Probability and Statistics (1st half) 2023: MATH10013

Professor Oliver Johnson
o.johnson@bristol.ac.uk
bristoliver.substack.com

School of Mathematics, University of Bristol

"You're coming of age in the 21st century. A century in which I promise you mathematics is going to play a starring role."
– President Josiah Bartlet, The West Wing *Series 1 Episode 17*

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, understand medical tests, study algorithms etc.
- The world is full of randomness and uncertainty: we need to understand it!

Course outline

- 20+2 lectures, 6 exercise classes (odd weeks), 6 mandatory HW sets (even weeks).
- 2 online quizzes (Weeks 4, 8) count 5% towards final module mark.
- **IT IS YOUR RESPONSIBILITY TO KEEP UP WITH 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**: 12pm Tuesdays, G83 Fry Building (Other times, I may be unavailable – but just email `maotj@bristol.ac.uk` to fix an appointment).
-

Contents

Textbook

- 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
- 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 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
- A *sample point* or *elementary outcome* ω is the result of a trial:
  - the number on the roulette wheel
  - the number on the dice
- The *sample space* Ω is the set of all possible elementary outcomes ω.

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 (r, g), such as

$$\omega = (6, 3)$$

  where r is the score on the red die and g 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.

Informal description	Formal description
A occurs or B occurs (or both)	$A \cup B$
A and B both occur	$A \cap B$
A does not occur	A^c
A occurs implies B occurs	$A \subseteq B$
A and B cannot both occur together (*disjoint* or *mutually exclusive*)	$A \cap B = \emptyset$

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

Section 1.2: Axioms of probability

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

- The probability $\mathbb{P}$ captures the intuitive idea that some events are more likely than others.
- We will give three axioms of probability ...
- ... and 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.

Axioms

Definition 1.6.

- Let $\mathbb{P}$ be a map from events $A \subseteq \Omega$ to the real numbers $\mathbb{R}$.
- For each event A (each subset of Ω) there is a number $\mathbb{P}(A)$.
- Then $\mathbb{P}$ is a *probability measure* if it satisfies:

  Axiom 1 $0 \leq \mathbb{P}(A) \leq 1$ for every event A.

  Axiom 2 $\mathbb{P}(\Omega) = 1$.

  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

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Deductions from the axioms

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

Lemma 1.7.

1. $\mathbb{P}(\emptyset) = 0$
2. *Axiom 3 implies a 'finite' version of the same result for* **disjoint** *events* $A_1, \ldots, A_n$, *("Property 2")*

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

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

Deductions (continued)

Proof.

1. Take $A_i = \emptyset$, then $\cup_{i=1}^{\infty} A_i = \emptyset$, so Axiom 3 gives

$$\mathbb{P}(\emptyset) = \sum_{i=1}^{\infty} \mathbb{P}(\emptyset),$$

   and hence $\mathbb{P}(\emptyset) = 0$.
2. This follows from Axiom 3 by taking $A_i = \emptyset$ for $i \geq n+1$.
3. To prove the complement result:
   - 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 $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$.

□

Some simple applications of the axioms (cont.)

Lemma 1.9.

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.
- Draw a Venn diagram!
- 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 $\mathbb{P}(B) \geq \mathbb{P}(A)$.

□

Inclusion–exclusion principle $n = 2$

Lemma 1.10.

Let A and B be any two events (not necessarily disjoint). 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
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c) \quad \text{(Property 2).} \tag{1.1}$$
- $B = (B \cap A) \cup (B \cap A^c)$ is a disjoint union, so
$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c) \quad \text{(Property 2).} \tag{1.2}$$
- Subtracting (1.2) from (1.1) we have
$\mathbb{P}(A \cup B) - \mathbb{P}(B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$.

□

More general inclusion–exclusion principle

Theorem 1.11.

For three events $A_1, \ldots, A_3$, we can write

$$\begin{aligned}\mathbb{P}\left(A_1 \bigcup A_2 \bigcup A_3\right) &= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \mathbb{P}(A_3) \\ &\quad - \mathbb{P}(A_1 \cap A_2) - \mathbb{P}(A_2 \cap A_3) - \mathbb{P}(A_3 \cap A_1) \\ &\quad + \mathbb{P}(A_1 \cap A_2 \cap A_3).\end{aligned}$$

Proof.

Not proved here – can you see the result for general n?

□

Boole's inequality – 'union bound'

Proposition 1.12 (Boole's inequality).

For any events $A_1, A_2, \ldots, A_n$, the $\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i)$.

Proof.

- Proof by induction. When $n = 2$, by Lemma 1.10:

$$\mathbb{P}\left(A_1 \cup A_2\right) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

- Now suppose true for n. Then

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{n+1} A_i\right) &= \mathbb{P}\left(\left(\bigcup_{i=1}^{n} A_i\right) \cup A_{n+1}\right) \leq \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) + \mathbb{P}(A_{n+1}) \\ &\leq \sum_{i=1}^{n} \mathbb{P}(A_i) + \mathbb{P}(A_{n+1}) = \sum_{i=1}^{n+1} \mathbb{P}(A_i). \end{aligned}$$

□

Key idea: de Morgan's Law

Theorem 1.13.

For any events A and B:

$$\begin{aligned} (A \cup B)^c = A^c \cap B^c \quad &\Longrightarrow \quad 1 - \mathbb{P}\left(A \cup B\right) = \mathbb{P}\left(A^c \cap B^c\right), \quad (1.3) \\ (A \cap B)^c = A^c \cup B^c \quad &\Longrightarrow \quad 1 - \mathbb{P}\left(A \cap B\right) = \mathbb{P}\left(A^c \cup B^c\right). \quad (1.4) \end{aligned}$$

Proof.

Draw a Venn diagram. □

Remark 1.14.

- *Swapping A and A^c, and B and B^c,* (1.3) *and* (1.4) *are equivalent.*
- (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'*
- *By a similar argument, can extend* (1.3) *and* (1.4) *to n events.*

Example

Example 1.15.

- 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 (1.3),

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

  since $\mathbb{P}\left(A^c \cap B^c\right) = \mathbb{P}(A^c)\mathbb{P}(B^c) = (1 - \mathbb{P}(A))(1 - \mathbb{P}(B))$
- Caution: This final step only works because two rolls are 'independent' (see later for much more on this!!)

Section 1.3: Equally likely sample points

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

- A common case is where each sample point has the same probability.
- e.g. symmetry says dice rolls have equal probability.
- Assume that
  - Ω, the sample space, is finite
  - all sample points are equally likely
- Then by Axiom 2 and Property 2, considering the disjoint union

$$1 = \mathbb{P}(\Omega) = \mathbb{P}\left(\bigcup_{\omega \in \Omega} \{\omega\}\right) = \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = |\Omega| \mathbb{P}(\{\omega\})$$

  we can see that

$$\mathbb{P}(\{\omega\}) = \frac{1}{\text{Number of points in } \Omega} = \frac{1}{|\Omega|}.$$

- Also, if $A \subseteq \Omega$, then $\mathbb{P}\left(\bigcup_{\omega \in A} \{\omega\}\right) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) = |A| \mathbb{P}(\{\omega\})$ so:

$$\mathbb{P}(A) = \frac{\text{Number of points in } A}{\text{Number of points in } \Omega} = \frac{|A|}{|\Omega|}.$$

Example: red and green dice.

Example 1.16.

- Return to the red and green dice, Example 1.2
- $$\Omega = \{(1,1),(1,2),\ldots,(6,6)\}$$
  with 36 sample points.
- By symmetry, assume that $\mathbb{P}(\{\omega\}) = \frac{1}{36}$ for each ω (i.e. equally likely outcomes).
- For each i, let A_i be the event that the sum of the scores is i:
  $$A_5 = \{(1,4),(2,3),(3,2),(4,1)\} \text{ so } |A_5| = 4 \text{ and } \mathbb{P}(A_5) = \frac{4}{36} = \frac{1}{9}.$$
- **Exercise:** Show that
  $$\mathbb{P}(A_4) = \frac{1}{12}, \qquad \mathbb{P}(A_3) = \frac{1}{18}, \qquad \mathbb{P}(A_2) = \frac{1}{36}.$$

Section 1.4: Permutations and combinations

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

Definition 1.17.

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

Example 1.18.

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

- Gold medal winner can be chosen in 8 ways.
- For each gold medal winner, the silver medal can go to one of the other 7 swimmers, so there are 8×7 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 1.19.

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

Remark 1.20.

Check the special cases:

$r = n$: ${}^nP_n = \frac{n!}{(n-n)!} = \frac{n!}{1} = n!$, *so there are* $n!$ *ways of ordering* n *objects.*

$r = 1$: ${}^nP_1 = \frac{n!}{(n-1)!} = n$, *so there are* n *ways of choosing 1 of* n *objects.*

BANANA example[1]

Can extend this analysis to situations with multiple objects of the same type:

Example 1.21.

- 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.
- (If you like, think about labelling A_1, A_2 and A_3)
- So each word is produced by $3! \times 2!$ orderings of letters A and N.
- So the total number of distinct words is

$$\frac{6!}{3!2!1!} = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{3 \times 2 \times 1 \times 2 \times 1 \times 1} = \frac{6 \times 5 \times 4}{2} = 60.$$

[1] This kind of analysis was first performed by al-Farahidi in Iraq in the 8th Century

Combinations

Definition 1.22.

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

Example 1.23.

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 1.24.

- *More generally, think about choosing r where the order is important: this can be done in ${}^nP_r = \frac{n!}{(n-r)!}$ 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*

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

different ways if order doesn't matter.

- *At school many of you will have written ${}^nC_r$ for this binomial coefficient. Please use this new notation from now onwards.*

Example

Example 1.25.

- 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

$$\binom{52}{5} = \frac{52!}{47! \times 5!} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1}$$

distinct hands.

Properties of binomial coefficients

Proposition 1.26.

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

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}.$$

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:* $2^n = \sum_{r=0}^{n} \binom{n}{r}$.

[a] In fact, dates back to Indian mathematician Pingala, 2nd century B.C.

Proof.

1. Choosing r objects to be included is the same as choosing $(n-r)$ objects to be excluded.
2. Consider choosing r objects out of n, and paint one object red. Either
   - the red object *is* chosen, and the remaining $r-1$ objects need to be picked out of $n-1$, or
   - the red object *is not* chosen, and all r 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 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.

□

Section 1.5: Counting examples

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

Example 1.27.

- A fair coin is tossed n times.
- Represent the outcome of the experiment by, e.g. $(H, T, T, \dots, H, T)$.
- $\Omega = \{(s_1, s_2, \dots, s_n) : s_i = H \text{ or } T, i = 1, \dots, n\}$ 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 $\omega = (s_1, s_2, \dots, s_n)$ with exactly r of the s_i being a head.
- There are $\binom{n}{r}$ different ways to choose the r elements of ω to be a head, so $|A_r| = \binom{n}{r}$.

Example 1.27.

- Therefore $\mathbb{P}(\text{Exactly } r \text{ heads}) = \mathbb{P}(A_r) = \frac{\binom{n}{r}}{2^n}$.
-
$$\sum_{r=0}^{n} \mathbb{P}(A_r) = \sum_{r=0}^{n} \frac{\binom{n}{r}}{2^n} = \frac{1}{2^n} \sum_{r=0}^{n} \binom{n}{r} = \frac{1}{2^n} 2^n = 1,$$
using the Binomial Theorem, Proposition 1.26.4.
- Example of binomial distribution . . . see Definition 3.10 later.

Example: Bridge hand

Example 1.28.

- We 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. $|\Omega| = \binom{52}{13}$.
- We assume these are equally likely.

Example 1.28.

- Now we calculate the number of hands containing the JQKA of spades.
- Each of these hands contains those four cards, and 9 other cards from the remaining 48 cards in the pack.
- So there are $|A| = \binom{48}{9}$ different hands containing JQKA of spades.
-

$$\begin{aligned}\mathbb{P}(\text{JQKA spades}) &= \frac{\binom{48}{9}}{\binom{52}{13}} = \frac{\frac{48!}{9!39!}}{\frac{52!}{13!39!}} = \frac{48!13!}{52!9!} \\ &= \frac{13 \times 12 \times 11 \times 10}{52 \times 51 \times 50 \times 49} = \frac{17160}{6497400} \simeq 0.00264.\end{aligned}$$

- Roughly 0.2% chance, or 1 in 400 hands.

Example: Birthdays

Example 1.29.

- 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 ith person have a birthday on day a_i, and assume $1 \leq a_i \leq 365$.
- The m-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 1.29.

- 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

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

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

- For example,

$$\begin{aligned} \mathbb{P}(B_{23}) &\approx 0.493 \\ \mathbb{P}(B_{40}) &\approx 0.109 \\ \mathbb{P}(B_{60}) &\approx 0.006 \end{aligned}$$

Section 2: 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 2.1: Motivation and definitions

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

Remark 2.1.

- *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 2.2.

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

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

(Sometimes also call this the ‘probability of A conditioned on B’)

Example: Sex of children

Example 2.3.

