Chapter 2

Probability spaces

2.1 From set theory to probability

In this course, we are studying **random phenomena**: things whose outcome cannot be known with certainty. Examples of random phenomenon include:

- the outcome of a coin flip;
- the outcome of a die roll;
- the outcome of a presidential election;
- whether or not a basketball player makes a free throw shot;
- whether or not my soufflé falls in the oven;
- the birth weight of a newborn child;
- the number of costumers that will arrive at a store or restaurant on a given day;
- when and where a hurricane will make landfall;
- the time until an unstable particle will decay;
- the bid-ask spread of Google stock at 2:37 pm ET next Wednesday.

These are all things that, for one reason or another (physical reality, lack of information, imperfect measurement), we do not believe we can perfectly predict. To study random phenomena mathematically, we introduce a definition:

Definition 2.1. A **probability space** is a triple $(S,\ A \subseteq S,\ P)$:

- (**sample space**) A nonempty set S containing all of the possible outcomes of the phenomenon;
- (**events**) Subsets $A \subseteq S$ of the sample space – sets of possible outcomes;

- (**probability measure**) A function P that takes an event A and assigns to it a a real number called $P(A)$. This is the probability that the ultimate outcome of the random phenomenon is an element of the subset A.

Remark 2.1. In mathematics, a **space** is a set together with some "extra structure." A *metric space* is a set together with a function that computes distance between elements. A *topological space* is a set together with a notion of what it means for a subset to be "open." So a probability space is a set (the sample space) together with a function that tell you how to compute probabilities.

2.1.1 Sample spaces

Here are examples of sample spaces:

Phenomenon	Sample space S
Flip two coins in order	$\{HH,\ HT,\ TT,\ TH\}$
Roll a single die	$\{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$
Card dealt from a shuffled deck	$\{2\clubsuit,\ 3\clubsuit,\ 4\clubsuit,\ ...\}$
Winning party in US presidential election	{Republican, Democrat, Libertarian, Green, ... finitely-many other goofy parties that will never win}
A person's blood sodium level in mEq/L	$\mathbb{R}_+ = (0,\ \infty)$
How many insurance claims in a week?	$\mathbb{N}$
Return on a risky asset	$\mathbb{R}$

2.1.2 Events

An event is a set of possible outcomes – a subset of the sample space. We assign probabilities to events. You may have thought that we would be assigning probabilities to individual outcomes, and we are doing that too. You can have an event $A = \{s\}$ containing a single outcome. But we are going further than that.

When talking informally about a "random event," we usually describe it in words. But this qualitative description can always be written as a subset of the sample space.

Description	Event A
"the first of two coin flips is a head"	$\{HH,\ HT\}$
"the die is even"	$\{2,\ 4,\ 6\}$
"dealt a four"	$\{4\clubsuit,\ 4\heartsuit,\ 4\spadesuit,\ 4\diamondsuit\}$
"right-wing party wins"	{Republican, Libertarian, ...}
"blood sodium in healthy range"	$[133,\ 145]$
"over a thousand claims"	$\{1001, 1002, 1003, ...\}$
"your investment loses money"	$(-\infty,\ 0)$

2.1.3 Translating set theory to probability

Since events are *subsets* of a sample space, we will use set theory to work with them. See the slides from 9/3 for elaboration:

Probability	Set theory
A or B occur	$A \cup B$
A and B occur	$A \cap B$
A does not occur	A^c, with S as the reference set
A happening implies B happened too	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \varnothing$ (disjoint)

2.2 Probability axioms

We model random phenomena using the idea of probability space, which is a triple $(S,\ A \subseteq S,\ P)$ (Definition 2.1). In order for the probability measure P to be valid, we require that it satisfy the following axioms:

