die is rolled once.

Questions

- Q1) What is the probability of $E = \{\text{the die shows a } \square\}$?
- Q2) What is the probability of $E = \{\text{the die shows a } \square\}$ given we know $F = \{\text{the die shows an odd number}\}$?

Solutions

S1) $P(E) = \frac{\text{Number of ways a } \square \text{ can come up}}{\text{Total number of possible outcomes}} = \frac{1}{6}.$

S2) Now it is given that the set of possible outcomes is just $F = \{\square, \blacksquare, \boxtimes\}$. So $P(E|F) = \frac{\text{Number of ways a } \square \text{ can come up}}{\text{Total number of possible outcomes}} = \frac{1}{3}.$

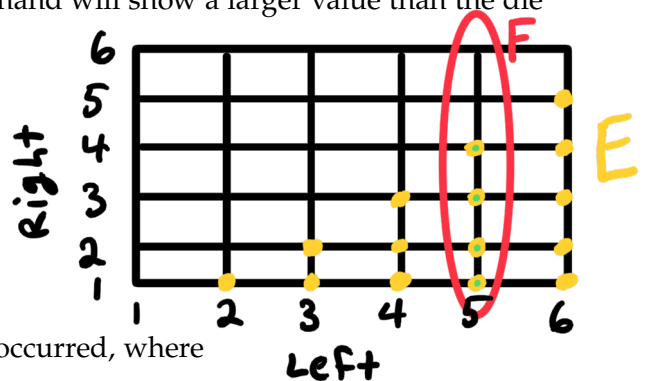
Note $P(F) = \frac{1}{2}$ and $E \cap F = E$, and hence we have $P(E|F) = \frac{P(E \cap F)}{P(F)}.$

Now suppose we roll two normal dice, one from each hand. Then the sample space comprises all of the ordered pairs of dice values

$$S = \{(\square, \square), (\square, \square), \dots, (\blacksquare, \blacksquare)\}.$$

Let E be the event that the die thrown from the left hand will show a larger value than the die thrown from the right hand.

$$P(E) = \frac{\text{\# outcomes with left value} > \text{right}}{\text{total \# outcomes}} = \frac{15}{36}.$$



Suppose we are now informed that an event F has occurred, where

$$F = \{\text{the value of the left hand die is } \blacksquare\}$$

How does this change the probability of E occurring?

Well since F has occurred, the only sample space elements which could have possibly occurred are exactly those elements in $F = \{(\blacksquare, \square), (\blacksquare, \square), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}.$

Similarly the only sample space elements in E that could have occurred now must be in $E \cap F = \{(\blacksquare, \square), (\blacksquare, \square), (\blacksquare, \blacksquare), (\blacksquare, \blacksquare)\}.$

So our revised probability is

$$\frac{\text{\# outcomes with left value } \blacksquare \text{ and left} > \text{right}}{\text{total \# outcomes } (\blacksquare, \cdot)} = \frac{4}{6} = \frac{P(E \cap F)}{P(F)} \equiv P(E|F).$$

■

In both examples, we considered the probability of an event E , and then reconsidered what this probability would be if we were given the knowledge that F had occurred. What we did was replace the sample space S by F , and the event E was replaced by $E \cap F$. I.e. in our new sample space (F) the events are the intersections of the original events $E_i \in \mathcal{F}$ with F (i.e. the events are *restricted* to F). All the rules of probability hold when the sample space is restricted to F .

1. $P(E | F) \geq 0$ for any event E

$P(E | F) = \frac{P(E \cap F)}{P(F)}$. The numerator is greater than or equal to zero and the denominator is greater than zero. Hence the ratio is greater than or equal to 0.

2. $P(F | F) = 1$

$$P(F | F) = \frac{P(F \cap F)}{P(F)} = \frac{P(F)}{P(F)} = 1, \text{ since } F \cap F = F.$$

3. If the events E_1, E_2, \dots are pairwise disjoint² then

$$P(E_1 \cup E_2 \cup \dots | F) = P(E_1 | F) + P(E_2 | F) + \dots$$

For any collection of sets, $\left(\bigcup_i E_i \right) \cap F = \bigcup_i (E_i \cap F)$ (distributive property).

