Contents

Chapter 7

Probability

7.1 Introduction to Probability

7.1.1 What is probability?

Probability is how we quantify uncertainty; it is the *extent to which an event is likely to occur.* We use it to study events whose outcomes we do not (yet) know, whether this is because they have not happened yet, or because we have not yet observed them.

We quantify this uncertainty by assigning each event a number between 0 and 1. The higher the probability of an event, the more likely it is to occur.

Historically, the early theory of probability was developed in the context of gambling. In the seventeenth century, Blaise Pascal, Pierre de Fermat, and the Chevalier de Méré were interested in questions like "If I roll a six-sided die four times, how likely am I to get at least one six?" and "if I roll a *pair* of dice twenty-four times, how likely am I to get at least one pair of sixes?" Many of the examples we'll see in this course still use situations like rolling dice, drawing cards, or sticking your hand into a bag filled with differently-coloured tokens.

Nowadays, probability theory helps us to understand how the world around us works, such as in the study of genetics and quantum mechanics; to model complex systems, such as population growth and financial markets, and to analyse data, via the theory of *statistics*.

We'll see a bit of statistical theory at the end of this chapter, but will mostly stay on the probabilistic side of that line.

7.1.2 Events

As we noted above, we use probability theory to describe scenarios in which we don't know what the outcome will be. We call these scenarios **experiments** or **trials**.

The set of all possible outcomes of an experiment is its **sample space**, S. Subsets of S are called **events**, and may contain several different outcomes.

Example 7.1.2.1 *In the experiment in which we roll a single six-sided die, we have:*

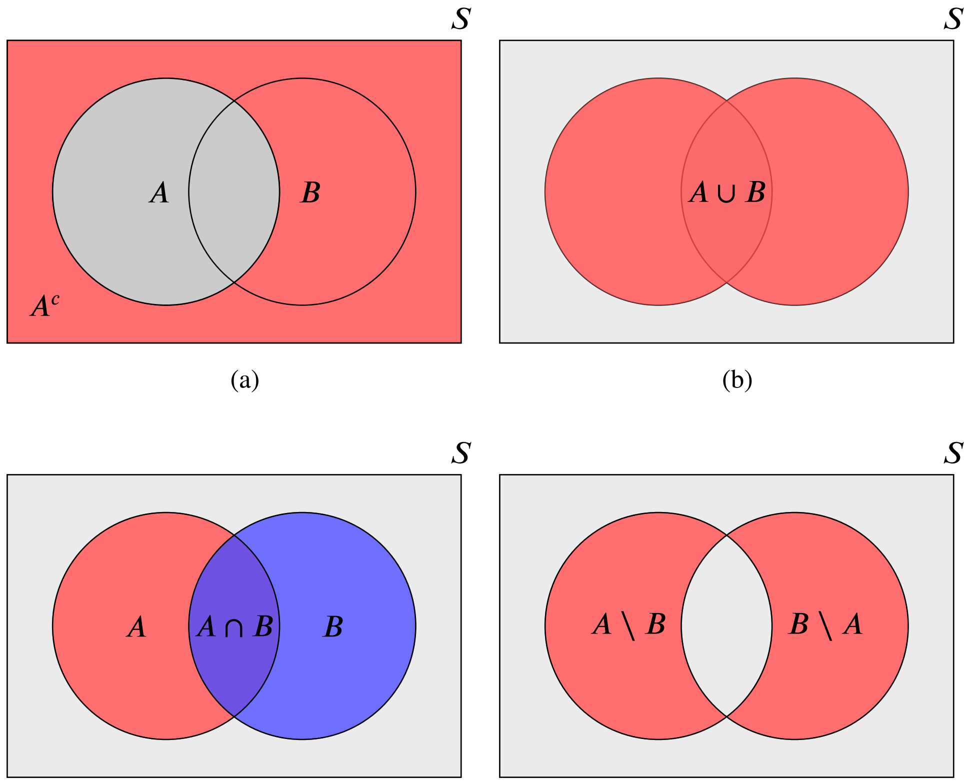
- *The sample space is* $S = \{1, 2, 3, 4, 5, 6\}$
- *An example of a possible outcome is 5 (or "we roll a five")*
- *An example of an event is* $A = \{2, 4, 6\}$ *(or "we roll an even number").*

Because events are subsets of the sample space, we can treat them as sets.

Set operations

There are three basic operations we can use to combine and manipulate sets. If A and B are events, then

- The event *not* A, which we write A^c (the c is for *complement*), is the set of all outcomes in S which are not in A.
- The event A *or* B, which we write $A \cup B$ and call the *union* of A and B, is the set of all outcomes which are in at least one of A and B.
- The event A *and* B, which we write $A \cap B$ and call the *intersection* of A and B, is the set of all outcomes which are in both A and B.



