Probability I

David Puerta

Based on Lectures by Debleena Thacker 2024

Abstract

These set of notes will cover the content of the 2024/25 Probability 1 course given by Debleena Thacker. The aim is for these notes to be a *condensed*[1] version of the lecture notes provided with less of a focus on a copious amount examples and more on the key concepts with additional comments and further explanation when needed.

[1]Short, to the point notes that prioritises an audience who has comfortably already seen the material, needing only a reminder of the key points.

Contents

1 Axioms **4**
- 1.1 Sample space and events . . . 4
- 1.2 Axioms of probability . . . 5
- 1.3 Consequences of the axioms . . . 5

2 Equally likely outcomes and counting principles **8**
- 2.1 Classical Probability . . . 8
- 2.2 Counting Principles . . . 10

3 Conditional Probability and Independence **11**
- 3.1 Conditional Probability . . . 11
- 3.2 Independence of events . . . 14

4 Random variables **15**
- 4.1 Definition . . . 15
- 4.2 Discrete random variables . . . 16
- 4.3 Discrete distributions . . . 17
- 4.4 Continuous random variables . . . 18
- 4.5 Continuous distributions . . . 21
- 4.6 Functions of random variables . . . 21

5 Multiple random variables **23**
- 5.1 Bivariate random variables . . . 23
- 5.2 Jointly distributed discrete random variables . . . 23
- 5.3 Jointly continuously distributed random variables . . . 27
- 5.4 Functions of multiple random variables . . . 30

6 Expectation **31**
- 6.1 Definition . . . 31
- 6.2 Expectation of functions of random variables . . . 31
- 6.3 Linearity of expectation . . . 33
- 6.4 Variance and covariance . . . 34
- 6.5 Conditional expectation . . . 38
- 6.6 Expectation and probability inequalities . . . 42

7 Moment generating functions **44**
- 7.1 Definition . . . 44
- 7.2 Properties, theorems and uses . . . 44

8 Limit theorems **47**
- 8.1 The weak law of large numbers . . . 47
- 8.2 The central limit theorem . . . 48

1 Axioms

1.1 Sample space and events

So far you most likely have seen probability in the context of assigning numerical values to events (the probability of a fair coin landing on heads), random variables, probability distributions and even abstract formulae such as Bayes' Theorem.
All of these are typically taught using intuition as the foundation/ justification when introducing the concepts. We will now formalise these concepts in probability, taking an axiomatic approach to probability.
In every scenario where we wish to use probability we have an experiment and outcomes of said experiment. An experiment can be anything from rolling a die to determining whether a patient has a life threatening disease. We will denote the outcome of an experiment ω.

Definition 1.1 (Sample space)**.** A sample space for a given experiment is the set Ω of all possible outcomes ω of the experiment.

Definition 1.2 (Event)**.** An event A is a subset of Ω. The collection $\mathcal{F}$ is the set of all subsets of the sample space Ω.

Definition 1.3 (Disjoint events)**.** Two events A and B are disjoint or mutually exclusive if $A \cap B = \emptyset$.

Example 1.1. Consider rolling a fair six sided die and noting the score.

The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$ with the event of rolling an even number being $A = \{2, 4, 6\}$ and the event of rolling an odd number being $B = \{1, 3, 5\}$.

We see that A and B are mutually exclusive as $A \cap B = \emptyset$.

Theorem 1.1 (De Morgan's Laws)**.** For a collection of events A_i we have,

$$\Big(\bigcup_i A_i\Big)^c = \bigcap_i A_i^c \quad \text{and} \quad \Big(\bigcap_i A_i\Big)^c = \bigcup_i A_i^c$$

with A^c denoting the compliment of the event A.

Proof. Suppose we have

$$x \in \Big(\bigcup_i A_i\Big)^c.$$

Then we have that

$$x \in \Big(\bigcup_i A_i\Big)^c \Leftrightarrow \quad x \notin A_i \text{ for all } i \quad \Leftrightarrow \quad x \in A_i^c \text{ for all } i \quad \Leftrightarrow x \in \bigcap_i A_i^c.$$

Then using the fact that $(A^c)^c = A$ we have that

$$\Big(\Big(\bigcup_i A_i\Big)^c\Big)^c = \Big(\bigcap_i A_i^c\Big)^c \Longleftrightarrow \Big(\bigcap_i A_i\Big)^c = \bigcup_i A_i^c$$

□

1.2 Axioms of probability

We will now define the space in which we will work in when using probability; this is similar to defining the 2D Cartesian space so that we can make sense of functions such as $y = x$.

Definition 1.4 (Probability measure)**.** The map $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called the probability of our probability space and satisfies the following:

(A1) $\mathbb{P}(A) \geq 0$ for every $A \in \mathcal{F}$.

(A2) $\mathbb{P}(\Omega) = 1$.

(A3) For $A_1, A_2, \cdots$ with $A_i \cap A_j = \emptyset$ for all i, j we have

$$\mathbb{P}\Big(\bigcup_i A_i\Big) = \sum_i \mathbb{P}(A_i)$$

.

Definition 1.5 (Sigma-algebra)**.** The collection of all subsets of the sample space $\mathcal{F}$ is called a σ-algebra and satisfies the following:

(S1) $\Omega \in \mathcal{F}$.

(S2) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.

(S3) $A_1, A_2, \cdots \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Definition 1.6 (Probability Space)**.** A probability space is the triple $(\Omega, \mathcal{F}, \mathbb{P})$, this is the sample space, the sigma-algebra and the probability (measure).

The probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for an experiment consists of the possible outcomes Ω, a 'nice' collection of subsets $\mathcal{F}$ to work in and a function $\mathbb{P}$ that assigns a number (the probability) to an event $A \in \mathcal{F}$. If Ω is finite we can take $\mathcal{F} = 2^{\Omega}$. Although it may seem cumbersome and unintuitive to start probability with abstract definitions, these definitions lead rise to all the properties we are used to in probability.

1.3 Consequences of the axioms

The axioms A1-3 result in the following results.

Theorem 1.2.

$$\mathbb{P}(\emptyset) = 0.$$

Proof. Noting that $\Omega = \Omega \cup \emptyset$ and $\Omega \cap \emptyset = \emptyset$ then we can use A3 to get

$$\mathbb{P}(\Omega \cup \emptyset) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) = \mathbb{P}(\Omega) \Rightarrow \mathbb{P}(\emptyset) = 0.$$

□

This can be thought of the event that nothing happens $\emptyset$ is impossible in an experiment where something must happen - an impossible event.

Theorem 1.3. For any event A,

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

Proof. Using the fact that $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$ for any event A, then

$$\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) = \mathbb{P}(\Omega) = 1 \Rightarrow \mathbb{P}(A^c) = 1 - \mathbb{P}(A).$$

□

Theorem 1.4. For two events A and B,

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

Proof. We see that $\{B \setminus A\} \cup \{A \cap B\} = B$ and $\{B \setminus A\} \cap \{A \cap B\} = \emptyset$ then by A3,

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

□

Theorem 1.5. If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

Proof. Let $A \subseteq B$, then $A \cap B = A$. Applying Theorem 1.4 we have

$$\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A) \geq 0$$

by A1. □

Theorem 1.6. For any event A,

$$\mathbb{P}(A) \leq 1.$$

Proof. For any event A we have that $A \subseteq \Omega$, hence using Theorem 1.5

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

□

Theorem 1.7. For any two events A and B,

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

Proof. Observe that $\{B \setminus A\} \cap A = \emptyset$ and $\{B \setminus A\} \cup A = A \cup B$. Hence by A3 we have

$$\mathbb{P}\Big(\{B \setminus A\} \cup A\Big) = \mathbb{P}(A \cup B) = \mathbb{P}(B \setminus A) + \mathbb{P}(A).$$

Finally, applying Theorem 1.4 we are left with

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

□

Theorem 1.8. For any events $A_1, A_2, \cdots$ we have

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Proof. Let $A_1, A_2, \cdots$ be any events. If $A_1, A_2, \cdots$ are pairwise disjoint then we have

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

from A3. Else, we have that $A_j \cap A_k$ is non-empty for some j, k. Thus, (without loss of generality) $A_j \subseteq A_k$. Let $B = \Big\{\bigcup_{i\neq j,k} A_i\Big\} \cap \{A_j \cup A_k\}$. We have that

$$\mathbb{P}\Big(\bigcup_{i} A_i\Big) = \mathbb{P}\Big(\bigcup_{i\neq j,k} A_i \cup A_j \cup A_k\Big)$$

and by Theorem 1.7 and A1,

$$\mathbb{P}\Big(\Big\{\bigcup_{i\neq j,k} A_i\Big\} \cup \{A_j \cup A_k\}\Big) = \mathbb{P}\Big(\bigcup_{i\neq j,k} A_i\Big) + \mathbb{P}(A_j \cup A_k) - \mathbb{P}(B) \leq \mathbb{P}\Big(\bigcup_{i\neq j,k} A_i\Big) + \mathbb{P}(A_j \cup A_k).$$

Finally, using Theorem 1.7 and A1 again on $\mathbb{P}(A_j \cup A_k)$ we have that $\mathbb{P}(A_j \cup A_k) = \mathbb{P}(A_j) + \mathbb{P}(A_k) - \mathbb{P}(A_j \cap A_k) \leq \mathbb{P}(A_j) + \mathbb{P}(A_k)$ and we arrive at

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) \leq \mathbb{P}\Big(\bigcup_{i\neq j,k} A_i\Big) + \mathbb{P}(A_j \cup A_k) \leq \mathbb{P}\Big(\bigcup_{i\neq j,k} A_i\Big) + \mathbb{P}(A_j) + \mathbb{P}(A_k) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

□

Theorem 1.9. If the events $A_1 \subseteq A_2 \subseteq \cdots$ is an increasing sequence of events then,

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

If the events $A_1 \supseteq A_2 \supseteq \cdots$ is an decreasing sequence of events then,

$$\mathbb{P}\Big(\bigcap_{i=1}^{\infty} A_i\Big) = \lim_{i\to\infty} \mathbb{P}(A_i).$$

Proof. Firstly we note that as $A_1 \subseteq A_2 \subseteq \cdots$ then

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} A_i \setminus A_{i-1}$$

with $A_0 = \emptyset$. We see that the events $B_n = A_n \setminus A_{n-1}$ are mutually exclusive, thus by A3 we have

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i \setminus A_{i-1}\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i \setminus A_{i-1}) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{P}(A_i \setminus A_{i-1}).$$

Applying A3 again to the finite collection of events $B_1, B_2, \cdots, B_n$ we see that

$$\sum_{i=1}^{n} \mathbb{P}(B_n) = \sum_{i=1}^{n} \mathbb{P}(A_i \setminus A_{i-1}) = \mathbb{P}\Big(\bigcup_{i=1}^{n} A_i \setminus A_{i-1}\Big) = \mathbb{P}(A_n).$$

Hence we arrive at

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

Define an increasing sequence of events $A_1^c \subseteq A_2^c \subseteq \cdots$. We find that

$$\Big(\bigcup_{i=1}^{\infty} A_i^c\Big)^c = \bigcap_{i=1}^{\infty} A_i$$

by De Morgan's law. In addition, we have that $A_1 \supseteq A_2 \supseteq \cdots$ by complementation. Using the above result with Theorem 1.3 we see that

$$\mathbb{P}\Big(\bigcap_{i=1}^{\infty} A_i\Big) = \mathbb{P}\Big(\Big[\bigcup_{i=1}^{\infty} A_i^c\Big]^c\Big) = 1 - \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_n^c\Big) = 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) = 1 - \lim_{n\to\infty} \Big(1 - \mathbb{P}(A_n)\Big) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

□

Definition 1.7 (Partition)**.** The events $E_1, E_2, \cdots \in \mathcal{F}$ form a partition of the sample space Ω if:

(a) $\mathbb{P}(E_i) > 0$ for all i.

(b) $E_i \cap E_j = \emptyset$ for $i \neq j$.

(c) $\bigcup_{i=1}^{\infty} E_i = \Omega$.

Theorem 1.10. If the events $E_1, E_2, \cdots$ form a partition then

$$\sum_{i=1}^{\infty} \mathbb{P}(E_i) = 1.$$

Proof. Using A3, A2 and the definition of a partition $E_1, E_2, \cdots$ we have

$$\sum_{i=1}^{\infty} \mathbb{P}(E_i) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \mathbb{P}(\Omega) = 1.$$

□

2 Equally likely outcomes and counting principles

2.1 Classical Probability

In Chapter 1 we discussed the abstract axioms of probability. The simplest scenario to apply these axioms is an experiment with a finite number of outcomes, all of which are equally likely; this is what is called Classical Probability.

If we have an experiment with the sample space $\Omega = \{\omega_1, \cdots, \omega_n\}$ with the outcomes ω_i equally likely then we define the probability $\mathbb{P} : \mathcal{F} \to [0, 1]$ to be the map

$$\mathbb{P}(\omega_i) = \frac{1}{|\Omega|} = \frac{1}{n}$$

for all $1 \leq i \leq n$. If we have an event A then we can define

$$\mathbb{P}(A) := \frac{|A|}{|\Omega|}$$

with $\mathcal{F} = 2^\Omega$.

Lemma 2.1. The triple $(\Omega, \mathcal{F}, \mathbb{P})$ defined as above is a probability space.