Also, for $i \neq j$, $(E_i \cap F)$ and $(E_j \cap F)$ are disjoint because E_i and E_j are. Therefore,

$$\begin{aligned} P(E_1 \cup E_2 \cup \dots | F) &= \frac{P((E_1 \cup E_2 \cup \dots) \cap F)}{P(F)} = \frac{P((E_1 \cap F) \cup (E_2 \cap F) \cup \dots)}{P(F)} \\ &= \frac{P(E_1 \cap F) + P(E_2 \cap F) + \dots}{P(F)} \\ &= \frac{P(E_1 \cap F)}{P(F)} + \frac{P(E_2 \cap F)}{P(F)} + \dots \\ &= P(E_1 | F) + P(E_2 | F) + \dots \end{aligned}$$

We can think about our original probability space as conditioned on S . The conditional probability formula still makes sense when conditioned on S :

$$P(E|S) = \frac{P(E \cap S)}{P(S)} = P(E) \quad (\text{since } E \cap S = E, \text{ and } P(S) = 1 \text{ by def. of a prob measure.})$$

²Actually, all that is required is the weaker condition that $(E_1 \cap F), (E_2 \cap F), \dots$ are pairwise disjoint

Warning! It is generally the case that $P(E|F) \neq P(F|E)$. People are confused by this all the time. For example, the probability of spots given you have measles is 1, but the probability that you have measles given that you have spots is not 1. In this case, the difference between $P(E|F)$ and $P(F|E)$ is obvious but there are cases where it is less obvious.

Note We will often deal with probabilities of single events, intersections of events and conditional events. To this end we will refer to

- probabilities of the form $P(E | F)$ as **conditional probabilities** — “The probability of E given F ”;
- probabilities of the form $P(E \cap F)$ as **joint probabilities** — “The probability of E and F ”;
- probabilities of the form $P(E)$ as **marginal probabilities** — “The probability of E ”.

Summary of Conditional Probability

1. If $P(F) > 0$ then

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$

2. Read $P(E | F)$ as “The probability of E **given** F .”
3. $P(\cdot|F)$ satisfies the axioms of probability, for fixed F . However, in general, $P(E|\cdot)$ does not satisfy the axioms of probability, for fixed E .
4. In general, $P(E|F) \neq P(F|E)$.

4.7 Independent Events

If we flip a fair coin twice, then the probability of two heads is $\frac{1}{2} \times \frac{1}{2}$. We multiply the probabilities because we regard the two tosses as independent. The definition of independence is as follows:

Definition 4.7.1. Two events E and F are **independent** if and only if

$$P(E \cap F) = P(E)P(F).$$

Extension: The events E_1, \dots, E_k are independent if, for **every** subset of events of size $\ell \leq k$, indexed by $\{i_1, \dots, i_\ell\}$, say,

$$P\left(\bigcap_{j=1}^{\ell} E_{i_j}\right) = \prod_{j=1}^{\ell} P(E_{i_j})$$

Independence can arise in two distinct ways. Sometimes, we explicitly assume that two events are independent. For example, in tossing a coin twice we usually assume the tosses are independent. In other instances, we derive independence by verifying that $P(E \cap F) = P(E)P(F)$ holds true.

Example Toss a fair coin 10 times. Let A be the event {at least one head}. Let T_j be the event {tails occurs on the j th toss}. Then

$$\begin{aligned} P(A) &= 1 - P(\bar{A}) \\ &= 1 - P(\text{all tails}) \\ &= 1 - P(T_1 \cap T_2 \cap \dots \cap T_{10}) \\ &= 1 - P(T_1)P(T_2) \cdots P(T_{10}) \quad \text{by indep.} \\ &= 1 - \left(\frac{1}{2}\right)^{10} \end{aligned}$$

■

Example Suppose that the events E and F are independent. Show that E and \bar{F} are also independent.

$$P(E) = P((E \cap F) \cup (E \cap \bar{F})) = \underbrace{P(E \cap F) + P(E \cap \bar{F})}_{\text{by axiom 3}} = \underbrace{P(E)P(F) + P(E \cap \bar{F})}_{\text{by independence}}.$$

Therefore,

$$P(E \cap \bar{F}) = P(E) - P(E)P(F) = P(E)(1 - P(F)) = P(E)P(\bar{F}).$$

■

Example Suppose that E and F are disjoint events, and that $P(E) > 0$ and $P(F) > 0$. Can the events E and F be independent?

