Probability

Adam Kelly

February 8, 2021

This set of notes is a work-in-progress account of the course ‘Probability’, originally lectured by Dr Perla Sousi in Lent 2020 at Cambridge. These notes are not a transcription of the lectures, but they do roughly follow what was lectured (in content and in structure).

These notes are my own view of what was taught, and should be somewhat of a superset of what was actually taught. I frequently provide different explanations, proofs, examples, and so on in areas where I feel they are helpful. Because of this, this work is likely to contain errors, which you may assume are my own. If you spot any or have any other feedback, I can be contacted at ak2316@cam.ac.uk.

Contents

1 Basic Concepts

Most of the phenomena in everyday lives involve randomness. What we try to do in probability is model this randomness in a mathematical way. It's likely that you have studied some probability before, but the difference in the treatment here is that we will try to be somewhat more rigerous.

We will define the notion of a probability space, where 'our experiments take place'. Then we will discuss discrete and continuous random variables. In the discrete setting, we will find that there is no real subtleties, and we can be quite rigorous. In the continuous setting however we will have to take some things for granted (but rigour will return in the Part II course).

> *"Probability theory has a right and a left hand. On the right is the rigorous foundational work using the tools of measure theory. The left hand 'thinks probabilistically,' reduces problems to gambling situations, coin-tossing, motions of a physical particle."*

In this course, we will need both hands.

§1.1 Probability Space

Probability is the mathematical formulation of randomness. So in order to study random phenomena in a rigorous way, we first need to set out a rigorous mathematical framework.

The first notion that we will define is that of a *probability space*.

Definition 1.1.1 (σ-Algebra)

Suppose Ω is a set and $\mathcal{F}$ is a collection of subsets of Ω. We call $\mathcal{F}$ a **σ-algebra** if the following properties are satisfied.

(i) $\Omega \in \mathcal{F}$.

(ii) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, the compliment of A.

(iii) For any countable collection $A_1, A_2, \ldots$ with $A_i \in \mathcal{F}$ for all i, we must also have that $\bigcup_{i \geq 1} A_i \in \mathcal{F}$.

Definition 1.1.2 (Probability Measure)

Suppose that $\mathcal{F}$ is a σ-algebra on Ω. Then a function $\mathbb{P} : \mathcal{F} \to [0,1]$ is called a **probability measure** if the following are true.

(i) $\mathbb{P}(\Omega) = 1$.

(ii) For any countable disjoint collection $A_1, A_2, \ldots$ with $A_i \in \mathcal{F}$ for all i, we have

$$\mathbb{P}\left(\bigcup_{n \geq 1} A_n\right) = \sum_{n \geq 1} \mathbb{P}(A_n).$$

We say that $\mathbb{P}(A)$ is the **probability** of A.

Definition 1.1.3 (Probability Space)

If Ω is a σ-algebra on Ω and $\mathbb{P}$ is a probability measure, then $(\Omega, \mathcal{F}, \mathbb{P})$ is a **probability space**.

When the set Ω is countable, we take $\mathcal{F}$ to be all subsets of Ω.

Definition 1.1.4 (Outcomes and Events)

The elements of Ω are called **outcomes**, and the elements of $\mathcal{F}$ are called **events**.

Note that $\mathbb{P}$ is defined on $\mathcal{F}$, so it is defined on the *events*, not the *outcomes*.

Let's look at some properties of the probability measure (which follow immediately from the definition).

Proposition 1.1.5 (Properties of $\mathbb{P}$)

If $\mathbb{P}$ is a probability measure then

- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(\emptyset) = 0$
- If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
- $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

Proof Sketch. Check definitions. □

Example 1.1.6 (Examples of Probability Spaces)

Some examples of probability spaces are given below.

- *Rolling a fair die.* Consider rolling a fair die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $\mathcal{F}$ is all subsets of Ω. Then $\mathbb{P}(\{w\}) = \frac{1}{6}$ for all $\omega \in \Omega$ and if $A \subseteq \Omega$, then $\mathbb{P}(A) = \frac{|A|}{6}$.
- *Equally likely outcomes.* Let Ω be a finite set, $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, and $\mathcal{F}$ be all subsets of Ω. Then define $\mathbb{P} : \mathcal{F} \to [0, 1]$ by $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$. In classical probability, this models picking a random element of Ω. Note that $\mathbb{P}(\{\omega_i\}) = \frac{1}{|\Omega|}$ for all $\omega_i \in \Omega$.
- *Picking balls from a bag.* Suppose we have n balls with n labels from $\{1, \dots, n\}$ that are all indistinguishable by touch. Picking $k \leq n$ balls at random[a] without replacement. Then we take $\Omega = \{A \subseteq \{1, 2, \dots, n\} \mid |A| = k\}$, and $\mathcal{F}$ be all subsets of Ω. Then $|\Omega| = \frac{n}{k}$, and for $\omega \in \Omega$, $\mathbb{P}(\{\omega\}) = \frac{1}{|\Omega|}$.
- *Deck of cards.* Take a well-shuffled deck of 52 cards. Then let Ω be the set of all permutations of the cards, and note $|\Omega| = 52!$. Then we have $\mathbb{P}(\text{top 2 cards are aces}) = \frac{4 \cdot 3 \times 50!}{52!} = \frac{1}{221}$.

