Why study probability?

- Probability began from the study of gambling and games of chance.
- It took hundreds of years to be placed on a completely rigorous footing.
- Now probability is used to analyse physical systems, model financial markets, study algorithms etc.
- The world is full of randomness and uncertainty: we need to understand it!
Course outline

- 22+2 lectures, 12 exercise classes, 11 mandatory HW sets.
- HW on two weeks is assessed and counts 10% towards final mark.
- Notes have gaps, model solutions will only be handed out on paper:
  **IT IS YOUR RESPONSIBILITY TO ATTEND LECTURES AND TO ENSURE YOU HAVE A FULL SET OF NOTES AND SOLUTIONS**

- **Course webpage** for notes, problem sheets, links etc:
  https://people.maths.bris.ac.uk/~maotj/prob.html

- **Drop-in sessions**: 1pm-2pm Mondays. Just turn up to Room 3.17 in these times. (Other times, I may be out or busy - but just email maotj@bristol.ac.uk to fix an appointment).

- This material is copyright of the University unless explicitly stated otherwise. It is provided exclusively for educational purposes at the University and is to be downloaded or copied for your private study only.

Contents

1. Introduction
2. Section 1: Elementary probability
3. Section 2: Counting arguments
4. Section 3: Conditional probability
5. Section 4: Discrete random variables
6. Section 5: Expectation and variance
7. Section 6: Joint distributions
8. Section 7: Properties of mean and variance
9. Section 8: Continuous random variables I
10. Section 9: Continuous random variables II
11. Section 10: Conditional expectation
12. Section 11: Moment generating functions
The recommended textbook for the unit is: 
*A First Course in Probability* by S. Ross.

Copies are available in the Queens Building library.

Section 1: Elementary probability

**Objectives:** by the end of this section you should be able to
- Define events and sample spaces, describe them in simple examples
- Describe combinations of events using set-theoretic notation
- List the axioms of probability
- State and use simple results such as inclusion–exclusion and de Morgan’s Law
Section 1.1: Random events

[This material is also covered in Sections 2.1 and 2.2 of the course book]

**Definition 1.1.**
- **Random experiment or trial.** Examples:
  - spin of a roulette wheel
  - throw of a dice
  - London stock market running for a day
- A **sample point or elementary outcome** $\omega$ is the result of a trial:
  - the number to come up on the roulette wheel
  - the number on the dice
  - the observed position of the stock market at the end of the day
- The **sample space** $\Omega$ is the set of all possible elementary outcomes $\omega$.

---

**Red and green dice**

**Example 1.2.**
- Consider the experiment of throwing a red die and a green die.
- Represent an elementary outcome as a pair, such as
  \[
  \omega = (6, 3)
  \]
  where the first number is the score on the red die and the second number is the score on the green die.
- Then the sample space
  \[
  \Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}
  \]
  has 36 sample points.

Note we use set notation: this will be key for us.
**Events**

**Definition 1.3.**
- An *event* is a set of outcomes specified by some condition.
- Note that events are subsets of the sample space, denoted $A \subseteq \Omega$.
- We say that event $A$ *occurs* if the elementary outcome of the trial lies in the set $A$, denoted $\omega \in A$.

**Example 1.4.**

In the red and green dice example, Example 1.2, let $A$ be the event that the sum of the scores is 5:

$$A = \{(1,4), (2,3), (3,2), (4,1)\}.$$

**Two special cases**

**Remark 1.5.**

*There are two special events:*
- $A = \emptyset$, the empty set. *This event never occurs, since we can never have* $\omega \in \emptyset$.
- $A = \Omega$, the whole sample space. *This event always occurs, since we always have* $\omega \in \Omega$. 
Combining events.

Given two events $A$ and $B$, we can combine them together, using standard set notation.

<table>
<thead>
<tr>
<th>Informal description</th>
<th>Formal description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ occurs or $B$ occurs (or both)</td>
<td>$A \cup B$</td>
</tr>
<tr>
<td>$A$ and $B$ both occur</td>
<td>$A \cap B$</td>
</tr>
<tr>
<td>$A$ does not occur</td>
<td>$A^c$</td>
</tr>
<tr>
<td>$A$ occurs implies $B$ occurs</td>
<td>$A \subseteq B$</td>
</tr>
<tr>
<td>$A$ and $B$ cannot both occur together</td>
<td>$A \cap B = \emptyset$</td>
</tr>
<tr>
<td>(disjoint or mutually exclusive)</td>
<td></td>
</tr>
</tbody>
</table>

You may find it useful to represent combinations of events using Venn diagrams.

Example: draw a lottery ball.

**Example 1.6.**

- The trial is to draw one ball in the lottery.
- An elementary outcome is $\omega = k$, where $k$ is the observed number drawn. $\Omega = \{1, 2, \ldots, 59\}$.

**Example events**

<table>
<thead>
<tr>
<th>Event</th>
<th>Informal description</th>
<th>Formal description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$</td>
<td>an odd number is drawn</td>
<td>$A = {1, 3, 5, \ldots, 59}$</td>
</tr>
<tr>
<td>$B$</td>
<td>an even number is drawn</td>
<td>$B = {2, 4, 6, \ldots, 58}$</td>
</tr>
<tr>
<td>$C$</td>
<td>number is divisible by 3</td>
<td>$C = {3, 6, 9, \ldots, 57}$</td>
</tr>
</tbody>
</table>

- We then have

<table>
<thead>
<tr>
<th>Event</th>
<th>Informal description</th>
<th>Formal description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ or $C$</td>
<td>odd or divisible by 3</td>
<td>$A \cup C = {1, 3, 5, 6, 7, 9, \ldots, 57, 59}$</td>
</tr>
<tr>
<td>$B$ and $C$</td>
<td>div by 2 and 3</td>
<td>$B \cap C = {6, 12, \ldots, 54}$</td>
</tr>
</tbody>
</table>
Section 1.2: Axioms of probability

[This material is also covered in Section 2.3 of the course book.]

- We have an intuitive idea that some events are more likely than others.
- Tossing a head is more likely than winning the lottery.
- The probability \( P \) captures this.

Axioms of probability

**Definition 1.7.**

- Suppose we have a sample space \( \Omega \).
- Let \( P \) be a map from events \( A \subseteq \Omega \) to the real numbers \( \mathbb{R} \).
- For each event \( A \) (each subset of \( \Omega \)) there is a number \( P(A) \).
- Then \( P \) is a *probability measure* if it satisfies:
  
  **Axiom 1** \( 0 \leq P(A) \leq 1 \) for every event \( A \).
  
  **Axiom 2** \( P(\Omega) = 1 \).
  
  **Axiom 3** Let \( A_1, A_2, \ldots \) be an infinite collection of disjoint events (so \( A_i \cap A_j = \emptyset \) for all \( i \neq j \)). Then
  
  \[
  P \left( \bigcup_{i=1}^{\infty} A_i \right) = P(A_1) + P(A_2) + \cdots = \sum_{i=1}^{\infty} P(A_i).
  \]
Deductions from the axioms

- These three axioms form the basis of all probability theory.
- We will develop the consequences of these axioms as a rigorous mathematical theory, using only logic.
- We show that it matches our intuition for how we expect probability to behave.

**Example 1.8.**

From these 3 axioms it is simple to prove the following:

**Property 1** \( P(\emptyset) = 0 \)

**Property 2** For a finite collection of disjoint events \( A_1, \ldots, A_n \),

\[
P \left( \bigcup_{i=1}^{n} A_i \right) = P(A_1) + P(A_2) + \cdots + P(A_n) = \sum_{i=1}^{n} P(A_i).
\]

Property 2 follows from Axiom 3 by taking \( A_i = \emptyset \) for \( i \geq n + 1 \).

Motivating the axioms

- Suppose we have a sample space \( \Omega \), and an event \( A \subseteq \Omega \).
- For example, if we roll a 6-sided dice then \( \Omega = \{1, 2, \ldots, 6\} \), and consider the event \( A = \{6\} \) (the roll gives a 6).
- Suppose the trial is repeated infinitely many times, and let \( a_n \) be the number of times \( A \) occurs in the first \( n \) trials.
- We might expect \( \frac{a_n}{n} \) to converge to a limit as \( n \to \infty \), which provisionally we might call \( \text{Prob}(A) \).
- Assuming it does converge, it is clear that
  - \( 0 \leq \text{Prob}(A) \leq 1 \)
  - \( \text{Prob}(\Omega) = 1 \)
  - \( \text{Prob}(\emptyset) = 0 \)
Motivating the axioms (cont.)

- Furthermore, if $A$ and $B$ are disjoint events, and $C = A \cup B$, for the first $n$ trials let
  - $a_n$ be the number of times $A$ occurs
  - $b_n$ be the number of times $B$ occurs
  - $c_n$ be the number of times $C$ occurs
- Then $c_n = a_n + b_n$ since $A$ and $B$ are disjoint.
- Therefore
\[
\frac{a_n}{n} + \frac{b_n}{n} = \frac{c_n}{n}.
\]
- Taking the limit as $n \to \infty$ we see that
\[
\text{Prob}(A) + \text{Prob}(B) = \text{Prob}(C) = \text{Prob}(A \cup B).
\]

Remark 1.9.
- However, how can we know if these limits exist?
- It seems intuitively reasonable, but we can’t be sure.
- The modern theory of probability works the other way round: we simply assume that for each event $A$ there exists a number $\mathbb{P}(A)$, where $\mathbb{P}$ is assumed to obey the axioms of Definition 1.7.
Section 1.3: Some simple applications of the axioms

[This material is also covered in Section 2.4 of the course book.]

**Lemma 1.10.**

For any event $A$, the complement satisfies $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

**Proof.**

- By definition, $A$ and $A^c$ are disjoint events: that is $A \cap A^c = \emptyset$.
- Further, $\Omega = A \cup A^c$, so $\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$ by Property 2.
- But $\mathbb{P}(\Omega) = 1$, by Axiom 2. So $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$.

---

**Lemma 1.11.**

Let $A \subseteq B$. Then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

**Proof.**

- We can write $B = A \cup (B \cap A^c)$, and $A \cap (B \cap A^c) = \emptyset$.
- That is, $A$ and $B \cap A^c$ are disjoint events.
- Hence by Property 2 we have $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$.
- But by Axiom 1 we have $\mathbb{P}(B \cap A^c) \geq 0$, so $\mathbb{P}(B) \geq \mathbb{P}(A)$. 

---
Lemma 1.12.

Let $A$ and $B$ be any two events. Then

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

Proof.

- $A \cup B = A \cup (B \cap A^c)$ is a disjoint union, so
  $$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \quad \text{(Property 2).} \quad (1.1)$$

- $B = (B \cap A) \cup (B \cap A^c)$ is a disjoint union, so
  $$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) \quad \text{(Property 2).} \quad (1.2)$$

Subtracting (1.2) from (1.1) we have

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

General inclusion–exclusion principle

Theorem 1.13.

For $n$ events $A_1, \ldots, A_n$, we can write

$$\mathbb{P}\left( \bigcup_{i=1}^{n} A_i \right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{S:|S|=k} \mathbb{P}\left( \bigcap_{j \in S} A_j \right),$$

where the sum over sets $S$ covers all subsets of $\{1, 2, \ldots, n\}$ of size $k$.

Proof.

Not proved here.
### Example 1.14.

- In the sports club, 
  36 members play tennis, 22 play tennis and squash, 
  28 play squash, 12 play tennis and badminton, 
  18 play badminton, 9 play squash and badminton, 
  4 play tennis, squash and badminton.
- How many play at least one of these games?
- Introduce probability by picking a random member out of those \( N \) enrolled to the club. Then

\[
T := \{ \text{that person plays tennis} \}, \\
S := \{ \text{that person plays squash} \}, \\
B := \{ \text{that person plays badminton} \}.
\]

\[
\begin{align*}
\text{Our answer is therefore } & 43 \text{ members.} \\
\text{Note that in some sense, we don’t need probability here – there is a} \\
\text{version of inclusion–exclusion just for set sizes.}
\end{align*}
\]
Boole’s inequality – ‘union bound’

**Proposition 1.15 (Boole’s inequality).**

For any events $E_1, E_2, \ldots, E_n$,

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

**Boole’s inequality proof**

**Proof.**

Proof by induction. When $n = 2$, by Lemma 1.12:

$$
\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2) \leq \mathbb{P}(E_1) + \mathbb{P}(E_2).
$$

Now suppose true for $n$. Then

$$
\mathbb{P}\left(\bigcup_{i=1}^{n+1} E_i\right) = \mathbb{P}\left(\left(\bigcup_{i=1}^{n} E_i\right) \cup E_{n+1}\right) \leq \mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) + \mathbb{P}(E_{n+1})
$$

$$
\leq \sum_{i=1}^{n} \mathbb{P}(E_i) + \mathbb{P}(E_{n+1}) = \sum_{i=1}^{n+1} \mathbb{P}(E_i).
$$
Key idea: de Morgan’s Law

**Theorem 1.16.** For any events $A$ and $B$:

\[
(A \cup B)^c = A^c \cap B^c \quad (1.3)
\]

\[
(A \cap B)^c = A^c \cup B^c \quad (1.4)
\]

**Proof.**

Draw a Venn diagram. Note that (swapping $A$ and $A^c$, and swapping $B$ and $B^c$), (1.3) and (1.4) are equivalent.

**Remark 1.17.**

- (1.3) ‘Neither $A$ nor $B$ happens’ same as ‘$A$ doesn’t happen and $B$ doesn’t happen’.
- (1.4) ‘$A$ and $B$ don’t both happen’ same as ‘either $A$ doesn’t happen, or $B$ doesn’t’

Key idea: de Morgan’s Law

**Theorem 1.18.** For any events $A$ and $B$:

\[
1 - \mathbb{P}(A \cup B) = \mathbb{P}(A^c \cap B^c) \quad (1.5)
\]

\[
1 - \mathbb{P}(A \cap B) = \mathbb{P}(A^c \cup B^c) \quad (1.6)
\]

**Proof.**

- Since the events on either side of (1.3) are the same, they must have the same probability.
- Further, we know from Lemma 1.10 that $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$ for any event $E$.
- A similar argument applies to (1.4).

- By a similar argument, can extend (1.5) and (1.6) to collections of $n$ events.
Example 1.19.

- Return to Example 1.2: suppose we roll a red die and a green die.
- What is the probability that we roll a 6 on at least one of them?
- Write $A = \{\text{roll a 6 on red die}\}$, $B = \{\text{roll a 6 on green die}\}$.
- Event ‘roll a 6 on at least one’ is $A \cup B$.
- Hence by (1.5),

$$P(A \cup B) = 1 - P(A^c \cap B^c) = 1 - \frac{5}{6} \cdot \frac{5}{6} = \frac{11}{36},$$

since $P(A^c \cap B^c) = P(A^c)P(B^c) = (1 - P(A))(1 - P(B))$

- Caution: This final step only works because two rolls are ‘independent’ (see later for much more on this!!)

Section 2: Counting arguments

Objectives: by the end of this section you should be able to

- Understand how to calculate probabilities when there are equally likely outcomes
- Describe outcomes in the language of combinations and permutations
- Count these outcomes using factorial notation
Section 2.1: Equally likely sample points

[This material is also covered in Section 2.5 of the course book]

- A common case that arises is where each of the sample points has the same probability.
- This can often be justified by physical symmetry and the use of Property 2.
- For example, think of rolling a dice.
- There are 6 disjoint sample outcomes, and symmetry says they have equal probability.
- Hence Property 2 tells us that \( P(i) = \frac{1}{6} \) for \( i = 1, \ldots, 6 \).
- In such cases calculating probability reduces to a counting problem.

More formally

- Assume that
  - \( \Omega \), the sample space, is finite
  - all sample points are equally likely
- Then by Axiom 2 and Property 2, we can see that
  \[
P(\{\omega\}) = \frac{1}{\text{Number of points in } \Omega} = \frac{1}{|\Omega|}.
\]
- Also, if \( A \subseteq \Omega \), then
  \[
P(A) = \frac{\text{Number of points in } A}{\text{Number of points in } \Omega} = \frac{|A|}{|\Omega|}.
\]
- Hence, we need to learn to count!
Example: red and green dice.

**Example 2.1.**

- A red die and a green die are rolled.
- A sample point is $\omega = (r, g)$ where $r$ is the score on the red dice and $g$ is the score on the green dice.

$$
\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}
$$

with 36 sample points.
- By arguments like the above, assume that $P(\{\omega\}) = \frac{1}{36}$ for each $\omega$ (i.e. equally likely outcomes).

Example: red and green dice. (cont)

**Example 2.1.**

- Let $A_5$ be the event that the sum of the scores is 5:

$$
A_5 = \{(1, 4), (2, 3), (3, 2), (4, 1)\} = \{(1, 4)\} \cup \{(2, 3)\} \cup \{(3, 2)\} \cup \{(4, 1)\}
$$

a disjoint union.
- Then

$$
\mathbb{P}(A_5) = \mathbb{P}(\{(1, 4)\}) + \mathbb{P}(\{(2, 3)\}) + \mathbb{P}(\{(3, 2)\}) + \mathbb{P}(\{(4, 1)\}) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{9}.
$$

**Exercise:** For each $i$, let $A_i$ be the event that the sum of the scores is $i$. Show that

$$
\mathbb{P}(A_4) = \frac{1}{12}, \quad \mathbb{P}(A_3) = \frac{1}{18}, \quad \mathbb{P}(A_2) = \frac{1}{36}.
$$
Example 2.1.

- Let $B$ be the event that the sum of the scores is less than or equal to 5. Then
  \[ B = A_2 \cup A_3 \cup A_4 \cup A_5 \]
  a disjoint union. So
  \[ P(B) = P(A_2) + P(A_3) + P(A_4) + P(A_5) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{10}{36} = \frac{5}{18} \]

- Let $C$ be the event that the sum of the scores is greater than 5. We could calculate $P(A_6), \ldots, P(A_{12})$. But it’s easier to spot that $C = B^c$, so
  \[ P(C) = 1 - P(B) = 1 - \frac{5}{18} = \frac{13}{18}. \]

Example: tossing a coin to obtain a head.

Example 2.2.

- A fair coin is tossed repeatedly until the first head is obtained.
- We will find the probability that an odd number of tosses is required.
- For this random experiment a sample point is just a positive integer corresponding to the total number of throws required to get the first head. Thus
  \[ \Omega = \{1, 2, 3, 4, \ldots\} \]

- Let $A_k$ denote the event that exactly $k$ tosses are required, i.e.
  \[ A_k = \{k\}. \]

- Assume that $P(A_k) = \frac{1}{2^k}$ (will see how to justify this properly later).
Example: tossing a coin to obtain a head (cont).

Example 2.2.

- Let $B$ denote the event that the number of tosses required is odd. Thus
  \[ B = A_1 \cup A_3 \cup A_5 \cup \ldots. \]
- Since the $A_k$’s are disjoint
  \[ P(B) = P(A_1) + P(A_3) + P(A_5) + \ldots. \]
- Thus
  \[ P(B) = \frac{1}{2} + \frac{1}{2^3} + \frac{1}{2^5} + \ldots. \]
- Hence
  \[ P(B) = \frac{1}{2} \left[ 1 + \frac{1}{4} + \frac{1}{4^2} + \ldots \right] = \frac{1}{2} \left[ \frac{1}{(1 - \frac{1}{4})} \right] = \frac{2}{3}. \]

Section 2.2: Permutations and combinations

[This material is also covered by Sections 1.1 - 1.4 of the course book.]

Definition 2.3.

A permutation is a selection of $r$ objects from $n \geq r$ objects when the ordering matters.
Example 2.4.

Eight swimmers in a race, how many different ways of allocating the three medals are there?

- First place can be chosen in 8 ways.
- For each winner of the gold medal, the silver medal can go to one of the other 7 swimmers, so there are $8 \times 7$ different options for gold and silver.
- For each choice of first and second place, the bronze medal can go to one of the other 6 swimmers, so there are $8 \times 7 \times 6$ different ways the medals can be handed out.

General theory

Lemma 2.5.

- In general there are $n_P_r = n(n-1)(n-2) \cdots (n-r+1)$ different ways.
- Note that we can write $n_P_r = \frac{n!}{(n-r)!}$.
- General convention: $0! = 1$

Remark 2.6.

Check the special cases:

- $r = n$: $n_P_n = \frac{n!}{(n-n)!} = \frac{n!}{1} = n!$, so there are $n!$ ways of ordering $n$ objects.
- $r = 1$: $n_P_1 = \frac{n!}{(n-1)!} = n$, so there are $n$ ways of choosing 1 of $n$ objects.
BRISTOL and BANANA examples

Example 2.7.
- How many four letter 'words' can be formed from distinct letters of the word BRISTOL? (i.e. no letter used more than once)
- There are 7 distinct letters in BRISTOL. Hence the first letter can be chosen in 7 ways, the second in 6 ways, etc.
- There are then $7 \times 6 \times 5 \times 4 = 840 = \frac{7!}{(7-4)!}$ words.

Example 2.8.
- In how many ways can the letters of the word BANANA be rearranged to produce distinct 6-letter “words”?
- There are 6! orderings of the letters of the word BANANA.
- But can order the 3 As in 3! ways, and order two Ns in 2! ways. So each word is produced by 3! × 2! orderings of the letters.
- So the total number of distinct words is
  $$\frac{6!}{3!2!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1} = \frac{6 \times 5 \times 4}{2} = 60.$$  

Combinations

Definition 2.9.
A combination is a selection of $r$ objects from $n \geq r$ objects when the order is not important.

