Cardiff University School of Mathematics

MA2500

Foundations of Probability and Statistics

Lecture Notes

Autumn 2018

D Evans Version: 15th January 2018

Contents

Preface

In *MA2500 Foundations of Probability and Statistics* we build a theory of probability and statistics from first principles and as such the module can be studied independently of the first year modules *MA1500 Introduction to Probability Theory* and *MA1501 Statistical Inference.* The module however does assume a good understanding of many fundamental mathematical ideas and techniques covered in *MA1005 Foundations of Mathematics I* and *MA1006 Foundations of Mathematics II.*

Timetable

Monday	16.10 – 17.00	Marking session	E/0.15
Tuesday	15.10 – 17.00	Lectures and discussion	E/0.15
Thursday	12.10 – 14.00	Lectures and discussion	E/0.15

Topics

Week 1	Elementary probability	Week 7	Estimation
Week 2	Probability spaces	Week 8	Likelihood
Week 3	Random variables	Week 9	Hypothesis testing
Week 4	Expectation	Week 10	Non-parametric methods
Week 5	Sums of random variables	Week 11	Bivariate analysis
Week 6	Joint distributions	Week 12	Random processes

Instructions

Read handout notes before lectures.

Attend all lectures and marking sessions.

Ask questions during lectures and marking sessions.

Work with peers on exercises and participate in marking sessions.

Ask for help when required.

Handouts

These handout notes contain only definitions, statements of theorems, exercise questions and partial examples. During lectures you should annotate these notes and use a separate notebook for proofs, sketches and longer examples.

You are expected to read the relevant parts of the handout notes before lectures.

Exercises

Learning mathematics takes exercise, which invariably involves a lot of hard work. When solutions to exercises are available we can be tempted to give up the struggle and look at the solutions. Struggling with exercises however is by far the best way to **understand mathematics**, and is the only way we can learn how to **do mathematics**. Nobody expects to get fit by watching someone working on an exercise bike: we need to get on the bike ourselves and start pedalling, which of course involves a lot of hard work. For this reason, **written solutions to exercises will not be provided** as a matter of course, but individual solutions will be provided on request, perhaps as short videos. Answers to exercises can be submitted at any time for assessment and feedback. This extends to any exercise contained in either of the recommended textbooks for the module. Every exercise is a challenge.

Marking sessions

The ability to criticise your own work is very important, but it does not always come easily. To develop our critical faculties we can first learn to criticise other people's work, before applying the same principles to our own. For this reason we will follow a weekly cycle of **peer marking**. Every week teams of 2–4 students will prepare written answers to selected exercise questions and bring these along to peer marking sessions. During these sessions, solutions will be presented on the board while and teams mark and provide feedback on each others' work using a standard marking form. The marked scripts will be collected at the end of each session and returned the following week with additional feedback included where necessary.

All answers should aim to meet the criteria shown in Table 1 and marked accordingly.

Clear	Statements are explicit and unambiguous.
Complete	All relevant details are included.
Concise	No irrelevant details are included.
Coherent	An appropriate narrative is provided.
Correct	Arguments are precise and logically sound.

Table 1: Assessment criteria for homework exercises.

Recommended textbooks

1. Probability and Random Processes (Third edition)
   G. R. Grimmett and D. R. Stirzaker
   Oxford University Press (2001)
   ISBN 0-19-857222-0

2. Introduction to Mathematical Statistics (Sixth edition)
   R. V. Hogg, J. W. McKean and A. T. Craig
   Prentice Hall (2005)
   ISBN 0-13-122605-3

A brief history of probability

In the 17th century, gambling was popular among the French aristocracy. A popular game involved rolling a pair of fair dice 24 times and betting even money on a "double six" appearing at least once. It was understood that betting on at least one six in four rolls of a single fair die was profitable, and it was believed that the same reasoning extended to betting on at least one double six in 24 rolls. A French nobleman, Chevalier de Méré, knew from experience that when betting on a six appearing in four rolls of a single die he won more times than he lost, but when betting on a double six appearing in 24 rolls of two dice, he lost more times than he won.

- In 1654 de Méré asked his friend, the mathematician Blaise Pascal (1623-1662), to explain this apparent paradox. The problem initiated an exchange of letters between Pascal and Pierre de Fermat (1601-1665), which led to the formulation of the classical principles of probability theory.
- In 1657 the Dutch scientist Christiaan Huygens (1629-1695) published the first book on probability, *De Ratiociniis in Ludo Aleae* (*On Reasoning in Games of Chance*) which introduced the concept of mathematical expectation.
- In 1713 the celebrated book *Ars Conjectandi* by Swiss mathematician Jakob Bernoulli (1654-1705) was published. It contains a theorem known as the law of large numbers, the first limit theorem in probability theory.
- In 1718 Abraham de Moivre (1667-1754) published *The Doctrine of Chances.* This was the first textbook on probability theory and is said to have been prized by gamblers.
- In 1812 Pierre de Laplace (1749-1827) introduced many new ideas and techniques in his book, *Théorie Analytique des Probabilités.* Laplace applied probabilistic ideas to many scientific and practical problems as well as to games of chance.
- In 1919 Richard von Mises (1883-1953) established an axiomatic approach to the subject based on relative frequency, and introduced the idea of sample spaces.
- In 1933 Andrey Kolmogorov (1903-1987) introduced the modern axiomatic theory of probability, which is part of a more general field known as **measure theory**.

Our study will be based on Kolmogorov's axiomatic theory.

Chapter 1 Introduction

1.1 Events

A set is a collection of distinct elements. If ω is an element of the set A, we denote this by $\omega \in A$. The set containing no elements is called the **empty set** and denoted by $\emptyset$.

Definition 1.1 (Set relations)
Let A and B be sets.

1. A is a **subset** of B if $\omega \in B$ for every $\omega \in A$. This is denoted by $A \subseteq B$.
2. A is **equal** to B if $A \subseteq B$ and $B \subseteq A$. This is denoted by $A = B$.
3. A is a **proper subset** of B if $A \subseteq B$ and $A \neq B$. This is denoted by $A \subset B$.
4. A and B are **disjoint** (or **mutually exclusive**) if $A \cap B = \emptyset$.

Definition 1.2 (Set operations)
Let Ω be a set, and let A and B be subsets of Ω.

1. The **complement** of A (relative to Ω) is the set $A^c = \{\omega \in \Omega : \omega \notin A\}$.
2. The **union** of A and B is the set $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$.
3. The **intersection** of A and B is the set $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$.

Table 1.1 shows the connection between set theory and logic.

Set Theory		Logic		
Union	$A \cup B$	Disjunction	OR	$\vee$
Intersection	$A \cap B$	Conjunction	AND	$\wedge$
Complement	A^c	Negation	NOT	$\neg$

Table 1.1: Correspondence between set operations and logical connectives.

Definition 1.3 (Set algebra)
The union and intersection operations have the following properties.

- The **commutative** property: $A \cup B = B \cup A$ and $A \cap B = B \cap A$.
- The **associative** property: $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$.
- The **distributive** property: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ and $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

Note that a statement such as $A \cup B \cap C$ is ambiguous.

Exercise 1.4

1. The **set difference** $A \setminus B$ is the set containing those elements of A that are not contained in B. Express $A \setminus B$ using only the intersection and complementation operations.

2. Prove De Morgan's laws: $(A \cup B)^c = A^c \cap B^c$ and $(A \cap B)^c = A^c \cup B^c$.

1.1.1 Countable unions and intersections

Let $A_1, A_2, \ldots$ be a countable family of sets. The union and intersection of this countable family are defined by

$$\textstyle\bigcup_{i=1}^{\infty} A_i = \{\omega : \omega \in A_i \text{ for some } i = 1, 2, \ldots\}$$

$$\textstyle\bigcap_{i=1}^{\infty} A_i = \{\omega : \omega \in A_i \text{ for all } i = 1, 2, \ldots\}$$

Exercise 1.5

Show that De Morgan's laws hold for countable unions and intersections:

$$\left(\textstyle\bigcup_{i=1}^{\infty} A_i\right)^c = \textstyle\bigcap_{i=1}^{\infty} A_i^c \quad \text{and} \quad \left(\textstyle\bigcap_{i=1}^{\infty} A_i\right)^c = \textstyle\bigcup_{i=1}^{\infty} A_i^c.$$

Definition 1.6

1. A **random experiment** is any process of observation or measurement whose result is uncertain.

2. The result of a random experiment is called the **outcome**.

3. The **sample space** of a random experiment is the set of all possible outcomes.

4. A **random event** is a subset of the sample space.

Sample spaces can be **finite**, **countably infinite** or **uncountable** sets.

Example 1.7

1. A coin is tossed 10 times. The outcome is the total number of heads.

   - The sample space is a finite set: $\Omega = \{0, 1, 2, \ldots, 10\}$.
   - $A = \{1, 3, 5, 7, 9\}$ is the event that the total is odd.

2. A coin is tossed repeatedly until the first head occurs. The outcome is the number of times the coin is tossed.

   - The sample space is a countably infinite set: $\Omega = \{1, 2, 3, \ldots\}$.
   - $A = \{1, 3, 5, \ldots\}$ is the event that the number is odd.
   - $B_n = \{n, n+1, n+2, \ldots\}$ is the event that the coin is tossed at least n times.

3. A spinning top is spun and comes to rest at a random angle relative to the horizontal axis.

   - The sample space is an uncountable set: $\Omega = [0, 2\pi)$.
   - $A = [0, \pi/2]$ is the event that the angle lies in the positive quadrant.

Definition 1.8

Let Ω be the sample space of some random experiment and let $A \subseteq \Omega$ be an event. Suppose we perform the experiment and observe the outcome ω. If $\omega \in A$ we say that event A **occurs**, otherwise we say that A **does not occur**.

Example 1.9
A coin is tossed 10 times. Let the outcome of the experiment be the total number of heads, and let A be the event that the total is odd.

- If we observe a total of 5 heads then A occurs.
- If we observe a total of 6 heads, A does not occur.

Exercise 1.10
Let A and B be two random events.

1. Show that if A occurs and $A \subseteq B$ then B also occurs.
2. Show that if A occurs and $A \cap B = \emptyset$ then B does not occur.

Table 1.2 shows the correspondence between terms used in set theory and those used in probability theory.

Notation	Set theory	Probability theory
Ω	Universal set	Sample space, certain event
$\omega \in \Omega$	Element	Elementary event, outcome
$A \subseteq \Omega$	Subset	Event A
$A \subseteq B$	Inclusion	If A occurs, then B occurs
A^c	Complement	A does not occur
$A \cap B$	Intersection	A and B both occur
$A \cup B$	Union	A or B (or both) occur
$A \setminus B$	Set difference	A occurs, but B does not
$\emptyset$	Empty set	Impossible event

Table 1.2: Set theory and probability theory (Grimmett & Stirzaker 2001).

1.2 Elementary probability

What is the probability that a random event A occurs? One answer is that this is a number between 0 and 1 that indicates how likely A is to occur. Putting aside the question of how we might find such a number, this is not a convincing definition because we must now explain what we mean by "how likely". It turns out that probability is difficult to define precisely.

Suppose for the moment however that probabilities have been assigned to the individual outcomes of a random experiment. This can be represented by a **probability mass function**,

$$\begin{aligned} p: \quad \Omega &\longrightarrow [0,1] \\ \omega &\mapsto p(\omega). \end{aligned}$$

Because we want to **quantify** how likely A is to occur, it is reasonable to require that $p(\omega) \geq 0$ for all $\omega \in \Omega$. If the sample space is a countable set, it is also reasonable to define the probability of an event A to be equal the sum of the probabilities of the outcomes it contains:

$$P(A) = \sum_{\omega \in A} p(\omega). \tag{*}$$

This is a **probability distribution function**, which assigns probabilities to **subsets** of the sample space.

Intuitively, probability should somehow reflect a sense of **proportion**, between how likely A is to occur and how likely A is not to occur.

- The ratio $P(A)/P(A^c)$ is called the **odds on** A occurring.
- The ratio $P(A^c)/P(A)$ is called the **odds against** A occuring.

Odds suffer by the fact that they can take arbitrarily large values. A better definition is obtained by expressing how likely A is to occur as a proportion of how likely **any** event is to occur, and make the latter equal to 1:

$$\sum_{\omega\in\Omega} p(\omega) = 1.$$

Exercise 1.11
Show that probability distribution functions as defined in (*) have the following properties, where A and B are subsets of the sample space:

1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
2. Complementarity: $P(A^c) = 1 - P(A)$.
3. Monotonicity: if $A \subseteq B$ then $P(A) \leq P(B)$.
4. Additivity: if A and B are disjoint then $P(A \cup B) = P(A) + P(B)$.
5. Addition formula: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.2.1 Conditional probability

Suppose we know that event B occurs. Then event A occurs if and only if $A \cap B$ occurs, and the probability that A occurs should now represent how likely $A \cap B$ is to occur as a proportion of how likely B was to occur in the first place. Thus we define the **conditional probability of** A **given** B to be

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

provided that $P(B) > 0$.

To analyse random events it is often useful to break then into disjoint components and analyse each component separately.

Definition 1.12
A family of sets $\{A_1, A_2, \dots A_n\}$ is called a **partition** of set B if

1. $B \subseteq \bigcup_{i=1}^{n} A_i$ and
2. $A_i \cap A_j = \emptyset$ for all $i \neq j$.

Theorem 1.13 (Partition Theorem)
If $\{A_1, A_2, \dots, A_n\}$ is a partition of B then $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$.

Proof: Because the A_i are disjoint and B is contained in $\subseteq \cup_{i=1}^{n} A_i$, we can represent B as a disjoint union,

$$B = (B \cap A_1) \cup (B \cap A_2) \cup \ldots = \bigcup_{i=1}^{\infty} (B \cap A_i)$$

By the additivity property (see Exercise 1.11)

$$P(B) = P\left(\bigcup_{i=1}^{\infty} (B \cap A_i)\right) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i),$$

as required.

If we perform a random experiment and observe that B occurs, how does this change what we know about event A_i? This is answered by Bayes' theorem.

Theorem 1.14 (Bayes' Theorem)
If $\{A_1, A_2, \dots, A_n\}$ is a partition of B and $P(B) > 0$, then $P(A_i|B) = \dfrac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$.

Proof: Set intersection is commutative so

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

Hence,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}$$

where the last equality follows by the partition theorem, as required.

Example 1.15
One in every 100 people has a certain disease. People having the disease test positively with probability 0.92; people not having the disease test negatively with probability 0.97. If a person tests positively for the disease, what is the probability that the person actually has the disease?

Solution: Let A be the event that the person has the disease. Let B be the event that the person tests positively:

$$P(A) = 0.01,\ P(A^c) = 0.99,\ P(B|A) = 0.92 \text{ and } P(B|A^c) = 0.03$$

The set $\{A, A^c\}$ is a partition of B, so by the partition theorem,

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c) = (0.92 \times 0.01) + (0.03 \times 0.99) = 0.029403.$$

By Bayes' theorem,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.92 \times 0.01}{0.029403} = 0.3129 \text{ approx.}$$

Approximately one third of positive tests are "true positives", the remainder are "false positives".

Exercise 1.16

1. **Bertrand's box paradox**. Suppose we have a box containing two gold coins, a box containing two silver coins, and a box containing one gold coin and one silver coin. A box is chosen at random, and a coin is chosen at random from the box. If the chosen coin is a gold coin, what is the probability that the box contains another gold coin?

   Answer: Let A be the event that a gold coin is chosen. By Bayes theorem (with the obvious notation),

   $$\begin{aligned} P(GG|A) &= \frac{P(A|GG)P(GG)}{P(A|GG)P(GG) + P(A|SS)P(SS) + P(A|GS)P(GS)} \\ &= \frac{1 \times 1/3}{(1 \times 1/3) + (0 \times 1/3) + (1/2 \times 1/3)} = \frac{2}{3}. \end{aligned}$$

   This is related to the famous **Monty Hall problem**.

2. **Galton's paradox**. Three fair coins are tossed independently. Given that at least two are alike, the probability that they are all alike is 1/2. Do you agree?

   Answer: No. For any outcome at least two coins will be alike, so this provides no additional information. In fact,

   $$P(\text{All are alike}|\text{Two are alike}) = P(\text{All are alike}) = 1/4.$$

1.2.2 Independence

If the probability that event A occurs is not affected by whether or not event B occurs (and vice versa), we say that A and B are **independent**.

Definition 1.17
Two events A and B are said to be **independent** if $P(A|B) = P(A)$ or equivalently

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

Example 1.18
A fair die is rolled once. Let A be the event that the outcome is even, and let B be the event that the outcome is divisible by 3. Are A and B independent?

Solution:

$$P(A) = P(\{2,4,6\}) = 1/2; \quad P(B) = P(\{3,6\}) = 1/3; \quad P(A \cap B) = P(\{6\}) = 1/6.$$

Thus A and B are independent because $P(A \cap B) = P(A)P(B)$.

The notion of independence can be extended to three or more events.

Definition 1.19
A family of events $\{A_1, A_2, \ldots\}$ is said to be

1. **pairwise independent** if $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j$.
2. **totally independent** if for every finite sub-family $\{B_1, B_2, \ldots, B_m\} \subset \{A_1, A_2, \ldots\}$,

$$P(B_1 \cap B_2 \cap \ldots \cap B_m) = P(B_1)P(B_2)\cdots P(B_m).$$

Note that total independence implies pairwise independence, but not vice versa.

Example 1.20
Let $\Omega = \{1,2,3,4\}$ where each outcome is equally likely, and consider the events $A = \{1,2\}$, $B = \{1,3\}$ and $C = \{1,4\}$. Show that $\{A,B,C\}$ is pairwise independent but not totally independent.

Solution:

- A and B are independent: $P(A) = P(B) = 1/2$ and $P(A \cap B) = 1/4$ so $P(A \cap B) = P(A)P(B)$.
- Similary A and C are independent, and B and C are indpendent, so the set $\{A,B,C\}$ is pairwise independent.
- In contrast, $P(A \cap B \cap C) = 1/4$ but $P(A)P(B)P(C) = 1/8$, so the set $\{A,B,C\}$ is not totally independent.

Exercise 1.21

1. Let A and B be two events with $P(A) = 0.2$, $P(B) = 0.5$ and $P(A \cap B) = 0.1$. Compute the conditional probability $P(A|B)$ and decide whether or not A and B are independent.

   Answer:

   - $P(A|B) = P(A \cap B)/P(B) = 0.1/0.5 = 0.25$.
   - $P(A|B) = P(A)$ so A and B are independent.

2. Show that if A and B are independent then A and B^c are also independent.

   Answer:

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \quad \text{by independence,} \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^c). \end{aligned}$$

3. **de Méré's paradox**. Solve de Méré's paradox by computing the probability that at least one six appears in 4 rolls of a single fair die, and the probability that at least one double-six appears in 24 rolls of two fair dice.

   Answer: Assuming that the rolls are independent:

$$\begin{aligned} P(\text{at least one six in 4 rolls of a single die}) &= 1 - P(\text{no sixes obtained in 4 rolls}) \\ &= 1 - (5/6)^4 \quad \text{(by independence)} \\ &= 0.5177 \\ P(\text{at least one double six in 24 rolls of two dice}) &= 1 - P(\text{no double-sixes in 24 rolls}) \\ &= 1 - (35/36)^{24} \quad \text{(by independence)} \\ &= 0.4914 \end{aligned}$$

4. Two fair dice are rolled independently. Let A be the event that the first die shows 3, B that the second die shows 4 and and C that the total shown on the two dice is 7.

   (a) Define a sample space for the experiment and identify the subsets corresponding to events A, B and C.

   Answer:

   $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$
   $A = \{(3,j) : 1 \leq j \leq 6\}$
   $B = \{(i,4) : 1 \leq i \leq 6\}$
   $C = \{(i,j) : 1 \leq i,j \leq 6 \text{ and } i + j = 7\}$

   Note that because each outcome is equally likely, $P(A) = P(B) = P(C) = 1/6$.

   (b) Show that $\{A, B, C\}$ is pairwise independent but not (totally) independent.

   Answer: To show that the set $\{A, B, C\}$ is pairwise independent, we need to show that any two events chosen from the set are independent of each other.

   - For A and B, their intersection $A \cap B$ is the event $\{(3,4)\}$, consisting of the single outcome $(3,4)$. This means that $P(A \cap B) = 1/36$ and hence $P(A \cap B) = P(A)P(B)$.
   - Similarly, $A \cap C = \{(3,4)\}$ so $P(A \cap C) = P(A)P(C)$, and $B \cap C = \{(3,4)\}$ so $P(B \cap C) = P(B)P(C)$. Hence, the set $\{A, B, C\}$ is pairwise independent.
   - However, the intersection of all three events is also $\{(3,4)\}$ so $P(A \cap B \cap C) = 1/36$. Hence $P(A \cap B \cap C) \neq P(A)P(B)P(C)$ so the set $\{A, B, C\}$ is not totally independent.

5. A random number N of dice are rolled independently. Suppose that $P(N = k) = 2^{-k}$ for $k \in \{1, 2, \ldots\}$ (and zero otherwise). Let S be the sum of the scores shown on the dice. Show that

(a) $P(N = 2|S = 4) = 432/2197$,

Answer: The event $S = 4$ can only occur when

- $N = 1$ and the die shows 4 (which occurs with probability 1/6),
- $N = 2$ and the dice show $(1, 3)$, $(2, 2)$ or $(3, 1)$ (which occurs with probability 3/36),
- $N = 3$ and the dice show $(1, 1, 2)$, $(1, 2, 1)$ or $(2, 1, 1)$ (which occurs with probability 3/216),
- $N = 4$ and all four dice show 1 (which occurs with probability 1/1296).

By Bayes' theorem,

$$\begin{aligned}P(N = 2|S = 4) &= \frac{P(S = 4|N = 2)P(N = 2)}{\sum_{k=1}^{\infty} P(S = 4|N = k)P(N = k)} \\ &= \frac{3/36 \times 1/4}{(1/6 \times 1/2) + (3/36 \times 1/4) + (3/216 \times 1/8) + (1/1296 \times 1/16)} \\ &= \frac{432}{1728 + 432 + 36 + 1} = \frac{432}{2197}.\end{aligned}$$

(b) $P(S = 4|N \text{ is even}) = 433/6912$,

Answer: Using the formula for a geometric series with first term 1/4 and common raio 1/4, we have

$$P(N \text{ even}) = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots = \frac{1}{3}.$$

If N is even, then $S = 4$ only if

- $N = 2$ and the dice show $(1, 3)$, $(2, 2)$ or $(3, 1)$ (which occurs with probability 3/36),
- $N = 4$ and all four dice show 1 (which occurs with probability 1/1296).

Hence,

$$\begin{aligned}P(S = 4|N \text{ even}) &= \frac{P(S = 4 \text{ and } N \text{ even})}{P(N \text{ even})} \\ &= \frac{P(S = 4|N = 2)P(N = 2) + P(S = 4|N = 4)P(N = 4)}{P(N \text{ even})} \\ &= \frac{(3/36 \times 1/4) + (1/1296 \times 1/16)}{1/3} = \frac{433}{6912}.\end{aligned}$$

(c) $P(N = 2|\text{S=4 and the first die shows 1}) = 144/169$,

Answer: Let D be the score on the first die. If $D = 1$, then $S = 4$ only if

- $N = 2$ and the other die shows 3 (which occurs with probability 1/6),
- $N = 3$ and the other two dice show $(1, 2)$ or $(2, 1)$ (which occurs with probability 2/36), or
- $N = 4$ and the other three dice all show 1 (which occurs with probability 1/216).

Hence,

$$\begin{aligned}P(N = 2|S = 2, D = 1) &= \frac{P(N = 2, S = 4, D = 1)}{P(S = 4, D = 1)} \\ &= \frac{1/6 \times 1/4}{(1/6 \times 1/4) + (2/36 \times 1/8) + (1/216 \times 1/16)} = \frac{144}{169}.\end{aligned}$$

(d) $P(\text{Largest number is } m) = m/(12-m)$ for every $m \in \{1,2,3,4,5,6\}$.
Answer: Let M be the maximum number shown on the dice. For $m \in \{1,2,3,4,5,6\}$, the probability that m is the maximum number shown on a single die is $m/6$. Assuming that the dice are independent, the probability that m is the maximum number shown on k dice is $(m/6)^k$. Hence,

$$\begin{aligned} P(M \le m) &= \sum_{k=1}^{\infty} P(M \le m | N = k) P(N = k) \\ &= \sum_{k=1}^{\infty} \left(\frac{m}{6}\right)^k \frac{1}{2^k} = \sum_{k=1}^{\infty} \left(\frac{m}{12}\right)^k = \frac{m}{12-m}, \end{aligned}$$

where we have used the formula for a geometric series whose first term and common ratio are both equal to $m/12$.

1.3 Classical probability

We now return to the question of how we might reasonably assign probabilities to random events. Classical probability deals with finite sample spaces in which all outcomes are equally likely. This means that the probability of an event A is proportional to the number of outcomes it contains, which is called the **cardinality** of the set and denoted by $|A|$. The classical definition of probability can therefore be written as

$$P(A) = |A|/|\Omega|.$$

Problems can be solved by counting the number of ways different events can occur. For example, if a game has n possible outcomes of which m correspond to winning, then the probability of winning is m/n.

Example 1.22
An urn contains 3 white balls and 5 black balls. If two balls are drawn at random from the urn, what is the probability that they are both white?

Solution:

- The sample space Ω is the set of all possible pairs: $|\Omega| = \binom{8}{2} = \frac{8!}{6!2!} = 28$.
- Let A be the event that both balls are white: $|A| = \binom{3}{2} = \frac{3!}{1!2!} = 3$.

The probability that both balls are white is $P(A) = |A|/|\Omega| = 3/28$.

Exercise 1.23 (The Division Paradox)
A **fair game** is one in which the probability of winning is equal to the probability of losing. Two players A and B decide to play a sequence of fair games until one of the players wins 6 games, but they stop when the score is 5:3 in favour of player A. How should the prize money be fairly divided?

Answer: Assume that the players carried on playing the sequence of games. The maximum number of additional games is 3, and the sample space can be expressed as

$$\{AAA, AAB, ABA, BAA, ABB, BAB, BBA, BBB\}.$$

The games are fair, so all outcomes are equally likely.

- Only one outcome is in favour of player B, the other seven are in favour of A.
- The prize money should therefore be divided in the ratio 7 : 1 in favour of A.

1.4 Relative frequency

Classical probability exploits the symmetry that exists in many random experiments, for example those involving dice, coins and cards, to choose sensible values for the probabilities of different events. How might we define probability in more general situations? If a random experiment can be repeated many times under the same conditions it is natural to think of probability as the number of times an event occurs as a proportion of the total number times that the experiment is repeated.

Definition 1.24
Let N be the number of times an experiment is repeated and let $N(A)$ be the number of times event A occurs during these N repetitions. The ratio $N(A)/N$ is called the **relative frequency** of A.

Definition 1.25
Under the **frequentist model**, the probability of A is the limit of its relative frequency as the number of trials increases to infinity:

$$\mathbb{P}(A) = \lim_{N\to\infty} \frac{N(A)}{N}.$$

Exercise 1.26
Let Ω be a finite sample space. Show that probability as defined by the frequentist model has the following properties:

1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
2. Complementarity: if $A \subseteq B$ then $P(A^c) = 1 - P(A)$.
3. Monotonicity: $P(A) \leq P(B)$.
4. Additivity: If A and B are disjoint, $P(A \cup B) = P(A) + P(B)$.

Show also that the conditional probability of A given B is $P(A \cap B)/P(B)$ whenever $P(B) > 0$.

Answer:

1. $N(\emptyset) = 0$ and $N(\Omega) = N$ for any number of repetitions N, so $P(\emptyset) = 0$ and $P(\Omega) = 1$.
2. A^c occurs if and only if A does not, so $N(A^c) = N - N(A)$ and thus
$$P(A^c) = \lim_{N\to\infty} \frac{N(A^c)}{N} = \lim_{N\to\infty} \frac{N - N(A)}{N} = 1 - \lim_{N\to\infty} \frac{N(A)}{N} = 1 - P(A^c).$$
3. If $A \subseteq B$ then A occurs whenever B occurs, so $N(A) \leq N(B)$ and hence
$$P(A) = \lim_{N\to\infty} \frac{N(A)}{N} \leq \lim_{N\to\infty} \frac{N(B)}{N} = P(B).$$
4. If A and B are disjoint, they cannot both occur together, so $N(A \cup B) = N(A) + N(B)$ and thus
$$P(A \cup B) = \lim_{N\to\infty} \frac{N(A \cup B)}{N} = \lim_{N\to\infty} \frac{N(A) + N(B)}{N} = \lim_{N\to\infty} \frac{N(B)}{N} + \lim_{N\to\infty} \frac{N(B)}{N} = P(A) + P(B).$$

Let $P(A|B)$ denote the conditional probability of A given B. This is the number of trials in which A and B both occur expressed as a proportion of the number of trials in which B occurs. If $N(A, B)$ is the number of times A and B both occur then

$$\mathbb{P}(A|B) = \lim_{N\to\infty} \frac{N(A, B)}{N(B)} = \lim_{N\to\infty} \frac{N(A, B)/N}{N(B)/N} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

as required.