(a) (b) (c) (d)

Figure 7.1: Some nice pictures illustrating: (a) A^c, (b) $A \cup B$, (c) $A \cap B$, (d) $A \setminus B$.

Working with events

When we want to consider all the outcomes in an event A which are *not* in B, we write $A \cap B^c = A \setminus B$.

We say that two events are **disjoint** (or incompatible, or mutually exclusive) if they cannot occur at the same time; in other words, if A and B are disjoint, then $A \cap B$ contains no outcomes.

We write $A \cap B = \emptyset$, and we call $\emptyset$ the empty set.

If every outcome in an event A is also in an event B, we say that A is a *subset* of B, and we write $A \subseteq B$. For example, since all Single Maths students are fans of probability,

$$\{\text{Single Maths students}\} \subseteq \{\text{Fans of probability}\}.$$

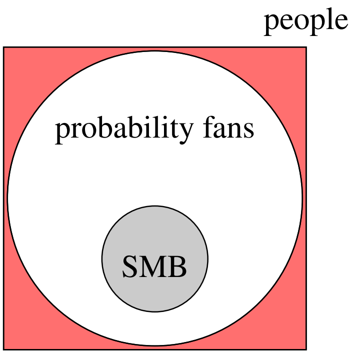


Figure 7.2: Notice that the circle of “probability fans“ takes up quite a lot of the state space.

The following set of basic rules will be helpful when working with events.

Commutativity:

$$A \cup B = B \cup A, \qquad A \cap B = B \cap A$$

Associativity:

$$(A \cup B) \cup C = A \cup (B \cup C), \qquad (A \cap B) \cap C = A \cap (B \cap C)$$

Distributivity:

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C), \qquad (A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

De Morgan’s laws:

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c$$

For example, if $A = \{\text{Dinner is on time}\}$ and $B = \{\text{Dinner is delicious}\}$, then

$$(A \cap B)^c = \{\text{Dinner is either late or disappointing}\},$$

and

$$(A \cup B)^c = \{\text{Dinner is } \textit{both} \text{ late } \textit{and} \text{ disappointing}\}.$$

7.1.3 Axioms of Probability

Once we have decided what our experiment (and hence our sample space) should be, we assign a probability to each event $A \subseteq S$. This probability is a number, which we write $\mathbb{P}(A)$.

Remember that A is an event, which is a set, and that $\mathbb{P}(A)$ is a probability, which is a number. It makes sense to take the union of sets, or to add numbers together - but not the other way around!

We need a system of rules (the *axioms*) for how the probabilities are assigned, to make sure everything stays consistent. There are lots of such systems, but we will use Kolmogorov's axioms, from 1933. There's no particular reason to choose one system over another, but these are a popular choice.

The axioms are:

1. The probability of any event is a real number in the interval $[0, 1]$: $0 \leq \mathbb{P}(A) \leq 1$.
2. The probability that *something* in S happens is 1: $\mathbb{P}(S) = 1$.
3. If A and B are disjoint events, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

We can use set operations to see some immediate consequences of the axioms:

- Since A and A^c are disjoint, we have $\mathbb{P}(A^c) = \mathbb{P}(S) - \mathbb{P}(A) = 1 - \mathbb{P}(A)$.
- Impossible events have probability zero: $\mathbb{P}(\emptyset) = 0$.
- For (not necessarily disjoint) events A and B, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
- If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Suggested exercises: Q1 – Q10.

7.1.4 Counting principles

When our experiment has m outcomes, each of which is *equally likely*, then for any event A we have

$$\mathbb{P}(A) = \frac{|A|}{m} = \frac{\text{number of ways A can occur}}{\text{total no. of outcomes}}.$$

In this section, we look at some different ways to count the number of outcomes in an event, when the events are more complex than, say, a roll of a die.

The multiplication principle

If our experiment can be broken down into r parts, in which

- the first part has m_1 equally-likely outcomes
- the second part has m_2 equally-likely outcomes
- $\cdots$
- the rth part has m_r equally-likely outcomes,

then there are

$$m_1 \times m_2 \times \cdots \times m_r = \prod_{j=1}^{r} m_j$$

possible, equally-likely, outcomes for the whole experiment.

Example 7.1.4.1

- *If there are four different routes from Newcastle to Durham, and three different routes from Durham to York, how many different routes are there from Newcastle to York?*
- *If I toss six coins (1p, 2p, 5p, 10p, 15p, and 20p), how many different ways are there to get one 'heads' and five 'tails'?*
- *In general, sampling r times with replacement from m options gives m^r different possiblities.*

Permutations

When we select r items from a group of size n, in order and without replacement, we call the result a **permutation of size r from n.**

The number of permutations of size r from n is

$$n \times (n-1) \times \cdots \times (n-r+1) = \frac{n!}{(n-r)!}.$$

A special case is when we want to arrange the whole list. Then, there are

$$r \times (r-1) \times \cdots \times 1 = \frac{r!}{0!} = r!$$

different permutations.