Proof. (A1) We see that for any $A \in 2^\Omega$ we have

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \geq 0$$

as $|A| \geq 0$. (A2) For Ω we have

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

(A3) Let $A_1, A_2, \cdots, A_k \in 2^\Omega$ be disjoint (this is finite as Ω is finite), then we have that

$$\mathbb{P}\Big(\bigcup_{i=1}^{k} A_i\Big) = \frac{\left|\bigcup_{i=1}^{k} A_i\right|}{|\Omega|} = \sum_{i=1}^{k} \frac{|A_i|}{|\Omega|}$$

as $\left|\bigcup_{i=1}^{k} A_i\right| = \sum_{i=1}^{k} |A_i|$ due to the events A_i being disjoint. Hence we have

$$\mathbb{P}\Big(\bigcup_{i=1}^{k} A_i\Big) = \sum_{i=1}^{k} \mathbb{P}(A_i).$$

(S1-3) We are given that 2^Ω is a sigma algebra. □

Example 2.1. Continuing from Example 1.1, consider the same fair six sided die. The probability of rolling an even number is the probability of the event $A = \{2, 4, 6\}$ which by our definition is

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{3}{6} = \frac{1}{2}$$

which agrees with our intuitive definition of what probability is in this scenario.

Example 2.2. Consider rolling two fair dice, noting their score as the pair (i, j) where i is the score on the first die and j the second.
The probability of rolling the same number on both die is the event $C = \{(i, i) : \text{for } i = 1, 2, 3, 4, 5, 6\}$ with the sample space being $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$. Thus we have that

$$\mathbb{P}(C) = \frac{|C|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}.$$

2.2 Counting Principles

As we have defined the probability of an event in the equally likely outcomes setting as the cardinality of a set, to find the probability of an event we are simply counting the number of ways it can occur. We can use the following facts when counting the number of ways an event can occur.

Counting Principle 2.1. Suppose we make k choices in succession with the number of possible selections for each choice being m_k and each choice is independent. The total number of choices is

$$m_1 \times m_2 \times \cdots \times m_k = \prod_{i=1}^{k} m_i.$$

Counting Principle 2.2. Suppose we have m objects and we select r objects *with* replacement. The total number of possibilities is then

$$\underbrace{m \times m \times \cdots \times m}_{r-\text{times}} = m^r$$

Counting Principle 2.3. Suppose we have m objects and we select r objects *without* replacement. The total number of possibilities is then

$$(m)_r := m \times (m-1) \times \cdots \times (m-r+1) = \frac{m!}{(m-r)!}$$

where $(m)_r$ is the falling factorial.

Counting Principle 2.4. Suppose we have m objects and wish to chose a subset of size r without replacement. The number of possibilities is then

$$\binom{m}{r} := \frac{(m)_r}{r!} = \frac{m!}{r!(m-r)!}$$

with $\binom{m}{r}$ being the binomial coefficient.

Counting Principle 2.5. Suppose we have m objects, r of type-1 and $m-r$ of type 2. The number of distinct ordered choices of the m objects is

$$\binom{m}{r}.$$

Counting Principle 2.6. Suppose we have m objects and wish to divide them into k distinct groups. The total number of ways is

$$\binom{m+k-1}{m} = \binom{m+k-1}{k-1}.$$

Example 2.3. Suppose we have $n < 365$ people in a room. Let B be the event that at least two people have the same birthday. Using Theorem 1.3 we see that

$$\mathbb{P}(B) = 1 - \mathbb{P}(B^c)$$

where B^c is the event that no two people share a birthday. We see that for each person to have a different birthday each person has to have a different birthday from the last. The first person has 365 choices for their birthday, the second has 364 choices etc. As we have n people then we have

$$365 \times 364 \times \cdots \times (365 - n + 1) = (365)_n$$

total choices where each person has a different birthday. Hence we have that

$$\mathbb{P}(B^c) = \frac{(365)_n}{365^n} \Rightarrow \mathbb{P}(B) = 1 - \frac{(365)_n}{365^n}$$

as we have 365^n total choices.

When doing any counting it is well worth taking your time and not immediately fixating on an answer you believe is correct purely by intuition; this typically leads to an inability to understand the correct answer later as it clashes with your intuition.

3 Conditional Probability and Independence

3.1 Conditional Probability

The phrase "What are the chances of X happening given Y" in English is simple: we know that Y has happened and we want to know the chances of X happening. Often times the event Y is excluded when deciding on the chances of X. This is in direct contrast with the mathematical interpretation of the phrase: finding the probability that both X and Y occur then divide by the probability of Y occurring.
An example to illustrate this potential discrepancy is the following: suppose we have a fair die and wish to find the probability that we roll a 6 given that the number we roll is even. We could argue that as we are told we roll an even number ($2, 4$ or 6) then the probability we get a 6 is $1/3$ as they are equally likely. Yet this is not the correct answer to the mathematical interpretation of the question, as we have calculated the probability that we roll and even number *and* we roll a 6. The probability that we roll an even number in the first place is $1/2$ thus the answer would be $2/3$. This leads us to the definition of the mathematical meaning of "What are the chances of X happening given Y".

Definition 3.1 (Conditional Probability)**.** For events $A, B \subseteq \Omega$ with $\mathbb{P}(B) \neq 0$, we define

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

as the conditional probability of A given B.

It is critical to understand that the conditional probability of A given B *does not* represent an event.

Theorem 3.1. For any event $B \subseteq \Omega$ with $\mathbb{P}(B) > 0$ we have that $\mathbb{P}(\cdot|B)$ is a probability measure on Ω.

Proof. (A1) For any event $A \in \mathcal{F}$ we have that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \geq 0$$

as $A \cap B \in \mathcal{F}$ and as $\mathbb{P}$ is a probability measure on Ω then $\mathbb{P}(A \cap B) \geq 0$. (A2) We see that

$$\mathbb{P}(\Omega|B) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

(A3) Let $A_1, A_2, \cdots \in \mathcal{F}$ be disjoint, then

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i|B\Big) = \frac{\mathbb{P}\Big(\{\bigcup_{i=1}^{\infty} A_i\} \cap B\Big)}{\mathbb{P}(B)}.$$

We see that

$$\Big\{\bigcup_{i=1}^{\infty} A_i\Big\} \cap B = \bigcup_{i=1}^{\infty}\{A_i \cap B\}$$

thus we arrive at

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i|B\Big) = \frac{\mathbb{P}\Big(\bigcup_{i=1}^{\infty}\{A_i \cap B\}\Big)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i|B)$$

as $\mathbb{P}$ is a probability over Ω. □

Theorem 3.2. For events $A, B, C \subseteq \Omega$ with $\mathbb{P}(B \cap C) > 0$,

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

Proof. From the definition,

$$\mathbb{P}(B|C)\mathbb{P}(A|B \cap C) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} \cdot \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B|C)} = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(C)} = \mathbb{P}(A \cap B|C).$$

□

Theorem 3.3. For any events $A_0, A_1, \cdots, A_n$ with $\mathbb{P}\Big(\bigcap_{i=0}^{n-1} A_i\Big) > 0$,

$$\mathbb{P}\Big(\bigcap_{i=1}^{n} A_i|A_0\Big) = \prod_{m=0}^{n-1} \mathbb{P}\Big(A_{m+1}\Big|\bigcap_{i=0}^{m} A_i\Big)$$

Proof. Using induction we have our base case $n = 1$:

$$\mathbb{P}\Big(\bigcap_{i=1}^{1} A_i|A_0\Big) = \mathbb{P}(A_1|A_0) = \prod_{m=0}^{0} \mathbb{P}\Big(A_{m+1}\Big|\bigcap_{i=0}^{m} A_i\Big)$$

thus true for $n = 1$. Suppose

$$\mathbb{P}\Big(\bigcap_{i=1}^{k} A_i | A_0\Big) = \prod_{m=0}^{k-1} \mathbb{P}\Big(A_{m+1}\Big|\bigcap_{i=0}^{m} A_i\Big)$$

for $n = k$. Consider $n = k + 1$,

$$\mathbb{P}\Big(\bigcap_{i=1}^{k+1} A_i | A_0\Big) = \mathbb{P}\Big(A_{k+1} \cap \Big\{\bigcap_{i=1}^{k} A_i\Big\}\Big| A_0\Big)$$

by Theorem 3.2 we have

$$\mathbb{P}\Big(A_{k+1} \cap \Big\{\bigcap_{i=1}^{k} A_i\Big\}\Big| A_0\Big) = \mathbb{P}\Big(\bigcap_{i=1}^{k} A_i | A_0\Big)\mathbb{P}\Big(A_{k+1}\Big|\bigcap_{i=0}^{k} A_i\Big) = \mathbb{P}\Big(A_{k+1}\Big|\bigcap_{i=0}^{k} A_i\Big) \prod_{m=0}^{k-1} \mathbb{P}\Big(A_{m+1}\Big|\bigcap_{i=0}^{m} A_i\Big)$$

using the induction hypothesis. Hence,

$$\mathbb{P}\Big(\bigcap_{i=1}^{k+1} A_i | A_0\Big) = \prod_{m=0}^{k} \mathbb{P}\Big(A_{m+1}\Big|\bigcap_{i=0}^{m} A_i\Big)$$

and true for $n = k + 1$. As true for $n = 1$, $n = k + 1$ and assumed true for $n = k$ then statement is true for all $n \geq 1$. □

Theorem 3.4. Let $E_1, E_2, \cdots$ form a partition on Ω. For any event A,

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)\mathbb{P}(A|E_i)$$

and for any event B with $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A|B) = \sum_{i=1}^{\infty} \mathbb{P}(E_i|B)\mathbb{P}(A|E_i \cap B).$$

Proof. As $E_1, E_2, \cdots$ forms a partition we have $\bigcup_{i=1}^{\infty} E_i = \Omega$. Using the fact that $A = A \cap \Omega$,

$$A = A \cap \Omega = A \cap \Big\{\bigcup_{i=1}^{\infty} E_i\Big\} = \bigcup_{i=1}^{\infty} \{A \cap E_i\}.$$

From the definition of conditional probability, $\mathbb{P}(A \cap E_i) = \mathbb{P}(E_i)\mathbb{P}(A|E_i)$. Hence,

$$\mathbb{P}(A) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} \{A \cap E_i\}\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap E_i)$$

by A3 as the events $\{A \cap E_i\} \cap \{A \cap E_j\} = \emptyset$ for $i \neq j$. Hence we arrive at

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

Similarly, using the definition of conditional probability

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1}{\mathbb{P}(B)} \sum_{i=1}^{\infty} \mathbb{P}(E_i)\mathbb{P}(A \cap B|E_i).$$

using the first part on the event $A \cap B$. Simplifying,

$$\frac{1}{\mathbb{P}(B)} \sum_{i=1}^{\infty} \mathbb{P}(E_i)\mathbb{P}(A \cap B|E_i) = \sum_{i=1}^{\infty} \frac{\mathbb{P}(E_i)}{\mathbb{P}(B)} \cdot \frac{\mathbb{P}(A \cap B \cap E_i)}{\mathbb{P}(E_i)} = \sum_{i=1}^{\infty} \frac{\mathbb{P}(A \cap B \cap E_i)}{\mathbb{P}(B)}.$$

We see that

$$\frac{\mathbb{P}(A \cap B \cap E_i)}{\mathbb{P}(B)} = \frac{\mathbb{P}(E_i \cap B)}{\mathbb{P}(B)} \cdot \frac{\mathbb{P}(A \cap B \cap E_i)}{\mathbb{P}(E_i \cap B)} = \mathbb{P}(E_i|B)\mathbb{P}(A|E_i \cap B)$$

as $\mathbb{P}(E_i \cap B) \neq 0$, thus

$$\mathbb{P}(A|B) = \sum_{i=1}^{\infty} \mathbb{P}(E_i|B)\mathbb{P}(A|E_i \cap B).$$

□

Theorem 3.5 (Bayes' Theorem)**.** For any events A and B with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$,

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

Proof. From the definition,

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

as $A \cap B = B \cap A$. □

Typically it is also useful to partition the denominator of Bayes' Theorem. Let $E_1, E_2, \cdots$ be a partition on Ω, then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\sum_{i=1}^{\infty} \mathbb{P}(E_i)\mathbb{P}(B|E_i)}$$

which is often used when $\mathbb{P}(B)$ is difficult to calculate.

3.2 Independence of events

Definition 3.2 (Independence of events)**.** Two events A and B are independent if

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

and they are conditionally independent given a third event C with $\mathbb{P}(C) > 0$ if

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

Immediately from the definition we see that two events A and B being independent is also the same as

$$\mathbb{P}(A|B) = \mathbb{P}(A) \quad \text{or} \quad \mathbb{P}(B|A) = \mathbb{P}(B)$$

by using the definition of the conditional probability.

Definition 3.3 (Mutually independent events)**.** A collection of events $\mathcal{A} \subseteq \mathcal{F}$ are mutually independent if for every finite subset $\mathcal{C} \subseteq \mathcal{A}$,

$$\mathbb{P}\Big(\bigcap_{A\in\mathcal{C}} A\Big) = \prod_{A\in\mathcal{C}} \mathbb{P}(A).$$

Similarly, a collection of events $\mathcal{A} \subseteq \mathcal{F}$ are mutually conditionally independent given the event B if for every finite subset $\mathcal{C} \subseteq \mathcal{A}$,

$$\mathbb{P}\Big(\bigcap_{A\in\mathcal{C}} A\Big|B\Big) = \prod_{A\in\mathcal{C}} \mathbb{P}(A|B).$$

Example 3.1. The events A, B and C are mutually independent if,

$$\begin{aligned}
&\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C);\\
&\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B);\\
&\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C);\\
&\mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C).
\end{aligned}$$

4 Random variables

4.1 Definition

Definition 4.1 (Random variables)**.** A random variable $X : \Omega \to X(\Omega)$ on a sample space Ω is a mapping which maps outcomes $\omega \to X(\omega)$.