The frequentist model is the basis of how probability theory is applied to real-world problems, and dominates in many areas of science (e.g. medical trials). It does however have some serious practical and philosophical drawbacks.

- Not all experiments can be repeated many times under the same conditions. For example, it is reasonable to consider the probability that Wales will win the World Cup in 2018. The frequentist model does not provide an adequate definition of such probabilities.
- In practical applications, an experiment is repeated finitely many times and the relative frequency of an event is taken as an approximation of its "true" probability. If we accept that we can only measure probability with some error of measurement, we find that an error of measurement can itself only be expressed as a probability, which is the very concept we are trying to define.
- The frequentist model is also limited when dealing with infinite sample spaces. For example, consider a random experiment where a coin is tossed repeatedly until the first head occurs, and whose the outcome is the total number of times the coin is tossed. The sample space is the countably infinite set $\{1, 2, 3, \ldots\}$ but no matter how many times we repeat the experiment, there are outcomes which cannot be observed (e.g if we repeat the experiment N times, we cannot observe runs of length $N + 1, N + 2, \ldots$).

Frequentist probability is best regarded an informal theory which is useful in many practical applications, but which lacks the mathematical clarity offered by Kolmogorov's axiomatic theory.

Chapter 2 Probability spaces

2.1 Probability measures

Definition 2.1 (Kolmogorov's axioms)
A **probability measure** on a set Ω is a function which maps subsets of Ω to numbers in the interval $[0, 1]$ such that $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$ and for any sequence of pairwise disjoint events $A_1, A_2, \ldots$,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i) \qquad \text{(countable additivity).}$$

The pair $(\Omega, \mathbb{P})$ is called a **probability space**.

Theorem 2.2 (Properties of probability measures)
Probability measures have the following properties.

1. Complementarity: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. Monotonicity: if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
3. Addition rule: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof:

1. Because $A \cup A^c = \Omega$ is a disjoint union and $\mathbb{P}(\Omega) = 1$, it follows by additivity that
$$1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c).$$
2. Let $A \subseteq B$. We can write B as the disjoint union $A \cup (B \setminus A)$, so by additivity
$$\mathbb{P}(B) = \mathbb{P}\big[A \cup (B \setminus A)\big] = \mathbb{P}(A) + \mathbb{P}(B \setminus A).$$
Thus because $\mathbb{P}(B \setminus A) \geq 0$ we have $\mathbb{P}(B) \geq \mathbb{P}(A)$.
3. We can write the three sets as disjoint unions:
    - $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$
    - $A = (A \setminus B) + (A \cap B)$
    - $B = (B \setminus A) + (A \cap B)$

    By additivity,

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

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

Theorem 2.3 (Continuity of probability measures)
Let $\mathbb{P}$ be a probability measure on Ω.

1. If $A_1 \subseteq A_2 \subseteq \ldots$ is an increasing sequence of subsets then $\mathbb{P}\left(\bigcup_{n=1}^\infty A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n)$.
2. If $B_1 \supseteq B_2 \supseteq \ldots$ is a decreasing sequence of subsets then $\mathbb{P}\left(\bigcap_{n=1}^\infty B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n)$.

Proof:

1. Let $A = \bigcup_{n=1}^\infty A_n$. We can write A as a disjoint union:
$$A = A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \cup \ldots$$
Because the sets $A_{n+1} \setminus A_n$ are disjoint, by countable additivity we have
$$\mathbb{P}(A) = \mathbb{P}(A_1) + \mathbb{P}(A_2 \setminus A_1) + \mathbb{P}(A_3 \setminus A_2) + \ldots \tag{*}$$
Furthermore, $A_n \subseteq A_{n+1}$ means that $A_{n+1} = (A_{n+1} \setminus A_n) \cup A_n$ is a disjoint union, so
$$\mathbb{P}(A_{n+1} \setminus A_n) = \mathbb{P}(A_{n+1}) - \mathbb{P}(A_n).$$
Substituting this into Eq. (*),
$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A_1) + \big[\mathbb{P}(A_2) - \mathbb{P}(A_1)\big] + \big[\mathbb{P}(A_3) - \mathbb{P}(A_2)\big] + \ldots \\ &= \big[\mathbb{P}(A_1) - \mathbb{P}(A_1)\big] + \big[\mathbb{P}(A_2) - \mathbb{P}(A_2)\big] + \big[\mathbb{P}(A_3) - \mathbb{P}(A_3)\big] + \ldots \\ &= \lim_{n\to\infty} \mathbb{P}(A_n).\end{aligned}$$
2. Let $B = \bigcap_{i=1}^\infty B_i$. Then $B_1^c \subseteq B_2^c \subseteq \ldots$ and by De Morgan's laws,
$$B^c = \bigcup_{n=1}^\infty B_n^c.$$
By the first part of the theorem, $\mathbb{P}(B^c) = \lim_{n\to\infty} \mathbb{P}(B_n^c)$, so
$$\mathbb{P}(B) = 1 - \mathbb{P}(B^c) = 1 - \lim_{n\to\infty} \mathbb{P}(B_n^c) = \lim_{n\to\infty} [1 - \mathbb{P}(B_n^c)] = \lim_{n\to\infty} \mathbb{P}(B_n)$$
as required.

Example 2.4
A fair coin is tossed repeatedly. Show that a head eventually occurs with probability one.

Solution: Let A_n be the event that a head occurs during the first n tosses, and let A be the event that a head eventually occurs. Then $A_1 \subset A_2 \subset A_3, \ldots$ is an increasing sequence with
$$A = \bigcup_{n=1}^\infty A_n.$$
By the continuity property of probability measures,
$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{n=1}^\infty A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty}\left(1 - \frac{1}{2^n}\right) = 1,$$
where we have assumed that the tosses are independent, so that
$$\mathbb{P}(A_n) = 1 - \mathbb{P}(\text{no heads in the first } n \text{ tosses}) = 1 - (1/2)^n.$$

Exercise 2.5

1. (a) For any finite collection of events $A_1, A_2, \ldots, A_n$ (where $n \geq 2$), use proof by induction to show that

$$\mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) + \ldots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n).$$

This is called the **inclusion-exclusion principle**.

Answer: By Theorem 2.2 the result is true for $n = 2$. Let $m \geq 2$ and suppose the result is true for all $n \leq m$. Because Theorem 2.2 result holds for pairs of events it holds for the pair $\cup_{i=1}^n A_i$ and A_{n+1} so

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{m+1} A_i\right) &= \mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right) + \mathbb{P}(A_{m+1}) - \mathbb{P}\left[\left(\bigcup_{i=1}^{m} A_i\right) \cap A_{m+1}\right] \\ &= \mathbb{P}\left(\bigcup_{i=1}^{m} A_i\right) + \mathbb{P}(A_{m+1}) - \mathbb{P}\left[\bigcup_{i=1}^{m} (A_i \cap A_{m+1})\right].\end{aligned}$$

By the inductive hypothesis, we can expand the first and last terms on the right-hand side to obtain the result.

(b) Suppose that at least one of the events $A_1, A_2, \ldots, A_n$ is certain to occur, and that more than two will certainly not occur. If $\mathbb{P}(A_i) = p$ and $\mathbb{P}(A_i \cap A_j) = q$ for $i \neq j$, use the inclusion-exclusion principle to show that $p \geq 1/n$ and $q \leq 2/n$.

Answer: We know that $\mathbb{P}(\cup_{i=1}^n A_i) = 1$ and $\mathbb{P}(A_i \cap A_j \cap A_k) = 0$ whenever $i \neq j \neq k$. By the inclusion-exclusion principle,

$$1 = \mathbb{P}\left(\bigcup_{i=1}^n A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) = np - \frac{1}{2}n(n-1)q.$$

The second term is negative so we must have $np \geq 1$, which shows that $p \geq 1/n$. Hence

$$\frac{1}{2}n(n-1)q = np - 1 \leq n - 1, \text{ so } \frac{nq}{2} \leq 1 \text{ and thus } q \leq 2/n, \text{ as required.}$$

2. For any countable collection of events $A_1, A_2, \ldots$ show that

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

This property is called (countable) **subadditivity**.

Answer: We will find a sequence of sets $B_1, B_2, \ldots$ whose union coincides with the union of $A_1, A_2, \ldots$, but which are disjoint so that the countable additivity property of probability measures can be applied.

Let $B_1 = A_1$, $B_2 = A_2 \setminus A_1$, $B_3 = A_3 \setminus (A_1 \cup A_2)$, and in general

$$B_i = A_i \setminus \left(\bigcup_{j=1}^{i-1} A_j\right).$$

Claim 1: $\cup_{i=1}^\infty A_i = \cup_{i=1}^\infty B_i$. To see this, let $\omega \in \cup_{i=1}^\infty B_i$. Then there exists some B_j with $\omega \in B_j$, which implies that $\omega \in A_j$ and hence $\omega \in \cup_{i=1}^\infty A_i$. Thus $\cup_{i=1}^\infty B_i \subseteq \cup_{i=1}^\infty A_i$. Conversely, let $\omega \in \cup_{i=1}^\infty A_i$. Then there exists some A_j with $\omega \in A_j$, which implies that $\omega \in \cup_{i=1}^j B_i$ and hence $\omega \in \cup_{i=1}^\infty B_i$. Thus $\cup_{i=1}^\infty A_i \subseteq \cup_{i=1}^\infty B_i$.

- This means that $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\cup_{i=1}^{\infty} B_i\right)$.

Claim 2: $B_i \subseteq A_i$. This follows because

$$B_i = A_i \setminus \left(\bigcup_{j=1}^{i-1} A_j\right) = A_i \cap \left(\bigcup_{j=1}^{i-1} A_j\right)^c \subseteq A_i.$$

By the monotonicity of probability measures, this means that $\mathbb{P}(B_i) \leq \mathbb{P}(A_i)$ for all $i = 1, 2, \ldots$.

Claim 3: the sets $B_1, B_2, \ldots$ are pairwise disjoint. To see this, consider any two sets B_i and B_j and suppose, without loss of generality, that $i < j$. Then for any $\omega \in B_i$ it follows that $\omega \in A_i$, so $\omega \notin B_j$ (because $i < j$, and B_j excludes all outcomes contained in $A_1, A_2, \ldots, A_{j-1}$). Hence B_i and B_j are disjoint.

Thus, by the countable additivity of probability measures,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i),$$

as required.

3. A fair coin is tossed repeatedly. Using the continuity of probability measures show that

(a) a head eventually occurs with probability one,

Answer: Let B_n be the event that no heads occur in the first n tosses, and let B be the event that no heads occur at all. Then $B_1 \supseteq B_2 \supset \ldots$ is a decreasing sequence and $B = \cap_{i=1}^{\infty} B_n$. THence by continuity,

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \left(\frac{1}{2}\right)^n = 0,$$

Thus we are certain of eventually observing a head.

(b) a sequence of 10 consecutive tails eventually occurs with probability one, and

Answer: Let us think of the first $10n$ tosses as disjoint groups of consecutive outcomes, each group of length 10. The probability any one of the n groups consists of 10 consecutive tails is 2^{-10}, independently of the other groups. The event that one of the groups consists of 10 consecutive tails is a subset of the event that a sequence of 10 consecutive tails appears anywhere in the first $10n$ tosses. Hence, using the continuity of probability measures,

$$\begin{aligned}\mathbb{P}(10T \text{ eventually appears}) &= \lim_{n\to\infty} \mathbb{P}(10T \text{ occurs somewhere in the first } 10n \text{ tosses}) \\ &\geq \lim_{n\to\infty} \mathbb{P}(10T \text{ occurs as one of the first } n \text{ groups of } 10) \\ &= 1 - \lim_{n\to\infty} \mathbb{P}(10T \text{ does not occur as one of the first } n \text{ groups of } 10) \\ &= 1 - \lim_{n\to\infty} \left(1 - \frac{1}{2^{10}}\right)^n = 1.\end{aligned}$$

Thus we are certain of eventually observing sequence of 10 consecutive tails.

(c) any finite sequence of heads and tails eventually occurs with probability one.

Answer: Let s be a fixed sequence of length k. As in the previous part, we think of the first kn tosses as n distinct groups of length k. The event that the one of these groups is exactly equal to s is a subset of the event that first kn tosses contains at least one instance of s. Hence

$$\begin{aligned}
\mathbb{P}(s \text{ eventually appears}) &= \lim_{n\to\infty} \mathbb{P}(s \text{ occurs somewhere in the first } kn \text{ tosses}) \\
&\geq \lim_{n\to\infty} \mathbb{P}(s \text{ occurs as one of the first } n \text{ groups of } k) \\
&= 1 - \lim_{n\to\infty} \mathbb{P}(s \text{ does not occur as one of the first } n \text{ groups of } k) \\
&= 1 - \lim_{n\to\infty} \left(1 - \frac{1}{2^k}\right)^n = 1.
\end{aligned}$$

Thus we are certain of eventually observing the sequence s.

- In an infinite sequence of coin tosses, anything that can happen, does happen!

2.2 Conditional probability

Recall from elementary probability that the conditional probability of A given B is defined by

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

provided that $P(B) > 0$. The following theorem shows that this defines a probability measure on subsets of Ω.

Theorem 2.6
Let $\mathbb{P}$ be a probability measure on subsets of Ω and let B be an event with $\mathbb{P}(B) > 0$. Then the function

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

is also probability measure on subsets of Ω.

Proof: To show that $\mathbb{Q}$ is a probability measure, we need to verify that $\mathbb{Q}(\emptyset) = 1$, $\mathbb{Q}(\Omega) = 1$ and that $\mathbb{Q}$ is countably additive. Firstly,

$$\begin{aligned}
\mathbb{Q}(\emptyset) &= \mathbb{P}(\emptyset|B) = \mathbb{P}(\emptyset \cap B)/\mathbb{P}(B) = \mathbb{P}(\emptyset)/\mathbb{P}(B) = 0; \\
\mathbb{Q}(\Omega) &= \mathbb{P}(\Omega|B) = \mathbb{P}(\Omega \cap B)/\mathbb{P}(B) = \mathbb{P}(B)/\mathbb{P}(B) = 1.
\end{aligned}$$

To prove that $\mathbb{Q}$ is countable additive, let $A_1, A_2, \ldots$ be pairwise disjoint subsets of Ω. Then using the fact that $\mathbb{P}$ is countably additive,

$$\begin{aligned}
\mathbb{Q}(\textstyle\bigcup_{i=1}^\infty A_i) = \mathbb{P}(\textstyle\bigcup_{i=1}^\infty A_i \,|\, B) &= \frac{\mathbb{P}\big[(\bigcup_{i=1}^\infty A_i) \cap B\big]}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}\big[\bigcup_{i=1}^\infty (A_i \cap B)\big]}{\mathbb{P}(B)} \quad \text{(because intersection is distributive over union),} \\
&= \frac{\sum_{i=1}^\infty \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \quad \text{(because the sets } A_i \cap B \text{ are disjoint),} \\
&= \sum_{i=1}^\infty \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^\infty \mathbb{P}(A_i|B) = \sum_{i=1}^\infty \mathbb{Q}(A_i) \quad \text{as required.}
\end{aligned}$$

Chapter 3 Random variables

3.1 Random variables

Let $(\Omega, \mathbb{P})$ be a probability space associated with some random experiment. Random experiments with numerical outcomes lend themselves to mathematical analysis; for more abstract sample spaces we introduce the notion of **random variables**, which map sample spaces to the real numbers. Random variables are typically represented by uppercase letters.

Definition 3.1
A **random variable** is a function which maps a sample space Ω to the real numbers,

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

We are often not directly interested in the outcome of a random experiment, but rather in some consequence of the outcome: a gambler might be more interested in his losses than in the outcomes of the individual games which led to them. Random variables can be used to pick out particular features of an experiment that are of interest.

Example 3.2
A fair coin is tossed until a head is observed. Suppose we win £3 if the coin is tossed an odd number of times and lose £5 if the coin is tossed an even number of times. The sample space is the countably infinite set $\Omega = \{H, TH, TTH, TTTH, \ldots\}$. We win if event $A = \{H, TTH, TTTTH, \ldots\}$ occurs and lose if event $A^c = \{TH, TTTH, TTTTH, \ldots\}$ occurs. Because the coin is fair,

$$\begin{array}{lll} \mathbb{P}(A) & = \frac{1}{2} + \frac{1}{8} + \frac{1}{32} + \ldots & = \frac{2}{3}, \\ \mathbb{P}(A^c) & = \frac{1}{4} + \frac{1}{16} + \frac{1}{64} + \ldots & = \frac{1}{3}. \end{array}$$

Our situation can be represented more concisely by the random variable $X : \Omega \to \mathbb{R}$ defined by

$$X(\omega) = \begin{cases} +3 & \text{if } \omega \in \{H, TTH, TTTTH, \ldots\}, \\ -5 & \text{if } \omega \in \{TH, TTTH, TTTTTH, \ldots\}. \end{cases}$$

From here it follows that $\mathbb{P}(X = 3) = 2/3$ and $\mathbb{P}(X = -5) = 1/3$.

3.1.1 Distributions

Let B be a subset of $\mathbb{R}$. We use the notation $\{X \in B\}$ as shortand for the event $\{\omega : X(\omega) \in B\}$ and $\mathbb{P}(X \in B)$ as shorthand for the probability of this event. Thus $\{X \in B\}$ is the event consisting of those outcomes that are mapped by X into B, and $\mathbb{P}(X \in B)$ is therefore the probability that X takes a value in the set B.

Definition 3.3
The **distribution** of X is the function $\mathbb{P}_X$ defined on subsets of $\mathbb{R}$ by

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

Theorem 3.4
$\mathbb{P}_X$ is a probability measure on subsets of $\mathbb{R}$.

Proof: First we check that $\mathbb{P}_X(\mathbb{R}) = 1$:

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

Next we show that $\mathbb{P}_X$ is countably additive. Let $B_1, B_2, \ldots$ be a sequence of pairwise disjoint subsets of $\mathbb{R}$. Then

$$\begin{aligned}
\mathbb{P}_X\big(\textstyle\bigcup_{i=1}^\infty B_i\big) &= \mathbb{P}\big(\{\omega : X(\omega) \in \textstyle\bigcup_{i=1}^\infty B_i\}\big) \\
&= \mathbb{P}\big(\textstyle\bigcup_{i=1}^\infty \{\omega : X(\omega) \in B_i\}\big) \\
&= \sum_{i=1}^\infty \mathbb{P}\big(\{\omega : X(\omega) \in B_i\}\big) \quad \text{because the } B_i \text{ are disjoint,} \\
&= \sum_{i=1}^\infty \mathbb{P}_X(B_i),
\end{aligned}$$

as required.

3.1.2 Indicator variables

Every event can be represented by its **indicator** variable, which *indicates* whether or not the event occurs. Indicator variables provide an explicit link between random events and random variables.

Definition 3.5
The **indicator variable** of an event A is the function $I_A : \Omega \to \mathbb{R}$ defined by

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Exercise 3.6
Let A and B be any two events. Show that $I_{A^c} = 1 - I_A$, $I_{A\cap B} = I_A I_B$ and $I_{A\cup B} = I_A + I_B - I_{A\cap B}$. Note that two functions are equal if and only if they are equal at every point.

3.2 CDFs

A probability distribution $\mathbb{P}_X$ is uniquely determined by the values it takes on intervals of the form $(-\infty, x]$, and hence by its **cumulative distribution function** (CDF).

Definition 3.7
The CDF of a random variable X is the function $F : \mathbb{R} \to [0,1]$ given by

$$F(x) = \mathbb{P}(X \le x).$$

We can use the properties of probability measures to derive some properties of CDFs.

Theorem 3.8 (Properties of CDFs)
Let $F : \mathbb{R} \to [0,1]$ be a CDF. Then

1. F is an increasing function,
2. $F(x) \to 0$ as $x \to -\infty$,

3. $F(x) \to 1$ as $x \to +\infty$, and
4. $F(x+h) \to F(x)$ as $h \downarrow 0$ (right continuity).

Proof: Let $X : \Omega \to \mathbb{R}$ be a random variable and let F be its CDF.

1. To show that F is increasing, let $x < y$ and consider the events $A = \{X \leq x\}$ and $B = \{X \leq x'\}$, i.e.
$$A = \{\omega : X(\omega) \leq x\} \quad \text{and} \quad B = \{\omega : X(\omega) \leq x'\}.$$
By construction, $F(x) = \mathbb{P}(A)$ and $F(x') = \mathbb{P}(B)$, and because $x < x'$ we have $A \subseteq B$. By the monotonicity of probability measures, $\mathbb{P}(A) \leq \mathbb{P}(B)$ or equivalently $F(x) \leq F(x')$, so F is an increasing function.

2. To show that $F(x) \to 0$ as $x \to -\infty$, let $B_n = \{X \leq -n\}$ for $n = 1, 2, \ldots$. Then $F(-n) = \mathbb{P}(B_n)$, and $B_1, B_2, \ldots$ is a decreasing sequence ($B_{n+1} \subseteq B_n$), with
$$\bigcap_{n=1}^{\infty} B_n = \emptyset.$$
(This is because for any x, there exists an n such that $x \notin (-\infty, -n]$.) By the continuity of probability measures,
$$\lim_{n\to\infty} F(-n) = \lim_{n\to\infty} \mathbb{P}(B_n) = \mathbb{P}\left(\bigcap_{n=1}^{n} B_n\right) = \mathbb{P}(\emptyset) = 0.$$
Because $F(x)$ is an increasing function, we conclude that $\lim_{x\to-\infty} F(x) = 0$.

3. To show that $F(x) \to 1$ as $x \to \infty$, let $A_n = \{X \leq n\}$ for $n = 1, 2, \ldots$. Then $F(n) = \mathbb{P}(A_n)$, and $A_1, A_2, \ldots$ is an increasing sequence ($A_n \subseteq A_{n+1}$), with
$$\bigcup_{n=1}^{\infty} A_n = \Omega.$$
(This is because for any x, there exists an n such that $x \in (-\infty, n]$). By the continuity of probability measures,
$$\lim_{n\to\infty} F(n) = \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \mathbb{P}(\Omega) = 1,$$
Because $F(x)$ is an increasing function, we conclude that $\lim_{x\to\infty} F(x) = 1$.

4. To show that $F(x)$ is right-continuous, let $B_n = \{X \leq x + 1/n\}$ for $n = 1, 2, \ldots$. Then $F\left(x + 1/n\right) = \mathbb{P}(B_n)$, and $B_1, B_2, \ldots$ is a decreasing sequence ($B_{n+1} \subseteq B_n$), with
$$\bigcap_{n=1}^{\infty} B_n = (-\infty, x].$$
By the continuity of probability measures,
$$F(x) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} F\left(x + \frac{1}{n}\right).$$
Because $F(x)$ is an increasing function, we conclude that $\lim_{h\downarrow 0} F(x+h) = F(x)$.

Exercise 3.9
It can be shown that any function $F : \mathbb{R} \to [0,1]$ which has the four properties listed in Theorem 3.8 is a CDF. With this in mind, show that if F and G are CDFs and $0 < \lambda < 1$ is a constant, then $\lambda F + (1-\lambda)G$ is also a CDF.

Answer: Let $H(x) = \lambda F(x) + (1-\lambda)G(x)$.

1. if $x < x'$ then
$$H(x) = \lambda F(x) + (1-\lambda)G(x) \leq \lambda F(x') + (1-\lambda)G(x') = H(x').$$

2. $$\lim_{x\to-\infty} H(x) = \lim_{x\to-\infty}[\lambda F(x) + (1-\lambda)G(x)] = \lambda \lim_{x\to-\infty} F(x) + (1-\lambda)\lim_{x\to-\infty} G(x) = 0.$$

3. $$\lim_{x\to\infty} H(x) = \lim_{x\to\infty}[\lambda F(x) + (1-\lambda)G(x)] = \lambda \lim_{x\to\infty} F(x) + (1-\lambda)\lim_{x\to\infty} G(x) = \lambda + (1-\lambda) = 1.$$

4. $$\begin{aligned}\lim_{\epsilon\downarrow 0} H(x+\epsilon) &= \lim_{\epsilon\downarrow 0}[\lambda F(x+\epsilon) + (1-\lambda)G(x+\epsilon)] \\ &= \lambda \lim_{\epsilon\downarrow 0} F(x+\epsilon) + (1-\lambda)\lim_{\epsilon\downarrow 0} G(x+\epsilon) \\ &= \lambda F(x) + (1-\lambda)G(x) = H(x).\end{aligned}$$

Thus H satisfies the four properties of Theorem 3.8, and is therefore a CDF.

3.3 PMFs and PDFs

Definition 3.10
A random variable is called

1. **simple** if it takes only finitely many values;
2. **discrete** if it takes only countably many values;
3. **continous** if it takes uncountably many values (and satisfies some other conditions).

Because finte sets are countable, simple random variables are also discrete random variables.

3.3.1 Discrete distributions

Discrete random variables are completely described by their PMFs.

Definition 3.11
The **probability mass function** (PMF) of a discrete random variable X is a function

$$\begin{array}{rccl} f: & \mathbb{R} & \to & [0,1] \\ & x & \mapsto & \mathbb{P}(X = x), \end{array}$$

with the property that $\sum_{i=1}^{\infty} x_i f(x_i) = 1$, where $\{x_1, x_2, \ldots\}$ is the range of X.

Example 3.12
Two fair dice are rolled independently. Each die has two faces coloured red, two coloured blue and two coloured green. Suppose we win £10 if both dice show the same colour but lose £5 if they show different colours. Let X be the amount won.

- The sample space can be expressed as $\Omega = \{RR, RB, RG, BR, BB, BG, GR, GB, GG\}$.
- The random variable $X : \Omega \to \mathbb{R}$ can then be defined by

$$X(\omega) = \begin{cases} -5 & \text{if } \omega \in \{RB, RG, BR, BG, GR, GB\}, \\ 10 & \text{if } \omega \in \{RR, BB, GG\}. \end{cases}$$

- The range of X is the set $\{-5, 10\}$.
- The PMF and CDF of X are the functions

$$f(x) = \begin{cases} 2/3 & \text{if } x = -5 \\ 1/3 & \text{if } x = 10 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0 & \text{if } x < -5 \\ 2/3 & \text{if } -5 \leq x < 10 \\ 1 & \text{if } x \geq 10 \end{cases} \quad \text{respectively.}$$

Fundamental discrete distributions

A simple random variable taking all values in its range with equal probability is said to have a (discrete) **uniform** distribution. For example, if X represents the score on a fair die, we say that X is uniformly distributed over the set $\{1, 2, 3, 4, 5, 6\}$.

The **Bernoulli** distribution is the distribution of an indicator variable. For example, suppose a coin has probability p of landing on heads. If X is the indicator variable of this event we say that X has the Bernoullii(p) distribution.

Three other fundamental discrete distributions are derived from the following convergent series,

$$\begin{aligned} &\sum_{k=0}^{n} p^k (1-p)^{n-k} = 1 && \text{(binomial theorem)}, \\ &\sum_{k=0}^{\infty} (1-p)^k = 1/p \quad \text{for } |p| < 1 && \text{(geometric series)}, \\ &\sum_{k=0}^{\infty} \lambda^k / k! = e^{\lambda} && \text{(exponential function)}. \end{aligned}$$

which yield the binomial, geometric, and Poisson distributions respectively.

Distribution	PMF	Range
$X \sim$ Uniform(n)	$1/n$	$\{1, 2, \ldots, n\}$
$X \sim$ Bernoulli(p)	$p^x(1-p)^{1-x}$	$\{0, 1\}$
$X \sim$ Binomial(n, p)	$\binom{n}{x} p^x (1-p)^{n-x}$	$\{0, 1, \ldots, n\}$
$X \sim$ Geometric(p)	$(1-p)^x p$	$\{0, 1, \ldots\}$
$X \sim$ Poisson(λ)	$\lambda^x e^{-\lambda}/x!$	$\{0, 1, \ldots\}$

Table 3.1: Fundamental discrete distributions

Exercise 3.13
For each distribution shown in the Table 3.1 verify that the expressions given for their PMFs are indeed PMFs.

3.3.2 Continuous distributions

Consider a random experiment which involves measuring the lifetime of a lightbulb. Let X be the time elapsed between the start of the experiment and the point at which the lightbulb fails. Then X can take any value in the non-negative real numbers $[0,\infty)$, which is an **uncountable** set. The distribution of X cannot therefore be specified by a probability mass function. Subject to a condition on its CDF however, the distribution of X can be specified by a **probability density function**.