- *Largest digit.* Consider a string of n random digits from $0, \dots, 9$. Then $\Omega = \{0, 1, \dots, 9\}^n$, and $|\Omega| = 10^n$. Now define $A_k = \{\text{no digit exceeds } l\}$ and $B_k = \{\text{largest digit is } k\}$. Then $\mathbb{P}(B_k) = \frac{|B_k|}{|\Omega|}$. Notice that $B_k = A_k \backslash A_{k-1}$, and $|A_k| = (k+1)^n$, so $|B_k| = (k+1)^n - k^n$, thus $\mathbb{P}(B_k) = \frac{(k+1)^n - k^n}{10^n}$.
- *Birthday Problem.* There are n people. What is the probability that at least two of them share the same birthday? We can assume nobody is born on 29/02, and that each birthday is equally likely. So $\Omega = \{1, \dots, 365\}^n$, and $\mathcal{F}$ is all subsets of Ω. As we assumed all outcomes are equally likely, we take $\mathbb{P}(\{\omega\}) = \frac{1}{365^n}$ with $\omega \in \Omega$. Letting $A = \{\text{at least 2 people share a birthday}\}$, then $A^c = \{\text{all } n \text{ birthdays are different}\}$, and since $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$, it suffices to calculate $\mathbb{P}(A^c)$. Now $\mathbb{P}(A^c) = \frac{|A^c|}{|\Omega|} = \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n}$, and hence $\mathbb{P}(A) = 1 - \frac{365 \times 364 \times \cdots \times (365-n+1)}{365^n}$. If $n = 23$, then this probability is approximately 0.507.

[a] That is, with all outcomes equally likely.

§1.2 Combinatorial Analysis

Suppose we have some finite set Ω, and that $|\Omega| = n$. We want to partition Ω into k disjoint subsets $\Omega_1, \Omega_2, \dots, \Omega_k$ with $|\Omega_i| = n_i$ and $\sum_{i=1}^k n_i = n$. How many ways is there to do this?

If M is the number of ways, then

$$M = \binom{n}{n_1}\binom{n-n_1}{n_2} \cdots \binom{n-(n_1+\cdots+n_{k-1})}{n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.$$

Definition 1.2.1 (Multinomial Coefficient)

We define the **multinomial coefficient** $\binom{n}{n_1, \dots, n_k}$ to be the number of ways of partitioning a set with n elements into k subsets of size $n_1, n_2, \dots, n_k$. We have

$$\binom{n}{n_1, \dots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.$$

Now let's think about the following question: How many strictly increasing and increasing functions are there between two sets?

If we have $f : \{1, \dots, k\} \to \{1, \dots, n\}$, we say it's *strictly increasing* if whenever $x < y$, then $f(x) < f(y)$. We say that it's *increasing* if $x < y$ implies $f(x) \leq f(y)$.

Any such function is uniquely determined by its range which is a subset of $\{1, \dots, n\}$ of size k. There are $\binom{n}{k}$ such subsets, and hence $\binom{n}{k}$ strictly increasing functions.

We define a bijection from $\{f : \{1, \dots, k\} \to \{1, \dots, n\} \mid f \text{ increasing}\}$ to $\{g : \{1, \dots, k\} \to \{1, \dots, n+k-1\} \mid f \text{ strictly increasing}\}$. For each f, we define $g(i) = g(i) + i - 1$. Then g is strictly increasing and takes values in $\{1, \dots, n+k-1\}$. Then g is strictly increasing and takes values in $\{1, \dots, n+k-1\}$.

So the total number of increasing functions $f : \{1, \dots, k\} \to \{1, \dots, n\}$ is $\binom{n+k-1}{k}$.

§1.3 Stirling's Formula

Frequently in probability it will help to have some bounds/an asymptotic expression for the factorial.

Notation. Let (a_n) and (b_n) be two sequences. We will write $a_n \sim b_n$ if $\frac{a_n}{b_n} \to 1$ as $n \to \infty$.

We will first prove a weak approximation of the factorial.

Proposition 1.3.1

$\log(n!) \sim n \log n$ as $n \to \infty$.

Proof. For $x \in \mathbb{R}$, we write $\lfloor x \rfloor$ for the *integer part* of x. Then we have $\log\lfloor x \rfloor \leq \log x \leq \log\lfloor x+1 \rfloor$. Integrating this from 1 to n, we get

$$\begin{aligned} \sum_{k=1}^{n-1} \log k \leq \int_1^n \log x \, \mathrm{d}x \leq \sum_{k=1}^{n} \log k \\ \implies \log(n-1)! \leq n \log n - n + 1 \leq \log n! \end{aligned}$$

Using this, we get that

$$n \log n - n + 1 \leq \log n! \leq (n+1)\log(n+1) - (n+1) + 1,$$

and dividing through by $n \log n$ we get

$$\frac{\log n!}{n \log n} \to 1 \qquad n \to \infty.$$

□

Now let's prove Stirling's formula. Note that the proof is non-examinable.

Theorem 1.3.2 (Stirling Approximation)

$n! \sim n^n \sqrt{2\pi n}\, e^{-n}$ as $n \to \infty$.

Proof (Non-Examinable). For any function f that is twice differentiable, and $a < b$, then

$$\int_a^b f(x) \, \mathrm{d}x = \frac{f(a) + f(b)}{2}(b-a) - \frac{1}{2}\int_a^b (x-a)(b-x) f''(x) \, \mathrm{d}x.$$

This follows by integrating by parts.

Now take $f(x) = \log x$, and $a = k$, $b = k+1$. Then substituting into the formula, we get

$$\begin{aligned} \int_k^{k+1} \log x \, \mathrm{d}x &= \frac{\log k + \log(k+1)}{2} - \frac{1}{2}\int_k^{k+1} \frac{(x-k)(k+1-x)}{x^2} \, \mathrm{d}x \\ &= \frac{\log k + \log(k+1)}{2} + \frac{1}{2}\int_0^1 \frac{x(1-x)}{(x+k)^2} \, \mathrm{d}x. \end{aligned}$$