Example 2.10.
Eight swimmers in a club, how many different ways are there to select a team of three of them?
- We saw before that there are $8 \times 7 \times 6$ ways to choose 3 people in order.
- The actual ordering is unimportant in terms of who gets in the team.
- Each team could be formed from $3! = 6$ different allocations of the medals.
- So the number of distinct teams is $\frac{8 \times 7 \times 6}{6}$. 
General result

**Lemma 2.11.**

- More generally, think about choosing \( r \) where the order is important: this can be done in \( \binom{n^r}{(n-r)!} \) different ways.
- But \( r! \) of these ways result in the same set of \( r \) objects, since ordering is not important.
- Therefore the \( r \) objects can be chosen in 
  \[
  \frac{n^r}{r!} = \frac{n!}{(n - r)!r!}
  \]
  different ways if order doesn’t matter.
- We write this binomial coefficient as 
  \[
  \binom{n}{r} = \frac{n!}{(n - r)!r!}.
  \]
- At school many of you will have used \( \binom{n}{r} \). Please use this new notation from now onwards.

Example

**Example 2.12.**

- How many hands of 5 can be dealt from a pack of 52 cards?
- Note that the order in which you are dealt the cards is assumed to be unimportant here.
- Thus there are 
  \[
  \binom{52}{5} = \frac{52!}{47!5!} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1}
  \]
  distinct hands.
Properties of binomial coefficients

**Proposition 2.13.**

1. For any $n$ and $r$:
   \[
   \binom{n}{r} = \binom{n}{n-r}.
   \]

2. **[Pascal’s Identity]** For any $n$ and $k$:
   \[
   \binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.
   \]

3. **[Binomial theorem]** For any real $a$, $b$:
   \[
   (a + b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}.
   \]

4. For any $n$, we know: $2^n = \sum_{r=0}^{n} \binom{n}{r}$.

**Proof.**

1. Choosing $r$ objects to be included is the same as choosing $(n - r)$ objects to be excluded.

2. Consider choosing $k$ objects out of $n$, and imagine painting one object red. Either
   - the red object is chosen, and the remaining $k - 1$ objects need to be picked out of $n - 1$, or
   - the red object is not chosen, and all $k$ objects need to be picked out of $n - 1$.

3. Write $(a + b)^n = (a + b)(a + b) \cdots (a + b)$ and imagine writing out the expansion. You choose an $a$ or a $b$ from each term of the product, so to get $a^r b^{n-r}$ you need to choose $r$ brackets to take an $a$ from (and $n - r$ to take a $b$ from). There are $\binom{n}{r}$ ways to do this.

4. Simply take $a = b = 1$ in 3.
Remark 2.14.

- We can generalize the binomial theorem to obtain the multinomial theorem.
- We can define the multinomial coefficient

\[
\binom{n}{n_1, n_2, \ldots, n_r} := \frac{n!}{n_1! \cdot n_2! \cdots n_r!},
\]

where \( n = n_1 + \cdots + n_r \)

- Check this is the number of permutations of \( n \) objects, \( n_1 \) of which are of type 1, \( n_2 \) of type 2 etc.
- For any real \( x_1, x_2, \ldots, x_r \):

\[
(x_1 + x_2 + \cdots + x_r)^n = \sum_{n_1, \ldots, n_r \geq 0 \atop n_1 + \cdots + n_r = n} \binom{n}{n_1 n_2 \ldots n_r} \cdot x_1^{n_1} \cdot x_2^{n_2} \cdots x_r^{n_r}.
\]

Section 2.3: Counting examples

[This material is also covered in Section 2.5 of the course book.]

Example 2.15.

- A fair coin is tossed \( n \) times.
- Represent the outcome of the experiment by, e.g.

\( (H, T, T, \ldots, H, T) \).

- \( \Omega = \{(s_1, s_2, \ldots, s_n) : s_i = H \text{ or } T, i = 1, \ldots, n\} \) so that \( |\Omega| = 2^n \).
- If the coin is fair and tosses are independent then all \( 2^n \) outcomes are equally likely.
- Let \( A_r \) be the event “there are exactly \( r \) heads”.
- Each element of \( A_r \) is a sample point \( \omega = (s_1, s_2, \ldots, s_n) \) with exactly \( r \) of the \( s_i \) being a head.
- There are \( \binom{n}{r} \) different ways to choose the \( r \) elements of \( \omega \) to be a head, so \( |A_r| = \binom{n}{r} \).
Example 2.15.

- Therefore $\Pr(\text{Exactly } r \text{ heads}) = \Pr(A_r) = \binom{n}{r} 2^{-n}$.
- Take $n = 5$, so that $|\Omega| = 2^5 = 32$.

<table>
<thead>
<tr>
<th>$r$</th>
<th>$A_r$</th>
<th>$\Pr(A_r)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>$\frac{1}{32}$</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>$\frac{5}{32}$</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>$\frac{10}{32}$</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>$\frac{10}{32}$</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>$\frac{5}{32}$</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>$\frac{1}{32}$</td>
</tr>
</tbody>
</table>

- This peaks in the middle because e.g. there are more ways to have 2 heads than 0 heads.

Remark 2.16.

- **Note:** $A_0, A_1, \ldots, A_5$ are disjoint, and also $\bigcup_{r=1}^{5} A_r = \Omega$, so we know that $\sum_{r=0}^{5} \Pr(A_r) = 1$.
- **This can easily be verified for the case $n = 5$ from the numbers above.**
- **More importantly we can show this generally.**
- **Since** $\Pr(A_r) = \binom{n}{r} 2^{-n}$, so

$$
\sum_{r=0}^{n} \Pr(A_r) = \sum_{r=0}^{n} \frac{\binom{n}{r}}{2^n} = \frac{1}{2^n} \sum_{r=0}^{n} \binom{n}{r} = \frac{1}{2^n} 2^n = 1,
$$

using the Binomial Theorem, Proposition 2.13.3.
Example 2.17.

- The experiment is to deal a bridge hand of 13 cards from a pack of 52.
- What is the probability of being dealt the JQKA of spades?
- A sample point is a set of 13 cards (order not important).
- Hence the number of sample points is the number of ways of choosing 13 cards from 52, i.e. \( \binom{52}{13} \).
- We assume these are equally likely.

\[
P(\text{JQKA spades}) = \frac{\binom{48}{9}}{\binom{52}{13}} = \frac{\frac{48!}{9!39!}}{\frac{52!}{13!39!}} = \frac{48!13!}{52!9!}\]
\[
= \frac{13 \times 12 \times 11 \times 10}{52 \times 51 \times 50 \times 49}
= \frac{17160}{6497400}
\approx 0.00264.
\]
Example: Birthdays

Example 2.18.

- There are $m$ people in a room.
- What is the probability that no two of them share a birthday?
- Label the people 1 to $m$. Let the $i$th person have a birthday on day $a_i$, and assume $1 \leq a_i \leq 365$.
- The $n$-tuple $(a_1, a_2, \ldots, a_m)$ specifies everyone’s birthday.
- So

$$\Omega = \{(a_1, a_2, \ldots, a_m) : a_i = 1, 2, \ldots, 365, i = 1, 2, \ldots, m\}$$

and $|\Omega| = 365^m$.
- Let $B_m$ be the event “no 2 people share the same birthday”.
- An element of $B_m$ is a point $(a_1, \ldots, a_m)$ with each $a_i$ different.

Example 2.18.

- Need to choose $m$ birthdays out of the 365 days, and ordering is important. (If Alice’s birthday is 1 Jan and Bob’s is 2 Jan, that is a different sample point to if Alice’s is 2 Jan and Bob’s is 1 Jan.)
- So

$$|B_m| = \binom{365}{m} = \frac{365!}{(365 - m)!}$$

$$P(B_m) = \frac{|B_m|}{|\Omega|} = \frac{365!}{365^m(365 - m)!}.$$ 

- Not easy to calculate directly.
Example 2.18.

- Use Stirling’s formula.
  \[ n! \approx \sqrt{2\pi n^{n+\frac{1}{2}}} e^{-n}. \]

\[ \mathbb{P}(B_m) \approx e^{-m} \left( \frac{365}{365 - m} \right)^{365.5 - m} \]

For example,

\[ \mathbb{P}(B_{23}) \approx 0.493 \]
\[ \mathbb{P}(B_{40}) \approx 0.109 \]
\[ \mathbb{P}(B_{60}) \approx 0.006 \]

Example: fixed points in random permutations

Can combine all these kinds of arguments together:

Example 2.19.

- \( n \) friends go clubbing, and check their coats in on arrival.
- When they leave, they are given a coat at random.
- What is the probability that none of them get the right coat?
- Write \( A_i = \{ \text{person } i \text{ is given the right coat} \} \).
- By de Morgan (1.5) extended to \( n \) sets:

\[ \mathbb{P}(\text{nobody has right coat}) = \mathbb{P}\left( \bigcap_{i=1}^{n} A_i^c \right) = 1 - \mathbb{P}\left( \bigcup_{i=1}^{n} A_i \right). \]
Example 2.19.

- Apply inclusion–exclusion, Theorem 1.13.

\[
P(\text{nobody has right coat}) = 1 - \sum_{k=1}^{n} (-1)^{k+1} \sum_{S: |S|=k} P \left( \bigcap_{j \in S} A_j \right)
\] (2.1)

- Consider a set \( S \) of size \( k \): the event \( \bigcap_{j \in S} A_j \) means that a specific set of \( k \) people get the right coat. The remainder may or may not.

- This can happen in \((n-k)!\) ways, so

\[
P \left( \bigcap_{j \in S} A_j \right) = \frac{(n-k)!}{n!}.
\]

Example: fixed points in random permutations (cont.)

Example 2.19.

- Since there are \( \binom{n}{k} \) sets of size \( k \), in (2.1) this gives

\[
1 - \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} = 1 + \sum_{k=1}^{n} (-1)^{k} \frac{1}{k!}
\]

\[
= \sum_{k=0}^{n} \frac{(-1)^{k}}{k!}.
\]

- Notice that as \( n \to \infty \), this tends to

\[
\sum_{k=0}^{\infty} \frac{(-1)^{k}}{k!} = e^{-1} \approx 0.3679\ldots
\]
Section 3: Conditional probability

Objectives: by the end of this section you should be able to
- Define and understand conditional probability.
- State and prove the partition theorem and Bayes' theorem
- Put these results together to calculate probability values
- Understand the concept of independence of events

[This material is also covered in Sections 3.1 - 3.3 of the course book.]

Section 3.1: Motivation and definitions

- An experiment is performed, and two events are of interest.
- Suppose we know $B$ has occurred.
- What information does this give us about whether $A$ occurred in the same experiment?

Remark 3.1.
- Intuition: repeat the experiment infinitely often.
- $B$ occurs a proportion $\mathbb{P}(B)$ of the time.
- $A$ and $B$ occur together a proportion $\mathbb{P}(A \cap B)$ of the time.
- So when $B$ occurs, $A$ also occurs a proportion

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

of the time.
Conditional probability

This motivates the following definition.

**Definition 3.2.**

Let $A$ and $B$ be events, with $\mathbb{P}(B) > 0$. The *conditional probability of $A$ given $B*$, denoted $\mathbb{P}(A \mid B)$, is defined as

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

**Example: Sex of children**

**Example 3.3.**

- Choose a family at random from the set of all families with two children
- Given the family has at least one boy, what is the probability that the other child is also a boy?
- Assume equally likely sample points:
  $$\Omega = \{(b, b), (b, g), (g, b), (g, g)\}.$$ 

  $$A = \{(b, b)\} = \text{“both boys”}$$ 
  $$B = \{(b, b), (b, g), (g, b)\} = \text{“at least one boy”}$$ 
  $$A \cap B = \{(b, b)\}$$ 
  $$\mathbb{P}(A \cap B) = \frac{1}{4}$$ 
  $$\mathbb{P}(B) = \frac{3}{4}$$ 
  $$\mathbb{P}(A \mid B) = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$
Section 3.2: Reduced sample space

- A good way to understand this is via the idea of a reduced sample space.

**Example 3.4.**

- Return to the red and green dice, Example 1.2.
- Suppose I tell you that the sum of the dice is 5: what is the probability the red dice scored 2?
- Write $A = \{\text{red dice scored 2}\}$ and $B = \{\text{sum of dice is 5}\}$.
- Remember from Example 2.1 that $P(B) = \frac{4}{36}$.
- Clearly $A \cap B = \{(2, 3)\}$, so $P(A \cap B) = \frac{1}{36}$.
- Hence

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1/36}{4/36} = \frac{1}{4}.$$  

Reduced sample space

**Example 3.4.**

- When we started in Example 1.2, our sample space was

$$\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\},$$

with 36 sample points.
- However, learning that $B$ occurred means that we can rule out a lot of these possibilities.
- We have reduced our world to the event

$$B = \{(1, 4), (2, 3), (3, 2), (4, 1)\}.$$  
- Conditioning on $B$ means that we just treat $B$ as our sample space and proceed as before.
- The set $B$ is a reduced sample space.
- We simply work in this set to figure out the conditional probabilities given this event.
Conditional probabilities are well-behaved

**Proposition 3.5.**

For a fixed $B$, the conditional probability $\mathbb{P}(\cdot \mid B)$ is a probability measure (it satisfies the axioms):

1. the conditional probability of any event $A$ satisfies $0 \leq \mathbb{P}(A \mid B) \leq 1$, 
2. the conditional probability of the sample space is one: $\mathbb{P}(\Omega \mid B) = 1$, 
3. for any finitely or countably infinitely many disjoint events $A_1, A_2, \ldots$, 

$$
\mathbb{P}\left(\bigcup_i A_i \mid B\right) = \sum_i \mathbb{P}(A_i \mid B).
$$

**Sketch proofs**

1. By Axiom 1 and Lemma 1.11, we know that $0 \leq \mathbb{P}(A \cap B) \leq \mathbb{P}(B)$, and dividing through by $\mathbb{P}(B)$ the result follows.
2. Since $\Omega \cap B = B$, we know that $\mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = \mathbb{P}(B)/\mathbb{P}(B) = 1$.
3. Applying Axiom 3 to the (disjoint) events $A_i \cap B$, we know that

$$
\mathbb{P}\left(\left(\bigcup_i A_i\right) \cap B\right) = \mathbb{P}\left(\bigcup_i (A_i \cap B)\right) = \sum_i \mathbb{P}(A_i \cap B),
$$

and again the result follows on dividing by $\mathbb{P}(B)$. 
Deductions from the axioms

- Since (for fixed $B$) Proposition 3.5 shows that $P(\cdot | B)$ is a probability measure, all the results we deduced in Chapter 1 continue to hold true.
- This is a good advert for the axiomatic method.

**Corollary 3.6.**

For example
- $P(A^c | B) = 1 - P(A | B)$.
- $P(\emptyset | B) = 0$.
- $P(A \cup C | B) = P(A | B) + P(C | B) - P(A \cap C | B)$.

**Remark 3.7.**

*WARNING: DON’T CHANGE THE CONDITIONING:* e.g. $P(A | B)$ and $P(A | B^c)$ have nothing to do with each other.

---

Section 3.3: Partition theorem

**Definition 3.8.**

A collection of events $B_1, B_2, \ldots, B_n$ is a disjoint partition of $\Omega$, if
- $B_i \cap B_j = \emptyset$ if $i \neq j$, and
- $\bigcup_{i=1}^{n} B_n = \Omega$.

In other words, the collection is a disjoint partition iff every sample point lies in exactly one of the events.

**Theorem 3.9 (Partition Theorem).**

Let $A$ be an event. Let $B_1, B_2, \ldots, B_n$ be a disjoint partition of $\Omega$ with $P(B_n) > 0$ for all $n$. Then

$$P(A) = \sum_{i=1}^{n} P(A | B_i)P(B_i).$$
Proof of Partition theorem, Theorem 3.9

Proof.

1. Write $C_i = A \cap B_i$.
2. Then for $i \neq j$ the $C_i \cap C_j = (A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = \emptyset$.
3. Also $\bigcup_{i=1}^n C_i = \bigcup_{i=1}^n (A \cap B_i) = A \cap (\bigcup_{i=1}^n B_i) = A \cap \Omega = A$.
4. So $\mathbb{P}(A) = \mathbb{P}(\bigcup_{i=1}^n C_i) = \sum_{i=1}^n \mathbb{P}(C_i)$ since the $C_i$ are disjoint.
5. Note: In the proof of Lemma 1.12, we saw that $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$, which is exactly the same thing as we’re seeing here.
6. But $\mathbb{P}(C_i) = \mathbb{P}(A \cap B_i) = \mathbb{P}(A | B_i) \mathbb{P}(B_i)$ by the definition of conditional probability, so

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

Example: Diagnostic test

Example 3.10.

1. A test for a disease gives positive results 90% of the time when a disease is present, and 20% of the time when the disease is absent.
2. It is known that 1% of the population have the disease.
3. In a randomly selected member of the population, what is the probability of getting a positive test result?
4. Let $B_1$ be the event “has disease”: $\mathbb{P}(B_1) = 0.01$.
5. Let $B_2 = B_1^c$ be the event “no disease”: $\mathbb{P}(B_2) = 0.99$.
6. Let $A$ be the event “positive test result”.
7. We are told: $\mathbb{P}(A | B_1) = 0.9$ $\mathbb{P}(A | B_2) = 0.2$.
8. Therefore

$$\mathbb{P}(A) = \sum_{i=1}^2 \mathbb{P}(A | B_i) \mathbb{P}(B_i) = 0.9 \times 0.01 + 0.2 \times 0.99 = 0.207.$$
Important advice

**Remark 3.11.**
- With questions of this kind, always important to be methodical.
- Write a list of named events.
- Write down probabilities (conditional or not)?
- Will get a lot of credit in exam for just that step.
- Seems too obvious to bother with, but leaving it out can lead to serious confusion.
- Obviously need to do final calculation as well.

Section 3.4: Bayes’ theorem

- We have seen in Definition 3.2 that $P(A \cap B) = P(A | B)P(B)$.
- We also have $P(A \cap B) = P(B \cap A) = P(B | A)P(A)$.
- So $P(A | B)P(B) = P(B | A)P(A)$ and therefore

**Theorem 3.12 (Bayes’ theorem).**

$$P(B | A) = \frac{P(A | B)P(B)}{P(A)}.$$  (3.1)

- This very simple observation forms the basis of large parts of modern statistics.
- If $A$ is an observed event, and $B$ is some hypothesis about how the observation was generated, it allows us to switch

$$P(\text{observation} | \text{hypothesis}) \leftrightarrow P(\text{hypothesis} | \text{observation}).$$
Alternative form of Bayes’

**Theorem 3.13 (Bayes’ theorem – partition form).**

Let $A$ be an event, and let $B_1, B_2, \ldots, B_n$ be a partition of $\Omega$. Then for any $k$:

$$
P(B_k \mid A) = \frac{P(A \mid B_k)P(B_k)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)}.
$$

**Proof.**

- We have already seen in (3.1) that

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

- The partition theorem (Theorem 3.9) tells us that

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

- The result follows immediately.

**Example: Diagnostic test revisited**

- In Example 3.10, the observation is the positive test result, and the hypothesis is that you have the disease.

**Example 3.14.**

- Return to the setting of Example 3.10
- A person receives a positive test result. What is the probability they have the disease?
- $A$ is the event “positive test result” and $B_1$ is the event “has disease”.
- Use the formulation (3.1), since we already know $P(A) = 0.207$.
- So $P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{P(A)} = \frac{0.9 \times 0.01}{0.207} = 0.0435$ (3.s.f.)
Example: Prosecutor’s fallacy

Example 3.15.
- A crime is committed, and some DNA evidence is discovered.
- The DNA is compared with the national database and a match is found.
- In court, the prosecutor tells the jury that the probability of seeing this match if the suspect is innocent is 1 in 1,000,000.
- How strong is the evidence that the suspect is guilty?
- Let $E$ be the event that the DNA evidence from the crime scene matches that of the suspect.
- Let $G$ be the event that the suspect is guilty.

\[
P(E \mid G) = 1, \quad P(E \mid G^c) = 10^{-6}.
\]

Oliver Johnson (maotj@bris.ac.uk)

Example 3.15.
- We want to know $P(G \mid E)$, so use Bayes' theorem.
- We need to know $P(G)$.
- Suppose that only very vague extra information is known about the suspect, so there is a pool of $10^7$ equally likely suspects, except for the DNA data: $P(G) = 10^{-7}$.
- Hence

\[
P(G \mid E) = \frac{P(E \mid G)P(G)}{P(E \mid G)P(G) + P(E \mid G^c)P(G^c)} = \frac{1 \times 10^{-7}}{1 \times 10^{-7} + 10^{-6} \times (1 - 10^{-7})} \approx \frac{1}{11}.
\]

This is a much lower probability of guilt than you might think, given the DNA evidence.

Oliver Johnson (maotj@bris.ac.uk)
Section 3.5: Independence of events

Motivation: Events are independent if the occurrence of one does not affect the occurrence of the other i.e.

\[ P(A \mid B) = P(A) \iff \frac{P(A \cap B)}{P(B)} = P(A) \iff P(A \cap B) = P(A)P(B). \]

**Definition 3.16.**

Two events \( A \) and \( B \) are independent if and only if \( P(A \cap B) = P(A)P(B) \).

**Definition 3.17.**

Events \( A_1, \ldots, A_n \) are independent if and only if

\[ P(A_s \cap \cdots \cap A_t) = P(A_s) \cdots P(A_t) \]

for any subset \( \{s, \ldots, t\} \subseteq \{1, \ldots, n\} \).

**Example 3.18.**

- Throw a fair dice repeatedly, with the throws independent.
- What is \( P(1\text{st six occurs on 4th throw})? \)
- Let \( A_i \) be the event that a 6 is thrown on the \( i \)th throw of the dice.
  - \[
    \{ \text{1st six occurs on 4th throw} \} = \{ \text{1st throw not 6 AND 2nd throw not 6 AND 3rd throw not 6 AND 4th throw 6} \} = A_1^c \cap A_2^c \cap A_3^c \cap A_4.
    
    By independence,
    
    \[
    P(A_1^c \cap A_2^c \cap A_3^c \cap A_4) = P(A_1^c)P(A_2^c)P(A_3^c)P(A_4) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{5^3}{6^4}.\n    
    \]
Lemma 3.19.