Definition 3.14
Let X be a random variable and let F denote its CDF. If there exists an integrable function $f:\mathbb{R}\to[0,\infty)$ such that

$$F(x)=\int_{-\infty}^{x} f(t)\,dt \quad \text{for all} \quad x\in\mathbb{R},$$

then X is called a **continuous random variable** and f is called its **probability density function** (PDF).

By the second fundamental theorem of calculus, $f(x)=F'(x)$. Note also that if X is a continuous random variable then $\mathbb{P}(X=x)=0$ for any $x\in\mathbb{R}$ (this is analagous to saying that the length of a point is zero). The following result is sometimes called the **law of total probability**.

Proposition 3.15
For any PDF,

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

Proof: By Theorem 3.8,

$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{x\to\infty}\int_{-\infty}^{x} f(t)\,dt = \lim_{x\to\infty} F(x) = 1.$$

Fundamental continuous distributions

The fundamental continuous distributions are derived from the following definite inegrals,

$$\int_0^1 dx = 1, \qquad \int_0^{\infty} e^{-x}\,dx = 1, \qquad \int_{-\infty}^{\infty} e^{-x^2} = \sqrt{\pi},$$

which yield the standard uniform, exponential and normal distributions respectively. These standard distributions can be scaled and shifted to yield the parameterized families of distributions shown in Table 3.2.

Distribution	PDF	Range
Uniform(a,b)	$1/(b-a)$	$[a,b]$
Exponential(λ)	$\lambda e^{-\lambda x}$	$[0,\infty)$
Normal(μ,σ^2)	$\exp\big[-(x-\mu)^2/2\sigma^2\big]/\sigma\sqrt{2\pi}$	$(-\infty,\infty)$

Table 3.2: The fundamental continuous distributions

Two other notable continuous distributions are based on the following definite integrals,

$$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt \qquad \text{and} \qquad \Gamma(\alpha)=\int_0^{\infty} t^{\alpha-1}e^{-t}\,dt.$$

These are special functions known the **beta function** and the **gamma function** respectively, and have been widely studied. Unsurprisingly the corresponding distriubtions are known as the beta distribution and the gamma distributions, as shown in Table 3.3.

Distribution	PDF	Range
Beta(α,β)	$x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$	$[0,1]$
Gamma(α,β)	$\beta^\alpha x^{\alpha-1}e^{-\beta x}/\Gamma(\alpha)$	$[0,\infty)$

Table 3.3: The beta and gamma distributions

Exercise 3.16
For each distribution shown in Tables 3.2 and 3.3 verify that the expressions given for their PDFs are indeed PDFs.

Exercise 3.17

1. The PDF of a continuous random variable X is given by

$$f(x) = \begin{cases} cx^2 & 1 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the value of the constant c, and sketch the PDF of X.
Answer: The PDF must integrate to 1:

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_1^2 cx^2\,dx = \left[\frac{cx^3}{3}\right]_1^2 = \frac{7c}{3} = 1$$

so $c = 3/7$. (The sketch is a quadratic curve between $x=1$ and $x=2$.)

(b) Show that $\mathbb{P}(X > 3/2) = 37/56$.
Answer:

$$\mathbb{P}(X > 3/2) = \int_{3/2}^2 \frac{3x^2}{7}\,dx = \left[\frac{x^3}{7}\right]_{3/2}^2 = \frac{37}{56}$$

(c) Find the CDF of X.
Answer: For $1 \leq x \leq 2$,

$$F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_1^x \frac{3x^2}{7}\,dx = \left[\frac{x^3}{7}\right]_1^x = \frac{x^3-1}{7}$$

so the CDF of X is

$$F(x) = \begin{cases} 0 & x < 1 \\ (x^3-1)/7 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

2. Consider the function

$$f(x) = \begin{cases} c/x^d & x > 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Why can f be a PDF only when $d > 1$?
Answer: The function $f(x) = c/x^d$ is only integrable when $d > 1$. in which case

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_1^{\infty} \frac{c}{x^d}\,dx = \left[\frac{-c}{(d-1)x^{d-1}}\right]_1^{\infty} = \frac{c}{d-1}$$

(b) If $d > 1$, find the value of c and the corresponding CDF.
Answer: If $d > 1$,

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_1^{\infty} \frac{c}{x^d}\,dx = \left[\frac{-c}{(d-1)x^{d-1}}\right]_1^{\infty} = \frac{c}{d-1}$$

If f is a PDF, we need that $\int_{-\infty}^{\infty} f(x)\,dx = 1$, so we must have that $c = d - 1$. The corresponding CDF is

$$F(x) = \int_{-\infty}^{x} f(u)\,du = \int_1^{\infty} \frac{d-1}{u^d}\,du = \left[\frac{-1}{x^{d-1}}\right]_1^{x} = 1 - \frac{1}{x^{d-1}}$$

for $x > 1$, and zero otherwise.

3.4 Transformations

Random variables can be transformed into other random variables, and the distribution of a transformed variable can be deduced from the distribution of the original variable. Transformations of discrete distributions are relatively straightforward: here we focus on transformations of continuous distributions.

Applying a transformation $g : \mathbb{R} \to \mathbb{R}$ to a random variable $X : \Omega \to \mathbb{R}$ involves the **composition** of these two functions,

$$\begin{aligned} g(X) : \quad \Omega &\to \mathbb{R} \\ \omega &\mapsto g\big[X(\omega)\big], \end{aligned}$$

This can be interpreted in two ways:

1. $g(X)$ is a random variable on the probability space $(\Omega, \mathbb{P})$;
2. g is a random variable on the probability space $(\mathbb{R}, \mathbb{P}_X)$,

where $\mathbb{P}_X$ is the distribution of X. Here we focus on the first interpretation with $Y = g(X)$ denoting the transformed variable.

3.4.1 Support

Many PDFs are non-zero only over certain subsets of $\mathbb{R}$. When we transform continuous distributions we need only consider these subsets, provided we ensure that those over which PDFs are zero are carried over correctly into the transformed space. For technical reasons, it is not quite enough to focus on the **range** of the random variable in question: instead we must consider the smallest closed set that contains the range (a closed set is one that contains all its limit points). This is called the **support** of the associated PDF.

Definition 3.18
The **support** of an arbitrary function $h : \mathbb{R} \to \mathbb{R}$, denoted by $\text{supp}(h)$, is the smallest closed set for which $h(x) = 0$ for all $x \notin \text{supp}(h)$.

Let $Y = g(X)$ where $g : \mathbb{R} \to \mathbb{R}$ is a transformation, and let f_X and f_Y denote the PMFs/PDFs of X and Y respectively. The support of the transformed variable $Y = g(X)$ is given by

$$\operatorname{supp}(f_Y) = \big\{g(x) : x \in \operatorname{supp}(f_X)\big\}.$$

In fact, $\operatorname{supp}(f_Y)$ should be defined as the **closure** of this set, which is obtained by adding in its limit points where necessary. We will not pursue such matters here.

3.4.2 Linear transformations

Let X be a random variable and let $Y = aX + b$ where $a \neq 0$.

The CDF of of the transformed variable Y can be expressed in terms of the CDF of X as follows:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) = \begin{cases} \mathbb{P}\left(X \leq \dfrac{y-b}{a}\right) = \quad F_X\left(\dfrac{y-b}{a}\right) & \text{if } a > 0, \\ \mathbb{P}\left(X > \dfrac{y-b}{a}\right) = 1 - F_X\left(\dfrac{y-b}{a}\right) & \text{if } a < 0. \end{cases}$$

Using the chain rule, the PDF of Y can then be expressed in terms of the PDF of X:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \left\{\begin{array}{ll} \dfrac{1}{a}f_X\left(\dfrac{y-b}{a}\right) & \text{if } a > 0, \\ -\dfrac{1}{a}f_X\left(\dfrac{y-b}{a}\right) & \text{if } a < 0. \end{array}\right\} = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right).$$

Example 3.19
Let $X \sim \text{Uniform}[0,1]$. Find the distribution of the random variable $Y = 3X + 7$.

Solution: Let $g(x) = 3x + 7$ denote the transformation. The PDF of $X \sim \text{Uniform}[0,1]$ is

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

First we see that $\operatorname{supp}(f_X)$ is transformed as follows:

$$\operatorname{supp}(f_Y) = \{g(x) : x \in \operatorname{supp}(f_X)\} = \{3x + 7 : x \in [0,1]\} = [7, 10].$$

From the above discussion,

$$f_Y(y) = \frac{1}{3}f_X\left(\frac{y-7}{3}\right) = \begin{cases} 1/3 & \text{if } 7 \leq y \leq 10, \\ 0 & \text{otherwise,} \end{cases}$$

so $Y \sim \text{Uniform}[7, 10]$. We see that the original distribution has been scaled by a factor of 3 and shifted 7 units to the right.

These ideas can be extended to any one-to-one transformation of X.

3.4.3 Transformations of CDFs

Theorem 3.20
If $g : \mathbb{R} \to \mathbb{R}$ is one-to-one over $\operatorname{supp}(f_X)$ the CDF of $Y = g(X)$ is

$$F_Y(y) = \begin{cases} F_X\left[g^{-1}(y)\right] & \text{if } g \text{ is increasing, and} \\ 1 - F_X\left[g^{-1}(y)\right] & \text{if } g \text{ is decreasing.} \end{cases}$$

Proof:

1. If g is increasing, $g(x) \leq y$ implies that $x \leq g^{-1}(y)$ so

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\big[g(X) \leq y\big] = \mathbb{P}\big[X \leq g^{-1}(y)\big] = F_X\big[g^{-1}(y)\big].$$

2. If g is decreasing, $g(x) \leq y$ implies that $x \geq g^{-1}(y)$ so

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}\big[g(X) \leq y\big] &= \mathbb{P}\big[X \geq g^{-1}(y)\big] \\ &= 1 - \mathbb{P}\big[X \leq g^{-1}(y)\big] \quad \text{(because X is a continuous r.v.)} \\ &= 1 - F_X\big[g^{-1}(y)\big]. \end{aligned}$$

Example 3.21
Let $X \sim \text{Uniform}[0,1]$ and let $Y = -\dfrac{1}{\lambda}\log X$ where $\lambda > 0$. Show that $Y \sim \text{Exponential}(\lambda)$ where λ is a rate parameter.

Solution: The CDF and PDF of $X \sim \text{Uniform}[0,1]$ are, respectively,

$$F_X(x) = \begin{cases} 0 & x < 0, \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases} \quad \text{and} \quad f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let $g(x) = -\log x/\lambda$ denote the transformation. Because $\lambda > 0$ we see that $\text{supp}(f_X)$ is transformed to

$$\text{supp}(f_Y) = \{g(x) : x \in \text{supp}(f_X)\} = \{-\log/\lambda : x \in [0,1]\} = [0,\infty).$$

The transformation is strictly decreasing over $\text{supp}(f_X)$, and its inverse is $g^{-1}(y) = e^{-\lambda y}$ over $\text{supp}(f_Y)$. Hence, by Theorem 3.20,

$$F_Y(y) = 1 - F_X\big[g^{-1}(y)\big] = 1 - F_X(e^{-\lambda y}) = 1 - e^{-\lambda y} \quad \text{for} \quad y \geq 0.$$

This is the CDF of the Exponential(λ) distribution (where λ is a rate parameter).

3.4.4 Transformations of PDFs

Theorem 3.22
If $g : \mathbb{R} \to \mathbb{R}$ is one-to-one over $\text{supp}(f_X)$ the PDF of $Y = g(X)$ is

$$f_Y(y) = f_X\big[g^{-1}(y)\big]\left|\frac{d}{dy}g^{-1}(y)\right|$$

Proof: For clarity of notation, let $h(y)$ denote the inverse function $g^{-1}(y)$.

1. If g is increasing then $F_Y(y) = F_X\big[g^{-1}(y)\big]$, so by the chain rule (and using the fact that h is also increasing),

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y(y) &= \frac{d}{dy}F_X\big[h(y)\big] \\ &= \frac{d}{dh(y)}F_X\big[h(y)\big] \cdot \frac{dh(y)}{dy} \\ &= f_X\big[h(y)\big]\left|\frac{dh(y)}{dy}\right|, \qquad \text{because } \frac{dh(y)}{dy} > 0 \text{ over } \text{supp}(f_Y). \end{aligned}$$

2. If g decreasing, $F_Y(y) = 1 - F_X\big[g^{-1}(y)\big]$, so by the chain rule (and using the fact that h is also decreasing),

$$\begin{aligned}
f_Y(y) = \frac{d}{dy}F_Y(y) &= \frac{d}{dy}\big[1 - F_X[h(y)]\big] \\
&= 0 - \frac{d}{dh(y)}F_X\big[h(y)\big]\cdot\frac{dh(y)}{dy} \\
&= -f_X\big[h(y)\big]\frac{dh(y)}{dy} \\
&= f_X\big[h(y)\big]\left|\frac{dh(y)}{dy}\right|, \qquad \text{because}\frac{dh(y)}{dy} < 0 \text{ over } \operatorname{supp}(f_Y).
\end{aligned}$$

Remark 3.23
The term $\left|\frac{d}{dy}g^{-1}(y)\right|$ in Theorem 3.22 is a **scale factor**, which ensures that f_Y integrates to one.

Example 3.24
Let X be a continuous random variable with the following PDF,

$$f_X(x) = \begin{cases} 1/x^2 & \text{for } x > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the PDF of $Y = 1/X$.

Solution: Let $g(x) = 1/x$. Since $\operatorname{supp}(f_X) = [x, \infty)$, the support of f_Y is

$$\operatorname{supp}(f_Y) = \{g(x) : x \in \operatorname{supp}(f_X)\} = \{1/x : x \in [0,\infty)\} = [0,1].$$

The transformation is one-to-one over $\operatorname{supp}(f_X)$, and its inverse is $g^{-1}(y) = 1/y$ over $\operatorname{supp}(f_Y)$. Hence, by Theorem 3.22, the PDF of Y is given by

$$\begin{aligned}
f_Y(y) &= f_X\big[g^{-1}(y)\big]\left|\frac{d}{dy}g^{-1}(y)\right| \\
&= f_X\left(\frac{1}{y}\right)\left|\frac{d}{dy}\left(\frac{1}{y}\right)\right| \\
&= y^2\left|-\frac{1}{y^2}\right| = \begin{cases} 1 & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Thus $Y \sim \text{Uniform}(0,1)$.

3.4.5 The probability integral transform

What happens when a random variable is transformed using its own CDF?

Theorem 3.25
Let X be a continuous random variable, and suppose that the inverse of its CDF exists for all $x \in \mathbb{R}$. Then the random variable $U = F(X)$ has the uniform distribution on $[0,1]$.

Proof: Because $F(x) = P(X \le x)$ is a CDF we know that $F(x) \in [0,1]$ for all $x \in \mathbb{R}$. In particular, $\mathbb{P}(U < 0) = 0$ and $\mathbb{P}(U > 1) = 0$. For $u \in [0,1]$, because the inverse F^{-1} exists for all $x \in \mathbb{R}$ we have that

$$\begin{aligned}
F_U(u) = P(U \le u) &= P\big(F(X) \le u\big) \\
&= P\big(X \le F^{-1}(u)\big) \\
&= F\big(F^{-1}(u)\big) \\
&= u,
\end{aligned}$$

which is the CDF of the continuous uniform distribution on $[0, 1]$.

Corollary 3.26
Let F be a CDF whose inverse exists for all $x \in \mathbb{R}$, and let $U \sim \text{Uniform}[0, 1]$. Then F is the CDF of the random variable $X = F^{-1}(U)$.

Although it is difficult to generate truly random numbers, there are fast deterministic algorithms which generate numbers that are approximately random - such numbers are called **pseudo-random numbers**. Many such algorithms can generate numbers that are approximately uniformly distributed in $[0, 1]$. Using the probability integral transform we can convert these into pseudo-random numbers from other continuous distributions.

1. First we obtain a uniformly distributed pseudo-random number $u \in [0, 1]$.
2. The number $x = F^{-1}(u)$ is then a pseudo-random number from the distribution F.

Example 3.27
Given an algorithm which generates uniformly distributed pseudo-random numbers in the range $[0, 1]$, show how to obtain a pseudo-random number from the exponential distribution having rate parameter 2.

Solution: The CDF of the exponential distribution with rate parameter 2 is

$$F(x) = \begin{cases} 1 - e^{-2x} & x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

First we invert F:

$$\begin{aligned} u = 1 - e^{-2x} &\iff e^{-2x} = 1 - u \\ &\iff e^{x} = \frac{1}{\sqrt{1-u}} \\ &\iff x = \log\left(\frac{1}{\sqrt{1-u}}\right) \end{aligned}$$

Then we generate a pseudo-random number u from the Uniform$[0, 1]$ distribution and

$$x = \log\left(\frac{1}{\sqrt{1-u}}\right)$$

which is a pseudo-random number from the Exponential(λ) distribution.

Exercise 3.28

1. Let $X \sim \text{Uniform}(-1, 1)$. Find the CDF and PDF of X^2.

   Answer: The PDF of X is

   $$f_X(x) = \begin{cases} 1/2 & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   For $x \in [-1, 1]$,

   $$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt = \int_{-1}^{x} \frac{1}{2}\, dt = \left[\frac{t}{2}\right]_{-1}^{x} = \frac{1}{2}(x + 1).$$

   The CDF of X is:

   $$F(x) = \begin{cases} 0 & x < -1, \\ \frac{1}{2}(x+1) & -1 \leq x \leq 1, \\ 1 & x > 1. \end{cases}$$

Let $Y = X^2$. For $0 \leq y \leq 1$ we have

$$\begin{aligned}\mathbb{P}(Y \leq y) = \mathbb{P}(X^2 \leq y) &= \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= \mathbb{P}(X \leq \sqrt{y}) - \mathbb{P}(X \leq -\sqrt{y}) \\ &= \sqrt{y}.\end{aligned}$$

Hence the CDF of Y is

$$F_Y(y) = \begin{cases} 0 & y < 0, \\ \sqrt{y} & 0 \leq y \leq 1, \\ 1 & y > 1. \end{cases}$$

and the PDF of Y is

$$f_Y(y) = \begin{cases} \frac{1}{2}y^{-1/2} & 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

2. Suppose that X has the exponential distribution with rate parameter $\lambda > 0$. (The PDF of X is $f(x) = \lambda\exp(-\lambda x)$ for $x \geq 0$ and zero otherwise.) Find the PDFs of $Y = X^2$ and $Z = e^X$.

Answer:

1. The transformation $g(x) = x^2$ is one-to-one and increasing over $[0, \infty)$; its inverse function is

$$g^{-1}(y) = \sqrt{y}, \quad \text{which has first derivative} \quad \frac{d}{dy}g^{-1}(y) = \frac{1}{2\sqrt{y}}.$$

Since $\text{supp}(f_X) = [0, \infty)$ it follows immediately that $\text{supp}(f_Y) = [0, \infty)$.
For $y > 0$,

$$f_Y(y) = f_X\left[g^{-1}(y)\right]\left|\frac{d}{dy}g^{-1}(y)\right| = \lambda\exp(\lambda\sqrt{y})\left|\frac{1}{2\sqrt{y}}\right| = \frac{\lambda}{2\sqrt{y}}\exp(-\lambda\sqrt{y}).$$

Hence the PDF of $Y = X^2$ is given by

$$f_Y(y) = \begin{cases} \dfrac{\lambda}{2\sqrt{y}}\exp(-\lambda\sqrt{y}) & y \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

2. The transformation $g(x) = e^x$ is one-to-one and increasing over $[0, \infty)$; its inverse function is

$$g^{-1}(z) = \log y \quad \text{and} \quad \frac{d}{dy}g^{-1}(z) = \frac{1}{z}.$$

Since $\text{supp}(f_X) = [0, \infty)$ it follows immediately that $\text{supp}(f_Z) = [1, \infty)$.
For $z \geq 1$,

$$f_Z(z) = f_X\left[g^{-1}(z)\right]\left|\frac{d}{dz}g^{-1}(z)\right| = \lambda\exp(-\lambda\log z)\left|\frac{1}{z}\right| = \lambda z^{-(\lambda+1)}.$$

Hence the PDF of $Z = e^X$ is given by

$$f_Z(z) = \begin{cases} \lambda z^{-(\lambda+1)} & z \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

3. A continuous random variable U has PDF $f(u) = 12u^2(1-u)$ for $0 < u < 1$ and zero otherwise. Find the PDF of $V = (1-U)^2$.

   Answer:

   - The transformation $g(u) = (1-u)^2$ is one-to-one and decreasing over $[0,1]$.
   - The inverse transformation is $g^{-1}(v) = 1 - v^{1/2}$, for which $\dfrac{d}{dv}g^{-1}(v) = -\dfrac{1}{2v^{1/2}}$.
   - Since $\text{supp}(f_U) = (0,1)$ it follows that $\text{supp}(f_V) = (0,1)$.

   Hence for $0 < v < 1$ the PDF of V is

$$\begin{aligned} f_V(v) &= f_U\left[g^{-1}(v)\right]\left|\frac{d}{dv}g^{-1}(v)\right| \\ &= 12(1-v^{1/2})^2 v^{1/2}\left|-\frac{1}{2v^{1/2}}\right| \\ &= 6(1-v^{1/2})^2, \end{aligned}$$

   and zero otherwise.

4. The CDF of a random variable X is $F(x) = 1 - 1/x^3$ for $x \geq 1$ and zero otherwise. Find the CDF of the random variable $Y = 1/X$, then describe how a pseudo-random number from the distribution of Y can be obtained using an algorithm that generates uniformly distributed pseudo-random numbers in the range $[0,1]$.

   Answer: Let $g(x) = 1/x$ denote the transformation.

   - $\text{supp}(f_X) = [1,\infty] \Rightarrow \text{supp}(f_Y) = [0,1]$.
   - The inverse transformation: $g^{-1}(y) = 1/y$.

   Because $g(x)$ is a decreasing function over $\text{supp}(f_X)$,

$$F_Y(y) = 1 - F_X\left[g^{-1}(y)\right] = 1 - F_X\left(\frac{1}{y}\right) = \begin{cases} 0 & y < 0 \\ y^3 & 0 \leq y \leq 1 \\ 1 & y > 1. \end{cases}$$

   For the second part we use the fact that $F_Y(Y) \sim \text{Uniform}(0,1)$. First we invert F_Y by letting $u = F_Y(y)$ from which we obtain.

$$y = F_Y^{-1}(u) = u^{1/3}.$$

   Next we obtain a pseudo-random number u from the Uniform$[0,1]$ distribution, then compute

$$y = u^{1/3},$$

   which is a pseudo-random number from the distribution of Y.

3.5 More transformations

Example 3.29
Let $Z \sim N(0,1)$. Find the CDF of $X = \mu + \sigma Z$ in terms of the CDF of Z.

Solution: Let F_X and F_Z respectively denote the CDFs of X and Z.

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\mu + \sigma Z \leq x) = \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) = F_Z\left(\frac{x-\mu}{\sigma}\right).$$

Example 3.30 (The chi-squared distribution)
Let $X \sim N(0,1)$. Find the CDF of $Y = X^2$.

Solution: The PDF of $X \sim N(0,1)$ is given by $f_X(x) = \dfrac{1}{\sqrt{2\pi}}e^{-z^2/2}$. Hence,

$$\mathbb{P}(Y \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\,dx = \frac{2}{\sqrt{2\pi}}\int_0^{\sqrt{y}} e^{-x^2/2}\,dx.$$

Using a change-of-variable $t = x^2$, we obtain

$$\mathbb{P}(Y \leq y) = \int_0^y \left(\frac{1}{\sqrt{2\pi t}}e^{-t/2}\right) dt.$$

This is the CDF of the **chi-squared** distribution with one degree of freedom.

Example 3.31
Let $X \sim \text{Exponential}(\alpha)$ where α is a rate parameter, and let $Y = \theta e^X$, where $\theta > 0$ is a constant. Show that Y has the so-called **Pareto**(θ, α) distribution, whose CDF is given by

$$F_Y(y) = \begin{cases} 1 - \left(\frac{\theta}{y}\right)^\alpha & \text{for } y > \theta \\ 0 & \text{otherwise.} \end{cases}$$

Solution: Consider the transformation $g(x) = \theta e^x$.

- The CDF of X is $F_X(x) = 1 - e^{-\alpha x}$ for $x > 0$ (and zero otherwise).
- $\text{supp}(f_X) = [0, \infty)$.
- $\text{supp}(f_Y) = \{\theta e^x : x \geq 0\} = [\theta, \infty)$.
- $g(x)$ is an increasing function over $\text{supp}(f_X)$;
- the inverse transformation is $g^{-1}(y) = \log(y/\theta)$.

By Theorem 3.20,

$$F_Y(y) = F_X\left[\log\left(\frac{y}{\theta}\right)\right] = 1 - \exp\left[-\alpha\log\left(\frac{y}{\theta}\right)\right] = \begin{cases} 1 - \left(\dfrac{\theta}{y}\right)^\alpha & \text{for } y > \theta, \\ 0 & \text{otherwise.} \end{cases}$$

as required.

Remark: Compare the upper-tail probabilities of $X \sim \text{Exponential}(\alpha)$ and $Y \sim \text{Pareto}(\theta, \alpha)$:

$$\mathbb{P}(X > x) = e^{-\alpha x} \qquad \text{and} \qquad \mathbb{P}(Y > y) = \theta^\alpha y^{-\alpha}.$$

In both cases, the rate at which the tail probabilities converge to zero is controlled by the parameter α. We can see however that $\mathbb{P}(X > x) \to 0$ relatively quickly as $x \to \infty$, the rate of convergence depending "exponentially" on x, while $\mathbb{P}(Y > y) \to 0$ more slowly as $y \to \infty$, with the rate of convergence depending "polynomially" on y. Consequently, the Pareto distribution belongs to the class of so-called **heavy-tailed** distributions.

Example 3.32 (The standard normal distribution)
Let $X \sim N(\mu, \sigma^2)$, and define $Z = (X - \mu)/\sigma$. Find the PDF of Z.

Solution: Let $g(X) = \dfrac{X - \mu}{\sigma}$.

- The PDF of X is $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right]$.
- Because $\sigma > 0$, we see that $g(x)$ is increasing over $\operatorname{supp}(f_X) = (-\infty, \infty)$.
- $g(x)$ has inverse function $g^{-1}(z) = \mu + \sigma z$, whose first derivative is $\dfrac{d}{dz}g^{-1}(z) = \sigma$.

The PDF of Z is therefore given by

$$\begin{aligned} f_Z(z) &= f_X\left[g^{-1}(z)\right]\left|\frac{d}{dz}g^{-1}(z)\right| \\ &= \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{(\mu+\sigma z)-\mu}{\sigma}\right)^2\right]\sigma = \frac{1}{\sqrt{2\pi}}e^{-z^2/2} \end{aligned}$$

Example 3.33 (The lognormal distribution)
If $X \sim (\mu, \sigma^2)$, then $Y = e^X$ is said to have **lognormal** distribution. Find the PDF of Y.

Solution: Let $g(X) = e^X$.

- Recall that $f_X(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right]$.
- $g(x)$ is an increasing function over $\operatorname{supp}(f_X) = (-\infty, \infty)$.
- $g(x) \in [0, \infty]$ for $x \in \mathbb{R}$ so $\operatorname{supp}(f_Y) = [0, \infty)$.
- $g(x)$ has inverse function $g^{-1}(y) = \log y$, for which $\dfrac{d}{dy}g^{-1}(y) = \dfrac{1}{y}$.

Thus

$$\begin{aligned} f_Y(y) &= f_X\left[g^{-1}(y)\right]\left|\frac{d}{dy}g^{-1}(y)\right| \\ &= f_X(\log y)\left|\frac{1}{y}\right| = \frac{1}{y\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\log y-\mu}{\sigma}\right)^2\right] \end{aligned}$$

for $y > 0$, and zero otherwise.

Example 3.34 (The logistic distribution)
The Lomax(θ, α) distribution, also known as the Pareto Type II distribution or the shifted Pareto distribution, is a continuous distribution with PDF

$$f_X(x) = \frac{\alpha}{\theta}\left(1 + \frac{x}{\theta}\right)^{-(\alpha+1)} \quad \text{for } x > 0 \text{ and zero otherwise.}$$

Let $X \sim \text{Lomax}(1, 1)$. Show that the CDF of $Z = \log X$ is given by

$$F_Z(z) = \frac{e^z}{1+e^z}.$$

This is the CDF of the **standard logistic distribution**.

Solution: Taking $\alpha = 1$ and $\theta = 1$, the PDF of X is

$$f_X(x) = \frac{1}{(1+x)^2}$$

Consider the transformation $g(x) = \log x$.