- Choose a family at random from 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: $\Omega = \{(b,b),(b,g),(g,b),(g,g)\}$.

$$\begin{aligned} A &= \{(b,b)\} = \text{"both boys"} \\ B &= \{(b,b),(b,g),(g,b)\} = \text{"at least one boy"} \\ A \cap B &= \{(b,b)\} \\ \mathbb{P}(A \cap B) &= 1/4 \\ \mathbb{P}(B) &= 3/4 \\ \mathbb{P}(A \mid B) &= \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} \end{aligned}$$

Section 2.2: Reduced sample space

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

Example 2.4.

- 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 1.16 that $\mathbb{P}(B) = \frac{4}{36}$.
- Clearly $A \cap B = \{(2,3)\}$, so $\mathbb{P}(A \cap B) = \frac{1}{36}$.
- Hence

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

Reduced sample space

Example 2.4.

- 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 2.5.

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

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

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

Sketch proofs

1. By Axiom 1 and Lemma 1.9, 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 2.5 shows that $\mathbb{P}(\cdot \mid B)$ 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 2.6.

For example for fixed set B:

- $\mathbb{P}(A^c \mid B) = 1 - \mathbb{P}(A \mid B)$.
- $\mathbb{P}(\emptyset \mid B) = 0$.
- $\mathbb{P}(A \cup C \mid B) = \mathbb{P}(A \mid B) + \mathbb{P}(C \mid B) - \mathbb{P}(A \cap C \mid B)$.

Remark 2.7.

WARNING: DON'T CHANGE THE CONDITIONING: e.g. $\mathbb{P}(A \mid B)$ and $\mathbb{P}(A \mid B^c)$ have nothing to do with each other.

Section 2.3: Partition theorem

Definition 2.8.

A collection of events $B_1, B_2, \ldots, B_n$ is a disjoint partition of Ω, if

- $B_i \cap B_j = \emptyset$ if $i \neq j$, and
- $\bigcup_{i=1}^n B_i = \Omega$.

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

Theorem 2.9 (Partition Theorem).

Let A be an event. Let $B_1, B_2, \ldots, B_n$ be a disjoint partition of Ω with $\mathbb{P}(B_i) > 0$ for all i. Then

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

Proof of Partition theorem, Theorem 2.9

Proof.

- Write $C_i = A \cap B_i$.
- Then for $i \neq j$ the $C_i \cap C_j = (A \cap B_i) \cap (A \cap B_j) = A \cap (B_i \cap B_j) = \emptyset$.
- Also $\bigcup_{i=1}^n C_i = \bigcup_{i=1}^n (A \cap B_i) = A \cap (\bigcup_{i=1}^n B_i) = A \cap \Omega = A$.
- 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
- But $\mathbb{P}(C_i) = \mathbb{P}(A \cap B_i) = \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$ by the definition of conditional probability, so

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

□

- Note: In the proof of Lemma 1.10, we saw that $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$, just as here. In fact, B and B^c is a disjoint partition of Ω.

Example: Diagnostic test

Example 2.10.

- A test for a disease gives a positive result 90% of the time when the disease is present, and 20% of the time when it is absent.
- It is known that 1% of the population have the disease.
- In a randomly selected member of the population, what is the probability of getting a positive test result?
- Let B_1 be the event "has disease": $\mathbb{P}(B_1) = 0.01$.
- Let $B_2 = B_1^c$ be the event "no disease": $\mathbb{P}(B_2) = 0.99$.
- Let A be the event "positive test result".
- We are told: $\mathbb{P}(A \mid B_1) = 0.9 \qquad \mathbb{P}(A \mid B_2) = 0.2$.
- Therefore

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

Important advice

Remark 2.11.

- *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 2.4: Bayes' theorem

- We saw in Definition 2.2 that $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$.
- We also have $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\mathbb{P}(A)$.
- So $\mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$ and therefore

Theorem 2.12 (Bayes' theorem).

For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}. \tag{2.1}$$

- 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

$$\mathbb{P}(\text{observation} \mid \text{hypothesis}) \leftrightarrow \mathbb{P}(\text{hypothesis} \mid \text{observation}).$$

Alternative form of Bayes'

Theorem 2.13 (Bayes' theorem – partition form).

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

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

Proof.

- We have already seen in (2.1) that

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

- The partition theorem (Theorem 2.9) tells us that $\mathbb{P}(A) = \sum_{i=1}^n \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$.
- The result follows immediately. □

Example: Diagnostic test revisited

- In Example 2.10, the observation is the positive test result, and the hypothesis is that you have the disease.

Example 2.14.

- A test for a disease gives positive results 90% of the time when the disease is present, and 20% of the time when it is absent.
- It is known that 1% of the population have the disease.
- A randomly chosen 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 (2.1), since we already know $\mathbb{P}(A) = 0.207$.
- So $\mathbb{P}(B_1 \mid A) = \frac{\mathbb{P}(A \mid B_1)\mathbb{P}(B_1)}{\mathbb{P}(A)} = \frac{0.9 \times 0.01}{0.207} = 0.0435 \quad (3.s.f.)$

Example: Prosecutor's fallacy

Example 2.15.

- 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.
- $$\mathbb{P}(E \mid G) = 1, \qquad \mathbb{P}(E \mid G^c) = 10^{-6}.$$

Example 2.15.

- We want to know $\mathbb{P}(G \mid E)$, so use Bayes' theorem.
- We need to know $\mathbb{P}(G)$.
- 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: $\mathbb{P}(G) = 10^{-7}$.
- Hence

$$\begin{aligned}\mathbb{P}(G \mid E) &= \frac{\mathbb{P}(E \mid G)\mathbb{P}(G)}{\mathbb{P}(E \mid G)\mathbb{P}(G) + \mathbb{P}(E \mid G^c)\mathbb{P}(G^c)} \\ &= \frac{1 \times 10^{-7}}{1 \times 10^{-7} + 10^{-6} \times (1 - 10^{-7})} = \frac{1}{1 + 10 \times (1 - 10^{-7})} \\ &\approx \frac{1}{11}.\end{aligned}$$

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

Section 2.5: Independence of events

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

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

Definition 2.16.

1. Two events A and B are independent if and only if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.
2. Events $A_1, \ldots, A_n$ are independent if and only if for any subset $S \subseteq \{1, \ldots, n\}$

$$\mathbb{P}\left(\bigcap_{i \in S} A_i\right) = \prod_{i \in S} \mathbb{P}(A_i)$$

Lemma 2.17.

If events A and B are independent, so are events A and B^c.

Example

Example 2.18.

- Throw a fair dice repeatedly, with the throws independent.
- What is $\mathbb{P}$(first six occurs on 4th throw)?
- Let A_i be the event that a 6 is thrown on the ith throw of the dice.
- Event of interest is

$$\begin{aligned}&\{\text{ first six occurs on 4th throw}\}\\&\quad= \{\text{ 1st throw not 6 AND 2nd throw not 6}\\&\qquad\quad\text{AND 3rd throw not 6 AND 4th throw is 6}\}\\&\quad= A_1^c \cap A_2^c \cap A_3^c \cap A_4.\end{aligned}$$

- By independence,

$$\mathbb{P}(A_1^c \cap A_2^c \cap A_3^c \cap A_4) = \mathbb{P}(A_1^c)\mathbb{P}(A_2^c)\mathbb{P}(A_3^c)\mathbb{P}(A_4) = \frac{5}{6}\cdot\frac{5}{6}\cdot\frac{5}{6}\cdot\frac{1}{6} = \frac{5^3}{6^4}.$$

Chain rule

Lemma 2.19.

Chain rule / Multiplication rule

1. *For any two events A and B with $\mathbb{P}(B) > 0$,*

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

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

$$\begin{aligned}&\mathbb{P}(A_1 \cap \cdots \cap A_n)\\&= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2)\cdots\mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).\end{aligned} \tag{2.2}$$

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 (2.2) as

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

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

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

- Hence as required the RHS of (2.2) is equal to

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

□

Example: bridge hand (revisited – see Example 1.28)

Example 2.20.

- 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 = "dealt J spades"
  - A_2 = "dealt Q spades"
  - A_3 = "dealt K spades"
  - A_4 = "dealt A spades"
- Note $\mathbb{P}(A_1) = \mathbb{P}(A_2) = \mathbb{P}(A_3) = \mathbb{P}(A_4) = \frac{13}{52} = \frac{1}{4}$, but these events are not independent.

Example 2.20.

- $$\begin{aligned}\mathbb{P}(A_2 \mid A_1) &= \frac{\mathbb{P}(A_1 \cap A_2)}{\mathbb{P}(A_1)} \\ &= \frac{\binom{50}{11}/\binom{52}{13}}{\binom{51}{12}/\binom{52}{13}} \quad \left(= \frac{\text{number of hands with J and Q}}{\text{number of hands with J}}\right) \\ &= \frac{12}{51} \quad \text{(or see this directly?)}\end{aligned}$$
- 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 $\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) = \frac{10}{49}$.
- Deduce (as before) that

$$\begin{aligned}&\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) \\ &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2)\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) \\ &= \frac{13}{52} \cdot \frac{12}{51} \cdot \frac{11}{50} \cdot \frac{10}{49}.\end{aligned}$$

Section 3: 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 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 3.1: Motivation and definitions

- A *trial* selects an *outcome* ω from a *sample space* Ω.
- Often we are interested in a number associated with the outcome, not the outcome itself.

Example 3.1.

- Throw two fair dice. Look at the total score.
- Let $X(\omega)$ be the total score when the outcome is ω.
- Remember we write the sample space as

$$\Omega = \{(a, b) \, : \, a, b = 1, \ldots, 6\}.$$

- So $X((a, b)) = a + b$.

Formal definition

Definition 3.2.

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

Remark 3.3.

- *For any set $B \subseteq \mathbb{R}$, we use the notation $\mathbb{P}(X \in B)$ as shorthand for*

$$\mathbb{P}(\{\omega \in \Omega \, : \, X(\omega) \in B\}).$$

- *E.g. X is the sum of the scores of two fair dice, $\mathbb{P}(X \leq 3)$ is shorthand for*

$$\mathbb{P}\Big(\{\omega \in \Omega \, : \, X(\omega) \leq 3\}\Big) = \mathbb{P}\Big(\{(1,1), (1,2), (2,1)\}\Big) = \frac{3}{36}.$$

Probability mass functions

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

Definition 3.4.

- 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

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}).$$

Remark 3.5.

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

In fact, any function with these properties can be thought of as a pmf of some random variable.

Example 3.6.

X is the sum of the scores on 2 fair dice

x	$=$	2	3	4	5	6	7	$\ldots$
$\lvert\{\omega : X(\omega) = x\}\rvert$	$=$	1	2	3	4	5	6	$\ldots$
$p_X(x)$	$=$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\ldots$

x	$=$	8	9	10	11	12
$\lvert\{\omega : X(\omega) = x\}\rvert$	$=$	5	4	3	2	1
$p_X(x)$	$=$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Section 3.2: Bernoulli distribution

- This is the building block for many distributions.

Definition 3.7.

- Think of an experiment with two outcomes: success or failure.

$$\Omega = \{\text{success}, \text{failure}\}$$

- 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 $\mathbb{P}(X = 1) = \mathbb{P}(\{\text{success}\}) = p$.
- Then

$$\mathbb{P}(X = 0) = \mathbb{P}(\{\text{failure}\}) = 1 - \mathbb{P}(\{\text{success}\}) = 1 - p.$$

- We say that X has a *Bernoulli distribution with parameter* p.

Bernoulli distribution notation

Remark 3.8.

- *Notation:* $X \sim \text{Bernoulli}(p)$
- *X has pmf*

$$\begin{aligned} p_X(0) &= 1 - p, \\ p_X(1) &= p, \\ p_X(x) &= 0 \text{ for } x \notin \{0, 1\}. \end{aligned}$$

- *Equivalently,* $p_X(x) = (1-p)^{1-x} p^x$ *for* $x = 0, 1$.

Example: Indicator functions

Example 3.9.

- Let A be an event, and let random variable I be defined by

$$I(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

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

$$\begin{array}{lll} \mathbb{P}(I = 1) & = \mathbb{P}(\{\omega : I(\omega) = 1\}) & = \mathbb{P}(A) \\ \mathbb{P}(I = 0) & = \mathbb{P}(\{\omega : I(\omega) = 0\}) & = \mathbb{P}(A^c) \end{array}$$

- 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 3.3: Binomial distribution

Definition 3.10.

- 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 3.11.