If $X(\Omega) \subseteq \mathbb{R}$ then X is a real-valued random variable (univariate random variable). If $X(\Omega) \subseteq \mathbb{R}^n$ then X is a vector-valued random variable (multivariate random variable).

Example 4.1. Continuing from Example 2.2, consider rolling two fair dice and noting the score (i, j) with the sample space

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}.$$

We will define the random variable X by

$$X\Big((i, j)\Big) := i + j$$

which is the sum of the two scores on the dice. In this case we see that $X(\Omega) = \{2, 3, \cdots, 11, 12\}$ as they are the possible total scores.

At a glance, the definition of a random variable may seem to not actually define what a random variable is. A random variable X is a mapping! Just as you can define as many mapping on the Cartesian plane as you want (e.g $f : \mathbb{R} \to \mathbb{R}$ with $x \to x^2$), you can define as many random variables on the sample space as you want with some more useful when answering questions than others.

Theorem 4.1. The function $\mathbb{P}_X : X(\Omega) \to [0,1]$ defined by

$$\mathbb{P}_X(B) := \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$$

for $B \subseteq X(\Omega)$ is a probability measure on $X(\Omega)$.

Proof. (A1) Let $B \subseteq X(\Omega)$,

$$\mathbb{P}_X(B) = \mathbb{P}(X \in B) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}) \geq 0$$

as $\{\omega \in \Omega : X(\omega) \in B\} \in \Omega$ and $\mathbb{P}$ is a probability on Ω. (A2) Let $B = X(\Omega)$,

$$\mathbb{P}_X\Big(X(\Omega)\Big) = \mathbb{P}\Big(X \in X(\Omega)\Big) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in X(\Omega)\}) = \mathbb{P}(\Omega) = 1$$

as $\mathbb{P}$ is a probability on Ω. (A3) Let $B_1, B_2, \cdots$ be pairwise disjoint events,

$$\mathbb{P}_X\Big(\bigcup_{i=1}^{\infty} B_i\Big) = \mathbb{P}\Big(X \in \bigcup_{i=1}^{\infty} B_i\Big) = \mathbb{P}\Big(\bigcup_{i=1}^{\infty} \{X \in B_i\}\Big).$$

As the events are pairwise disjoint we have $\{X \in B_i\} \cap \{X \cap B_j\} = \emptyset$ for $i \neq j$, thus

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} \{X \in B_i\}\Big) = \sum_{i=1}^{\infty} \mathbb{P}(X \in B_i) = \sum_{i=1}^{\infty} \mathbb{P}_X(B_i)$$

due to $\mathbb{P}$ being a probability on Ω. □

4.2 Discrete random variables

Definition 4.2 (Discrete random variable)**.** A random variable $X : \Omega \to X(\Omega)$ is said to be discrete when there is a countable[2] set $\mathcal{X} \subseteq X(\Omega)$ such that $\mathbb{P}_X(\mathcal{X}) = \mathbb{P}(X \in \mathcal{X}) = 1$.

Definition 4.3 (Probability mass function)**.** The probability mass function of a discrete random variable is the function $p : \mathcal{X} \to [0,1]$ given by

$$p(x) = \mathbb{P}(X = x)$$

for all $x \in \mathcal{X}$.

Theorem 4.2. Suppose that the discrete random variable X has the probability mass function $p : \mathcal{X} \to [0,1]$, then

$$\mathbb{P}(X \in B) = \sum_{x \in B} p(x)$$

for all $B \subseteq \mathcal{X}$. In addition,

$$\sum_{x \in \mathcal{X}} p(x) = 1.$$

[2] a set is countable if there exists a bijection between that set and a subset of the natural numbers $\mathbb{N}$.

Proof. Let $B \subseteq \mathcal{X}$,

$$\mathbb{P}(X \in B) = \mathbb{P}\Big(\bigcup_{x \in B} \{X = x\} \Big) = \sum_{x \in B} \mathbb{P}(X = x) = \sum_{x \in B} p(x)$$

using A3. Letting $B = \mathcal{X}$,

$$\mathbb{P}(X \in \mathcal{X}) = \sum_{x \in \mathcal{X}} p(x) = \mathbb{P}(\Omega) = 1.$$

□

4.3 Discrete distributions

We will look at four different types of discrete random variables, each uniquely defined by their probability mass function. These distributions can be used in different scenarios which will be briefly outlined.

Suppose we have an experiment where we have:

- Two distinct outcomes: success and failure
- The probability of success is p and the probability of failure is $1 - p$

We can then define the random variable X to be the number of successes; this random variable is called a Bernoulli random variable.

Definition 4.4 (Bernoulli distribution)**.** A discrete random variable X is a Bernoulli random variable with parameter $p \in [0, 1]$ when $\mathcal{X} = \{0, 1\}$ and

$$p(x) = p^x (1 - p)^{1-x} \quad \text{for all } x \in \{0, 1\}.$$

We write $X \sim \text{Ber}(p)$.

Suppose we have an experiment where we have:

- A fixed number n of trials
- Each trial is independent of the others
- Each trial has two distinct outcomes: success and failure
- The probability of success is p and the probability of failure is $1 - p$

We can then define the random variable X to be the number of successes; this random variable is called a binomial random variable.

Definition 4.5 (Binomial distribution)**.** A discrete random variable X is a binomially distributed with parameter $p \in [0, 1]$ when $\mathcal{X} = \{0, 1, 2, \cdots, n\}$ and

$$p(x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad \text{for all } x \in \{0, 1, 2, \cdots, n\}.$$

We write $X \sim \text{Bin}(n, p)$.

Suppose we have an experiment where we have:

- An unlimited number of trials
- Each trial is independent of the others
- Each trial has two distinct outcomes: success and failure
- The probability of success is p and the probability of failure is $1 - p$
- We repeat each trial until a success is obtained

We can then define the random variable X to be the number of trials until a success; this random variable is called a geometric random variable.

Definition 4.6 (Geometric distribution)**.** A discrete random variable X is geometrically distributed with parameter $p \in (0, 1]$ when $\mathcal{X} = \mathbb{N} = \{1, 2, 3, \cdots\}$ and

$$p(x) = (1 - p)^{x-1} p \quad \text{for all } x \in \mathbb{N}.$$

We write $X \sim \text{Geo}(p)$.

Suppose we have an experiment where we wish to count the number of events that occur in a given time, where:

- There is an infinite number of possible events.
- Each event occurs independently of the others
- Events happen at a constant average rate λ
- Each trial has two distinct outcomes: success and failure

We can then define the random variable X to be the number of events that occur in a given time; this random variable is called a Poisson random variable.

Definition 4.7 (Poisson distribution)**.** A discrete random variable X is Poisson distributed with parameter λ when $\mathcal{X} = \{0, 1, 2, \cdots\}$ and

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!} \quad \text{for all } x \in \{0, 1, 2, \cdots\}.$$

We write $X \sim \text{Po}(\lambda)$.

4.4 Continuous random variables

Definition 4.8 (Continuous random variable & Probability density function)**.** A real-valued random variable $X : \Omega \to \mathbb{R}$ is said to be Continuous if there is a non-negative function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(t) \, \mathrm{d}t$$

for all $[a, b] \subseteq \mathbb{R}$. Such a function f is called the probability density function of X.

Often times, the function f will be piecewise defined on $\mathbb{R}$. Also note from the definition that $\mathbb{P}(X = a) = \int_a^a f(t)\,\mathrm{d}t = 0$ for any $a \in \mathbb{R}$; hence $\mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in (a, b))$.

Theorem 4.3. If X is a continuous random variable with probability density function f, then for any $B \subseteq \mathbb{R}$ that is a finite union of intervals:

$$\mathbb{P}(X \in B) = \int_B f(t)\,\mathrm{d}t.$$

Proof. As B is a finite union of intervals,

$$B = \bigcup_i U_i$$

for some intervals U_i with $U_i \cap U_j = \emptyset$ for $i \neq j$.[3] From the definition,

$$\mathbb{P}(X \in U_i) = \int_{U_i} f(t)\,\mathrm{d}t.$$

Using A3,

$$\mathbb{P}(X \in B) = \mathbb{P}\Big(X \in \bigcup_i U_i\Big) = \mathbb{P}\Big(\bigcup_i \{X \in U_i\}\Big) = \sum_i \mathbb{P}(X \in U_i) = \sum_i \int_{U_i} f(t)\,\mathrm{d}t$$

as the intervals U_i are pairwise disjoint. Hence, by properties of integration

$$\mathbb{P}(X \in B) = \sum_i \int_{U_i} f(t)\,\mathrm{d}t = \int_{\bigcup_i U_i} f(t)\,\mathrm{d}t = \int_B f(t)\,\mathrm{d}t.$$

□

Theorem 4.4. Let X be a continuous random variable with probability density function f, then

$$\mathbb{P}(X \in [a, \infty)) = \int_a^\infty f(t)\,\mathrm{d}t \quad \text{and} \quad \mathbb{P}(X \in (-\infty, b]) = \int_{-\infty}^b f(t)\,\mathrm{d}t$$

Proof. Firstly, notice that

$$\mathbb{P}(X \in [a, \infty)) = \mathbb{P}\Big(\bigcup_{n=1}^\infty \{X \in [a, a + n)\}\Big)$$

and using Theorem 1.9,

$$\mathbb{P}\Big(\bigcup_{n=1}^\infty \{X \in [a, a + n)\}\Big) = \lim_{n\to\infty} \mathbb{P}(X \in [a, a + n))$$

as the events $\{X \in [a, a + n)\}$ form an increasing sequence of events. Hence,

$$\mathbb{P}(X \in [a, \infty)) = \lim_{n\to\infty} \mathbb{P}(X \in [a, a + n)) = \lim_{n\to\infty} \int_a^{a+n} f(t)\,\mathrm{d}t = \int_a^\infty f(t)\,\mathrm{d}t.$$

[3] We can construct such a union as if we chose intervals V_i with $B = \bigcup_i V_i$ and $V_j \cap V_k$ is non-empty for some j, k, then we can replace V_j and V_k with $U_i = V_j \setminus V_k$, $U_j = V_j \cap V_k$ and $U_k = V_k \setminus V_j$.

Similarly,

$$\mathbb{P}(X \in (-\infty, b]) = \mathbb{P}\Big(\bigcup_{n=1}^{\infty} \{X \in (b-n, b]\}\Big)$$

and using Theorem 1.9,

$$\mathbb{P}\Big(\bigcup_{n=1}^{\infty} \{X \in (b-n, b]\}\Big) = \lim_{n\to\infty} \mathbb{P}(X \in (b-n, b])$$

as the events $\{X \in (b-n, b]\}$ form an increasing sequence of events. Hence,

$$\mathbb{P}(X \in (-\infty, b]) = \lim_{n\to\infty} \mathbb{P}(X \in (b-n, b]) = \lim_{n\to\infty} \int_{b-n}^{b} f(t)\,\mathrm{d}t = \int_{-\infty}^{b} f(t)\,\mathrm{d}t.$$

□

Corollary 4.1. Let X be a continuous random variable with probability density function f, then

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1.$$

Proof.

$$\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = \mathbb{P}(X \in (-\infty, \infty)) = 1.$$

□

Definition 4.9 (Cumulative distribution)**.** For any real valued random variable X, the function $F : \mathbb{R} \to [0, 1]$ defined by

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

is called the cumulative distribution function of X.

Immediately from Theorem 1.4,

$$\mathbb{P}(X \in (a, b]) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a).$$

Notice that each cumulative distribution function $F(x)$ definiens a unique random variable X; this means that the cumulative distribution function is unique to the random variable X.

Theorem 4.5. Suppose X is a continuous random variable on $\mathbb{R}$ with probability density function f. Then the cumulative density function F is a continuous function on $\mathbb{R}$ and for all $x \in \mathbb{R}$

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t \quad \text{and} \quad f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}(x) \text{ when } F \text{ is continuous at } x.$$

Proof. Straight from the definition we have that

$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x]) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t.$$

Using the fundamental theorem of calculus we have that F is continuous on $\mathbb{R}$ and the second equality. □

4.5 Continuous distributions

We will look at three different types of continuous random variables, each uniquely defined by their probability density function. These distributions can be used in different scenarios which will be briefly outlined.

Consider an experiment when a parameter of interest is equally likely to be any real number in the interval $[a, b]$ for real a, b with $a < b$.

Definition 4.10 (Uniform distribution)**.** Let a and b with $a < b$. A continuous random variable X is uniformly distributed on $[a, b]$ when

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for all } x \in [a, b] \\ 0 & \text{elsewhere.} \end{cases}$$

We write $X \sim \mathrm{U}[a, b]$.

Definition 4.11 (Exponential distribution)**.** Let $\beta > 0$. A continuous random variable X is exponentially distributed with parameter β when

$$f(x) = \begin{cases} \beta e^{-\beta x} & \text{for all } x \geq 0, \\ 0 & \text{elsewhere.} \end{cases}$$

We write $X \sim \mathrm{Exp}(\beta)$.

Definition 4.12 (Normal distribution)**.** Let μ and σ be real numbers with $\sigma > 0$. A continuous random variable is normally distributed with parameters μ and σ^2 when

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big\{ -\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2 \Big\} \quad \text{for all } x \in \mathbb{R}.$$

We write $X \sim \mathrm{N}(\mu, \sigma^2)$.

The above distribution describes a family of distributions, all of which are normal distributions. We also have a special type of normal distribution called the standard normal distribution.

Definition 4.13 (Standard normal distribution)**.** A continuous random variable Z is standard normally distributed when $Z \sim \mathrm{N}(0, 1)$.