- $g(x)$ is an increasing function over $\text{supp}(f_X) = (0, \infty)$.
- $\text{supp}(f_Z) = \{g(x) : x \in \text{supp}(f_X)\} = \{\log x : x \in (0, \infty)\} = (-\infty, \infty)$.
- The inverse transformation is $g^{-1}(z) = e^z$, and its first derivative is $\dfrac{d}{dz}g^{-1}(z) = e^z$.

The PDF of $Z = \log X$ is therefore given by

$$f_Z(z) = f_X\left[g^{-1}(z)\right]\left|\frac{d}{dz}g^{-1}(z)\right| = \frac{e^z}{(1+e^z)^2},$$

from which we the required CDF follows by integration.

Chapter 4 Expectation

Expectation is to random variables what probability is to events.

- Probability quantifies the 'size' of a random event.
- Expectation quantifies the 'size' of a random variable.

Elementary probability

Let Ω be a finite sample space and let $p : \Omega \to [0,1]$ be a probability mass function on Ω. In elementary probability the **expectation** of a random variable $X : \Omega \to \mathbb{R}$ is defined to be

$$\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega)$$

This is a sum over the **domain** of X. Expectation can also be defined over the **range** of X.

4.1 Simple variables

Definition 4.1 (Expectation of simple random variables)
The expectation of a simple random variable $X : \Omega \to \mathbb{R}$ is

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i f(x_i).$$

where f is the PMF of X and $\{x_1, x_2, \dots, x_n\}$ is the range of X.

In particular, the expectation of an indicator variable $I_A : \Omega \to \mathbb{R}$ is

$$\mathbb{E}(I_A) = \mathbb{P}(A)$$

which shows the connection between an event and its indicator function.

Definition 4.2
Let X and Y be random variables.

1. We say that X is **non-negative** if $X(\omega) \geq 0$ for all $\omega \in \Omega$. This is denoted by $X \geq 0$.
2. We say that X **dominates** Y if $X(\omega) \geq Y(\omega)$ for all $\omega \in \Omega$. This is denoted by $X \geq Y$.

Theorem 4.3 (Properties of expectation for simple random variables)
Let X and Y be simple random variables.

1. **Linearity**. For every $a, b \in \mathbb{R}$, $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.
2. **Positivity**. If $X \geq 0$ then $\mathbb{E}(X) \geq 0$.

3. **Monotonicity**. If $X \geq Y$ then $\mathbb{E}(X) \geq \mathbb{E}(Y)$.

Proof:

1. **Linearity**. Let $\{x_1, x_2, \ldots, x_m\}$ be the range of X and $\{y_1, y_2, \ldots, y_n\}$ be the range of Y. We define two partitions $\{A_1, A_2, \ldots, A_m\}$ and $\{B_1, B_2, \ldots, B_n\}$ of the underlying sample space Ω by
$$A_i = \{\omega : X(\omega) = x_i\} \quad \text{and} \quad B_j = \{\omega : Y(\omega) = y_j\}.$$
These can be combined to produce another partition of Ω:
$$\{A_i \cap B_j : i = 1, 2, \ldots, m;\ j = 1, 2, \ldots, n\}.$$
The composite variable $aX + bY$ takes the value $ax_i + by_j$ on the set $A_i \cap B_j$. Because there are only finitely many such sets, it follows that $aX + bY$ is also a simple random variable, so its expectation is given by
$$\begin{aligned}\mathbb{E}(aX + bY) &= \sum_{i=1}^{m}\sum_{j=1}^{n}(ax_i + by_j)\mathbb{P}(A_i \cap B_j) \\ &= a\sum_{i=1}^{m} x_i \sum_{j=1}^{n}\mathbb{P}(A_i \cap B_j) + b\sum_{j=1}^{n} y_j \sum_{i=1}^{m}\mathbb{P}(A_i \cap B_j).\end{aligned}$$
By the additivity of probability measures:
   - $\{A_i \cap B_j\}_{j=1}^{n}$ is a partition of A_i, so $\sum_{j=1}^{n}\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)$.
   - $\{A_i \cap B_j\}_{i=1}^{m}$ is a partition of B_j, so $\sum_{i=1}^{m}\mathbb{P}(A_i \cap B_j) = \mathbb{P}(B_j)$.

   Hence,
$$\mathbb{E}(aX + bY) = a\sum_{i=1}^{n} x_i\mathbb{P}(A_i) + b\sum_{j=1}^{m} y_j\mathbb{P}(B_j) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

2. **Positivity**. If $X(\omega) \geq 0$ for all ω we must have that each $x_i \geq 0$. Thus $\mathbb{E}(X) = \sum_{i=1}^{n} x_i\mathbb{P}(X = x_i)$ is a sum of non-negative terms, so $\mathbb{E}(X) \geq 0$.

3. **Monotonicity**. If $X \geq Y$ then $X - Y \geq 0$, so
   - by positivity, $\mathbb{E}(X - Y) \geq 0$;
   - by linearity, $\mathbb{E}(X) - \mathbb{E}(Y) \geq 0$, so $\mathbb{E}(X) \geq \mathbb{E}(Y)$.

For a transformed variable, we need not compute its PMF to compute its expectation. The following result is sometimes called the **law of the unconscious statistician**.

Theorem 4.4 (Expectation of transformed simple random variables)
If X be a simple random variable and $g : \mathbb{R} \to \mathbb{R}$ is a transformation, the expected value of the transformed variable $g(X)$ is
$$\mathbb{E}\big[g(X)\big] = \sum_{i=1}^{n} g(x_i)f(x_i).$$

Proof: Let $Y = g(X)$. Because X is simple, Y must also be simple. Let $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$ be the range of X and Y respectively and let C_j be the set containing those x_i that are mapped to y_j:
$$C_j = \{x_i : g(x_i) = y_j\} \qquad \text{for } i = 1, 2, \ldots, m.$$

Clearly, $\mathbb{P}(Y=y_j)=\sum_{x_i\in C_j}\mathbb{P}(X=x_i)$, so

$$\begin{aligned}\mathbb{E}(Y)&=\sum_{j=1}^{n}y_j\mathbb{P}(Y=y_j)=\sum_{j=1}^{n}y_j\left(\sum_{x_i\in C_j}\mathbb{P}(X=x_i)\right)\\&=\sum_{j=1}^{n}\sum_{x_i\in C_j}y_j\mathbb{P}(X=x_i)\\&=\sum_{j=1}^{n}\sum_{x_i\in C_j}g(x_i)\mathbb{P}(X=x_i)\quad\text{because } y_j=g(x_i)\text{ whenever } x_i\in C_j,\\&=\sum_{i=1}^{m}g(x_i)\mathbb{P}(X=x_i),\end{aligned}$$

where the last equality follows because $\{C_1,C_2,\ldots,C_m\}$ is a partition of $\{x_1,\ldots,x_n\}$.

4.2 Non-negative variables

Next we define the expectation for random variables that take only non-negative values.

Definition 4.5 (Expectation of non-negative random variables)

1. The expectation of a non-negative **discrete** random variable X is
$$\mathbb{E}(X)=\sum_{i=1}^{\infty}x_i f(x_i),$$
where f is the PMF of X and $\{x_1,x_2,\ldots\}$ is the range of X.
2. The expectation of a non-negative **continuous** random variable X is
$$\mathbb{E}(X)=\int_{-\infty}^{\infty}xf(x)\,dx,$$
where f is the PDF of X.

Non-negative random variables can have **infinite expectation**.

Example 4.6
A long line of athletes $k=0,1,2,\ldots$ make throws of a javelin to distances $D_0,D_1,D_2,\ldots$ respectively. Assume that the distances are independent and have the same distribution, and that the probability of any two throws being exactly the same distance is equal to zero. Let X be the number of throws until the initial distance D_0 is surpassed for the first time. Find the PMF of X, and show that $\mathbb{E}(X)$ is infinite.

Solution:

Example 4.7 (Pareto distrubtion)
Let X be a continuous random variable with the following PDF:

$$f(x) = \begin{cases} 1/x^2 & \text{for} \quad x > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\mathbb{E}(X)$ is infinite.

Solution:

4.3 Signed variables

Random variables that take both positive and negative values are called **signed variables**.

Definition 4.8
The **positive part** and **negative part** of a random variable X are

$$X^+(\omega) = \begin{cases} X(\omega) & \text{if } X(\omega) \geq 0, \\ 0 & \text{if } X(\omega) < 0, \text{ and} \end{cases}$$

$$X^-(\omega) = \begin{cases} -X(\omega) & \text{if } X(\omega) < 0, \\ 0 & \text{if } X(\omega) \geq 0. \end{cases}$$

respectively.

Note that X^+ and X^- are both non-negative random variables with

$$X = X^+ - X^- \qquad \text{and} \qquad |X| = X^+ + X^-.$$

Definition 4.9 (Expectation of signed random variables)
The **expectation** of a signed random variable X is

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$$

provided that $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ are not both infinite.

Remark 4.10
If $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ are both infinite, the expectation of X is **undefined** because we cannot make sense of the expression "$\infty - \infty$". In this case we say that the expectation of X **does not exist**.

Example 4.11
Let X be a discrete random variable with the following PMF:

$$f(k) = \begin{cases} \dfrac{3}{\pi^2 k^2} & \text{if } \ k \in \{\pm 1, \pm 2, \dots\} \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\mathbb{E}(X)$ is undefined.

Solution:

Example 4.12 (Cauchy distribution)
Let X be a continuous random variable having the following PDF,

$$f(x) = \frac{1}{\pi(1+x^2)} \qquad \text{for all } x \in \mathbb{R}.$$

Show that $\mathbb{E}(X)$ is undefined.

Solution:

4.4 Moments

4.4.1 Expectation of transformed variables

Let X be a random variable and let $g:\mathbb{R}\to\mathbb{R}$ be a transformation.

1. If g is a **non-negative** function, the expected value of $g(X)$ is

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \displaystyle\sum_{i=1}^{\infty} g(x_i)f(x_i) & \text{if } g(X) \text{ is discrete, and} \\ \displaystyle\int_{-\infty}^{\infty} g(x)f(x)\,dx & \text{if } g(X) \text{ is continuous.} \end{cases}$$

2. If g is a **signed function**, the expected value of $g(X)$ is

$$\mathbb{E}\big[g(X)\big] = \mathbb{E}\big[g^+(X)\big] - \mathbb{E}\big[g^-(X)\big],$$

provided $\mathbb{E}\big[g^+(X)\big]$ and $\mathbb{E}\big[g^-(X)\big]$ are not both infinite, where g^+ and g^- are respectively the positive and negative parts of g,

$$g^+(x) = \begin{cases} g(x) & \text{if } g(x) \geq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad g^-(x) = \begin{cases} -g(x) & \text{if } g(x) < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Example 4.13
For $X \sim \text{Uniform}[-1, 1]$ find $\mathbb{E}(1/X^2)$ and $\mathbb{E}(1/X)$.

Solution:

4.4.2 Moments

We now consider transformations of the form $g(x) = x^k$ for various values of $k \in \mathbb{Z}$.

Definition 4.14
The expectation $\mathbb{E}(X^k)$ is called the k**th moment about the origin** of X.

- $\mathbb{E}(X^0) = 1$,
- $\mathbb{E}(X)$ is the **mean** of X, often denoted by μ,
- $\mathbb{E}(X^2)$ is the **mean-square** of X.

Definition 4.15
The expectation $\mathbb{E}\big[(X-\mu)^k\big]$ is called the k**th moment about the mean** of X.

- $\mathbb{E}\big[(X-\mu)^0\big] = 1$,
- $\mathbb{E}\big[(X-\mu)\big] = 0$,
- $\mathbb{E}\big[(X-\mu)^2\big]$ is the **variance** of X, often denoted by σ^2.
- $\mathbb{E}\big[(X-\mu)^2\big]^{1/2}$ is the **standard deviation** of X, often denoted by σ.

4.4.3 Location, scale and shape

When trying to describe a distribution, it is natural to look for its **location**, **scale** (size) and **shape**.

Location To locate X, we look for a point $c \in \mathbb{R}$ such that the expected squared deviation $\mathbb{E}\big[(X-c)^2\big]$ around this point is minimum. By the linearity of expectation,

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

To find the value of c that minimises the expected squared deviation, we differentiate the right-hand side with respect to c and set the resulting expression to zero. This yields $c = \mathbb{E}(X)$, so the location of X is described by its **expectation** (or first moment about the origin), μ.

Scale The size of a distribution should not depend on its location, so we consider the **centred** variable

$$Y = X - \mathbb{E}(X),$$

which has the property $\mathbb{E}(Y) = 0$. The expected squared deviation of X around its mean $\mathbb{E}(X)$ is its **variance** (or second moment about the mean), so the size of X is described by its **standard deviation**, σ.

Shape The shape of a distribution should not depend on its location nor its scale. Thus we consider the higher moments of the so-called **standardised** variable,

$$Z = \frac{X-\mu}{\sigma},$$

which has the properties $\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$.

Definition 4.16
The **skewness** of a random variable X is defined to be

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

Skewness is a measure of **asymmetry**:

- Negative skew ($\gamma_1 < 0$): long tail on the left, mass concentrated on the right.
- Positive skew ($\gamma_1 > 0$): long tail on the right, mass concentrated on the left.

Example 4.17
If $X \sim \text{Binomial}(n,p)$, some tedious algebra shows that the skewness of X is

$$\gamma_1 = \frac{1-2p}{\sqrt{np(1-p)}} = \begin{cases} < 0 & \text{if } p > \frac{1}{2} \\ = 0 & \text{if } p = \frac{1}{2} \\ > 0 & \text{if } p < \frac{1}{2} \end{cases}$$

Definition 4.18
The **excess kurtosis** of a random variable X is defined to be

$$\gamma_2 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3.$$

Kurtosis is a measure of **peakiness**. The normal distribution has $\gamma_2 = 0$, so excess kurtosis provides a measure of peakiness relative to that of the normal distribution.

- Negative excess kurtosis ($\gamma_2 < 0$): tall and peaky with heavy tails (**leptokurtic**),
- Positive excess kurtosis ($\gamma_2 > 0$): low and wide with light tails (**platykurtic**).

Example 4.19
If $X \sim \text{Binomial}(n,p)$ some more tedious algebra shows that the excess kurtosis of X is

$$\gamma_2 = \frac{1-6p(1-p)}{np(1-p)} = \begin{cases} < 0 & \text{if } \left|p-\frac{1}{2}\right| < \frac{1}{2\sqrt{3}} \\ > 0 & \text{if } \left|p-\frac{1}{2}\right| > \frac{1}{2\sqrt{3}} \end{cases}$$

Exercise 4.20

1. Let X be a continuous random variable with uniform density on the interval $[-1,1]$,

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

Compute the moments $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathbb{E}(X^3)$, $\mathbb{E}(1/X)$ and $\mathbb{E}(1/X^2)$.

Answer: Since $xf(x)$, $x^2f(x)$ and $x^3f(x)$ are continuous and bounded over $\text{supp}(f) = [-1,1]$, the first three expectations can be computed using definite integrals. This is not the case for$\mathbb{E}(1/X)$ and $\mathbb{E}(1/X^2)$.

1. $\mathbb{E}(X) = \frac{1}{2}\int_{-1}^{1} x\,dx = 0.$
2. $\mathbb{E}(X^2) = \frac{1}{2}\int_{-1}^{1} x^2\,dx = \frac{1}{3}.$
3. $\mathbb{E}(X^3) = \frac{1}{2}\int_{-1}^{1} x^3\,dx = 0.$
4. Let $g(x) = 1/x$. This is a signed function, so

$$\begin{aligned} \mathbb{E}\left(\frac{1}{X}\right) &= \int_{-\infty}^{\infty} g^+(x)f(x)\,dx - \int_{-\infty}^{\infty} g^-(x)f(x)\,dx \\ &= \frac{1}{2}\int_0^1 \frac{1}{x}\,dx - \frac{1}{2}\int_{-1}^{0} \frac{-1}{x}\,dx \\ &= \frac{1}{2}\int_0^1 \frac{1}{x}\,dx - \frac{1}{2}\int_0^1 \frac{1}{x}\,dx = \infty - \infty \end{aligned}$$

so $\mathbb{E}(1/X)$ is undefined.

5. Let $g(x) = 1/x^2$. This is a non-negative function, so

$$\begin{aligned} \mathbb{E}\left(\frac{1}{X^2}\right) &= \int_{-\infty}^{\infty} g(x)f(x)\,dx \\ &= \frac{1}{2}\int_{-1}^{1} \frac{1}{x^2}\,dx = \int_0^1 \frac{1}{x^2}\,dx = \infty \end{aligned}$$

so $\mathbb{E}(1/X^2)$ is infinite.

2. Let X be a continuous random variable with the following PDF:

$$f(x) = \begin{cases} 1-|x| & \text{if } x \in [-1,1] \\ 0 & \text{otherwise.} \end{cases}$$

For what values of $k \in \mathbb{Z}$ do the moments $\mathbb{E}(X^k)$ exist?

Answer: For $k > 0$,

$$\mathbb{E}(X^k) = \int_{-1}^{0} x^k(1+x)\,dx + \int_{0}^{1} x^k(1-x)\,dx < \infty$$

Let $k < 0$. If k is even then X^k is non-negative, so

$$\mathbb{E}(X^k) = \mathbb{E}\big((X^+)^k\big) = +\infty$$

If k is odd,

$$\mathbb{E}(X^k) = \mathbb{E}\big((X^+)^k\big) - \mathbb{E}\big((X^-)^k\big) = \infty - \infty$$

so in this case the moment $\mathbb{E}(X^k)$ does not exist.

3. Let X be a discrete random variable with the following PMF:

$$\mathbb{P}(X = x) = \frac{45}{\pi^4 x^4} \qquad \text{for } x = \pm1, \pm2, \pm3\ldots \text{ (and zero otherwise).}$$

(a) For what values of $k \in \mathbb{Z}$ do the moments $\mathbb{E}(X^k)$ exist?

Answer: Let $c = 45/\pi^4$.

- If $k \leq 2$ then $\mathbb{E}(X^k)$ exists and is finite because

$$\mathbb{E}(X^k) = c\sum_{x\neq 0}\frac{x^k}{x^4} = 2c\sum_{x=1}^{\infty}\frac{1}{x^{4-k}} < \infty.$$

- If $k > 2$ and k is even, $\mathbb{E}(X^k)$ is infinite because

$$\mathbb{E}(X^k) = c\sum_{x\neq 0}\frac{x^k}{x^4} = c\sum_{x=1}^{\infty}\frac{1}{x^{4-k}} = \infty.$$

If $k > 2$ and k is odd, $\mathbb{E}(X^k)$ does not exist because

$$\mathbb{E}(X^k) = c\sum_{x\neq 0}\frac{x^k}{x^4} = c\sum_{x=1}^{\infty}\frac{1}{x^{4-k}} - c\sum_{x=1}^{\infty}\frac{1}{x^{4-k}} = \infty - \infty.$$

(b) Compute the variance of X and its first negative moment $\mathbb{E}(X^{-1})$.

Answer: The first two moments of X are

$$\mathbb{E}(X) = c\sum_{x\neq 0}\frac{1}{x^3} = c\sum_{x=1}^{\infty}\frac{1}{x^3} - c\sum_{x=1}^{\infty}\frac{1}{x^3} = 0$$

$$\mathbb{E}(X^2) = c\sum_{x\neq 0}\frac{1}{x^2} = 2c\sum_{x=1}^{\infty}\frac{1}{x^2} = 2\left(\frac{45}{\pi^4}\right)\left(\frac{\pi^2}{6}\right) = \frac{15}{\pi^2},$$

so $\mathrm{Var}(X) = 15/\pi^2$. Similarly,

$$\begin{aligned}\mathbb{E}\left(\frac{1}{X}\right) &= c\sum_{x\neq 0}^{\infty}\frac{1}{x^5} \\ &= c\sum_{n=1}^{\infty}\frac{1}{n^5} - c\sum_{x=1}^{\infty}\frac{1}{x^5} \\ &= 0.\end{aligned}$$

In fact, all negative odd-integer moments are zero.

4.5 The Markov and Chebyshev inequalities

4.5.1 Markov's inequality

If the distribution of a random variable is not known, probabilities can be estimated using the moments of the distribution. A simple upper bound on the tail probabilities of non-negative random variables is provided by **Markov's inequality**.

Theorem 4.21 (Markov's inequality)
Let X be a non-negative random variable. Then for every $a > 0$,

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

Proof: Let I_A be the indicator function of event $A = \{\omega : X(\omega) \geq a\}$.

$$I_A(\omega) = \begin{cases} 0 & \text{if } X(\omega) < a, \\ 1 & \text{if } X(\omega) \geq a. \end{cases}$$

- If $\omega \in A$ then $X(\omega) \geq a = aI_A(\omega)$.
- If $\omega \notin A$ then $X(\omega) \geq 0 = aI_A(\omega)$.

In either case we have $X(\omega) \geq aI_A(\omega)$ so by the monotonicity of expectation,

$$\mathbb{E}(X) \geq a\mathbb{E}(I_A) = a\mathbb{P}(A) \equiv a\mathbb{P}(X \geq a).$$

Hence $\mathbb{P}(X \geq a) \leq \mathbb{E}(X)/a$ as required.

Example 4.22
A fair die is rolled once. Use Markov's inequality to find an upper bound on the probability that we observe a score of at least 5.

Solution:

4.5.2 Chebyshev's inequality

An upper bound on the absolute deviation of a random variable from its mean is provided by **Chebyshev's inequality**.

Corollary 4.23 (Chebyshev's inequality)
Let X be any random varible with finite expectation. Then for all $\epsilon > 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\operatorname{Var}(X)}{\epsilon^2}.$$

Proof: Take the non-negative random variable $\big[X - \mathbb{E}(X)\big]^2$ in Markov's inequality with $a = \epsilon^2$:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \epsilon) = \mathbb{P}\big[\big(X - \mathbb{E}(X)\big)^2 \geq \epsilon^2\big] \leq \frac{\mathbb{E}\big[\big(X - \mathbb{E}(X)\big)^2\big]}{\epsilon^2} = \frac{\operatorname{Var}(X)}{\epsilon^2},$$

as required.

Example 4.24
Suppose that $\mathbb{E}(X) = 0$ and $\operatorname{Var}(X) = 1$. Find an integer value k such that $\mathbb{P}(|X| \geq k) \leq 0.01$.

Solution:

Example 4.25
Let X be a continuous random variable with expected value 3.6 and standard deviation 1.2.

Show that $\mathbb{P}(1.2 \leq X \leq 6.0) \geq 0.75$.

Solution:

Exercise 4.26

1. What does Chebyshev's inequality tell us about the probability that the value taken by a random variable deviates from its expectation by five or more standard deviations?

   Answer: For any random variable X with finite variance σ^2,

$$\mathbb{P}(|X-\mu| \geq 5\sigma) \leq \frac{\sigma^2}{(5\sigma)^2} = \frac{1}{25} = 0.04.$$

   Thus for any distribution, we are very unlikely to observe values that are more than 5 standard deviations away from the mean.

2. A fair coin is tossed n times. Does Chebyshev's inequality ensure that the observed number of heads will not deviate from $n/2$ by more than 100 with a probability of at least 0.99, provided that n is sufficiently large?

   Answer: No. Let $X_i = 1$ if a head occurs on the ith toss and $X_i = 0$ otherwise, and let $S_n = \sum_{i=1}^n X_i$ be the total number of heads after n tosses. Because the coin is fair, $\mathbb{E}(S_n) = n/2$ and $\text{Var}(S_n) = n/4$. Applying Chebyshev's inequality to S_n with $\epsilon = 100$:

$$\mathbb{P}(|S_n - n/2| > 100) \leq \text{Var}(S_n)/100^2$$

   But $\text{Var}(S_n) \to \infty$ as $n \to \infty$, so the probability that S_n deviates from $n/2$ by a fixed amount cannot be bounded using Chebyshev's inequality.

3. Let X be a random variable with mean $\mu \neq 0$ and variance σ^2. The **relative deviation** of X from its mean is defined by $D = \left|\frac{X-\mu}{\mu}\right|$. Use Chebyshev's inequality to show that $\mathbb{P}(D \geq \epsilon) \leq \left(\frac{\sigma}{\mu\epsilon}\right)^2$.

   Answer: By Chebyshev's inequality,

$$\mathbb{P}(D \geq \epsilon) = \mathbb{P}\left(\left|\frac{X-\mu}{\mu}\right| \geq \epsilon\right) = \mathbb{P}\big(|X-\mu| \geq |\mu|\epsilon\big) \leq \frac{\sigma^2}{\mu^2\epsilon^2}$$

4. (a) Let $X_1, X_2, \ldots, X_n$ be independent with each $X_i \sim \text{Bernoulli}(p)$. Using the fact that $p(1-p) \leq 1/4$ for all $0 < p < 1$, show that for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - p\right| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}.$$

   Answer: For the Binomial(n,p) distribution, Chebyshev's inequality yields

$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}$$

   and the result follows because $p(1-p) \leq 1/4$.

(b) Let A be an event associated with a random experiment. Suppose that the experiment is repeated n times. Let X_i be the indicator variable of the event that A occurs during the ith trial. Then $X_i \sim \text{Bernoulli}(p)$ where $p = \mathbb{P}(A)$, and the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ is the **relative frequency** of A over these n trials. What does the upper bound derived in part 1 say about the relative frequency of event A as $n \to \infty$.

Answer: The probability that the relative frequency $\displaystyle\frac{1}{n}\sum_{i=1}^{n} X_i$ of A differs from its true probability $p = \mathbb{P}(A)$ by more than a fixed amount (however small), tends to zero as the number of trials increases to infinity.

Chapter 5 Joint distributions

5.1 Random points in the plane

Let X and Y be random variables on the same probability space.

Definition 5.1

1. The **joint distribution** of X and Y is the function $\mathbb{P}_{X,Y}$ defined on pairs of subsets of $\mathbb{R}$ by
$$\mathbb{P}_{X,Y}(C,D) = \mathbb{P}(X \in C, Y \in D).$$
2. The distributions $\mathbb{P}_X(C) = \mathbb{P}(X \in C)$ and $\mathbb{P}_Y(D) = \mathbb{P}(Y \in D)$ are called the **marginal distributions** of X and Y respectively.

Definition 5.2

1. The **joint CDF** of X and Y is the function $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ given by
$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y).$$
2. The CDFs $F_X(x) = \mathbb{P}(X \leq x)$ and $F_Y(y) = \mathbb{P}(Y \leq y)$ are called the **marginal CDFs** of X and Y respectively.

Definition 5.3

1. X and Y are called **jointly discrete** if the random vector (X,Y) takes only countably many values in $\mathbb{R}^2$, in which case they are described by their **joint PMF**,
$$\begin{array}{rccl} f_{X,Y}: & \mathbb{R}^2 & \to & [0,1] \\ & (x,y) & \mapsto & \mathbb{P}(X = x, Y = y). \end{array}$$
2. The PMFs $f_X(x) = \mathbb{P}(X = x)$ and $f_Y(y) = \mathbb{P}(Y = y)$ are called the **marginal PMFs** of X and Y respectively.

Definition 5.4

1. X and Y are called **jointly continuous** if their joint CDF can be written as
$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,du\,dv \qquad \text{for all } x,y \in \mathbb{R},$$
for some integrable function $f_{X,Y} : \mathbb{R}^2 \to [0,\infty)$ called the **joint PDF** of X and Y.
2. The PDFs $f_X(x) = F'_X(x)$ and $f_Y(y) = F'_Y(y)$ are called the **marginal PDFs** of X and Y respectively, where $F_X(x)$ and $F_Y(y)$ are their marginal CDFs. Note that
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \qquad \text{and} \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx.$$

5.1.1 Independence

Recall that two events A and B are **independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Definition 5.5
Two random variables X and Y are said to be **independent** if the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for all $x, y \in \mathbb{R}$.

Lemma 5.6

1. X and Y are independent if and only if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$.
2. If X and Y are jointly discrete, they are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.
3. If X and Y are jointly continuous, they are independent if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$ and $\text{supp}(f_{X,Y})$ is a rectangular region in $\mathbb{R}^2$.

Proof: The first two parts follow directly from the definitions. For the jointly continuous case,

- because X and Y are independent, $F_{X,Y}(x, y) = F_X(x)F_Y(y)$, and differentiating both sides with respect to x and y we get $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.
- If the value taken by X affects the range of values taken by Y, then Y clearly dependes on X. Thus for X and Y to be independent we need that
$$\text{supp}(f_{X,Y}) = \text{supp}(f_X) \times \text{supp}(f_Y),$$
which is a rectangular region in $\mathbb{R}^2$.