Example 7.1.4.2

- *How many different ways are there to arrange six books on a shelf?*
- *In a society with twenty members, which must choose one president and one secretary, how many different ways can these roles be filled?*

- *If six (six-sided) dice are rolled, what is the probability that each of the numbers 1-6 appears exactly once?*

Combinations

When we select r items from a group of size n, without replacement, but not in any particular order, then we have a *combination of size r from n.*

There are

$$\binom{n}{r} = \frac{(n!)}{(n-r)!r!}$$

different ways to choose a combination of size r from n objects.

Two useful ways of thinking about combinations:

- You might notice that $\binom{n}{r} = \binom{n}{n-r}$. This is because we can also look at the combination of items we *don't* pick. It's much easier (psychologically, at least) to list the different ways to leave 3 cards in the deck than it is to list the different ways to draw 49 cards!

- There is a relationship between combinations and permutations:

  $$\text{the number of combinations} = \frac{1}{r!} \times \text{ the number of permutations.}$$

  This is because each combination counted when the order *doesn't* matter comes up $r!$ different times when the order *does* matter.

Example 7.1.4.3 1. *How many different ways are there to form a subcommittee of eight people, from a group of twenty?*

2. *If I have n points on the circumference of a circle, how many different triangles can I form with vertices among these points?*

Remember: If we're allowed repeated values, the only tool we need is the multiplication principle.

If there can be no repeats (sampling without replacement), then we use permutations if the objects are all distinct, and combinations if they are not. Usually if we're dealt a hand of cards, or draw a bunch of things out of a bag, then they're indistinguishable. But if we're rolling several dice, or assigning objects to people, then we can (hopefully) tell the dice or people apart.

You might find the flowchart in Figure 7.3 helpful.

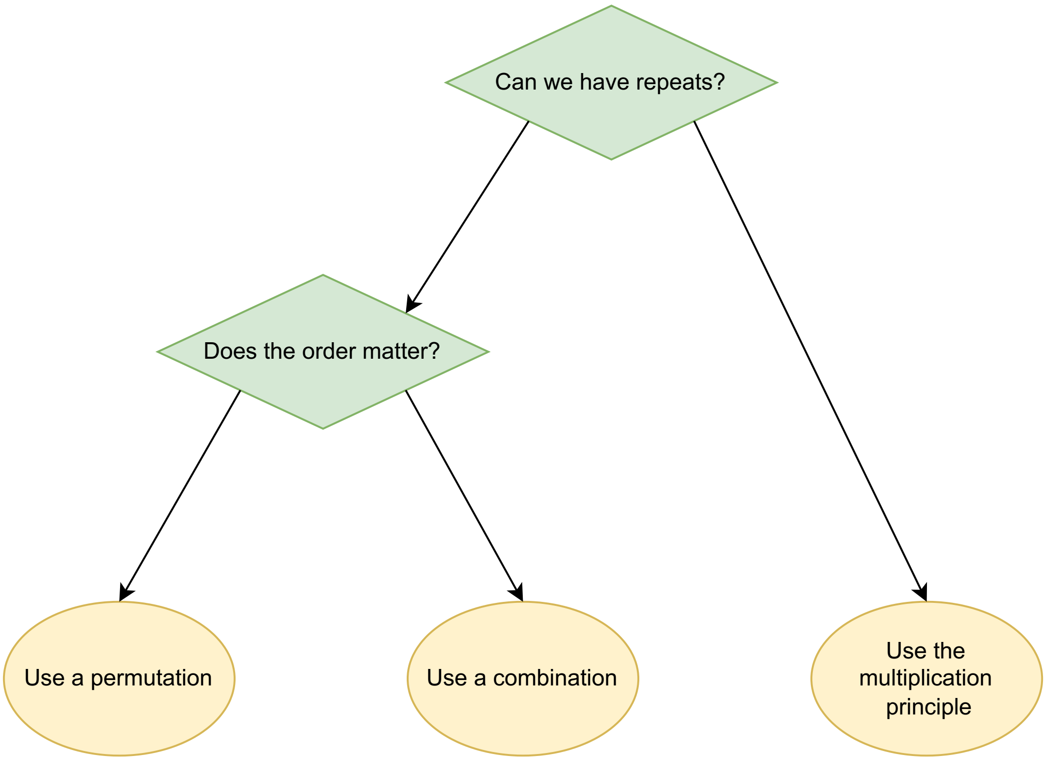


Figure 7.3: A decision-making flowchart for permutations, combinations, and the multplication principle.