- (**total measure 1**) $P(S) = 1$ ("something in the sample space is guaranteed to happen");
- (**nonnegativity**) $P(A) \geq 0$ for any event $A \subseteq S$;
- (**countable additivity**) If $A_1,\ A_2,\ A_3,\ ...$ is a countably infinite sequence of pairwise disjoint events (meaning $A_i \cap A_j = \varnothing$ for any distinct $i,\ j$), then

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

Notational Point 2.1. The infinite union notation we just introduced simply means

$$\bigcup_{i=1}^{\infty} A_i = A_1 \cup A_2 \cup A_3 \cup ...$$

It is meant in the exact same spirit as if we had an infinite sequence of numbers a_1, a_2, a_3, ..., and we wrote

$$\sum_{i=1}^{\infty} a_i = a_1 + a_2 + a_3 + ...$$

We will not prove it, but you are free to use the fact that all of our algebraic properties from Section 1.6 extend to the infinite case. So for instance,

$$B \cap \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} (B \cap A_i) \qquad \text{(distributive property)}$$

$$\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c. \qquad \text{(De Morgan)}$$

Feel free to use these things.

Remark 2.2. The axiom of countable additivity is stated in terms of an infinite sequence of pairwise disjoint events, but this implies that it holds for any finite collection $A_1,\ A_2,\ A_3,\ ...,\ A_n$ as well. If you wanted to see that beyond a shadow of a doubt, you do the following. Take your finite collection $A_1, A_2, A_3, \ldots, A_n$ and artificially extend it out so that for all $i > n$, $A_i = \varnothing$. Then you have

$$
\begin{aligned}
\bigcup_{j=1}^{\infty} A_j &= \left(\bigcup_{j=1}^{n} A_j\right) \cup \left(\bigcup_{j=n+1}^{\infty} A_j\right) \\
&= \left(\bigcup_{j=1}^{n} A_j\right) \cup \left(\bigcup_{j=n+1}^{\infty} \varnothing\right) \\
&= \left(\bigcup_{j=1}^{n} A_j\right) \cup \varnothing \\
&= \bigcup_{j=1}^{n} A_j.
\end{aligned}
$$

So you've turned the finite collection into an infinite one, but you haven't actually changed the contents of anything. Now, applying the axiom exactly as stated, you get:

$$
\begin{aligned}
P\left(\bigcup_{j=1}^{n} A_j\right) &= P\left(\bigcup_{j=1}^{\infty} A_j\right) \\
&= \sum_{j=1}^{\infty} P(A_j) \\
&= \sum_{j=1}^{n} P(A_j) + \sum_{j=n+1}^{\infty} P(A_j) \\
&= \sum_{j=1}^{n} P(A_j) + \sum_{j=n+1}^{\infty} P(\varnothing) \\
&= \sum_{j=1}^{n} P(A_j) + \sum_{j=n+1}^{\infty} 0 \\
&= \sum_{j=1}^{n} P(A_j) + 0 \\
&= \sum_{j=1}^{n} P(A_j).
\end{aligned}
$$

Wasn't that fun?

Remark 2.3. Figure 2.1 displays a cartoon of a countably infinite sequence of pairwise disjoint sets. So, an infinite parade of blobs in the plane that do not overlap at all. The axiom of countable additivity says that total probability of all of these sets taken together (their union) is equal to the sum of the individual probabilities. You should note that this is exactly how area works as well. If you wanted the total area covered by all of those blobs, you would just add them up. This is intuition that you can take to the bank. Probability behaves exactly the same way that length, area,