Chain rule / Multiplication rule

For any two events $A$ and $B$ with $P(B) > 0$,

$$P(A \cap B) = P(A \mid B)P(B).$$

More generally, if $A_1, \ldots, A_n$ are events with $P(A_1 \cap \cdots \cap A_{n-1}) > 0$, then

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

(3.2)

Chain rule (proof)

Proof.

To ease notation, let $B_i = A_1 \cap A_2 \cap \cdots \cap A_i$. Note that $B_1 \supseteq B_2 \supseteq \cdots \supseteq B_n$.

We can write the RHS of (3.2) as

$$P(B_1)P(A_2 \mid B_1)P(A_3 \mid B_2) \cdots P(A_n \mid B_{n-1}).$$

But $A_{i+1} \cap B_i = B_{i+1}$, so by definition:

$$P(A_{i+1} \mid B_i) = \frac{P(A_{i+1} \cap B_i)}{P(B_i)} = \frac{P(B_{i+1})}{P(B_i)}.$$

Hence the RHS of (3.2) is equal to

$$P(B_1)\frac{P(B_2)}{P(B_1)} \frac{P(B_3)}{P(B_2)} \cdots \frac{P(B_n)}{P(B_{n-1})} = P(B_n),$$

as required.
Example 3.20.

- You are dealt 13 cards at random from a pack of cards.
- What is the probability that you are dealt a JQKA of spades? Let
  - $A_1 = \text{"dealt J spades"}$
  - $A_2 = \text{"dealt Q spades"}$
  - $A_3 = \text{"dealt K spades"}$
  - $A_4 = \text{"dealt A spades"}$
- Note $P(A_1) = P(A_2) = P(A_3) = P(A_4) = \frac{1}{4}$, but these events are not independent.

\[
\begin{align*}
\mathbb{P}(A_2 \mid A_1) &= \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} \\
&= \frac{\binom{50}{11} \cdot \binom{52}{13}}{\binom{51}{12} \cdot \binom{52}{13}} \\
&= \frac{12}{51}.
\end{align*}
\]

- This is not equal to $\mathbb{P}(A_2) = \frac{1}{4}$.
- Similarly $\mathbb{P}(A_3 \mid A_1 \cap A_2) = \frac{11}{50}$ and $\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) = \frac{10}{49}$.

\[
\begin{align*}
\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) &= \mathbb{P}(A_1) \mathbb{P}(A_2 \mid A_1) \mathbb{P}(A_3 \mid A_1 \cap A_2) \mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) \\
&= \frac{13}{52} \frac{12}{51} \frac{11}{50} \frac{10}{49}.
\end{align*}
\]

- NB same result as before.
Section 4: Discrete random variables

Objectives: by the end of this section you should be able to
- To build a mathematical model for discrete random variables
- To understand the probability mass function, expectation and variance of such variables
- To get experience in working with some of the basic distributions (Bernoulli, Binomial, Poisson, Geometric)

[The material for this Section is also covered in Chapter 4 of the course book.]

Section 4.1: Motivation and definitions

- A trial selects an outcome \( \omega \) from a sample space \( \Omega \).
- Often we are interested in a number associated with the outcome \( \omega \).

Example 4.1.
- Let \( \omega \) be the positions of the stock market at the end of one day.
- You are interested only in the amount of money you make.
- That is a number that depends on the outcome \( \omega \).

Example 4.2.
- Throw two fair dice. Look at the total score.
- Let \( X(\omega) \) be the total score when the outcome is \( \omega \).
- Remember we write the sample space as
  \[ \Omega = \{(a, b) : a, b = 1, \ldots, 6\}. \]
- So \( X((a, b)) = a + b \).
Formal definition

**Definition 4.3.**
- Let $\Omega$ be a sample space.
- A random variable $X$ is a function $X : \Omega \to \mathbb{R}$.
- That is, $X$ assigns a value $X(\omega)$ to each outcome $\omega$.

**Remark 4.4.**
- For any set $B$, we use the notation $P(X \in B)$ as shorthand for $P(\{\omega \in \Omega : X(\omega) \in B\})$.
- E.g. $X$ is the sum of the scores of two fair dice, $P(X \leq 3)$ is shorthand for $P(\{\omega \in \Omega : X(\omega) \leq 3\}) = P(\{(1,1),(1,2),(2,1)\}) = \frac{3}{36}$.

Probability mass functions

- In this chapter we look at discrete random variables $X$, which are those such that $X(\omega)$ takes a discrete set of values $S = \{x_1, x_2, \ldots\}$.
- This avoids certain technicalities we will worry about in due course.

**Definition 4.5.**
- Let $X$ be a discrete r.v. taking values in $S = \{x_1, x_2, \ldots\}$.
- The probability mass function (pmf) of $X$ is the function $p_X$ given by $p_X(x) = P(X = x)$.

**Remark 4.6.**
If $p_X$ is a p.m.f. then
- $0 \leq p_X(x) \leq 1$ for all $x$
- $\sum_{x \in S} p_X(x) = 1$ (since $P(\Omega) = 1$).
In fact, any function with these properties can be thought of as a pmf of some random variable.
Example 4.7.

$X$ is the sum of the scores on 2 fair dice

\[
\begin{align*}
x & = 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad \ldots \\
|\{\omega : X(\omega) = x\}| & = 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad \ldots \\
p_X(x) & = \frac{1}{36} \quad \frac{2}{36} \quad \frac{3}{36} \quad \frac{4}{36} \quad \frac{5}{36} \quad \frac{6}{36} \quad \ldots \\
x & = 8 \quad 9 \quad 10 \quad 11 \quad 12 \\
|\{\omega : X(\omega) = x\}| & = 5 \quad 4 \quad 3 \quad 2 \quad 1 \\
p_X(x) & = \frac{5}{36} \quad \frac{4}{36} \quad \frac{3}{36} \quad \frac{2}{36} \quad \frac{1}{36}
\end{align*}
\]

Section 4.2: Bernoulli distribution

- This is the building block for many distributions.
- Named after Jakob Bernoulli, part of a very famous family of mathematicians.

Definition 4.8.

- Think of a trial with two outcomes: success or failure.
  \[\Omega = \{\text{success, failure}\}\]
- This is called a Bernoulli trial.
- Let $X(\text{failure}) = 0$ and $X(\text{success}) = 1$, so that $X$ counts the number of successes in the trial.
- Suppose that $P(\{\text{success}\}) = p$, so that $P(\{\text{failure}\}) = 1 - P(\{\text{success}\}) = 1 - p$.
- Then we say that $X$ has a Bernoulli distribution with parameter $p$. 
Remark 4.9.

- **Notation**: $X \sim \text{Bernoulli}(p)$
- $X$ has pmf $p_X(0) = 1 - p$, $p_X(1) = p$, $p_X(x) = 0$ for $x \notin \{0, 1\}$.
- Equivalently, $p_X(x) = (1 - p)^{1-x} p^x$ for $x = 0, 1$.

Example: Indicator functions

Example 4.10.

- Let $A$ be an event, and let
  
  $$I(\omega) = \begin{cases} 
  1 & \omega \in A \\
  0 & \omega \notin A 
  \end{cases}$$

- $I$ is called the *indicator function* of $A$.

  - $\mathbb{P}(I = 1) = \mathbb{P}(\{\omega : I(\omega) = 1\}) = \mathbb{P}(A)$
  - $\mathbb{P}(I = 0) = \mathbb{P}(\{\omega : I(\omega) = 0\}) = \mathbb{P}(A^c)$

- That is $p_I(1) = \mathbb{P}(A)$ and $p_I(0) = 1 - \mathbb{P}(A)$.
- Thus $I \sim \text{Bernoulli}(\mathbb{P}(A))$. 
Section 4.3: Binomial distribution

**Definition 4.11.**
- Consider $n$ independent Bernoulli trials
- Each trial has probability $p$ of success
- Let $T$ be the total number of successes
- Then $T$ is said to have a *binomial distribution with parameters* $(n, p)$
- Notation: $T \sim \text{Bin}(n, p)$.

Binomial distribution example

**Example 4.12.**
- Take $n = 3$ trials with $p = \frac{1}{3}$
- $\Omega = \{FFF, FFS, FSF, SFF, FSS, SFS, SSF, SSS\}$

\[
\begin{align*}
\mathbb{P}\{\text{FFF}\} &= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{2}{3} = \frac{8}{27} \\
\mathbb{P}\{\text{FFS}\} = \mathbb{P}\{\text{FSF}\} = \mathbb{P}\{\text{SFF}\} &= \frac{2}{3} \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{27} \\
\mathbb{P}\{\text{FSS}\} = \mathbb{P}\{\text{SFS}\} = \mathbb{P}\{\text{SSF}\} &= \frac{2}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{2}{27} \\
\mathbb{P}\{\text{SSS}\} &= \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{1}{3} = \frac{1}{27}
\end{align*}
\]
Example 4.12.

- Hence
  - \( \{ T = 0 \} = \{ FFF \} \) so that \( P(T = 0) = \frac{8}{27} \)
  - \( \{ T = 1 \} = \{ FFS, FSF, SFF \} \) so that \( P(T = 1) = 3 \times \frac{4}{27} = \frac{12}{27} \)
  - \( \{ T = 2 \} = \{ FSS, SFS, SSF \} \) so that \( P(T = 2) = 3 \times \frac{2}{27} = \frac{6}{27} \)
  - \( \{ T = 3 \} = \{ SSS \} \) so that \( P(T = 3) = \frac{1}{27} \)

- Thus \( T \) has pmf

\[
p_T(0) = \frac{8}{27}, \quad p_T(1) = \frac{12}{27}, \quad p_T(2) = \frac{6}{27}, \quad p_T(3) = \frac{1}{27}
\]

with \( p_T(x) = 0 \) otherwise.

General binomial distribution pmf

Lemma 4.13.

In general if \( T \sim Bin(n, p) \) then claim that

\[
p_T(x) = P(T = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n.
\]

Proof.

- There are \( \binom{n}{x} \) sample points with \( x \) successes from the \( n \) trials.
- Each of these sample points has probability \( p^x (1 - p)^{n-x} \).

Exercise Verify that \( \sum_{x=0}^{n} p_X(x) = 1 \) in this case (Hint: use Proposition 2.13.3).
Example 4.14.

- 40% of a large population vote Labour.
- A random sample of 10 people is taken.
- What is the probability that not more than 2 people vote Labour?
- Let $T$ be the number of people that vote Labour. So $T \sim \text{Bin}(10, 0.4)$.

\[
P(T \leq 2) = p_T(0) + p_T(1) + p_T(2) = \binom{10}{0}(0.4)^0(0.6)^{10} + \binom{10}{1}(0.4)^1(0.6)^9 + \binom{10}{2}(0.4)^2(0.6)^8
\]
\[
= 0.167
\]

Section 4.4: Geometric distribution

Definition 4.15.

- Carry out independent Bernoulli trials until we obtain first success.
- Let $X$ be the number of the trial when we see the first success.
- Suppose the probability of a success on any one trial is $p$, then

\[
P(X = x) = (1 - p)^{x-1}p, \quad x = 1, 2, 3, \ldots
\]

- Hence the mass function is

\[
p_X(x) = P(X = x) = p(1 - p)^{x-1}, \quad x = 1, 2, 3, \ldots
\]

with $p_X(x) = 0$ otherwise.
- $X$ is said to be a geometric distribution with parameter $p$
- Notation: $X \sim \text{Geom}(p)$

Exercise: Verify that $\sum_{x=1}^{\infty} p_X(x) = 1$. 
Example 4.16.

- Consider a call-centre which has 10 incoming phone lines.
- Each time an operative is free, they answer a random line.
- Let $X$ be the number of people served (up to and including yourself) from the time that you get through.
- Each time the operative serves someone there is a probability $\frac{1}{10}$ that it will be you.
- So $X \sim \text{Geom}(\frac{1}{10})$.


\[
\begin{array}{cccccccc}
 x & = & 1 & 2 & 3 & 4 & 5 & 6 & \cdots \\
 \mathbb{P}(X = x) & = & 0.1 & 0.09 & 0.081 & 0.0729 & 0.06561 & 0.059049 & \cdots
\end{array}
\]

Geometric tail distribution

Lemma 4.17.

If $X \sim \text{Geom}(p)$ then $\mathbb{P}(X > x) = (1 - p)^x$.

Proof.

Write $q = 1 - p$. Then

\[
\begin{align*}
\mathbb{P}(X > x) &= \mathbb{P}(X = x + 1) + \mathbb{P}(X = x + 2) + \mathbb{P}(X = x + 3) + \cdots \\
&= pq^x + pq^{x+1} + pq^{x+2} + \cdots \\
&= pq^x(1 + q + q^2 + \cdots) \\
&= pq^x \frac{1}{1 - q} \\
&= pq^x \frac{1}{p} \\
&= q^x
\end{align*}
\]

by summing a geometric progression to infinity.
Remark 4.18.

Lemma 4.17 is easily seen by thinking about waiting for successes: the probability of waiting more than $x$ for a success is the probability that you get failures on the first $x$ trials, which has probability $(1 - p)^x$.

- If waiting at the call-centre (Example 4.16),
  \[ P(X > 10) = 0.9^{10} = 0.349\text{ (s.f.)} \]

Lack-of-memory property

Lemma 4.19.

Lack of memory property

\[ P(X = x + n \mid X > n) = P(X = x). \]

Remark 4.20.

- In the call-centre example (Example 4.16) this tells us for example that
  \[ P(X = 5 + x \mid X > 5) = P(X = x). \]
- The fact that you have waited for 5 other people to get served doesn’t mean you are more likely to get served quickly than if you have just joined the queue.
Section 4.5: Poisson distribution

**Definition 4.21.**

- Let \( \lambda > 0 \) be a real number.
- A r.v. \( X \) has a Poisson distribution with parameter \( \lambda \) if \( X \) takes values in the range 0, 1, 2, \ldots and has pmf
  \[
  p_X(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots
  \]
- Notation: \( X \sim \text{Poi}(\lambda) \).

**Exercise:** verify \( \sum_{x=0}^{\infty} p_X(x) = 1 \).

**Two motivations**

**Remark 4.22.**

*If \( X \sim \text{Bin}(n, p) \) with \( n \) large and \( p \) small then*

\[
\mathbb{P}(X = x) \approx e^{-np} \frac{(np)^x}{x!}
\]

*i.e. \( X \) is distributed approximately the same as a \( \text{Poi}(\lambda) \) random variable where \( \lambda = np \).*

**Remark 4.23.**

*In the second year Probability 2 course you can see that the Poisson distribution is a natural distribution for the number of arrivals of something in a given time period: telephone calls, internet traffic, disease incidences, nuclear particles.*
Example: airline tickets

Example 4.24.
- An airline sells 403 tickets for a flight with 400 seats.
- On average 1% of purchasers fail to turn up.
- What is the probability that there are more passengers than seats (someone is bumped)?
- Let \( X \) = number of purchasers that fail to turn up.
- True distribution \( X \sim \text{Bin}(403, 0.01) \)
- Approximately \( X \sim \text{Poi}(4.03) \)

Example: airline tickets (cont.)

Example 4.24.
- \( \mathbb{P}(X = x) \approx e^{-4.03} \frac{4.03^x}{x!} \)
  
  \[
  \begin{array}{cccccc}
  x & 0 & 1 & 2 & 3 & 4 & 5 \\
  \mathbb{P}(X = x) & 0.0178 & 0.0716 & 0.144 & 0.1939 & 0.1953 & 0.1574 \\
  \end{array}
  \]
- We can deduce that
  
  \[
  \mathbb{P}(\text{at least one passenger bumped}) = \mathbb{P}(X \leq 2) = px(0) + px(1) + px(2) \\
  \approx 0.2334.
  \]
Section 5: Expectation and variance

Objectives: by the end of this section you should be able to

- To understand where random variables are centred and how dispersed they are
- To understand basic properties of mean and variance
- To use results such as Chebyshev’s theorem to bound probabilities

[The material for this Section is also covered in Chapter 4 of the course book.]

Section 5.1: Expectation

- We want some concept of the average value of a r.v. $X$ and the spread about this average.
- Some of this will seem slightly arbitrary for now, but we will prove some results that help motivate why it is convenient to make these definitions after we have introduced jointly-distributed r.v.s.

Definition 5.1.

- Let $X$ be a random variable taking the values in a discrete set $S$.
- The expected value (or expectation) of $X$, denoted $\mathbb{E}(X)$, is defined as

$$
\mathbb{E}(X) = \sum_{x \in S} xp_X(x).
$$

- This is well-defined so long as $\sum_{x \in S} |x|p_X(x)$ converges.
Example

**Remark 5.2.**

- $\mathbb{E}(X)$ is also sometimes called the mean of the distribution of $X$.

**Example 5.3.**

- Consider a Bernoulli random variable.
- Recall from Remark 4.9 that if $X \sim \text{Bernoulli}(p)$ then $X$ has pmf $p_X(0) = 1 - p$, $p_X(1) = p$, $p_X(x) = 0$ for $x \notin \{0, 1\}$.
- Hence in Definition 5.1

  $$
  \mathbb{E}(X) = 0(1 - p) + 1 \cdot p = p.
  $$

- Note that in general the random variable $X$ won’t be equal to $\mathbb{E}X$.

Motivation

**Remark 5.4.**

- Do not confuse $\mathbb{E}(X)$ with the the mean of a collection of observed values, which is referred to as the sample mean.
- However, there is a relationship between $\mathbb{E}(X)$ and sample mean which motivates the definition.
- Perform an experiment and observe the random variable $X$ which takes values in the discrete set $S$.
- Repeat the experiment infinitely often, and observe outcomes $X_1$, $X_2$, $\ldots$
- Consider the limit of the sample means

  $$
  \lim_{n \to \infty} \frac{X_1 + \cdots + X_n}{n}.
  $$
Remark 5.5.

- Let \( a(x, n) \) be the number of times the outcome is \( x \) in the first \( n \) trials. Then reordering the sum, we know that

\[
X_1 + X_2 + \cdots + X_n = \sum_{x \in S} xa(x, n).
\]

- We expect (but have not yet proved) that

\[
\frac{a(x, n)}{n} \to p_X(x) \quad \text{as} \quad n \to \infty.
\]

- If so then

\[
\frac{X_1 + \cdots + X_n}{n} = \frac{\sum_{x \in S} xa(x, n)}{n} = \sum_{x \in S} \frac{a(x, n)}{n} \to \sum_{x \in S} xp_X(x).
\]

This motivates Definition 5.1.

Examples

Example 5.6.

- Let \( X \) take the integer values \( 1, \ldots, n \).

\[
p_X(x) = \begin{cases} 
\frac{1}{n} & x = 1, \ldots, n \\
0 & \text{otherwise}
\end{cases}
\]

- 

\[
\mathbb{E}(X) = \sum_{x=1}^{n} x \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x = \frac{1}{n} \frac{1}{2} n(n + 1) = \frac{n + 1}{2}.
\]

- Hence for example if \( n = 6 \), the expected value of a dice roll is \( 7/2 \).
Example: binomial distribution

Example 5.7.

- \( X \sim \text{Bin}(n, p) \) (see Definition 4.11).
- \( P(X = x) = \left\{ \begin{array}{ll} \binom{n}{x} p^x (1 - p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{array} \right. \)
- \( \mathbb{E}(X) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1 - p)^{n-x} \)
  \[ = np \sum_{x=0}^{n} \binom{n-1}{x-1} p^{x-1} (1 - p)^{(n-1)-(x-1)} \]
  \[ = np. \]

Here we use the fact that \( x\binom{n}{x} = n \binom{n-1}{x-1} \) (check directly?) and apply the Binomial Theorem 2.13.3.
- There are easier ways — see later.

Example: Poisson distribution

Example 5.8.

- \( X \sim \text{Poi}(\lambda) \) (see Definition 4.21).
- \( P(X = x) = \left\{ \begin{array}{ll} e^{-\lambda} \frac{\lambda^x}{x!} & x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{array} \right. \)
- \( \mathbb{E}(X) = \sum_{x=0}^{\infty} xe^{-\lambda} \frac{\lambda^x}{x!} \)
  \[ = \sum_{x=1}^{\infty} xe^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^x}{(x-1)!} \]
  \[ = e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y+1}}{y!} = \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{\lambda^{y}}{y!} = \lambda e^{-\lambda} e^\lambda. \]
- So \( \mathbb{E}(X) = \lambda. \)
Example: geometric distribution

Example 5.9.

- $X \sim \text{Geom}(p)$ (see Definition 4.15).
- Recall that $\mathbb{P}(X = x) = (1 - p)^{x-1}p$, so that
  
  $$
  \mathbb{E}(X) = \sum_{x=1}^{\infty} (1 - p)^{x-1}px
  = p \sum_{x=1}^{\infty} (1 - p)^{x-1}x
  = p \frac{1}{(1 - (1 - p))^2} = \frac{1}{p}.
  $$

- Here we use the standard result that $\sum_{x=1}^{\infty} t^{x-1}x = 1/(1 - t)^2$ (differentiate sum of geometric progression?)

Section 5.2: Expectation of a function of a r.v.

- Consider a random variable taking values $x_1, x_2, \ldots$
- Take a function $g : \mathbb{R} \to \mathbb{R}$.
- Define a new r.v. $Z(\omega) = g(X(\omega))$.
- Then $Z$ takes values in the range $z_1 = g(x_1), z_2 = g(z_2), \ldots$
- Look at $\mathbb{E}(Z)$.
- By definition $\mathbb{E}(Z) = \sum_i z_i p_Z(z_i)$ where $p_Z$ is the pmf of $Z$ which we could in principle work out.
- But it’s often easier to use

Theorem 5.10.

Let $Z = g(X)$. Then

$$
\mathbb{E}(Z) = \sum_i g(x_i)p_X(x_i).
$$
Proof.