No. As $P(E) > 0$ and $P(F) > 0$ we have that $P(E)P(F) > 0$. However, since $E \cap F = \emptyset$ we also have that $P(E \cap F) = P(\emptyset) = 0$. Since $P(E \cap F) \neq P(E)P(F)$ we conclude that the events are not independent.



Note Conditional probability gives us another way of understanding independence. If E and F are independent, then

$$P(E | F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E),$$

and similarly, $P(E | \bar{F}) = \frac{P(E \cap \bar{F})}{P(\bar{F})} = \frac{P(E)P(\bar{F})}{P(\bar{F})} = P(E).$

I.e. the probability of event E is independent of whether or not F occurs. Similarly, $P(F | E) = P(F)$ and $P(F | \bar{E}) = P(F)$.

We can go the other way too and show if $P(E | F) = P(E)$ then E and F are independent.

Summary of Independence

1. E and F are independent if and only if $P(E \cap F) = P(E)P(F)$.
2. Independence is sometimes assumed and sometimes derived.
3. Disjoint events with positive probability are not independent.
4. E and F are independent if and only if $P(E|F) = P(E)$.

4.7.1 More Examples

Example Which of these two events is more likely?

$E = \{4 \text{ rolls of a die yield at least one } 1\}$; or

$F = \{24 \text{ rolls of two dice yield at least one pair of } (1,1)\}.$

We calculate $P(E)$ and $P(F)$.

1. Each roll of the die is independent from the other rolls, and so there are 6^4 equally likely outcomes. Of these, 5^4 show no 1s.

So the probability of no 1 showing is $\frac{5^4}{6^4} \approx 0.4823$.

So $P(E)$, the probability of at least one 1 showing, is $1 - \frac{5^4}{6^4} \approx 1 - 0.4823 = 0.5177$.

2. There are 36^{24} equally likely outcomes here. Of these, 35^{24} don't show a $\{\text{111}\}$.

So the probability of no $\{\text{111}\}$ is $\frac{35^{24}}{36^{24}} \approx 0.5086$

So $P(F)$, the probability of at least one $\{\text{111}\}$, is $\approx 1 - 0.5086 = 0.4914$

Hence $P(E) \approx 0.5177 > \frac{1}{2} > 0.4914 \approx P(F)$. ■

Example There is a 1% probability for a hard drive to crash. Therefore, we make two backups, each having a 2% probability of crashing, and all three components are independent of each other. The stored information is lost only in the event that all three devices crash. What is the probability that the information is lost?

Start by organizing and labelling the events. Denote

$M = \{\text{main hard drive crashes}\}$

$B_1 = \{\text{first backup crashes}\}$

$B_2 = \{\text{second backup crashes}\}$

In the wording, we are given that M , B_1 , and B_2 are independent and

$$P(M) = 0.01, \quad P(B_1) = 0.02, \quad P(B_2) = 0.02$$

Then, applying rules of complements and intersection for independent events we have

$$\begin{aligned} P(\text{lost}) &= P(M \cap B_1 \cap B_2) \\ &= P(M)P(B_1)P(B_2) \\ &= (0.01)(0.02)(0.02) = 0.000004 = 0.0004\% \end{aligned}$$
■

4.7.2 Conditional Independence

We can combine the concepts of independence with conditional probabilities to define the idea of **conditional independence**.

For three events E_1 , E_2 and F , the pair of events E_1 and E_2 are said to be *conditionally independent given F* if and only if $P(E_1 \cap E_2 | F) = P(E_1 | F)P(E_2 | F)$.³

³This is sometimes written $E_1 \perp E_2 \mid F$.







4.7.3 Joint events

We can think about experiments that are the “product”, or combination of two or more experiments.

For example, consider tossing a coin and rolling a die. We might consider each of the 12 possible combinations of Head/Tail and die value as equally likely.