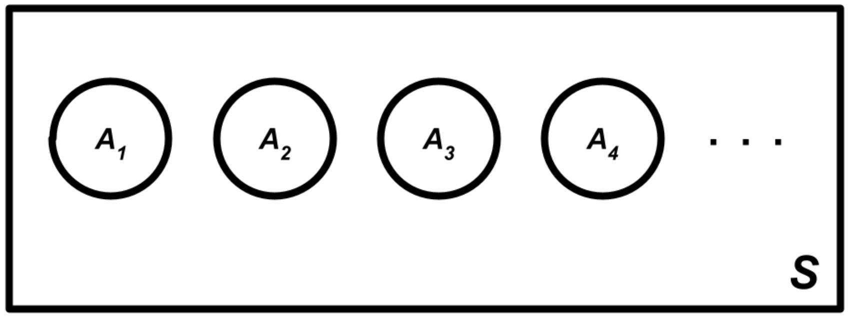

Figure 2.1

and volume behave on the number line, the plane, and three-dimensional space, respectively. This is because these are all special cases of the same abstract mathematical structure. So anything you know about how length/area/volume behave can probably be applied one-for-one with probability, and you will get the right answer.

Remark 2.4. In some sense, probability theory is “just” a branch of mathematical analysis. In particular, it is a branch of measure theory, and probability spaces are special cases of measure spaces (as are Euclidean spaces with their measures of length, area, volume, etc). If you are interested in really going deep on the mathematics of probability (something you will have to do if you are considering graduate work in statistics), I suggest you take courses in real analysis.

2.3 Basic probability rules

We can deduce from the probability axioms several basic "rules" that will be fundamental going forward. To succeed in the course, you need to become fluent in the use of both the axioms and these basic rules.

2.3.1 Complement rule

The first such rule is the complement rule which says that for any event $A \subseteq S$, its probability can be rewritten as $P(A) = 1 - P(A^c)$. This will come in handy because quite often $P(A^c)$ is easier to compute than $P(A)$, and so the complement rule gives us permission to do the easier thing and then apply a simple formula to arrive at the probability we actually care about.

Proof. Let $A \subseteq S$. As we saw in Section 1.7.3, $A \cup A^c = S$, and since an element cannot be both in A and not in A at the same time, it is a disjoint union: $A \cap A^c = \varnothing$. Because $A \cup A^c$ is equal to S, their probabilities are also equal, and so we have

$$\begin{aligned}
P(A \cup A^c) &= P(S) && (A \cup A^c = S) \\
P(A \cup A^c) &= 1 && \text{(total measure one)} \\
P(A) + P(A^c) &= 1 && \text{(countable additivity)} \\
P(A) &= 1 - P(A^c).
\end{aligned}$$

Bada bing. □

With this result, we can perform a cute sanity check:

Corollary 2.1. $P(\varnothing) = 0$.

Proof. Note that

$$\begin{aligned}
\varnothing^c &= \{x \in S : x \notin \varnothing\} = S \\
S^c &= \{x \in S : x \notin S\} = \varnothing.
\end{aligned}$$

So

$$\begin{aligned}
P(\varnothing) &= P(S^c) \\
&= 1 - P(S) && \text{(complement rule)} \\
&= 1 - 1 && \text{(total measure 1)} \\
&= 0.
\end{aligned}$$

□

Riveting.

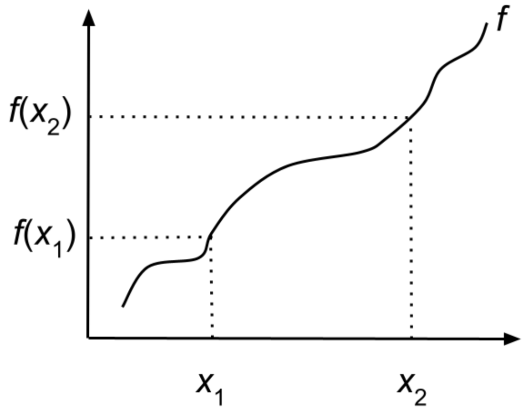

Figure 2.2: f is a monotone increasing function, so if the inputs are ordered $x_1 \leq x_2$, then the outputs have the same ordering: $f(x_1) \leq f(x_2)$. So we say f is an *order-preserving* map. The order of the original x-values was preserved after the function f was applied to them.

2.3.2 Monotonicity

Theorem 2.2. If $A \subseteq B$, then $P(A) \leq P(B)$.