Recall that \( p_Z(z_i) = \mathbb{P}(Z = z_i) = \mathbb{P}(\{\omega \in \Omega : Z(\omega) = z_i\}) \).

Notice that \( \{\omega \in \Omega : Z(\omega) = z_i\} = \bigcup_{j : g(x_j) = z_i} \{\omega : X(\omega) = x_j\} \),

which is a disjoint union.

So \( p_Z(z_i) = \sum_{j : g(x_j) = z_i} p_X(x_j) \).

Therefore

\[
\mathbb{E}(Z) = \sum_i z_i p_Z(z_i) = \sum_i z_i \left( \sum_{j : g(x_j) = z_i} p_X(x_j) \right) = \sum_i \left( \sum_{j : g(x_j) = z_i} g(x_j) p_X(x_j) \right) = \sum_j g(x_j) p_X(x_j).
\]

Example 5.11.

\[ p_X(x) = \begin{cases} \frac{1}{n} & x = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases} \]

Consider \( Z = X^2 \) so \( Z \) takes the values 1, 4, 9, \ldots, \( n^2 \) each with probability \( \frac{1}{n} \). We have \( g(x) = x^2 \).

By Theorem 5.10

\[
\mathbb{E}(Z) = \sum_{x=0}^{n} g(x) p_X(x) = \sum_{x=0}^{n} x^2 \frac{1}{n} = \frac{1}{n} \sum_{x=0}^{n} x^2 = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} = \frac{1}{6}(n+1)(2n+1)
\]
Linearity of expectation

**Lemma 5.12.**

Let $a$ and $b$ be constants. Then $E(aX + b) = aE(X) + b$.

**Proof.**

Let $g(x) = ax + b$. From Theorem 5.10 we know that

$$E(g(X)) = \sum_i g(x_i)p_X(x_i) = \sum_i (ax_i + b)p_X(x_i)$$

$$= a\sum_i x_ip_X(x_i) + b\sum_i p_X(x_i) = aE(X) + b.$$

Section 5.3: Variance

- This is the standard measure for the spread of a distribution.

**Definition 5.13.**

- Let $X$ be a r.v., and let $\mu = E(X)$.
- Define the **variance of $X$**, denoted by $\text{Var} (X)$, by

$$\text{Var} (X) = E \left( (X - \mu)^2 \right).$$

- Notation: $\text{Var} (X)$ is often denoted $\sigma^2$.
- The **standard deviation of $X$** is $\sqrt{\text{Var} (X)}$. 
Example of spread

Example 5.14.

- Define

\[
Y = \begin{cases} 
1, \text{ wp. } \frac{1}{2}, \\
-1, \text{ wp. } \frac{1}{2},
\end{cases}
\]

\[
Z = \begin{cases} 
2, \text{ wp. } \frac{1}{5}, \\
-\frac{1}{2}, \text{ wp. } \frac{4}{5},
\end{cases}
\]

\[
U = \begin{cases} 
10, \text{ wp. } \frac{1}{2}, \\
-10, \text{ wp. } \frac{1}{2}.
\end{cases}
\]

- Notice \( \mathbb{E}Y = \mathbb{E}Z = \mathbb{E}U = 0 \), the expectation does not distinguish between these rv.’s.

- Yet they are clearly different, and the variance captures this.

\[
\text{Var}(Y) = \mathbb{E}((Y - 0)^2) = 1^2 \cdot \frac{1}{2} + (-1)^2 \cdot \frac{1}{2} = 1,
\]

\[
\text{Var}(Z) = \mathbb{E}((Z - 0)^2) = 2^2 \cdot \frac{1}{5} + \left(-\frac{1}{2}\right)^2 \cdot \frac{4}{5} = 1,
\]

\[
\text{Var}(U) = \mathbb{E}((U - 0)^2) = 10^2 \cdot \frac{1}{2} + (-10)^2 \cdot \frac{1}{2} = 100.
\]

Useful lemma

Lemma 5.15.

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

Sketch proof: see Theorem 7.6.

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

\[
= \mathbb{E}(X^2 - 2\mu X + \mu^2)
\]

\[
= \mathbb{E}(X^2) - 2\mu \mathbb{E}(X) + \mu^2 \quad \text{ (will prove this step later)}
\]

\[
= \mathbb{E}(X^2) - 2\mu^2 + \mu^2
\]

\[
= \mathbb{E}(X^2) - \mu^2
\]
Example: Bernoulli random variable

Example 5.16.

- Recall from Remark 4.9 and Example 5.3 that if $X \sim \text{Bernoulli}(n, p)$ then $p_X(0) = 1 - p$, $p_X(1) = p$ and $\mathbb{E}X = p$.
- We can calculate $\text{Var}(X)$ in two different ways:
  1. $\text{Var}(X) = \mathbb{E}(X - \mu)^2 = \sum_x p_X(x)(x - \mu)^2 = (1 - p)(-p)^2 + p(1 - p)^2 = (1 - p)p(p + 1 - p) = p(1 - p)$.
  2. Alternatively:
     $$\mathbb{E}X^2 = \sum_x p_X(x)x^2 = (1 - p)0^2 + p1^2 = p,$$
     so that $\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = p - p^2$.

Remark 5.17.

We will see in Example 7.10 below that if $X \sim \text{Bin}(n, p)$ (see Definition 4.11) then $\text{Var}(X) = np(1 - p)$.

Example

Example 5.18.

- $$p_X(x) = \begin{cases} \frac{1}{n} & x = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

- Know from Example 5.6 that $\mathbb{E}(X) = \frac{n+1}{2}$ and from Example 5.11 that $\mathbb{E}(X^2) = \frac{1}{6}(n + 1)(2n + 1)$.

- Thus
  $$\text{Var}(X) = \frac{1}{6}n(2n + 1) - \left( \frac{n+1}{2} \right)^2 = \frac{n+1}{12} \left( 4n + 2 - 3(n + 1) \right)$$
  $$= \frac{(n+1)}{12}(n-1) = \frac{(n^2 - 1)}{12}.$$
Example 5.19.

Consider $X \sim \text{Poi}(\lambda)$ (see Definition 4.21).

Recall that $P(X = x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!} & x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$ and $E(X) = \lambda$.

We show (see next page) that $E(X^2) = \lambda^2 + \lambda$.

Thus $\text{Var}(X) = E(X^2) - (E(X))^2 = (\lambda^2 + \lambda) - (\lambda)^2 = \lambda$.

Example: Poisson (cont.)

Example 5.19.

\[
E(X^2) = \sum_{x=0}^{\infty} x^2 e^{-\lambda} \frac{\lambda^x}{x!} = e^{-\lambda} \sum_{x=1}^{\infty} \frac{x^2}{x!} \lambda^x
\]

\[
= e^{-\lambda} \sum_{x=1}^{\infty} \frac{x}{(x-1)!} \lambda^x = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{x}{(x-1)!} \lambda^{x-1}
\]

\[
= \lambda e^{-\lambda} \sum_{y=0}^{\infty} \frac{y+1}{y!} \lambda^y = \lambda e^{-\lambda} \left[ \sum_{y=0}^{\infty} \frac{y \lambda^y}{y!} + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right]
\]

\[
= \lambda e^{-\lambda} \left[ \lambda \sum_{z=0}^{\infty} \frac{\lambda^z}{z!} + \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right]
\]

\[
= \lambda e^{-\lambda} \left[ \lambda e^\lambda + e^\lambda \right] = \lambda^2 + \lambda
\]
Non-linearity of expectation

We now state (and prove later) an important result concerning variances, which is the counterpart of Lemma 5.12:

**Lemma 5.20.**

Let $a$ and $b$ be constants. Then $\text{Var}(aX + b) = a^2\text{Var}(X)$.

---

Section 5.4: Chebyshev’s inequality

- Let $X$ be any random variable with finite mean $\mu$ and variance $\sigma^2$, and let $c$ be any constant.
- Define the *indicator variable* $I(\omega) = \begin{cases} 1 & \text{if } |X(\omega) - \mu| > c \\ 0 & \text{otherwise} \end{cases}$
- Calculate $E(I) = 0\cdot P(I = 0) + 1\cdot P(I = 1) = P(I = 1) = P(|X - \mu| > c)$.
- Define also $Z(\omega) = \frac{(X(\omega) - \mu)^2}{c^2}$, so that
  $$E(Z) = E\left(\frac{(X - \mu)^2}{c^2}\right) = \frac{\text{Var}(X)}{c^2} = \frac{\sigma^2}{c^2}$$
- This last step uses Lemma 5.12 with $a = 1/c^2$ and $b = 0$.
- Notice that $I(\omega) \leq Z(\omega)$ for any $\omega$. (plot a graph?)
- So $E(I) \leq E(Z)$, and we deduce that . . .
**Theorem 5.21 (Chebyshev’s inequality).**

For any random variable $X$ with finite mean $\mu$ and variance $\sigma^2$, and any constant $c$:

$$P(|X - \mu| > c) \leq \frac{\sigma^2}{c^2}.$$ 

**Remark 5.22.**

- We have not made any assumptions about the distribution of $X$ (other than finite mean and variance).
- We have shown that the probability that $X$ is further than some constant from $\mu$ is bounded by some quantity that increases with the variance $\sigma^2$ and decreases with the distance from $\mu$.
- This shows that the axioms and definitions we have made give us something that at least fits with our intuition.

**Application of Chebyshev’s inequality**

**Example 5.23.**

- A fair coin is tossed $10^4$ times.
- Let $T$ denote the total number of heads.
- Then since $T \sim \text{Bin}(10^4, 0.5)$ we have $E(T) = 5000$ and $\text{Var}(T) = 2500$ (see Example 5.7 and Remark 5.17).
- Thus by taking $c = 500$ in Chebyshev’s inequality (Theorem 5.21) we have
  $$P(|T - 5000| > 500) \leq 0.01,$$
  so that
  $$P(4500 \leq T \leq 5500) \geq 0.99.$$
- We can also express this as
  $$P\left(0.45 \leq \frac{T}{10^4} \leq 0.55\right) \geq 0.99.$$
Objectives: by the end of this section you should be able to

- Understand the joint probability mass function
- Know how to use relationships between joint, marginal and conditional probability mass functions
- Use convolutions to calculate mass functions of sums.

[This material is also covered in Chapter 6 of the course book.]

Section 6.1: The joint probability mass function

- Often we want to measure two attributes, $X$ and $Y$, in the same experiment.
- For example
  - height $X$ and weight $Y$ of a randomly chosen person
  - the DNA profile $X$ and the cancer type $Y$ of a randomly chosen person.
- Up to now we have only been able to consider a single random variable at once, but clearly we need to be able to consider related random variables.
Joint probability mass function

- Recall that random variables are functions of the underlying outcome \( \omega \) in sample space \( \Omega \).
- Hence two random variables are simply two different functions of \( \omega \) in sample space.
- In particular, consider discrete random variables \( X, Y : \Omega \mapsto \mathbb{R} \).

**Definition 6.1.**

The *joint pmf* for \( X \) and \( Y \) is \( p_{X,Y} \), defined by

\[
p_{X,Y}(x,y) = P(X = x, Y = y) = P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\})
\]

- We can define the joint pmf of random variables \( X_1, \ldots, X_n \) in an analogous way.

Example: coin tosses

**Example 6.2.**

A fair coin is tossed 3 times. Let

- \( X = \) number of heads in first 2 tosses
- \( Y = \) number of heads in all 3 tosses

We can display the joint pmf in a table

<table>
<thead>
<tr>
<th>( p_{X,Y}(x,y) )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
<th>( y = 2 )</th>
<th>( y = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>( 0 )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>( 0 )</td>
<td>( \frac{1}{4} )</td>
<td>( \frac{1}{4} )</td>
<td>( 0 )</td>
</tr>
<tr>
<td>( x = 2 )</td>
<td>( 0 )</td>
<td>( 0 )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
</tr>
</tbody>
</table>
Section 6.2: Marginal pmfs

Continue the set-up from above: imagine we have two random variables $X$ and $Y$. Then:

**Definition 6.3.**
- The *marginal pmf* for $X$ is $p_X$, defined by
  \[ p_X(x) = \mathbb{P}(X = x) = \mathbb{P}\{\omega : X(\omega) = x\} \].
- Similarly the *marginal pmf* for $Y$ is $p_Y$, defined by
  \[ p_Y(y) = \mathbb{P}(Y = y) = \mathbb{P}\{\omega : Y(\omega) = y\} \].

Joint pmf determines the marginals
- Suppose $X$ takes values $x_1, x_2, \ldots$ and $Y$ takes values $y_1, y_2, \ldots$.
- Then for each $x_i$:
  \[ \{X = x_i\} = \bigcup_j \{X = x_i, Y = y_j\} \quad \text{(disjoint union)}, \]
  \[ \implies \mathbb{P}(X = x_i) = \sum_j \mathbb{P}(X = x_i, Y = y_j) \quad \text{(Axiom 3.)}. \]
- Hence (and with a corresponding argument for $\{Y = y_j\}$) we deduce that summing over the joint distribution determines the marginals:

**Theorem 6.4.**
For any random variables $X$ and $Y$:
\[ p_X(x_i) = \sum_j p_{X,Y}(x_i, y_j) \]
\[ p_Y(y_j) = \sum_i p_{X,Y}(x_i, y_j), \]
Example 6.5.

A fair coin is tossed 3 times. Let
- \( X \) = number of heads in first 2 tosses
- \( Y \) = number of heads in all 3 tosses

We can display the joint and marginal pmfs in a table:

<table>
<thead>
<tr>
<th>( p_{X,Y}(x,y) )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
<th>( y = 2 )</th>
<th>( y = 3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{1}{8} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>0</td>
<td>( \frac{1}{4} )</td>
<td>( \frac{1}{4} )</td>
<td>0</td>
</tr>
<tr>
<td>( x = 2 )</td>
<td>0</td>
<td>0</td>
<td>( \frac{1}{8} )</td>
<td>1</td>
</tr>
</tbody>
</table>

We calculate marginals for \( X \) by summing the rows of the table. We calculate marginals for \( Y \) by summing the columns.

Marginal pmfs don’t determine joint

Example 6.6.

- Consider tossing a fair coin once.
- Let \( X \) be the number of heads, and let \( Y \) be the number of tails.
- Write the joint pmf in a table:

<table>
<thead>
<tr>
<th>( p_{X,Y}(x,y) )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>0</td>
<td>( \frac{1}{2} )</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>( \frac{1}{2} )</td>
<td>0</td>
</tr>
</tbody>
</table>

- Either write down the marginals directly, or calculate

\[
p_X(0) = p_{X,Y}(0,0) + p_{X,Y}(0,1) = \frac{1}{2}, \quad p_X(1) = 1 - p_X(0) = \frac{1}{2}
\]

and similarly \( p_Y(0) = p_Y(1) = \frac{1}{2} \).
Example 6.7.

- Now toss a fair coin twice.
- Let $X$ is the number of heads on the first throw, and $Y$ be the number of tails on the second throw.
- Write the joint pmf in a table:

<table>
<thead>
<tr>
<th>$p_{X,Y}(x,y)$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 0$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
</tbody>
</table>

- Summing rows and columns we see that
  $p_X(0) = p_X(1) = p_Y(0) = p_Y(1) = \frac{1}{2}$, just as in Example 6.6.

Comparing Examples 6.6 and 6.7 we see that the marginal pmfs don’t determine the joint pmf.

Section 6.3: Conditional pmfs

Definition 6.8.

- The conditional pmf for $X$ given $Y = y$ is $p_{X|Y}$, defined by

$$p_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y).$$

(This is only well-defined for $y$ for which $\mathbb{P}(Y = y) > 0$.)

- Similarly the conditional pmf for $Y$ given $X = x$ is $p_{Y|X}$, defined by

$$p_{Y|X}(y|x) = \mathbb{P}(Y = y \mid X = x).$$
Calculating conditional pmfs

Remark 6.9.

- *Notice that*

\[
p_{X|Y}(x|y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} = \frac{p_{X,Y}(x, y)}{p_Y(y)}. \tag{6.1}
\]

- *Similarly*

\[
p_{Y|X}(y|x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.
\]

Conditional pmfs are probability mass functions

Remark 6.10.

- *We can check (in the spirit of Remark 4.6) that for any fixed y, the p_{X|Y}(\cdot|y) is a pmf.*

- *That is, for any x, since (6.1) expresses it as a ratio of probabilities, clearly p_{X|Y}(\cdot|y) \geq 0.*

- *Similarly using Theorem 6.4 we know that p_Y(y) = \sum_x p_{X,Y}(x, y).*

- *This means that (by (6.1))

\[
\sum_x p_{X|Y}(x|y) = \sum_x \frac{p_{X,Y}(x, y)}{p_Y(y)} = \frac{1}{p_Y(y)} \sum_x p_{X,Y}(x, y) = \frac{1}{p_Y(y)} p_Y(y) = 1,
\]

*as required.*
Example 6.11.

- Condition on $X = 2$:
  
  $$p_{Y|X}(y|2) = \frac{p_{X,Y}(2,y)}{p_X(2)} = 4p_{X,Y}(2,y)$$

  $y$ | 0 | 1 | 2 | 3
  --- | --- | --- | --- | ---
  $p_{Y|X}(y|2)$ | 0 | 0 | 4 | 2

- Condition on $Y = 1$

  $$p_{X|Y}(x|1) = \frac{p_{X,Y}(x,1)}{p_Y(1)} = \frac{8}{3}p_{X,Y}(x,1)$$

  $x$ | 0 | 1 | 2
  --- | --- | --- | ---
  $p_{X|Y}(x|1)$ | 1 | 4 | 0

Example: Inviting friends to the pub

Example 6.12.

- You decide to invite every friend you see today to the pub tonight.
- You have 3 friends (!)
- You will see each of them with probability $\frac{1}{2}$.
- Each invited friend will come with probability $\frac{2}{3}$ independently of the others.
- Find the distribution of the number of friends you meet in the pub.
- Let $X$ be the number of friends you invite.
  - $X \sim \text{Bin}(3, \frac{1}{2})$ so $p_X(x) = \binom{3}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{3-x} = \binom{3}{x} \frac{1}{8}$ for $0 \leq x \leq 3$.
- Let $Y$ be the number of friends who come to the pub.
  - $Y|X = x \sim \text{Bin}(x, \frac{2}{3})$ so $p_{Y|X}(y|x) = \binom{x}{y} \left(\frac{2}{3}\right)^y \left(\frac{1}{3}\right)^{x-y}$ for $0 \leq y \leq x$. 
Example: Inviting friends to the pub (cont.)

Example 6.12.

So \( p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y|x) \) = \[ \frac{1}{8} \binom{3}{x} \left( \frac{2^y}{3^y} \right) \]

\[
\begin{array}{cccc}
  & y = 0 & 1 & 2 & 3 \\
 x = 0 & \frac{1}{8} & 0 & 0 & 0 \\
 1 & \frac{3}{8} \times \frac{1}{3} = \frac{1}{8} & \frac{3}{8} \times \frac{2}{3} = \frac{1}{4} & 0 & 0 \\
 2 & \frac{3}{8} \times \frac{1}{9} = \frac{1}{24} & \frac{3}{8} \times \frac{4}{9} = \frac{1}{6} & \frac{3}{8} \times \frac{4}{9} = \frac{1}{6} & 0 \\
 3 & \frac{1}{18} = \frac{8}{27} & \frac{1}{18} = \frac{8}{27} & \frac{1}{18} = \frac{8}{27} & \frac{1}{27} \\
\end{array}
\]

Therefore \( E(Y) = 0 \times \frac{8}{27} + 1 \times \frac{12}{27} + 2 \times \frac{6}{27} + 3 \times \frac{1}{27} = 1 \). There is a much easier way to calculate \( E(Y) \) - see Chapter 10.

Section 6.4: Independent random variables

Definition 6.13.

- Two random variables are independent if
  \[ p_{X,Y}(x, y) = p_X(x)p_Y(y), \quad \text{for all } x \text{ and } y. \]

- In general, random variables \( X_1, \ldots, X_n \) are independent if
  \[ p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i), \quad \text{for all } x_i. \]
Properties of independent random variables

**Remark 6.14.**

- Consistent with Definition 3.16 (independence of events).
- We require that the events \( \{X = x\} \) and \( \{Y = y\} \) are independent for any \( x \) and \( y \).
- In fact this is equivalent to requiring events \( \{X \in A\} \) and \( \{Y \in B\} \) independent for any \( A \) and \( B \).
- Important: if \( X \) and \( Y \) are independent, so are \( g(X) \) and \( h(Y) \) for any functions \( g \) and \( h \).

**IID random variables**

**Definition 6.15.**

- We say that random variables \( X_1, \ldots, X_n \) are IID (independent and identically distributed) if they are independent, and all their marginals \( p_{X_i} \) are the same, so

\[
p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i),
\]

for some fixed \( p_X \).

- Here we obtain marginals \( p_{X_1}(x_1) = \sum_{x_2,\ldots,x_n} p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \) etc.
Example

Example 6.16.

- Again return to Example 1.2, rolling red and green dice.
- Let $X$ be the number on the red dice, $Y$ on the green dice.
- Then every pair of numbers have equal probability:
  \[ p_{X,Y}(x,y) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p_X(x) \cdot p_Y(y) \]  
  for all $x, y = 1, \ldots, 6$.
- We see that these variables are independent (in fact IID as well).

Discrete convolution

Proposition 6.17.