- Take $n = 3$ trials with $p = \frac{1}{3}$
- $\Omega = \{FFF, FFS, FSF, SFF, FSS, SFS, SSF, SSS\}$
- $$\begin{aligned}
\mathbb{P}(\{FFF\}) &= \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{2}{3} = \frac{8}{27} \\
\mathbb{P}(\{FFS\}) = \mathbb{P}(\{FSF\}) = \mathbb{P}(\{SFF\}) &= \frac{2}{3}\cdot\frac{2}{3}\cdot\frac{1}{3} = \frac{4}{27} \\
\mathbb{P}(\{FSS\}) = \mathbb{P}(\{SFS\}) = \mathbb{P}(\{SSF\}) &= \frac{2}{3}\cdot\frac{1}{3}\cdot\frac{1}{3} = \frac{2}{27} \\
\mathbb{P}(\{SSS\}) &= \frac{1}{3}\cdot\frac{1}{3}\cdot\frac{1}{3} = \frac{1}{27}
\end{aligned}$$

Binomial distribution example (cont.)

Example 3.11.

- Hence
  - $\{T = 0\} = \{FFF\}$ so that $\mathbb{P}(T = 0) = \frac{8}{27}$
  - $\{T = 1\} = \{FFS, FSF, SFF\}$ so that $\mathbb{P}(T = 1) = 3 \times \frac{4}{27} = \frac{12}{27}$
  - $\{T = 2\} = \{FSS, SFS, SSF\}$ so that $\mathbb{P}(T = 2) = 3 \times \frac{2}{27} = \frac{6}{27}$
  - $\{T = 3\} = \{SSS\}$ so that $\mathbb{P}(T = 3) = \frac{1}{27}$
- 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 3.12.

In general if $T \sim Bin(n, p)$ *then*

$$p_T(x) = \mathbb{P}(T = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

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 $\sum_{x=0}^{n} p_T(x) = 1$ in this case (Hint: use Proposition 1.26.3).

Binomial distribution example

Example 3.13.

- 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 $T \sim \text{Bin}(10, 0.4)$.

$$\begin{aligned}
\mathbb{P}(T \leq 2) &= p_T(0) + p_T(1) + p_T(2) \\
&= \binom{10}{0}(0.4)^0(0.6)^{10} + \binom{10}{1}(0.4)^1(0.6)^9 \\
&\quad + \binom{10}{2}(0.4)^2(0.6)^8 \\
&= 0.167
\end{aligned}$$

Section 3.4: Geometric distribution

Definition 3.14.

- 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

$$\mathbb{P}(X = x) = (1-p)^{x-1} p, \qquad x = 1, 2, 3, \ldots$$

- Hence the mass function is

$$p_X(x) = \mathbb{P}(X = x) = p(1-p)^{x-1}, \qquad x = 1, 2, 3, \ldots$$

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

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

Example: call-centre

Example 3.15.

- Consider a call-centre with 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 $X \sim \text{Geom}(\frac{1}{10})$.

x	=	1	2	3	4	5	6	$\cdots$
$\mathbb{P}(X = x)$	=	0.1	0.09	0.081	0.0729	0.06561	0.05905	$\cdots$

Geometric tail distribution

Lemma 3.16.

If $X \sim Geom(p)$ then $\mathbb{P}(X > x) = (1-p)^x$ for any integer $x \geq 0$.

Proof.

Write $q = 1 - p$. Then by summing a geometric progression to infinity:

$$\begin{aligned}
\mathbb{P}(X > x) &= \mathbb{P}(X = x+1) + \mathbb{P}(X = x+2) + \mathbb{P}(X = x+3) + \cdots \\
&= pq^x + pq^{x+1} + pq^{x+2} + \cdots \\
&= pq^x(1 + q + q^2 + \cdots) \\
&= pq^x \frac{1}{1-q} \\
&= q^x,
\end{aligned}$$

since $p/(1-q) = 1$. □

Waiting time formulation

Remark 3.17.

Lemma 3.16 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 3.15),

$$\mathbb{P}(X > 10) = 0.9^{10} = 0.349 \quad \text{(to 3 s.f.)}.$$

Lack-of-memory property

Lemma 3.18.

Lack of memory property *If $X \sim Geom(p)$ then for any $x \geq 1$:*

$$\mathbb{P}(X = x + n \,|\, X > n) = \mathbb{P}(X = x).$$

Remark 3.19.

- *In the call-centre example (Example 3.15) this tells us for example that*
$$\mathbb{P}(X = 5 + x \,|\, X > 5) = \mathbb{P}(X = x).$$
- *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 3.5: Poisson distribution

Definition 3.20.

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

- **Exercise:** verify that $\sum_{x=0}^{\infty} p_X(x) = 1$.
- Hint: see later in Analysis that
$$\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}.$$

Two motivations

Remark 3.21.

If $X \sim Bin(n, p)$ 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 $Poi(\lambda)$ random variable where $\lambda = np$.

Remark 3.22.

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 3.23.

- 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 3.23.

- $\mathbb{P}(X = x) \approx e^{-4.03} \frac{4.03^x}{x!}$
- For example

$$\begin{array}{cccccccc} x & = & 0 & 1 & 2 & 3 & 4 & \cdots \\ \mathbb{P}(X = x) & \approx & 0.0178 & 0.0716 & 0.144 & 0.1939 & 0.1953 & \cdots \end{array}$$

- We can deduce that

$$\begin{aligned} & \mathbb{P}(\text{at least one passenger bumped}) \\ = \; & \mathbb{P}(X \leq 2) = p_X(0) + p_X(1) + p_X(2) \\ \approx \; & 0.2334. \end{aligned}$$

Section 4: 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 4.1: Expectation

- We want some concept of the average value of a r.v. X and the spread about this average.
- Key insight is that the average should weight the outcomes by probability.

Definition 4.1.

- 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

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

- This is well-defined so long as $\sum_{x \in S} |x| p_X(x)$ converges.
- $\mathbb{E}(X)$ is also sometimes called the *mean* of the distribution of X.

Example: Bernoulli random variable

Example 4.2.

- Recall from Remark 3.8 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 4.1

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

- Note that for $p \neq 0, 1$ this random variable X won't ever equal $\mathbb{E}(X)$.

Motivation

Remark 4.3.

- *Do not confuse $\mathbb{E}(X)$ with 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, ...*
- *Consider the limit of the sample means*

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

Motivation (cont.)

Remark 4.4.

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

$$X_1 + X_2 + \cdots + X_n = \sum_{x\in S} x a_n(x).$$

- *We expect (but have not yet proved) that*

$$\frac{a_n(x)}{n} \to p_X(x) \quad \text{as} \quad n \to \infty.$$

- *If so then*

$$\frac{X_1 + \cdots + X_n}{n} = \frac{\sum_{x\in S} x a_n(x)}{n} = \sum_{x\in S} x \frac{a_n(x)}{n} \to \sum_{x\in S} x p_X(x).$$

- *This motivates Definition 4.1.*

Section 4.2: Examples

Example 4.5 (Uniform random variable).

- 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 4.6.

- $X \sim \text{Bin}(n, p)$ (see Definition 3.10).
- $\mathbb{P}(X = x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$
- $$\begin{aligned} \mathbb{E}(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\ &= np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} \\ &= np. \end{aligned}$$

- Here we use the fact that $x\binom{n}{x} = n\binom{n-1}{x-1}$ (check directly?) and apply the Binomial Theorem 1.26.3.
- There are easier ways — see later.

Example: Poisson distribution

Example 4.7.

- $X \sim \text{Poi}(\lambda)$ (see Definition 3.20).
- $\mathbb{P}(X = x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$
- $$\begin{aligned} \mathbb{E}(X) &= \sum_{x=0}^{\infty} x e^{-\lambda}\frac{\lambda^x}{x!} \\ &= \sum_{x=1}^{\infty} x e^{-\lambda}\frac{\lambda^x}{x!} = \lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda}\sum_{y=0}^{\infty}\frac{\lambda^y}{y!} = \lambda e^{-\lambda}e^{\lambda}. \end{aligned}$$
- So $\mathbb{E}(X) = \lambda$.

Example: geometric distribution

Example 4.8.

- $X \sim \text{Geom}(p)$ (see Definition 3.14).
- Recall that $\mathbb{P}(X = x) = (1-p)^{x-1}p$, so that

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x=1}^{\infty}(1-p)^{x-1}px \\ &= p\sum_{x=1}^{\infty}(1-p)^{x-1}x \\ &= p\frac{1}{(1-(1-p))^2} = \frac{1}{p}. \end{aligned}$$

- 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 4.3: Expectation of a function of a r.v.

- Consider a random variable X taking values $x_1, x_2, \ldots$
- Take a function $g : \mathbb{R} \to \mathbb{R}$ and define a new r.v. $Z(\omega) = g(X(\omega))$.
- Then Z takes values in the range $z_1 = g(x_1)$, $z_2 = g(x_2)$. . . .
- 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 4.9.

Let $Z = g(X)$. Then

$$\mathbb{E}(Z) = \mathbb{E}g(X) = \sum_i g(x_i)p_X(x_i) = \sum_{x \in S} g(x)p_X(x).$$

Proof.

(you are not required to know this 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

$$\begin{aligned}\mathbb{E}(Z) &= \sum_i z_i p_Z(z_i) = \sum_i z_i \left(\sum_{j : g(x_j) = z_i} p_X(x_j) \right) \\ &= \sum_i \left(\sum_{j : g(x_j) = z_i} g(x_j) p_X(x_j) \right) = \sum_j g(x_j) p_X(x_j).\end{aligned}$$

□

Example 4.10.

- Returning to Example 4.5:

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

- 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 4.9

$$\begin{aligned} \mathbb{E}(Z) &= \sum_{x=1}^{n} g(x) p_X(x) \\ &= \sum_{x=1}^{n} x^2 \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x^2 \\ &= \frac{1}{n}\frac{1}{6} n(n+1)(2n+1) = \frac{1}{6}(n+1)(2n+1) \end{aligned}$$

Linearity of expectation

Lemma 4.11.

Let a and b be constants. Then $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.

Proof.

Let $g(x) = ax + b$. From Theorem 4.9 we know that

$$\begin{aligned} \mathbb{E}(g(X)) &= \sum_i g(x_i) p_X(x_i) = \sum_i (a x_i + b) p_X(x_i) \\ &= a \sum_i x_i p_X(x_i) + b \sum_i p_X(x_i) = a\mathbb{E}(X) + b. \end{aligned}$$

□

Section 4.4: Variance

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

Definition 4.12.

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

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

- Notation: $\mathrm{Var}\,(X)$ is often denoted σ^2.
- The *standard deviation of* X is $\sqrt{\mathrm{Var}\,(X)}$.

Example of spread

Example 4.13.

- Define random variables each with mean zero $\mathbb{E}Y = \mathbb{E}Z = \mathbb{E}U = 0$

$$Y = \begin{cases} 1, \text{ wp. } \frac{1}{2}, \\ -1, \text{ wp. } \frac{1}{2}, \end{cases} \quad U = \begin{cases} 10, \text{ wp. } \frac{1}{2}, \\ -10, \text{ wp. } \frac{1}{2}, \end{cases} \quad Z = \begin{cases} 2, \text{ wp. } \frac{1}{5}, \\ -\frac{1}{2}, \text{ wp. } \frac{4}{5}. \end{cases}$$

- Notice the expectation does not distinguish between these rv.'s.
- Yet they are clearly different, and the variance helps capture this.

$$\mathrm{Var}\,(Y) = \mathbb{E}(Y-0)^2 = 1^2 \cdot \frac{1}{2} + (-1)^2 \cdot \frac{1}{2} = 1,$$
$$\mathrm{Var}\,(U) = \mathbb{E}(U-0)^2 = 10^2 \cdot \frac{1}{2} + (-10)^2 \cdot \frac{1}{2} = 100,$$
$$\mathrm{Var}\,(Z) = \mathbb{E}(Z-0)^2 = 2^2 \cdot \frac{1}{5} + \left(-\frac{1}{2}\right)^2 \cdot \frac{4}{5} = 1.$$

Useful lemma

Lemma 4.14.

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

Sketch proof: see Theorem 6.15.

$$\begin{aligned}\mathrm{Var}\,(X) &= \mathbb{E}((X-\mu)^2)\\ &= \mathbb{E}(X^2 - 2\mu X + \mu^2)\\ &= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 \quad \text{(will prove this step later)}\\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2\\ &= \mathbb{E}(X^2) - \mu^2\end{aligned}$$

□

Example: Bernoulli random variable

Example 4.15.

- Recall from Remark 3.8 and Example 4.2 that if $X \sim \text{Bernoulli}(p)$ then $p_X(0) = 1-p$, $p_X(1) = p$ and $\mathbb{E}X = p$.
- We can calculate $\mathrm{Var}\,(X)$ in two different ways:
  1. $\mathrm{Var}\,(X) = \mathbb{E}(X-\mu)^2 = \sum_x p_X(x)(x-p)^2 = (1-p)(-p)^2 + p(1-p)^2 = (1-p)p(p+1-p) = p(1-p)$.
  2. Alternatively:

$$\mathbb{E}(X^2) = \sum_x p_X(x)x^2 = (1-p)0^2 + p1^2 = p,$$

so that $\mathrm{Var}\,(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = p - p^2 = p(1-p)$.

Remark 4.16.

We will see in Example 6.17 below that if $X \sim Bin(n,p)$ (see Definition 3.10) then $\mathrm{Var}\,(X) = np(1-p)$. (Need to know this formula)

Uniform Example

Example 4.17.