"Monotonicity" means that the probability measure P is an order-preserving function on sets: if A is a "smaller" set than B in the sense that $A \subseteq B$, then correspondingly $P(A) \leq P(B)$. So the order is preserved after the P function is applied. This is similar to a monotone functions of real numbers like the one displayed in Figure 2.2.

Proof. To get the job done, we need to express B in terms of A. Recalling Theorem 1.5 and Figure 1.6, we know that $B = A \cup (B \cap A^c)$ and that this union is disjoint. As such:

$$\begin{aligned}
P(A \cup (B \cap A^c)) &= P(B) \\
P(A) + P(B \cap A^c) &= P(B) && \text{(countable additivity)} \\
P(A) &= P(B) - \underbrace{P(B \cap A^c)}_{\geq 0} \\
P(A) &\leq P(B). && \text{(axiom of nonnegativity)}
\end{aligned}$$

We showed that $P(A)$ was equal to $P(B)$ minus a quantity that is necessarily nonnegative by our second axiom. So if $P(B \cap A^c)$ is exactly zero, $P(A) = P(B)$. If it is positive, $P(A) < P(B)$. Those are the only options, so either way, we have the result. □

Just like the complement rule, this result has momentous consequences:

Corollary 2.3. For any $A \subseteq S$, $0 \leq P(A) \leq 1$.

Proof. Let $A \subseteq S$. Our axioms tell us already that $0 \leq P(A)$ and $P(S) = 1$. Monotonicity gives us that $P(A) \leq P(S)$, and so the transitivity of $\leq$ gives that $0 \leq P(A) \leq 1$. □

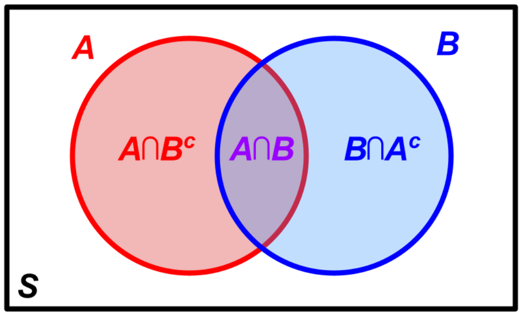

Figure 2.3

We surely already believed that probability was a number between zero and one, but this was not a property we explicitly imposed with an axiom. Instead, we assumed nonnegativity and total measure 1, and deduced this property as a consequence.

2.3.3 Law of inclusion/exclusion

Theorem 2.4. If $A, B \subseteq S$, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

This law is about the probability of a union of two events. We already have an axiom (countable additivity) addressing this in the special case of two *disjoint* events, but not all events are disjoint, and so this rule completes the story by telling us what to do if there is overlap.

Proof. In all of our rules so far, the strategy has been to rewrite the set of interest as a disjoint union, apply countable additivity, and then do some algebra. So it shall be here. Inspecting Figure 2.3, we see that the sets A, B, and $A \cup B$ can all be rewritten as disjoint unions:

$$\begin{aligned} A &= (A \cap B) \cup (A \cap B^c) \\ B &= (A \cap B) \cup (B \cap A^c) \\ A \cup B &= (A \cap B^c) \cup (A \cap B) \cup (B \cap A^c). \end{aligned}$$

As such, we can apply countable additivity to each of these:

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) && (2.1) \\ P(B) &= P(A \cap B) + P(B \cap A^c) && (2.2) \\ P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(B \cap A^c). && (2.3) \end{aligned}$$

Rearranging (2.1) and (2.2) gives

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ P(B \cap A^c) &= P(B) - P(A \cap B), \end{aligned}$$