Let $X$ and $Y$ be independent, integer valued random variables with respective mass functions $p_X$ and $p_Y$. Then

\[ p_{X+Y}(k) = \sum_{i=-\infty}^{\infty} p_X(k - i) \cdot p_Y(i), \quad \text{for all } k \in \mathbb{Z}. \]

This formula is called the discrete convolution of the mass functions $p_X$ and $p_Y$. 
Discrete convolution proof

Proof.
Using independence we know that

\[ p_{X+Y}(k) = \mathbb{P}(X + Y = k) = \sum_{i=-\infty}^{\infty} \mathbb{P}(X + Y = k, Y = i) \]

\[ = \sum_{i=-\infty}^{\infty} \mathbb{P}(X = k - i, Y = i) = \sum_{i=-\infty}^{\infty} \mathbb{P}(X = k - i) \mathbb{P}(Y = i) \]

\[ = \sum_{i=-\infty}^{\infty} p_X(k - i) \cdot p_Y(i). \]

Convolution of Poissons gives a Poisson

Theorem 6.18.

- Recall the definition of the Poisson distribution from Definition 4.21.
- Let \( X \sim \text{Poi}(\lambda) \) and \( Y \sim \text{Poi}(\mu) \) be independent.
- Then \( X + Y \sim \text{Poi}(\lambda + \mu) \).
Proof of Theorem 6.18

Proof.

Using Proposition 6.17, since $X$ and $Y$ only take positive values we know

$$p_{X+Y}(k) = \sum_{i=0}^{k} p_X(k-i)p_Y(i)$$

$$= \sum_{i=0}^{k} \left(e^{-\lambda} \frac{\lambda^{k-i}}{(k-i)!} \right) \left(e^{-\mu} \frac{\mu^{i}}{i!}\right)$$

$$= e^{-(\lambda+\mu)} \frac{1}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} \mu^{i}$$

$$= e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^{k}}{k!},$$

where we use the Binomial Theorem, Proposition 2.13.3.

---

Section 7: Properties of mean and variance

**Objectives:** by the end of this section you should be able to

- To explore further properties of expectations of a single and multiple variables.
- To understand and use the Law of Large Numbers.
- To define covariance, and use it for computing variances of sums.
- To calculate and interpret correlation coefficients.

[This material is also covered in Sections 7.1 to 7.3 of the course book]
Section 7.1: Properties of expectation

Theorem 7.1.

1. Let $X$ be a constant r.v. with $\mathbb{P}(X = c) = 1$. Then $\mathbb{E}(X) = c$.
2. Let $a$ and $b$ be constants and $X$ be a r.v. Then $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.
3. Let $X$ and $Y$ be r.v.s. Then $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.

Proof.

1. If $\mathbb{P}(X = c) = 1$ then $\mathbb{E}(X) = c\mathbb{P}(X = c) = c$.
2. This is Lemma 5.12.

Proof of Theorem 7.1 (cont).

Proof.

Let $Z = X + Y$, i.e. $Z = g(X, Y)$ where $g(x, y) = x + y$. Then

$$\mathbb{E}Z = \sum_{x_i} \sum_{y_j} g(x_i, y_j) p_{X,Y}(x_i, y_j) \quad \text{by extension of Theorem 5.10}$$

$$= \sum_{x_i} \sum_{y_j} (x_i + y_j) p_{X,Y}(x_i, y_j)$$

$$= \sum_{x_i} \sum_{y_j} \{x_i p_{X,Y}(x_i, y_j) + y_j p_{X,Y}(x_i, y_j)\}$$

$$= \sum_{x_i} x_i \left\{ \sum_{y_j} p_{X,Y}(x_i, y_j) \right\} + \sum_{y_j} y_j \left\{ \sum_{x_i} p_{X,Y}(x_i, y_j) \right\}$$

$$= \sum_{x_i} x_i p_X(x_i) + \sum_{y_j} y_j p_Y(y_j)$$

$$= \mathbb{E}(X) + \mathbb{E}(Y)$$
Additivity of expectation

**Corollary 7.2.**

If $X_1, \ldots, X_n$ are r.v.s then

$$\mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n).$$

**Proof.**

Use Theorem 7.1.3 and induction on $n$.

Combining this with Theorem 7.1.2, we can also show more generally that:

**Theorem 7.3.**

If $a_1, \ldots, a_n$ are constants and $X_1, \ldots, X_n$ are r.v.s then

$$\mathbb{E}(a_1X_1 + \cdots + a_nX_n) = a_1\mathbb{E}(X_1) + \cdots + a_n\mathbb{E}(X_n).$$

---

**Example: Bernoulli trials**

**Example 7.4.**

- Let $T$ be the number of successes in $n$ independent Bernoulli trials.
- Each trial has probability $p$ of success, so $T \sim \text{Bin}(n, p)$.
- Can represent $T$ as $X_1 + \cdots X_n$ where
  $$X_i = \begin{cases} 
  0 & \text{if } i \text{th trial a failure} \\
  1 & \text{if } i \text{th trial a success.}
  \end{cases}$$
- For each $i$, $\mathbb{E}(X_i) = (1-p).0 + p.1 = p$
- So $\mathbb{E}(T) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$ by the corollary.
- So $\mathbb{E}(T) = np$.

- This is simpler (and more general) than Example 5.7.
- Argument extends to Bernoulli trials $X_i$ with probabilities $p_i$ varying with $i$.
- In general $\mathbb{E}T = \sum_{i=1}^{n} p_i$. 

---
Example: BitTorrent problem

Example 7.5.

- Every pack of cornflakes contains a plastic monster drawn at random from a set of \( k \) different monsters.
- Let \( N \) be the number of packs bought in order to obtain a full set.
- Find the expected value of \( N \).
- Let \( X_i \) be the number of packs you need to buy to get from \( i - 1 \) distinct monsters to \( i \) distinct monsters. So
  \[
  N = X_1 + X_2 + \cdots + X_k.
  \]
- Then \( X_1 = 1 \) (i.e. when you do not have any monsters it takes one pack to get the first monster).
- For \( 2 \leq r \leq k \) we have \( X_r \sim \text{Geom}(p_r) \) where
  \[
  p_r = \frac{\text{number of monsters we don’t have}}{\text{number of different monsters}} = \frac{k - (r - 1)}{k}.
  \]

Oliver Johnson (maotj@bris.ac.uk)

Example: BitTorrent problem (cont.)

Example 7.5.

- Therefore (see Example 5.9) \( \mathbb{E}(X_r) = \frac{1}{p_r} = \frac{k}{k - r + 1} \).
- Hence
  \[
  \mathbb{E}(N) = \sum_{r=1}^{k} \mathbb{E}(X_r) = \sum_{r=1}^{k} \frac{k}{k - r + 1}
  \]
  \[
  = k\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}\right) \approx k \ln k.
  \]
- To illustrate this result we have:

<table>
<thead>
<tr>
<th>( k )</th>
<th>( \mathbb{E}(N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>11.4</td>
</tr>
<tr>
<td>10</td>
<td>29.3</td>
</tr>
<tr>
<td>20</td>
<td>80.0</td>
</tr>
</tbody>
</table>
Section 7.2: Properties of variance

Theorem 7.6.

1. \( \text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \)
2. Let \( a \) and \( b \) be constants. Then \( \text{Var}(aX + b) = a^2 \text{Var}(X) \).
3. If \( X \) and \( Y \) are independent r.v.s then
   \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \).

Important: Note that if \( X \) and \( Y \) are not independent, then 3. is not usually true - we return to this in Section 7.3.

Proof of Theorem 7.6

Proof.

1. Seen before as Lemma 5.15 — now we can justify all the steps in that proof. Key is to observe that
   \[ \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu \mathbb{E}(X) + \mu^2, \]
   by Theorem 7.1.2.
2. Set \( Z = aX + b \). We know \( \mathbb{E}(Z) = a\mathbb{E}(X) + b \), so
   \[ (Z - \mathbb{E}(Z))^2 = ((aX + b) - (a\mathbb{E}(X) + b))^2 \]
   \[ = (a(X - \mathbb{E}(X)))^2 = a^2(X - \mathbb{E}(X))^2. \]
   Thus
   \[ \text{Var}(Z) = \mathbb{E}((Z - \mathbb{E}(Z))^2) = a^2\mathbb{E}((X - \mathbb{E}(X))^2) = a^2\text{Var}(X). \]
Proof of Theorem 7.6 (cont).

Proof.

- Set $T = X + Y$. We know that $\mathbb{E}(T) = \mathbb{E}(X) + \mathbb{E}(Y)$, so

$$
(\mathbb{E}(T))^2 = (\mathbb{E}(X))^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + (\mathbb{E}(Y))^2. 
$$

(7.1)

- Need to calculate

$$
\mathbb{E}(T^2) = \mathbb{E}(X^2 + 2XY + Y^2) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2). 
$$

(7.2)

- Hence subtracting (7.1) from (7.2) and rearranging, we obtain:

$$
\text{Var}(T) = \mathbb{E}(T^2) - (\mathbb{E}(T))^2
= (\mathbb{E}(X^2) - (\mathbb{E}(X))^2) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y))
+ (\mathbb{E}(Y^2) - (\mathbb{E}(Y))^2)
= \text{Var}(X) + 2(\mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y)) + \text{Var}(Y). 
$$

(7.3)

- But from Lemma 7.7 below $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, since $X$ and $Y$ are independent, and the result follows.

Useful lemma

**Lemma 7.7.**

Let $X$ and $Y$ be independent r.v.s. Then

$$
\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).
$$

**Proof:**

$$
\mathbb{E}(XY) = \sum_i \sum_j x_i y_j p_{X,Y}(x_i, y_j)
= \sum_i \sum_j x_i y_j p_X(x_i) p_Y(y_j) \quad \text{by independence}
= \sum_i \left[ x_i p_X(x_i) \sum_j y_j p_Y(y_j) \right]
= \sum_i [x_i p_X(x_i) \mathbb{E}(Y)]
= \mathbb{E}(X) \mathbb{E}(Y).
$$
Corollary 7.8.

If $X$ and $Y$ are independent then (by Remark 6.14)
\[ \mathbb{E} g(X)h(Y) = \mathbb{E} g(X)h(Y), \]
for any functions $g$ and $h$.

Corollary 7.9.

Let $X_1, X_2, \ldots$ be independent. Then
\[ \text{Var} (X_1 + X_2 + \cdots + X_n) = \text{Var} (X_1) + \text{Var} (X_2) + \cdots + \text{Var} (X_n). \]

Proof.

Induction on $n$. \qed

Important: If $X_i$ are not independent then the situation is more complicated – see Section 7.3 below.

Example: Bernoulli trials – see Example 7.4

Example 7.10.

- Recall from Example 7.4 that $T \sim \text{Bin}(n, p)$.
- Can write $T = X_1 + \cdots + X_n$ where the $X_i$ are independent Bernoulli($p$) r.v.s.
- Recall from Example 5.16 that $\mathbb{E}(X_i) = 0 \times (1 - p) + 1 \times p = p$ and $\mathbb{E}(X_i^2) = 0^2 \times (1 - p) + 1^2 \times p = p$
- So $\text{Var} (X_i) = p - p^2 = p(1 - p)$.
- Hence by independence
\[ \text{Var} (T) = \text{Var} (X_1 + \cdots + X_n) = \text{Var} (X_1) + \cdots + \text{Var} (X_n) = np(1 - p). \]

Note: much easier than trying to sum this directly!
Application: Sample means

**Theorem 7.11.**
- Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed (IID) random variables with common mean $\mu$ and variance $\sigma^2$.
- Let the sample mean $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.
- Then

$$\mathbb{E}(\bar{X}) = \mu, \quad \text{Var} (\bar{X}) = \frac{\sigma^2}{n}$$

**Proof.**
- Then (see also Theorem 7.3)

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \mathbb{E}(X_1 + \cdots + X_n)$$

$$= \frac{1}{n} (\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) = \frac{1}{n} (\mu + \cdots + \mu) = \mu.$$
Example 7.12.

- For example, toss a fair coin repeatedly, and let
  \[ X_i = \begin{cases} 
  1 & \text{if } i\text{th throw head} \\
  0 & \text{if } i\text{th throw tail} 
  \end{cases} \]
- Then \( \bar{X} \) is the proportion of heads in the first \( n \) tosses.
- \( \mathbb{E}(\bar{X}) = \mathbb{E}(X_i) = \frac{1}{2} \).
- \( \text{Var}(X_i) = \frac{1}{2}(1 - \frac{1}{2}) = \frac{1}{4} \), so
  \[ \text{Var}(\bar{X}) = \frac{1}{4n}. \]

The weak law of large numbers

- Let \( Y \) be any r.v. and let \( c > 0 \) be a positive constant.
- Recall Chebychev’s inequality (Theorem 5.21):
  \[ \mathbb{P}(|Y - \mathbb{E}(Y)| > c) \leq \frac{\text{Var}(Y)}{c^2}. \]
- We know that \( \mathbb{E}(\bar{X}) = \mu \) and \( \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \).
- So taking \( Y = \bar{X} \) in Chebyshev we deduce:
  \[ \mathbb{P}(|\bar{X} - \mu| > c) \leq \frac{\sigma^2}{nc^2}. \]

**Theorem 7.13 (Weak law of large numbers).**

Let \( X_1, X_2, \ldots \) be a sequence of independent identically distributed (IID) random variables with common mean \( \mu \) and variance \( \sigma^2 \). Then for any \( c > 0 \):
\[ \mathbb{P}(|\bar{X} - \mu| > c) \to 0 \text{ as } n \to \infty. \]
Application to coin tossing

Example 7.14.

- As in Example 7.12, let $\bar{X}$ be the proportion of heads in first $n$ tosses.
- Then $\mu = \frac{1}{2}$ and $\sigma^2 = \frac{1}{4}$. Thus

$$\mathbb{P}\left(\left|\bar{X} - \frac{1}{2}\right| > c\right) \leq \frac{1}{4nc^2}.$$

- So for example

$$\mathbb{P}(0.49 < \bar{X} < 0.51) \geq 1 - \frac{2500}{n}.$$

- This tends to one as $n \to \infty$.
- In fact the inequalities are very conservative here.

Axioms and definitions match our intuitive beliefs about probability.

Closely related to central limit theorem (see later) and other laws of large numbers (see Probability 2).

Section 7.3: Covariance

- When does $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$?
- We know from Lemma 7.7 that it does if $X$ and $Y$ are independent.
- We first note that it is not generally true that $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Example 7.15.

- Let $X$ and $Y$ be r.v.s with

$$X = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases} \quad \text{and} \quad Y = X.$$

- We have $\mathbb{E}(X) = \mathbb{E}(Y) = \frac{1}{2}$.

- Let $Z = XY$, so $Z = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases}$ and $\mathbb{E}(Z) = \frac{1}{2}$.

- We see that in this case

$$\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y).$$
Covariance definition

**Definition 7.16.**
The covariance of $X$ and $Y$ is defined as

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))].$$

Covariance measures how the two random variables vary together.

**Remark 7.17.**
- For any random variable $X$ we have $\text{Cov}(X, X) = \text{Var}(X)$.
- $\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$
- $\text{Cov}(X, bY) = b \text{Cov}(X, Y)$
- $\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z)$

*The proofs are an exercise.*

**Lemma 7.18.**

$$\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

**Proof.**
Write $\mu = \mathbb{E}(X)$ and $\nu = \mathbb{E}(Y)$. Then

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mu)(Y - \nu)]$$
$$= \mathbb{E}[XY - \nu X - \mu Y + \mu \nu]$$
$$= \mathbb{E}(XY) - \nu \mathbb{E}(X) - \mu \mathbb{E}(Y) + \mu \nu$$
$$= \mathbb{E}(XY) - \mathbb{E}(Y)\mathbb{E}(X) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)$$
Lemma 7.19.

For any random variables \( X \) and \( Y \),

\[
\text{Var} (X + Y) = \text{Var} (X) + 2\text{Cov} (X, Y) + \text{Var} (Y).
\]

Proof.

- In the proof of Theorem 7.6, equation (7.3) shows that

\[
\text{Var} (X + Y) = \text{Var} (X) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) + \text{Var} (Y),
\]

without assuming independence.

- The term in the middle is the covariance.

Delicate issue

We can rephrase Lemmas 7.18 and 7.7 to deduce that

Lemma 7.20.

Let \( X \) and \( Y \) be independent. Then \( \text{Cov} (X, Y) = 0 \).

Example 7.21.

If \( \text{Cov} (X, Y) = 0 \), we cannot deduce that \( X \) and \( Y \) are independent.

- Consider

\[
p_{X,Y}(-1,0) = p_{X,Y}(1,0) = p_{X,Y}(0,-1) = p_{X,Y}(0,1) = 1/4.
\]

- Then (check): \( XY \equiv 0 \) so \( \mathbb{E}(XY) = 0 \), and by symmetry \( \mathbb{E}X = \mathbb{E}Y = 0 \).

- Hence \( \text{Cov} (X, Y) = 0 \), but clearly \( X \) and \( Y \) are dependent.

Important: to understand the direction of implication of these statements.
Example 7.22.

- An urn contains two biased coins.
- Coin 1 has a probability $\frac{1}{3}$ of showing a head.
- Coin 2 has a probability $\frac{2}{3}$ of showing a head.
- A coin is selected at random and the same coin is tossed twice.
- Let $X = \begin{cases} 1 & \text{if 1st toss is H} \\ 0 & \text{if 1st toss is T} \end{cases}$
  and $Y = \begin{cases} 1 & \text{if 2nd toss is H} \\ 0 & \text{if 2nd toss is T} \end{cases}$
- Let $W = X + Y$ be the total number of heads. Find $\text{Cov}(X, Y)$, $\mathbb{E}(W)$, $\text{Var}(W)$.

Urn example (cont.)

Example 7.22.

- $\mathbb{P}(X = 1, Y = 1) = \mathbb{P}(X = 1, Y = 1 | \text{coin 1}) \mathbb{P}(\text{coin 1})$
  $+ \mathbb{P}(X = 1, Y = 1 | \text{coin 2}) \mathbb{P}(\text{coin 2})$
  $= \left( \frac{1}{3} \right)^2 \frac{1}{2} + \left( \frac{2}{3} \right)^2 \frac{1}{2} = \frac{5}{18}$
- Similarly for the other values

<table>
<thead>
<tr>
<th>$p_{X,Y}(x,y)$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
<th>$p_X(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 0$</td>
<td>$\frac{5}{18}$</td>
<td>$\frac{4}{18}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
</tbody>
</table>

- $X$ and $Y$ are Bernoulli($\frac{1}{2}$) r.v.s, so $\mathbb{E}(X) = \mathbb{E}(Y) = \frac{1}{2}$ and $\text{Var}(X) = \text{Var}(Y) = \frac{1}{4}$. 
Example 7.22.

\[
E(XY) = 0 \times 0 \times p_{X,Y}(0,0) + 0 \times 1 \times p_{X,Y}(0,1) \\
+ 1 \times 0 \times p_{X,Y}(1,0) + 1 \times 1 \times p_{X,Y}(1,1) = \frac{5}{18}.
\]

Thus \( \text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{5}{18} - \left(\frac{1}{2}\right)^2 = \frac{1}{36} \).

\[
\text{Var}(W) = \text{Var}(X) + 2\text{Cov}(X, Y) + \text{Var}(Y) = \frac{1}{4} + \frac{2}{36} + \frac{1}{4} = \frac{5}{9}.
\]

Compare with Bin(2, \( \frac{1}{2} \)) when variance = \( \frac{1}{2} \).

Section 7.4: Correlation coefficient

If \( X \) and \( Y \) tend to increase (and decrease) together \( \text{Cov}(X, Y) > 0 \) (e.g. age and salary).

If one tends to increase as the other decreases then \( \text{Cov}(X, Y) < 0 \) (e.g. hours of training, marathon times).

If \( X \) and \( Y \) are independent then \( \text{Cov}(X, Y) = 0 \).

Definition 7.23.

The correlation coefficient of \( X \) and \( Y \) is

\[
\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.
\]

Note that it can be shown that

\[-1 \leq \rho(X, Y) \leq 1.\]

This is essentially the Cauchy-Schwarz inequality from linear algebra.

\( \rho \) is a measure of how dependent the random variables are, and doesn’t depend on the scale of either r.v.
Further example

In the previous urn example (Example 7.22) we had \( \text{Var} (X) = \frac{1}{4}, \text{Var} (Y) = \frac{1}{4} \) and \( \text{Cov} (X, Y) = \frac{1}{36} \). So \( \rho(X, Y) = \frac{1}{9} \).

Example 7.24.

- A fair coin is tossed 10 times.
- Let \( X \) be the number of heads in the first 5 tosses and let \( Y \) be the total number of heads.
- We will find \( \rho(X, Y) \).
- First note that since \( X \) and \( Y \) are both binomially distributed we have

\[
\text{Var} (X) = \frac{5}{4}, \\
\text{Var} (Y) = \frac{5}{2}.
\]

Further example (cont.)

Example 7.24.

- To find the covariance of \( X \) and \( Y \) it is convenient to set \( Z = Y - X \).
- Note that \( Z \) is the number of heads in the last 5 tosses.
- Thus \( X \) and \( Z \) are independent. This implies that \( \text{Cov} (X, Z) = 0 \).

Thus

\[
\text{Cov} (X, Y) = \text{Cov} (X, X + Z) = \text{Cov} (X, X) + \text{Cov} (X, Z) = \text{Var} (X) + 0 = \frac{5}{4}.
\]

Thus

\[
\rho(X, Y) = \frac{\text{Cov} (X, Y)}{\sqrt{\text{Var} (X)\text{Var} (Y)}} = \frac{1}{\sqrt{2}}.
\]
Section 8: Continuous random variables

**Objectives:** by the end of this section you should be able to

- Understand continuous random variables.
- Interpret density and distribution functions.
- Know how to calculate means and variances of continuous random variables.
- Understand the basic properties of the exponential distribution.