Then taking the sum for $k = 1, \dots, n-1$ of the equality above, we get

$$\begin{aligned}\int_1^n \log x \,\mathrm{d}x &= \frac{\log(n-1)! + \log n!}{2} + \frac{1}{2}\sum_{k=1}^{n-1}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x \\ \implies \log n - n + 1 &= \log(n!) - \frac{\log n}{2} + \sum_{k=1}^{n-1} a_k,\end{aligned}$$

where we set

$$a_k = \frac{1}{2}\int_0^1 \frac{x(1-x)}{(x+k)^2}\,\mathrm{d}x.$$

But then

$$\begin{aligned}\log n! &= n\log n - n + \frac{\log n}{2} + 1 - \sum_{k=1}^{n-1} a_k \\ \implies n! &= n^n e^{-n} \cdot \sqrt{n}\exp\left(1 - \sum_{k=1}^{n-1} a_k\right).\end{aligned}$$

Note that $a_k \leqslant \frac{1}{2}\int_0^1 \frac{x(1-x)}{k^2} dx = \frac{1}{12k^2}$, so $\sum a_k < \infty$. We set $A = \exp\left(1 - \sum_{k=1}^{\infty} a_k\right)$. Then

$$n! = n^n \cdot e^{-n}\sqrt{n} \cdot A \cdot \exp\left(\sum_{k=n}^{\infty} a_k\right),$$

and as $\exp\left(\sum_{k=n}^{\infty} a_k\right) \to 1$ (since the argument goes to 0), we have proved that

$$\frac{n!}{n^n e^{-n}\sqrt{n}} \to A, \qquad \text{as } n \to \infty,$$

which means that $n! \sim n^n e^{-n}\sqrt{n} \cdot A$ as $n \to \infty$.

To finish the proof, we need to show that $A = \sqrt{2\pi}$. Knowing that $n! \sim n^n e^{-n}\sqrt{n} \cdot A$ as $n \to \infty$, we have

$$2^{-2n} \cdot \binom{2n}{n} = 2^{-2n}\frac{(2n)!}{n! \cdot n!} \sim \frac{2^{-2n} \cdot (2n)^{2n} \cdot \sqrt{2n} \cdot A \cdot e^{-2n}}{n^n \cdot e^{-n} \cdot \sqrt{n} \cdot A \cdot n^n \cdot e^{-n} \cdot \sqrt{n} \cdot A} = \frac{\sqrt{2}}{A\sqrt{n}}.$$

Using a different method we will prove that

$$2^{2n}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}},$$

which will force $A = \sqrt{2\pi}$.

Consider the integral

$$I_n = \int_0^{2\pi} (\cos\theta)^n \,\mathrm{d}\theta, \qquad n \geq 0.$$

So $I_0 = \pi/2$ and $I_1 = 1$. Then integrating by parts, we get $I_n = \frac{n-1}{n} T_{n-2}$, and thus

$$I_{2n} = \frac{2n-1}{2n} \cdot I_{2n-2} = \frac{(2n-1)(2n-3)\cdots 3 \cdot 1}{2n \cdot (2n-2) \cdots 2} I_0 = \frac{(2n)!}{2^{2n} n! \cdot n!} \cdot \frac{\pi}{2},$$

so

$$I_{2n} = 2^{-2n} \binom{2n}{n} \frac{\pi}{2}.$$

In the same way we get

$$I_{2n+1} = \frac{2n \cdots 4 \cdot 2}{(2n+1) \cdots 3 \cdot 1} I_1 = \frac{1}{2n+1} \left(2^{-2n} \binom{2n}{n} \right)^{-1}.$$

From $I_n = \frac{n-1}{n} I_{n-2}$ we get $\frac{I_n}{I_{n-2}} \to 1$ as $n \to \infty$, and what we want is $\frac{I_{2n}}{I_{2n+1}} \to 1$ as $n \to \infty$.

Note that I_n is a decreasing function of n, therefore

$$\frac{I_{2n}}{I_{2n+1}} \le \frac{I_{2n-1}}{I_{2n+1}} \to 1,$$

and also

$$\frac{I_{2n}}{I_{2n+1}} \ge \frac{I_{2n}}{I_{2n-2}} \to 1,$$

thus $\frac{I_{2n}}{I_{2n+1}} \to 1$ as $n \to \infty$, which means

$$\frac{2^{-2n}\binom{2n}{n}\frac{\pi}{2}}{\left(2^{-2n}\binom{2n}{n}\right)^{-1} \frac{1}{2n+1}} \to 1$$
$$\implies \left(2^{-2n} \binom{2n}{n} \right)^2 \frac{\pi}{2} (2n+1) \to 1,$$

thus

$$\left(2^{-2n} \binom{2n}{n} \right)^2 \sim \frac{2}{\pi(2n+1)} \sim \frac{1}{\pi n},$$

thus $A = \sqrt{2\pi}$, which completes the proof. □

2 Properties of Probability Measures

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Recall that the probability measure $\mathbb{P}$ is a function

$$\mathbb{P} : \mathcal{F} \to [0,1]$$

with $\mathbb{P}(\Omega) = 1$ which is *countable additive*, that is, for any countable disjoint collection A_1, A_2, ... with $A_i \in \mathcal{F}$ for all i,

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) = \sum_{n\geq 1} \mathbb{P}\left(A_n\right).$$

§2.1 Countable Subadditivity

When the sequence is not necessarily disjoint, this equality becomes an inequality.

Proposition 2.1.1 (Countable Subadditivity)
Let (A_n) be a sequence of events with $A_i \in \mathcal{F}$ for all i. Then we have

$$\mathbb{P}\left(\bigcup_{n\geq 1} A_n\right) \leq \sum_{n\geq 1} \mathbb{P}\left(A_n\right).$$

Proof. Define $B_1 = A_1$ and $B_n = A_n \backslash (A_1 \cup \cdots \cup A_{n-1})$ for all $n \geq 2$. Then (B_n) is a disjoint sequence of events in $\mathcal{F}$, and $\bigcup_{n\geq 1} B_n = \bigcup_{n\geq 1} A_n$. So $\mathbb{P}(\bigcup A_n) = \mathbb{P}(\bigcup B_n)$. By countable additivity for (B_n),

$$\mathbb{P}\left(\bigcup_{n\geq 1} B_n\right) = \sum_{n\geq 1} \mathbb{P}(B_n).$$

But $B_n \subseteq A_n$, so $\mathbb{P}(B_n) \leq \mathbb{P}(A_n)$ for all n. Therefore

$$\mathbb{P}(\bigcup A_n) = \mathbb{P}(\bigcup B_n) = \sum \mathbb{P}(B_n) \leq \sum_{n\geq 1} \mathbb{P}(A_n).$$

□

§2.2 Continuity of Probability Measures

We have continuity for probability measures as follows.