We can use a **probability table** to specify our probability measure. The probability table shows the probability of all the elementary events when the sample space has a cartesian product structure. In the case of the coin and die experiment:

	b_1	b_2	b_3	\dots
a_1	\ddots			
a_2	$P(\{(a_i, b_j)\})$			
a_3			\ddots	
\vdots				

							
H	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
T	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

(For the left table above, $S = A \times B$, the a_i and b_j are the elements of A and B respectively, and so $\{(a_i, b_j)\}$ is an elementary event of S .)

From this table we can calculate the probability of *any* event we might be interested in, simply by adding up the probabilities of all the elementary events it contains.

For example, the event of getting a head on the coin

$$\{H\} = \{(H, \text{1 dot}), (H, \text{2 dots}), \dots, (H, \text{6 dots})\}$$




has probability

$$\begin{aligned} P(\{H\}) &= P(\{(H, \text{1 dot})\}) + P(\{(H, \text{2 dots})\}) + \dots + P(\{(H, \text{6 dots})\}) \\ &= \frac{1}{12} + \frac{1}{12} + \dots + \frac{1}{12} \\ &= \frac{1}{2}. \end{aligned}$$

Notice the two experiments satisfy our probability definition of independence, since for example

$$P(\{(H, \text{6 dots})\}) = \frac{1}{12} = \frac{1}{2} \times \frac{1}{6} = P(\{H\}) \times P(\{\text{6 dots}\}).$$

A crooked die has the same faces on opposite sides.

Suppose we have two dice, one normal and one crooked. The crooked die has faces numbered , , or .

Now suppose we first flip the coin. If it comes up heads, we roll the normal die; tails, and we roll the crooked one.

We can use the same sample space structure as before. But now the elementary events (T, even roll) never occur so they get probability 0. To calculate the probability of (T, odd roll) easily, we notice that this is equivalent to the previous game using one normal die except with the change that after tails a roll of \square is relabelled as \boxtimes , $\boxtimes \rightarrow \square$, $\boxminus \rightarrow \square$. So we can just merge those probabilities in the tails row.

	\square	\square	\square	\boxtimes	\boxtimes	\boxminus	
H	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{2}$
T	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{6}$	0	$\frac{1}{2}$
	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{12}$	

The probabilities of the different outcomes of the dice change according to the outcome of the coin toss. And note, for example,

$$P(\{(H, \boxminus)\}) = \frac{1}{12}$$

$$\text{but } P(\{H\}) \times P(\{\boxminus\}) = \frac{1}{2} \times \frac{1}{12} = \frac{1}{24},$$

so the two events $\{H\}$ and $\{\boxminus\}$ are no longer independent.

Example A medical test for a disease D has outcomes $+$ and $-$. The probabilities are

	D	\bar{D}
$+$	0.009	0.099
$-$	0.001	0.891

Using the definition of conditional probability, we have

$$P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{0.009}{0.009 + 0.001} = 0.9$$

and

$$P(-|\bar{D}) = \frac{P(- \cap \bar{D})}{P(\bar{D})} = \frac{0.891}{0.891 + 0.099} = 0.9$$

Apparently, the test is fairly accurate. Sick people yield a positive 90% of the time and healthy people yield a negative 90% of the time.

Now suppose you go for a test and get a positive. What is the probability you have the disease? Most people would answer 0.9. The correct answer is

$$P(D|+) = \frac{P(+ \cap D)}{P(+)} = \frac{0.009}{0.009 + 0.099} = 0.08$$

which is much less than 0.9.

The lesson here is that you need to compute the answer numerically. Do not trust your intuition!

Also note that $P(+|D) \neq P(D|+)$



4.8 Bayes's Theorem

The famous Bayes's Theorem tells us how to reverse the positions of the E and F in a conditional probability.

Theorem 4.9 (Bayes's Theorem).

If $P(F) > 0$ and $P(E) > 0$ then we have

$$P(E | F) = \frac{P(F | E) P(E)}{P(F)}. \quad (4.5)$$

Proof. The proof is simple: just write out $P(E \cap F)$ in two different ways using conditional probability,

$$\begin{aligned} P(E \cap F) &= P(E | F) P(F), \\ &= P(F | E) P(E), \end{aligned}$$

and then divide both sides by $P(F)$. □

It is very common to apply Bayes's Theorem to a partition of the sample space, i.e. a collection of events E_1, E_2, \dots where we are guaranteed that exactly one of them occurs. (Recall that E_1, E_2, \dots is a partition of S if $E_i \cap E_j = \emptyset$ for $i \neq j$ and $\bigcup E_i = S$.)