[This material is also covered in Sections 5.1, 5.2, 5.3 and 5.5 of the course book]

Section 8.1: Motivation and definition

**Remark 8.1.**

- So far we have studied r.v.s that take a discrete (finite) set of values.
- Many r.v.s take a continuum of values e.g. height, weight, time, temperature are real-valued.
- Let $X$ be the time till an atom decays (in seconds). Then $\mathbb{P}(X = \pi) = 0$.
- But we expect
  
  $\mathbb{P}(\pi \leq X \leq \pi + \delta) \approx \text{const} \times \delta$

  for $\delta$ small.
- In general $\mathbb{P}(X = x) = 0$ for any particular $x$ and we expect
  
  $\mathbb{P}(x \leq X \leq x + \delta) \approx f_X(x)\delta$

  for $\delta$ small.
- Think of $f_X(x)$ as an ‘intensity’ – won’t generally be 0.
Remark 8.1.

- Consider an interval $[a, b]$.
- Divide it up into $n$ segments
  
  $$a = x_0 < x_1 < \cdots < x_n = b$$

  with $\delta_i = x_i - x_{i-1} = (b - a)/n$ for $i = 1, \ldots, n$.

- Then
  
  $$\mathbb{P}(a \leq X < b) = \mathbb{P}\left(\bigcup_{i=1}^{n} \{ x_{i-1} \leq X < x_i \} \right)$$
  
  $$= \sum_{i=1}^{n} \mathbb{P}(x_{i-1} \leq X < x_i) \approx \sum_{i=1}^{n} f_X(x_i) \delta_i.$$

- As $n \to \infty$, $\sum_{i=1}^{n} f_X(x_i) \delta_i \to \int_{a}^{b} f_X(x) \, dx$.
- So we expect $\mathbb{P}(a \leq X < b) = \int_{a}^{b} f_X(x) \, dx$.

Continuous random variables

Definition 8.2.

A random variable $X$ has a *continuous distribution* if there exists a function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{P}(a \leq X < b) = \int_{a}^{b} f_X(x) \, dx \quad \text{for all } a, b \text{ with } a < b.$$

The function $f_X(x)$ is called the *probability density function* (pdf) for $X$. 
Remark 8.3.

Suppose that $X$ is a continuous r.v., then

- $P(X = x) = 0$ for all $x$, so
  \[ P(a \leq X < b) = P(a \leq X \leq b). \]

- Special case:
  \[ P(X \leq b) = P(X < b) = \lim_{a \to -\infty} P(a \leq X \leq b) = \int_{-\infty}^{b} f_X(x) \, dx. \]
  
  Since $P(-\infty < X < \infty) = 1$ we have
  \[ \int_{-\infty}^{\infty} f_X(x) \, dx = 1. \]

- $f_X(x)$ is not a probability. In particular we can have $f_X(x) > 1$.

- However $f_X(x) \geq 0$.

Section 8.2: Mean and variance

Definition 8.4.

Let $X$ be a continuous r.v. with pdf $f_X(x)$. The mean or expectation of $X$ is

\[ \mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx. \]

Lemma 8.5.

Let $X$ be a continuous r.v. with pdf $f_X(x)$ and $Z = g(X)$ for some function $g$. Then

\[ \mathbb{E}(Z) = \mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) \, dx. \]

- Note that $x$ is a dummy variable.
- Note that in general we need to integrate over $x$ from $-\infty$ to $\infty$.
- However (see e.g. Example 8.7) we only need to consider the range where $f_X(x) > 0$. 
Variance

Definition 8.6.
The variance of $X$ is

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

where $\mu$ is shorthand for $\mathbb{E}(X)$. As before we can show that

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

Uniform distribution

Example 8.7.

- Suppose the density $f_X(x) = 1$ for $0 \leq x \leq 1$ and 0 otherwise.
- May be best to represent this with an indicator function $\mathbb{I}$.
- Can write $f_X(x) = \mathbb{I}(0 \leq x \leq 1)$.
- We know that this is a valid density function since

$$\int_{-\infty}^{\infty} f_X(x) \, dx = \int_{-\infty}^{\infty} \mathbb{I}(0 \leq x \leq 1) = \int_{0}^{1} 1 \, dx = 1.$$  

- We call this the Uniform distribution on $[0, 1]$.
- Generalize: given $a < b$, uniform distribution on $[a, b]$ has density

$$f_Y(y) = \frac{1}{b-a} \mathbb{I}(a \leq y \leq b).$$
- Write $Y \sim U(a, b)$.  

Uniform distribution

**Example 8.7.**

- If $X$ is uniform on $[0, 1]$:

$$
\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_0^1 x \, dx = \left[ \frac{x^2}{2} \right]_0^1 = \frac{1}{2},
$$

$$
\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx = \int_0^1 x^2 \, dx = \left[ \frac{x^3}{3} \right]_0^1 = \frac{1}{3},
$$

so that $\text{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1}{3} - \frac{1}{2^2} = \frac{1}{12}$.

- Similarly if $Y$ is uniform on $[a, b]$:

$$
\mathbb{E}(Y) = \int_{-\infty}^{\infty} x f_Y(x) \, dx = \int_a^b \frac{x}{b-a} \, dx = \frac{1}{b-a} \left[ \frac{x^2}{2} \right]_a^b = \frac{b^2 - a^2}{2(b-a)} = \frac{a + b}{2},
$$

Section 8.3: The distribution function

**Definition 8.8.**

For any r.v. $X$, the (cumulative) distribution function of $X$ is defined as the function $F_X : \mathbb{R} \to [0, 1]$ given by

$$
F_X(x) = \mathbb{P}(X \leq x) \text{ for } x \in \mathbb{R}.
$$

**Lemma 8.9.**

In fact, these hold for any r.v. whether discrete, continuous or other:

- $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$
- $F_X$ is increasing
- $F_X(x) \to 0$ as $x \to -\infty$
- $F_X(x) \to 1$ as $x \to \infty$
Distribution and density function

**Lemma 8.10.**

Let $X$ have a continuous distribution. Then (again $\xi$ is a dummy variable)

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(\xi) \, d\xi \quad \text{for all } x \in \mathbb{R}.$$  

- $\mathbb{P}(X \leq x)$ is the area under the density function left of $x$.
- Hence we have that $F_X'(x) = f_X(x)$.
- Note that when $X$ is continuous, $\mathbb{P}(X = x) = 0$ for all $x$ so $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X < x)$.

**Example 8.11.**

In the setting of Example 8.7, the

$$F_X(x) = \begin{cases} 
0 & \text{for } x \leq 0, \\
 x & \text{for } 0 \leq x \leq 1, \\
1 & \text{for } 1 \leq x.
\end{cases}$$

**Example 8.12.**

- Suppose $X$ has a continuous distribution with density function

$$f_X(x) = \begin{cases} 
0 & x \leq 1 \\
\frac{2}{x^3} & x > 1
\end{cases}$$

- Find $F_X$.
- Let $x \leq 1$. Then

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi) \, d\xi = \int_{-\infty}^{x} 0 \, d\xi = 0.$$
Example 8.12.

- Let $x > 1$. Then

$$F_X(x) = \int_{-\infty}^{x} f_X(\xi) \, d\xi = \int_{-\infty}^{1} 0 \, d\xi + \int_{1}^{x} \frac{2}{\xi^3} \, d\xi = 0 + \left[ \frac{-1}{\xi^2} \right]_{1}^{x} = \frac{-1}{x^2} - \frac{-1}{1} = 1 - \frac{1}{x^2}.$$

- So $F_X(x) = \begin{cases} 0 & x \leq 1 \\ 1 - \frac{1}{x^2} & x > 1 \end{cases}$

Note: the integrals have limits. Don’t write $F_X(x) = \int f_X(\xi) \, d\xi$ without limits then determine $C$. It is both confusing and sloppy!

Section 8.4: Examples of continuous random variables

- Let $T$ be the time to wait for an event e.g. a bus to arrive, or a radioactive decay to occur.

- Suppose that if the event has not happened by $t$ then the probability that it happens in $(t, t + \delta)$ is $\lambda \delta + o(\delta)$ (i.e. it doesn’t depend on $t$).

- Then (for $t > 0$) $F_T(t) = \mathbb{P}(T \leq t) = 1 - e^{-\lambda t}$ and $f_T(t) = \lambda e^{-\lambda t}$. See why in Probability 2.

Definition 8.13.

- A r.v. $T$ has an exponential distribution with rate parameter $\lambda$ if it has a continuous distribution with density

$$f_T(t) = \begin{cases} 0 & t \leq 0 \\ \lambda e^{-\lambda t} & t > 0 \end{cases}$$

- Notation $T \sim \text{Exp}(\lambda)$. 


- \( P(T > t) = 1 - P(T \leq t) = e^{-\lambda t} \)

\[ E(T) = \int_{-\infty}^{\infty} tf_T(t) \, dt = \int_{0}^{\infty} t\lambda e^{-\lambda t} \, dt = \left[ -te^{-\lambda t} \right]_0^\infty + \int_0^\infty e^{-\lambda t} \, dt = 0 + \frac{1}{\lambda} \]

\[ \text{Var}(T) = \frac{1}{\lambda^2} \quad (Exercise) \]

Exponential distribution properties (cont.)


- Continuous analogue of the geometric distribution. In particular it has the lack of memory property (cf Lemma 4.19):

\[
P(T > t + s \mid T > s) = \frac{P(T > t + s \text{ and } T > s)}{P(T > s)} = \frac{P(T > t + s)}{P(T > s)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T > t).
\]
Section 9: Continuous random variables II

Objectives: by the end of this section you should be able to

- Describe the gamma distribution and its properties.
- Understand transformations of continuous random variables.
- Describe normal random variables and use tables to calculate probabilities.
- Consider jointly distributed continuous random variables.

[This material is also covered in Sections 5.4, 5.7 and 6.1 of the course book]

Section 9.1: Gamma distributions

Definition 9.1.

For $\alpha > 0$ define the gamma function

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x} \, dx.$$ 

We will see that this is a generalisation of the(shifted) factorial function.
Gamma function properties

Remark 9.2.

- Note that
  \[ \Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x} \, dx \]
  \[ = \left[-x^{\alpha-1}e^{-x}\right]_0^\infty + (\alpha - 1) \int_0^\infty x^{\alpha-2}e^{-x} \, dx \]
  \[ = 0 + (\alpha - 1)\Gamma(\alpha - 1) \]

  for general \( \alpha \).

- Also
  \[ \Gamma(1) = \int_0^\infty x^{1-1}e^{-x} \, dx = [-e^{-x}]_0^\infty = 1. \]

- So by induction for integer \( n \), the \( \Gamma(n) = (n - 1)! \) since
  \[ \Gamma(n) = (n - 1)\Gamma(n - 1) = (n - 1)(n - 2)! = (n - 1)!. \]

Gamma distribution

Definition 9.3.

- A random variable has a gamma distribution with shape parameter \( \alpha \) and rate parameter \( \lambda \) if it has a continuous distribution with density proportional to
  \[ x^{\alpha-1}e^{-\lambda x}, \]
  for \( x > 0 \).

- Note that for \( \alpha = 1 \) this reduces to the exponential distribution of Definition 8.13

- We find the normalization constant in Lemma 9.4 below.

- Notation: \( X \sim \text{Gamma}(\alpha, \lambda) \).

Warning: sometimes gamma and exponential distributions are reported with different parameterisations, using a mean \( \mu = 1/\lambda \) instead of a rate \( \lambda \).
Lemma 9.4.

Let $X \sim \Gamma(\alpha, \lambda)$. Then

$$f_X(x) = \left\{ \begin{array}{ll} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x > 0, \\ 0 & x \leq 0. \end{array} \right.$$ 

Proof.

For $x > 0$, $f_X(x) = C x^{\alpha-1} e^{-\lambda x}$ for some constant $C$. We know

$$1 = \int_{\infty}^{\infty} f_X(x) \, dx$$

$$= \int_{0}^{\infty} C x^{\alpha-1} e^{-\lambda x} \, dx$$

$$= C \int_{0}^{\infty} \left( \frac{y}{\lambda} \right)^{\alpha-1} e^{-y} \frac{dy}{\lambda} \quad \text{setting } y = \lambda x$$

$$= \frac{C}{\lambda^{\alpha}} \int_{0}^{\infty} y^{\alpha-1} e^{-y} \, dy = \frac{C}{\lambda^{\alpha}} \Gamma(\alpha).$$

Gamma distribution properties

Remark 9.5.

- If $\alpha = 1$ then $f_X(x) = \left\{ \begin{array}{ll} 0 & x \leq 0 \\ \lambda e^{-\lambda x} & x > 0. \end{array} \right.$

  I.e. if $X \sim \Gamma(1, \lambda)$ then $X \sim \text{Exp}(\lambda)$.

- In Lemma 11.13 we will see that (for integer $\alpha$) a $\Gamma(\alpha, \lambda)$ r.v. has the same distribution as the sum of $\alpha$ independent $\text{Exp}(\lambda)$ r.v.s
Remark 9.5.

\[ \mathbb{E}(X) = \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \, dx \]
\[ = \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{\lambda x} \, dx \]
\[ = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 1)}{\lambda^{\alpha+1}} \int_0^\infty \frac{\lambda^{\alpha+1}}{\Gamma(\alpha + 1)} x^{\alpha} e^{\lambda x} \, dx \]
\[ = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + 1)}{\lambda^{\alpha+1}} \int_0^\infty f_Y(x) \, dx \quad \text{where } Y \sim \text{Gamma}(\alpha + 1, \lambda) \]
\[ = \frac{\alpha \Gamma(\alpha)}{\lambda \Gamma(\alpha)} \times 1 = \frac{\alpha}{\lambda} \]

Similarly \( \text{Var}(X) = \frac{\alpha}{\lambda^2} \).

Section 9.2: Change of variables

- Let \( X \) be a r.v. with a known distribution.
- Let \( g : \mathbb{R} \to \mathbb{R} \), and define a new r.v. \( Y \) by \( Y = g(X) \).
- What is the distribution of \( Y \)?
- Note we already know how to calculate \( \mathbb{E}(Y) = \mathbb{E}(g(X)) \) using Theorem 5.10.
Example: scaling uniforms

**Example 9.6.**

- Suppose that $X \sim U(0, 1)$, so that
  
  \[ f_X(x) = \begin{cases} 
  1 & 0 < x < 1 \\
  0 & \text{otherwise} 
  \end{cases} \]

- Suppose that $g(x) = a + (b - a)x$ with $b > a$, so $Y = a + (b - a)X$.
- Note that $0 \leq X \leq 1 \implies a \leq Y \leq b$.
- For $a \leq y \leq b$ we have
  
  \[
  F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(a + (b - a)X \leq y) = \mathbb{P}(X \leq \frac{y - a}{b - a})
  \]
  
  \[
  = \frac{y - a}{b - a} \quad \text{since } X \sim U(0, 1).
  \]

- Thus $f_Y(y) = F'_Y(y) = \frac{1}{b - a}$ if $a < y < b$. Also $f_Y(y) = 0$ otherwise.
- So $Y \sim U(a, b)$.

---

**General case**

**Lemma 9.7.**

Let $X$ be a r.v. taking values in $I \subseteq \mathbb{R}$. Let $Y = g(X)$ where $g : I \to J$ is strictly monotonic and differentiable on $I$ with inverse function $h = g^{-1}$. Then

\[
 f_Y(y) = \begin{cases} 
  f_X(h(y))|h'(y)| & y \in J \\
  0 & y \notin J 
  \end{cases}
\]

**Proof.**

- $X$ takes values in $I$, and $g : I \to J$, so $Y$ takes values in $J$.
- Therefore $f_Y(y) = 0$ for $y \notin J$.
- **Case 1** Assume first that $g$ is strictly increasing. For $y \in J$

  \[
  F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq h(y)) = F_X(h(y)).
  \]

- So $f_Y(y) = F'_Y(y) = F'_X(h(y))h'(y) = f_X(h(y))h'(y)$ by chain rule.
Proof of Lemma 9.7 (cont.)

Proof.

- **Case 2** Now assume $g$ is strictly decreasing. For $y \in J$

  \[
  F_Y(y) = P(g(X) \leq y) = P(X \geq h(y)) = 1 - P(X < h(y)) = 1 - F_X(h(y)).
  \]

- So $f_Y(y) = -f_X(h(y))h'(y)$.
- But $g$ (and therefore $h$) are strictly decreasing, so $h'(y) < 0$, and $-h'(y) = |h'(y)|$.

Simulation of random variables

- In general computers can give you $U(0,1)$ random numbers and nothing else.
- You need to transform these $U(0,1)$ to give you something useful.

**Example 9.8.**

- Let $X \sim U(0,1)$ and let $Y = \frac{1}{\lambda} \log \left( \frac{1}{1-X} \right)$.
- What is the distribution of $Y$?
- Define $g : (0,1) \to (0,\infty)$ by $g(x) = \frac{1}{\lambda} \log \left( \frac{1}{1-x} \right)$.
- To find the inverse of the function $g$ set

  \[
  y = \frac{1}{\lambda} \log \left( \frac{1}{1-x} \right) \quad \implies -\lambda y = \log(1-x) \quad \implies x = 1 - e^{-\lambda y}
  \]
Example 9.8.

- That is, the inverse function $h$ is given by $h(y) = 1 - e^{-\lambda y}$.
- The image of the function $g$ is $J = (0, \infty)$, so $f_Y(y) = 0$ for $y \leq 0$.
- Let $y > 0$. Then $f_Y(y) = f_X(h(y))|h'(y)| = 1 \times \lambda e^{-\lambda y}$.
- So $Y \sim \text{Exp}(\lambda)$.
- To generate $\text{Exp}(\lambda)$ random variables, you take the $U(0, 1)$ r.v.s given by the computer and apply $g$.

Exercise Let $F_Y$ be the distribution function of a continuous r.v. and let $g = F_Y^{-1}$. Take $X \sim U(0, 1)$ and let $Y = g(X)$. What is the distribution of $Y$?

Section 9.3: The normal distribution

Definition 9.9.

A r.v. $Z$ has the standard normal distribution if it is continuous with pdf

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

Notation: $Z \sim \mathcal{N}(0, 1)$.

Lemma 9.10.

For $Z \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}(Z) = 0,$$

$$\text{Var}(Z) = \mathbb{E}(Z^2) = 1.$$
Proof of Lemma 9.10

Proof.

- $f_Z(z)$ is symmetric about 0. So
  
  $\mathbb{E}(Z) = \int_{-\infty}^{\infty} zf_Z(z) \, dz = 0$.

- Alternatively, notice that $zf_Z(z) = -\frac{d}{dz}f_Z(z)$ so that
  
  $\mathbb{E}(Z) = \int_{-\infty}^{\infty} zf_Z(z) \, dz = -\int_{-\infty}^{\infty} \left( \frac{d}{dz}f_Z(z) \right) \, dz = [f_Z(z)]_{-\infty}^{\infty} = 0$.

- Similarly, integration by parts gives
  
  $\mathbb{E}(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\left(ze^{-\frac{z^2}{2}}\right) \, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \, dz = 1$.

- So $\text{Var}(Z) = \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2 = 1$.

General normal distribution properties

Remark 9.11.

- Often $f_Z(z)$ is denoted $\phi(z)$ and $F_Z(z)$ is denoted $\Phi(z)$.

- It is not possible to write down a formula for $\Phi(z)$ for most values of $z$. Instead values of $\Phi(z)$ are in tables, or can be calculated using a computer program. See Statistics 1.

Definition 9.12.

A r.v. $X$ has a normal distribution with mean $\mu$ and variance $\sigma^2$ if it is continuous with pdf

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

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$. 
General normal distribution properties (cont.)

Lemma 9.13.
Let \( X \sim \mathcal{N}(\mu, \sigma^2) \) and define \( Z = \frac{X - \mu}{\sigma} \). Then \( Z \sim \mathcal{N}(0, 1) \).

Proof.
- \( Z = g(X) \) where \( g(x) = \frac{x - \mu}{\sigma} \).
- If \( z = g(x) = \frac{x - \mu}{\sigma} \) then \( x = \mu + \sigma z \) so \( h(z) = \mu + \sigma z = g^{-1}(z) \).
- Therefore by Lemma 9.7
  \[
  f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(h(z) - \mu)^2}{2\sigma^2} \right\} \times \sigma = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{(\mu + \sigma z - \mu)^2}{2\sigma^2} \right\} = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{z^2}{2} \right\}.
  \]

General normal distribution properties (cont.)

Let \( X \sim \mathcal{N}(\mu, \sigma^2) \). Then \( \mathbb{E}(X) = \mu \) and \( \text{Var}(X) = \sigma^2 \).

Proof.
- We know that \( Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1) \), so \( \mathbb{E}(Z) = 0 \) and \( \text{Var}(Z) = 1 \).
- So \( 0 = \mathbb{E}(Z) = \mathbb{E} \left( \frac{X - \mu}{\sigma} \right) = \frac{\mathbb{E}(X) - \mu}{\sigma} \) and \( \mathbb{E}(X) = \mu \).
- Also \( 1 = \text{Var}(Z) = \text{Var} \left( \frac{X - \mu}{\sigma} \right) = \frac{\text{Var}(X)}{\sigma^2} \).