Proposition 2.2.1 (Continuity of Probability Measures)
Let (A_n) be an increasing sequence in $\mathcal{F}$, so that $A_1 \subseteq A_2 \subseteq \cdots$. We know that $\mathbb{P}(A_n) \leq \mathbb{P}(A_{n+1})$. So $\mathbb{P}(A_n)$ converges as $n \to \infty$, and $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_n A_n\right)$.

Proof. Set $B_1 = A_1$ and for $n \geq 2$ $B_n = A_n \backslash (A_1 \cup \cdots \cup A_{n-1})$. Then

$$\bigcup_{k=1}^{n} B_k = A_n, \quad \text{and} \quad \bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k.$$

So $\mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{k=1}^{n} B_k\right) = \sum_{k=1}^{n} \mathbb{P}(B_k) \to \sum_{k=1}^{\infty} \mathbb{P}(B_k)$ as $n \to \infty$.

It remains to prove that $\sum_{k=1}^{\infty} \mathbb{P}(B_k) = \mathbb{P}(\bigcup A_n)$. Since $\bigcup_{k=1}^{\infty} B_k = \bigcup_{k=1}^{\infty} A_k$, we get $\mathbb{P}(\bigcup A_n) = \mathbb{P}(\bigcup B_n) = \sum_n \mathbb{P}(B_n)$. □

Similarly, if (A_n) is a decreasing sequence in $\mathcal{F}$, that is, $A_1 \supseteq A_2 \supseteq \cdots$, then $\mathbb{P}(A_n) \to \mathbb{P}(\cap_n A_n)$ as $n \to \infty$.

§2.3 Inclusion-Exclusion Formula

Suppose that $A, B \in \mathcal{F}$. Then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$. If we also have $C \in \mathcal{F}$, then repeatedly applying the previous we have

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

In general, we have the *inclusion-exclusion formula.*

Proposition 2.3.1 (Inclusion-Exclusion Formula)
Let $A_1, \ldots, A_n \in F$. Then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k}) \right).$$

Proof. We will use induction. For $n = 2$ it holds by definition. Now assume it holds for $n-1$ events. We have

$$\mathbb{P}((A_1 \cup \cdots \cup A_{n-1}) \cup A_n) = \mathbb{P}(A_1 \cup \cdots \cup A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}((A_1 \cup \cdots \cup A_{n-1}) \cap A_n),$$

and we can rewrite the intersection term as $\mathbb{P}((A_1 \cap A_n) \cup \cdots \cup (A_{n-1} \cap A_n))$. Setting $B_i = A_i \cap A_n$, then by the inductive hypothesis we have

$$\mathbb{P}(A_1 \cup \cdots A_{n-1}) = \sum_{k=1}^{n-1} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \cdots < i_k \leq n-1} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k}) \right),$$

and

$$\mathbb{P}(B_1 \cup \cdots B_{n-1}) = \sum_{k=1}^{n-1} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \cdots < i_k \leq n-1} \mathbb{P}(B_{i_1} \cap B_{i_2} \cap \cdots B_{i_k}) \right),$$

Plugging these into our expression gives us the required claim. □

Let $(\Omega, \mathcal{F}, \mathbb{P})$ with Ω finite be a probability space, with $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ for all $A \in F$. Let

$A_1, \dots, A_n \in \mathcal{F}$. Then

$$|A_1 \cup \cdots \cup A_n| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} |A_{i_1} \cap \cdots \cap A_{i_k}|.$$

We can also think about what happens if we truncate the inclusion exclusion formula at some point. For example, for two events we have

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \implies \mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B),$$

which is an upper bound and for three events we have

$$\mathbb{P}(A \cup B \cup C) \geq \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C),$$

which is a lower bound. We can generalize this as follows.

Proposition 2.3.2 (Bonferroni Inequalities)

Truncating the sum in the inclusion-exclusion formula at the rth term gives an overestimate if r is odd, and an underestimate if r is even.

Proof. We will again use induction. For $n = 2$ we know $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. Now assume the claim holds for $n - 1$ events. Suppose that r is odd. Then $\mathbb{P}(A_1 \cup \dots A_n) = \mathbb{P}(A_1 \cup \dots A_{n-1}) + \mathbb{P}(A_n) - \mathbb{P}(B_1 \cap \cdots \cap B_{n-1})$, where $B_i = A_i \cap B_n$.

Sine r is odd, apply the inductive hypothesis to $\mathbb{P}(A_1 \cup \cdots A_{n-1})$ to get

$$\mathbb{P}(A_1 \cup \cdots A_{n-1}) \leq \sum_{k=1}^{r} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \cdots < i_k \leq n-1} \mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots A_{i_k}) \right),$$

and since $r - 1$ is even, we can apply the inductive hypothesis to $\mathbb{P}(B_1 \cup \cdots B_{n-1})$ to get

$$\mathbb{P}(B_1 \cup \cdots B_{n-1}) \geq \sum_{k=1}^{r-1} (-1)^{k+1} \left(\sum_{1 \leq i_1 < \cdots < i_k \leq n-1} \mathbb{P}(B_{i_1} \cap B_{i_2} \cap \cdots B_{i_k}) \right).$$

Substituting both upper bounds in the original expression, we get an overestimate. The case for r even follows analogously. □

We can use the inclusion-exclusion to count combinatorially.