We require an intermediate result which is very important in its own right.

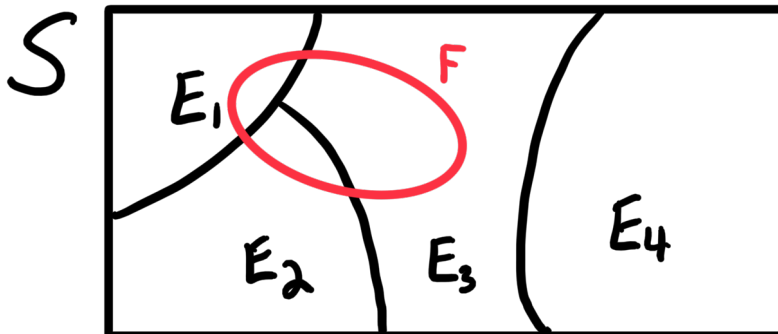
Theorem 4.10 (The Law of Total Probability). *Let E_1, E_2, \dots be a partition of S . Then, for any event $F \subseteq S$, we have*

$$P(F) = \sum_i P(F | E_i) P(E_i)$$

Proof. Define $C_i = F \cap E_i$. Note that C_1, C_2, \dots are disjoint and that $F = \bigcup_i C_i$. Hence,

$$P(F) = P\left(\bigcup_i C_i\right) = \sum_i P(C_i) = \sum_i P(F \cap E_i) = \sum_i P(F | E_i) P(E_i),$$

where the second equality is by the third probability axiom, and the last equality is from the definition of conditional probability. □



A simple use of this theorem is as follows: for any events E and F in S , note that $\{E, \bar{E}\}$ form a partition of S . So by the law of total probability we have

$$\begin{aligned} P(F) &= P(F \cap E) + P(F \cap \bar{E}) \\ &= P(F | E) P(E) + P(F | \bar{E}) P(\bar{E}). \end{aligned}$$

Theorem 4.11 (Bayes's Theorem (applied to a partition)). Let E_1, E_2, \dots be a partition on S such that $P(E_i) > 0$ for each i . If $P(F) > 0$, then for each i we have

$$P(E_i | F) = \frac{P(F | E_i) P(E_i)}{P(F)} = \frac{P(F | E_i) P(E_i)}{\sum_j P(F | E_j) P(E_j)}. \quad (4.6)$$

Proof. The first equality is just Bayes's Theorem (Eq. 4.5) applied to the events E_i and F . The second equality just expands the denominator using the Law of Total Probability. \square

Why is Bayes' Theorem so important? Oftentimes we are really interested in the probability $P(A | B)$ but it is easier to calculate or reason about $P(B | A)$. Recall the previous example about diagnostic testing – a person might be most interested in the probability they have a disease given that they tested positive. But scientific study of the test focuses on establishing the probability that the test is positive given the fact that someone has the disease (called the *sensitivity*) and the probability that the test comes out negative given that someone does not have the disease (called the *specificity*).

Bayes's theorem takes center stage in subjective Bayesian inference. There H is a statement about a hypothesis we are interested in and D is a statement about the data collected by an experiment. It is often conceptually straightforward to calculate $P(D | H)$, the probability of obtaining the data assuming that a hypothesis is true. Bayes' theorem allows one to invert the conditional probability and obtain a degree of belief in the hypothesis given the fact that we obtained certain data in the experiment, $P(H | D)$.

A handwritten diagram illustrating Bayes' Theorem. The equation is written as $P(H_i | D) = \frac{P(D | H_i) P(H_i)}{\sum_i P(D | H_i) P(H_i)}$. The terms are labeled with handwritten text and arrows:

- $P(H_i | D)$ is labeled "posterior" in red, with an arrow pointing to it from the word "hypothesis i" above.
- $P(D | H_i)$ is labeled "likelihood" in blue, with an arrow pointing to it from the word "data" above.
- $P(H_i)$ is labeled "prior" in green, with an arrow pointing to it from the word "prior" to its right.
- $P(D)$ is labeled "evidence" in yellow, with an arrow pointing to it from the word "evidence" below.
- The denominator $\sum_i P(D | H_i) P(H_i)$ is labeled with a yellow arrow pointing to it from the word "evidence" below.