- Again consider the uniform random variable (from Example 4.5)

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

- Know from Example 4.5 that $\mathbb{E}(X) = \frac{n+1}{2}$ and from Example 4.10
- that $\mathbb{E}(X^2) = \frac{1}{6}(n+1)(2n+1)$.

$$\begin{aligned} \mathrm{Var}\,(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \\ &= \frac{1}{6}(n+1)(2n+1) - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{n+1}{12}\left(4n+2-3(n+1)\right) \\ &= \frac{(n+1)}{12}(n-1) = \frac{(n^2-1)}{12}. \end{aligned}$$

Example: Poisson random variable

Example 4.18.

- Consider $X \sim \mathrm{Poi}(\lambda)$ (see Definition 3.20).
- Recall that $\mathbb{P}(X = x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & x = 0, 1, \ldots \\ 0 & \text{otherwise} \end{cases}$ and $\mathbb{E}(X) = \lambda$.
- We show (see next page) that $\mathbb{E}(X^2) = \lambda^2 + \lambda$.
- Thus $\mathrm{Var}\,(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = (\lambda^2 + \lambda) - (\lambda)^2 = \lambda$.

Example: Poisson (cont.)

Example 4.18.

Key is that $x^2 = x(x-1) + x$, so again changing the range of summation:

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{x=0}^{\infty} x^2 e^{-\lambda} \frac{\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} x(x-1) e^{-\lambda} \frac{\lambda^x}{x!} + \sum_{x=0}^{\infty} x e^{-\lambda} \frac{\lambda^x}{x!} \\
&= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\
&= \lambda^2 e^{-\lambda} \left(\sum_{z=0}^{\infty} \frac{\lambda^z}{z!} \right) + \lambda e^{-\lambda} \left(\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \right)
\end{aligned}$$

which equals $\lambda^2 + \lambda$ since each bracketed term is precisely e^{λ} as before.

Non-linearity of variance

We now state (and prove later) an important result concerning variances, which is the counterpart of Lemma 4.11:

Lemma 4.19.

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

Section 4.5: Chebyshev's inequality

- Let X be any random variable with finite mean μ and variance σ^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 $\mathbb{E}(I) = 0 \cdot \mathbb{P}(I = 0) + 1 \cdot \mathbb{P}(I = 1) = \mathbb{P}(I = 1) = \mathbb{P}(|X - \mu| > c)$.
- Define also $Z(\omega) = (X(\omega) - \mu)^2/c^2$, so that

$$\mathbb{E}(Z) = \mathbb{E}\left(\frac{(X-\mu)^2}{c^2}\right) = \frac{\mathbb{E}((X-\mu)^2)}{c^2} = \frac{\sigma^2}{c^2}$$

- This last step uses Lemma 4.11 with $a = 1/c^2$ and $b = 0$.
- Notice that $I(\omega) \leq Z(\omega)$ for any ω. (plot a graph?)
- So $\mathbb{E}(I) \leq \mathbb{E}(Z)$, and we deduce that ...

Theorem 4.20 (Chebyshev's inequality).

For any random variable X with finite mean μ and variance σ^2, and any constant c:

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

Remark 4.21.

- *We only need to assume that X has finite mean and variance.*
- *Inequality says the probability that X is far from μ is bounded by a quantity that increases with the variance σ^2 and decreases with the distance from μ.*
- *In particular makes sense to take c a multiple of σ.*
- *This shows that our axioms and definitions give us something that fits with our intuition.*

Application of Chebyshev's inequality

Example 4.22.

- 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 $\mathbb{E}(T) = 5000$ and $\text{Var}\,(T) = 2500$ (see Example 4.6 and Remark 4.16).
- Thus by taking $c = 500$ in Chebyshev's inequality (Theorem 4.20) we have
$$\mathbb{P}(|T - 5000| > 500) \leq 0.01,$$
so that
$$\mathbb{P}(4500 \leq T \leq 5500) \geq 0.99.$$
- We can also express this as
$$\mathbb{P}\left(0.45 \leq \frac{T}{10^4} \leq 0.55\right) \geq 0.99.$$

Section 5: Joint distributions

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 5.1: The joint probability mass function

- Up to now we have only considered a single random variable at once, but now consider related random variables.
- 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.

Joint probability mass function

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

Definition 5.1.

The *joint pmf for* X *and* Y is $p_{X,Y}$, defined by

$$\begin{aligned} p_{X,Y}(x,y) &= \mathbb{P}(X = x, Y = y) \\ &= \mathbb{P}\big(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}\big) \end{aligned}$$

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

Example: coin tosses

Example 5.2.

- 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

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$
$x=0$	$1/8$	$1/8$	0	0
$x=1$	0	$1/4$	$1/4$	0
$x=2$	0	0	$1/8$	$1/8$

Section 5.2: Marginal pmfs

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

Definition 5.3.

- 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 $y_1, y_2, \ldots$.
- 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 5.4.

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: coin tosses (return to Example 5.2)

Example 5.5.

- 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

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$y=2$	$y=3$	
$x=0$	1/8	1/8	0	0	1/4
$x=1$	0	1/4	1/4	0	1/2
$x=2$	0	0	1/8	1/8	1/4
	1/8	3/8	3/8	1/8	

- 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 5.6.

- 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:

$p_{X,Y}(x,y)$	$y=0$	$y=1$
$x=0$	0	1/2
$x=1$	1/2	0

- Either write down the marginals directly, or calculate

$$p_X(0) = p_{X,Y}(0,0) + p_{X,Y}(0,1) = 1/2,$$

and $p_X(1) = 1 - p_X(0) = 1/2$ and similarly $p_Y(0) = p_Y(1) = 1/2$.

Marginal pmfs don't determine joint (cont.)

Example 5.7.

- 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:

$p_{X,Y}(x,y)$	$y=0$	$y=1$
$x=0$	1/4	1/4
$x=1$	1/4	1/4

- Summing rows and columns we see that $p_X(0) = p_X(1) = p_Y(0) = p_Y(1) = 1/2$, just as in Example 5.6.

Comparing Examples 5.6 and 5.7 we see that the marginal pmfs don't determine the joint pmf.

Section 5.3: Conditional pmfs

Definition 5.8.

- 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

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

Calculating conditional pmfs

Remark 5.9.

- *Notice that ('scale column by its sum')*

$$\begin{aligned} p_{X|Y}(x|y) &= \mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)} \\ &= \frac{p_{X,Y}(x,y)}{p_Y(y)}. \end{aligned} \tag{5.1}$$

- *Similarly ('scale row by its sum')*

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

Conditional pmfs are probability mass functions

Remark 5.10.

- *We can check (in the spirit of Remark 3.5) that for any fixed y, the* $p_{X|Y}(\cdot \,|\, y)$ *is a pmf.*
- *That is, for any* x*, since* (5.1) *expresses it as a ratio of probabilities, clearly* $p_{X|Y}(\cdot \,|\, y) \geq 0$.
- *Similarly using Theorem 5.4 we know that* $p_Y(y) = \sum_x p_{X,Y}(x, y)$.
- *This means that (by* (5.1)*)*

$$\begin{aligned}\sum_x p_{X|Y}(x \,|\, y) &= \sum_x \frac{p_{X,Y}(x,y)}{p_Y(y)} \\ &= \frac{1}{p_Y(y)} \sum_x p_{X,Y}(x,y) = \frac{1}{p_Y(y)} p_Y(y) = 1,\end{aligned}$$

as required.

Example 5.2 continued

Example 5.11.

- 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	1/2	1/2

- 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
$p_{X\|Y}(x\|1)$	1/3	2/3	0

Example: Inviting friends to the pub

Example 5.12.

- 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 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.
- $X \sim \mathrm{Bin}(3, \frac{1}{2})$ 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.
- $Y|X = x \sim \mathrm{Bin}(x, \frac{2}{3})$ 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 5.12.

So $p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) = \begin{cases} \frac{1}{8}\binom{3}{x}\binom{x}{y}\frac{2^y}{3^x} & 0 \leq y \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$

	$y = 0$	1	2	3
$x = 0$	$\frac{1}{8} \times 1 = \frac{1}{8}$	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	$\frac{1}{8} \times \frac{1}{27} = \frac{1}{216}$	$\frac{1}{8} \times \frac{6}{27} = \frac{1}{36}$	$\frac{1}{8} \times \frac{12}{27} = \frac{1}{18}$	$\frac{1}{8} \times \frac{8}{27} = \frac{1}{27}$
	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$

Therefore $\mathbb{E}(Y) = 0 \times \frac{8}{27} + 1 \times \frac{12}{27} + 2 \times \frac{6}{27} + 3 \times \frac{1}{27} = \frac{12+12+3}{27} = 1.$

There is a much easier way to calculate $\mathbb{E}(Y)$ - see Section 9.

Section 5.4: Independent random variables

Definition 5.13.

- Two random variables are *independent* if

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \quad \text{for all } x \text{ and } y.$$

- Equivalently if

$$p_{X|Y}(x|y) = p_X(x), \quad \text{for all } x \text{ and } y.$$

- 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 5.14.

1. *Consistent with Definition 2.16 (independence of events).*
2. *We require that the events $\{X = x\}$ and $\{Y = y\}$ are independent for any x and y.*
3. *In fact this is equivalent to requiring events $\{X \in A\}$ and $\{Y \in B\}$ independent for any A and B.*
4. **Important**: *if X and Y are independent, so are $g(X)$ and $h(Y)$ for any functions g and h.* [a]

---

[a] Proof (not examinable): For any u, v

$$\begin{aligned}
\mathbb{P}(g(X) = u, h(Y) = v) &= \mathbb{P}\left(\{X \in g^{-1}(u)\} \bigcap \{Y \in h^{-1}(v)\}\right) \\
&= \mathbb{P}\left(\{X \in g^{-1}(u)\}\right) \mathbb{P}\left(\{Y \in h^{-1}(v)\}\right) \\
&= \mathbb{P}(g(X) = u)\mathbb{P}(h(Y) = v).
\end{aligned}$$

IID random variables

Definition 5.15.

- 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

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_X(x_i),$$

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 5.16.

- 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:

$$p_{X,Y}(x,y) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = p_X(x) \cdot p_Y(y) \qquad \text{for all } x, y = 1, \ldots, 6.$$

- We see that these variables are independent (in fact IID as well).

Discrete convolution

Proposition 5.17.

- *Let X and Y be independent, integer-valued random variables with respective mass functions p_X and p_Y.*
- *Then random variable $X+Y$ is also integer-valued and has mass function satisfying*

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

- *This formula is called the discrete convolution of the mass functions p_X and p_Y.*

Discrete convolution proof

Proof.

Using independence, and since it is a disjoint union, we know that

$$\begin{aligned}
p_{X+Y}(k) &= \mathbb{P}(X+Y=k) = \mathbb{P}\left(\bigcup_{i=-\infty}^{\infty} \{X+Y=k, Y=i\} \right) \\
&= \sum_{i=-\infty}^{\infty} \mathbb{P}(X+Y=k,\ Y=i) \\
&= \sum_{i=-\infty}^{\infty} \mathbb{P}(X=k-i,\ Y=i) = \sum_{i=-\infty}^{\infty} \mathbb{P}(X=k-i)\mathbb{P}(Y=i) \\
&= \sum_{i=-\infty}^{\infty} p_X(k-i) \cdot p_Y(i).
\end{aligned}$$

□

Convolution of Poissons gives a Poisson

Theorem 5.18.

- *Recall the definition of the Poisson distribution from Definition 3.20.*
- *Let $X \sim Poi(\lambda)$ and $Y \sim Poi(\mu)$ be independent.*
- *Then $X + Y \sim Poi(\lambda + \mu)$.*

Proof of Theorem 5.18

Proof.

Using Proposition 5.17, since X and Y only take positive values we know

$$
\begin{aligned}
p_{X+Y}(k) &= \sum_{i=0}^{k} p_X(k-i) p_Y(i) \\
&= \sum_{i=0}^{k} \left(e^{-\lambda} \frac{\lambda^{k-i}}{(k-i)!} \right) \left(e^{-\mu} \frac{\mu^i}{i!} \right) \\
&= e^{-(\lambda+\mu)} \frac{1}{k!} \sum_{i=0}^{k} \binom{k}{i} \lambda^{k-i} \mu^i \\
&= e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!},
\end{aligned}
$$

where we use the Binomial Theorem, Proposition 1.26.3. □

Section 6: 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 6.1: Properties of expectation $\mathbb{E}$

Theorem 6.1.

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 4.11.

□

Proof of Theorem 6.1 (cont).

Proof.