Multinomial coefficients

When we want to separate a group of size n into $k \geq 2$ groups of possibly different sizes, we use *multinomial coefficients*. If the group sizes are $n_1, n_2, \ldots, n_k$, with $n_1 + n_2 + \cdots + n_k = n$, then the number of different ways to arrange the groups is given by the multinomial coefficient

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

To see how this works, think about choosing the groups in order. There are $\binom{n}{n_1}$ ways to choose the first group; then, there are $\binom{n-n_1}{n_2}$ ways to choose the second group from the remaining objects. Continuing like this until all the groups are selected, by the multiplication principle there are

$$\binom{n}{n_1} \times \binom{n-n_1}{n_2} \times \binom{n-n_1-n_2}{n_3} \times \cdots \times \binom{n_{k-1}+n_k}{n_{k-1}} \times \binom{n_k}{n_k}$$

ways to choose all the groups. Writing each binomial coefficient in terms of factorials, and doing (lots of nice) cancelling, we end up with our expression for the multinomial coefficient.

Example 7.1.4.4 - *In how many different (that is, distinguishable) ways can you arrange the letters in STATISTICS?*

- *If you arrange the letters S,S,S,T,T,T,I,I,A,C in a random order, what is the probability that they spell 'Statistics'?*

Suggested exercises: Q11 – Q17.

7.1.5 Conditional Probability and Bayes' Theorem

Sometimes, knowing whether or not one event has occurred can change the probability of another event. For example, if we know that the score on a die was even, there is a one in three chance that we rolled a two (rather than one in six). Gaining the knowledge that our score is even affects how likely it is that we got each possible score.

We write $\mathbb{P}(A \mid B)$ for the *conditional probability of A, given B*; it is defined by

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

We can rearrange this expression to get

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

which leads to **Bayes' theorem**:

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

Writing conditional probabilities in this way allows us to "invert" them; quite often, one of $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$ is easier to spot than the other.

7.1.6 Independence

We say that two events are **independent** if the occurrence of one has no bearing on the occurrence of the other, that is,

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

Example 7.1.6.1 - *The scores obtained from rolling two separate dice are independent*

- *Height and shoe size of people are usually not independent*

- *Lecture attendance and exam grades are not independent!*

When events A and B are independent, we have

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

7.1.7 Partitions

Suppose we can separate our state space into *n disjoint events* $E_1, E_2, \dots, E_n$: we know that exactly one of these events must happen. We call the collection $\{E_1, E_2, \dots, E_n\}$ a **partition**, and we can use it to break down the probabilities of different events $A \subseteq S$.

First, we can write

$$A = (A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_n),$$

so that

$$\mathbb{P}(A) = \mathbb{P}(A \cap E_1) + \mathbb{P}(A \cap E_2) + \cdots + \mathbb{P}(A \cap E_n).$$

We can also introduce conditional probability, to get **the partition theorem**:

$$\mathbb{P}(A) = \mathbb{P}(A \mid E_1)\,\mathbb{P}(E_1) + \mathbb{P}(A \mid E_2)\,\mathbb{P}(E_2) + \cdots + \mathbb{P}(A \mid E_n)\,\mathbb{P}(E_n).$$

The partition theorem is useful whenever we can break an event down into cases, each of which is straightforward.

Example 7.1.7.1 *One of the most well-known (especially recently!) examples of the partition theorem is in testing for diseases.*

Suppose that a disease affects one in 10,000 people. We have a test for this disease which correctly identifies 90% of people who do *have the disease (so gives false negatives to 10% of people with the disease), and gives false positives to 1% of people who* do not *have the disease.*

If a random person is tested, what is the probability that their test result is positive?

Given that the test result is positive, what is the probability that they have the disease?

Suggested exercises: Q18 – Q26.

7.2 Random variables

A **random variable** is a variable which takes different numerical values, according to the different possible outcomes of an experiment.

Example 7.2.0.1 *If the* experiment *is "toss four coins", then some of the elements of the state space are HHHH, HHHT, HHTH, HHTT,... . One random variable we can define is*

$$X = \textit{Number of heads.}$$

Then if our outcome *is HHTT, we have* $X = 2$.

We say that a random variable is **discrete** if we can list its possible values, or **continuous** if it can take any value in a range.

7.2.1 Discrete random variables

To define a discrete random variable, we need to know its **probability distribution**, which is sometimes called a probability mass function.