4.12 More Examples

Example A box contains 5000 VLSI chips, 1000 from company X and 4000 from company Y. 10% of the chips made by X are defective and 5% of those made by Y are defective. If a randomly chosen chip is found to be defective, find the probability that it came from company X.

Let C_X = “chip was made by company X”;

let D = “chip is defective”.

First of all, which probabilities have we been given?

The statement

“A box contains 5000 VLSI chips, 1000 from company X and 4000 from Y.”

$$\implies P(C_X) = \frac{1000}{5000} = 0.2, \quad \text{and} \quad P(\overline{C_X}) = \frac{4000}{5000} = 0.8.$$

and

“10% of the chips made by X are defective and 5% of those made by Y are defective.”

$$\implies P(D | C_X) = 10\% = 0.1, \quad \text{and} \quad P(D | \overline{C_X}) = 5\% = 0.05.$$

We have enough information to construct the probability table

	C_X	$\overline{C_X}$	
D	0.02	0.04	0.06
\overline{D}	0.18	0.76	0.94
	0.2	0.8	

The law of total probability has enabled us to extract the marginal probabilities $P(D)$ and $P(\overline{D})$ as 0.06 and 0.94 respectively.

So by Bayes Theorem we can calculate the conditional probabilities. In particular, we want

$$P(C_X | D) = \frac{P(C_X \cap D)}{P(D)} = \frac{0.02}{0.06} = \frac{1}{3}.$$

■

Example Kidney stones are small (< 2cm diam) or large (> 2cm diam). Treatment can succeed or fail. The following data were collected from a sample of 700 patients with kidney stones.

	Success (S)	Failure (\bar{S})	
Large (L)	247	96	343
Small (\bar{L})	315	42	357
Total	562	138	700

For a patient randomly drawn from this sample, what is the probability that the outcome of treatment was successful, given the kidney stones were large?

Clearly we can get the answer directly from the table by ignoring the small stone patients

$$P(S | L) = \frac{247}{343}$$

or we can go the long way around:

$$P(L) = \frac{343}{700}, \quad P(S \cap L) = \frac{247}{700},$$

$$P(S | L) = \frac{P(S \cap L)}{P(L)} = \frac{\frac{247}{700}}{\frac{343}{700}} = \frac{247}{343}.$$

■

Example A multiple choice question has c available choices. Let p be the probability that a student knows the right answer, and $1 - p$ that they do not. When they don't know, they guess an answer at random. Given that the answer the student chooses is correct, what is the probability that the student actually knew the correct answer?

Let A be the event that the question is answered correctly.

Let K be the event that the student knew the correct answer.

Then we are looking for $P(K | A)$.

By Bayes's Theorem

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

and we know $P(A | K) = 1$ and $P(K) = p$, so it remains to find $P(A)$.

By the partition rule (aka the theorem of total probability), we have

$$P(A) = P(A | K)P(K) + P(A | \bar{K})P(\bar{K})$$

and since $P(A | \bar{K}) = \frac{1}{c}$, this gives

$$P(A) = 1 \times p + \frac{1}{c} \times (1 - p).$$

Hence

$$P(K | A) = \frac{p}{p + \frac{1-p}{c}} = \frac{cp}{cp + 1 - p}.$$

Notice that the larger c is, the greater the probability that the student knew the answer, given that they answered correctly. ■

Example Measurements at the North Carolina Super Computing Center (NCSC) on a certain day showed that 15% of the jobs came from Duke, 35% from UNC, and 50% from NC State University. Suppose that the probabilities that each of these jobs is a multitasking job is 0.01, 0.05, and 0.02 respectively.

Questions

- Q1) Find the probabilities that a job chosen at random is a multitasking job.
- Q2) Find the probability that a randomly chosen job comes from UNC, given that it is a multitasking job.

Solutions Let

U_i = "job is from university i ", $i = 1, 2, 3$ for Duke, UNC, NC State respectively; and

M = "job uses multitasking".