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

$$\begin{aligned}
\mathbb{E}(Z) &= \sum_{x_i}\sum_{y_j} g(x_i, y_j) p_{X,Y}(x_i, y_j) \text{ by extension of Theorem 4.9} \\
&= \sum_{x_i}\sum_{y_j} (x_i + y_j) p_{X,Y}(x_i, y_j) \\
&= \sum_{x_i}\sum_{y_j} \{x_i p_{X,Y}(x_i, y_j) + y_j p_{X,Y}(x_i, y_j)\} \\
&= \sum_{x_i} x_i \left\{\sum_{y_j} p_{X,Y}(x_i, y_j)\right\} + \sum_{y_j} y_j \left\{\sum_{x_i} p_{X,Y}(x_i, y_j)\right\} \\
&= \sum_{x_i} x_i p_X(x_i) + \sum_{y_j} y_j p_Y(y_j) \\
&= \mathbb{E}(X) + \mathbb{E}(Y).
\end{aligned}$$

□

Additivity of expectation

Corollary 6.2.

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 6.1.3 and induction on n. □

Combining this with Theorem 6.1.2, we can also show more generally that:

Theorem 6.3.

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

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

Example: Bernoulli trials

Example 6.4.

- 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 indicator
$X_i = \begin{cases} 0 & \text{if } i\text{th trial a failure} \\ 1 & \text{if } i\text{th trial a success.} \end{cases}$
- For each i, $\mathbb{E}(X_i) = (1-p) \cdot 0 + p \cdot 1 = p$
- So $\mathbb{E}(T) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$ by Corollary 6.2.
- So $\mathbb{E}(T) = np$.

- This is simpler (and more general) than Example 4.6.
- 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 6.5.

- 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_r be the number of packs you need to buy to get from $r-1$ distinct monsters to r 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

$$p_r = \frac{\text{number of monsters we don't have}}{\text{number of different monsters}} = \frac{k-(r-1)}{k}$$

Example: BitTorrent problem (cont.)

Example 6.5.

- Therefore (see Example 4.8) $\mathbb{E}(X_r) = \frac{1}{p_r} = \frac{k}{k-r+1}$.
- Hence

$$\begin{aligned} \mathbb{E}(N) &= \sum_{r=1}^{k} \mathbb{E}(X_r) = \sum_{r=1}^{k} \frac{k}{k-r+1} \\ &= k(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{k}) \approx k \ln k. \end{aligned}$$

- To illustrate this result we have:

k	$\mathbb{E}(N)$
5	11.4
10	29.3
20	80.0

Section 6.2: Covariance

- We've seen that $\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
- But when does $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$?
- We will see in Lemma 6.10 that it holds 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 6.6.

- 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 6.7.

The *covariance* of X and Y is defined as

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

Covariance measures how the two random variables vary together.

Remark 6.8.

- *For any random variable X we have $\mathrm{Cov}\,(X,X) = \mathrm{Var}\,(X)$.*
- *Further (the proofs are an exercise):*

$$\begin{aligned}\mathrm{Cov}\,(aX,Y) &= a\mathrm{Cov}\,(X,Y)\\ \mathrm{Cov}\,(X,bY) &= b\mathrm{Cov}\,(X,Y)\\ \mathrm{Cov}\,(X,Y+Z) &= \mathrm{Cov}\,(X,Y) + \mathrm{Cov}\,(X,Z)\end{aligned}$$

Alternative expression for covariance

Lemma 6.9.

For any random variables X and Y $\mathrm{Cov}\,(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Proof.

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

$$\begin{aligned}\mathrm{Cov}\,(X,Y) &= \mathbb{E}[(X-\mu)(Y-\nu)]\\ &= \mathbb{E}[XY - \nu X - \mu Y + \mu\nu]\\ &= \mathbb{E}(XY) - \nu\mathbb{E}(X) - \mu\mathbb{E}(Y) + \mu\nu\\ &= \mathbb{E}(XY) - \mathbb{E}(Y)\mathbb{E}(X) - \mathbb{E}(X)\mathbb{E}(Y) + \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

□

Useful lemma

Lemma 6.10.

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

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

Proof:

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_i \sum_j x_i y_j p_{X,Y}(x_i, y_j) \\
&= \sum_i \sum_j x_i y_j p_X(x_i) p_Y(y_j) \quad \text{by independence} \\
&= \sum_i x_i p_X(x_i) \sum_j y_j p_Y(y_j) \\
&= \sum_i x_i p_X(x_i) \mathbb{E}(Y) \\
&= \mathbb{E}(X)\mathbb{E}(Y)
\end{aligned}$$

Delicate issue

We can rephrase Lemmas 6.9 and 6.10 to deduce that

Lemma 6.11.

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

Example 6.12.

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.

Corollary of Lemma 6.11

Corollary 6.13.

If X and Y are independent then (by Remark 5.14)

$$\mathbb{E}\left(g(X)h(Y)\right) = \left(\mathbb{E}g(X)\right) \cdot \left(\mathbb{E}h(Y)\right),$$

for any functions g and h.

Correlation coefficient

- If X and Y tend to increase (and decrease) together $\mathrm{Cov}\,(X, Y) > 0$ (e.g. age and salary).
- If one tends to increase as the other decreases then $\mathrm{Cov}\,(X, Y) < 0$ (e.g. hours of training, marathon times).
- If X and Y are independent then $\mathrm{Cov}\,(X, Y) = 0$

Definition 6.14.

The correlation coefficient of X and Y is

$$\rho(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sqrt{\mathrm{Var}\,(X)\mathrm{Var}\,(Y)}}.$$

- Note that it can be shown that $-1 \leq \rho(X, Y) \leq 1$.
- This is essentially the Cauchy–Schwarz inequality from linear algebra.
- ρ is a measure of how dependent the random variables are, and doesn't depend on the scale of either r.v.

Section 6.3: Properties of variance

Theorem 6.15.

1. $\mathrm{Var}\,(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$
2. *Let a and b be constants. Then* $\mathrm{Var}\,(aX + b) = a^2 \mathrm{Var}\,(X)$.
3. *For any random variables* X *and* Y*,*

$$\mathrm{Var}\,(X + Y) = \mathrm{Var}\,(X) + 2\mathrm{Cov}\,(X, Y) + \mathrm{Var}\,(Y).$$

4. *If* X *and* Y *are independent r.v.s then*

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

Important: Note that if X and Y are not independent, then 4. is not usually true.

Proof of Theorem 6.15

Proof.

1. Seen before as Lemma 4.14 — 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 6.1.2.

2. Set $Z = aX + b$. We know $\mathbb{E}(Z) = a\mathbb{E}(X) + b$, so

$$\begin{aligned}(Z - \mathbb{E}(Z))^2 &= ((aX + b) - (a\mathbb{E}(X) + b))^2 \\ &= (a(X - \mathbb{E}(X)))^2 = a^2(X - \mathbb{E}(X))^2.\end{aligned}$$

Thus

$$\mathrm{Var}\,(Z) = \mathbb{E}\left((Z - \mathbb{E}(Z))^2\right) = a^2\mathbb{E}((X - \mathbb{E}(X))^2) = a^2\mathrm{Var}\,(X).$$

□

Proof of Theorem 6.15 (cont).

Proof.

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

$$(\mathbb{E}(T))^2 = (\mathbb{E}(X))^2 + 2\mathbb{E}(X)\mathbb{E}(Y) + (\mathbb{E}(Y))^2. \tag{6.1}$$

   - Need to calculate

$$\mathbb{E}(T^2) = \mathbb{E}(X^2 + 2XY + Y^2) = \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2). \tag{6.2}$$

   - Hence subtracting (6.1) from (6.2) and rearranging, we obtain:

$$\begin{aligned} \mathrm{Var}\,(T) &= \mathbb{E}(T^2) - (\mathbb{E}(T))^2 \\ &= (\mathbb{E}(X^2) - (\mathbb{E}(X))^2) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) \\ &\quad +(\mathbb{E}(Y^2) - (\mathbb{E}(Y))^2) \\ &= \mathrm{Var}\,(X) + 2(\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)) + \mathrm{Var}\,(Y). \\ &= \mathrm{Var}\,(X) + 2\mathrm{Cov}\,(X, Y) + \mathrm{Var}\,(Y) \end{aligned} \tag{6.3}$$

4. Part 4. follows using Lemma 6.11. □

General variance formula for independent X_i

Corollary 6.16.

Let $X_1, X_2, \ldots$ be independent. Then

$$\mathrm{Var}\,(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}\,(X_1) + \mathrm{Var}\,(X_2) + \cdots + \mathrm{Var}\,(X_n).$$

Proof.

Induction on n, using Theorem 6.15.4. □

Important: If X_i are not independent then the situation is more complicated, will have covariance terms as well.

Section 6.4: Examples and Law of Large Numbers

Example 6.17.

- Recall from Example 6.4 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 4.15 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 $\mathrm{Var}\,(X_i) = p - p^2 = p(1-p)$.
- Hence by independence and Corollary 6.16

$$\begin{aligned} \mathrm{Var}\,(T) &= \mathrm{Var}\,(X_1 + \cdots + X_n) \\ &= \mathrm{Var}\,(X_1) + \cdots + \mathrm{Var}\,(X_n) = np(1-p). \end{aligned}$$

Note: much easier than trying to sum this directly!

Application: Sample means

Theorem 6.18.

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

$$\begin{aligned} \mathbb{E}(\bar{X}) &= \mu \\ \mathrm{Var}\,(\bar{X}) &= \sigma^2/n \end{aligned}$$

Proof.

- Then (see also Theorem 6.3)

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \frac{1}{n}\mathbb{E}(X_1 + \cdots + X_n) \\ &= \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)) = \frac{1}{n}(\mu + \cdots + \mu) = \mu.\end{aligned}$$

-

$$\begin{aligned}\mathrm{Var}\,(\bar{X}) &= \mathrm{Var}\,\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \\ &= \left(\frac{1}{n}\right)^2 \mathrm{Var}\,(X_1 + \cdots + X_n) \quad \text{by Theorem 6.15} \\ &= \frac{1}{n^2}(\mathrm{Var}\,(X_1) + \cdots + \mathrm{Var}\,(X_n)) \quad \text{by Corollary 6.16} \\ &= \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}\end{aligned}$$

□

Example: coin toss (slight return)

Example 6.19.

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

$$\mathrm{Var}\,(\bar{X}) = \frac{1}{4n}.$$

The weak law of large numbers

- Let Y be any r.v. and let $c > 0$ be a positive constant.
- Recall Chebyshev's inequality (Theorem 4.20):

$$\mathbb{P}(|Y - \mathbb{E}(Y)| > c) \leq \frac{\mathrm{Var}\,(Y)}{c^2}.$$

- We know that $\mathbb{E}(\bar{X}) = \mu$ and $\mathrm{Var}\,(\bar{X}) = \frac{\sigma^2}{n}$.
- So taking $Y = \bar{X}$ in Chebyshev we deduce:

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

Theorem 6.20 (Weak law of large numbers).

Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed (IID) random variables with common mean μ and variance σ^2. Then for any $c > 0$:

$$\mathbb{P}(|\bar{X} - \mu| > c) \to 0 \text{ as } n \to \infty.$$

Application to coin tossing

Example 6.21.

- As in Example 6.19, 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 taking $c = 0.01$:

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

- 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).

Section 6.5: Examples

Example 6.22.

- 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 $\mathrm{Cov}\,(X, Y)$, $\mathbb{E}(W)$, $\mathrm{Var}\,(W)$.

Urn example (cont.)

Example 6.22.

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

$p_{X,Y}(x,y)$	$y=0$	$y=1$	$p_X(x)$
$x=0$	$5/18$	$4/18$	$1/2$
$x=1$	$4/18$	$5/18$	$1/2$
$p_Y(y)$	$1/2$	$1/2$	

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

Urn example (cont.)

Example 6.22.

- $$\mathbb{E}(XY) = 0 \times 0 \times p_{X,Y}(0,0) + 0 \times 1 \times p_{X,Y}(0,1) + 1 \times 0 \times p_{X,Y}(1,0) + 1 \times 1 \times p_{X,Y}(1,1) = \frac{5}{18}.$$
- Thus $\mathrm{Cov}\,(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{5}{18} - (\frac{1}{2})^2 = \frac{1}{36}$.
- Further, since $\mathrm{Var}\,(X) = \frac{1}{4}$, $\mathrm{Var}\,(Y) = \frac{1}{4}$, we know $\rho(X,Y) = \frac{1}{9}$.
- $$\mathrm{Var}\,(W) = \mathrm{Var}\,(X) + 2\mathrm{Cov}\,(X,Y) + \mathrm{Var}\,(Y) = \frac{1}{4} + \frac{2}{36} + \frac{1}{4} = \frac{5}{9}.$$
- Compare with $\mathrm{Bin}(2, \frac{1}{2})$ when variance $= \frac{1}{2}$.

Further example

Example 6.23.

- 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

$$\mathrm{Var}\,(X) = \frac{5}{4}$$

$$\mathrm{Var}\,(Y) = \frac{5}{2}.$$

Further example (cont.)

Example 6.23.