Following from the definition we have

$$\phi(z) := f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{and} \quad \Phi(z) := \mathrm{F}(z) = \int_{-\infty}^{z} \phi(t)\,\mathrm{d}t.$$

In addition, as $\phi(z) = \phi(-z)$ we have $\Phi(z) = 1 - \Phi(-z)$.

4.6 Functions of random variables

Suppose we have a random variable $X : \Omega \to X(\Omega)$ and a function $g : X(\Omega) \to \mathcal{G}$. Then $g(X)$ is also a random variable, namely the outcome to a 'new experiment' obtained by running the 'old experiment' to produce a value x for X, and then evaluating $g(x)$.

Definition 4.14 (Function of a random variable)**.** Let $X : \Omega \to X(\Omega)$ be a random variable and $g : X(\Omega) \to \mathcal{G}$ be a function. The function $g(X) : \Omega \to \mathcal{G}$ is a random variable defined by $g(X) = g \circ X$.

From the definition,

$$\mathbb{P}(g(X) \in B) = \mathbb{P}(\{\omega \in \Omega : g(X(\omega)) \in B\})$$

for all $B \subseteq \mathcal{G}$.

There is one particularly important function which enables us to get the cumulative distribution function of any normally distributed random variable, using just the the standard normal tables.

Theorem 4.6 (Standardizing the normal distribution)**.** Suppose we have random variables $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Z \sim \mathrm{N}(0, 1)$, then

$$\frac{X - \mu}{\sigma} \sim \mathrm{N}(0, 1) \quad \text{and} \quad \sigma Z + \mu \sim \mathrm{N}(\mu, \sigma^2).$$

Proof. We will show that $\mathrm{F}_{(X-\mu)/\sigma}(z) = \mathrm{F}_Z(z)$ and that $\mathrm{F}_{\sigma Z+\mu}(x) = \mathrm{F}_X(x)$, then use the fact that the cumulative distribution function defines a unique random variable X.
For any $z \in \mathbb{R}$,

$$\mathrm{F}_{(X-\mu)/\sigma}(z) = \mathbb{P}((X - \mu)/\sigma \leq z) = \mathbb{P}(X \leq \sigma z + \mu) = \mathrm{F}_X(\sigma z + \mu)$$

where $\mathrm{F}_X(x)$ is the cumulative distribution function for X. Hence,

$$\mathrm{F}_X(\sigma z + \mu) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\sigma z+\mu} e^{-(x-\mu)^2/2\sigma^2} \, \mathrm{d}x$$

and with $t = \frac{x-\mu}{\sigma}$,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\sigma z+\mu} e^{-(x-\mu)^2/2\sigma^2} \, \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} \, \mathrm{d}t = \mathrm{F}_Z(z).$$

Similarly, for any $x \in \mathbb{R}$,

$$\mathrm{F}_{\sigma Z+\mu}(x) = \mathbb{P}(\sigma Z + \mu \leq x) = \mathbb{P}(Z \leq (x - \mu)/\sigma) = \mathrm{F}_Z\Big(\frac{x - \mu}{\sigma}\Big)$$

where $\mathrm{F}_Z(z)$ is the cumulative distribution function for Z. Hence,

$$\mathrm{F}_Z\Big(\frac{x - \mu}{\sigma}\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2} \, \mathrm{d}t$$

and with $x = \sigma t + \mu$,

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-t^2/2} \, \mathrm{d}t = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(x-\mu)^2/2\sigma^2} \, \mathrm{d}x = \mathrm{F}_X(x).$$

□

5 Multiple random variables

5.1 Bivariate random variables

Definition 5.1 (Bivariate random variable)**.** Let X and Y be two random variables on the sample space Ω given by $X : \Omega \to X(\Omega)$ and $Y : \Omega \to Y(\Omega)$. The mapping $(X, Y) : \Omega \to (X, Y)(\Omega)$ defined by

$$(X, Y)(\omega) := (X(\omega), Y(\omega))$$

is then a bivariate random variable on Ω.

Just as we discussed that a random variable X is a mapping from the sample space Ω to the image $X(\Omega)$, so too is the bivariate random variable (X, Y); this is a mapping from the sample space Ω to the image $(X, Y)(\Omega)$.

We can also express the image of a random variable X as $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$ and the image of a bivariate random variable (X, Y) as $(X, Y)(\Omega) = \{(X(\omega), Y(\omega)) : \omega \in \Omega\}$.

Example 5.1. Continuing from Example 2.1, rolling a fair die and noting the score has the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$. We will define two random variables X and Y as

$$X(\omega) = 2\omega \quad \text{and} \quad Y(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is even} \\ 1 & \text{if } \omega \text{ is odd} \end{cases}$$

for all $\omega \in \Omega$. We see that

ω	1	2	3	4	5	6
$X(\omega)$	2	4	6	8	10	12
$Y(\omega)$	1	0	1	0	1	0

and thus the bivariate random variable (X, Y) is

ω	1	2	3	4	5	6
$(X, Y)(\omega)$	$(2, 1)$	$(4, 0)$	$(6, 1)$	$(8, 0)$	$(10, 1)$	$(12, 0)$

For any $x \in X(\Omega)$ and $y \in Y(\Omega)$, we write $\{X = x, Y = y\}$ to mean

$$\{X = x, Y = y\} := \{(X, Y) \in (x, y)\} = \{\omega \in \Omega : X(\omega) = x \text{ and } Y(\omega) = y\}.$$

Definition 5.2 (Independence of two random variables)**.** Two random variables X and Y on the same sample space Ω are independent if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

for all $A \subseteq X(\Omega)$ and $B \subseteq Y(\Omega)$.

5.2 Jointly distributed discrete random variables

Definition 5.3 (Discrete bivariate random variable)**.** A bivariate random variable $(X, Y) : \Omega \to (X, Y)(\omega)$ is said to be discrete if there is a countable set $\mathcal{Z} \subseteq (X, Y)(\Omega)$ such that $\mathbb{P}((X, Y) \in \mathcal{Z}) = 1$.

Definition 5.4 (Joint probability mass function)**.** The joint probability mass function of a discrete random variable is the function $p : \mathcal{Z} \to 1$ given by

$$p(x, y) := \mathbb{P}(X = x, Y = y)$$

for all $(x, y) \subseteq \mathcal{Z}$.

Definition 5.5 (Marginal probability mass function)**.** Suppose we have a bivariate random variable (X, Y) on the sample space Ω with a probability mass function $p(x, y)$.

We define the marginal probability mass function of X as the probability mass function of X, also written as $p_X(x)$.

Similarly, the marginal probability mass function of Y is the probability mass function of Y, also written as $p_Y(y)$.

Theorem 5.1. Let X and Y be two random variables on the same sample space Ω. Then the bivariate random variable (X, Y) is discrete if and only if both X and Y are discrete.

Proof. Suppose that X and Y are discrete. Then there exists countable sets $\mathcal{X}$ and $\mathcal{Y}$ such that

$$\mathbb{P}(X \in \mathcal{X}) = \mathbb{P}(Y \in \mathcal{Y}) = 1.$$

Taking $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \times \mathcal{Y}$ denotes the Cartesian product of the sets $\mathcal{X}$ and $\mathcal{Y}$ means that the bivariate random variable (X, Y) is discrete. This is as $\mathcal{X} \times \mathcal{Y}$ will be countable as $\mathcal{X}$ and $\mathcal{Y}$ are countable.
Suppose that (X, Y) is discrete. Then there exists a countable set $\mathcal{Z} \subseteq (X, Y)(\Omega)$ such that

$$\mathbb{P}((X, Y) \in \mathcal{Z}) = 1.$$

Taking $\mathcal{X} = \{x : (x, y) \in \mathcal{Z}\}$ we see that $\mathcal{X}$ is countable with

$$\mathbb{P}(X \in \mathcal{X}) = \mathbb{P}((X, Y) \in \mathcal{Z}) = 1$$

so X is discrete. Similar argument for Y. □

Theorem 5.2. If two random variables X and Y, defined on the same sample space Ω, are discrete with $\mathbb{P}(X \in \mathcal{X}) = \mathbb{P}(Y \in \mathcal{Y}) = 1$, for countable $\mathcal{X}, \mathcal{Y}$, then their marginal probability mass functions are given in terms of the joint probability mass function by

$$p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y) \quad \text{and} \quad p_Y(y) = \sum_{x \in \mathcal{X}} p(x, y)$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ respectively.

Proof. Using the definition of $\mathcal{X}$ as $\mathcal{X} = \{x : (x, y) \in \mathcal{Z}\}$, the fact that $\mathbb{P}(X \in \mathcal{X}) = 1$ and A3,

$$\mathbb{P}(Y = y) = \mathbb{P}(X \in \mathcal{X}, Y = y) = \mathbb{P}\Big(\bigcup_{x \in \mathcal{X}} \{X = x, Y = y\} \Big) = \sum_{x \in \mathcal{X}} \mathbb{P}(X = x, Y = y) = \sum_{x \in \mathcal{X}} p(x, y)$$

as the events $\{X = x_i, Y = y\}$ and $\{X = x_j, Y = y\}$ are disjoint for all $i \neq j$. Similar argument for $\mathbb{P}(X = x) = p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y)$. □

Example 5.2. Consider rolling two fair six-sided dice. Let X be the number of 6's rolled and Y be the number of 1's and 2's. We find that:

$p(x,y)$	$x=0$	$x=1$	$x=2$	$p_Y(y)$
$y=0$	9/36	6/36	1/36	16/36
$y=1$	12/36	4/36	0	16/36
$y=2$	4/36	0	0	4/36
$p_X(x)$	25/36	10/36	1/36	1

We see that if we add the rows or columns we arrive at the marginal probability mass functions $p_Y(y)$ and $p_X(x)$ respectively. In addition, the bottom right entry in our table is 1; this is a good check to see if you have calculated the probabilities correctly!

Corollary 5.1. Let X and Y be discrete random variables with $\mathbb{P}((X,Y) \in \mathcal{Z}) = 1$ for some countable $\mathcal{Z} \subseteq (X,Y)(\Omega)$. Then we have

$$\mathbb{P}((X,Y) \in A) = \sum_{(x,y)\in A} p(x,y)$$

for all $(x,y) \subseteq \mathcal{Z}$ and

$$\sum_{(x,y)\in\mathcal{Z}} p(x,y) = 1.$$

Proof. This is a direct application of Theorem 4.2 to the case of a discrete random variable (X,Y) whose possible values are ordered pairs (x,y). □

Definition 5.6 (Conditional probability mass function)**.** Let X and Y be jointly distributed discrete random variables. For $y \in \mathcal{Y}$ with $p_Y(y) > 0$, the conditional probability mass function of X given $Y = y$ is defined by

$$p_{X|Y}(x|y) := \mathbb{P}(X = x|Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

where $p_{X,Y}(x,y)$ denotes the joint probability mass function of X and Y. Similarly, the conditional probability mass function of Y given $X = x$ is defined by

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

for $x \in \mathcal{X}$ with $p_X(x) > 0$.

Example 5.3. Continuing from Example 5.2, we find that

$p_{X\|Y}(x,y)$	$x=0$	$x=1$	$x=2$
$y=0$	9/16	6/16	1/16
$y=1$	12/16	4/16	0
$y=2$	1	0	0

using our definition for $p_{X|Y}(x,y)$.

Theorem 5.3. Let X and Y be discrete random variables, then

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x|y) p_Y(y).$$

Proof. From Theorem 5.2,

$$p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

and using the definition of the conditional probability mass function,

$$p_X(x) = \sum_{y \in \mathcal{Y}} p(x, y) = \sum_{y \in \mathcal{Y}} p_{X|Y}(x, y) p_Y(y).$$

□

Theorem 5.4 (Independence of discrete random variables)**.** Two discrete random variables X and Y on the same sample space Ω are independent if and only if

$$p_{X,Y}(x, y) = p_X(x) p_Y(y)$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Proof. Suppose that X and Y are independent, from Definition 5.2,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \Leftrightarrow p_{X,Y}(x, y) = p_X(x) p_Y(y).$$

Suppose that $p_{X,Y}(x, y) = p_X(x) p_Y(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, from Corollary 5.1 we have that

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}((X, Y) \in A \times B) = \sum_{(x,y) \in A \times B} p(x, y)$$

for all sets $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$. We see that

$$\sum_{(x,y) \in A \times B} p(x, y) = \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$$

thus X and Y are independent. □

Example 5.4. Continuing from Example 5.2, notice that X and Y are *not* independent as

$$p(0, 0) = 9/36 \neq p_X(0) p_Y(0) = 25/36 \ \times \ 16/36.$$

When showing that two random variables are not independent you only have to find one example when $p(x, y) \neq p_X(x) p_Y(y)$ and when you want to show that they are independent you have to show $p(x, y) = p_X(x) p_Y(y)$ *for all* $x \in X(\Omega)$ and $y \in Y(\Omega)$.

5.3 Jointly continuously distributed random variables

Definition 5.7 (Joint probability density function)**.** Let $X : \Omega \to X(\Omega)$ and $Y : \Omega \to Y(\Omega)$ be two continuos random variables on the same sample space. Both random variables are said to be jointly continuously distributed if there is a non-negative function $f : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{P}(X \in [a,b], Y \in [c,d]) = \int_a^b \Big(\int_c^d f(x,y)\,\mathrm{d}y \Big)\,\mathrm{d}x$$

for all $[a,b] \times [c,d] \subseteq \mathbb{R}^2$. We call the function f the joint probability density function of X and Y.

Definition 5.8 (Marginal probability density function)**.** Let $X : \Omega \to X(\Omega)$ and $Y : \Omega \to Y(\Omega)$ be two continuos random variables with joint probability density function $f_{X,Y}(x,y)$.