- S1) We want to find $P(M)$. Since U_1, U_2, U_3 form a partition we have

$$\begin{aligned} P(M) &= P(M | U_1)P(U_1) + P(M | U_2)P(U_2) + P(M | U_3)P(U_3) \\ &= 0.01 \times 0.15 + 0.05 \times 0.35 + 0.02 \times 0.5 = 0.029. \end{aligned}$$

- S2) We want to find the conditional probability $P(U_2 | M)$.

$$P(U_2 | M) = \frac{P(M | U_2)P(U_2)}{P(M)} = \frac{0.05 \times 0.35}{0.029} = 0.603.$$

■

Example A new covid-19 test is claimed to correctly identify 95% of people who are really covid-positive and 98% of people who are really covid- negative. Is this acceptable?

If only 1 in a 1000 of the population are infected, what is the probability that a randomly selected person who tests positive actually has the disease?

Solution: Let

I = “has a covid infection”; and

T = “test is positive”

We have been given $P(T | I) = 0.95$, $P(\bar{T}|\bar{I}) = 0.98$ and $P(I) = 0.001$. We wish to find $P(I | T)$.

$$\begin{aligned} P(I | T) &= \frac{P(T | I)P(I)}{P(T)} = \frac{P(T | I)P(I)}{P(T|I)P(I) + P(T|\bar{I})P(\bar{I})} \\ &= \frac{0.95 \times 0.001}{0.95 \times 0.001 + 0.02 \times 0.999} \\ &= 0.045 \end{aligned}$$

Therefore, less than 5% of those who test positive really have covid.

If the test shows a positive result, the individual might wish to retake the test. Suppose that the results of a person retaking the covid test are conditionally independent given infection status (clearly two results of the test would certainly not be *unconditionally* independent).

If the test again gives a positive result, what is the probability that the person actually has covid?

Solution: Let T_i = “ i^{th} test is positive”.

$$\begin{aligned} P(I | T_1 \cap T_2) &= \frac{P(T_1 \cap T_2 | I)P(I)}{P(T_1 \cap T_2)} \\ &= \frac{P(T_1 \cap T_2 | I)P(I)}{P(T_1 \cap T_2 | I)P(I) + P(T_1 \cap T_2 | \bar{I})P(\bar{I})} \\ &= \frac{P(T_1 | I)P(T_2 | I)P(I)}{P(T_1 | I)P(T_2 | I)P(I) + P(T_1 | \bar{I})P(T_2 | \bar{I})P(\bar{I})} \end{aligned}$$

by conditional independence.

Since $P(T_i | I) = 0.95$ and $P(T_i | \bar{I}) = 0.02$,

$$P(I | T_1 \cap T_2) = \frac{0.95 \times 0.95 \times 0.001}{0.95 \times 0.95 \times 0.001 + 0.02 \times 0.02 \times 0.999} \approx 0.693.$$

So almost a 70% chance after taking the test twice and both times showing as positive. For three times, this goes up to 99%. Always remember, though, the assumptions that go into

a model (e.g. conditional independence of test results, constant values of sensitivity and specificity that are the same for every person). To the extent the assumptions are wrong, the inferences are unreliable. It is always important to think about ways to test our models – to see if the assumptions hold up. ■

Example *Question:* I divide my emails into 3 categories: $A_1 = \text{“spam”}$, $A_2 = \text{“reply today”}$ and $A_3 = \text{“reply later”}$. From previous experience I find that $P(A_1) = 0.5$, $P(A_2) = 0.1$ and $P(A_3) = 0.4$. Let B be the event that the email contains the word “trial”. From previous experience, I find that $P(B|A_1) = 0.9$, $P(B|A_2) = 0.05$ and $P(B|A_3) = 0.05$. I receive an email with the word “trial”. What is the probability that it is spam?

Solution: Using Bayes’s theorem we have

$$\begin{aligned} P(A_1|B) &= \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)} \\ &= \frac{(0.9)(0.5)}{(0.9)(0.5) + (0.05)(0.1) + (0.05)(0.4)} = \frac{0.45}{0.45 + 0.005 + 0.02} = \frac{0.45}{0.475} = 0.947 \end{aligned}$$

■