- Many quantities have an approximate normal distribution.
- For example heights in a population, measurement errors.
- There are good theoretical reasons for this (see Theorem 11.21).
- The normal distribution is very important in statistics.
Normal distribution tables

\[ \Phi(z) \]

<table>
<thead>
<tr>
<th>z</th>
<th>0.00</th>
<th>0.01</th>
<th>0.02</th>
<th>0.03</th>
<th>0.04</th>
<th>0.05</th>
<th>0.06</th>
<th>0.07</th>
<th>0.08</th>
<th>0.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.5000</td>
<td>0.5040</td>
<td>0.5080</td>
<td>0.5120</td>
<td>0.5160</td>
<td>0.5199</td>
<td>0.5239</td>
<td>0.5279</td>
<td>0.5319</td>
<td>0.5359</td>
</tr>
<tr>
<td>0.1</td>
<td>0.5398</td>
<td>0.5438</td>
<td>0.5478</td>
<td>0.5517</td>
<td>0.5557</td>
<td>0.5596</td>
<td>0.5636</td>
<td>0.5675</td>
<td>0.5714</td>
<td>0.5753</td>
</tr>
<tr>
<td>0.2</td>
<td>0.5793</td>
<td>0.5832</td>
<td>0.5871</td>
<td>0.5910</td>
<td>0.5949</td>
<td>0.5987</td>
<td>0.6026</td>
<td>0.6064</td>
<td>0.6103</td>
<td>0.6141</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6179</td>
<td>0.6217</td>
<td>0.6255</td>
<td>0.6293</td>
<td>0.6331</td>
<td>0.6368</td>
<td>0.6406</td>
<td>0.6443</td>
<td>0.6480</td>
<td>0.6517</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6554</td>
<td>0.6591</td>
<td>0.6628</td>
<td>0.6664</td>
<td>0.6700</td>
<td>0.6736</td>
<td>0.6772</td>
<td>0.6808</td>
<td>0.6844</td>
<td>0.6879</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6915</td>
<td>0.6950</td>
<td>0.6985</td>
<td>0.7019</td>
<td>0.7054</td>
<td>0.7088</td>
<td>0.7123</td>
<td>0.7157</td>
<td>0.7190</td>
<td>0.7224</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7257</td>
<td>0.7291</td>
<td>0.7324</td>
<td>0.7357</td>
<td>0.7389</td>
<td>0.7422</td>
<td>0.7454</td>
<td>0.7486</td>
<td>0.7517</td>
<td>0.7549</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7580</td>
<td>0.7611</td>
<td>0.7642</td>
<td>0.7673</td>
<td>0.7704</td>
<td>0.7734</td>
<td>0.7764</td>
<td>0.7794</td>
<td>0.7823</td>
<td>0.7852</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7881</td>
<td>0.7910</td>
<td>0.7939</td>
<td>0.7967</td>
<td>0.7995</td>
<td>0.8023</td>
<td>0.8051</td>
<td>0.8078</td>
<td>0.8106</td>
<td>0.8133</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8159</td>
<td>0.8186</td>
<td>0.8212</td>
<td>0.8238</td>
<td>0.8264</td>
<td>0.8289</td>
<td>0.8315</td>
<td>0.8340</td>
<td>0.8365</td>
<td>0.8389</td>
</tr>
<tr>
<td>1.0</td>
<td>0.8413</td>
<td>0.8438</td>
<td>0.8461</td>
<td>0.8485</td>
<td>0.8508</td>
<td>0.8531</td>
<td>0.8554</td>
<td>0.8577</td>
<td>0.8599</td>
<td>0.8621</td>
</tr>
<tr>
<td>1.1</td>
<td>0.8643</td>
<td>0.8665</td>
<td>0.8686</td>
<td>0.8708</td>
<td>0.8729</td>
<td>0.8749</td>
<td>0.8770</td>
<td>0.8790</td>
<td>0.8810</td>
<td>0.8830</td>
</tr>
<tr>
<td>1.2</td>
<td>0.8849</td>
<td>0.8869</td>
<td>0.8888</td>
<td>0.8907</td>
<td>0.8925</td>
<td>0.8944</td>
<td>0.8962</td>
<td>0.8980</td>
<td>0.8997</td>
<td>0.9015</td>
</tr>
<tr>
<td>1.3</td>
<td>0.9032</td>
<td>0.9049</td>
<td>0.9066</td>
<td>0.9082</td>
<td>0.9099</td>
<td>0.9115</td>
<td>0.9131</td>
<td>0.9147</td>
<td>0.9162</td>
<td>0.9177</td>
</tr>
<tr>
<td>1.4</td>
<td>0.9192</td>
<td>0.9207</td>
<td>0.9222</td>
<td>0.9236</td>
<td>0.9251</td>
<td>0.9265</td>
<td>0.9279</td>
<td>0.9292</td>
<td>0.9306</td>
<td>0.9319</td>
</tr>
<tr>
<td>1.5</td>
<td>0.9332</td>
<td>0.9345</td>
<td>0.9357</td>
<td>0.9370</td>
<td>0.9382</td>
<td>0.9394</td>
<td>0.9406</td>
<td>0.9418</td>
<td>0.9429</td>
<td>0.9441</td>
</tr>
<tr>
<td>1.6</td>
<td>0.9452</td>
<td>0.9463</td>
<td>0.9474</td>
<td>0.9484</td>
<td>0.9494</td>
<td>0.9505</td>
<td>0.9515</td>
<td>0.9525</td>
<td>0.9535</td>
<td>0.9545</td>
</tr>
<tr>
<td>1.7</td>
<td>0.9554</td>
<td>0.9564</td>
<td>0.9573</td>
<td>0.9582</td>
<td>0.9591</td>
<td>0.9599</td>
<td>0.9608</td>
<td>0.9616</td>
<td>0.9625</td>
<td>0.9633</td>
</tr>
<tr>
<td>1.8</td>
<td>0.9641</td>
<td>0.9649</td>
<td>0.9656</td>
<td>0.9664</td>
<td>0.9671</td>
<td>0.9678</td>
<td>0.9686</td>
<td>0.9693</td>
<td>0.9699</td>
<td>0.9706</td>
</tr>
<tr>
<td>1.9</td>
<td>0.9713</td>
<td>0.9719</td>
<td>0.9726</td>
<td>0.9732</td>
<td>0.9738</td>
<td>0.9744</td>
<td>0.9750</td>
<td>0.9756</td>
<td>0.9761</td>
<td>0.9767</td>
</tr>
<tr>
<td>2.0</td>
<td>0.9772</td>
<td>0.9778</td>
<td>0.9783</td>
<td>0.9788</td>
<td>0.9793</td>
<td>0.9798</td>
<td>0.9803</td>
<td>0.9808</td>
<td>0.9812</td>
<td>0.9817</td>
</tr>
<tr>
<td>2.1</td>
<td>0.9821</td>
<td>0.9826</td>
<td>0.9830</td>
<td>0.9834</td>
<td>0.9838</td>
<td>0.9842</td>
<td>0.9846</td>
<td>0.9850</td>
<td>0.9854</td>
<td>0.9857</td>
</tr>
<tr>
<td>2.2</td>
<td>0.9861</td>
<td>0.9864</td>
<td>0.9868</td>
<td>0.9871</td>
<td>0.9875</td>
<td>0.9878</td>
<td>0.9881</td>
<td>0.9884</td>
<td>0.9887</td>
<td>0.9890</td>
</tr>
<tr>
<td>2.3</td>
<td>0.9893</td>
<td>0.9896</td>
<td>0.9898</td>
<td>0.9901</td>
<td>0.9904</td>
<td>0.9906</td>
<td>0.9909</td>
<td>0.9911</td>
<td>0.9913</td>
<td>0.9916</td>
</tr>
<tr>
<td>2.4</td>
<td>0.9918</td>
<td>0.9920</td>
<td>0.9922</td>
<td>0.9925</td>
<td>0.9927</td>
<td>0.9929</td>
<td>0.9931</td>
<td>0.9932</td>
<td>0.9934</td>
<td>0.9936</td>
</tr>
</tbody>
</table>

Normal convergence

**Theorem 9.15 (DeMoivre-Laplace).**

Fix \( p \), and let \( X_n \sim \text{Bin}(n, p) \). Then for every fixed \( a < b \),

\[
\lim_{n \to \infty} \mathbb{P} \left( a < \frac{X_n - np}{\sqrt{np(1-p)}} \leq b \right) = \Phi(b) - \Phi(a).
\]

- That is, take \( X \sim \text{Bin}(n, p) \) with large \( n \) and fixed \( p \).
- Then \( \frac{X - np}{\sqrt{np(1-p)}} \) is approximately \( N(0, 1) \) distributed.
- This is a special case of the Central Limit Theorem, Theorem 11.21.
Example 9.16.

The height of a randomly selected male student at Bristol has a normal distribution with mean 1.75m and standard deviation 0.05m.

A student is chosen at random. What is the probability his height is greater than 1.86m?

Let $X$ be the height of the student, so $X \sim N(1.75, 0.05^2)$.

Let $Z = \frac{X - 1.75}{0.05}$ so that $Z \sim N(0, 1)$.

$$\mathbb{P}(X > 1.86) = \mathbb{P}\left(\frac{X - 1.75}{0.05} > \frac{1.86 - 1.75}{0.05}\right) = \mathbb{P}(Z > 2.2) = 1 - \mathbb{P}(Z \leq 2.2) = 1 - \Phi(2.2)$$

Can find $\Phi(2.2)$ from tabulated values, or (when you’ve done Stats 1) using a computer language called R.

The value is $\Phi(2.2) = 0.9861$. So $\mathbb{P}(Z > 1.86) = 0.0139$.

Fact, proved in Section 11.4

Lemma 9.17.

If $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\nu, \tau^2)$ are independent then

$$X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2).$$

Very few random variables have this property that you can add them and still get a distribution in the same family.

Compare with the addition of Poissons in Theorem 6.18.

See Lemma 11.18 for full proof.
Definition 9.18.

- Let $X$ and $Y$ be continuous r.v.s. They are jointly distributed with density function $f_{X,Y}(x,y)$ if for any region $A \subset \mathbb{R}^2$

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

- Marginal density for $X$ is $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dy$.

- Conditional for $X$ given $Y = y$ is $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$.

- Similarly for $Y$.

- $X$ and $Y$ are independent iff $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.

Time to wait for a lift while hitchhiking

Example 9.19.

- You choose a site to hitchhike at random.
- Let $X$ be the site type and assume $X \sim \text{Exp}(1)$.
- If the site type is $x$ it takes an $\text{Exp}(x)$ amount of time to get a lift (so large $x$ is good).
- We have been given

$$
 f_X(x) = e^{-x} \quad x > 0 \\
 f_{T|X}(t|x) = xe^{-xt} \quad x, t > 0
$$

- Thus $f_{X,T}(x,t) = f_{T|X}(t|x)f_X(x) = xe^{-(t+1)x}$ for $x, t > 0$.
- Hence

$$
 f_T(t) = \int_{-\infty}^{\infty} f_{X,T}(x,t) \, dx = \int_0^{\infty} xe^{-(t+1)x} \, dx = \frac{\Gamma(2)}{(t+1)^2} = \frac{1}{(t+1)^2}.
$$

- Finally, $\mathbb{P}(T > t) = \int_t^{\infty} f_T(\tau) \, d\tau = \int_t^{\infty} \frac{1}{(\tau+1)^2} \, d\tau = \left[ -\frac{1}{\tau+1} \right]_t^{\infty} = \frac{1}{t+1}$. 

\begin{align*}
\frac{\Gamma(2)}{(t+1)^2} &= \frac{1}{(t+1)^2} \\
\int_t^{\infty} \frac{1}{(\tau+1)^2} \, d\tau &= \left[ -\frac{1}{\tau+1} \right]_t^{\infty} = \frac{1}{t+1}.
\end{align*}
Section 10: Conditional expectation

Objectives: by the end of this section you should be able to

- Calculate conditional expectations.
- Understand the difference between function $\mathbb{E}[X|Y = y]$ and random variable $\mathbb{E}[X|Y]$.
- Perform calculations with these quantities.
- Use conditional expectations to perform calculations with random sums.

[This material is also covered in Section 7.4 of the course book]

Section 10.1: Introduction

- We have a pair of r.v.s $X$ and $Y$.
- Recall that we define

<table>
<thead>
<tr>
<th></th>
<th>pmf (discrete)</th>
<th>pdf (continuous)</th>
</tr>
</thead>
<tbody>
<tr>
<td>joint</td>
<td>$p_{X,Y}(x,y)$</td>
<td>$f_{X,Y}(x,y)$</td>
</tr>
<tr>
<td>marg.</td>
<td>$p_X(x) = \sum_y p_{X,Y}(x,y)$</td>
<td>$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) , dy$</td>
</tr>
<tr>
<td></td>
<td>$p_Y(y) = \sum_x p_{X,Y}(x,y)$</td>
<td>$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) , dx$</td>
</tr>
<tr>
<td>cond.</td>
<td>$p_{X</td>
<td>Y}(x</td>
</tr>
<tr>
<td></td>
<td>$p_{Y</td>
<td>X}(y</td>
</tr>
</tbody>
</table>
Definition 10.1.
Define $\mathbb{E}(X \mid Y = y)$ to be the expected value of $X$ using the conditional distribution of $X$ given that $Y = y$:

\[
\mathbb{E}(X \mid Y = y) = \begin{cases} 
\sum_x x p_{X,Y}(x \mid y) & X \text{ discrete} \\
\int_{-\infty}^{\infty} x f_{X,Y}(x \mid y) \, dx & X \text{ continuous}
\end{cases}
\]

Example

Example 10.2.

- $X, Y$ discrete

\[
\begin{array}{c|cccc}
p_{X,Y}(x, y) & y = 0 & 1 & 2 & 3 \\
\hline
x = 0 & \frac{1}{3} & 0 & 0 & 0 \\
1 & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\
2 & \frac{1}{3} & \frac{2}{3} & \frac{1}{3} & 0 \\
3 & \frac{1}{3} & \frac{3}{3} & \frac{3}{3} & \frac{1}{3} \\
\hline
p_Y(y) & \frac{32}{32} & \frac{32}{32} & \frac{32}{32} & \frac{32}{32}
\end{array}
\]

- For $\mathbb{E}(X \mid Y = 0)$: $p_{X,Y}(x \mid 0) = \frac{p_{X,Y}(x, 0)}{p_Y(0)} = \frac{32}{15} p_{X,Y}(x, 0)$ so

\[
\begin{array}{c|cccc}
x & 0 & 1 & 2 & 3 \\
\hline
p_{X,Y}(x \mid 0) & \frac{8}{15} & \frac{4}{15} & \frac{4}{15} & \frac{1}{15}
\end{array}
\]

So $\mathbb{E}(X \mid Y = 0) = 0 \times \frac{8}{15} + 1 \times \frac{4}{15} + 2 \times \frac{4}{15} + 3 \times \frac{1}{15} = \frac{11}{15}$. 

Example (cont.)

**Example 10.2.**

Similarly

\[
\begin{align*}
\mathbb{E}(X \mid Y = 1) &= 0 \times 0 + 1 \times \frac{4}{11} + 2 \times \frac{4}{11} + 3 \times \frac{3}{11} = \frac{21}{11} \\
\mathbb{E}(X \mid Y = 2) &= 0 \times 0 + 1 \times 0 + 2 \times \frac{2}{5} + 3 \times \frac{3}{5} = \frac{13}{5} \\
\mathbb{E}(X \mid Y = 3) &= 0 \times 0 + 1 \times 0 + 2 \times 0 + 3 \times 1 = 3
\end{align*}
\]

**Remark 10.3.**

*It is vital to understand that:*

- \( \mathbb{E}(X) \) is a number
- \( \mathbb{E}(X \mid Y = y) \) is a function – specifically a function of \( y \) (call it \( A(y) \)).
- We also define random variable \( \mathbb{E}(X \mid Y) = A(Y) \) (pick value of \( Y \) randomly).
- Good to spend time thinking which type of object is which.

---

**Section 10.2: Expectation of a conditional expectation**

**Theorem 10.4 (Tower law aka Law of Total Expectation).**

*For any random variables \( X \) and \( Y \), the \( \mathbb{E}[X \mid Y] \) is a random variable, with*

\[
\mathbb{E}(X) = \mathbb{E}(\mathbb{E}[X \mid Y])
\]

**Remark 10.5.**

- For \( Y \) discrete

\[
\mathbb{E}(\mathbb{E}[X \mid Y]) = \sum_y \mathbb{E}(X \mid Y = y) \Pr(Y = y).
\]

- For \( Y \) continuous

\[
\mathbb{E}(\mathbb{E}[X \mid Y]) = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y) \, dy.
\]
Proof of Theorem 10.4

Proof. 

(For discrete $Y$)

\[
\mathbb{E}(X) = \sum_x x \mathbb{P}(X = x)
\]

\[
= \sum_x x \sum_y \mathbb{P}(X = x \mid Y = y) \mathbb{P}(Y = y) \quad \text{by Partition Theorem 3.9}
\]

\[
= \sum_y \left[ \sum_x x \mathbb{P}(X = x \mid Y = y) \right] \mathbb{P}(Y = y)
\]

\[
= \sum_y \mathbb{E}(X \mid Y = y) \mathbb{P}(Y = y)
\]

For the continuous case, replace the sums with integrals and $\mathbb{P}(Y = y)$ with $f_Y(y)$.

Example 10.6.

Recall from Example 10.2

<table>
<thead>
<tr>
<th>$y$</th>
<th>$0$</th>
<th>$1$</th>
<th>$2$</th>
<th>$3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_Y(y)$</td>
<td>$\frac{15}{32}$</td>
<td>$\frac{11}{32}$</td>
<td>$\frac{5}{32}$</td>
<td>$\frac{1}{32}$</td>
</tr>
<tr>
<td>$\mathbb{E}(X \mid Y = y)$</td>
<td>$\frac{11}{15}$</td>
<td>$\frac{11}{11}$</td>
<td>$\frac{21}{13}$</td>
<td>$\frac{5}{3}$</td>
</tr>
</tbody>
</table>

Hence

\[
\mathbb{E}(X) = \frac{11 \times 15}{32} + \frac{21 \times 11}{32} + \frac{13 \times 5}{32} + 3 \frac{1}{32} = \frac{48}{32} = \frac{3}{2}.
\]

Direct calculation from $p_X(x)$ confirms

\[
\mathbb{E}(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{4} + 2 \times \frac{1}{4} + 3 \times \frac{1}{4} = \frac{3}{2}.
\]
Example 10.7.

A disoriented miner finds himself in a room of the mine with three doors:
- The first door brings him to safety after a 3 hours long hike.
- The second door takes him back to the same room after 5 hours of climbing.
- The third door takes him again back to the same room after 7 hours of exhausting climbing.

The disoriented miner chooses one of the three doors with equal chance independently each time he is in that room. What is the expected time after which the miner is safe?

Example 10.7.

Let $X$ be the time to reach safety, and $Y$ the initial choice of a door ($= 1, 2, 3$). Then using Theorem 10.4

\[
E(X) = E(E(X \mid Y)) \\
= E(X \mid Y = 1) \cdot P(Y = 1) + E(X \mid Y = 2) \cdot P(Y = 2) \\
+ E(X \mid Y = 3) \cdot P(Y = 3) \\
= 3 \cdot \frac{1}{3} + (E(X) + 5) \cdot \frac{1}{3} + (E(X) + 7) \cdot \frac{1}{3},
\]

which we rearrange as

\[
3E(X) = 15 + 2E(X); \quad E(X) = 15.
\]
Example

Example 10.8.

- Nuts in a wood have an intrinsic hardness $H$, a non-negative integer random variable.
- If a nut has hardness $H = h$ a squirrel takes a geometric $\frac{1}{h+1}$ number of attempts to crack the nut.
- The hardness of a randomly selected nut has a Poi(1) distribution.
- What is the expected number of attempts taken to crack a randomly selected nut?
- Let $X$ be the number of attempts. We want $\mathbb{E}(X)$.
- Given $H = h$, $X \sim \text{Geom}(\frac{1}{h+1})$, so $\mathbb{E}(X \mid H = h) = \frac{1}{h+1} = h + 1$.
- Therefore
  \[ \mathbb{E}(X) = \sum_{h=0}^{\infty} (h + 1) \mathbb{P}(H = h) = \mathbb{E}(H + 1) = \mathbb{E}(H) + 1 = 1 + 1 = 2. \]

Important notation

Remark 10.9.

- Remember from Remark 10.3 $\mathbb{E}(X \mid Y = y)$ is a function of $y$.
- Set $A(y) = \mathbb{E}(X \mid Y = y)$.
- Then the tower law Theorem 10.4 gives
  \[ \mathbb{E}(X) = \sum_{y} \mathbb{E}(X \mid Y = y) \mathbb{P}(Y = y) = \sum_{y} A(y) \mathbb{P}(Y = y) = \mathbb{E}(A(Y)). \]
- Remember $A(Y)$ is a random variable that we often write as $\mathbb{E}(X \mid Y)$.

Example 10.10.

- In the nut example, Example 10.8 $A(h) = \mathbb{E}(X \mid H = h) = h + 1$.
- Hence $A(H) = H + 1$ i.e. $\mathbb{E}(X \mid H) = H + 1$ [NB FUNCTION OF H].
- Therefore $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid H)) = \mathbb{E}(H + 1) = \mathbb{E}(H) + 1 = 2.$
Section 10.3: Conditional variance

**Definition 10.11.**
The *conditional variance of $X$, given $Y$* is

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

- No surprise here, just use conditionals everywhere in the definition of variance.
- Notice that $\text{Var}(X | Y)$ is again a function of $Y$ (a random variable).
- If we write $A(Y)$ for $[\mathbb{E}(X | Y)]$ then we can rewrite Definition 10.11 as
  $$\text{Var}(X | Y) = \mathbb{E}(X^2 | Y) - A(Y)^2. \quad (10.1)$$

**Law of total variance**

**Proposition 10.12.**
The *law of total variance* holds:

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

- In words: the variance is the expectation of the conditional variance plus the variance of the conditional expectation.
- Note that since $\text{Var}(X | Y)$ and $\mathbb{E}(X | Y)$ are random variables, it makes sense to take their mean and variance.
- They are both functions of $Y$, so implicitly these are taken over $Y$. 
Proof of Proposition 10.12 (not examinable)

Proof.