We define the marginal probability density function of X as the probability mass function of X, also written as $f_X(x)$.

Similarly, the marginal probability mass function of Y is the probability mass function of Y , also written as $f_Y(y)$.

Recall from a calculus class that our iterated integral can be conceived as a double integral

$$\int_a^b \Big(\int_c^d f(x,y)\,\mathrm{d}y \Big)\,\mathrm{d}x = \iint\limits_{[a,b]\times[c,d]} f(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

over the region $[a,b] \times [c,d]$. We can thus generalise to the following.

Definition 5.9. If X and Y are jointly continuously distributed then for any nice[4] region $A \subseteq \mathbb{R}^2$,

$$\mathbb{P}((X,Y) \in A) = \iint\limits_A f(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

Corollary 5.2. If X and Y are jointly continuously distributed then

$$\iint\limits_{\mathbb{R}^2} f(x,y)\,\mathrm{d}y\,\mathrm{d}x = 1$$

Proof. Taking $A = \mathbb{R}^2$ in our definition,

$$\mathbb{P}((X,Y) \in \mathbb{R}^2) = \iint\limits_{\mathbb{R}^2} f(x,y)\,\mathrm{d}y\,\mathrm{d}x = 1$$

□

Example 5.5. Consider random variables X and Y with joint probability density function

$$f(x,y) = \begin{cases} x+y & \text{if } (x,y) \in [0,1]^2 \\ 0 & \text{otherwise} \end{cases}$$

[4]This region A in fact has to be a finite union of x or y simple regions, see a calculus course for more detail but for our uses this will be sufficient

To find $\mathbb{P}(1/4 < X < 3/4, 0 < Y < 1/2)$,

$$\mathbb{P}(1/4 < X < 3/4, 0 < Y < 1/2) = \int_{1/4}^{3/4} \int_{0}^{1/2} x + y \,\mathrm{d}y \,\mathrm{d}x = \int_{1/4}^{3/4} \frac{1}{2}x + \frac{1}{8} \,\mathrm{d}x = \frac{3}{16}.$$

If we needed $\mathbb{P}(X^2 < Y < X)$,

$$\mathbb{P}(X^2 < Y < X) = \int_{0}^{1} \int_{x^2}^{x} x + y \,\mathrm{d}y \,\mathrm{d}x = \int_{0}^{1} \frac{3}{2}x^2 - x^3 - \frac{1}{4}x^4 \,\mathrm{d}x = \frac{3}{20}.$$

Theorem 5.5. Let X and Y be jointly continuously distributed random variables. Then both X and Y are continuously distributed with

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \,\mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \,\mathrm{d}x.$$

Proof. Suppose that X and Y are jointly continuously distributed random variables. Then there exists a non-negative function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{P}(X \in [a,b], Y \in [c,d]) = \int_a^b \Big(\int_c^d f_{X,Y}(x,y) \,\mathrm{d}y \Big) \,\mathrm{d}x$$

for all $[a,b] \times [c,d] \in \mathbb{R}^2$. Using $\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \in [a,b], Y \in (-\infty, \infty))$,

$$\mathbb{P}(X \in [a,b]) = \mathbb{P}(X \in [a,b], Y \in (-\infty,\infty)) = \int_a^b \int_{-\infty}^{\infty} f_{X,Y}(x,y) \,\mathrm{d}y \,\mathrm{d}x$$

which gives us a function $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \,\mathrm{d}y$ such that

$$\mathbb{P}(X \in [a,b]) = \int_a^b f_X(x) \,\mathrm{d}x$$

thus X is continuously distributed. Similar argument for Y. □

Definition 5.10 (Conditional probability density function)**.** Let X and Y be jointly continuously distributed random variables. For all $x, y \in \mathbb{R}$ such that $f_Y(y) > 0$, the conditional probability density function of X at $Y = y$ is defined by

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

where $f_{X,Y}(x,y)$ is the joint probability density function of X and Y. Similarly, the conditional probability mass function of Y at $X = x$ is defined by

$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x,y)}{f_X(x)}$$

for all $x, y \in \mathbb{R}$ such that $f_X(x) > 0$.

Theorem 5.6. Let X and Y be jointly continuously distributed random variables, then

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\,\mathrm{d}y.$$

Proof. From Theorem 5.5,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y$$

and using the definition of the conditional probability density function,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\,\mathrm{d}y.$$

□

Theorem 5.7. Two jointly continuously distributed random variables X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

for all $x, y \in \mathbb{R}$.

Proof. Suppose that X and Y are independent. Then from Definition 5.2,

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \Leftrightarrow f_{X,Y}(x,y) = f_X(x) f_Y(y).$$

Suppose that $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ for all $x, y \in \mathbb{R}$. Then from Definition 5.9,

$$\mathbb{P}((X,Y) \in A) = \iint_A f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \iint_A f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y$$

and choosing $A = A_1 \times A_2$[5],

$$\mathbb{P}((X,Y) \in A) = \mathbb{P}(X \in A_1, Y \in A_2) = \int_{A_1} f_X(x)\,\mathrm{d}x \int_{A_2} f_Y(y)\,\mathrm{d}y = \mathbb{P}(X \in A_1)\mathbb{P}(Y \in A_2).$$

□

Example 5.6. Continuing from Example 5.5, we see that

$$f_X(x) = \int_0^1 x + y\,\mathrm{d}y = x + \frac{1}{2}$$

and also that

$$f_Y(y) = \int_0^1 x + y\,\mathrm{d}x = y + \frac{1}{2}.$$

Hence we have that X and Y are not independent as

$$f_{X,Y}(x,y) = x + y \neq f_X(x) f_Y(y) = \left(x + \frac{1}{2}\right)\left(y + \frac{1}{2}\right)$$

[5]Which we can do as A is a nice region

5.4 Functions of multiple random variables

Similar to the single variable case, suppose we have random variables $X : \Omega \to X(\Omega)$ and $Y : \Omega \to Y(\Omega)$ with a function $g : X(\Omega) \times Y(\Omega) \to \mathcal{G}$. Then $g(X, Y)$ is also a random variable, namely the outcome to a 'new experiment' obtained by running the 'old experiment' to produce a value (x, y) for (X, Y), and then evaluating $g(x, y)$.

Definition 5.11 (Function of a bivariate random variable)**.** Let $X : \Omega \to X(\Omega)$ and $Y : \Omega \to Y(\Omega)$ be random variables with a function $g : X(\Omega) \times Y(\Omega) \to \mathcal{G}$. The function $g(X, Y) : \Omega \to \mathcal{G}$ is a random variable defined by $g(X, Y) = g \circ (X, Y)$.

From the definition,

$$\mathbb{P}(g(X, Y) \in B) = \mathbb{P}(\{\omega \in \Omega : g(X(\omega), Y(\omega)) \in B\})$$

for all $B \subseteq \mathcal{G}$.

Example 5.7. Continuing from Example 5.5, we will define a new random variable $G = X + Y$.

Notice that $\mathbb{P}(0 \leq G \leq 2) = 1$ so our cumulative density function for G will look a bit odd. Let us find $\mathrm{F}_G(g)$ by considering if $0 \leq g \leq 1$ first:

$$\mathrm{F}_G(g) = \mathbb{P}(X + Y \leq g) = \int_0^g \Big(\int_0^{g-y} x + y \, \mathrm{d}x \Big) \, \mathrm{d}y$$

as $\mathbb{P}(X + Y \leq g) = \mathbb{P}(0 \leq Y \leq 1, X \leq g - Y)$. Hence,

$$\mathrm{F}_G(g) = \frac{1}{2} \int_0^g (g^2 - y^2) \, \mathrm{d}y = \frac{1}{3} g^3 \quad \text{for } 0 \leq g \leq 1.$$

Let $1 \leq g \leq 2$:

$$\mathrm{F}_G(g) = \mathbb{P}(X + Y \leq g) = \int_0^{g-1} \Big(\int_0^1 x + y \, \mathrm{d}x \Big) \, \mathrm{d}y + \int_{g-1}^1 \Big(\int_0^{g-y} x - y \, \mathrm{d}x \Big) \, \mathrm{d}y$$

as $\mathbb{P}(X + Y \leq g) = \mathbb{P}\Big(\{0 \leq Y \leq g - 1 \ , \ 0 \leq X \leq 1\} \cup \{g - 1 \leq Y \leq 1 \ , \ 0 \leq X \leq g - Y\} \Big)$. Thus,

$$\mathrm{F}_G(g) = \mathbb{P}(X + Y \leq g) = \int_0^{g-1} y + \frac{1}{2} \, \mathrm{d}y + \frac{1}{2} \int_{g-1}^1 g^2 - y^2 \, \mathrm{d}y = -\frac{1}{3} + g^2 - \frac{1}{3} g^3$$

for $1 \leq g \leq 2$. Hence we have that

$$\mathrm{F}_G(g) = \begin{cases} \frac{1}{3} + g^2 - \frac{1}{3} g^3 & \text{for } 1 \leq g \leq 2 \\ \frac{1}{3} g^3 & \text{for } 0 \leq g \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and G is therefore a continuous random variable.

When tackling multiple random variable problems (especially continuous random variables) you will see that the questions are tricky often because it is simply a multivariable calculus problem in disguise.

A good thing to remember is when trying to show that something is a random variable or a specific random variable (such as a uniform random variable) it is useful to show it in terms of the cumulative (mass or density) function: if you want to show that the random variable U is continuously distributed and in particular a uniform random variable, show that the cumulative density function is the corresponding function for the uniform distribution!

Note: Everything we have covered for the bivariate case can be extended to finitely many random variables.

6 Expectation

6.1 Definition

Definition 6.1 (Expectation of a discrete random variable)**.** Let X be a discrete random variable taking values in $\mathcal{X}$. The expectation of X, denoted as $\mathbb{E}[X]$, is defined as

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} xp(x)$$

provided that the sum converges absolutely[6].

Definition 6.2 (Expectation of a continuous random variable)**.** Let X be a continuous random variable. The expectation of X, denoted as $\mathbb{E}[X]$, is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$$

provided that the integral converges absolutely[7].

The expectation $\mathbb{E}[X]$ can be thought of as a weighted average of X - taking the possible values of X and multiplying them by the possibility of those values occurring.

6.2 Expectation of functions of random variables

Theorem 6.1. Let X be a discrete random variable taking values in $\mathcal{X}$. For any function $g : \mathcal{X} \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

provided that the sum converges absolutely. Similarly, if X is a continuous valued random variable then for any function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$$

provided that the integral converges absolutely.

[6] If you have not taken any analysis courses (or have forgotten) this means that $\sum_{x \in \mathcal{X}} |xp(x)| < \infty$

[7] Similar to the above, this means $\int_{-\infty}^{\infty} |xf(x)|\,\mathrm{d}x < \infty$

Proof. Suppose that X is a discrete random variable taking values in $\mathcal{X}$ and $g : \mathcal{X} \to \mathbb{R}$ is a function. From Definition 4.14 we know that $g(X)$ is also a discrete random variable and hence should have an expectation. We see that

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{G}} x p_{g(X)}(x)$$

where $\mathcal{G} = g(\mathcal{X})$ (the image of $\mathcal{X}$ under g) and $p_{g(X)}(x)$ is the probability mass function of the random variable $g(X)$.
Now looking at $p_{g(X)}(y)$,

$$p_{g(X)}(y) = \mathbb{P}(g(X) = y).$$

We start with a outcome ω, that outcome is mapped by X to the value $X(\omega)$. The value $X(\omega)$ is then renamed as x (i.e $X(\omega) = x$) and is mapped by $g(X)$ to the value $g(x) = g(X(\omega))$. Finally, we rename the value $g(x)$ as y (i.e $g(x) = y$). Thus we have two random variables $X : \Omega \to X(\Omega)$ and $g(X) : \Omega \to g(\mathcal{X})$ who share the same sample space.

Using Theorem 5.3,

$$\mathbb{P}(g(X) = y) = \sum_{x \in \mathcal{X}} \mathbb{P}(g(X) = y \mid X = x)\mathbb{P}(X = x) = \sum_{x \in \mathcal{X} : g(x) = y} p(x)$$

as if $X = x$ and $g(x) = y$ then $g(X) = y$ is fulfilled whist if $X = x$ and $g(x) \neq y$ then $g(X) = y$ is not fulfilled; this leads to the set $\{x \in \mathcal{X} : g(x) = y\}$ being the one which we sum over as for these values $\mathbb{P}(g(X) = y \mid X = x) = 1$.

It follows that

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{G}} x p_{g(X)}(x) = \sum_{y \in \mathcal{G}} y p_{g(X)}(y) = \sum_{y \in \mathcal{G}} y \Big(\sum_{x \in \mathcal{X} : g(x) = y} p(x) \Big) = \sum_{y \in \mathcal{G}} \Big(\sum_{x \in \mathcal{X} : g(x) = y} y p(x) \Big)$$

as the inner sum is independent of y.

Moreover,

$$\sum_{y \in \mathcal{G}} \Big(\sum_{x \in \mathcal{X} : g(x) = y} y p(x) \Big) = \sum_{x \in \mathcal{X}} \Big(\sum_{y \in \mathcal{G} : y = g(x)} y \Big) p(x)$$

which is due to the following.

The left hand-side reads: sum all $x \in \mathcal{X}$ such that $g(x) = y$, then sum all $y \in \mathcal{G}$. Which is the same as: sum all $y \in \mathcal{G}$ such that $y = g(x)$, then sum all $x \in \mathcal{X}$. Doing as such changes the sums into the right hand-side. In addition, the inner sum is independent of x so we can sift the $p(x)$ outside the inner sum.