Example 2.3.3 (Number of Surjective Functions)

We will find the number of surjections $f : \{1, \dots, n\} \to \{1, \dots, m\}$.

Let Ω be the set of functions from $\{1, \dots, n\} \to \{1, \dots, m\}$, and A be the subset of surjective functions in Ω. We wish to find $|A|$.

For all $i \in \{1, \dots, m\}$, we define $A_i = \{f \in \Omega \mid i \notin \{f(1), \dots, f(n)\}\}$. Then $A = A_1^c \cap A_2^c \cap \cdots \cap A_m^c = (A_1 \cup \cdots A_m)^c$. Thus $|A| = |\Omega| - |A_1 \cup \cdots \cup A_m| =$

$m^n - |A_1 \cup \cdots \cup A_m|$. Now we have (by inclusion-exclusion)

$$|A_1 \cup \cdots \cup A_m| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \cdots < i_k \le n} |A_{i_1} \cap \cdots \cap A_{i_k}|.$$

We can count $|A_{i_1} \cap \cdots \cap A_{i_k}| = (m-k)^n$. Thus

$$|A_1 \cup \cdots \cup A_m| = \sum_{k=1}^{m} (-1)^{k+1} \binom{m}{k} (m-k)^n$$

So $|A| = \sum_{k=0}^{m} (-1)^k \binom{m}{k} (m-k)^n$.

Counting Derangements

A derangement is a permutation that has no fixed points.

Let Ω be the set of permutations of $\{1, 2, \ldots, n\}$, and A be the set of derangements, $A = \{f \in \Omega \mid f(i) \neq i \text{ for } i = 1, 2, \ldots, n\}$. We pick a permutation at random, and we want to know the probability that it is in A.

Define $A_i = \{f \in \Omega \mid f(i) = i\}$. Then $A = A_1^c \cap \cdots \cap A_n^c = (\bigcup_{i=1}^n A_i)^c$. So $\mathbb{P}(A) = 1 - \mathbb{P}(\bigcup_{i=1}^n A_i)$. By inclusion exclusion,

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \cdot \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \frac{n!}{k! \cdot (n-k)!} \cdot \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!}.
\end{aligned}$$

Thus $\mathbb{P}(A) = 1 - \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} = \sum_{k=0}^{n} \frac{(-1)^k}{k!}$, and as $n \to \infty$, we have $\mathbb{P}(A) \to e^{-1} \approx 0.3678$.

3 Independence and Conditional Probability

We will now look at cases where events are 'independent', and also how we can work with probabilities, given that we know some condition is true.

§3.1 Independence

If we have some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have the notion of *independence*.

Definition 3.1.1 (Independence)

Let $A, B \in \mathcal{F}$. They are called **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.

A countable collection of events (A_n) is said to be **independent** if for all distinct $i_1, i_2, \dots, i_k$, we have

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \prod_{j=1}^{k} \mathbb{P}(A_{i_j}).$$

Note that pairwise independence does not imply independence.

Example 3.1.2 (Pairwise Independence is not Independence)

If we toss a fair coin twice, we have $\Sigma = \{(0,0), (0,1), (1,0), (1,1)\}$, and $\mathbb{P}(\{\omega\}) = 1/4$ for all $\omega \in \Omega$.

Define $A = \{(0,0), (0,1)\}$, $B = \{(0,0), (1,0)\}$ and $C = \{(1,0), (0,1)\}$. Then $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$. Also $\mathbb{P}(A \cap B) = \mathbb{P}(\{(0,0)\}) = 1/4 = 1/2 \cdot 1/2 = \mathbb{P}(A) \cdot \mathbb{P}(B)$. Thus A and B are independent. Similarly, B and C are independent, and A and C are independent.

However, $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\emptyset) = 0 \neq \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C)$, so A, B and C are not independent.

Proposition 3.1.3

If A is independent of B, then A is also independent of B^c.

Proof. $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A) \cdot \mathbb{P}(B)$, by the independence of A and B. Then this is $\mathbb{P}(A) \cdot (1 - \mathbb{P}(B)) = \mathbb{P}(A) \cdot \mathbb{P}(B^c)$. □

§3.2 Conditional Probability

We can now think of this idea of probability based on conditions.

Definition 3.2.1 (Conditional Probability)

Suppose we had some event $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, and let $A \in \mathcal{F}$. We define the **conditional probability** of A given B and write $\mathbb{P}(A \mid B)$ to be

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

If A and B are independent, then $\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A)$. So in this case, $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

We can also generalise this slightly.

Proposition 3.2.2

Suppose (A_n) is a disjoint sequence in $\mathcal{F}$. Then

$$\mathbb{P}\left(\bigcup A_n \mid B\right) = \sum_n \mathbb{P}(A_n \mid B).$$

Proof. By countable additivity, we have

$$\begin{aligned} \mathbb{P}(\bigcup A_n \mid B) &= \frac{\mathbb{P}((\bigcup A_n) \cap B)}{\mathbb{P}(B)} \\ &= \sum_n \frac{\mathbb{P}(A_n \cap B}{\mathbb{P}(B)} \\ &= \sum_n \mathbb{P}(A_n \mid B). \end{aligned}$$

□

We will also use the following result frequently.