- Again we write $A(Y)$ for $\mathbb{E}(X \mid Y)$.
- Taking the expectation (over $Y$) of Equation (10.1) and applying the tower law Theorem 10.4 gives

\[
\mathbb{E}(\text{Var}(X \mid Y)) = \mathbb{E}(\mathbb{E}(X^2 \mid Y) - A(Y)^2) = \mathbb{E}(X^2) - \mathbb{E}(A(Y)^2) \quad (10.2)
\]

- Similarly, since Theorem 10.4 gives $\mathbb{E}(A(Y)) = \mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}(X)$:

\[
\text{Var}(\mathbb{E}(X \mid Y)) = \text{Var}(A(Y)) = \mathbb{E}(A(Y)^2) - (\mathbb{E}(A(Y)))^2 = \mathbb{E}(A(Y)^2) - (\mathbb{E}(X))^2. \quad (10.3)
\]

- Notice that first term of (10.3) is minus the second term of (10.2).

Proof of Proposition 10.12 (cont.)

Proof.

- Hence adding (10.2) and (10.3) together, cancellation occurs and we obtain:

\[
\mathbb{E}\text{Var}(X \mid Y) + \text{Var}\mathbb{E}(X \mid Y) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \text{Var}(X).
\]
Section 10.4: Random sum

Definition 10.13.

- Let \( X_1, X_2, \ldots \) be IID random variables with the same distribution as a random variable \( X \).
- Let \( N \) be a non-negative integer valued random variable which is independent of \( X_1, X_2, \ldots \).
- Let \( S = \begin{cases} 0 & \text{if } N = 0 \\ X_1 + X_2 + \cdots + X_N & \text{if } N \geq 1. \end{cases} \)
- We call \( S \) a random sum.

Random sum examples


- Look at the total value of insurance claims made in one year.
- Let \( N \) be the number of claims, and \( X_i \) be the value of the \( i \)th claim.
- Then \( S = X_1 + X_2 + \cdots + X_N \) is the total value of claims.
- Does it make sense that \( N \) and \( X \) are independent?

Example 10.15 (Inviting friends to a party).

- Let \( N \) be the number of friends invited
- Let \( X_i = \begin{cases} 0 & \text{if the } i\text{th invited person does not come} \\ 1 & \text{if the } i\text{th invited person does come} \end{cases} \)
- Then \( S = X_1 + \cdots + X_N \) is the total number of people at the party.
Example 10.16 (Number of grandchildren).

- Suppose a rabbit produces $N$ offspring.
- The $i$th child rabbit then produces $X_i$ offspring itself, so that these offspring are the grandchildren of the original rabbit.
- Then $S = X_1 + \cdots + X_N$ is the total number of grandchildren.

Random sum theorem

Theorem 10.17.

For any random sum of the form of Definition 10.13

$$\mathbb{E}(S) = \mathbb{E}(N)\mathbb{E}(X).$$

Proof.

- $N$ is random, but we can condition on its value.
- Let $A(n) = \mathbb{E}(S \mid N = n)$. Then

$$A(n) = \mathbb{E}(X_1 + \cdots + X_N \mid N = n)$$
$$= \mathbb{E}(X_1 + \cdots + X_n \mid N = n)$$
$$= \mathbb{E}(X_1 + \cdots + X_n) \text{ since the } X_i \text{ are independent of } N$$
$$= n\mathbb{E}(X)$$

So $A(N) = \mathbb{E}(S \mid N) = N\mathbb{E}(X)$.

Therefore $\mathbb{E}(S) = \mathbb{E}(\mathbb{E}(S \mid N)) = \mathbb{E}(N\mathbb{E}(X)) = \mathbb{E}(X)\mathbb{E}(N)$. 

Oliver Johnson (maotj@bris.ac.uk)  Probability 1: @BristOliver  TB 1 ©UoB 2017  239 / 267
Section 11: Moment generating functions

**Objectives:** by the end of this section you should be able to
- Define and calculate the moment generating function of a random variable.
- Manipulate the moment generating function to calculate moments.
- Find the moment generating function of sums of independent random variables.
- Use moment generating functions to work with random sums.
- Know the moment generating function of the normal.
- Understand the sketch proof of the Central Limit Theorem.

[This material is also covered in Sections 7.6 and 8.3 of the course book]

Section 11.1: MGF definition and properties

**Definition 11.1.**
Let $X$ be a random variable. The *moment generating function* (MGF) $M_X : \mathbb{R} \to \mathbb{R}$ of $X$ is given by

$$M_X(t) = \mathbb{E}(e^{tX})$$
(defined for all $t$ such that $\mathbb{E}(e^{tX}) < \infty$).

- So $M_X(t) = \left\{ \begin{array}{ll} \sum_i e^{tx_i} p_X(x_i) & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx & X \text{ cts} \end{array} \right.$
- The moment generating function is a way of encoding the information in the original pmf or pdf.
- In this Section we will see ways in which this encoding is useful.
Example: geometric

Example 11.2.

- Consider $X \sim \text{Geom}(p)$

$$M_X(t) = \sum_{x=1}^{\infty} e^{tx} p(1 - p)^{x-1}$$

$$= \sum_{x=1}^{\infty} pe^t ((1 - p)e^t)^{x-1}$$

$$= pe^t \sum_{y=0}^{\infty} ((1 - p)e^t)^y$$

$$= \frac{pe^t}{1 - (1 - p)e^t} \text{ defined for } (1 - p)e^t < 1$$

Example: Poisson

Example 11.3.

- Consider $X \sim \text{Poi}(\lambda)$

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!}$$

$$= e^{-\lambda} \sum_{x=0}^{\infty} \frac{1}{x!} (\lambda e^t)^x$$

$$= e^{\lambda(e^t-1)}.$$
Example: exponential

**Example 11.4.**

- Consider $X \sim \text{Exp}(\lambda)$

\[
M_X(t) = \int_0^\infty e^{tx} \lambda e^{-\lambda x} \, dx
\]

\[
= \lambda \int_0^\infty e^{-(\lambda-t)x} \, dx
\]

\[
= \frac{\lambda}{\lambda - t} \left[ -e^{-(\lambda-t)} \right]_0^\infty
\]

\[
= \frac{\lambda}{\lambda - t} \text{ defined for } t < \lambda
\]
$M_X$ uniquely defines the distribution of $X$.

**Theorem 11.6.**

**Uniqueness of the MGF.**

- Consider random variables $X, Y$ such that that $M_X(t)$ and $M_Y(t)$ are finite on an interval $I \subseteq \mathbb{R}$ containing the origin.
- Suppose that
  $$M_X(t) = M_Y(t) \quad \text{for all } t \in I.$$
- Then $X$ and $Y$ have the same distribution.

**Proof.**

Not given.

---

**Moments**

**Definition 11.7.**

The $r$th moment of $X$ is $E(X^r)$.

**Lemma 11.8.**

For any random variable $X$ and for any $t$:

$$M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots = \sum_{r=0}^\infty \frac{t^r}{r!}E(X^r)$$

i.e. $M_X$ "generates" the moments of $X$. 
Proof of Lemma 11.8

Proof.

For any $t$, using the linearity of expectation:

\[
M_X(t) = \mathbb{E}(e^{tX})
= \mathbb{E}
\left[
1 + (tX) + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots
\right]
= 1 + t\mathbb{E}(X) + \frac{t^2}{2!}\mathbb{E}(X^2) + \frac{t^3}{3!}\mathbb{E}(X^3) + \cdots
\]

Note that $M_X(0) = \mathbb{E}(e^0) = 1$, as we’d expect.

Recovering moments of exponential

We can recover the moments of $X$ from $M_X(t)$ in two ways:

**Method 1** Expand $M_X(t)$ as a power series in $t$. The coefficient of $t^k$ is $\mathbb{E}(X^k)/k!$.

**Method 2** $M_X^{(k)}(0) = \mathbb{E}(X^k)$, where $M_X^{(k)}$ denotes the $k$th derivative of $M_X$.

To see this, note that

\[
M'_X(t) = \mathbb{E}(X) + t\mathbb{E}(X^2) + \frac{t^2}{2!}\mathbb{E}(X^3) + \cdots
\]

\[
M'_X(0) = \mathbb{E}(X)
\]

\[
M''_X(t) = \mathbb{E}(X^2) + t\mathbb{E}(X^3) + \frac{t^2}{2!}\mathbb{E}(X^4) + \cdots
\]

\[
M''_X(0) = \mathbb{E}(X^2)
\]

etc
Recovering moments of exponential: example

**Example 11.9.**

- Consider \( X \sim \text{Exp}(\lambda) \).
- We know from Example 11.4 that \( M_X(t) = \frac{\lambda}{\lambda - t} \).
- To find \( E(X^r) \) use Method 1.
- \( M_X(t) = \frac{1}{1 - \frac{t}{\lambda}} = 1 + \frac{t}{\lambda} + \left( \frac{t}{\lambda} \right)^2 + \left( \frac{t}{\lambda} \right)^3 + \cdots \)
- Compare with \( M_X(t) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \cdots \)
- We see that \( \frac{E(X^k)}{k!} = \frac{1}{\lambda^k} \)
- Hence \( E(X^k) = \frac{k!}{\lambda^k} \).

Recovering moments of gamma

**Example 11.10.**

- Recall from Example 11.5 that \( M_X(t) = \lambda^\alpha (\lambda - t)^{-\alpha} \).
- To find \( E(X^r) \) use Method 2:
  \[
  M'_X(t) = \lambda^\alpha \alpha (\lambda - t)^{-\alpha-1} \\
  E(X) = M'_X(0) = \frac{\alpha}{\lambda} \\
  M''_X(t) = \lambda^\alpha \alpha (\alpha + 1) (\lambda - t)^{-(\alpha+2)} \\
  E(X^2) = M''_X(0) = \frac{\alpha(\alpha + 1)}{\lambda^2} \\
  
  This can be continued, but notice that with minimal work we can now see that
  \[
  \text{Var}(X) = E(X^2) - (E(X))^2 = \frac{\alpha(\alpha + 1)}{\lambda^2} - \left( \frac{\alpha}{\lambda} \right)^2 = \frac{\alpha}{\lambda^2}.
  \]
Section 11.2: Sums of random variables

**Theorem 11.11.**

Let \( X_1, X_2, \ldots, X_n \) be independent rvs and let \( Z = \sum_{i=1}^{n} X_i \). Then

\[
M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t).
\]

**Proof.**

- Since \( X_i \) are independent, then for fixed \( t \) so are \( e^{tX_i} \) (by Remark 6.14).

\[
M_Z(t) = \mathbb{E}(e^{tZ}) = \mathbb{E}(e^{t(\sum_{i=1}^{n} X_i)}) = \mathbb{E}\left(\prod_{i=1}^{n} e^{tX_i}\right)
\]

\[
= \prod_{i=1}^{n} \mathbb{E}(e^{tX_i}) \quad \text{by independence}
\]

\[
= \prod_{i=1}^{n} M_{X_i}(t).
\]

---

**Example: adding Poissons**

**Example 11.12 (cf Theorem 6.18).**

- If \( X \sim \text{Poi}(\lambda) \) and \( Y \sim \text{Poi}(\mu) \), we deduce using Example 11.3 and Theorem 11.11 that \( Z = X + Y \) has moment generating function

\[
M_Z(t) = M_X(t)M_Y(t) = e^{\lambda(e^t-1)} \cdot e^{\mu(e^t-1)}
\]

\[
= e^{(\lambda+\mu)(e^t-1)},
\]

- We deduce that (since it has the same MGF) \( Z \sim \text{Poi}(\lambda + \mu) \) by Theorem 11.6.
### Lemma 11.13.

- Let $X_1, X_2, \ldots, X_n$ be independent $\text{Exp}(\lambda)$ rvs, and let $Z = X_1 + \cdots + X_n$.
- Then $M_{X_i}(t) = \frac{\lambda}{\lambda - t}$ for each $i = 1, \ldots, n$.
- Thus by Theorem 11.11:

$$M_Z(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$$

and $Z \sim \text{Gamma}(n, \lambda)$ by the uniqueness theorem (Theorem 11.6)

---

### Section 11.3: Random sums

- In Theorem 10.17 we saw how to calculate the expectation of a random sum $S$.
- e.g. insurance company cares about the distribution of the total claims in a year.
- What if we want the full distribution of $S$?

### Theorem 11.14.

Consider $X_1, X_2, \ldots$ iid with distribution the same as $X$, and $N$ is a non-negative integer-valued rv independent of the $X_i$. Then

$$S = \begin{cases} 0 & N = 0 \\ X_1 + \cdots + X_N & N > 0 \end{cases}$$

has MGF satisfying

$$M_S(t) = M_N(\log M_X(t))$$
Proof of Theorem 11.14

Proof.

Let \( A(n) = \mathbb{E}(e^{tS} \mid N = n) \)
\[ = \mathbb{E}(e^{t(X_1 + \cdots + X_N)} \mid N = n) \]
\[ = \mathbb{E}(e^{t(X_1 + \cdots + X_n)} \mid N = n) \]
\[ = \mathbb{E}(e^{tX_1 + \cdots + X_n}) \quad \text{since the } X_i \text{s are independent of } N \]
\[ = \mathbb{E}(e^{tX_1}) \cdots \mathbb{E}(e^{tX_n}) \quad \text{since the } X_i \text{s are independent} \]
\[ = (M_X(t))^n \]
\[ = e^{n \log M_X(t)} \]

Thus \( \mathbb{E}(e^{tS} \mid N) = A(N) = e^{N \log M_X(t)} \) and by Theorem 10.4

\[ M_S(t) = \mathbb{E}(e^{tS}) = \mathbb{E}(\mathbb{E}(e^{tS} \mid N)) = \mathbb{E}(e^{N \log M_X(t)}) = M_N(\log M_X(t)) \]

Example

Example 11.15.

- Suppose the number of insurance claims in one year is \( N \sim \text{Poi}(\lambda) \).
- Suppose claims are IID \( X_i \sim \text{Exp}(1) \), and these are independent of \( N \).
- Let \( S = X_1 + X_2 + \cdots + X_N \) be the total claim.
- Start by calculating

\[ M_N(t) = \sum_{n=0}^{\infty} e^{tn} \frac{\lambda^n e^{-\lambda}}{n!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}. \]

We also know that \( M_X(t) = \frac{1}{1-t} \). So

\[ M_S(t) = M_N(\log M_X(t)) = e^{\lambda(e^{\log M_X(t)-1})} = e^{\lambda(M_X(t)-1)} \]
\[ = e^{\lambda(\frac{1}{1-t}-1)} = e^{\lambda(\frac{t}{1-t})}. \]

- From this we can calculate \( \mathbb{E}(S) \), \( \text{Var}(S) \), etc.
Example 11.16.

- Let $X \sim \mathcal{N}(0,1)$.
- So $M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx$.
- Let $y = x - t$:

$$M_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(y+t)} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y+t)^2}{2}} \, dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[-2ty-2t^2+y^2+yt+t^2]} \, dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[y^2-t^2]} \, dy$$

$$= e^{\frac{1}{2}t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} \, dy$$

$$= e^{\frac{1}{2}t^2}$$

MGF of the general normal

Example 11.17.

- Now let $Y \sim \mathcal{N}(\mu, \sigma^2)$
- Set $X = \frac{Y-\mu}{\sigma^2}$ so $X \sim \mathcal{N}(0,1)$ by Lemma 9.13.
- Then $Y = \mu + \sigma X$ and

$$M_Y(t) = \mathbb{E}(e^{ty}) = \mathbb{E}(e^{t(\mu+\sigma X)})$$

$$= \mathbb{E}(e^{\mu t} e^{\sigma t X}) = e^{\mu t} \mathbb{E}(e^{(\sigma t) X})$$

$$= e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{1}{2}(\sigma t)^2}$$

$$= e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$
Normal distribution properties

**Lemma 11.18.**

1. If \( X \sim \mathcal{N}(\mu, \sigma^2) \) and \( c \) is a constant then \( X + c \sim \mathcal{N}(\mu + c, \sigma^2) \).
2. If \( X \sim \mathcal{N}(\mu, \sigma^2) \) and \( \beta \) is a constant then \( \beta X \sim \mathcal{N}(\beta \mu, \beta^2 \sigma^2) \).
3. If \( X \) and \( Y \) are independent with \( X \sim \mathcal{N}(\mu_X, \sigma_X^2) \) and \( Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \) then
   \[
   X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2).
   \]

Note: Properties 1 and 2 can easily be shown using transformation of variables. We use MGFs to prove all three here.

**Proof.**

1. Let \( Y = X + c \). Then
   \[
   M_Y(t) = \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t(X+c)}) = e^{tc} \mathbb{E}(e^{tX}) = e^{tc} M_X(t) = e^{tc} e^{\mu t + \frac{1}{2} \sigma^2 t^2} = e^{(\mu+c)t + \frac{1}{2} \sigma^2 t^2}
   \]
   So \( Y \sim \mathcal{N}(\mu + c, \sigma^2) \) by uniqueness, Theorem 11.6.

2. Let \( Y = \beta X \). Then
   \[
   M_Y(t) = \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t\beta X}) = M_X(\beta t) = e^{\mu \beta t + \frac{1}{2} \sigma^2 \beta^2 t^2} = e^{\mu \beta t + \frac{1}{2} \beta^2 \sigma^2 t^2}
   \]
   So \( Y \sim \mathcal{N}(\beta \mu, \beta^2 \sigma^2) \) by uniqueness, Theorem 11.6.

3. Let \( Z = X + Y \). Then by Theorem 11.11
   \[
   M_Z(t) = M_X(t)M_Y(t) = e^{\mu_X t + \frac{1}{2} \sigma_X^2 t^2} e^{\mu_Y t + \frac{1}{2} \sigma_Y^2 t^2} = e^{(\mu_X + \mu_Y) t + \frac{1}{2} (\sigma_X^2 + \sigma_Y^2) t^2}
   \]
   So \( Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \) by uniqueness, Theorem 11.6.
Example: heights

Example 11.19.

- Heights of male students are $\mathcal{N}(175, 33)$ and heights of female students are $\mathcal{N}(170, 25)$.
- One female and three male students are chosen at random.
- What is the probability that the female is taller than the average height of the three males?
- Let $X_1, X_2, X_3$ be the height of the three male students, and $Y$ be the height of the female student.
- We have $X_i \sim \mathcal{N}(175, 33)$ and $Y \sim \mathcal{N}(170, 25)$.
- By Lemma 11.18.3,
  $$X_1 + X_2 + X_3 \sim \mathcal{N}(175 + 175 + 175, 33 + 33 + 33).$$

Example: heights (cont.)

Example 11.19.

- Let $W = \frac{X_1 + X_2 + X_3}{3}$ be the average height of the male students. By Lemma 11.18.2
  $$W \sim \mathcal{N}\left(\frac{1}{3}(3 \times 175), \left(\frac{1}{3}\right)^2(3 \times 33)\right) = \mathcal{N}(175, 11).$$
- Let the difference $D = Y - W = Y + (-W)$.
- We know $Y \sim \mathcal{N}(170, 25)$, and $(-W) \sim \mathcal{N}(-175, 11)$ by Lemma 11.18.2.
- So $D \sim \mathcal{N}(170 + (-175), 25 + 11)$ by Lemma 11.18.3, i.e.
  $$D \sim \mathcal{N}(-5, 36) \text{ or } \frac{D+5}{6} \sim \mathcal{N}(0, 1).$$
- We want to know $P(D > 0) = P\left(\frac{D+5}{6} > \frac{5}{6}\right) = 1 - \Phi\left(\frac{5}{6}\right)$. Using tables or R we can find $\Phi\left(\frac{5}{6}\right) = 0.7976$, so $P(D > 0) = 1 - 0.7976 = 0.2024$. 
Section 11.5: Central Limit Theorem

- Consider IID $X_1, \ldots, X_n$ with mean $\mu$ and $\sigma^2$.
- The Weak Law of Large Numbers (Theorem 7.13) tells us that $\frac{1}{n} (X_1 + \ldots + X_n) \sim \mu$ or $X_1 + \ldots + X_n - n\mu \sim 0$.
- The Central Limit Theorem tells us how close these two quantities are (the approximate distribution of the difference).

We start with an auxiliary proposition without proof.

**Proposition 11.20.**

- Suppose $M_{Z_n}(t) \rightarrow M_Z(t)$ for every $t$ in an open interval containing 0.
- Then distribution functions converge: $F_{Z_n}(z) \rightarrow F_Z(z)$.

Central Limit Theorem

**Theorem 11.21 (Central Limit Theorem (CLT)).**

Let $X_1, X_2, \ldots$ be IID random variables with both their mean $\mu$ and variance $\sigma^2$ finite. Then for every real $a < b$:

$$\lim_{n \to \infty} \mathbb{P} \left( a < \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} < b \right) = \Phi(b) - \Phi(a).$$

**Remark 11.22.**

- Notice that $X_1 + \ldots + X_n$ has mean $n\mu$ and variance $n\sigma^2$.
- CLT implies that for large $n$, the $X_1 + \ldots + X_n \sim N(n\mu, n\sigma^2)$ or equivalently $\frac{1}{\sqrt{n\sigma^2}} (X_1 + \ldots + X_n) \sim N(0, 1)$.
- If $X_i \sim \text{Bernoulli}(p)$ this reduces to the de Moivre-Laplace Theorem 9.15.
Sketch proof.

- Will just consider the case $\mu = 0, \sigma^2 = 1$ for brevity.
- Write $M_X$ for the MGF of each $X_i$.
- Know that $M_X(t) = 1 + \frac{1}{2}t^2 + O(t^3)$.
- Consider $T_n := \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}}$. Its moment generating function is

$$M_{T_n}(t) = \mathbb{E}e^{t\sum_{i=1}^{n} \frac{X_i}{\sqrt{n}}} = \mathbb{E}\prod_{i=1}^{n} e^{t\frac{X_i}{\sqrt{n}}} = \prod_{i=1}^{n} \mathbb{E}e^{\left(\frac{t}{\sqrt{n}}\right)X_i}$$

$$= \prod_{i=1}^{n} M_{X_i}\left(\frac{t}{\sqrt{n}}\right) = \left[M_X\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

$$= \left(1 + \frac{1}{2} \frac{t^2}{n} + O(n^{-3/2})\right)^n \rightarrow e^{t^2/2},$$

as required.