Exercise 5.7
A fair die is rolled once. Let ω denote the outcome, and consider the random variables

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is odd,} \\ 2 & \text{if } \omega \text{ is even,} \end{cases} \quad \text{and} \quad Y(\omega) = \begin{cases} 1 & \text{if } \omega \leq 3, \\ 2 & \text{if } \omega \geq 4. \end{cases}$$

Find the joint PMF of X and Y. Are X and Y independent?

Solution:

Example 5.8
Let X and Y be jointly continuous random variables with the following joint PDF:

$$f_{X,Y}(x, y) = \begin{cases} c(1-x)y & \text{for } 0 \leq y \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

1. Sketch the support of $f_{X,Y}$.

2. Are X and Y independent?

3. Find the marginal PDFs of X and Y, and show that $c = 24$.

Solution:

5.2 Correlation

5.2.1 Product moments

Definition 5.9
The **product** of X and Y is the random variable

$$\begin{aligned} XY: \quad \Omega &\to \mathbb{R} \\ \omega &\mapsto X(\omega)Y(\omega). \end{aligned}$$

Definition 5.10

1. If X and Y are jointly discrete, the **product moment** of X and Y is
$$\mathbb{E}(XY)=\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}x_iy_j\,f_{X,Y}(x_i,y_j)$$
whenever this sum exists.
2. If X and Y are jointly continuous, the **product moment** of X and Y is
$$\mathbb{E}(XY)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xy\,f_{X,Y}(x,y)\,dx\,dy$$
whenever this integral exists.

5.2.2 Covariance

The covariance of X and Y is the product moment of the **centred** variables $X-\mathbb{E}(X)$ and $Y-\mathbb{E}(Y)$.

Definition 5.11
The **covariance** of X and Y is
$$\operatorname{Cov}(X,Y)=\mathbb{E}\big(\big[X-\mathbb{E}(X)\big]\big[Y-\mathbb{E}(Y)\big]\big).$$

Remark 5.12
Note that $\operatorname{Cov}(X,Y)=\operatorname{Cov}(Y,X)$ and $\operatorname{Cov}(X,X)=\operatorname{Var}(X)$.

In the same way that $\operatorname{Var}(X)=\mathbb{E}(X^2)-\mathbb{E}(X)^2$ we have the following convenient expression for $\operatorname{Cov}(X,Y)$.

Lemma 5.13
$\operatorname{Cov}(X,Y)=\mathbb{E}(XY)-\mathbb{E}(X)\mathbb{E}(Y)$.

Proof: Expand the product in Definition 5.11 then apply the linearity of expectation.

Lemma 5.14
For random variables $X_1,X_2,\ldots,X_n$,
$$\operatorname{Var}\left(\sum_{i=1}^{n}X_i\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}(X_i,X_j)$$

Proof: Let $Y=\sum_{i=1}^{n}X_i$. Then
$$\begin{aligned}\operatorname{Var}(Y)=\mathbb{E}(Y^2)-\mathbb{E}(Y)^2&=\mathbb{E}\left[\left(\sum_{i=1}^{n}X_i\right)^2\right]-\left[\mathbb{E}\left(\sum_{i=1}^{n}X_i\right)\right]^2\\&=\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\right]-\left[\sum_{i=1}^{n}\mathbb{E}(X_i)\right]^2\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(X_iX_j)-\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(X_i)\mathbb{E}(X_j)\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}(X_i,X_j)\end{aligned}$$

Exercise 5.15
Show that covariance is a **bilinear** operator, in the sense that

$$\mathrm{Cov}(aX_1 + bX_2, cY_1 + dY_2) = ac\mathrm{Cov}(X_1, Y_1) + ad\mathrm{Cov}(X_1, Y_2) + bc\mathrm{Cov}(X_2, Y_1) + cd\mathrm{Cov}(X_2, Y_2).$$

5.2.3 Correlation

Definition 5.16
X and Y are said to be **correlated** if $\mathrm{Cov}(X, Y) \neq 0$ or equivalently if

$$\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y),$$

otherwise they are said to be **uncorrelated**.

Lemma 5.17
If X and Y are independent, they are uncorrelated.

Proof: Let X and Y be jointly continuous (the discrete case is similar).

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

$$\begin{aligned}\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x, y)\, dx\, dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_X(x)f_Y(y)\, dx\, dy \quad \text{(by independence)}, \\ &= \left(\int_{-\infty}^{\infty} x\, f_X(x)\, dx\right)\left(\int_{-\infty}^{\infty} y\, f_Y(y)\, dy\right) \\ &= \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Lemma 5.18
If $X_1, X_2, \ldots, X_n$ are pairwise uncorrelated,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Proof: Because the X_i are pairwise uncorrelated, $\mathrm{Cov}(X_i, X_j) = 0$ whenever $i \neq j$, so by Lemma 5.14,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Cov}(X_i, X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Example 5.19
Let $X_1, \ldots, X_r$ be independent with each $X_i \sim \mathrm{Geometric}(p)$, the distribution of the number of failures before the first success in a sequence of independent Bernoulli trials where the probability of success is p. Find the mean and variance of $Y = \sum_{i=1}^{r} X_i$.

Solution:

5.2.4 Correlation coefficient

The correlation coefficient is the product moment of the **standardized** variables $\dfrac{X-\mathbb{E}(X)}{\sqrt{\operatorname{Var}(X)}}$ and $\dfrac{Y-\mathbb{E}(Y)}{\sqrt{\operatorname{Var}(Y)}}$.

Definition 5.20
The **correlation coefficient** of X and Y is

$$\rho(X,Y)=\mathbb{E}\left[\left(\frac{X-\mathbb{E}(X)}{\sqrt{\operatorname{Var}(X)}}\right)\left(\frac{Y-\mathbb{E}(Y)}{\sqrt{\operatorname{Var}(Y)}}\right)\right]$$

By the linearity of expectation, we have the following convenient expression for $\rho(X,Y)$.

Lemma 5.21
The correlation coefficient of X and Y can be written as

$$\rho(X,Y)=\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\cdot\operatorname{Var}(Y)}}$$

Proof: Follows easily from Lemma 5.13.

Note that $\rho(X,Y)=0$ whenever X and Y are uncorrelated. In fact the correlation coefficient satisfies the inequality $|\rho(X,Y)|\leq 1$ and thus provides a **standardized** measure of the (linear) dependence between X and Y. To prove this we need the following result from mathematical analysis.

Theorem 5.22 (Cauchy-Schwarz inequality for random variables)
For any two random variables X and Y,

$$\mathbb{E}(XY)^2\leq\mathbb{E}(X^2)\mathbb{E}(Y^2)$$

with equality if and only if $\mathbb{P}(Y=aX)=1$ for some $a\in\mathbb{R}$.

Theorem 5.23
The correlation coefficient satisfies the inequality

$$|\rho(X,Y)| \leq 1,$$

with equality if and only if $\mathbb{P}(Y = aX + b) = 1$ for some $a, b \in \mathbb{R}$.

Proof: Apply the Cauchy-Schwarz inequality to $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$:

$$\begin{aligned}\mathrm{Cov}(X,Y)^2 &= \mathbb{E}\big((X-\mathbb{E}X)(Y-\mathbb{E}Y)\big) \\ &\leq \mathbb{E}\big((X-\mathbb{E}X)^2\big)\mathbb{E}\big((Y-\mathbb{E}Y)^2\big) \\ &= \mathrm{Var}(X)\mathrm{Var}(Y),\end{aligned}$$

with equality if and only if there exists $a \in \mathbb{R}$ such that

$$\mathbb{P}\big[Y - \mathbb{E}Y = a(X - \mathbb{E}X)\big] = 1.$$

Hence,

$$|\rho(X,Y)| = \left|\frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}\right| \leq 1$$

with equality if and only if $\mathbb{P}(Y = aX + b) = 1$, where $b = \mathbb{E}Y - a\mathbb{E}X$.

Exercise 5.24

1. Let X and Y be two random variables having the same distribution but which are not necessarily independent. Show that $\mathrm{Cov}(X+Y, X-Y) = 0$ provided that their distribution has finite mean and variance.

   Answer: Perhaps the simplest method is the following: let $U = X + Y$ and $V = X - Y$. Then

$$\begin{aligned}\mathrm{Cov}(X+Y,X-Y) &= \mathbb{E}(UV) - \mathbb{E}(U)\mathbb{E}(V) \\ &= \mathbb{E}\big[(X+Y)(X-Y)\big] - \mathbb{E}(X+Y)\mathbb{E}(X-Y) \\ &= \mathbb{E}(X^2 - Y^2) - \big[\mathbb{E}(X) + \mathbb{E}(Y)\big]\big[\mathbb{E}(X) - \mathbb{E}(Y)\big] \qquad \text{(by the linearity of expectation)} \\ &= \mathbb{E}(X^2) - \mathbb{E}(Y^2) - \mathbb{E}(X)^2 + \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(Y)\mathbb{E}(X) + \mathbb{E}(Y)^2 \quad \text{(by linearity again)} \\ &= \big[\mathbb{E}(X^2) - \mathbb{E}(X)^2\big] - \big[\mathbb{E}(Y^2) - \mathbb{E}(Y)^2\big] \\ &= \mathrm{Var}(X) - \mathrm{Var}(Y).\end{aligned}$$

   Since X and Y have the same distribution, their variances must be equal, so $\mathrm{Cov}(X+Y, X-Y) = 0$.

2. Consider a fair six-sided die whose faces show the numbers $-2, 0, 0, 1, 3, 4$. The die is independently rolled four times. Let X be the average of the four numbers that appear, and let Y be the product of these four numbers. Compute $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathbb{E}(Y)$ and $\mathrm{Cov}(X,Y)$.

   Answer: Let X_1, X_2, X_3, X_4 be independent discrete random variables on the set $\{-2, 0, 0, 1, 3, 4\}$. Each X_i is identically distributed according to the following PMF:

k	-2	0	1	3	4
$\mathbb{P}(X_i = k)$	$1/6$	$1/3$	$1/6$	$1/6$	$1/6$

   Hence for $i = 1, 2, 3, 4$,

$$\begin{aligned}\mathbb{E}(X_i) &= \frac{1}{6}(-2+0+0+1+3+4) = 1, \\ \mathbb{E}(X_i^2) &= \frac{1}{6}(4+0+0+1+9+16) = \frac{30}{6} = 5, \\ \mathrm{Var}(X_i) &= \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2 = 4.\end{aligned}$$

Let $X = \frac{1}{4}(X_1 + X_2 + X_3 + X_4)$. Then

$$\mathbb{E}(X) = \frac{1}{4}\big(\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_4)\big) = 1$$

By independence,

$$\mathrm{Var}(X) = \frac{1}{16}\big(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_4)\big) = 1$$

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

and $Y = X_1X_2X_3X_4$. By independence,

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(X_1X_2X_3X_4)\\ &= \mathbb{E}(X_1)\mathbb{E}(X_2)\mathbb{E}(X_3)\mathbb{E}(X_4)\\ &= 1\end{aligned}$$

and because the X_i are identically distributed,

$$\begin{aligned}\mathbb{E}(XY) &= \frac{1}{4}\mathbb{E}\big(\mathbb{E}(X_1 + X_2 + X_3 + X_4)X_1X_2X_3X_4\big)\\ &= \mathbb{E}(X_1^2)\mathbb{E}(X_2)\mathbb{E}(X_3)\mathbb{E}(X_4)\\ &= \mathbb{E}(X_1^2) = 5\end{aligned}$$

so $\mathrm{Cov}(XY) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 4$.

3. A fair die is rolled twice. Let U denote the number obtained on the first roll, let V denote the number obtained on the second roll, let $X = U + V$ denote their sum and let $Y = U - V$ denote their difference. Compute the mean and variance of X and Y, and compute $\mathbb{E}(XY)$. Check whether X and Y are uncorrelated. Check whether X and Y are independent.

   Answer: Let $U, V \sim \text{Uniform}\{1, 2, 3, 4, 5, 6\}$ be independent (and identically distributed) random variables, and define $X = U + V$ and $Y = U - V$.

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(U) + \mathbb{E}(V) = 7\\ \mathbb{E}(Y) &= \mathbb{E}(U) - \mathbb{E}(V) = 0\end{aligned}$$

   By independence,

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(U) + \mathrm{Var}(V) = 35/6\\ \mathrm{Var}(Y) &= \mathrm{Var}(U) + \mathrm{Var}(V) = 35/6\end{aligned}$$

   Because U and V are identically distributed, and

$$XY = (U + V)(U - V) = U^2 - V^2$$

   it follows that

$$\mathbb{E}(XY) = \mathbb{E}(U^2) - \mathbb{E}(V^2) = 0$$

   X and Y are uncorrelated, since

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$$

   However X and Y are not independent, because (for example)

$$\mathbb{P}(Y = 0) \neq \mathbb{P}(Y = 0|X = 12) = 1$$

5.3 Conditional distributions

Recall that for any two events A and B with $\mathbb{P}(A) > 0$, the **conditional probability** of B given A is

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

This notion extends to random variables: if C and D are subsets of $\mathbb{R}$, the conditional probability of $\{Y \in D\}$ given $\{X \in C\}$ is

$$\mathbb{P}(Y \in D | X \in C) = \frac{\mathbb{P}(X \in C, Y \in D)}{\mathbb{P}(X \in C)}.$$

5.3.1 Conditioning on events

Let A be an event and let Y be a random variable.

Definition 5.25

1. The **conditional distribution** of $Y|A$ is the function $\mathbb{P}_{Y|A}$ defined on subsets of $\mathbb{R}$ by

$$\mathbb{P}_{Y|A}(D) = \mathbb{P}(Y \in D|A) = \frac{\mathbb{P}\big(\{Y \in D\} \cap A\big)}{\mathbb{P}(A)} \qquad \text{for all } D \subseteq \mathbb{R}.$$

2. The **conditional CDF** of $Y|A$ is the function $F_{Y|A}$ defined on $\mathbb{R}$ by

$$F_{Y|A}(y) = \mathbb{P}(Y \leq y|A) = \frac{\mathbb{P}\big(\{Y \leq y\} \cap A\big)}{\mathbb{P}(A)} \qquad \text{for all } y \in \mathbb{R}.$$

3. If Y is discrete the, **conditional PMF of** $Y|A$ is the function

$$f_{Y|A}(y) = \mathbb{P}(Y = y|A) = \frac{\mathbb{P}\big(\{Y = y\} \cap A\big)}{\mathbb{P}(A)}.$$

4. If Y is continuous the **conditional PDF of** $Y|A$ is the function

$$f_{Y|A}(y) = F'_{Y|A}(y).$$

5. The **conditional expectation of** $Y|A$ is taken with respect to the conditional distribution of $Y|A$:

$$\mathbb{E}\big(Y|X\big) = \begin{cases} \displaystyle\sum_{i=1}^{\infty} y_i f_{Y|A}(y_i) & \text{if } Y \text{ is discrete;} \\ \displaystyle\int_{-\infty}^{\infty} y f_{Y|A}(y)\,dy & \text{if } Y \text{ is continuous.} \end{cases}$$

Example 5.26
A fair coin is tossed repeatedly until a head occurs. Find the conditional PMF of the number of times the coin is tossed, given that the coin is tossed an odd number of times.

Solution:

Example 5.27
Let $Y \sim \text{Uniform}[0,1]$. Find the conditional expectation of Y given that $1/2 \leq Y \leq 3/4$.

Solution:

5.3.2 Conditioning on random variables

Let X and Y be two random variables defined on the same probability space.

Definition 5.28

1. The conditional distribution of $Y|X$ is the function $\mathbb{P}_{Y|X}$ defined on pairs of subsets of $\mathbb{R}$ by

$$\mathbb{P}_{Y|X}(C,D) = \mathbb{P}(Y \in D|X \in C) = \frac{\mathbb{P}(X \in C, Y \in D)}{\mathbb{P}(X \in C)} \qquad \text{for all } C, D \subseteq \mathbb{R}.$$

   This is completely determined by the following.

2. The conditional CDF of $Y|X$ is the function $F_{Y|X}$ defined on $\mathbb{R}^2$ by

$$F_{Y|X}(x,y) = \mathbb{P}(Y \leq y|X \leq x) = \frac{\mathbb{P}(X \leq x, Y \leq y)}{\mathbb{P}(X \leq x)} \qquad \text{for all } x, y \in \mathbb{R}.$$

Lemma 5.29
The conditional CDF of $Y|X$ satisfies

$$F_{Y|X}(x,y) = \frac{F_{X,Y}(x,y)}{F_X(x)}$$

where $F_{X,Y}$ is the joint CDF of X and Y, and F_X is the marginal CDF of X.

Proof:

$$F_{Y|X}(x,y) = \mathbb{P}(Y \leq y \,|\, X \leq x) = \frac{\mathbb{P}(X \leq x, Y \leq y)}{\mathbb{P}(X \leq x)} = \frac{F_{X,Y}(x,y)}{F_X(x)}.$$

Definition 5.30
If $f_X(x) > 0$ the **conditional PMF/PDF of** $Y|X = x$ is

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

where $f_{X,Y}$ is the joint PMF/PDF of X and Y.

Recall the partition theorem for random events: if $\{A_1, A_2, \ldots\}$ is a partition of B then

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

We have the following version of the partition theorem for random variables.

Theorem 5.31
The marginal PMF/PDF of Y satisfies

$$\begin{aligned} f_Y(y) &= \sum_{i=1}^{\infty} f_{Y|X=x_i}(y) f_X(x_i) && \text{if } X \text{ is discrete, and} \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{Y|X=x}(y) f_X(x)\,dx && \text{if } X \text{ is continuous.} \end{aligned}$$

Proof: For the continuous case (the discrete case is similar),

$$\int f_{Y|X=x}(y) f_Y(y)\,dx = \int \left(\frac{f_{X,Y}(x,y)}{f_Y(y)}\right) f_Y(y)\,dx = \int f_{X,Y}(x,y)\,dx = f_Y(y).$$

Definition 5.32
Let x be a fixed value. The **conditional expectation of** $Y|X=x$ is

$$\begin{aligned} \mathbb{E}(Y|X=x) &= \sum_{j=1}^{\infty} y_j\, f_{Y|X=x}(y_j) && \text{if } Y \text{ is discrete, or} \\ \mathbb{E}(Y|X=x) &= \int_{-\infty}^{\infty} y\, f_{Y|X=x}(y)\,dy && \text{if } Y \text{ is continuous.} \end{aligned}$$

5.3.3 Conditional expectation

For any fixed value of x the conditional expectation $\mathbb{E}(Y|X=x)$ is just a number. Let us now think of x as a variable quantity, and consider the transformation

$$\begin{aligned} g:\ \mathbb{R} &\to \mathbb{R} \\ x &\mapsto \mathbb{E}(Y|X=x) \end{aligned}$$

This transformation of X yields a new random variable.

Definition 5.33
The **conditional expectation of** $Y|X$ is the random variable

$$\begin{aligned} \mathbb{E}(Y|X):\ \Omega &\to \mathbb{R} \\ \omega &\mapsto \mathbb{E}\big[Y|X=X(\omega)\big]. \end{aligned}$$

The distribution of $\mathbb{E}(Y|X)$ depends only on the distribution of X and its expectation is given by

$$\begin{aligned} \mathbb{E}\big[\mathbb{E}(Y|X)\big] &= \sum_{i=1}^{\infty} \mathbb{E}(Y|X=x_i) f_X(x_i) && \text{if } X \text{ is discrete, or} \\ \mathbb{E}\big[\mathbb{E}(Y|X)\big] &= \int_{-\infty}^{\infty} \mathbb{E}(Y|X=x) f_X(x)\,dx && \text{if } X \text{ is continuous.} \end{aligned}$$

Theorem 5.34 (Law of total expectation)
Let X and Y be random variables defined on the same probability space. Then

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

Proof: For discrete random variables (the continuous case is similar):

$$\begin{aligned}\mathbb{E}\big[\mathbb{E}(Y|X)\big] = \sum_x \mathbb{E}(Y|X=x)f_X(x) &= \sum_x \left(\sum_y y\, f_{Y|X}(y|x)\right) f_X(x) \\ &= \sum_x \left(\sum_y y\, \frac{f_{X,Y}(x,y)}{f_X(x)}\right) f_X(x) \\ &= \sum_x \left(\sum_y y f_{X,Y}(x,y)\right) \\ &= \sum_x y \left(\sum_y f_{X,Y}(x,y)\right) \\ &= \sum_x y f_Y(y) = \mathbb{E}(Y).\end{aligned}$$

Theorem 5.35 (Law of total variance)
If X and Y are random variables defined on the same probability space,

$$\mathrm{Var}(Y) = \mathbb{E}\big[\mathrm{Var}(Y|X)\big] + \mathrm{Var}\big[\mathbb{E}(Y|X)\big]$$

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

Proof: By the law of total expectation,

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 \\ &= \mathbb{E}\big[\mathbb{E}(Y^2|X)\big] - \mathbb{E}\big[\mathbb{E}(Y|X)\big]^2\end{aligned}$$

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

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

Hence, by the linearity of expectation,

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}\big[\mathrm{Var}(Y|X)\big] + \Big(\mathbb{E}\big[\mathbb{E}(Y|X)^2\big] - \mathbb{E}\big[\mathbb{E}(Y|X)\big]^2\Big) \\ &= \mathbb{E}\big[\mathrm{Var}(Y|X)\big] + \mathrm{Var}[\mathbb{E}(Y|X)\big].\end{aligned}$$

Example 5.36
Let the joint PDF of the continuous random variables X and Y be

$$f_{X,Y}(x,y) = \begin{cases} cxy & \text{for } x,y \geq 0 \text{ with } x+y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

1. Sketch the support of $f_{X,Y}$
2. Show that $c = 24$.
3. Compute the conditional expectation $\mathbb{E}(Y|X)$.
4. Verify the identity $\mathbb{E}\big[\mathbb{E}(Y|X)\big] = \mathbb{E}(Y)$.

Solution:

Exercise 5.37

1. Let X and Y be jointly continuous random variables having the following joint PDF,

$$f_{X,Y}(x,y) = \begin{cases} \frac{21}{4}x^2 y & x^2 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

1. Sketch the support of $f_{X,Y}$.
2. Find the marginal PDFs of X and Y.
3. Find the mean and variance of Y.
4. Find the conditional PDF of Y given $X = x$.
5. Are X and Y independent?
6. Find the conditional expectation of Y given $X = x$.
7. Find the conditional expectation of Y given X.
8. Verify that $\mathbb{E}(Y) = \mathbb{E}\big[\mathbb{E}(Y|X)\big]$.

Answer:

1. The support of the joint PDF $f(x,y)$ is the set $\{(x,y) : x^2 < y < 1\}$. This is the region of the plane between the vertical lines $x = -1$ and $x = +1$, bounded above by the horizontal line $y = 1$ and below by parabola $y = x^2$. In particular,
   - For fixed $x \in [-1,1]$, $f_{X,Y}(x,y) \neq 0$ only for $y \in [x^2, 1]$.
   - For fixed $y \in [0,1]$, $f_{X,Y}(x,y) \neq 0$ only for $x \in [-\sqrt{y}, +\sqrt{y}]$.
2. The marginal distributions are computed as follows:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy = \int_{x^2}^{1} \frac{21}{4}x^2 y\,dy = \frac{21}{4}x^2\left[\frac{y^2}{2}\right]_{x^2}^{1} = \begin{cases} \frac{21}{8}x^2(1-x^4) & -1 < x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$[2ex] f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx = \int_{-\sqrt{y}}^{\sqrt{y}} \frac{21}{4}x^2 y\,dx = \frac{21}{4}y\left[\frac{x^3}{3}\right]_{-\sqrt{y}}^{\sqrt{y}} = \begin{cases} \frac{7}{2}y^{5/2} & 0 < y < 1, \\ 0 & \text{otherwise.} \end{cases}$$

3. The expected value and variance of Y are computed as follows:

$$\begin{aligned} \mathbb{E}(Y) &= \int_{-\infty}^{\infty} y\, f_Y(y)\,dy = \int_0^1 y\left(\frac{7y^{5/2}}{2}\right)dy = \frac{7}{9}, \\ \mathbb{E}(Y^2) &= \int_{-\infty}^{\infty} y^2\, f_Y(y)\,dy = \int_0^1 y^2\left(\frac{7y^{5/2}}{2}\right)dy = \frac{7}{11}, \\ \mathrm{Var}(Y) &= \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \frac{7}{11} - \frac{49}{81} = \frac{28}{891}. \end{aligned}$$

4. The conditional PDF of Y given $X = x$ is
$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{(21/4)x^2y}{(21/8)x^2(1-x^4)} = \begin{cases} \dfrac{2y}{1-x^4} & x^2 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$
5. X and Y are clearly not independent, because the support of $f_{X,Y}$ is not a rectangular region, and moreover, the conditional PDF of Y given $X = x$ depends on x.
6. The conditional expected value of Y given that $X = x$ is
$$\begin{aligned} \mathbb{E}(Y|X=x) &= \int_{-\infty}^{\infty} y f_{Y|X=x}(y\,|\,x)\,dy \\ &= \int_{x^2}^{1} y\left(\frac{2y}{1-x^4}\right)dy = \frac{2}{1-x^4}\left[\frac{y^3}{3}\right]_{x^2}^{1} = \frac{2(1-x^6)}{3(1-x^4)} \end{aligned}$$
7. The conditional expectation of Y given X is the random variable
$$\mathbb{E}(Y|X) = \frac{2(1-X^6)}{3(1-X^4)}$$
8. The expected value of $\mathbb{E}(Y|X)$ is
$$\begin{aligned} \mathbb{E}\big[\mathbb{E}(Y|X)\big] &= \int_{-\infty}^{\infty} \mathbb{E}(Y|X=x) f_X(x)\,dx \\ &= \frac{2}{3}\int_{-1}^{1}\left(\frac{1-x^6}{1-x^4}\right)\left(\frac{21}{8}x^2(1-x^4)\right)dx \\ &= \frac{7}{4}\int_{-1}^{1} x^2(1-x^6)\,dx = \frac{7}{9}. \end{aligned}$$
Thus $\mathbb{E}\big[\mathbb{E}(Y|X)\big] = \mathbb{E}(Y)$ as required.

2. A man puts his house for sale and decides to accept the first offer that exceeds the reserve price of r. Let $Y_1, Y_2, \ldots$ represent the sequence of offers received, and suppose that the Y_i are independent and identically distributed random variables, each having exponential distribution with rate parameter λ.

(a) Show that the expected number of offers received before the house is sold is $e^{\lambda r}$.

Answer: Let N be the number of offers received before the house is sold. Then $\{N = k\}$ is the event that the first $k-1$ offers are at most r, each occurring independently with probability $F(r)$, and the kth offer exceeds r, which occurs with probability $1 - F(r)$. Thus N has **geometric** distribution, with 'probability of success' equal to $1 - F(r)$ (where 'success' corresponds to the sale of the house). Hence the expected number of offers received before the house is sold is
$$\mathbb{E}(N) = \frac{1}{1-F(r)} = e^{\lambda r}.$$

(b) Show that the expected the expected selling price of the house is $r + 1/\lambda$.

Answer: Let $Y \sim \text{Exponential}(\lambda)$ and let $A = \{Y > r\}$. The con Let F_S be the conditional CDF of Y_i given that $X_i > r$:
$$F_Y|A(y) = \frac{\mathbb{P}(r < Y \leq y)}{\mathbb{P}(Y > r)} = \frac{F(y) - F(r)}{1 - F(r)} = \begin{cases} 1 - e^{-\lambda(y-r)} & y > r, \\ 0 & \text{otherwise.} \end{cases}$$
A straightforward calculation yields $\mathbb{E}(Y|Y > r) = r + 1/\lambda$.

3. **Compound distributions**. If $Y \sim \text{Poisson}(\lambda)$ then $\mathbb{E}(Y) = \lambda$ and $\text{Var}(Y) = \lambda$. The Poisson distribution has only a single parameter so we cannot shift and scale the distribution by different ammounts, which limits its usefulness in certain practical applications. This problem can be addressed by allowing the parameter itself to be a random variable.

   (a) Let $Y \sim \text{Poisson}(X)$ where $X \sim \text{Exponential}(\theta)$ and $\theta > 0$ is a fixed scale parameter. Use the laws of total expectation and total variance, and the fact that $\mathbb{E}(X) = \theta$ and $\text{Var}(X) = \theta^2$, to show that $\mathbb{E}(Y) = \theta$ and $\text{Var}(Y) = \theta(1+\theta)$.