Proposition 3.2.3 (Law of Total Probability)

Suppose (B_n) is a disjoint collection in $\mathcal{F}$, and $\bigcup B_n = \Omega$, and $\mathbb{P}(B_n) > 0$ for all n. Let $A \in \mathcal{F}$. Then

$$\mathbb{P}(A) = \sum_n \mathbb{P}(A \mid B_n)\mathbb{P}(B_n).$$

Proof. $\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A \cap (\bigcup_n B_n))$, and by countable additivity of $\mathbb{P}$, $\sum_n \mathbb{P}(A \cap B_n) = \sum_n \mathbb{P}(A \mid B_n) \cdot \mathbb{P}(B_n)$. □

Proposition 3.2.4 (Bayes' Formula)

Let (B_n) be a disjoint collection of events, with $\cup B_n = \Omega$ and $\mathbb{P}(B_n) > 0$ for all n. Then

$$\mathbb{P}(B_n \mid A) = \frac{\mathbb{P}(A \mid B_n) \cdot \mathbb{P}(B_n)}{\sum_k \mathbb{P}(A \mid B_k) \cdot \mathbb{P}(B_k)}.$$

Proof. We have

$$\mathbb{P}(B_n \mid A) = \frac{\mathbb{P}(B_n \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A \mid B_n) \cdot \mathbb{P}(B_n)}{\mathbb{P}(A)},$$

and by the law of total probability, $\mathbb{P}(A) = \sum_k \mathbb{P}(A \mid B_k) \cdot \mathbb{P}(B_k)$. □

This formula is the basis of Bayesian statistics.

We know the probabilities of the events (B_k), and we have a model which gives us the conditional probabilities $\mathbb{P}(A \mid B_n)$. Bayes' formula tells us how to calculate the posterior probabilities of B_n given that the event A occurs.

Example 3.2.5 (False Positive of a Rare Diease)

Suppose that some rare disease A affects 0.1% of the population. We have a medical test that is positive for 98% of the affected population and 1% of those unaffected by the disease. Suppose we picked an individual at random. What is the probability that they suffer from the disease A given that they tested positive?

We define $A = \{\text{individual suffers from } A\}$ and $P = \{\text{individual tested positive}\}$. We want $\mathbb{P}(A \mid P)$.

We have $\mathbb{P}(A) = 0.001$, and $\mathbb{P}(P \mid A) = 0.98$, and $\mathbb{P}(A \mid A^c) = 0.01$. Then

$$\begin{aligned}\mathbb{P}(A \mid P) &= \frac{\mathbb{P}(P \mid A) \cdot \mathbb{P}(A)}{\mathbb{P}(P \mid A) \cdot \mathbb{P}(A) + \mathbb{P}(P \mid A^c) \cdot \mathbb{P}(A^c)} \\ &= \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.01 \cdot 0 \cdot 999} = 0.089 \cdots \approx 0.09.\end{aligned}$$

So $\mathbb{P}(A \mid P) = 0.09$.

The reason why this is so low is that $\mathbb{P}(A \mid A^c)$ is much larger than $\mathbb{P}(A)$.

Example 3.2.6 (Extra Knowledge Gives Surprising Results)

Consider the following three statements:

(a) I have two children, one of which is a boy.

(b) I have two children, and the eldest one is a boy.

(c) I have two children, one of whom is a boy born on a Thursday.

In each case, we want to know $\mathbb{P}(\text{I have 2 boys} \mid a)$ (or b or c).

Since no further information is given, we take all outcomes to be equally likely.

Define the event BG where the eldest is a boy, youngest is a girl. Also define the event GB where the eldest is a girl and youngest is a boy. Lastly define events BB, GG for two boys or two girls respectively.

Now consider the various statements

(a) $\mathbb{P}(BB \mid BB \cup BG \cup GB) = \frac{1}{3}$.

(b) $\mathbb{P}(BB \mid BB \cup BG) = \frac{1}{2}$.

(c) Define the event GT where the eldest is a girl and the youngest is a boy born on a Thursday. Also define TN where the eldest is a boy born on a Thursday, and the youngest is a boy not born on a Thursday. Similarity define TT, TG, and NT.

Then $\mathbb{P}((TT \cup TN \cup NT) \mid (GT \cup TG \cup TT \cup TN \cup NT)) = \frac{13}{27}$.

Example 3.2.7 (Simpson's Paradox)

Consider a program that has 100 applicants, 50 of which are women and 50 of which are men. The table below shows the probability of applicants getting into the program based on what type of school they went to.

All applicants	Admitted	Rejected	% Admitted
State	25	25	50%
Independent	28	22	56%

Now the next two tables show this information for men only and women only.

Men only	Admitted	Rejected	% Admitted
State	15	22	41%
Independent	5	8	38%

Women only	Admitted	Rejected	% Admitted
State	10	3	77%
Independent	23	14	62%

Note that in both the men only and women only, the percentage admitted from independent schools was *lower* than from state schools, but in the total applicants the percentage admitted from independent schools was *higher*.

This phenomenon is called *confounding* in statistics, and arises when we aggregate data from disparate populations.

Define A to be the event that an individual is admitted, B that they are a man, B^c that they are a woman, C that they come from a state school, and C^c that they come from an independent school. Then we see that

$$\mathbb{P}(A \mid B \cap C) > \mathbb{P}(A \mid B \cap C^c), \quad \text{and} \quad \mathbb{P}(A \mid B^c \cap C) > \mathbb{P}(A \mid B^c \cap C^c),$$

but in the example above we have $\mathbb{P}(A \mid C^c) > \mathbb{P}(A \mid C)$.

So

$$\begin{aligned} PP(A \mid C) &= \mathbb{P}(A \cap B \mid C) + \mathbb{P}(A \cap B^c \mid C) \\ &= \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} + \frac{\mathbb{P}(A \cap B^c \cap C)}{\mathbb{P}(C)} \\ &= \mathbb{P}(A \mid B \cap C) + \mathbb{P}(A \mid B^c \cap C) \cdot \mathbb{P}(B^c \mid C) \\ &> \mathbb{P}(A \mid B \cap C^c) \cdot \mathbb{P}(B \mid C) + \mathbb{P}(A \mid B^c \cap C^c)\mathbb{P}(B^c \mid C). \end{aligned}$$

Assuming that $\mathbb{P}(B \mid C) = \mathbb{P}(B \mid C^c)$, though this wasn't the case in the example, then

$$\begin{aligned}\mathbb{P}(A \mid C) &> \mathbb{P}(A \mid B \cap C^c) \cdot \mathbb{P}(B \mid C^c) + \mathbb{P}(A \mid B^c \cap C^c) \cdot \mathbb{P}(B^c \mid C^c) \\ &= \mathbb{P}(A \mid C^c).\end{aligned}$$

So under this extra assumption, we would get that indeed $\mathbb{P}(A \mid C) > \mathbb{P}(A \mid C^c)$.