The probability distribution is often displayed in a table, which shows the different values X can take, along with the associated probabilities:

values	x_1	x_2	$\ldots$	x_n
probabilities	$\mathbb{P}(X = x_1)$	$\mathbb{P}(X = x_2)$	$\ldots$	$\mathbb{P}(X = x_n)$

In a probability distribution, the values must be *non-negative* and must *sum to 1.* To find the probability that X lies in an interval $[a, b]$, we have

$$\mathbb{P}(a \leq X \leq b) = \sum_{a \leq x_i \leq b} \mathbb{P}(X = x_i).$$

Joint and marginal distributions

When we have two (or more) discrete random variables, X and Y (and Z and...), the **joint probability distribution** is the table of every possible (x, y) value for X and Y, with the associated probabilities $\mathbb{P}(X = x, Y = y)$:

	x_1	$\dots$	x_n
y_1	$\mathbb{P}(X = x_1, Y = y_1)$	$\dots$	$\mathbb{P}(X = x_n, Y = y_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
y_m	$\mathbb{P}(X = x_1, Y = y_m)$	$\dots$	$\mathbb{P}(X = x_n, Y = y_m)$

We can find the **marginal probability distributions** of X and Y from the joint distribution, by summing across the rows or columns:

$$\mathbb{P}(X = x_k) = \sum_j \mathbb{P}(X = x_k, Y = y_j),$$
$$\mathbb{P}(Y = y_j) = \sum_k \mathbb{P}(X = x_k, Y = y_j).$$

Example 7.2.1.1 *Let X be the random variable which takes value 3 when a fair coin lands heads up, and takes value 0 otherwise. Let Y be the value shown after rolling a fair die. Write down the distributions of X, and Y, and the joint distribution of (X, Y). You may assume that X and Y are independent. Use your table to find the probability that $X > Y$.*

7.2.2 Continuous random variables

When our random variable is continuous, we can't describe it using a list of probabilities. Instead. we use a **probability density function** (pdf), $f_X(x)$. The pdf describes a curve over the possible values taken by the random variable.

In a density function, the values must be *non-negative* and *integrate to 1.* To find the probability that X lies in an interval $[a, b]$, we have

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

Remember that the density $f_X(x)$ is not the same thing as $\mathbb{P}(X = x)$. In fact, for every x, we have $\mathbb{P}(X = x) = 0$.

Another way of specifying the distribution of a continuous random variable is through its **cumulative distribution function**, or cdf, given by

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(t)dt.$$

Joint and marginal distributions

When we have two (or more) continuous random variables, X and Y (and Z and...), we describe them via their **joint probability density function** $f_{X,Y}(x, y)$. As it is a density, $f_{X,Y}$ is non-negative, and must integrate to 1. The probability that X and Y are in a region A of the xy-plane is

$$\mathbb{P}((X, Y) \in A) = \int\int_A f_{X,Y}(x, y)dxdy.$$

We can find the **marginal probability distributions** of X and Y from the joint distribution, by integrating out one of the variables:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx. \end{aligned}$$

Example 7.2.2.1 *Let X be a continuous random variable with probability density function:*

$$f_X(x) = \begin{cases} \beta e^{-\beta x} & \textit{for } x > 0, \\ 0 & \textit{for } x \leq 0. \end{cases}$$

Check that $f_X(x)$ is a valid probability density function when $\beta > 0$. Find the cdf of X, and hence $\mathbb{P}(X > 3)$.

Suggested exercises: Q27 – Q32.

7.3 Expectation and Variance

While the probability distribution or probability density function tells us everything about a random variable, this can often be too much information. Summaries of the distribution can be useful to convey information about our random variable without trying to describe it in its entirity.

Summaries of a distribution include the expectation, the variance, the skewness and the kurtosis. In this course, we're only interested in the expectation, which tells us about the *location* of the distribution, and the variance, which tells us about its *spread*.

7.3.1 Expectation

The **expectation** of a random variable X is given by

$$\mathbb{E}[X] = \sum_x x\,\mathbb{P}(X = x) \qquad \text{or} \qquad \mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\,dx.$$

The expectation is sometimes called the mean or the average of X.

Properties of Expectation

Linearity: If X is a random variable and a and b are (real) constants, then

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b.$$

Additivity: If $X_1, X_2, \dots, X_n$ are random variables, then

$$\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

Positivity: If X is a positive random variable ($\mathbb{P}(X \geq 0) = 1$), then $\mathbb{E}[X] \geq 0$.

Independence: If X and Y are independent random variables, then

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

Expectation of a function: If X is a random variable and r is a (nice[1]) function of X, then

$$\mathbb{E}[r(X)] = \sum_x r(x)\mathbb{P}(X = x) \qquad \textit{or} \qquad \mathbb{E}[r(X)] = \int_{\mathbb{R}} r(x) f_X(x) dx.$$

7.3.2 Variance

For a random variable X with expectation $\mathbb{E}[X] = \mu$, the **variance** of X is given by

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2].$$

By expanding out the brackets and using the linearity of the expectation, we can rewrite the variance as

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

[1] Here 'nice' actually means 'measurable'. It's possible to come up with functions r for which this doesn't work; luckily for us, they're usually quite weird and we won't run into any of them.