Finally,

$$\sum_{y \in \mathcal{G} : y = g(x)} y = g(x)$$

as $y \in \mathcal{G} : y = g(x)$ is always satisfied so we only have a single term of $y = g(x)$.

Hence,

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) p(x).$$

The continuous case follows similarly from the discrete case. □

Theorem 6.2. Let X and Y be jointly distributed discrete random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$. For any function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$,

$$\mathbb{E}[g(X,Y)] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} g(x,y)p(x,y)$$

provided that the sum converges absolutely. Similarly, if X and Y be jointly distributed continuous random variables, then for any function $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y)f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

provided that the integral converges absolutely.

Proof. Proof follows similarly from Theorem 6.1. □

6.3 Linearity of expectation

Theorem 6.3. If X is a random variable with $\alpha, \beta \in \mathbb{R}$, then

$$\mathbb{E}[\alpha X + \beta] = \alpha\mathbb{E}[X] + \beta.$$

Proof. Suppose that X is a discrete random variable. Using Theorem 6.1,

$$\mathbb{E}[\alpha X + \beta] = \sum_{x\in\mathcal{X}} (\alpha x + \beta)p(x) = \alpha \sum_{x\in\mathcal{X}} \Big\{ xp(x) \Big\} + \beta \sum_{x\in\mathcal{X}} p(x) = \alpha\mathbb{E}[X] + \beta$$

by linearity of the summation, Definition 6.1 and Theorem 4.2.

Suppose that X is a continuous random variable. Using Theorem 6.1,

$$\mathbb{E}[\alpha X + \beta] = \int_{-\infty}^{\infty} (\alpha x + \beta)f(x)\,\mathrm{d}x = \alpha \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x + \beta \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \alpha\mathbb{E}[X] + \beta$$

by linearity of the integral, Definition 6.2 and Corollary 4.1. □

Theorem 6.4. If $X_1, X_2, \cdots, X_n$ are random variables on the same sample space Ω, then

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

Before our proof we need the following Lemma.

Lemma 6.1. Let X and Y be two random variables on the same sample space Ω, then

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

Proof. Suppose that X and Y are discrete. As X and Y are random variables on the same sample space Ω we can look at the bivariate random variable (X, Y). Consider the function $g(x, y) = x + y$. By Theorem 6.2,

$$\mathbb{E}[X + Y] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (x + y)p(x,y) = \sum_{x\in\mathcal{X}} x \sum_{y\in\mathcal{Y}} p(x,y) + \sum_{y\in\mathcal{Y}} y \sum_{x\in\mathcal{X}} p(x,y)$$

and using Theorem 5.2 with Definition 6.1,

$$\mathbb{E}[X+Y] = \sum_{x\in\mathcal{X}} x p_X(x) + \sum_{y\in\mathcal{Y}} y p_Y(y) = \mathbb{E}[X] + \mathbb{E}[Y].$$

The continuous case follows in the same way. □

Now we can prove Theorem 6.4.

Proof. Using induction, we have our base case $n = 1$:

$$\mathbb{E}\Big[\sum_{i=1}^{(1)} X_i\Big] = \mathbb{E}[X] = \sum_{i=1}^{(1)} \mathbb{E}[X_i]$$

thus true for $n = 1$.

Suppose

$$\mathbb{E}\Big[\sum_{i=1}^{k} X_i\Big] = \sum_{i=1}^{k} \mathbb{E}[X_i]$$

is true for some $n = k$.

Consider $n = k + 1$:

$$\mathbb{E}\Big[\sum_{i=1}^{k+1} X_i\Big] = \mathbb{E}\Big[\sum_{i=1}^{k} X_i + X_{k+1}\Big]$$

Extending Definition 5.11 to k-many random variables, we arrive at $\sum_{i=1}^{k} X_i$ being its own random variable. Hence we can write $Y = \sum_{i=1}^{k} X_i$ where Y is a random variable on the same sample space Ω as X_{k+1}. Now we are considering

$$\mathbb{E}[Y + X_{k+1}]$$

where Y and X_{k+1} are random variables. By Lemma 6.1 and the induction hypothesis,

$$\mathbb{E}[Y + X_{k+1}] = \mathbb{E}[Y] + \mathbb{E}[X_{k+1}] = \mathbb{E}\Big[\sum_{i=1}^{k} X_i\Big] + \mathbb{E}[X_{k+1}] = \sum_{i=1}^{k} \mathbb{E}[X_i] + \mathbb{E}[X_{k+1}] = \sum_{i=1}^{k+1} \mathbb{E}[X_i],$$

thus true for $n = k + 1$. As true for $n = 1$, $n = k + 1$ and assumed true for $n = k$ then statement is true for all $n \geq 1$. □

6.4 Variance and covariance

Definition 6.3 (Variance of a random variable)**.** Let X be a random variable. The variance of X is defined as

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

Definition 6.4 (Standard deviation of a random variable)**.** Let X be a random variable. The standard deviation of X is defined as

$$\mathrm{SD}[X] := \sqrt{\mathrm{Var}[X]}.$$

Using Theorem 6.1 we also have

$$\mathrm{Var}[X] = \sum_{x\in\mathcal{X}} (x - \mathbb{E}[X])^2 p(x)$$

if X is a discrete random variable,

$$\mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f(x)\,\mathrm{d}x$$

if X is a continuous random variable - provided that the sum and integral converge absolutely.

The variance of a random variable X can be interpreted as the variability of X from the mean $\mathbb{E}[X]$. That is the difference in the value of X and the value of the expectation $\mathbb{E}[X]$.

Definition 6.5 (Covariance of two random variables)**.** Let X and Y be random variables on the same sample space. The covariance of X and Y is defined as

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

The covariance of two random variables is a real number which we use to determine the correlation between X and Y in the following way.

- If $\mathrm{Cov}[X, Y] > 0$ we say that X and Y are positivity correlated
- If $\mathrm{Cov}[X, Y] < 0$ we say that X and Y are negatively correlated
- If $\mathrm{Cov}[X, Y] = 0$ we say that X and Y are uncorrelated

Using Theorem 6.2 we also have

$$\mathrm{Cov}[X, Y] = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (x - \mathbb{E}[X])(y - \mathbb{E}[Y])p(x, y)$$

if X and Y are discrete,

$$\mathrm{Cov}[X, Y] = \iint_{\mathbb{R}^2} (x - \mathbb{E}[X])(y - \mathbb{E}[Y])f(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

if X and Y are continuos - provided that the sum and integral converge absolutely.

In addition,

$$\mathrm{Var}[X] = \mathrm{Cov}[X, X] \quad \text{and} \quad \mathrm{Cov}[X, Y] = \mathrm{Cov}[Y, X]$$

using the definitions, Theorem 6.3 and Theorem 6.4.

Theorem 6.5. Let X be a random variable and $\alpha, \beta \in \mathbb{R}$. Then,

$$\mathrm{Var}[\alpha X + \beta] = \alpha^2 \mathrm{Var}[X].$$

Proof. Observe that,

$$\mathrm{Var}[\alpha X+\beta]=\mathbb{E}\Big[(\alpha X+\beta-\mathbb{E}[\alpha X+\beta])^2\Big]=\mathbb{E}\Big[(\alpha X+\beta-\alpha\mathbb{E}[X]-\beta)^2\Big]=\mathbb{E}\Big[(\alpha X-\alpha\mathbb{E}[X])^2\Big]$$

by Theorem 6.3. Simplifying and applying Theorem 6.3 again,

$$\mathrm{Var}[\alpha X+\beta]=\mathbb{E}\Big[(\alpha X-\alpha\mathbb{E}[X])^2\Big]=\mathbb{E}\Big[\alpha^2(X-\mathbb{E}[X])^2\Big]=\alpha^2\mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big]=\alpha^2\mathrm{Var}[X].$$

□

Theorem 6.6. For random variables X, Y and Z on the same sample space with constants $\alpha,\beta,\gamma,\delta\in\mathbb{R}$ we have the following.

$$\begin{aligned}&\mathrm{Cov}[\alpha X+\beta,\gamma Y+\delta]=\alpha\gamma\mathrm{Cov}[X,Y],\\&\mathrm{Cov}[X+Y,Z]=\mathrm{Cov}[X,Z]+\mathrm{Cov}[Y,Z],\\&\mathrm{Cov}[X,Y+Z]=\mathrm{Cov}[X,Y]+\mathrm{Cov}[X,Z].\end{aligned}$$

This is showing that $\mathrm{Cov}[X,Y]$ is a bilinear form.

Proof. Observe that,

$$\mathrm{Cov}[\alpha X+\beta,\gamma Y+\delta]=\mathbb{E}\Big[(\alpha X+\beta-\mathbb{E}[\alpha X+\beta])(\gamma Y+\delta-\mathbb{E}[\gamma Y+\delta])\Big]$$

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

using Theorem 6.3 twice.

We see that,

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

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

by applying Theorem 6.4 twice.

Similarly,

$$\mathrm{Cov}[X,Y+Z]=\mathrm{Cov}[Y+Z,X]=\mathrm{Cov}[Y,X]+\mathrm{Cov}[Z,X]=\mathrm{Cov}[X,Y]+\mathrm{Cov}[X,Z]$$

by $\mathrm{Cov}[X,Y]=\mathrm{Cov}[Y,X]$ and the previous result. □

Theorem 6.7. For any random variable X we have

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

Proof. From the definition,

$$\mathrm{Var}[X]=\mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big]=\mathbb{E}\Big[X^2-2X\mathbb{E}[X]+(\mathbb{E}[X])^2\Big]$$

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

by Theorem 6.3. □

Theorem 6.8. For any random variables X and Y on the same sample space,

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

Proof. From the definition and Theorem 6.3,

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

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

□

In practice, Theorem 6.8 and Theorem 6.7 are typically the ways in which you calculate the variance and covariance although not always!

Theorem 6.9. For any random variables $X_1, X_2, \cdots, X_n$ on the same sample space,

$$\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{Cov}[X_i, X_j].$$

Before our proof we need the following Lemma.

Lemma 6.2. Let X and Y be two random variables on the same sample space, then

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y].$$

Proof. Observe that,

$$\mathrm{Var}[X+Y] = \mathrm{Cov}[X+Y, X+Y]$$

$$= \mathrm{Cov}[X,X] + \mathrm{Cov}[Y,Y] + 2\mathrm{Cov}[X,Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$$

by applying Theorem 6.6 twice. □

Now we can prove Theorem 6.9.

Proof. Using induction, we have our base case $n = 1$:

$$\mathrm{Var}\Big[\sum_{i=1}^{(1)} X_i\Big] = \mathrm{Var}[X_1] = \sum_{i=1}^{(1)} \mathrm{Var}[X_i] + 2\sum_{i=1}^{(1)-1}\sum_{j=i+1}^{(1)} \mathrm{Cov}[X_i, X_j]$$

using the convention that an empty sum is zero, thus true for $n = 1$.

Suppose

$$\mathrm{Var}\Big[\sum_{i=1}^{k} X_i\Big] = \sum_{i=1}^{k} \mathrm{Var}[X_i] + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \mathrm{Cov}[X_i, X_j]$$

is true for some $n = k$.

Consider $n = k+1$:

$$\mathrm{Var}\Big[\sum_{i=1}^{k+1} X_i\Big] = \mathrm{Var}\Big[\sum_{i=1}^{k} X_i + X_{k+1}\Big]$$

Extending Definition 5.11 to k-many random variables, we arrive at $\sum_{i=1}^{k} X_i$ being its own random variable. Hence we can write $Y = \sum_{i=1}^{k} X_i$ where Y is a random variable on the same sample space Ω as X_{k+1}. Now we are considering

$$\mathrm{Var}[Y + X]$$

where Y and X_{k+1} are random variables. By Lemma 6.2 and the induction hypothesis,

$$\mathrm{Var}[Y+X_{k+1}] = \mathrm{Var}[Y]+\mathrm{Var}[X_{k+1}]+2\mathrm{Cov}[Y, X_{k+1}] = \mathrm{Var}\Big[\sum_{i=1}^{k} X_i\Big]+\mathrm{Var}[X_{k+1}]+2\mathrm{Cov}\Big[\sum_{i=1}^{k} X_i, X_{k+1}\Big]$$

$$= \sum_{i=1}^{k} \mathrm{Var}[X_i] + 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \mathrm{Cov}[X_i, X_j] + \mathrm{Var}[X_{k+1}] + 2\mathrm{Cov}\Big[\sum_{i=1}^{k} X_i, X_{k+1}\Big]$$

$$= \sum_{i=1}^{k+1} \mathrm{Var}[X_i] + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k+1} \mathrm{Cov}[X_i, X_j]$$

as

$$2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \mathrm{Cov}[X_i, X_j] + 2\mathrm{Cov}\Big[\sum_{i=1}^{k} X_i, X_{k+1}\Big] = 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \mathrm{Cov}[X_i, X_j] + 2\sum_{i=1}^{k} \mathrm{Cov}\Big[X_i, X_{k+1}\Big]$$

$$= 2\sum_{i=1}^{k}\sum_{j=i+1}^{k+1} \mathrm{Cov}[X_i, X_j]$$

thus true for $n = k + 1$. As true for $n = 1$, $n = k + 1$ and assumed true for $n = k$ then true for all $n \geq 1$. □

6.5 Conditional expectation

Similar to how we defined the conditional probability of an event A given another event B, we will now define the expectation of a random variable X given an event B.