   Answer: The conditional expectation and conditional variance of Y are

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

   By the law of total expectation

$$\mathbb{E}(Y) = \mathbb{E}\big[\mathbb{E}(Y|X)\big] = \mathbb{E}(X) = \theta,$$

   and by the law of total variance,

$$\begin{aligned}\text{Var}(Y) &= \mathbb{E}\big[\text{Var}(Y|X)\big] + \text{Var}\big[\mathbb{E}(Y|X)\big] \\ &= \mathbb{E}(X) + \text{Var}(X) \\ &= \theta + \theta^2 = \theta(1+\theta)\end{aligned}$$

   because $\mathbb{E}(X) = \theta$ and $\text{Var}(X) = \theta^2$ for $X \sim \text{Exponential}(\theta)$.

   (b) Let $Y \sim \text{Poisson}(X)$ where $X \sim \text{Gamma}(\alpha, \beta)$ for some fixed parameters $\alpha, \beta > 0$. Find values for α and β such that $\mathbb{E}(Y) = 2$ and $\text{Var}(Y) = 9$.

   Answer: Because $Y \sim \text{Poisson}(X)$ we have $\mathbb{E}(Y|X) = X$ and $\text{Var}(Y|X) = X$, so both are distributed according to the $\text{Gamma}(\alpha, \beta)$ distribution, whose mean and variance are α/β and α/β^2 respectively.

   By the laws of total expectation and total variance,

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}\big[\mathbb{E}(Y|X)\big] = \mathbb{E}(X) = \alpha/\beta \\ \text{Var}(Y) &= \mathbb{E}\big[\text{Var}(Y|X)\big] + \text{Var}\big[\mathbb{E}(Y|X)\big] = \mathbb{E}(X) + \text{Var}(X) = \alpha(\beta+1)/\beta^2.\end{aligned}$$

   Solving $\alpha/\beta = 2$ and $\alpha(\beta+1)/\beta^2 = 9$ yields $\alpha = 4/7$ and $\beta = 2/7$.

   If α is an integer, Y has the so-called $\text{NegativeBinomial}(r, p)$ distribution with $r = \alpha$ and $p = 1/(1+\beta)$. This is the distribution of the number of successes up to the rth failure in a sequence of independent Bernoulli trials where each trial has probability of success p.

Chapter 6 Sums of random variables

6.1 Generating functions

A **power series** is an infinite series of the form

$$G(t) = \sum_{k=0}^{\infty} a_k t^k.$$

Power series can be used to represent sequences of real numbers $a_0, a_1, a_2, \ldots$, in which case they are often called **generating functions**. The terms of the sequence can be recovered by repeatedly differentiating the power series and evaluating these derivatives at $t = 0$:

$$G(0) = a_0, \quad G'(0) = a_1, \quad G''(0) = 2a_2, \quad G'''(0) = 6a_3 \quad \text{and so on.}$$

In this way, the power series **generates** the sequence.

6.1.1 Probability generating functions

Generating functions allow probability distributions to be represented as functions of a single variable.

Definition 6.1
Let X be a discrete random variable taking values in the range $\{0, 1, 2, \ldots\}$ and let f_X denote its PMF. The **probability generating function** (PGF) of X is the generating function of its PMF,

$$G_X(t) = \mathbb{E}(t^X) = \sum_{k=0}^{\infty} f_X(k) t^k.$$

Remark 6.2
Because f is a PMF, $G(t)$ converges for all $|t| \leq 1$ with $G(0) = 0$ and $G(1) = \sum_{k=0}^{\infty} f(k) = 1$.

As the following theorem shows, probability generating functions turn **sums** of independent random variables into **products** of independent random variables:

Theorem 6.3
If X and Y are independent then $G_{X+Y}(t) = G_X(t)G_Y(t)$.

Proof: If X and Y are independent, then t^X and t^Y are also independent, so

$$G_{X+Y}(t) = \mathbb{E}(t^{X+Y}) = \mathbb{E}(t^X t^Y) = \mathbb{E}(t^X)\mathbb{E}(t^Y) = G_X(t)G_Y(t).$$

Exercise 6.4
If $Y = a + bX$, show that $G_Y(t) = t^a G_X(t^b)$.

Answer:

$$G_Y(t) = \mathbb{E}\big(t^{a+bX}\big) = t^a\mathbb{E}\big((t^b)^X\big) = t^aG_X(t^b).$$

Corollary 6.5
If $X_1, X_2, \ldots, X_n$ are independent random variables taking values in the non-negative integers, the PGF of their sum is equal to the product of their individual PGFs,

$$G_{X_1+X_2+\ldots+X_n}(t) = G_{X_1}(t)G_{X_2}(t)\cdots G_{X_n}(t).$$

The PGFs of some fundamental discrete distributions on $\{0, 1, 2, \ldots\}$ are shown in Table 6.1.

	PMF	PGF	
$X \sim \text{Bernoulli}(p)$	$p^k(1-p)^{1-k}$	$1-p+pt$	
$X \sim \text{Binomial}(n, p)$	$\binom{n}{k}p^k(1-p)^{n-k}$	$(1-p+pt)^n$	
$X \sim \text{Geometric}(p)$	$(1-p)^kp$	$\dfrac{p}{1-(1-p)t}$	for $\lvert t\rvert < (1-p)^{-1}$
$X \sim \text{Poisson}(\lambda)$	$\lambda^ke^{-\lambda}/k!$	$e^{\lambda(t-1)}$	

Table 6.1: The PGFs of some fundamental discrete distributions.

Exercise 6.6
For each of the distributions shown in Table 6.1 derive the expressions for their PGFs from the corresponding PMFs.

Answer:

1. If $X \sim \text{Bernoulli}(p)$, then $\mathbb{P}(X = 0) = 1 - p$, $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = k) = 0$ for all $k \geq 2$ so
$$G_X(t) = \sum_{k=0}^{\infty} f_X(k)t^k = (1-p)t^0 + pt^1 = 1-p+pt.$$

2. If $X \sim \text{Binomial}(n, p)$, it can be written as $X = X_1 + X_2 + \ldots + X_n$ where each X_i is independent with $X_i \sim \text{Bernoulli}(p)$. Thus by Corollary 6.5,
$$G_X(t) = G_{X_1+X_2+\ldots+X_n}(t) = G_{X_1}(t)G_{X_2}(t)\cdots G_{X_n}(t) = (1-p+pt)^n.$$

3. If $X \sim \text{Geometric}(p)$, then $\mathbb{P}(X = k) = (1-p)^kp$ for $k = 0, 1, 2, \ldots$, so
$$G_X(t) = \sum_{k=0}^{\infty} f_X(k)t^k = \sum_{k=0}^{\infty}(1-p)^kpt^k = p\sum_{k=0}^{\infty}\big[(1-p)t\big]^k = \frac{p}{1-(1-p)t} \qquad \text{for all } |t| < \frac{1}{1-p}.$$
where we have used the fact that $\displaystyle\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for $|r| < 1$.

4. If $X \sim \text{Poisson}(\lambda)$, then $\mathbb{P}(X = k) = \lambda^ke^{-\lambda}/x!$ so
$$G_X(t) = \sum_{k=0}^{\infty} f_X(k)t^k = \sum_{k=0}^{\infty}\left(\frac{\lambda^ke^{-\lambda}}{k!}\right)t^k = e^{-\lambda}\sum_{i=1}^{\infty}\frac{(\lambda t)^k}{k!} = e^{-\lambda}e^{\lambda t} = e^{\lambda(t-1)}.$$

6.1.2 Moments

The moments of a distribution can be recoverd by repeatedly computing the derivatives of its PGF and evaluating the resulting expressions at $t = 1$.

Theorem 6.7
Let $G_X^{(n)}(t)$ denote the nth derivative of $G_X(t)$. Then

$$G_X^{(n)}(1) = \mathbb{E}\big[X(X-1)\dots(X-n+1)\big].$$

These are called the **factorial moments** of X.

Proof: Take $t < 1$, and compute the nth derivative of G to obtain

$$\begin{aligned}
G^{(n)}(t) &= \frac{d^n}{dt^n}\left[\sum_{k=0}^{\infty} t^k f_X(k)\right] \\
&= \sum_{k=0}^{\infty}\left[\frac{d^n}{dt^n}t^k\right] f_X(k) \\
&= \sum_{k=0}^{\infty} k(k-1)\cdots(k-n+1)t^{k-n} f_X(k) \\
&= \mathbb{E}\big[X(X-1)\cdots(X-n+1)t^{X-n}\big]
\end{aligned}$$

Thus $G^{(n)}(1) = \mathbb{E}\big[X(X-1)\cdots(X-n+1)\big]$, as required.

Exercise 6.8
Show that $\mathbb{E}(X) = G_X'(1)$ and $\text{Var}(X) = G_X''(1) + G_X'(1) - G_X'(1)^2$.

Answer:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&= \mathbb{E}\big[X(X-1) + X\big] - \mathbb{E}(X)^2 \\
&= \mathbb{E}\big[X(X-1)\big] + \mathbb{E}(X) - \mathbb{E}(X)^2 \\
&= G_X''(1) + G_X'(1) - G_X'(1)^2.
\end{aligned}$$

Sums with a random number of terms

Theorem 6.9
Let $X_1, X_2, \dots$ be independent and identically distributed random variables and let $G_X(t)$ denote their common PGF. Let N be another random variable taking values in the non-negative integers and independent of the X_i, and let $G_N(t)$ denote its PGF. Then the PGF of the sum $S_N = X_1 + X_2 + \dots + X_N$ is

$$G_{S_N}(t) = G_N\big[G_X(t)\big].$$

and its expectation satisfies $\mathbb{E}(S_N) = \mathbb{E}(N)\mathbb{E}(X)$. (This is known as Wald's identity.)

Proof:

$$\begin{aligned}
G_{S_N}(t) = \mathbb{E}(t^{S_N}) &= \mathbb{E}\big(\mathbb{E}(t^{S_N} \mid N)\big) \quad \text{(law of total expectation)} \\
&= \sum_{n=0}^{\infty} \mathbb{E}(t^{S_N} \mid N = n)\mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{E}(t^{S_n})\mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{E}(t^{X_1+X_2+\ldots+X_n} \mid N = n)\mathbb{P}(N = n) \\
&= \sum_{n=0}^{\infty} \mathbb{E}(t^{X_1})\mathbb{E}(t^{X_2})\cdots\mathbb{E}(t^{X_n})\mathbb{P}(N = n) \quad \text{(by independence)} \\
&= \sum_{n=0}^{\infty} G_X(t)^n\mathbb{P}(N = n) \\
&= G_N\big[G_X(t)\big].
\end{aligned}$$

To find $\mathbb{E}(S_N)$, we find $G_{S_N}(t)$ then evaluate its first derivative at $t = 1$.

$$\frac{d}{dt}\big[G_{S_N}(t)\big] = \frac{d}{dt}\big[G_N\big(G_X(t)\big)\big] = \frac{dG_N(u)}{du} \times \frac{du}{dt} \quad \text{where } u = G_X(t).$$

Setting $t = 1$ so that $u = G_X(1) = 1$,

$$\mathbb{E}(S_N) = G'_{S_N}(1) = \big[G'_N(1)\big]\big[G'_X(1)\big] = \mathbb{E}(N)\mathbb{E}(X).$$

Example 6.10
A hen lays N eggs where N has Poisson distribution with parameter λ. If each egg hatches independently with probability p show that the total number of chicks has Poisson distribution with parameter λp.

Solution:

Exercise 6.11

1. Let $X \sim \text{Binomial}(m,p)$ and $Y \sim \text{Binomial}(n,p)$ be independent. Show that $X+Y \sim \text{Binomial}(m+n,p)$.

   Answer: The PGFs of X and Y are

   $$G_X(t) = (1-p+pt)^m \quad \text{and} \quad G_Y(t) = (1-p+pt)^n$$

   Using the properties of PGFs,

   $$G_{X+Y}(t) = G_X(t)G_Y(t) = (1-p+pt)^m(1-p+pt)^n = (1-p+pt)^{m+n},$$

   which we recognise as the PGF of the Binomial$(m+n,p)$ distribution.

2. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent. Show that $X+Y \sim \text{Poisson}(\lambda+\mu)$.

   Answer: The PGFs of X and Y are respectively $G_X(t) = e^{\lambda(t-1)}$ and $G_Y(t) = e^{\mu(t-1)}$, so the PGF of $Z = X+Y$ is

   $$G_Z(t) = G_X(t)G_Y(t) = e^{(\lambda+\mu)(t-1)}.$$

   We recognise this as the PGF of a Poisson$(\lambda+\mu)$ random variable, so by the inversion theorem for PGFs, $Z \sim \text{Poisson}(\lambda+\mu)$.

3. Let $X \sim \text{Binomial}(n,p)$. Using the PGF of X, show that

   $$\mathbb{E}\left(\frac{1}{1+X}\right) = \frac{1-(1-p)^{n+1}}{(n+1)p}.$$

   Answer: Let $G(t)$ be the PGF of X. Then $G(t) = \mathbb{E}(t^X) = (q+pt)^n$ where $q = 1-p$. Now

   $$\int_0^1 t^x\,dt = \left[\frac{t^{1+x}}{1+x}\right]_0^1 = \frac{1}{1+x},$$

   so

   $$\mathbb{E}\left(\frac{1}{1+X}\right) = \mathbb{E}\left(\int_0^1 t^X\,dt\right) = \int_0^1 \mathbb{E}(t^X)\,dt = \int_0^1 (q+pt)^n\,dt = \frac{1-q^{n+1}}{(n+1)p}$$

6.1.3 Moment generating functions

Definition 6.12
The **moment generating function** (MGF) of a random variable X is the function

$$\begin{array}{rccl} M_X : & \mathbb{R} & \to & [0,\infty] \\ & t & \mapsto & \mathbb{E}(e^{tX}) \end{array}$$

Becaaus e^{tX} is non-negative, its expectation is well-defined and $\mathbb{E}(e^{tX}) \geq 0$. It may however be that $\mathbb{E}(e^{tX})$ is infinite, which limits the usefulness of MGFs in some cases.

Theorem 6.13
If X takes only non-negative integer values then $M_X(t) = G_X(e^t)$ where G_X is the PGF of X.

Proof:

$$M_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\big[(e^t)^X\big] = G_X(e^t),$$

Example 6.14
The MGFs of some fundamental discrete distributions on $\{0, 1, 2, \ldots\}$ can be computed as follows:

$$\begin{aligned}
&X \sim \text{Bernoulli}(p): && G_X(t) = 1 - p + pt && \Rightarrow \quad M_X(t) = 1 - p + pe^t \\
&X \sim \text{Binomial}(n, p): && G_X(t) = (1 - p + pt)^n && \Rightarrow \quad M_X(t) = (1 - p + pe^t)^n \\
&X \sim \text{Poisson}(\lambda): && G_X(t) = e^{\lambda(t-1)} && \Rightarrow \quad M_X(t) = e^{\lambda(e^t - 1)}
\end{aligned}$$

Theorem 6.15 (Properties of MGFs)

1. If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.
2. If $Y = a + bX$, then $M_Y(t) = e^{at}M_X(bt)$

Proof:

1. By independence,
$$M_{X+Y}(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX}e^{tY}) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY}) = M_X(t)M_Y(t).$$
2. For $Y = a + bX$,
$$M_Y(t) = \mathbb{E}\big(e^{t(a+bX)}\big) = e^{at}\mathbb{E}\big(e^{btX}\big) = e^{at}M_X(bt).$$

Corollary 6.16
If $X_1, X_2, \ldots, X_n$ are independent random variables,

$$M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t).$$

Proof: If $X_1, X_2, \ldots, X_n$ are independent, $e^{tX_1}, e^{tX_2}, \ldots, e^{tX_n}$ are also independent, so

$$\begin{aligned}
M_{X_1+X_2+\ldots+X_n}(t) &= \mathbb{E}(e^{t(X_1+x_2+\ldots+X_n)}) \\
&= \mathbb{E}(e^{tX_1})\mathbb{E}(e^{tX_2})\ldots\mathbb{E}(e^{tX_n}) \quad \text{(by independence)} \\
&= M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t)
\end{aligned}$$

Example 6.17 (Normal distribution)
By first considering the MGF of the $N(0, 1)$ distribution, show that the MGF of the $N(\mu, \sigma^2)$ distribution is given by

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

Solution:

Exercise 6.18

1. Let $X \sim \text{Geometric}(p)$ be the distribution of the number of failures before the first success in a sequence of independent Bernoulli trials in which the probability of success is p. Show that $M_X(t) = p/\big[1-(1-p)e^t\big]$ for $|t| < -\log(1-p)$.

   Answer: For $X \sim \text{Geometric}(p)$, $\mathbb{P}(X=k) = (1-p)^k p$ for $k = 0,1,2,\ldots$, so

   $$M_X(t) = \sum_{k=1}^{\infty} e^{tk}\mathbb{P}(X=k) = \sum_{k=1}^{\infty}(1-p)^k pe^{tk} = p\sum_{k=1}^{\infty}\big[(1-p)e^t\big]^k = \frac{pe^t}{1-(1-p)e^t} \qquad \text{for all } t < -\log(1-p).$$

   where we have used the fact that $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for $|r| < 1$.

2. Let $X \sim \text{Poisson}(\lambda)$. Show that $M_X(t) = e^{\lambda(e^t-1)}$.

   Answer: For $X \sim \text{Poisson}(\lambda)$, $\mathbb{P}(X=k) = \lambda^k e^{-\lambda}/x!$ so

   $$M_X(t) = \sum_{k=0}^{\infty} e^{tk}\mathbb{P}(X=k) = \sum_{k=0}^{\infty}\left(\frac{\lambda^k e^{-\lambda}}{k!}\right)e^{tk} = e^{-\lambda}\sum_{i=1}^{\infty}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda(e^t-1)}.$$

Theorem 6.19
If $M_X(t)$ converges in some neighbourhood of 0 then

$$M_X(t) = \sum_{k=0}^{\infty}\frac{\mathbb{E}(X^k)}{k!}t^k.$$

Proof: Using the series expansion of e^{tX} and the linearity of expectation,

$$M_X(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\left(\sum_{k=0}^{\infty}\frac{(tX)^k}{k!}\right) = \sum_{k=0}^{\infty}\frac{\mathbb{E}(X^k)}{k!}t^k$$

The moments of a distribution can be recovered from its MGF by repeated differentiation.

Corollary 6.20
Let X be a random variable and let $M_X(t)$ be its MGF. Then

$$\mathbb{E}(X^k) = M_X^{(k)}(0)$$

where $M_X^{(k)}(0)$ is the kth derivative of $M_X(t)$ evaluated at $t = 0$.

Example 6.21 (Exponential distribution)
Let $X \sim \text{Exponential}(\lambda)$ where $\lambda > 0$ is a rate parameter.

1. Show that $M_X(t) = \dfrac{\lambda}{\lambda - t}$.
2. Use $M_X(t)$ to find the mean and variance of X.

Solution:

Exercise 6.22

1. Let $X \sim \text{Gamma}(\alpha, \beta)$ where β is a rate parameter. Find the MGF of X and use this to show that $\mathbb{E}(X) = \alpha/\beta$ and $\text{Var}(X) = \alpha/\beta^2$.

   Answer: Let $X \sim \text{Gamma}(\alpha, \beta)$. The PDF is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\beta x} \quad \text{where} \quad \Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

   for $x > 0$, and zero otherwise. The MGF is

$$\begin{aligned} M(t) = \mathbb{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{tx} x^{\alpha-1} e^{-\beta x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-1)x}\, dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{(\beta - t)^\alpha} \int_0^\infty x^{\alpha-1} e^{-u}\, du \\ &= \left(\frac{\beta}{\beta - t}\right)^\alpha. \end{aligned}$$

   Elementary calculus yields

$$M'(t) = \frac{\alpha\beta^\alpha}{(\beta - t)^{\alpha+1}} \quad \text{and} \quad M''(t) = \frac{\alpha(\alpha+1)\beta^\alpha}{(\beta - t)^{\alpha+2}}.$$

   Evaluating these at $t = 0$, we get $\mathbb{E}(X) = \alpha/\beta$ and $\mathbb{E}(X^2) = \alpha(\alpha+1)/\beta^2$, and hence $\text{Var}(X) = \alpha(\alpha+1)/\beta^2$ as required.

2. Let $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$ be independent. Show that $X + Y \sim \text{Poisson}(\lambda + \mu)$.

   Answer: The PGFs of X and Y are respectively $G_X(t) = e^{\lambda(t-1)}$ and $G_Y(t) = e^{\mu(t-1)}$, so the PGF of $Z = X + Y$ is

   $$G_{X+Y}(t) = G_X(t)G_Y(t) = e^{\lambda(t-1)}e^{\mu(t-1)} = e^{(\lambda+\mu)(t-1)}.$$

   which we recognise as the MGF of the Poisson($\lambda + \mu$) distribution. (The result follows by the inversion theorem.)

3. Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent. Show that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

   Answer: Because X and Y are independent,

   $$\begin{aligned} M_{X+Y}(t) &= M_X(t)M_Y(t) \\ &= \left[e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2}\right]\left[e^{\mu_2 t + \frac{1}{2}\sigma_2^2 t^2}\right] \text{(as calculated in Example 6.17)} \\ &= e^{(\mu_1+\mu_2)t + \frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2} \end{aligned}$$

   We recognise this as the MGF of the normal distribution with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. By the inversion theorem for MGFs, we conclude that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

4. Let X and Y be independent random variables and let Z be equal to X with probability p, and equal to Y with probability $1 - p$. Use the law of total expectation to show that

   $$M_Z(t) = pM_X(t) + (1-p)M_Y(t).$$

   Answer: By the law of total expectation,

   $$\begin{aligned} M_Z(t) &= \mathbb{E}(e^{tZ}) \\ &= \mathbb{E}(e^{tZ}|X \text{ chosen})\mathbb{P}(X \text{ chosen}) + \mathbb{E}(e^{tZ}|Y \text{ chosen})\mathbb{P}(Y \text{ chosen}) \\ &= \mathbb{E}(e^{tX})\mathbb{P}(X \text{ chosen}) + \mathbb{E}(e^{tY})\mathbb{P}(Y \text{ chosen}) \\ &= pM_X(t) + (1-p)M_Y(t). \end{aligned}$$

6.1.4 Characteristic functions

MGFs can be useful but the expectations that define them may be infinite or not even defined (a notable example is the Cauchy distribution). Characteristic functions do not suffer this disadvantage.

Definition 6.23
The **characteristic function** (CF) of a random variable X is the function

$$\begin{array}{rccc} \phi_X : & \mathbb{R} & \to & \mathbb{C} \\ & t & \mapsto & \mathbb{E}(e^{itX}) \end{array}$$

Remark 6.24

- $\phi(t) = \mathbb{E}(\cos tX) + i\mathbb{E}(\sin tX)$.

- $\phi : \mathbb{R} \to \mathbb{C}$ exists for all $t \in \mathbb{R}$.
- $\phi(t) = M(it)$ provided the latter exists.

Theorem 6.25 (Properties of characteristic functions)

1. If X and Y are independent then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.
2. If $Y = a + bX$ then $\phi_Y(t) = e^{iat}\phi_X(bt)$

Proof:

1. $\phi(0) = \mathbb{E}(e^0) = \mathbb{E}(1) = 1$, and $|\phi(t)| = \left|\int e^{itx}\,dF(x)\right| \leq \int |e^{itx}|\,dF(x) = 1.$
2. By independence, $\phi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX}e^{itY})$, and since
$$e^{itX} = \cos(tX) + i\sin(tX) \quad \text{and} \quad e^{itY} = \cos(tY) + i\sin(tY),$$
we see that
$$e^{itX}e^{itY} = \big[\cos(tX)\cos(tY) - \sin(tX)\sin(tY)\big] + i\big[\cos(tX)\sin(tY) + \sin(tX)\cos(tY)\big]$$
Taking the expectation of both sides, and using the fact that X and Y are independent,
$$\begin{aligned}\phi_{X+Y}(t) &= \mathbb{E}\big(\cos(tX) + i\sin(tX)\big)\mathbb{E}\big(\cos(tY) + i\sin(tY)\big)\\ &= \mathbb{E}(e^{itX})\mathbb{E}(e^{itY})\\ &= \phi_X(t)\phi_Y(t).\end{aligned}$$
3. $\phi_Y(t) = \mathbb{E}\big(e^{it(a+bX)}\big) = \mathbb{E}\big(e^{iat}e^{i(bt)X}\big) = e^{iat}\mathbb{E}(e^{i(bt)X}) = e^{iat}\phi_X(bt).$

The next two theorems are needed to prove the central limit theorem.

The **inversion theorem** states that the distribution a random variable is uniquely determined by its CF, and if another random variable has the same CF then they both have the same distribution.

Theorem 6.26 (Inversion theorem for CFs)

1. ϕ_X uniquely determines the distribution of X.
2. If $\phi_X(t) = \phi_Y(t)$ in some neighbourhood of 0 then X and Y have the same distribution.

It is not always easy to recover a distribution by explicitly inverting the associated CF. Instead they are usually inverted by inspection, where we compare the CF in question with the CFs of various standard distributions.

Example 6.27
Let $X \sim \text{Binomial}(m, p)$ and $Y \sim \text{Binomial}(n, p)$ be independent. Show that

$X + Y \sim \text{Binomial}(m + n, p)$,

Solution:

The **continuity theorem** states that a sequence of distributions $F_1, F_2, \dots$ converges to a limiting distribution F if and only if the corresponding sequence of characteristic functions $\phi_1, \phi_2, \dots$ converges to the characteristic function of F.

Theorem 6.28 (Continuity theorem for CFs)
Let $F_1, F_2, \dots$ and F be CDFs and let $\phi_1, \phi_2, \dots$ and ϕ be the corresponding CFs. Then $F_n \to F$ if and only if $\phi_n \to \phi$ as $n \to \infty$.

6.2 Laws of large numbers

6.2.1 Convergence of random variables

Recall that a sequence of real numbers $x_1, x_2, \dots$ **converges** to the limit x as $n \to \infty$ if for all $\epsilon > 0$, there exists some $N > 0$ with $|x_n - x| < \epsilon$ for all $n > N$. This means that after a certain point in the sequence, all subsequent points are contained within an arbitrarily small neighbourhood of x.

What does it mean to say that a sequence of random variable $X_1, X_2, \dots$ converges to the random variable X?

This is a sequence of **functions** rather than a sequence of real numbers. If $X_n(\omega) \to X(\omega)$ for all outcomes ω in the underlying sample space, we say that X_n converges **pointwise** to X. This is rather restrictive: a more useful notion of convergence might be that $X_n(\omega) \to X(\omega)$ over a set of probability measure one (so that outcomes for which the sequence does not converge occur with probability zero).

There are several notions of convergence for sequences of random variables.

Definition 6.29
Let $X_1, X_2, \dots$ and X be random variables. We say that

1. $X_n \to X$ **almost surely** if $\mathbb{P}(X_n \to X \text{ as } n \to \infty) = 1$,
2. $X_n \to X$ **in mean square** if $\mathbb{E}(|X_n - X|^2) \to 0$ as $n \to \infty$,
3. $X_n \to X$ **in probability** if for all $\epsilon > 0$, $\mathbb{P}(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$,
4. $X_n \to X$ **in distribution** if $F_n(x) \to F(x)$ as $n \to \infty$ for every point x at which F is continuous.

Some of these notions of convergence are stronger than others. This is summarised in the following theorem (which we shall not prove).

Theorem 6.30

1. Convergence almost surely implies convergence in probability.
2. Convergence in mean square implies convergence in probability.
3. Convergence in probability implies convergence in distribution.

6.2.2 Laws of large numbers

Given a set of observations $X_1, X_2, \ldots, X_n$ from the distribution of X, a **law of large numbers** asserts that their sample mean $\bar{X}$ converges in some sense to the expectation $\mathbb{E}(X)$ of the distribution as the number of observations n increases to infinity.

Theorem 6.31 (The weak law of large numbers)
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables, whose common distribution has finite mean μ and finite variance. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{in probability as } n \to \infty.$$

Proof: Let σ^2 denote the variance of X.

- By the linearity of expectation, $\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu.$
- Because the X_i are independent, $\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{\sigma^2}{n}.$
- Applying Chebyshev's inequality, $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$ for all $\epsilon > 0$.

Thus, because σ^2 is finite, it follows that $\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu$ in probability as $n \to \infty$.

Remark 6.32 (Frequentist model)
Consider a random experiment with a finite sample space, and suppose that the experiment is repeated n times under the same conditions. Let A be some random event, and let X_i be the indicator variable of the event that A occurs on the ith trial. The random variables X_i are identically distributed, and their common expectation is $\mathbb{P}(A)$. Because the X_i are indicator variables, their sample mean corresponds to the **relative frequency** of event A over these n repetitions. By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{P}(A) \quad \text{(in some sense) as } n \to \infty.$$

This shows that the frequentist model, in which probability is defined to be the limit of relative frequency as the number of repetitions increases to infinity, is a reasonable one.