The variance is always positive, because it is the expectation of a positive random variable. The **standard deviation** is the square root of the variance:

$$SD(X) = \sqrt{\text{Var}(X)}.$$

Properties of Variance

Linear combinations: If X is a random variable and a and b are (real) constants, then

$$\text{Var}(aX + b) = a^2\text{Var}(X).$$

Independence: If X and Y are independent random variables, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Example 7.3.2.1 *Let X be a continuous random variable with probability density function:*

$$f_X(x) = \begin{cases} \beta e^{-\beta x} & \textit{for } x > 0, \\ 0 & \textit{for } x \leq 0. \end{cases}$$

What are the expectation and variance of X?

Example 7.3.2.2 *Let Y be a random variable with the following probability distribution:*

y	1	2	3
$\mathbb{P}(Y = y)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$

Find $\mathbb{E}[X]$, Var(X), and $\mathbb{E}\left[\frac{1}{X}\right]$.

Suggested exercises: Revisit Q30; Q33 – Q37.

7.4 The Binomial Distribution

The *Bernoulli distribution* is used to describe the following situation:

Our experiment consists of a fixed number (n) of trials, which either succeeds with probability p, or fails with probability $1 - p$.

If X is the number of successes (0 or 1), we say that X has a *Bernoulli distribution* with parameter p, and we write $X \sim \text{Bern}(p)$.

The expectation and variance of X are:

$$\mathbb{E}[X] = p$$
$$\text{Var}(X) = p(1-p).$$

Suppose we have n Bernoulli-style trials, which succeed or fail *independently* of each other. All trials have the same probability p of succeeding. We count the *total number of successes* across all the trials.

If Y is this total, we say that Y has a Binomial distribution with parameters n and p, and we write $Y \sim \text{Bin}(n, p)$.

If $0 \le k \le n$, we have

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

This is because each configuration of k successes and $n-k$ successes has probability $p^k(1-p)^{n-k}$, by the multiplication principple; and there are $\binom{n}{k}$ different ways of arranging the k successes and $n-k$ failures among the trials.

Exercise: Check that the probabilities in the Binomial distribution:

- are all non-negative
- sum to 1.

The expectation and variance of Y are:

$$\mathbb{E}[Y] = np$$
$$\text{Var}(Y) = np(1-p).$$

If $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, p)$, *and X and Y are independent*, then $X + Y \sim \text{Bin}(m+n, p)$.

Example 7.4.0.1

- *If I toss six coins, the total number of heads has a Bin*$(6, \frac{1}{2})$ *distribution.*
- *If each SMB student decides to skip a lecture with probability 0.2, then the number of students who turn up has a Bin*$(195, 0.8)$ *distribution (assuming you all decide independently of each other!)*

7.5 The Poisson Distribution

While the Binomial distribution is about counting successes in a *fixed* number of trials, the Poisson distribution lets us count how many times something happens without a fixed upper limit. This is useful in a lot of real-world contexts, for example:

- the number of people who visit a website
- the number of yeast cells in a sample (such as in experiments by Gossett at Guinness in the 1920s)
- the number of particles emitted from a radioactive sample.

The Poisson distribution is used to model scenarios in which events happen randomly, independently, and at a *constant rate* r. If X is the total number of these events that happen in a time period of length s, then X has a Poisson distribution with parameter $\lambda = rs$, and we write $X \sim \text{Po}(\lambda)$.

If $k \in \mathbb{N}$, we have

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Exercise: Check that the probabilities in the Poisson distribution:

- are all non-negative
- sum to 1.

The expectation and variance of X are

$$\mathbb{E}[X] = \text{Var}(X) = \lambda.$$

If $X \sim \text{Po}(\lambda)$ and $Y \sim \text{Po}(\mu)$, *and X and Y are independent*, then $X + Y \sim \text{Po}(\lambda + \mu)$.

7.5.1 Using the Poisson distribution to approximate the Binomial distribution

Instead of thinking about our time period $[0, s]$ as one long interval, we can split it up into n smaller ones (each one will have length $\frac{s}{n}$).

Suppose we count the number of sub-intervals in which events occur. If the sub-intervals are small enough, it is very unlikely that there will be multiple events in any of them, and the probability that there is one event will be $p \approx \frac{rs}{n} = \frac{\lambda}{n}$.

We can view the sub-intervals as n independent trials, and the total number of successes becomes Binomially-distributed.

This is a good approximation because the probabilities in the Binomial and Poisson distributions are similar:

$$\begin{aligned}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} &= \frac{n(n-1)\dots(n-k+1)}{k!} \times \frac{\lambda^k}{n^k} \times \left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{n^k} \times \left(1-\frac{\lambda}{n}\right)^{n-k} \times \frac{\lambda^k}{k!} \\ &\approx 1 \times e^{-\lambda} \times \frac{\lambda^k}{k!},\end{aligned}$$

as long as n is big enough.