Definition 6.6 (Indicator function)**.** Let A be an event. We define the indicator function of A as

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

Definition 6.7 (Conditional expectation with respect to an event)**.** Let X be a random variable and let $A \subseteq \mathcal{F}$ be an event with $\mathbb{P}(A) > 0$. The conditional expectation of X given A is defined as

$$\mathbb{E}[X|A] := \frac{\mathbb{E}[X\mathbb{I}_A]}{\mathbb{P}(A)}.$$

Just as when we defined conditional probability, the expectation of X given A *does not* represent a random variable.

The interpretation of the conditional expectation with respect to an event should be thought of similarly to the conditional probability - a definition.

Theorem 6.10. For any discrete random variable X and event $A \subseteq \mathcal{F}$ with $\mathbb{P}(A) > 0$, we have

$$\mathbb{E}[X|A] = \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x|A).$$

Proof. Letting $Y = X\mathbb{I}_A$ and using the definition of conditional probability,

$$\mathbb{P}(Y = x) = \mathbb{P}(\{X = x\} \cap A) = \mathbb{P}(X = x|A)\mathbb{P}(A).$$

Hence by the definition of expectation,

$$\mathbb{E}[X|A] = \frac{\mathbb{E}[X\mathbb{I}_A]}{\mathbb{P}(A)} = \frac{\mathbb{E}[Y]}{\mathbb{P}(A)} = \frac{1}{\mathbb{P}(A)} \sum_{x \in \mathcal{X}} x\mathbb{P}(Y = x) = \frac{1}{\mathbb{P}(A)} \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x|A)\mathbb{P}(A)$$

and thus

$$\mathbb{E}[X|A] = \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x|A).$$

□

Theorem 6.11. Let X be a random variable and $E_1, E_2, \cdots$ form a partition of the sample space. We have

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{E}[X|E_i]\mathbb{P}(E_i).$$

Proof. Since $E_1, E_2, \cdots$ form a partition $\bigcup_i E_i = \Omega$. Hence,

$$1 = \mathbb{I}_\Omega = \mathbb{I}_{\bigcup_i E_i} = \sum_{i=1}^{\infty} \mathbb{I}_{E_i}.$$

Using Theorem 6.4[8],

$$\mathbb{E}[X] = \mathbb{E}\Big[X \sum_{i=1}^{\infty} \mathbb{I}_{E_i}\Big] = \sum_{i=1}^{\infty} \mathbb{E}[X\mathbb{I}_{E_i}] = \sum_{i=1}^{\infty} \mathbb{E}[X|E_i]\mathbb{P}(E_i).$$

□

So far we have found results that mirror the results when we looked at conditional probability. We will now look at the conditional expectation of the random variable X given another random variable Y.

Definition 6.8 (Conditional expectation with respect to a random variable)**.** Let X and Y be two random variables. We will define a function $g : Y(\Omega) \to \mathbb{R}$ by

$$g(y) = \mathbb{E}[X|Y = y]$$

which is the conditional expectation of X given the event $\{Y = y\}$. The conditional expectation of X given Y is itself a random variable given by

$$\mathbb{E}[X|Y] := g(Y).$$

[8] We are taking $n \to \infty$ as our partition is infinite

It is important to notice that $\mathbb{E}[X|Y=y]$ is *not* a random variable whilst $\mathbb{E}[X|Y]$ *is* a random variable taking values $\mathbb{E}[X|Y=y]$ each with probability $\mathbb{P}(Y=y)$. You can also view $\mathbb{E}[X|Y]$ as a random variable of Y as for each value of Y we return the expectation of X given that value of Y.

Theorem 6.12. Let X and Y be random variables on the same sample space,

$$\mathbb{E}[X] = \mathbb{E}\Big[\mathbb{E}[X|Y]\Big].$$

Proof. We will provide a proof for the case when X and Y are discrete. Using that $\mathbb{E}[X|Y] = g(Y)$ is a random variable,

$$\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[g(Y)] = \sum_{y\in\mathcal{Y}} g(y)p_Y(y) = \sum_{y\in\mathcal{Y}} \mathbb{E}[X|Y=y]p_Y(y)$$

where $p(y)$ is the probability mass function of y. By Definition 6.7,

$$\sum_{y\in\mathcal{Y}} \mathbb{E}[X|Y=y]p_Y(y) = \sum_{y\in\mathcal{Y}} \Big(\sum_{x\in\mathcal{X}} x\mathbb{P}(X=x|Y=y)\Big)p_Y(y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} x\frac{p_{X,Y}(x,y)}{p_Y(y)}p_Y(y)$$

where $p_{X,Y}(x,y)$ is the joint probability mass function of X and Y. Changing the order of summation[9],

$$\sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} x\frac{p_{X,Y}(x,y)}{p_Y(y)}p_Y(y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} xp_{X,Y}(x,y) = \sum_{x\in\mathcal{X}} x\Big(\sum_{y\in\mathcal{Y}} p_{X,Y}(x,y)\Big) = \sum_{x\in\mathcal{X}} xp_X(x) = \mathbb{E}[X]$$

by Theorem 5.2.

The proof for the continuous case follows in a similar fashion, changing the sum to an integral and the probability mass function to a probability density function. □

This result is useful when attempting to find $\mathbb{E}[X]$ when it may be easier to work with $\mathbb{E}[X|Y]$.

Theorem 6.13. Let $X_1, X_2, \cdots, X_n$ be mutually independent random variables on the same sample space,

$$\mathbb{E}\Big[\prod_{i=1}^{n} X_i\Big] = \prod_{i=1}^{n} \mathbb{E}[X_i].$$

Before our proof we need the following Lemma.

Lemma 6.3. Let X and Y be independent random variables on the same sample space,

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

Proof. Suppose that both X and Y are discrete. As X and Y are both random variables on the same sample space we can look at the bivariate random variable (X,Y). Consider the function $g(x,y) = xy$. By Theorem 6.2,

$$\mathbb{E}[XY] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xyp_{X,Y}(x,y) = \sum_{x\in\mathcal{X}} xp_X(x)\sum_{y\in\mathcal{Y}} yp_Y(y) = \mathbb{E}[X]\mathbb{E}[Y]$$

as $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ (due to X and Y being independent).

The continuos case follows in the same way. □

[9]This is allowed as we are assuming these expectations exist and thus converge absolutely

Now we can prove Theorem 6.13.

Proof. Using induction we have our base case $n = 1$:

$$\mathbb{E}\Big[\prod_{i=1}^{(1)} X_i\Big] = \mathbb{E}[X_1] = \prod_{i=1}^{(1)} \mathbb{E}[X_i]$$

thus true for $n = 1$.

Suppose

$$\mathbb{E}\Big[\prod_{i=1}^{k} X_i\Big] = \prod_{i=1}^{k} \mathbb{E}[X_i]$$

is true for some $n = k$.

Consider $n = k + 1$:

$$\mathbb{E}\Big[\prod_{i=1}^{k+1} X_i\Big] = \mathbb{E}\Big[\prod_{i=1}^{k} X_i \cdot X_{k+1}\Big] = \mathbb{E}\Big[\prod_{i=1}^{k} X_i\Big]\mathbb{E}[X_{k+1}] = \prod_{i=1}^{k+1} \mathbb{E}[X_i].$$

by Lemma 6.3 and the fact that $\prod_{i=1}^{k} X_i$ and X_{k+1} are independent random variables. Thus true for $n = k + 1$.

As true for $n = 1$, $n = k + 1$ and assumed true for $n = k$ then true for all $n \geq 1$. □

Corollary 6.1. Let X and Y be independent random variables, then

$$\mathrm{Cov}[X, Y] = 0$$

Proof. From Theorem 6.8 and Theorem 6.13,

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

□

Corollary 6.2. Let $X, X_2, \cdots, X_n$ be mutually independent random variables on the same sample space and $f_1, f_2, \cdots, f_n$ be functions with $f_i : X_i(\Omega) \to \mathbb{R}$ for $1 \leq i \leq n$. We have

$$\mathbb{E}\Big[\prod_{i=1}^{n} f(X_i)\Big] = \prod_{i=1}^{n} \mathbb{E}[f(X_i)].$$

Proof. It follows that $f_i(X_i)$ for $1 \leq i \leq n$ form n mutually independent random variables, then use Theorem 6.13. □

Corollary 6.3. Let $X, X_2, \cdots, X_n$ be mutually independent random variables on the same sample space,

$$\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

Proof. From Theorem 6.9,

$$\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathrm{Cov}[X_i, X_j]$$

and using Theorem 6.1,

$$\mathrm{Var}\Big[\sum_{i=1}^{n} X_i\Big] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} 0 = \sum_{i=1}^{n} \mathrm{Var}[X_i].$$

□

6.6 Expectation and probability inequalities

Starting from a familiar place, we can build results so that we can bound probabilities using expectation and other probabilities.

Theorem 6.14. For any random variable X and $a \in \mathbb{R}$,

$$\text{If} \quad \mathbb{P}(X \geq a) = 1 \quad \text{then} \quad \mathbb{E}[X] \geq a.$$

This is the aforementioned familiar place, at a glance it looks obvious that if $\mathbb{P}(X \geq a) = 1$ then of course the expected value should be larger than a!

Proof. Let $\mathbb{P}(X \geq a) = 1$. We will define a new random variable $Y = X - a$. We see that

$$\mathbb{P}(X \geq a) = \mathbb{P}(Y \geq 0) = 1$$

thus $Y \geq 0$. In addition

$$\mathbb{E}[Y] = \sum_{y \in \mathcal{Y}} y p_Y(y) \geq 0$$

if Y is discrete as $y \geq 0$ when we take $\mathcal{Y} = \{0, 1, 2, \cdots\}$ and $p_Y(y) \geq 0$. In addition,

$$\mathbb{E}[Y] = \int_0^{\infty} y f_Y(y) \geq 0$$

if Y is continuous as $y \geq 0$ and $f_Y(y) \geq 0$ from Definition 4.8. Hence, $\mathbb{E}[Y] \geq 0$. Finally, using Theorem 6.3,

$$\mathbb{E}[Y] = \mathbb{E}[X - a] = \mathbb{E}[X] - a \geq 0 \Rightarrow \mathbb{E}[X] \geq a.$$

□

Corollary 6.4. For any random variable X we have that

$$\mathrm{Var}[X] \geq 0.$$

Proof. From Definition 6.3,

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

which is the expectation of the random variable $(X - \mathbb{E}[X])^2 \geq 0$. Thus by Theorem 6.14 we have $\mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big] \geq 0$. □

Theorem 6.15 (Markov's inequality)**.** If X is a random variable with $X \geq 0$ then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Proof. From Definition 6.6,

$$\mathbb{E}[\mathbb{I}_A] = 1 \cdot \mathbb{P}(A) + 0 \cdot \mathbb{P}(A^c) = \mathbb{P}(A).$$

Notice then that

$$\mathbb{E}\Big[X - a\mathbb{I}_{\{X \geq a\}}\Big] = \mathbb{E}[X] - a\mathbb{P}(X \geq a)$$

form Theorem 6.3. Moreover, we have that

$$X - a\mathbb{I}_{\{X \geq a\}} \geq 0$$

as if $X \geq a$ then,

$$X - a\mathbb{I}_{\{X \geq a\}} = X - a(1) \geq 0$$

and if $X < a$ then,

$$X - a\mathbb{I}_{\{X \geq a\}} = X - a(0) = X \geq 0.$$

Thus by Theorem 6.14 we have that

$$\mathbb{E}\Big[X - a\mathbb{I}_{\{X \geq a\}}\Big] \geq 0$$

which gives

$$\mathbb{E}[X] - a\mathbb{P}(X \geq a) \geq 0 \Leftrightarrow \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

□

Theorem 6.16 (Chebyshev's inequality)**.** Let X be a random variable and $a > 0$. We have,

$$\mathbb{P}\Big(\Big|X - \mathbb{E}[X]\Big| \geq a\Big) \leq \frac{\mathrm{Var[X]}}{a^2}.$$

Proof. Define the random variable $Y = (X - \mathbb{E}[X])^2$. Clearly $Y \geq 0$. Applying Theorem 6.15,

$$\mathbb{P}(Y \geq a^2) = \leq \frac{\mathbb{E}[Y]}{a^2}.$$

Noting that

$$\mathbb{E}[Y] = \mathbb{E}\Big[(X - \mathbb{E}[X])^2\Big] = \mathrm{Var}[X]$$

with $(X - \mathbb{E}[X])^2 \geq a^2 \Leftrightarrow \Big|X - \mathbb{E}[X]\Big| \geq a$ as $a > 0$,

$$\mathbb{P}(Y \geq a^2) = \mathbb{P}\Big((X - \mathbb{E}[X])^2 \geq a^2\Big) = \mathbb{P}\Big(\Big|X - \mathbb{E}[X]\Big| \geq a\Big) \leq \frac{\mathrm{Var}[X]}{a^2}.$$

□

These two inequalities (Markov's and Chebyshev's) can be used to bound probabilities with their expectation and variance. These will be further used in Chapter 8 in conjunction with the squeeze theorem to prove some very powerful results.

7 Moment generating functions

7.1 Definition

So far we have looked at probability mass (density) functions and the cumulative density function. These functions determine the random variable X we are interested in; if we have the probability mass (density) function or the cumulative density function we can say which random variable X is.

We will now look at the moment generating function of a continuous random variable X which uses these previous functions.

Definition 7.1 (Moment generating function)**.** For any random variable X, we define the function $\mathrm{M}_X : \mathbb{R} \to [0, \infty)$ given by

$$\mathrm{M}_X(t) := \mathbb{E}[e^{tX}]$$

to be the moment generating function of X.

If the expectation $\mathbb{E}[e^{tX}]$ does not exists then we say that the moment generating function of X is undefined.