Exercise 6.33

1. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having finite mean μ and finite variance. Show that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{in mean square as } n \to \infty.$$

Answer: By independence,

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty.$$

2. Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common mean $\mu < \infty$. Show that the MGF of their sample mean converges to the MGF of the constant random variable μ, and hence that

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{in distribution as} \quad n \to \infty.$$

Note that in contrast to Theorem 6.31, this result does **not** require that the X_i have bounded variance. The price we pay for this relaxation is to accept a weaker type of convergence.

Answer: By the inversion theorem for MGFs, we need only show that the MGF of $\bar{X}_n$ converges to the MGF of the constant μ as $n \to \infty$.

- Let M_X denote the common MGF of the X_i.
- Let $M_{\bar{X}_n}$ denote the MGF of $\bar{X}_n$.

Using the properties of MGFs,

$$\begin{aligned} M_{\bar{X}_n}(t) &= M_{\frac{1}{n}(X_1+X_2+\ldots+X_n)}(t) \\ &= M_{(X_1+X_2+\ldots+X_n)}\left(\frac{t}{n}\right) \\ &= \left[M_X\left(\frac{t}{n}\right)\right]^n \end{aligned}$$

By Theorem 6.19 (with $k = 1$),

$$M_X(t) = 1 + \mu t + o(t) \quad \text{as} \quad t \to 0,$$

so

$$M_{\bar{X}_n}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n = \left[1 + \frac{\mu t}{n} + o\left(\frac{t}{n}\right)\right]^n \to e^{\mu t} \quad \text{as} \quad n \to \infty.$$

where the last step follows by the fact that $e^x = \lim_{n\to\infty}(1 + x/n)^n$. This is the MGF of the constant μ, and the result follows by the inversion theorem for MGFs.

6.3 Central limit theorem

We will need the following definition of the exponential function e^x. This is one of several equivalent definitions.

Definition 6.34 (The compound interest formula)
The exponential function can be defined by

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

6.3.1 The Poisson limit theorem

The Poisson limit theorem asserts that the Binomial(n, p) distribution can be approximated by the Poisson(np) distribution when n is large and p is small.

Theorem 6.35 (Poisson limit theorem)
If $X_n \sim$ Binomial$(n, \lambda/n)$ then the distribution of X_n converges to the Poisson(λ) distribution as $n \to \infty$.

Proof: By the inversion theorem for CFs, it is enough to show that the CF of X_n converges to the CF of the Poisson(λ) distribution as $n \to \infty$.

Recall that

- the CF of the Binomial(n, p) distribution is $\phi(t) = (1 - p + pe^{it})^n$, and
- the CF of the Poisson(λ) distribution is $\phi(t) = e^{\lambda(e^{it}-1)}$.

The CF of $X_n \sim$ Binomial($n, \lambda/n$) is therefore

$$\phi_{X_n}(t) = \mathbb{E}(e^{itX_n}) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^{it}\right)^n = \left[1 + \frac{\lambda(e^{it}-1)}{n}\right]^n$$

By definition 6.34,

$$\phi_{X_n}(t) \to e^{\lambda(e^{it}-1)} \quad \text{as} \quad n \to \infty.$$

This is the CF of the Poisson(λ) distribution, and the result follows by the inversion theorem for CFs.

Let X be a random variable and let ϕ_X be its characteristic function (CF). We can use the series expansion of the exponential function to show that

$$\begin{aligned}\phi_X(t) = \mathbb{E}(e^{itX}) &= \sum_{k=0}^{\infty} \frac{\mathbb{E}(X^k)}{k!}(it)^k \\ &= 1 + i\mathbb{E}(X)t - \frac{\mathbb{E}(X^2)}{2}t^2 - i\frac{\mathbb{E}(X^3)}{6}t^3 + \ldots\end{aligned}$$

When t is small, t^2 is smaller and t^3 is smaller still: as k increases the corresponding terms thus contribute less and less to the value of the sum. We can therefore approximate $\phi(t)$ by taking only the first few terms of the sum, and this approximation becomes increasingly accurate as t tends to zero.

Linear (first order)	$\phi(t) = 1 + i\mathbb{E}(X)t + o(t)$	as $t \to 0$,
Quadratic (second order)	$\phi(t) = 1 + i\mathbb{E}(X)t - \frac{1}{2}\mathbb{E}(X^2)t^2 + o(t^2)$	as $t \to 0$,
Cubic (third order)	$\phi(t) = 1 + \mathbb{E}(X)t + \frac{1}{2}\mathbb{E}(X^2)t^2 + \frac{1}{6}\mathbb{E}(X^3)t^3 + o(t^3)$	as $t \to 0$.

Here $o(t^k)$ represents a quantity that converges to zero faster than t^k in the sense that $o(t^k)/t^k \to 0$ as $t \to 0$. Such quantities are represented in this way to indicate that they can be safely ignored when t is sufficiently small. A second order approximation will be needed to prove the central limit theorem.

6.3.2 The central limit theorem

Let $X_1, X_2, \ldots$ be independent and identically distributed random variables and consider the sequence of partial sums

$$S_n = X_1 + X_2 + \ldots + X_n.$$

By independence, $\mathbb{E}(S_n) = n\mu$ and $\text{Var}(S_n) = n\sigma^2$.

By the law of large numbers, S_n is approximately equal to its mean $n\mu$ when n is large. This however does not say much about the **distribution** of S_n when n is large.

It turns out that if the X_i have finite mean and variance then **irrespective of the distribution of the X_i** the distribution of the standardised sums

$$Z_n = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\text{Var}(S_n)}} \qquad \text{or equivalently} \qquad Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)$$

converges to the standard normal distribution $N(0,1)$ as $n \to \infty$.

Theorem 6.36 (Central limit theorem)
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed with common mean μ and variance σ^2. If μ and σ^2 are both finite, then the distribution of the standardised sums

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)$$

converges to the standard normal distribution $\mathrm{N}(0,1)$ as $n \to \infty$.

Proof: Let $Y_i = \dfrac{X_i - \mu}{\sigma}$. Then $\mathbb{E}(Y_i) = 0$ and $\mathrm{Var}(Y_i) = 1$, and

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

- Let $\phi_Y(t)$ denote the common CF of the Y_i.
- Let $\phi_{Z_n}(t)$ denote the CF of Z_n.

By Taylor's theorem,

$$\phi_Y(t) = \mathbb{E}(e^{itY}) = \sum_{j=0}^{k} \frac{\mathbb{E}(Y^j)}{j!}(it)^j + o((it)^k) \qquad \text{as} \quad t \to 0.$$

Taking the first three terms in the Taylor expansion of $\phi_Y(t)$, since $\mathbb{E}(Y) = 0$ and $\mathbb{E}(Y^2) = 1$ we get

$$\phi_Y(t) = 1 - \frac{1}{2}t^2 + o(t^2) \quad \text{as} \quad t \to 0$$

By the properties of CFs,

$$\begin{aligned}
\phi_{Z_n}(t) &= \phi_{\frac{1}{\sqrt{n}}(Y_1+Y_2+\ldots+Y_n)}(t) \\
&= \phi_{Y_1+Y_2+\ldots+Y_n}\left(\frac{t}{\sqrt{n}}\right) \\
&= \left[\phi_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n \\
&= \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \quad \text{as } t \to 0 \\
&\to e^{-\frac{1}{2}t^2} \quad \text{as } n \to \infty,
\end{aligned}$$

where the last step follows by definition 6.34. This is the CF of the $\mathrm{N}(0,1)$ distribution, and the result follows by the inversion theorem for CFs.

Example 6.37
The **Erlang distribution** with parameters $n \in \mathbb{N}$ and $\lambda > 0$ is defined to be the sum of n independent and identically distributed random variables $X_1, X_2, \ldots, X_n$, where each X_i is exponentially distributed with (rate) parameter λ. Show that if $S_n \sim \mathrm{Erlang}(n, \lambda)$, then the random variable

$$Z_n = \frac{\lambda S_n - n}{\sqrt{n}}$$

has approximately the standard normal distribution when n is large.

Solution:

Exercise 6.38

1. The continuous uniform distribution on (a, b) has the following PDF:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

Use the central limit theorem to deduce the approximate distribution of the sample mean of n independent observations from this distribution when n is large.

Answer: The mean is

$$\mu = \int_a^b \frac{x}{b-a}\,dx = \frac{a+b}{2},$$

and the second moment is

$$\mu_2 = \int_a^b \frac{x^2}{b-a}\,dx = \frac{a^2+ab+b^2}{3},$$

so the variance is

$$\sigma^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{(b-a)^2}{12}$$

By the central limit theorem, if X is a random variable with mean μ and variance σ^2, the distribution of the sample mean $\bar{X}$ of a random sample of n independent observations is approximately $N(\mu, \frac{\sigma^2}{n})$, the approximation being better for larger n. In this case, the approximate distribution of $\bar{X}$ is $N\left(\frac{a+b}{2}, \frac{(b-a)^2}{12n}\right)$.

2. The exponential distribution with scale parameter $\theta > 0$ has the following PDF:

$$f(x) = \begin{cases} \dfrac{1}{\theta}e^{-x/\theta} & x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Use the central limit theorem to deduce the approximate distribution of the sample mean of n independent observations from this distribution when n is large.

Answer:

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{\theta}\int_0^\infty xe^{-x/\theta}\,dx = \theta, \\ \mathbb{E}(X^2) &= \frac{1}{\theta}\int_0^\infty x^2e^{-x/\theta}\,dx = 2\theta^2 \\ \text{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \theta^2. \end{aligned}$$

By the CLT, the distribution of $\bar{X}$ is approximately $\text{N}(\mu, \frac{\sigma^2}{n})$, the approximation being better for larger n. In this case, the approximate distribution of $\bar{X}$ is $\text{N}\big(\theta, \theta^2/n\big)$.

3. We perform a sequence of independent Bernoulli trials, each with probability of success p, until a fixed number r of successes is obtained. The total number of failures Y (up to the rth succes) has the **negative binomial** distribution with parameters r and p, so the PMF of Y is

$$\mathbb{P}(Y = k) = \binom{k+r-1}{k}(1-p)^k p^r, \qquad k = 0, 1, 2, \ldots$$

Using the fact that Y can be written as the sum of r independent geometric random variables, show that this distribution can be approximated by a normal distribution when r is large.

Answer: If $Y \sim \text{NB}(r, p)$, we can write

$$Y = X_1 + X_2 + \ldots + X_r \qquad \text{where} \quad X_i \sim \text{Geometric}(p).$$

Let $X \sim \text{Geometric}(p)$. Since $\text{Var}(X) < \infty$, it follows by the central limit theorem that

$$\frac{Y - r\mathbb{E}(X)}{\sqrt{r\text{Var}(X)}} \to \text{N}(0, 1) \quad \text{in distribution as } r \to \infty.$$

In fact, since $\mathbb{E}(X) = (1-p)/p$ and $\text{Var}(X) = (1-p)/p^2$, we see that Y can be approximated by the $\text{N}\left(\dfrac{r(1-p)}{p}, \dfrac{r(1-p)}{p^2}\right)$ distribution as $r \to \infty$.

Chapter 7 Estimation

An important problem in statistics is to estimate the distribution of a random variable from a set of observations. For example, certain observations may be known (or assumed) to have an exponential distribution and we want to estimate the rate parameter λ of the distribution, or observations may be taken from a normal distribution and we want to estimate the mean μ and variance σ^2 of the distribution.

7.1 Random samples

First we define random vectors. These are simply vectors of random variables.

Definition 7.1
Let $X_1, X_2, \ldots, X_n$ be random variables defined on the same probability space. The vector-valued function

$$\begin{array}{rccl} \mathbf{X}: & \Omega & \longrightarrow & \mathbb{R}^n \\ & \omega & \mapsto & \big[X_1(\omega), X_2(\omega), \ldots, X_n(\omega)\big] \end{array}$$

is called a **random vector** of size n. The individual X_i are called the **component variables** of $\mathbf{X}$ and a vector $\mathbf{x} = (x_1, x_2 \ldots, x_n) \in \mathbb{R}^n$ is called a **realisation** of $\mathbf{X}$, where x_i is the value taken by the corresponding component variable X_i.

The behaviour of a random vector is completely described by the joint CDF of its component variables.

Definition 7.2
Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random vector.

1. The **joint distribution** of $\mathbf{X}$ is the function $\mathbb{P}_{\mathbf{X}}(B) = \mathbb{P}(\mathbf{X} \in B)$ defined on subsets of $\mathbb{R}^n$.

2. The **joint CDF** of $\mathbf{X}$ is the function $F_{\mathbf{X}} : \mathbb{R}^n \to [0, 1]$ given by

$$F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \leq \mathbf{x}) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

In general the component variables X_i might depend on each other and have different distributions. Our analysis is greatly simplified by assuming that the X_i are **independent** and **identically distributed**.

Definition 7.3
Let X be a random variable. A **random sample from the distribution of** X is a random vector with the property that the component variables X_i are independent and have the same distribution as X.

By independence the joint CDF of a random sample is just the product of its marginal CDFs and because each X_i has the same distribution as X, this can be expressed entirely in terms of the CDF of X.

Lemma 7.4
Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a random sample from the distribution of X. Then the joint CDF and joint PMF/PDF of $\mathbf{X}$ can be written respectively as

$$F_{\mathbf{X}}(\mathbf{x}) = \textstyle\prod_{i=1}^n F_X(x_i) \qquad \text{and} \qquad f_{\mathbf{X}}(\mathbf{x}) = \textstyle\prod_{i=1}^n f_X(x_i)$$

where F_X and f_X are the CDF and PMF/PDF of X respectively, and $\boldsymbol{x} = (x_1, x_2, \dots, x_n)$ is a realisation of the sample..

Proof:

$$\begin{aligned}
F_{\mathbf{X}}(\mathbf{x}) &= \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \\
&= \mathbb{P}(X_1 \leq x_1)\mathbb{P}(X_2 \leq x_2) \dots \mathbb{P}(X_n \leq x_n) \quad \text{(by independence)} \\
&= F_X(x_1)F_X(x_2) \cdots F_X(x_n) \quad \text{(because each } X_i \text{ has the same distribution as } X\text{)} \\
&= \textstyle\prod_{i=1}^n F_X(x_i).
\end{aligned}$$

7.2 Statistical models

To estimate the underlying distribution F_X from a random sample $X_1, X_2, \dots, X_n$ of observations from the distribution of X, let us assume that F_X belongs to some **parametric family** of distributions.

Definition 7.5
A **statistical model** is a parametric family of CDFs,

$$\mathcal{M} = \big\{F(x\,;\theta) : \theta \in \Theta\big\}$$

where θ is a vector of parameters, and Θ is called the **parameter space**.

Example 7.6

- The family of exponential distributions:

$$\mathcal{M} = \big\{F(x;\lambda) : \lambda > 0\big\} \text{ where } F(x;\lambda) = 1 - e^{-\lambda x} \text{ for } x > 0.$$

- The family of uniform distributions:

$$\mathcal{M} = \big\{F(x;a,b) : a < b\big\} \text{ where } F(x;a,b) = \frac{x-a}{b-a} \text{ for } a \leq x \leq b.$$

Remark 7.7
Let $\mathcal{M} = \big\{F(x\,;\theta) : \theta \in \Theta\big\}$ be a statistical model and suppose that $F_X \in \mathcal{M}$. Then estimating F_X amounts to estimating the "true" value of the parameter θ.

Definition 7.8

1. To estimate a particular value for θ is called **point estimation**.

2. To estimate a range of values for θ is called **interval estimation**.

3. To assert whether or not θ lies in some range is called **hypothesis testing**.

7.2.1 Estimators

Definition 7.9
Let X be a random variable on Ω and let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from the distribution of X. Any transformation $T : \mathbb{R}^n \to \mathbb{R}$ of $\mathbf{X}$ is called a **sample statistic**.

There are two interpretations:

1. $T(\boldsymbol{X})$ is a random variable on $(\Omega, \mathbb{P})$,

$$\begin{array}{rccc} T(\boldsymbol{X}): & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & T\big[\boldsymbol{X}(\omega)\big] \end{array}$$

2. T is a random variable on $(\mathbb{R}^n, \mathbb{P}_{\mathbf{X}})$,

$$\begin{array}{rccc} T: & \mathbb{R}^n & \to & \mathbb{R} \\ & \boldsymbol{x} & \mapsto & T(\boldsymbol{x}) \end{array}$$

where $\mathbb{P}_{\mathbf{X}}(B) = \mathbb{P}(\mathbf{X} \in B)$ is the distribution of $\mathbf{X}$.

For the latter interpretation to be useful, the distribution of the sample statistic T over subsets of $\mathbb{R}$ must be deduced from the distribution of the random sample $\mathbf{X}$ over subsets of $\mathbb{R}^n$.

Definition 7.10

1. A statistic $T(\mathbf{X})$ used to estimate an unknown parameter θ is called an **estimator** of θ.
2. For any particular sample realisation $\mathbf{x}$ the value $T(\boldsymbol{x})$ is called an **estimate** of θ.

Note that estimators of θ are often denoted by $\hat{\theta}$.

Example 7.11
A coin has an unknown probability θ of landing on heads. An appropriate statistical model for this experiment is the family of **Bernoulli** distributions:

$$\mathcal{M} = \big\{F(x\,;\theta) : 0 \le \theta \le 1\big\} \qquad \text{where} \quad F(x;\theta) = \begin{cases} 0 & x < 0, \\ 1-\theta & 0 \le x < 1, \\ 1 & x \ge 1. \end{cases}$$

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from the Bernoulli(θ) distribution. An estimator for θ is provided by the following statistic, called the **sample mean**:

$$\begin{array}{rccc} T: & \mathbb{R}^n & \to & \mathbb{R} \\ & \boldsymbol{x} & \mapsto & \displaystyle\frac{1}{n}\sum_{i=1}^n x_i. \end{array}$$

The distribution of T is determined by the common distribution of the individual observations X_i. In this case $X_i \sim$ Bernoulli(θ) so

$$\mathbb{P}\left(T = \frac{k}{n}\right) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \qquad \text{for } k = 0, 1, 2, \ldots, n.$$

7.3 The method of moments

Perhaps the simplest estimators are based on the **method of moments**.

Let X be a random variable, let $f(x;\theta)$ be its PMF/PDF and consider its kth moment:

$$\begin{aligned}\mathbb{E}(X^k) &= \sum_{i=1}^{\infty} x_i^k f(x_i;\theta) \qquad \text{(discrete case)},\\ \mathbb{E}(X^k) &= \int_{-\infty}^{\infty} x^k f(x;\theta)\,dx \quad \text{(continuous case)}.\end{aligned}$$

Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X. To estimate k parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$, we equate the expressions for the first k moments with the first k empirical moments, then solve the resulting system of equations with respect to $\theta_1, \theta_2, \ldots, \theta_k$. The number of equations must be equal to the number of unknown parameters we wish to estimate: these are known as the **moment equations**.

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \mathbb{E}(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \ldots, \quad \mathbb{E}(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Example 7.12
Let $X_1, X_2, \ldots, X_n$ be a random sample from the continuous Uniform$(0,\theta)$ distribution, where

$\theta > 0$ is unknown. Find an estimator of θ using the method of moments.

Solution:

Example 7.13
Let $X_1, X_2, \ldots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution. Find estimators of μ and σ^2 using the method of moments.

Solution:

Exercise 7.14

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli(θ) distribution. Find an estimator of θ using the method of moments.

   Answer: The first moment of the Bernoulli(θ) distribution is θ. Equating this to the first empirical moment (i.e. the sample mean),

   $$\hat{\theta}_{\text{MME}} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

   The MME of θ is thus the proportion of successes observed in n trials.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Exponential(λ) distribution, where $\lambda > 0$ is an unknown rate parameter. Find an estimator of λ using the method of moments.

   Answer: Let $X \sim \text{Exponential}(\lambda)$. Then $\mathbb{E}(X) = 1/\lambda$, so the first moment equation is

   $$\frac{1}{\lambda} = \bar{X}.$$

   Solving for λ, we obtain $\hat{\lambda}_{\text{MME}} = \bar{X}^{-1}$.

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson(λ) distribution, where $\lambda > 0$ is unknown. Find an estimator of λ using the method of moments.

   Answer: Let $X \sim \text{Poisson}(\lambda)$. The expected value of X is $\mathbb{E}(X) = \lambda$, so the first moment equation is simply

   $$\lambda = \bar{X}.$$

   Solving for λ, we obtain $\hat{\lambda}_{\text{MME}} = \bar{X}$.

7.4 Bias and mean squared error

Let X be a random variable and let $F_X(x;\theta)$ denote its CDF where $\theta \in \Theta$ is an unknown parameter. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from the distribution of X and let $T(\mathbf{X})$ be an estimator of θ. Intuitively, we would like the estimate $T(\mathbf{x})$ to be "close" to the true value θ for "most" sample realisations $\mathbf{x} \in \mathbb{R}^n$.

7.4.1 Bias

Definition 7.15
The **bias** of an estimator T is the expected difference between $T(\mathbf{X})$ and the true parameter value θ,

$$\text{Bias}(T;\theta) = \mathbb{E}\big[T(\mathbf{X}) - \theta\big].$$

If $\text{Bias}(T;\theta) = 0$ for all $\theta \in \Theta$ then T is said to be an **unbiased** estimator of θ.

Example 7.16
Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a random sample from the $N(\mu, \sigma^2)$ distribution and consider the sample mean estimator of σ^2 defined by

$$T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{where} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

1. Show that T is a biased estimator of σ^2 and find its bias.
2. Show how can T be modified to produce an unbiased estimator of σ^2.

Solution:

Exercise 7.17
Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ be a random sample from the Uniform$(0, \theta)$ distribution, where $\theta > 0$ is unknown.

1. Find the bias of $T_1(\mathbf{X}) = 2\bar{X}$ as an estimator of θ.

2. Find the bias of $T_2(\mathbf{X}) = \max\{X_1, X_2, \dots, X_n\}$ as an estimator of θ.
3. How can T_2 be modified to produce an unbiased estimator of θ?

Answer:

1. T_1 is an **unbiased** estimator of θ because
$$\text{Bias}(T_1) = \mathbb{E}(T_1 - \theta) = \mathbb{E}(T_1) - \theta = 2\mathbb{E}(\bar{X}) - \theta = 2(\theta/2) - \theta = 0.$$
2. Because the X_i are independent, the CDF of T_2 is
$$\mathbb{P}(T_2 \leq t) = \begin{cases} 0 & t \leq 0, \\ (t/\theta)^n & 0 < t < \theta, \\ 1 & t \geq \theta. \end{cases}$$
A simple calculation shows that
$$\mathbb{E}(T_2) = \left(\frac{n}{n+1}\right)\theta.$$
Hence T_2 is a **negatively biased** estimator of θ:
$$\text{Bias}(T_2) = \mathbb{E}(T_2 - \theta) = \left(-\frac{1}{n+1}\right)\theta.$$
3. A **bias correction** can be applied to T_2 to give an unbiased estimator $T_3 = \left(\dfrac{n+1}{n}\right) T_2$.

7.4.2 Mean squared error

- The bias of T as an estimator of θ is the **mean estimation error** $\mathbb{E}\big[T - \theta\big]$.
- The magnitude of the error can be quantified by the **mean squared estimation error** $\mathbb{E}\big[(T - \theta)^2\big]$.

Definition 7.18
The **mean squared error** of T as an estimator of θ is
$$\text{MSE}(T;\theta) = \mathbb{E}\big[\big(T(\mathbf{X}) - \theta\big)^2\big].$$

The accuracy of an unbiased estimator can be quantified by its variance. For biased estimators, the MSE also takes the bias into account. The following result is easily proved by writing $T - \theta$ as $[T - \mathbb{E}(T)] + [\mathbb{E}(T) - \theta]$ then expanding the square and applying the linearity of expectation.

Theorem 7.19
$\text{MSE}(T;\theta) = \text{Var}(T) + \text{Bias}(T;\theta)^2$.

As the following example shows, a biased estimator with small variance is often “better” than an unbiased estimator with large variance.

Example 7.20
Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. The sample mean $\bar{X}$ is an unbiased estimator of μ, while the following statistic is a biased estimator of μ:
$$T(X_1, X_2, \dots, X_n) = \frac{1}{n+1}\sum_{i=1}^{n} X_i.$$

1. Find $\text{MSE}(\bar{X};\mu)$ and $\text{MSE}(T;\mu)$.

2. Find a condition involving μ and σ under which $\text{MSE}(T;\mu) < \text{MSE}(\bar{X};\mu)$.

Solution:

Exercise 7.21

1. Let $X_1, X_2, ..., X_n$ be a random sample of observations from the Uniform$[\theta, \theta+1]$ distribution, where θ is an unknown parameter.

    1. Show that the sample mean $\bar{X}$ is a biased estimator of θ, and find its bias.
    2. Find the variance and mean squared error of $\bar{X}$.
    3. How can $\bar{X}$ be modified to produce an unbiased estimator of θ?

    Answer:

    1. The sample mean is biased because $\mathbb{E}(\bar{X}) = \theta + \frac{1}{2} \neq \theta$. The bias is given by

$$\text{Bias}(\bar{X}) = \mathbb{E}(\bar{X} - \theta) = \mathbb{E}(\bar{X}) - \theta = \frac{1}{2}.$$

2. The variance and mean squared error of $\bar{X}$ as an estimator of θ are

$$\begin{aligned}\operatorname{Var}(\bar{X}) &= \frac{\operatorname{Var}(X)}{n} = \frac{1}{12n}.\\ \operatorname{MSE}(\bar{X}) &= \operatorname{Var}(\bar{X}) + \operatorname{Bias}(\bar{X})^2 = \frac{1}{12n} + \frac{1}{4} = \frac{(1+3n)}{12n}.\end{aligned}$$

3. $\bar{X} - \frac{1}{2}$ is an unbiased estimator of θ, because $\mathbb{E}(\bar{X} - \frac{1}{2}) = \mathbb{E}(\bar{X}) - 1/2 = \theta$.

2. Let X be a continuous random variable with the following PDF, where $\alpha > 0$ is known but $\theta > 0$ is unknown.

$$f(x;\theta) = \begin{cases} \dfrac{\alpha x^{\alpha-1}}{\theta^\alpha} & 0 \leq x \leq \theta, \\ 0 & \text{otherwise.} \end{cases}$$

A random sample $X_1, X_2, \ldots, X_n$ is taken from the distribution of X. An estimator of θ is provided by

$$T = \max\{X_1, X_2, \ldots, X_n\}.$$

(a) Show that

$$\mathbb{E}(T) = \left(\frac{n\alpha}{n\alpha+1}\right)\theta \quad \text{and} \quad \operatorname{Var}(T) = \left(\frac{n\alpha}{n\alpha+2}\right)\theta^2 - \left(\frac{n\alpha}{n\alpha+1}\right)^2\theta^2.$$

Answer: To find $\mathbb{E}(T)$ and $\operatorname{Var}(T)$ we first need to find the distribution of T. Let F_X denote the CDF of X.

$$F_X(x;\theta) = \begin{cases} 0 & x < 0 \\ \left(\dfrac{x}{\theta}\right)^\alpha & 0 \leq x \leq \theta \\ 1 & x > \theta. \end{cases}$$

It is easy to show that the CDF of $T = \max\{X_1, X_2, \ldots, X_n\}$ is

$$F_T(v) = \mathbb{P}(T \leq t) = \big[F_X(t)\big]^n.$$

In this case,

$$F_T(t;\theta) = \begin{cases} 0 & t < 0, \\ \left(\dfrac{t}{\theta}\right)^{n\alpha} & 0 \leq t \leq \theta, \\ 1 & t > \theta, \end{cases}$$

Hence the PDF of T is

$$f_T(t;\theta) = \begin{cases} \dfrac{n\alpha}{\theta^{n\alpha}} t^{(n\alpha-1)} & 0 \leq t \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

The expected value and variance of $T = \max\{X_1, X_2, \ldots, X_n\}$ are computed as follows:

$$\begin{aligned}\mathbb{E}(T) &= \frac{n\alpha}{\theta^{n\alpha}} \int_0^\theta t^{n\alpha}\,dt = \frac{n\alpha}{\theta^{n\alpha}} \left[\frac{t^{n\alpha+1}}{(n\alpha+1)}\right]_0^\theta = \frac{n\alpha}{(n\alpha+1)}\theta,\\ \mathbb{E}(T^2) &= \int_0^\theta \frac{n\alpha}{\theta^{n\alpha}} t^{(n\alpha+1)}\,dt = \frac{n\alpha}{\theta^{n\alpha}} \left[\frac{t^{n\alpha+2}}{n\alpha+2}\right]_0^\theta = \frac{n\alpha}{(n\alpha+2)}\theta^2,\\ \operatorname{Var}(T) &= \frac{n\alpha}{(n\alpha+2)}\theta^2 - \frac{n^2\alpha^2}{(n\alpha+1)^2}\theta^2.\end{aligned}$$

(b) Show that T is a biased estimator of θ.
Answer: T is a biased estimator of θ because

$$\mathbb{E}(T) = \left(\frac{n\alpha}{n\alpha+1}\right)\theta,$$

so $\mathbb{E}(T) \neq \theta$. The bias is

$$\text{Bias}(T) = \mathbb{E}(T-\theta) = \mathbb{E}(T) - \theta = \frac{n\alpha}{(n\alpha+1)}\theta - \theta = \frac{-\theta}{(n\alpha+1)}$$

(c) Find a multiple of T that yields an unbiased estimator of θ.
Answer: The estimator $\left(\frac{n\alpha+1}{n\alpha}\right)T$ is an unbiased estimator of θ because

$$\mathbb{E}\left[\left(\frac{n\alpha+1}{n\alpha}\right)T\right] = \left(\frac{n\alpha+1}{n\alpha}\right)\mathbb{E}(T) = \left(\frac{n\alpha+1}{n\alpha}\right)\frac{n\alpha}{(n\alpha+1)}\theta = \theta.$$

(d) Find the mean squared error of T.
Answer: The mean squared error of T as an estimator of θ is

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{Var}(\hat{\theta}) + \text{Bias}(T)^2 \\
&= \frac{n\alpha}{(n\alpha+2)}\theta^2 - \frac{n^2\alpha^2}{(n\alpha+1)^2}\theta^2 + \frac{\theta^2}{(n\alpha+1)^2} \\
&= \frac{2\theta^2}{(n\alpha+2)(n\alpha+1)}
\end{aligned}$$

7.5 Large samples

Intuitively we would like an estimator to improve as the sample size increases. The **asymptotic** behaviour of a quantity refers to the way it behaves as another quantity tends to some limit. We now write T_n to represent an estimator defined on a random sample of size n and consider its behaviour as $n \to \infty$.