This approximation is good if $n \geq 20$ and $p \leq 0.05$, and excellent if $n \geq 100$ and $np \leq 10$. It is useful because calculating $e^{-\lambda}$ is often computationally much more efficient than calculating $\binom{n}{k}$, especially when n is large!

Suggested exercises: Revisit Q38 – Q41.

7.6 The Normal Distribution

Unlike the Binomial and Poisson distributions, the Normal (or Gaussian) distribution is continuous. It is one of the most used (and most useful) distributions. Random variables whose "large-scale" randomness comes from many small-scale contributions is usually Normally distributed: for example, people's heights are determined by many different genetic and environmental factors. All of these different factors have tiny impacts on your final height; overall, the distribution of the height of a random person is roughly Normal.

7.6.1 The standard Normal distribution

The first version of the Normal distribution we will meet is the *standard* Normal. We say that a random variable Z has a standard Normal distribution, and we write $Z \sim \mathcal{N}(0, 1)$, if

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad \forall x \in \mathbb{R}.$$

Properties of the standard Normal distribution

- The density of the Normal distribution is symmetric about 0; so the variables Z and $-Z$ have the same distribution.
- This symmetry also means that $xf_Z(x)$ is an odd function; so the expectation of Z is zero.
- The variance of Z is

$$\begin{aligned}\mathrm{Var}(Z) &= \mathbb{E}[Z^2] - 0\\ &= \int_{-\infty}^{\infty} x^2 f_Z(x)dx\\ &= \frac{1}{\sqrt{2\pi}}\int_{\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}dx\\ &= 1.\end{aligned}$$

(You can find this via integration by parts)

The cumulative distribution function for Z

The cumulative distribution function for Z is denoted $\Phi(z)$ and is given by

$$\Phi(z) = \mathbb{P}(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx.$$

There is no neat ("algebraic") expression for $\Phi(z)$: in practice, when we need to evaluate it we use numerical methods to get (usually very good) approximations. These values are traditionally recorded in tables but usually, they're built into computer software and some calculators.

Some useful properties of $\Phi(z)$, which reduce the number of values we need in the tables, are:

- Because $f_Z(x)$ is symmetric,

$$\mathbb{P}(Z \leq z) = \mathbb{P}(-Z \leq z) = \mathbb{P}(Z \geq -z);$$

so $\Phi(z) = 1 - \Phi(-z)$.

- We have $\Phi(0) = \frac{1}{2}$.
- $\mathbb{P}(a \leq Z \leq b) = \Phi(b) - \Phi(a)$.

Interpolation: When the value we need to find isn't in a table we have access to, we can interpolate. If $a \leq b \leq c$ and we know $\Phi(a)$ and $\Phi(b)$, we approximate:

$$\Phi(b) \approx \Phi(a) + \frac{b-a}{c-a}\left(\Phi(c) - \Phi(a)\right).$$

For example, most Normal tables only go to two decimal places, but $\Phi(0.553)$ will be approximately $3/10$ths of the way between $\Phi(0.55)$ and $\Phi(0.56)$.

7.6.2 General Normal Distributions

We say that X has a Normal distribution with parameters μ and σ^2, and we write $X \sim \mathcal{N}(\mu, \sigma^2)$, if the variable $Z = \frac{X-\mu}{\sigma}$ has a standard Normal distribution.

We can also write this in the other direction: $X \sim \mathcal{N}(\mu, \sigma^2)$ if $X = \mu + \sigma Z$. Since the distribution of Z is symmetric, we use the convention $\sigma > 0$.

Properties of general Normal distributions

- The expectation of X is

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[\mu + \sigma Z] \\ &= \mu + \sigma \mathbb{E}[Z] \\ &= \mu + 0 = \mu. \end{aligned}$$

- The variance of X is

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(\mu + \sigma Z) \\ &= \sigma^2 \mathrm{Var}(Z) \\ &= \sigma^2. \end{aligned}$$

- The density of X is

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}.$$

- The cdf of X is given by

$$\begin{aligned}\mathbb{P}(X \leq x) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right)\\ &= \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right)\\ &= \Phi\left(\frac{x-\mu}{\sigma}\right).\end{aligned}$$

  We can use the table for the standard Normal distribution to evaluate the cdf of *any* Normal distribution, by using this transformation.

Example 7.6.2.1 *1. If $X \sim \mathcal{N}(12, 25)$, what is $\mathbb{P}(X \leq 3)$?*

2. If $Y \sim \mathcal{N}(1, 4)$, what is $\mathbb{P}(-1 < Y < 2)$?

7.6.3 Using the Normal distribution to approximate the Binomial and Poisson distributions

Just as we can use the Poisson distribution to approximate *specific* probabilities in the Binomial distribution, we can use the Normal distribution to approximate *cumulative* probabilities. If n is large and $X \sim \text{Bin}(n, p)$, then approximately we have $X \sim \mathcal{N}(np, np(1-p))$.