From Theorem 6.1 we have that

$$\mathrm{M}_X(t) = \sum_{x \in \mathcal{X}} e^{tx} p(x)$$

if X is discrete, and

$$\mathrm{M}_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x) \, \mathrm{d}x$$

if X is continuous, provided that the sum or integral converge absolutely.

The moment generating function has various useful properties that we can use to calculate the expectation of a random variable X.

Definition 7.2 (Moment of a random variable)**.** For any random variable X, the n^{th} moment of X is defined as

$$n^{\text{th}} \text{ moment of } X \ := \mathbb{E}[X^n].$$

7.2 Properties, theorems and uses

Theorem 7.1. For any random variable X, the n^{th} moment of X is given by

$$\mathbb{E}[X^n] = \mathrm{M}_X^{(n)}(0)$$

where $\mathrm{M}_X^{(n)}(0)$ is the n^{th} derivative of $\mathrm{M}_X(t)$ evaluated at zero.

Proof. Recall from calculus that the Taylor series for e^{tX} about $t = 0$ is given by

$$e^{tX} = 1 + tX + \frac{1}{2!}(tX)^2 + \frac{1}{3!}(tX)^3 + \cdots$$

and converges for all $t \in \mathbb{R}$. Hence,

$$\mathrm{M}_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}\Big[1 + tX + \frac{1}{2!}(tX)^2 + \frac{1}{3!}(tX)^3 + \cdots\Big].$$

Applying Theorem 6.3 and Theorem 6.4,

$$\mathrm{M}_X(t) = 1 + t\mathbb{E}[X] + \frac{1}{2!}t^2\mathbb{E}[X^2] + \frac{1}{3!}t^3\mathbb{E}[X^3] + \cdots + \frac{1}{n!}t^n\mathbb{E}[X^n] + \cdots$$

and comparing this to the Taylor series[10] of $\mathrm{M}_X(t)$ about $t = 0$,

$$\mathrm{M}_X(t) = \mathrm{M}_X(0) + t\mathrm{M}'_X(0) + \frac{1}{2!}t^2\mathrm{M}''_X(0) + \frac{1}{3!}t^3\mathrm{M}'''_X(0) + \cdots + \frac{1}{n!}t^n\mathrm{M}^{(n)}_X(0) + \cdots$$

we see that

$$\mathbb{E}[X^n] = \mathrm{M}^{(n)}_X(0).$$

□

We will also assume the following theorem without proof as the proof goes beyond the scope of this course.

Theorem 7.2. Let X and Y be two random variables. If there is a $h > 0$ such that

$$\mathrm{M}_X(t) = \mathrm{M}_Y(t) < \infty$$

for all $t \in (-h, h)$, then

$$\mathrm{F}_X(x) = \mathrm{F}_Y(x)$$

for all $x \in \mathbb{R}$.

Conversely, if $\mathrm{F}_X(x) = \mathrm{F}_Y(x)$ for all $x \in \mathbb{R}$, then $\mathrm{M}_X(t) = \mathrm{M}_Y(t)$ for all $t \in \mathbb{R}$.

Theorem 7.3. Let X be a random variable and $\alpha, \beta \in \mathbb{R}$, then

$$\mathrm{M}_{\alpha X+\beta}(t) = e^{\beta t}\mathrm{M}_X(\alpha t).$$

Proof. Using Theorem 6.3,

$$\mathrm{M}_{\alpha X+\beta}(t) = \mathbb{E}[e^{t(\alpha X+\beta)}] = e^{t\beta}\ \mathbb{E}[e^{(t\alpha)X)}] = e^{\beta t}\mathrm{M}_X(\alpha t).$$

□

Theorem 7.4. Suppose that $X_1, \cdots, X_n$ are mutually independent random variables and let $Y = \sum_{i=1}^n X_i$.

Then

$$\mathrm{M}_Y(t) = \prod_{i=1}^n \mathrm{M}_{X_i}(t).$$

Proof. Using the definition of the moment generating function,

$$\mathrm{M}_Y(t) = \mathbb{E}[e^{tY}] = \mathbb{E}\Big[e^{t\sum_{i=1}^n X_i}\Big] = \mathbb{E}\Big[\prod_{i=1}^n e^{tX_i}\Big]$$

by the properties of exponentials. As $X_1, \cdots, X_n$ are mutually independent then so are $e^{X_1}, \cdots, e^{X_n}$. Thus by Theorem 6.13,

$$\mathbb{E}\Big[\prod_{i=1}^n e^{tX_i}\Big] = \prod_{i=1}^n \mathbb{E}[e^{tX_i}] = \prod_{i=1}^n \mathrm{M}_{X_i}(t).$$

□

[10] Some work is needed to show that $\mathrm{M}_X(t)$ is in fact infinitely differentiable about $t = 0$ which we omit

Similar to Theorem 7.2, we will assume the following theorem without proof.

Theorem 7.5. Suppose that $X_1, X_2, \cdots$ is an infinite sequence of random variables, and that X is a further random variable. If there is a $h > 0$ such that

$$\lim_{n\to\infty} \mathrm{M}_{X_n}(t) = \mathrm{M}_X(t) < \infty$$

for all $t \in (-h, h)$ then

$$\lim_{n\to\infty} \mathrm{F}_{X_n}(x) = \mathrm{F}_X(x)$$

for all $x \in \mathbb{R}$ where $\mathrm{F}_X(x)$ is continuous.

Using these theorems in conjunction with one another can allow for simple proofs of some tricky results.

Example 7.1. Let Z be a random variable such that $Z \sim \mathrm{N}(0, 1)$.

Using Definition 7.1,

$$\mathrm{M}_Z(t) = \mathbb{E}[e^{tZ}] = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,\mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-t)^2/2}\ e^{t^2/2}\,\mathrm{d}z$$

completing the square in the exponential.

Letting $y = z - t$,

$$\mathrm{M}_Z(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-t)^2/2}\ e^{t^2/2}\,\mathrm{d}z = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\,\mathrm{d}y = e^{t^2/2}.$$

Example 7.2. Let $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Z \sim \mathrm{N}(0, 1)$. From Theorem 4.6, $\sigma Z + \mu \sim \mathrm{N}(\mu, \sigma^2)$ thus $X = \sigma Z + \mu$. By Theorem 7.3,

$$\mathrm{M}_{\sigma Z+\mu}(t) = e^{\mu t}\ \mathrm{M}_Z(\sigma t) = e^{\mu t}\ e^{\sigma^2 t^2/2} = \exp\Big\{\frac{\sigma^2}{2}t^2 + \mu t\Big\}.$$

from Example 7.1.

Hence,

$$\mathrm{M}_X(t) = \exp\Big\{\frac{\sigma^2}{2}t^2 + \mu t\Big\}.$$

Example 7.3. Suppose we have mutually independent random variables $X_1, \cdots, X_n$ with $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. Define the random variable $Y = \sum_{i=1}^n X_i$. Using Theorem 7.4 and Example 7.2,

$$\mathrm{M}_Y(t) = \prod_{i=1}^n \mathrm{M}_{X_i}(t) = \prod_{i=1}^n \exp\Big\{\frac{\sigma_i^2}{2}t^2 + \mu_i t\Big\} = \exp\Big\{\frac{\sigma^2}{2}t^2 + \mu t\Big\}$$

where $\sigma^2 = \sum_{i=1}^n \sigma_i^2$ and $\mu = \sum_{i=1}^n \mu_i$. Hence by Theorem 7.2 we have $Y \sim \mathrm{N}(\mu, \sigma^2)$; this means that the sum of mutually independent normally distributed random variables is also a normally distributed random variable!

8 Limit theorems

8.1 The weak law of large numbers

To build some intuition for the law of large numbers we will look at the following scenario.

Suppose we toss a fair coin n-times and count the number of heads. Let X be the number of heads and $\mathrm{P}_n = X/n$ be the proportion of heads in n-tosses. As $X \sim \mathrm{Bin}(n, 1/2)$, $\mathbb{E}[X] = n/2$ and $\mathrm{Var}[X] = n/4$ then,

$$\mathbb{E}[\mathrm{P}_n] = \frac{\mathbb{E}[X]}{n} = \frac{1}{2} \quad \text{and} \quad \mathrm{Var}[\mathrm{P}_n] = \frac{\mathrm{Var}[X]}{n^2} = \frac{1}{4n}$$

using Theorem 6.3 and Theorem 6.5.

Applying Chebyshev's inequality, for any $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|\mathrm{P}_n - \frac{1}{2}\Big| > \varepsilon\Big) \leq \frac{1}{4n\varepsilon^2}.$$

Taking the limit as $n \to \infty$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\mathrm{P}_n - \frac{1}{2}\Big| > \varepsilon\Big) \leq \lim_{n\to\infty} \frac{1}{4n\varepsilon^2} = 0$$

so by the squeeze theorem

$$\lim_{n\to\infty} \mathbb{P}\Big(\Big|\mathrm{P}_n - \frac{1}{2}\Big| > \varepsilon\Big) = 0.$$

We can interpret this as the probability of the difference $|\mathrm{P}_n - 1/2|$ being non-zero goes to zero as $n \to \infty$. Which agrees with our intuition: as we increase the number of times we toss the coin, the proportion of heads, P_n, should get closer and closer to $1/2$.

This is in fact a special case of the weak law of large numbers.

Theorem 8.1 (The weak law of large numbers)**.** Suppose we have an infinite sequence $X_1, X_2, \cdots$ of mutually independent random variables, with

$$\mathbb{E}[X_i] = \mu \quad \text{and} \quad \mathrm{Var}[X_i] = \sigma^2$$

for all $i \in \mathbb{N}$. Define the sample average as $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Then, for every $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) = 0.$$

This can be summarised as $\overline{X}_n$ converges in probability to μ.

Proof. Using Theorem 6.3 and Theorem 6.4,

$$\mathbb{E}[\overline{X}_n] = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n X_i\Big] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}][X_i] = \frac{1}{n}\cdot n\mu = \mu.$$

Similarly, using Theorem 6.5 and Corollary 6.3,

$$\mathrm{Var}[\overline{X}_n] = \mathrm{Var}\Big[\frac{1}{n}\sum_{i=1}^n X_i\Big] = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i] = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}.$$

Applying Chebyshev's inequality, for any $\varepsilon > 0$,

$$0 \leq \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0$$

as $n \to \infty$. Hence by the squeeze theorem[11],

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| > \varepsilon) = 0.$$

□

8.2 The central limit theorem

Definition 8.1 (Independently and identically distributed random variables)**.** Let $X_1, X_2, \cdots$ be random variables. We say that $X_1, X_2, \cdots$ are independently and identically distributed (i.i.d) if they are all independent and have the same marginal distribution.

Theorem 8.2 (Central limit theorem)**.** Let $X_1, X_2, \cdots$ be i.i.d random variables with

$$\mathbb{E}[X_i] = \mu \quad \text{and} \quad \mathrm{Var}[X_i] = \sigma^2$$

for all $i \in \mathbb{N}$. Define the sample average $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and the random variable Z_n by

$$Z_n := \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Then for every $z \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) = \Phi(z)$$

where $\Phi(z)$ is the cumulative distribution function of the random variable $Z \sim \mathrm{N}(0,1)$.

This can be summarised as Z_n converges in distribution to Z.

This result is so strong because we do not need to know the distribution of the X_i's, they can have any distribution! In addition, we can use this result to approximate Z_n for large n by using $Z \sim \mathrm{N}(0,1)$.

Proof. Our plan is to show that $\mathrm{M}_{Z_n}(t) \to e^{t^2/2} = \mathrm{M}_Z(t)$ as $n \to \infty$ for all t in some interval containing 0 and then use Theorem 7.5 to obtain our result.

Recall that for any sequence a_n with $a_n \to a$ as $n \to \infty$, we have

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Observe that

$$Z_n = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \sum_{i=1}^n \frac{X_n - \mu}{\sigma\sqrt{n}} = \sum_{i=1}^n \frac{Y_i}{\sqrt{n}}$$

where we defined $Y_i = (X_i - \mu)/\sigma$. Let $\mathrm{M}_{Y_i}(t) = m(t)$ for all $i \in \mathbb{N}$, as all the Y_i's are i.i.d.

Using Theorem 7.3, the moment generating function of $Y_i/\sqrt{n}$ is $m(t/\sqrt{n})$.

[11] Recall from analysis or calculus that if $0 \leq x_n \leq y_n \to 0$ then $x_n \to 0$

Hence, by Theorem 7.4,

$$\mathrm{M}_{Z_n}(t) = \prod_{i=1}^{n} m\Big(\frac{t}{\sqrt{n}}\Big) = \bigg[m\Big(\frac{t}{\sqrt{n}}\Big)\bigg]^n.$$

Using Theorem 7.1,

$$m(0) = \mathbb{E}[Y_i^0] = 1, \quad m'(0) = \mathbb{E}[Y_i^1] = 0, \quad m''(0) = \mathbb{E}[Y_i^2] = 1.$$

Hence, by Taylor's theorem[12] there exists a function $h(u)$ with $h(u) \to 0$ as $u \to 0$ such that

$$m(u) = 1 + \frac{1}{2}u^2 + u^2 h(u).$$

Hence,

$$\mathrm{M}_{Z_n}(t) = \bigg[1 + \frac{1}{2}\frac{t^2}{n} + \frac{t^2}{n}h\Big(\frac{t}{\sqrt{n}}\Big)\bigg]^n \to e^{t^2/2}$$

as $n \to \infty$ for any $t \in \mathbb{R}$.

□

[12]This is a result typically found in an analysis 1 course and this follows from $m(u)$ being twice differentiable at $u = 0$. This is also known as a second order Taylor expansion of $m(u)$ about $u = 0$