- 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 $\mathrm{Cov}\,(X, Z) = 0$. Thus

$$\begin{aligned}\mathrm{Cov}\,(X, Y) &= \mathrm{Cov}\,(X, X + Z) = \mathrm{Cov}\,(X, X) + \mathrm{Cov}\,(X, Z) \\ &= \mathrm{Var}\,(X) + 0 = \frac{5}{4}.\end{aligned}$$

- Thus

$$\rho(X, Y) = \frac{\mathrm{Cov}\,(X, Y)}{\sqrt{\mathrm{Var}\,(X)\mathrm{Var}\,(Y)}} = \frac{1}{\sqrt{2}}.$$

Section 7: Continuous random variables I

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 and gamma distributions.

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

Section 7.1: Motivation and definition

Remark 7.1.

- *So far we studied r.v.s that take a discrete (countable) set of values.*
- *Many r.v.s take a continuum of values e.g. height, weight, time, temperature are real-valued.*
- *Let X be time in seconds until an atom decays. Then $\mathbb{P}(X = \pi) = 0$.*
- *But we expect for δ small that*

$$\mathbb{P}(\pi \leq X \leq \pi + \delta) \approx const \times \delta$$

- *In general $\mathbb{P}(X = x) = 0$ for any particular x but expect for δ small:*

$$\mathbb{P}(x \leq X \leq x + \delta) \approx f_X(x)\delta$$

- *Think of $f_X(x)$ as an 'intensity' – won't generally be 0.*
- *But $f_X(x)$ will be ≥ 0 (because probabilities are).*

Remark 7.1.

- *Consider an interval $[a, b]$.*
- *Divide it up into n segments of equal size*

$$a = x_0 < x_1 < \cdots < x_n = b$$

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

- *Then*

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

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

Continuous random variables

Definition 7.2.

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 7.3.

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

- $\mathbb{P}(X = x) = 0$ *for all x, so*

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

- *Special case:*
  $\mathbb{P}(X \leq b) = \mathbb{P}(X < b) = \lim_{a \to -\infty} \mathbb{P}(a \leq X \leq b) = \int_{-\infty}^b f_X(x)\,dx.$
- *Since* $\mathbb{P}(-\infty < X < \infty) = 1$ *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 7.2: Mean and variance

Definition 7.4.

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 7.5.

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 ∞.
- However (see e.g. Example 7.7) we only need to consider the range where $f_X(x) > 0$.

Variance

Definition 7.6.

The *variance* of X is

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

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

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

Uniform distribution

Example 7.7.

- 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

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} \mathbb{I}(0 \leq x \leq 1)dx = \int_0^1 1dx = 1.$$

- 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 7.7.

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

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

so that $\mathrm{Var}\,(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1}{3} - \frac{1}{2^2} = \frac{1}{12}$.
- Similarly if Y is uniform on $[a, b]$:

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

Section 7.3: The distribution function

Definition 7.8.

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 7.9.

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

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

Distribution and density function

Lemma 7.10.

Let X have a continuous distribution. Then (y is a dummy variable)

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

- $\mathbb{P}(X \leq x)$ is the area under the density function to the 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 7.11.

In the uniform random variable setting of Example 7.7, 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

Example 7.12.

- 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

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\,dy = \int_{-\infty}^{x} 0\,dy = 0.$$

Example (cont.)

Example 7.12.

- Let $x > 1$. Then

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(y)\,dy = \int_{-\infty}^{1} 0\,dy + \int_{1}^{x} \frac{2}{y^3}\,dy = 0 + \left[\frac{-1}{y^2}\right]_1^x \\ &= \frac{-1}{x^2} - \frac{-1}{1} = 1 - \frac{1}{x^2}. \end{aligned}$$

- 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 $F_X(x) = \int f_X(y)\,dy$ without limits then determine C. It is both confusing and sloppy!

Continuous convolution

Recall the discrete convolution formula (Proposition 5.17)

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

In a very similar way we state without proof the continuous convolution formula for densities:

Proposition 7.13.

Suppose X and Y are independent continuous random variables with respective densities f_X and f_Y. Then their sum is a continuous random variable with density

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-y) \cdot f_Y(y)\, dy, \ \text{for all } z \in \mathbb{R}.$$

Section 7.4: 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 7.14.

- A r.v. T has an exponential distribution with rate parameter λ 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 $T \sim \mathrm{Exp}(\lambda)$.

Exponential distribution properties

Remark 7.15.

- $\mathbb{P}(T > t) = 1 - \mathbb{P}(T \leq t) = e^{-\lambda t}$
-
$$\begin{aligned}\mathbb{E}(T) &= \int_{-\infty}^{\infty} t f_T(t)\, dt = \int_0^{\infty} t\lambda e^{-\lambda t}\, dt \\ &= \left[-te^{-\lambda t}\right]_0^{\infty} + \int_0^{\infty} e^{-\lambda t}\, dt \\ &= 0 + \frac{1}{\lambda}\end{aligned}$$
-
$$\mathrm{Var}\,(T) = \frac{1}{\lambda^2} \quad \textit{(Exercise)}.$$

Exponential distribution properties (cont.)

Remark 7.15.

- *Exponential is continuous analogue of the geometric distribution.*
- *In particular it has the lack of memory property (cf Lemma 3.18):*

$$\begin{aligned}\mathbb{P}(T > t + s \mid T > s) &= \frac{\mathbb{P}(T > t + s \text{ and } T > s)}{\mathbb{P}(T > s)} \\ &= \frac{\mathbb{P}(T > t + s)}{\mathbb{P}(T > s)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= \mathbb{P}(T > t).\end{aligned}$$

Section 7.5: Gamma distributions

Definition 7.16.

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 7.17.

- *Note that for $\alpha > 1$:*

$$\begin{aligned} \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} e^{-x}\, dx \\ &= \left[-x^{\alpha-1} e^{-x}\right]_0^\infty + (\alpha-1)\int_0^\infty x^{\alpha-2} e^{-x}\, dx \\ &= 0 + (\alpha-1)\Gamma(\alpha-1) \end{aligned}$$

  for general α.
- *Also*

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

- *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 7.18.

- A random variable has a gamma distribution with shape parameter α and rate parameter λ 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 7.14.
- We find the normalization constant in Lemma 7.19 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 λ.

Lemma 7.19.

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

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

Proof.

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

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(x)\,dx \\ &= \int_0^\infty Cx^{\alpha-1}e^{-\lambda x}\,dx \\ &= C\int_0^\infty \left(\frac{y}{\lambda}\right)^{\alpha-1} e^{-y}\frac{dy}{\lambda} = \frac{C}{\lambda^\alpha}\Gamma(\alpha). \end{aligned}$$

□

Gamma distribution properties

Remark 7.20.

- *If* $\alpha = 1$ *then* $f_X(x) = \begin{cases} 0 & x \le 0 \\ \lambda e^{-\lambda x} & x > 0. \end{cases}$
  I.e. if $X \sim \text{Gamma}(1, \lambda)$ *then* $X \sim \text{Exp}(\lambda)$.
- *In Proposition 7.21 (see also Lemma 10.13) we will see that (for integer* α*) a* $\text{Gamma}(\alpha, \lambda)$ *r.v. has the same distribution as the sum of* α *independent* $\text{Exp}(\lambda)$ *r.v.s*

Proposition 7.21.

If $X \sim \text{Gamma}(\alpha, \lambda)$ *and* $Y \sim \text{Gamma}(\beta, \lambda)$ *are independent, then their sum* $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$.

Proof of Proposition 7.21 (integer α, β)

Proof.

By Proposition 7.13 we know that the density of $X + Y$ is the convolution

$$\begin{aligned}
f_{X+Y}(z) &= \int_{-\infty}^{\infty} f_X(z-y) \cdot f_Y(y) dy \\
&= \int_0^z f_X(z-y) \cdot f_Y(y) dy \\
&= \int_0^z \frac{\lambda^\alpha}{\Gamma(\alpha)} (z-y)^{\alpha-1} e^{-\lambda(z-y)} \cdot \frac{\lambda^\beta}{\Gamma(\beta)} y^{\beta-1} e^{-\lambda y} dy \\
&= \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda z} \int_0^z (z-y)^{\alpha-1} y^{\beta-1} dy \\
&=: \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} e^{-\lambda z} I_{\alpha,\beta}.
\end{aligned}$$

□

Proof of Proposition 7.21 (integer α, β, cont.)

Proof.

- This integral, known as a beta integral, equals $I_{\alpha,\beta} = z^{\alpha+\beta-1}\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$, as required.
- This follows in integer case by induction, since we can write it as $z^{\alpha+\beta-1}(\alpha-1)!(\beta-1)!/(\alpha+\beta-1)!$
- Value found using integration by parts (since function vanishes at either end of support):

$$\begin{aligned} I_{\alpha,\beta} &= \int_0^z (z-y)^{\alpha-1} y^{\beta-1} dy \\ &= \int_0^z (\alpha-1)(z-y)^{\alpha-2}\frac{y^\beta}{\beta} dy = \frac{\alpha-1}{\beta} I_{\alpha-1,\beta+1}. \end{aligned}$$

- We use the fact that $I_{1,\beta} = \int_0^z y^{\beta-1} = \frac{z^\beta}{\beta}$.

□

Gamma distribution properties (cont.)

Remark 7.21.

- $$\begin{aligned} \mathbb{E}(X) &= \int_0^\infty x \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^\alpha e^{-\lambda x}\, dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \left(\int_0^\infty \frac{\lambda^{\alpha+1}}{\Gamma(\alpha+1)} x^\alpha e^{-\lambda x}\, dx \right) \\ &= \frac{\alpha\Gamma(\alpha)}{\lambda\Gamma(\alpha)} \times 1 = \frac{\alpha}{\lambda} \end{aligned}$$

  since the bracketed term is the integral of a Gamma$(\alpha+1,\lambda)$ *density, which equals 1.*
- *Similarly* $\mathrm{Var}\,(X) = \dfrac{\alpha}{\lambda^2}$.

Section 8: Continuous random variables II

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

- 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 8.1: 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 4.9.

Example: scaling uniforms

Example 8.1.

- 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

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(a + (b-a)X \leq y) = \mathbb{P}\left(X \leq \frac{y-a}{b-a}\right) \\ &= \frac{y-a}{b-a} \quad \text{since } X \sim U(0,1). \end{aligned}$$

- 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 8.2.

Let X take 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 8.2 (cont.)

Proof.

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

$$\begin{aligned} F_Y(y) &= \mathbb{P}(g(X) \le y) = \mathbb{P}(X \ge h(y)) \\ &= 1 - \mathbb{P}(X < h(y)) = 1 - F_X(h(y)). \end{aligned}$$

- 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 8.3.

- 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

$$\begin{aligned} y &= \frac{1}{\lambda}\log\left(\frac{1}{1-x}\right) \\ \implies -\lambda y &= \log(1-x) \\ \implies x &= 1 - e^{-\lambda y} \end{aligned}$$

Simulation of random variables (cont.)

Example 8.3.

- 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.

General simulation result

Lemma 8.4.

Let F be the distribution function of a continuous r.v. and let $g = F^{-1}$. Take $X \sim U(0,1)$, then $Y = g(X)$ has density F' and distribution function F.

Proof.

- Distribution functions are monotone increasing, so apply Lemma 8.2.
- Here $h = g^{-1} = F$.
- Hence by Lemma 8.2 the density of Y satisfies

$$f_Y(y) = f_X(h(y))|h'(y)| = 1 \cdot F'(y),$$

  as required.
- The form of the distribution function follows by integration.
- This generalizes Example 8.3. □

Section 8.2: The normal distribution

Definition 8.5.

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

- Not obvious $1/\sqrt{2\pi}$ is the right constant to make f_Z integrate to 1.
- (There's a nice proof involving polar coordinates).

Lemma 8.6.

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

$$\begin{aligned} \mathbb{E}(Z) &= 0, \\ \mathrm{Var}\,(Z) = \mathbb{E}(Z^2) &= 1. \end{aligned}$$

Proof of Lemma 8.6

Proof.

- $f_Z(z)$ is symmetric about 0. So

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

- Alternatively, notice that $z f_Z(z) = -\frac{d}{dz} f_Z(z)$ so that

$$\mathbb{E}(Z) = \int_{-\infty}^{\infty} z f_Z(z)\, dz = \int_{-\infty}^{\infty} -\frac{d}{dz} f_Z(z) dz = [-f_Z(z)]_{-\infty}^{\infty} = 0.$$

- Similarly, integration by parts gives

$$\mathbb{E}(Z^2) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z(z e^{-\frac{z^2}{2}})\, dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz = 1.$$

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

General normal distribution properties

Remark 8.7.

- *Often $f_Z(z)$ is denoted $\phi(z)$ and $F_Z(z)$ is denoted $\Phi(z)$.*
- *Not possible to write down a formula for $\Phi(z)$ using 'standard functions'.*
- *Instead values of $\Phi(z)$ are in tables, or can be calculated by computer. See second half of course.*

Definition 8.8.

A r.v. X has a normal distribution with mean μ and variance σ^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 8.9.