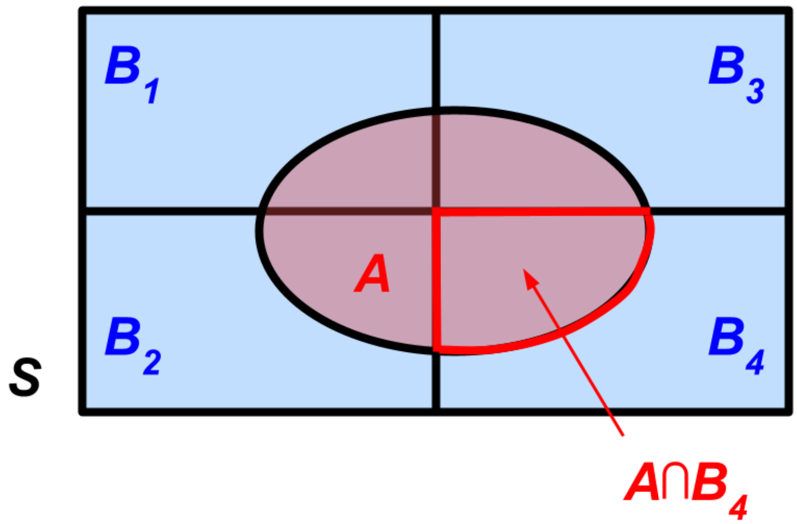

Figure 2.4

and plugging these expressions into (2.3) gives

$$\begin{aligned} P(A \cup B) &= P(A \cap B^c) + P(A \cap B) + P(B \cap A^c) \\ &= P(A) - \cancel{P(A \cap B)} + \cancel{P(A \cap B)} + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

□

The main idea of this proof is to avoid double counting the contribution of $A \cap B$ to the overall probability. The reason $P(A \cup B) = P(A) + P(B)$ does not remain the formula in the case of overlap is that $P(A)$ and $P(B)$ both "include" $P(A \cap B)$, and so the simple sum would incorrectly count it twice. To correct for this, we subtract it off once.

2.3.4 Law of total probability

Theorem 2.5. Let $B_1,\ B_2,\ B_3,\ ... \subseteq S$ be a partition of S, meaning that $\cup_{i=1}^{\infty} B_i = S$ and $B_i \cap B_j = \varnothing$ for all distinct i, j. Then for any $A \subseteq S$,

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i).$$

An alternative name for this could be the "divide-and-conquer" law. It says that, in order to compute the "total" probability of A, we have the option to compute the separate probabilities of the parts of A that overlap with the components of the partition. Once we have these, we just add them up to get the overall $P(A)$. As with the complement rule, the reason we care about this is because, sometimes, computing the separate probabilities is easier. Furthermore, recalling Remark 2.3, Figure 2.4 displays a cartoon demonstrating that the law of total probability is perfectly analogous with a "law of total area" (or length, or volume, etc): if you partition the plane, the total area of a blob in the plane can be alternatively expressed as the sum of the smaller areas in each component of the partition. All we're saying is that probability behaves exactly the same way.

Proof. Let $B_1, B_2, ... \subseteq S$ partition S, and let $A \subseteq S$. We can rewrite A as a disjoint union:

$$\begin{aligned}
A &= A \cap S \\
&= A \cap \left(\bigcup_{i=1}^{\infty} B_i\right) \\
&= A \cap \left(B_1 \cup B_2 \cup B_3 \cup \cdots\right) \\
&= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \cup \cdots \quad \text{(distributive property)} \\
&= \bigcup_{i=1}^{\infty} (A \cap B_i).
\end{aligned}$$

Because the B_i are pairwise disjoint, so are the $A \cap B_i$ (show this!), and so the result follows from countable additivity:

$$P(A) = P\left(\bigcup_{i=1}^{\infty} (A \cap B_i)\right) = \sum_{i=1}^{\infty} P(A \cap B_i).$$

$\square$

Remark 2.5. Be careful that you don't mix up operations on sets and operations on real numbers. $\cap$, $\cup$, and ^{c} are things you do to sets. $+$, $\times$, $-$, and $\div$ are things you do to real numbers like probabilities. So $P(A) \cup P(B)$ doesn't mean anything. And $A + B$ doesn't necessarily mean anything.