4 Discrete Probability Distribution

In this chapter we will discuss probability distributions that arise with a discrete random variable.

Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is finite or countable, so $\Omega = \{\omega_1, \omega_2, \dots\}$ and $\mathcal{F}$ is all subsets of Ω.

If we know $\mathbb{P}(\{\omega_i\})$ for all i, then this determines $\mathbb{P}$. Indeed, let $A \subseteq \Omega$. Then $\mathbb{P}(A) = \mathbb{P}(\bigcup_{\omega_i \in A} \{\omega_i\}) = \sum_{\omega_i \in A} \mathbb{P}(\{\omega_i\})$, by countable subadditivity. From this we get the discrete probability distribution.

Definition 4.0.1 (Discrete Probability Distribution)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where $\Omega = \{\omega_1, \omega_2, \dots\}$. We write $p_i \mathbb{P}(\{\omega_i\})$, and we call it a **discrete probability distribution**.

Notable, we have $p_i \geq 0$ for all i, and $\sum_i p_i = 1$

§4.1 Common Discrete Probability Distributions

In the following subsections, we will look at some common discrete probability distributions that arise regularly.

§4.1.1 Bernoulli Distribution

The Bernoulli distribution models 'the outcome of a coin toss'. Here, $\Omega = \{0, 1\}$, and $p_1 = \mathbb{P}(\{1\}) = p$, and $p_0 = \mathbb{P}(\{0\}) = 1 - p$.

§4.1.2 Binomial Distribution

The binomial distribution parameters, $B(N, p)$ where $n \in \mathbb{Z}^+$ and $p \in [0, 1]$. It models the toss of a p-coins (where the probability of heads is p) N times independently.

We get

$$\mathbb{P}(\text{we see } k \text{ heads}) = \binom{N}{k} p^k (1-p)^{N-k}.$$

So in this distribution, $\Omega = \{0, \dots, N\}$, and $p_i = \binom{N}{i} p^i (1-p)^{N-i}$.

§4.1.3 Multinomial Distribution

This distribution is $M(N, p_1, \dots, p_k)$ with $N \in \mathbb{Z}^+$, $p_1, \dots, p_k \geq 0$ and $\sum_i p_i = 1$.

It models having k boxes and N balls, where we throw the balls into random boxes with probabilities as given.

We have $\sigma = \{(n_1, \dots, n_k) \in N^k \mid \sum_{i=1}^k = N\}$, and

$$\mathbb{P}(n_1 \text{ balls in box } 1, \dots, n_k \text{ balls in box } k) = \binom{N}{n_1, \dots, n_k} p_1^{n_1} \cdots p_k^{n_k}.$$

§4.1.4 Geometric Distribution

This distribution corresponds to tossing a p-count until the first head appears.

Here $\Omega = \{1, 2, \dots\}$, and p_k is $\mathbb{P}(\text{tossed } K \text{ times until the first H}) = (1-p)^{k-1} \cdot p$. Also $\sum_{k=1}^\infty p_k = 1$.

We can also define it as $\Sigma = \{0, 1, \dots\}$ with $\mathbb{P}(k \text{ tails before the first } H) = (1-p)^k \cdot p$.

§4.1.5 Poisson Distribution

This distribution is used to model the number of occurrences of an event in a given interval of time. For instance, the number of customers that enter a shop in a day.

We have $\Omega = \{0, 1, 2, \dots\}$, and a parameter $\lambda > 0$ with $\lambda \in \mathbb{R}$. We define $p_k = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$ for all $k \in \omega$. We call this the Poisson distribution with parameter λ.

We note $\sum_{k=0}^\infty = e^{-\lambda} \sum_{k=1}^\infty \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1$, so this is indeed a probability distribution.

Suppose that customers arrive into a shop during $[0, 1]$. If we discretise $[0, 1]$, that is, subdivide it into N intervals $\left[\frac{i-1}{N}, \frac{i}{N}\right]$, $i = 1, \dots, N$. In each interval, a customer arrives with probability p, independent of other intervals, and nobody arrives with probability $1 - p$. So $\mathbb{P}(k \text{ customers arrived}) = \binom{N}{k} \cdot p^k (1-p)^{N-k}$.

Take $p = \lambda/N$. Then

$$\binom{N}{k} \cdot p^k \cdot (1-p)^{N-k} = \frac{N!}{k!(N-k)!} \left(\frac{\lambda}{N}\right)^k \cdot \left(1 - \frac{\lambda}{N}\right)^{N-k} = \frac{\lambda^k}{k!} \frac{N!}{N^k (N-k)!} \left(1 - \frac{\lambda}{N}\right)^{N-k}.$$

Keeping k fixed and letting $N \to \infty$, this gives us

$$\mathbb{P}(k \text{ customers arrived}) \to e^{-\lambda} \cdot \frac{\lambda^k}{k!}, \quad \text{as } N \to \infty.$$

Thus $B(N, p)$ with $o = \lambda/N$ converges to the Poisson distribution with parameter λ.

5 Random variables

§5.1 Definitions

Given some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable X is a function $X : \Omega \to \mathbb{R}$ satisfying $\{\omega \in \Omega \mid X(\omega) \leq< x\} \in \mathcal{F}$, for all $x \in \mathbb{R}$.

We will use the shorthand notation: suppose $A \subseteq \mathbb{R}$. Then $\{X \in A\} = \{\omega \mid X(\omega) \in A\}$.

Given $A \in \mathcal{F}$, define the indicator of A to be

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise} \end{cases}$$

Because $A \in \mathcal{F}$, 1_A is a random variable.