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 8.2

$$\begin{aligned} f_Z(z) &= f_X(h(z))|h'(z)| \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(h(z)-\mu)^2}{2\sigma^2}\right\} \times \sigma \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(\mu+\sigma z-\mu)^2}{2\sigma^2}\right\} = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}. \end{aligned}$$

□

General normal distribution properties (cont.)

Corollary 8.10.

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

Proof.

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

□

- Many quantities have an approximate normal distribution.
- For example heights in a population, measurement errors.
- There are good theoretical reasons for this (see Central Limit Theorem, Section 10).
- The normal distribution is *very* important in statistics.

Normal convergence

Theorem 8.11 (DeMoivre-Laplace).

Fix p, and let $X_n \sim Bin(n,p)$. 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 $X \sim \mathrm{Bin}(n,p)$ with large n and fixed p.
- Then $\frac{X_n - np}{\sqrt{np(1-p)}}$ is approximately $N(0,1)$ distributed.
- This is a special case of the *Central Limit Theorem*, see Section 10). .

Normal distribution tables

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
					$\Phi(z)$					
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936

Application

Example 8.12.

- 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 $X \sim \mathcal{N}(1.75, (0.05)^2)$.
- Let $Z = \frac{X-1.75}{0.05}$ so that $Z \sim \mathcal{N}(0,1)$.
- $$\begin{aligned} \mathbb{P}(X > 1.86) &= \mathbb{P}\left(\frac{X-1.75}{0.05} > \frac{1.86-1.75}{0.05}\right) \\ &= \mathbb{P}(Z > 2.2) = 1 - \mathbb{P}(Z \leq 2.2) = 1 - \Phi(2.2) \end{aligned}$$
- Can find $\Phi(2.2)$ from tabulated values, or (when you've done Stats part of course) using a computer language called R.
- The value is $\Phi(2.2) = 0.9861$. So $\mathbb{P}(X > 1.86) = 0.0139$.

Fact, proved in Section 10.4

Lemma 8.13.

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

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

- 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 5.18.
- See Lemma 10.18 for full proof.

Section 8.3: Jointly distributed continuous r.v.s

Definition 8.14.

- 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 $A \subset \mathbb{R}^2$

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

- Marginal density for X is $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy$.
- Conditional density for X given $Y = y$ is $f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$.
- 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 8.15.

- 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

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

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

Section 9: 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 9.1: Introduction

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

	pmf (discrete)	pdf (continuous)
joint	$p_{X,Y}(x,y)$	$f_{X,Y}(x,y)$
marg.	$p_X(x) = \sum_y p_{X,Y}(x,y)$ $p_Y(y) = \sum_x p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$ $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx$
cond.	$p_{X\|Y}(x\|y) = p_{X,Y}(x,y)/p_Y(y)$ $p_{Y\|X}(y\|x) = p_{X,Y}(x,y)/p_X(x)$	$f_{X\|Y}(x\|y) = f_{X,Y}(x,y)/f_Y(y)$ $f_{Y\|X}(y\|x) = f_{X,Y}(x,y)/f_X(x)$

Conditional expectation definition

Definition 9.1.

Define $\mathbb{E}(X \mid Y = y)$ to be the expected value of X using the conditional distribution of X given that $Y = y$:

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

Example

Example 9.2.

- X, Y discrete

$p_{X,Y}(x,y)$	$y=0$	1	2	3	$p_X(x)$
$x=0$	1/4	0	0	0	1/4
1	1/8	1/8	0	0	1/4
2	1/16	2/16	1/16	0	1/4
3	1/32	3/32	3/32	1/32	1/4
$p_Y(y)$	15/32	11/32	5/32	1/32	

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

x	0	1	2	3
$p_{X \mid Y}(x \mid 0)$	8/15	4/15	2/15	1/15

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

Example (cont.)

Example 9.2.

Similarly

$$
\begin{aligned}
\mathbb{E}(X \mid Y=1) &= 0 \times 0 + 1 \times \tfrac{4}{11} + 2 \times \tfrac{4}{11} + 3 \times \tfrac{3}{11} = \tfrac{21}{11} \\
\mathbb{E}(X \mid Y=2) &= 0 \times 0 + 1 \times 0 + 2 \times \tfrac{2}{5} + 3 \times \tfrac{3}{5} = \tfrac{13}{5} \\
\mathbb{E}(X \mid Y=3) &= 0 \times 0 + 1 \times 0 + 2 \times 0 + 3 \times 1 = 3
\end{aligned}
$$

Remark 9.3.

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, according to* p_Y*).*
- *Good to spend time thinking which type of object is which.*

Section 9.2: Expectation of a conditional expectation

Theorem 9.4 (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

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

Remark 9.5.

- *For Y discrete*

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

- *For Y continuous*

$$\mathbb{E}\left(\mathbb{E}[X \mid Y]\right) = \int_{-\infty}^{\infty} \mathbb{E}(X \mid Y = y) f_Y(y)\, dy.$$

Important notation

Remark 9.6.

- *Remember from Remark 9.3 that $\mathbb{E}(X \mid Y = y)$ is a function of y.*
- *Set $A(y) = \mathbb{E}(X \mid Y = y)$.*
- *Then the Tower Law (Theorem 9.4) 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 $\mathbb{E}(X \mid Y)$.*

Proof of Theorem 9.4

Proof.

For discrete Y, using the Partition Theorem 2.9 to expand $\mathbb{P}(X = x)$:

$$\begin{aligned}
\mathbb{E}(X) &= \sum_x x\mathbb{P}(X = x) \\
&= \sum_x x \left[\sum_y \mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y) \right] \\
&= \sum_y \left[\sum_x x\mathbb{P}(X = x \mid Y = y) \right] \mathbb{P}(Y = y) \\
&= \sum_y \mathbb{E}(X \mid Y = y)\mathbb{P}(Y = y)
\end{aligned}$$

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

Example 9.2 (cont.)

Example 9.7.

- Recall from Example 9.2

y	0	1	2	3
$p_Y(y)$	15/32	11/32	5/32	1/32
$\mathbb{E}(X \mid Y = y)$	11/15	21/11	13/5	3

- Hence

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

- Direct calculation from $p_X(x)$ confirms

$$\mathbb{E}(X) = 0 \times \frac{1}{4} + 1 \times \frac{1}{4} + 2 \times \frac{1}{4} + 3 \times \frac{1}{4} = \frac{3}{2}.$$

Tower law example

Example 9.8.

- A disoriented miner finds themselves in a room of the mine with three doors:
  - The first door brings them to safety after a 3 hours long hike.
  - The second door takes them back to the same room after 5 hours of climbing.
  - The third door takes them 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 they are in that room.
- What is the expected time after which the miner is safe?

Tower law example (cont.)

Example 9.8.

Let X be the time to reach safety, and Y the initial choice of a door ($= 1, 2, 3$). Then using Theorem 9.4

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}\left(\mathbb{E}(X \mid Y)\right) \\ &= \mathbb{E}(X \mid Y = 1) \cdot \mathbb{P}(Y = 1) + \mathbb{E}(X \mid Y = 2) \cdot \mathbb{P}(Y = 2) \\ &\quad + \mathbb{E}(X \mid Y = 3) \cdot \mathbb{P}(Y = 3) \\ &= 3 \cdot \frac{1}{3} + (\mathbb{E}X + 5) \cdot \frac{1}{3} + (\mathbb{E}X + 7) \cdot \frac{1}{3},\end{aligned}$$

which we rearrange as

$$3\mathbb{E}X = 15 + 2\mathbb{E}X; \qquad \mathbb{E}X = 15.$$

Example

Example 9.9.

- Nuts in a wood have an intrinsic hardness H, a non-negative integer random variable.
- The hardness H of a randomly selected nut has a Poi(1) distribution.
- If a nut has hardness $H = h$ a squirrel takes a geometric $\frac{1}{h+1}$ number of attempts to crack the nut.
- 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 $\mathbb{E}(X \mid H = h) = \frac{1}{\frac{1}{h+1}} = h + 1$.
- Therefore
  $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid H)) = \mathbb{E}(A(H)) = \mathbb{E}(H+1) = \mathbb{E}(H)+1 = 1+1 = 2.$

Important notation

Example 9.10.

- Remember we write A for the function $A(h) = \mathbb{E}(X \mid H = h)$.
- In the nut example, Example 9.9 $A(h) = \mathbb{E}(X \mid H = h) = h + 1$.
- Hence $A(H) = H + 1$ i.e. $\mathbb{E}(X \mid H) = H + 1$ [NB FUNCTION OF H].
- Therefore
  $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid H)) = \mathbb{E}(A(H)) = \mathbb{E}(H + 1) = \mathbb{E}(H) + 1 = 2.$

Section 9.3: Conditional variance

Definition 9.11.

The *conditional variance of* X*, given* Y is

$$\mathrm{Var}\,(X \mid Y) = \mathbb{E}\left[(X - \mathbb{E}(X \mid Y))^2 \mid Y\right] = \mathbb{E}(X^2 \mid Y) - \left[\mathbb{E}(X \mid Y)\right]^2.$$

- No surprise here, just use conditionals everywhere in the definition of variance.
- Notice that $\mathrm{Var}\,(X \mid Y)$ is again a function of Y (a random variable).
- If we write $A(Y)$ for $\left[\mathbb{E}(X \mid Y)\right]$ then we can rewrite Definition 9.11 as

$$\mathrm{Var}\,(X \mid Y) = \mathbb{E}(X^2 \mid Y) - A(Y)^2. \qquad (9.1)$$

Law of Total Variance

Proposition 9.12.

The Law of Total Variance holds:

$$\mathrm{Var}\,X = \mathbb{E}\left(\mathrm{Var}\,(X \mid Y)\right) + \mathrm{Var}\,\left(\mathbb{E}(X \mid Y)\right).$$

- In words: the variance is the expectation of the conditional variance plus the variance of the conditional expectation.
- Note that since $\mathrm{Var}\,(X \mid Y)$ and $\mathbb{E}(X \mid Y)$ 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 9.12 (not examinable)

Proof.

- Again we write $A(Y)$ for $\left[\mathbb{E}(X \mid Y)\right]$.
- Taking the expectation (over Y) of Equation (9.1) and applying the tower law Theorem 9.4 gives

$$\begin{aligned}\mathbb{E}\left(\mathrm{Var}\left(X \mid Y\right)\right) &= \mathbb{E}\left(\mathbb{E}(X^2 \mid Y) - A(Y)^2\right) \\ &= \mathbb{E}(X^2) - \mathbb{E}\left(A(Y)^2\right) \qquad (9.2)\end{aligned}$$

- Similarly, since Theorem 9.4 gives $\mathbb{E}(A(Y)) = \mathbb{E}\left(\mathbb{E}(X \mid Y)\right) = \mathbb{E}(X)$:

$$\begin{aligned}\mathrm{Var}\ \left(\mathbb{E}(X \mid Y)\right) &= \mathrm{Var}\ \left(A(Y)\right) \\ &= \mathbb{E}\left(A(Y)^2\right) - \left(\mathbb{E}(A(Y))\right)^2 \\ &= \mathbb{E}\left(A(Y)^2\right) - \left(\mathbb{E}(X)\right)^2. \qquad (9.3)\end{aligned}$$

- Notice that first term of (9.3) is minus the second term of (9.2). □

Proof of Proposition 9.12 (cont.)

Proof.

- Hence adding (9.2) and (9.3) together, cancellation occurs and we obtain:

$$\mathbb{E}\mathrm{Var}\ (X \mid Y) + \mathrm{Var}\ \mathbb{E}(X \mid Y) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \mathrm{Var}\ (X).$$

□

Section 9.4: Random sum

Definition 9.13.

- 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

Example 9.14 (Number of infections).

- Patient Zero infects N (a random number of) people with a virus.
- The ith infected person goes on to infect X_i people.
- Then $S = X_1 + \cdots + X_N$ is the total number of infected people in the second generation.

Example 9.15 (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.

Random sum examples (cont.)

Example 9.16.

- 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 ith 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?

Random sum theorem

Theorem 9.17.

For any random sum of the form of Definition 9.13

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

Proof.

- Condition on the (random) value of N. Let $A(n) = \mathbb{E}(S \mid N = n)$. Then

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

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

Section 10: 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 10.1: MGF definition and properties

Definition 10.1.

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 $M_X(t) = \begin{cases} \sum_i e^{tx_i} p_X(x_i) & X \text{ discrete} \\ \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx & X \text{ cts} \end{cases}$
- 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 10.2.

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

$$\begin{aligned} M_X(t) &= \sum_{x=1}^{\infty} e^{tx} p(1-p)^{x-1} \\ &= \sum_{x=1}^{\infty} pe^t \left((1-p)e^t\right)^{x-1} \\ &= pe^t \sum_{y=0}^{\infty} \left((1-p)e^t\right)^y \\ &= \frac{pe^t}{1-(1-p)e^t} \quad \text{defined for } (1-p)e^t < 1 \end{aligned}$$

Example: Poisson

Example 10.3.

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

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{1}{x!} \left(\lambda e^t\right)^x \\ &= e^{\lambda(e^t-1)}. \end{aligned}$$

Example: exponential

Example 10.4.