In particular,

$$\mathbb{P}(X \leq k) \approx \Phi\left(\frac{k - np}{\sqrt{np(1-p}}\right).$$

This is a useful approximation when both np and $np(1-p)$ are at least 10; as the two values increase, the approximation gets better.

Example 7.6.3.1 *A machine produces $n = 1500$ gadgets every day. Each individual gadget is defective with probability $p = 0.02$. Find (approximately) the probability that more than 40 of the items produced in one day are defective.*

Similarly, we can use the Normal distribution to approximate the cumulative probabilities in the Poisson distribution: if $X \sim \text{Po}(\lambda)$, then approximately we have $X \sim \mathcal{N}(\lambda, \lambda)$ and

$$\mathbb{P}(X \leq k) \approx \Phi\left(\frac{k-\lambda}{\sqrt{\lambda}}\right).$$

This is a useful approximation when λ is at least 5, and gets better as λ increases.

Suggested exercises: Q42–45.

7.7 The Central Limit Theorem

7.7.1 Experimental errors

When we are measuring a quantity whose "true value" is μ, our measurement takes the form $X = \mu + \varepsilon$, where ε is the *experimental error*. Before we do the experiment, we can think of both ε and X as random quantities. Afterwards, X is a fixed and known quantity, and μ and ϵ are fixed but unknown quantities (to us). Our goal is to use X to infer something about μ.

Assumption: We will assume that there are no *systematic errors* or bias in the experiment; in other words, $\mathbb{E}[\varepsilon] = 0$.

If the variance of ε is $\mathrm{Var}(\varepsilon) = \sigma^2$, then

$$\begin{aligned} \mathbb{E}[X] &= \mu + \mathbb{E}[\varepsilon] = \mu + 0 = \mu \\ \mathrm{Var}(X) &= 0 + \mathrm{Var}(\varepsilon) = \sigma^2. \end{aligned}$$

This means that, on average, the value of our measurement is a good estimate of the value of μ; however, if the variance of ε is large, our measurement will have quite a high probability of being far from the true value.

To improve our estimate, we can do one of two things:

- try to improve our measurement technique, to reduce the variance
- take more measurements!

7.7.2 The sample mean

We take n independent random variables $X_1, X_2, \dots, X_n$, which all have the same distribution – for example, we might repeat our measurement n times, or sample n people from a large population. We say that $X_1, X_2, \dots, X_n$ are *independent and identically distributed* (i.i.d.).

The sample mean is the average of the values $X_1, X_2, \ldots, X_n$:

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

Before we take our measurements, this is also a random variable; *afterwards*, it is just a number. To distinguish between the two situations, we use $\bar{X}$ for the random variable, and $\bar{x}$ for the number.

Assumption: We assume that $X_1, X_2, \ldots, X_n$ are i.i.d. with shared mean μ and variance σ^2.

Then

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[X_j] = \frac{n}{n}\mu = \mu \\ \mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}(X_j) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}.\end{aligned}$$

So the expectation of the sample mean is always μ: we call it an *unbiased estimator* for the mean. On the other hand, the variance is always smaller than σ^2, and decreases as we increase n. By taking a large enough sample size, we can get as small a variance as we want.

If n is large enough, the sample mean will give an accurate estimate for the true mean μ. This result is called the *Law of Large Numbers*, which says that $\bar{X}$ converges to μ as $n \to \infty$. (The word "converges" here is hiding quite a lot of probability theory).

7.7.3 The Central Limit Theorem

We know that the sample mean will be quite close to the true value μ on average. The Central Limit Theorem tells us more about the distribution of the error.

Assumption: We assume that $X_1, X_2, \ldots, X_n$ are i.i.d. with shared mean μ and variance σ^2.

Then the sample mean $\bar{X}$ is approximately Normally-distributed with mean μ and variance $\frac{\sigma^2}{n}$.

In other words, the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

has approximately a Standard Normal distribution.

Here, when we say that the distribution is approximately Normal, we mean that

$$\mathbb{P}(a \le \bar{X} \le b) \approx \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right),$$

whatever the values of a and b.

Example 7.7.3.1

- *If the random variables* $X_1, X_2, \dots, X_10$ *are independent, and all are uniformly distributed on the interval* $[0, 1]$*, use the Central Limit Theorem to estimate* $\mathbb{P}(\sum_{j=1}^{10} X_j > 7)$.
- *A manufacturing process is designed to produce bolts with a 0.5cm diameter. Once a day, a random sample of 36 bolts is selected and the diameters recorded. If the average of the 36 values is less than 0.49cm or greater than 0.51cm, then the process is shut down for inspection and adjustment. The standard deviation for individual diameters is 0.02cm. Find approximately the probability that the line will be shut down unnecessarily (i.e., if the true process mean really is 0.5cm).*

Suggested exercises: Q46–Q50.