Suppose X is a random variable. We define the probability distribution function of X to be $F_X(x) = \mathbb{P}(X \leq x)$. Note that $F_X : \mathbb{R} \to [0, 1]$.

Definition 5.1.1 (Random Variable in $\mathbb{R}^n$)
$(X_1, \dots, X_n)$ is a **random variable** in $\mathbb{R}^n$ if $(X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$ and for all $x_1, \dots, x_n \in R$, we have $\{X_1 \leq x_1, \dots, X_n \leq x_n\} \in \mathbb{F}$.

Note that this definition is equivalent to saying that $X_1, \dots, X_n$ are all random variables in $\mathbb{R}$.

§5.2 Discrete Random Variables

We are going to look at the specific case of a discrete random variable.

Definition 5.2.1 (Discrete Random Variable)
A random variable X is called **discrete** if it takes values in a countable set.

Suppose X takes values in the countable set S. Then for every $x \in S$, we write $p_x = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \mid X(\omega) = x\})$. We call $(p_x)_{x \in S}$ the **probability mass function** of X (pmf) or the distribution of X.

If (p_x) is Bernoulli, then we say that X is a Bernoulli random variable, or that X has the Bernoulli distribution. If (p_x) is geometric, then similarity we sau X is a geometric random variable.

Definition 5.2.2 (Independence)
Suppose that $X_1, \dots, X_n$ are discrete random variables taking values in $S_1, \dots, S_n$. We say that $X_1, \dots, X_n$ are **independent** if $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n)$.

Example 5.2.3

Toss a p-biased coin N times.

In this case, $\Omega = \{0,1\}^N$, corresponding to tails and heads. Then for $w \in \Omega$, $p_w = \prod p^{\omega_k}(1-p)^{1-\omega_k}$, for $\omega = (\omega_1, \dots, \omega_N)$.

Define $X_k(\omega) = \omega_k$ for every $k = 1, \dots, N$ and $\omega \in \Omega$. This is a discrete random variable, and gives the outcome of the k-th toss.

We have $\mathbb{P}(X_k = 1) = \mathbb{P}(\omega_k = 1) = p$, and $\mathbb{P}(X_k = 0) = \mathbb{P}(\omega_k = 0) = 1 - p$. So X_k has the Bernoulli distribution with parameter p.

We can show that $X_1, \dots, X_N$ are independent random variables. Let $x_1, \dots, x_N \in \{0, 1\}$. Then

$$\begin{aligned}\mathbb{P}(X_n = x_1, \dots, X_N = x_n) &= \mathbb{P}(\omega = (x_1, \dots, x_N)) \\ &= \prod_{k=1}^N p^{x_k}(1-p)^{1-x_k} = \prod_{k=1}^N \mathbb{P}(X_k = x_k).\end{aligned}$$

We can define $S_N(\omega) = X_1(\omega) + \cdots + X_N(\omega)$, which is the number of heads in N tosses. So $S_N : \Omega \to \{0, \dots, N\}$ and $\mathbb{P}(S_N = k) = \binom{N}{k} p^k (1-p)^{N-k}$. So S_N has the Binomial distribution of parameters N and p.

§5.3 Expectation

For some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and assuming that Ω is finite or countable. Let $X : \Omega \to \mathbb{R}$ be a discrete random variable. We say that X is **non-negative** if $X \geq 0$.

Definition 5.3.1 (Expectation)

We define the expectation of $X \geq 0$ so that

$$\mathbb{E}[X] = \sum_\omega X(\omega) \cdot \mathbb{P}(\{\omega\})$$

Writing $\Omega_X = \{X(\omega) \mid \omega \in \Omega\}$, so $\Omega = \bigcup_{x \in \Omega_X} \{X = x\}$, we can write this in the form

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} \sum_{\omega \in \{X = x\}} x \cdot \mathbb{P}(\{\omega\}) = \sum_{x \in \Omega_x} x \cdot \mathbb{P}(X = x).$$

So the expectation of X (or the average value) is an average of the values taken by X with weights given by $\mathbb{P}(X = x)$.

Example 5.3.2

Suppose X has the Binomial distribution with parameters N and p ($X \sim B(N, p)$).

Then computing we have

$$\mathbb{E}[X] = \sum_{k=0}^{N} k \cdot \mathbb{P}(X = k) = Np.$$

Example 5.3.3

Let X be a Poisson random variable with parameter $\lambda > 0$. Then

$$\mathbb{E}[X] = \lambda.$$

Now let X be a general (not necessarily non-negative) discrete random variable. We define $X_+ = \max(X, 0)$ and $X_- = \max(-X, 0)$. Then $X = X_+ - X_-$. We can then define $\mathbb{E}[X_+]$ and $\mathbb{E}[X_-]$. Then if at least one of them are finite, we have

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-].$$

If both are infinite, then we say the expectation of X is not defined. Whenever we write $\mathbb{E}[X]$, it is assumed to be well defined. If $\mathbb{E}[|X|] < \infty$, we say that X is **integrable**.

When $\mathbb{E}[X]$ is well defined, we again have that

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x \cdot \mathbb{P}(X = x).$$

Proposition 5.3.4 (Properties of Exectation)

For some discrete random variable X, we have the following.

(i) If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.

(ii) If $X \geq 0$ and $\mathbb{E}[X] = 0$, then $\mathbb{P}(X = 0) = 1$.

(iii) If $c \in \mathbb{R}$, then $\mathbb{E}[cX] = c\mathbb{E}[X]$ and $\mathbb{E}[c + X] = c + \mathbb{E}[X]$.

(iv) If X and Y are two random variables, then $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, where X and Y are both integrable.

(v) Suppose $c_1, \dots, c_n \in \mathbb{R}$ and $X_1, \dots, X_n$ are discrete random variables. Then

$$\mathbb{E}\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i].$$