- Consider $X \sim \text{Exp}(\lambda)$
- $$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx \\ &= \lambda \int_0^\infty e^{-(\lambda - t)x}\, dx \\ &= \frac{\lambda}{\lambda - t} \left[-e^{-(\lambda - t)x} \right]_0^\infty \\ &= \frac{\lambda}{\lambda - t} \quad \text{defined for } t < \lambda \end{aligned}$$

Example: gamma

Example 10.5.

- Consider $X \sim \text{Gamma}(\alpha, \lambda)$
- Taking $y = (\lambda - t)x$ so $dy = (\lambda - t)dx$:

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}\, dx \\ &= \left(\frac{\lambda}{\lambda - t} \right)^\alpha \int_0^\infty \frac{(\lambda - t)^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-(\lambda - t)x}\, dx \\ &= \left(\frac{\lambda}{\lambda - t} \right)^\alpha \int_0^\infty \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}\, dy \\ &= \left(\frac{\lambda}{\lambda - t} \right)^\alpha \quad \text{defined for } t < \lambda \end{aligned}$$

M_X uniquely defines the distribution of X.

Theorem 10.6.

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 10.7.

The rth *moment* of X is $\mathbb{E}(X^r)$.

Lemma 10.8.

For any random variable X and for any t:

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

i.e. M_X "generates" the moments of X.

Proof of Lemma 10.8

Proof.

For any t, using the linearity of expectation:

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

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 $\frac{\mathbb{E}(X^k)}{k!}$.

Method 2 $M_X^{(k)}(0) = \mathbb{E}(X^k)$, where $M_X^{(k)}$ denotes the kth derivative of M_X.
To see this, note that

$$\begin{aligned} M_X'(t) &= \mathbb{E}(X) + t\mathbb{E}(X^2) + \frac{t^2}{2!}\mathbb{E}(X^3) + \cdots \\ M_X'(0) &= \mathbb{E}(X) \end{aligned}$$

$$\begin{aligned} M_X''(t) &= \mathbb{E}(X^2) + t\mathbb{E}(X^3) + \frac{t^2}{2!}\mathbb{E}(X^4) + \cdots \\ M_X''(0) &= \mathbb{E}(X^2) \\ &\quad \text{etc} \end{aligned}$$

Recovering moments of exponential: example

Example 10.9.

- Consider $X \sim \text{Exp}(\lambda)$
- We know from Example 10.4 that $M_X(t) = \frac{\lambda}{\lambda - t}$.
- To find $\mathbb{E}(X^r)$ 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 $M_X(t) = 1 + t\mathbb{E}(X) + \frac{t^2}{2!}\mathbb{E}(X^2) + \frac{t^3}{3!}\mathbb{E}(X^3) + \cdots$
- We see that $\frac{\mathbb{E}(X^k)}{k!} = \frac{1}{\lambda^k}$
- Hence $\mathbb{E}(X^k) = \frac{k!}{\lambda^k}$.

Recovering moments of gamma

Example 10.10.

- Recall from Example 10.5 that $M_X(t) = \lambda^\alpha(\lambda - t)^{-\alpha}$.
- To find $\mathbb{E}(X^r)$ use Method 2:

$$\begin{aligned} M'_X(t) &= \lambda^\alpha \alpha(\lambda - t)^{-\alpha-1} \\ \mathbb{E}(X) &= M'_X(0) = \frac{\alpha}{\lambda} \\ \\ M''_X(t) &= \lambda^\alpha \alpha(\alpha+1)(\lambda - t)^{-(\alpha+2)} \\ \mathbb{E}(X^2) &= M''_X(0) = \frac{\alpha(\alpha+1)}{\lambda^2} \end{aligned}$$

- This can be continued, but notice that with minimal work we can now see that

$$\text{Var}\,(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \left(\frac{\alpha}{\lambda}\right)^2 = \frac{\alpha}{\lambda^2}.$$

Section 10.2: Sums of random variables

Theorem 10.11.

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 5.14).

$$\begin{aligned} M_Z(t) &= \mathbb{E}(e^{tZ}) = \mathbb{E}\left(\prod_{i=1}^n e^{tX_i}\right) \\ &= \prod_{i=1}^n \mathbb{E}\left(e^{tX_i}\right) = \prod_{i=1}^n M_{X_i}(t). \end{aligned}$$

□

Example: adding Poissons

Example 10.12 (cf Theorem 5.18).

- If $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$, we deduce using Example 10.3 and Theorem 10.11 that $Z = X + Y$ has moment generating function

$$\begin{aligned} M_Z(t) &= M_X(t) M_Y(t) = e^{\lambda(e^t - 1)} . e^{\mu(e^t-1)} \\ &= e^{(\lambda+\mu)(e^t-1)}, \end{aligned}$$

- We deduce that (since it has the same MGF) $Z \sim \text{Poi}(\lambda + \mu)$ by Theorem 10.6.

Application: adding exponentials

Lemma 10.13.

- *Let* $X_1, X_2, \ldots, X_n$ *be independent Exp*(λ) *rvs, and let* $Z = X_1 + \cdots + X_n$.
- *Then*

$$M_{X_i}(t) = \frac{\lambda}{\lambda - t} \quad \text{for each } i = 1, \ldots, n.$$

- *Thus by Theorem 10.11:*

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

and $Z \sim$ *Gamma*(n, λ) *by the uniqueness theorem (Theorem 10.6)*

Section 10.3: Random sums

- In Theorem 9.17 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 10.14.

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 10.14

Proof.

- Let
$$\begin{aligned} A(n) &= \mathbb{E}(e^{tS} \mid N = n) \\ &= \mathbb{E}(e^{t(X_1+\cdots+X_N)} \mid N = n) \\ &= \mathbb{E}(e^{t(X_1+\cdots+X_n)} \mid N = n) \\ &= \mathbb{E}(e^{t(X_1+\cdots+X_n)}) \quad \text{since the } X_i\text{s are independent of } N \\ &= \mathbb{E}(e^{tX_1} \cdots e^{tX_n}) \\ &= \mathbb{E}(e^{tX_1}) \cdots \mathbb{E}(e^{tX_n}) \quad \text{since the } X_i\text{s are independent} \\ &= (M_X(t))^n \\ &= e^{n \log M_X(t)} \end{aligned}$$
- Thus $\mathbb{E}(e^{tS} \mid N) = A(N) = e^{N \log M_X(t)}$ and by Theorem 9.4
$$M_S(t) = \mathbb{E}(e^{tS}) = \mathbb{E}(\mathbb{E}(e^{tS} \mid N)) = \mathbb{E}(e^{N \log M_X(t)}) = M_N(\log M_X(t))$$

□

Example

Example 10.15.

- 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.
- First by Example 10.3:
$$M_N(t) = e^{\lambda(e^t - 1)}.$$
- We also know that $M_X(t) = \frac{1}{1-t}$ (Example 10.4).
- So
$$\begin{aligned} M_S(t) &= M_N(\log M_X(t)) = e^{\lambda(e^{\log M_X(t)} - 1)} = e^{\lambda(M_X(t) - 1)} \\ &= e^{\lambda(\frac{1}{1-t} - 1)} = e^{\lambda(\frac{t}{1-t})}. \end{aligned}$$
- From this we can calculate $\mathbb{E}(S)$, $\mathrm{Var}\,(S)$, etc.

Section 10.4: MGF of the normal

Example 10.16.

- 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 $y = x - t$. Key is that $t(y+t) - \frac{(y+t)^2}{2} = -\frac{1}{2}\left[y^2 - t^2\right]$ so

$$\begin{aligned} M_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(y+t) - \frac{(y+t)^2}{2}}\, dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[y^2 - t^2]}\, dy \\ &= e^{\frac{1}{2}t^2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2}\, dy \\ &= e^{\frac{1}{2}t^2} \end{aligned}$$

MGF of the general normal

Example 10.17.

- Now let $Y \sim \mathcal{N}(\mu, \sigma^2)$
- Set $X = \frac{Y-\mu}{\sigma}$ so $X \sim \mathcal{N}(0,1)$ by Lemma 8.9.
- Then $Y = \mu + \sigma X$ and

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t(\mu + \sigma X)}) \\ &= \mathbb{E}(e^{\mu t} e^{\sigma t X}) = e^{\mu t}\mathbb{E}(e^{(\sigma t)X}) \\ &= e^{\mu t} M_X(\sigma t) = e^{\mu t} e^{\frac{1}{2}(\sigma t)^2} \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2} \end{aligned}$$

Normal distribution properties

Lemma 10.18.

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* β *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 of Lemma 10.18

Proof.

1. Let $Y = X + c$. Then

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t(X+c)}) = e^{tc}\mathbb{E}(e^{tX}) = e^{tc}M_X(t) \\ &= e^{tc}e^{\mu t + \frac{1}{2}\sigma^2 t^2} = e^{(\mu+c)t + \frac{1}{2}\sigma^2 t^2} \end{aligned}$$

So $Y \sim \mathcal{N}(\mu + c, \sigma^2)$ by uniqueness, Theorem 10.6.

□

Proof of Lemma 10.18 (cont).

Proof.

2. Let $Y = \beta X$. Then

$$\begin{aligned} M_Y(t) &= \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t\beta X}) = M_X(\beta t) \\ &= e^{\mu\beta t + \frac{1}{2}\sigma^2(\beta t)^2} = e^{\mu\beta t + \frac{1}{2}\beta^2\sigma^2 t^2} \end{aligned}$$

   So $Y \sim \mathcal{N}(\beta\mu, \beta^2\sigma^2)$ by uniqueness, Theorem 10.6.

3. Let $Z = X + Y$. Then by Theorem 10.11

$$\begin{aligned} M_Z(t) &= M_X(t) M_Y(t) \\ &= e^{\mu_X t + \frac{1}{2}\sigma_X^2 t^2} e^{\mu_Y t + \frac{1}{2}\sigma_Y^2 t^2} \\ &= e^{(\mu_X + \mu_Y)t + \frac{1}{2}(\sigma_X^2 + \sigma_Y^2)t^2} \end{aligned}$$

   So $Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ by uniqueness, Theorem 10.6.

□

Example: heights

Example 10.19.

- 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 10.18.3,
  $X_1 + X_2 + X_3 \sim \mathcal{N}(175 + 175 + 175, 33 + 33 + 33)$.

Example: heights (cont.)

Example 10.19.

- Let $W = \frac{X_1+X_2+X_3}{3}$ be the average height of the male students. By Lemma 10.18.2

$$W \sim \mathcal{N}\left(\frac{1}{3}(3 \times 175), \left(\frac{1}{3}\right)^2 (3 \times 33)\right) = \mathcal{N}(175, 11).$$

- 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 10.18.2.
- So $D \sim \mathcal{N}(170 + (-175), 25 + 11)$ by Lemma 10.18.3, i.e. $D \sim \mathcal{N}(-5, 36)$ or $\frac{D+5}{6} \sim \mathcal{N}(0, 1)$.
- We want to know $\mathbb{P}(D > 0) = \mathbb{P}(\frac{D+5}{6} > \frac{5}{6}) = 1 - \Phi(\frac{5}{6})$. Using tables or R we can find $\Phi(\frac{5}{6}) = 0.7976$, so $\mathbb{P}(D > 0) = 1 - 0.7976 = 0.2024$.

Section 10.5: Central Limit Theorem

- Consider IID $X_1, \ldots, X_n$ with mean μ and variance σ^2.
- The Weak Law of Large Numbers (Theorem 6.20) tells us that $\frac{1}{n}(X_1 + \ldots + X_n) \simeq \mu$ or $X_1 + \ldots + X_n - n\mu \simeq 0$.
- 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 10.20.

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

Central Limit Theorem

Theorem 10.21 (Central Limit Theorem (CLT)).

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

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

Remark 10.22.

- *Notice that $X_1 + \ldots + X_n$ has mean $n\mu$ and variance $n\sigma^2$.*
- *CLT implies that for large n, the $X_1 + \ldots + X_n \simeq N(n\mu, n\sigma^2)$ or equivalently $\frac{1}{\sqrt{n\sigma^2}}(X_1 + \ldots + X_n - n\mu) \simeq N(0,1)$.*
- *If $X_i \sim$ Bernoulli(p) this reduces to the de Moivre–Laplace Theorem 8.11*

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

$$\begin{aligned} M_{T_n}(t) &= \mathbb{E}\left(e^{\frac{t}{\sqrt{n}}\sum_{i=1}^n X_i}\right) = \mathbb{E}\left(\prod_{i=1}^n e^{\frac{t}{\sqrt{n}}X_i}\right) = \prod_{i=1}^n \mathbb{E}\left(e^{\frac{t}{\sqrt{n}}X_i}\right) \\ &= \prod_{i=1}^n M_{X_i}\left(\frac{t}{\sqrt{n}}\right) = \left[M_X\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &= \left(1 + \frac{1}{2}\frac{t^2}{n} + O(n^{-3/2})\right)^n \to e^{t^2/2}, \end{aligned}$$

as required. □