Definition 7.22
Let T_n be an esimator of some unknown parameter $\theta \in \Theta$.

1. T_n is said to be **asymptotically unbiased** if $\mathbb{E}(T_n) \to \theta$ as $n \to \infty$ for all $\theta \in \Theta$.
2. T_n is said to be **consistent** if $T_n \to \theta$ in probability as $n \to \infty$ for all $\theta \in \Theta$.
3. T_n is said to be **asymptotically normal** if the distribution of $\sqrt{n}(T_n - \theta)$ converges to a normal distribution as $n \to \infty$ for all $\theta \in \Theta$.

Lemma 7.23
If T_n is an unbiased estimator of θ and $\text{Var}(T_n) \to 0$ as $n \to \infty$ then T_n is a consistent estimator of θ.

Proof: This follows by Chebyshev's inequality: for all $\epsilon > 0$,

$$\mathbb{P}(|T_n - \mathbb{E}(T_n)| > \epsilon) \leq \frac{\text{Var}(T_n)}{\epsilon^2}.$$

Since T_n is unbiased, $\mathbb{E}(T_n) = \theta$ so $\mathbb{P}(|T_n - \theta| > \epsilon) \to 0$ as $n \to \infty$, as required.

Example 7.24
The reading on a voltmeter connected to a test circuit is a random variable X that has a uniform distribution over the interval $(\theta, \theta+1)$, where θ is unknown. Let $X_1, X_2, ..., X_n$ be a random sample of observations from the voltmeter. Show that

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{1}{2}\right)$$

is a consistent estimator for θ.

Solution:

Exercise 7.25

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Uniform$(0,\theta)$ distribution. Show that

$$T_n = \max\{X_1, X_2, \ldots, X_n\}$$

is an asymptotically unbiased estimator of θ.

Answer: The PDF of $T_n = \max\{X_1, X_2, \ldots, X_n\}$ is

$$f(t;\theta) = \frac{nt^{n-1}}{\theta^n} \quad \text{for} \quad 0 < t < \theta \quad \text{(zero otherwise).}$$

The expected value of T_n is

$$\mathbb{E}(T_n) = \int_0^\theta t\frac{nt^{n-1}}{\theta^n}\,dv = \frac{n}{(n+1)\theta^n}\left[t^{n+1}\right]_0^\theta = \frac{n\theta}{n+1} = \left(1 - \frac{1}{n+1}\right)\theta.$$

The bias of T_n is

$$\text{Bias}(T_n) = \mathbb{E}(T_n - \theta) = -\frac{\theta}{n+1} \to 0 \text{ as } n \to \infty.$$

and because this holds for all $\theta > 0$ it follows that T_n is an asymptotically unbiased estimator for θ.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli(θ) distribution. Show that the sample mean

$$T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is a consistent estimator of θ.

Answer: This follows by the weak law of large numbers.

In more detail, because $\mathbb{E}(T_n) = \theta$ and $\text{Var}(T_n) = \dfrac{\theta(1-\theta)}{n}$, it follows by Chebyshev's inequality that

$$\mathbb{P}(|T_n - \theta| > \epsilon) \leq \frac{\theta(1-\theta)}{n\epsilon^2}.$$

Since $\theta \in [0,1]$ is bounded, we have for every $\epsilon > 0$ that

$$\mathbb{P}(|T_n - \theta| > \epsilon) \to 0 \text{ as } n \to \infty.$$

Thus $T_n \to \theta$ in probability as $n \to \infty$. This holds for all $\theta \in [0,1]$, so T_n is a consistent estimator of θ.

3. Let X be a random variable with finite mean μ and finite variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X and consider the sample mean estimators of μ and σ^2,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

(a) Show that $\hat{\sigma}_n^2$ is an asymptotically unbiased estimator of σ^2.

Answer: As we have seen,

$$\mathbb{E}(\hat{\sigma}_n^2) = \left(\frac{n-1}{n}\right)\sigma^2 \quad \text{and} \quad \text{Bias}(\hat{\sigma}_n^2) = -\frac{\sigma^2}{n}.$$

Hence $\text{Bias}(\hat{\sigma}_n^2) \to 0$ as $n \to \infty$, and because this holds for all $\sigma^2 > 0$ we have that $\hat{\sigma}_n^2$ is an asymptotically unbiased estimator for σ^2.

(b) Show that $\bar{X}_n$ is a consistent and asymptotically normal estimator of μ.

Answer: Because σ^2 is finite, by the law of large numbers we have $\bar{X}_n \to \mu$ in probability as $n \to \infty$, and by the central limit theorem,

$$\sqrt{n}(\bar{X}_n - \mu) \to N(0, \sigma^2) \quad \text{in distribution as } n \to \infty.$$

These hold for all $\mu \in \mathbb{R}$, so we have shown that $\bar{X}_n$ is a consistent and asymptotically normal estimator for μ.

(c) If $\mathbb{E}(X^4) < \infty$ show that $\hat{\sigma}_n^2$ is a consistent and asymptotically normal estimator of σ^2.

Answer: For $\hat{\sigma}_n^2$ we write $X_i - \bar{X}_n = (X_i - \mu) - (\bar{X}_n - \mu)$, from which we obtain

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 + (\bar{X}_n - \mu)^2$$

For the first term we have $\mathbb{E}\big[(X_i - \mu)^2\big] = \sigma^2$ and $\text{Var}\big[(X_i - \mu)^2\big] = \mu_4 - \sigma^4$ where $\mu_4 = \mathbb{E}(X^4)$. By hypothesis these are both finite, so by the law of large numbers applied to the random variables $(X_i - \mu)^2$,

$$\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \to \sigma^2 \quad \text{in probability as } n \to \infty.$$

For the second term, by Markov's inequality and the fact that $\mathbb{E}(\bar{X}_n) = \mu$ and $\text{Var}(\bar{X}_n) = \sigma^2/n$,

$$\mathbb{P}((\bar{X}_n - \mu)^2 > \epsilon) \leq \frac{\mathbb{E}\big[(\bar{X}_n - \mu)^2\big]}{\epsilon} = \frac{\text{Var}(\bar{X}_n)}{\epsilon} = \frac{\sigma^2}{n\epsilon}.$$

Hence $(\bar{X}_n - \mu)^2 \to 0$ in probability as $n \to \infty$, and thus we have shown that $\hat{\sigma}_n^2 \to \sigma^2$ in probability as $n \to \infty$. Since this holds for every $\sigma^2 > 0$ we conclude that $\hat{\sigma}_n^2$ is a consistent estimator for σ^2.

To show that $\hat{\sigma}_n^2$ is asymptotically normal, let us write

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\right) - \sqrt{n}(\bar{X}_n - \mu)^2$$

As above, we can use Markov's inequality to show that the second term converges to zero in probability as $n \to \infty$:

$$\mathbb{P}(\sqrt{n}(\bar{X}_n - \mu)^2 > \epsilon) \leq \frac{\mathbb{E}\big[\sqrt{n}(\bar{X}_n - \mu)^2\big]}{\epsilon} = \frac{\sqrt{n}\mathrm{Var}(\bar{X}_n)}{\epsilon} = \frac{\sigma^2}{\sqrt{n}\epsilon} \to 0 \quad \text{as } n \to \infty.$$

For the first term, we have $\mathbb{E}\big[(X_i - \mu)^2\big] = \sigma^2$ and $\mathrm{Var}\big[(X_i - \mu)^2\big] = \mu_4 - \sigma^4$, both of which are finite, so by the central limit theorem applied to the random variables $(X_i - \mu)^2$

$$\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \to N(0, \mu_4 - \sigma^4) \quad \text{in distribution as } n \to \infty.$$

Because this holds for all $\sigma^2 > 0$ we conclude that $\hat{\sigma}_n^2$ is an asymptotically normal estimator for σ^2.

Chapter 8 Likelihood

8.1 Likelihood

Let $f(x;\theta)$ denote the PMF/PDF of a random variable X. We have so far considered $f(x;\theta)$ to be a function of x and θ as a fixed parameter. We now change the emphasis and regard x as a fixed observation and θ as a variable parameter. To underline this new emphasis we refer to $f(x;\theta)$ as the **likelihood function** of θ.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from the distribution of X, let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a realisation of the sample and let Θ denote the parameter space.

Definition 8.1
Given $\mathbf{X} = \mathbf{x}$, the **likelihood function** $L : \Theta \to [0, \infty)$ is

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta).$$

Note that the right-hand side is just the joint PMF/PDF of the random sample. Until now we have interpreted this as the probability that the observed sample is $\mathbf{x}$ with θ a fixed parameter. For statistical inference, we instead consider the likelihood $L(\theta; \mathbf{x})$ that θ is the true parameter value given the fixed observations $\mathbf{x}$: such observations are often referred to as **data**.

The value of θ which maximises the likelihood function provides an estimator of the true parameter value, subject to the following regularity conditions.

Condition 8.2
Our statistical model $\mathcal{M} = \{f(x;\theta) : \theta \in \Theta\}$ satisfies the following conditions.

1. PDFs are distinct, i.e. $\theta \neq \theta' \implies f(x;\theta) \neq f(x;\theta')$.
2. PDFs have a common support for all $\theta \in \Theta$.
3. The true value of θ is an interior point of the parameter space Θ.

We shall not consider these conditions in detail.

8.2 Maximum likelihood estimators

Definition 8.3
An estimator $T = T(\mathbf{X})$ is said to be a **maximum likelihood estimator** (MLE) of θ if

$$T = \operatorname*{argmax}_{\theta \in \Theta} L(\theta; \mathbf{X}),$$

which means that T is a value of θ which maximises the likelihood function $L(\theta, \mathbf{X})$ over the parameter space Θ.

To find a value of θ that maximises $L(\theta; \mathbf{X})$ we might differentiate it with respect to θ, then set the resulting expression to zero and solve for θ. Computing derivatives of products such as $\prod_{i=1}^{n} f(X_i; \theta)$ is not always straightforward however, so we work with the **log-likelihood** function whenever possible.

Definition 8.4
Given $\mathbf{X} = \mathbf{x}$, the **log-likelihood function** $\ell : \Theta \to [0, \infty)$ is

$$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^{n} \log f(x_i; \theta).$$

Because log is a strictly increasing function, a value of θ that maximizes $\ell(\theta; \mathbf{x})$ coincides with a value of θ that maximizes $L(\theta; \mathbf{x})$. A maximum likelihood estimate of $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ can therefore be obtained by solving the equations

$$\frac{\partial \ell}{\partial \theta_1} = 0, \quad \frac{\partial \ell}{\partial \theta_2} = 0, \quad \ldots \quad , \frac{\partial \ell}{\partial \theta_k} = 0.$$

Example 8.5
Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli(θ) distribution where $0 < \theta < 1$ is unknown. Find the MLE of θ.

Solution:

Example 8.6
Let $X_1, X_2, \ldots, X_n$ be a random sample from the Uniform$[0, \theta]$ distribution where $\theta > 0$ is unknown. Find the maximum likelihood estimator of θ.

Solution:

Example 8.7
Let $X_1, X_2, \dots, X_n$ be a random sample from the $N(\mu, \sigma^2)$ distribution, where μ and σ^2 are both unknown. Find the MLEs of μ and σ^2.

Solution:

Exercise 8.8

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Exponential(λ) distribution, where the rate parameter $\lambda > 0$ is unknown. Find the MLE of λ.

   Answer: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n\}$ be a realization of the sample. The PDF of the Exponential(λ) distribution is

$$f(x;\lambda) = \begin{cases} \lambda \exp(-\lambda x) & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

   Hence the likelihood function is

$$L(\lambda) = L(\lambda;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right).$$

   and the log-likelihood function is therefore

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

   The first derivative of the log-likelihood function (with respect to λ) is

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i,$$

   and setting this equal to zero we obtain

$$\lambda = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^{-1}.$$

   The second derivative of $\ell(\lambda)$ is

$$\ell''(\lambda) = -\frac{n}{\lambda^2} < 0 \quad \text{for all } \lambda > 0.$$

   Hence the turning point is a maximum, so the MLE is $\hat{\lambda}_{\text{MLE}} = \bar{X}^{-1}$.

2. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson(λ) distribution, where $\lambda > 0$ is unknown. Find the MLE of λ.

   Answer: Let $\mathbf{x} = (x_1, x_2, \ldots, x_n\}$ be a realization of the sample. The PMF of the Poisson(λ) distribution is

$$f(x;\lambda) = \begin{cases} \dfrac{\lambda^x \exp(-\lambda)}{x!} & \text{for } x = 0, 1, 2, 3, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

The likelihood function is

$$L(\lambda) = L(\lambda; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \lambda) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \exp(-n\lambda)}{\prod_{i=1}^{n} x_i!},$$

and the log-likelihood function is therefore

$$\ell(\lambda) = \left(\sum_{i=1}^{n} x_i\right) \log\lambda - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

The first derivative of $\ell(\lambda)$ with respect to λ is

$$\ell'(\lambda) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n.$$

Setting this equal to zero, we obtain

$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The second derivative is

$$\ell''(\lambda) = \frac{-\sum_{i=1}^{n} x_i}{\lambda^2} < 0 \quad \text{for all } \lambda > 0.$$

Hence the turning point is a maximum, so the MLE is $\hat{\lambda}_{\text{MLE}} = \bar{X}$.

3. Let X be a single observation from the Binomial(n, θ) distribution, where n is known but θ is unknown. Find the MLE of θ.

Answer: The likelihood function for the observation $X = k$ is

$$L(\theta) = L(\theta; k) = \binom{n}{\theta}\theta^k(1-\theta)^{n-k}$$

and the log-likelihood is

$$\ell(\theta) = \log\binom{n}{k} + k\log\theta + (n-k)\log(1-\theta).$$

Taking the derivative of $\ell(\theta)$ with respect to the θ,

$$\ell'(\theta) = \frac{k}{\theta} - \frac{(n-k)}{(1-\theta)}$$

Setting this to zero,

$$\frac{k(1-\theta) - (n-k)\theta}{\theta(1-\theta)} = 0 \quad \Rightarrow \quad \theta = \frac{k}{n}.$$

The second derivative of $\ell(\theta)$ is

$$\ell''(\theta) = \frac{-k}{\theta^2} - \frac{(n-k)}{(1-\theta)^2}.$$

Because $k \leq n$, this is always negative, so the turning point is a maximum. Hence the MLE of θ is

$$\hat{\theta}(X) = \frac{X}{n},$$

which is the observed proportion of successes.

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from the Geometric(θ) distribution. Find the MLE of θ.

   Answer: Let $X \sim$ Geometric(θ) distribution. The PMF of X is

$$f(x;\theta) = \theta(1-\theta)^{x-1} \quad x = 1,2,\ldots \quad \text{(and zero otherwise)}.$$

   Let $(x_1, x_2, \ldots, x_n)$ be a realisation of the sample. The likelihood function is

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i;\theta) = \prod_{i=1}^n \theta(1-\theta)^{x_i-1} = \theta^n(1-\theta)^{\sum_{i=1}^n (k_i-1)}.$$

   The log-likelihood function is

$$\ell(\theta) = n\log\theta + \log(1-\theta)\sum_{i=1}^n (k_i - 1).$$

   The first derivative of $\ell(\theta)$ is

$$\ell'(\theta) = \frac{n}{\theta} - \frac{\sum_{i=1}^n (k_i-1)}{1-\theta}.$$

   Setting this equal to zero, we obtain

$$\theta = \left(\frac{1}{n}\sum_{i=1}^n k_i\right)^{-1}.$$

   Hence the maximum likelihood estimator of θ is $\bar{X}^{-1}$. This makes sense: the longer we wait until the first success, the lower our estimate of the probability of success.

5. **Simple Linear Model**. Let X and Y be two random variables, and consider the simple linear model:

$$Y = \alpha + \beta X + \epsilon \quad \text{where} \quad \epsilon \sim N(0, \sigma^2)$$

   where α, β and $\sigma^2 > 0$ are unknown parameters and the **error variable** ϵ is independent of X. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample of observations from the joint distribution of X and Y.

   (a) Show that the maximum likelihood estimators of α and β are

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

   respectively.

   Answer: If we observe $X = x$ we have that $Y \sim N(\alpha + \beta x, \sigma^2)$, so

$$f(y;\alpha,\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{y - (\alpha + \beta x)}{\sigma}\right)^2\right]$$

   In particular,

$$\mathbb{E}(Y|X = x) = \alpha + \beta x \quad \text{and} \quad \text{Var}(Y|X = x) = \sigma^2.$$

   Let $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ be a realisation of the sample. The likelihood function and log-likelihood functions are

$$\begin{aligned} L(\alpha,\beta,\sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n \big[y_i - (\alpha + \beta x_i)\big]^2\right] . \ell(\alpha,\beta,\sigma^2) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n \big[y_i - (\alpha + \beta x_i)\big]^2. \end{aligned}$$

The MLE estimates of α and β are obtained by the **method of least squares**, which is to minimise the sum of squared errors

$$S(\alpha,\beta)=\sum_{i=1}^{n}\big[y_i-(\alpha+\beta x_i)\big]^2$$

The partial derivatives of $S(\alpha,\beta)$ with respect to α and β are

$$\begin{aligned}\frac{\partial S}{\partial\alpha}&=2\sum_{i=1}^{n}\big[y_i-(\alpha+\beta x_i)\big](-1)\\ \frac{\partial S}{\partial\beta}&=2\sum_{i=1}^{n}\big[y_i-(\alpha+\beta x_i)\big](-x_i)\end{aligned}$$

To find the MLEs of α and β, we set the partial derivatives to equal zero.

$$\begin{aligned}\frac{\partial H}{\partial\alpha}=0\ &\Rightarrow\ \sum_{i=1}^{n}y_i-n\alpha-\beta\sum_{i=1}^{n}x_i=0\\ &\Rightarrow\ n\alpha=\sum_{i=1}^{n}y_i-\beta\sum_{i=1}^{n}x_i\\ &\Rightarrow\ \alpha=\bar{y}-\beta\bar{x}.\end{aligned}$$

$$\begin{aligned}\frac{\partial H}{\partial\beta}=0\ &\Rightarrow\ \sum_{i=1}^{n}x_iy_i-\alpha\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}x_i^2=0\\ &\Rightarrow\ \sum_{i=1}^{n}x_iy_i-(\bar{y}-\beta\bar{x})\sum_{i=1}^{n}x_i-\beta\sum_{i=1}^{n}x_i^2=0\\ &\Rightarrow\ \sum_{i=1}^{n}x_i(y_i-\bar{y})-\beta\sum_{i=1}^{n}x_i(x_i-\bar{x})=0\\ &\Rightarrow\ \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})-\beta\sum_{i=1}^{n}(x_i-\bar{x})^2=0\\ &\Rightarrow\ \beta=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\end{aligned}$$

The maximum-likelihood estimators of α and β are therefore

$$\hat{\alpha}=\bar{Y}-\hat{\beta}\bar{X}\quad\text{and}\quad\hat{\beta}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}.$$

(b) Show that the MLE of the error variance σ^2 is

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2\quad\text{where}\quad\hat{\epsilon}_i=Y_i-(\hat{\alpha}+\hat{\beta}X_i).$$

Answer: Recall the log-likelihood function:

$$\ell(\alpha,\beta,\sigma^2)=\frac{n}{2}\log(2\pi\sigma^2)+\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big[y_i-(\alpha+\beta x_i)\big]^2.$$

The first partial derivative of $\ell(\alpha,\beta,\sigma^2)$ with respect to σ^2 is

$$\frac{\partial\ell}{\partial(\sigma^2)} = \frac{n}{2\sigma^2} - \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n \big[y_i - (\alpha+\beta x_i)\big]^2.$$

Setting this equal to zero,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n \big[y_i - (\alpha+\beta x_i)\big]^2.$$

Substituting our MLEs for α and β we obtain the MLE

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2 \quad \text{where} \quad \hat{\epsilon}_i = Y_i - (\hat{\alpha}+\hat{\beta}x_i)$$

as required.

8.3 Fisher information

Let X be a continuous random variable and let $\mathcal{M} = \{f(x;\theta) : \theta\in\Theta\}$ be a statistical model for its distribution. How can we quantify the amount of information that X carries about θ?

We assume that our statistical model $\mathcal{M}$ satisfies Condition 8.2 along with the following additional requirements:

Condition 8.9

4. The PDFs $f(x;\theta)$ are twice differentiable as functions of θ.
5. The integral $\displaystyle\int f(x;\theta)\,dx$ is twice differentiable under the integral sign as a function of θ.

We shalll not consider these in detail.

8.3.1 The score function

Let X be a single observation.

Definition 8.10
The first derivative of the log-likelihood function is called the **score function**,

$$u(\theta;x) = \frac{\partial}{\partial\theta}\log f(x;\theta)$$

- Given a fixed observation $X=x$, we think of $u(\theta;x)$ as a function of θ.
- Given a fixed value of θ, we think of $u(\theta;X)$ as a random variable. This is a transformation of X which we denote by
$$U = \frac{\partial}{\partial\theta}\log f(X;\theta)$$

Lemma 8.11
Under the regularity conditions stated above,

$$\mathbb{E}(U) = \mathbb{E}\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right) = 0.$$

Proof: We prove the lemma for the case where X is a continuous random variable. For every value of θ, the integral of $f(x;\theta)$ with respect to x satisfies

$$\int_{-\infty}^{\infty} f(x;\theta)\,dx = 1.$$

Taking the derivative with respect to θ (and applying the regularity conditions), we see that

$$\frac{\partial}{\partial\theta}\left(\int_{-\infty}^{\infty} f(x;\theta)\,dx\right) = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial\theta}\,dx = 0.$$

By the chain rule, the first partial derivative of $\log f(x;\theta)$ with respect to θ is given by

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta).$$

Hence

$$\mathbb{E}\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right) = \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)f(x;\theta)\,dx = \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x;\theta)\,dx = 0$$

as required.

Example 8.12
Find the score function of a single observation from the Geometric(θ) distribution whose PMF is

$\mathbb{P}(X = k) = \theta(1-\theta)^{k-1}$ for $k = 1, 2, 3, \ldots$ where $0 < \theta < 1$ is unknown.

Solution:

8.3.2 Fisher information

Definition 8.13
The variance of the score function is called the **Fisher information**,

$$I(\theta) = \mathrm{Var}(U) = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right).$$

The following lemma gives a conveient way of computing $I(\theta)$ in terms of the derivative of the score function.

Lemma 8.14
Under the regularity conditions stated above, the Fisher information satisfies

$$I(\theta) = -\mathbb{E}(U') = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right).$$

Proof: By Lemma 8.11, the expected value of the score function is zero:

$$\mathbb{E}(U) = \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right) f(x;\theta)\,dx = 0.$$

Taking the derivative of both sides with respect to θ,

$$\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right) f(x;\theta)\,dx + \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)\frac{\partial f(x;\theta)}{\partial\theta}\,dx = 0.$$

By the chain rule,

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta).$$

Thus we have

$$\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right) f(x;\theta)\,dx + \int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx = 0.$$

This can be written as

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right) + \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2\right] = 0.$$

Because the expected value of the score function is zero, the second term is the variance of the score function. Hence

$$\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right) + I(\theta) = 0$$

and the result follows.

Example 8.15
Find the Fisher information of an observation from the Bernoulli(θ) distribution.

Solution:

Example 8.16
Find the Fisher information of an observation from the $N(\theta, \sigma^2)$ distribution, whose mean θ is unknown but whose variance σ^2 is known.

Solution:

8.3.3 Random samples

The score function and Fisher information of a single observation extend naturally to random samples $X_1, X_2, \ldots, X_n$ from the distribution of X.

Definition 8.17
The **score function** of a random sample $X_1, X_2, \ldots, X_n$ is the first partial derivative of its log-likelihood function with respect to θ. By independence,

$$u(\theta; \mathbf{X}) = \frac{\partial}{\partial\theta} \log L(\theta; \mathbf{X}) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log L(\theta; X_i) = \sum_{i=1}^{n} u(\theta; X_i)$$

Definition 8.18
The **Fisher information** of a random sample $X_1, X_2, \ldots, X_n$ is the variance of its score function:

$$I_n(\theta) = \operatorname{Var}\left(\frac{\partial}{\partial\theta} \log L(\theta; \mathbf{X})\right) = \sum_{i=1}^{n} \operatorname{Var}\left(\frac{\partial}{\partial\theta} \log L(\theta; X_i)\right) = nI(\theta),$$

where $I(\theta)$ is the Fisher information of a single observation.

Remark 8.19
$I_n(\theta)$ increases with the sample size: the more data we have, the more information we have about θ.

Exercise 8.20

1. Find the score function of an observation from the Poisson(θ) distribution, and verify that its expected value is zero.

   Answer: The PMF of the Poisson(θ) distribution is

   $$f(x; \theta) = \frac{\theta^x \exp(-\theta)}{x!} \text{for } x = 0, 1, 2, 3, \ldots \text{ (zero otherwise)}$$

   Let $X \sim$ Poisson(θ). The score function for a single observation $X = x$ is

   $$u(\theta; x) = \frac{\partial}{\partial\theta} \log f(x, \theta) = \frac{\partial}{\partial\theta}(x \log\theta - \theta - \log x!) = \frac{x - \theta}{\theta}.$$

   Since $\mathbb{E}(X) = \theta$, by the linearity of expectation we have

   $$\mathbb{E}\big[u(\theta; X)\big] = \mathbb{E}\left(\frac{X - \theta}{\theta}\right) = \frac{\mathbb{E}(X) - \theta}{\theta} = \frac{\theta - \theta}{\theta} = 0 \quad \text{as required.}$$

2. Find the Fisher information $I(\theta)$ of an observation from the Geometric(θ) distribution, and find the value of θ for which $I(\theta)$ is minimum.

   Answer: From Example 8.12, the score function of X is

$$u(\theta; X) = \frac{1-\theta X}{\theta(1-\theta)}$$

   The derivative of $u(\theta; X)$ with respect to θ is

$$\frac{\partial}{\partial\theta}u(\theta, X) = \frac{\theta(1-\theta)(-X) - (1-\theta X)(1-2\theta)}{\theta^2(1-\theta)^2} = -\frac{1-2\theta+\theta^2 X}{\theta^2(1-\theta)^2}$$

   Because $\mathbb{E}(X) = 1/\theta$, the Fisher information is therefore

$$\begin{aligned} I(\theta) = -\mathbb{E}\left(\frac{\partial}{\partial\theta}u(\theta, X)\right) &= \frac{1-2\theta+\theta^2\mathbb{E}(X)}{\theta^2(1-\theta)^2} \\ &= \frac{1-\theta}{\theta^2(1-\theta)^2} \\ &= \frac{1}{\theta^2(1-\theta)} \end{aligned}$$

   To find the minimum, we differentiate $I(\theta)$ and set the resulting expression to zero:

$$I'(\theta) = \frac{3\theta-2}{\theta^3(1-\theta)^2} = 0.$$

   This shows that $I(\theta)$ has a turning point at $\theta = 2/3$. The second derivative of $I(\theta)$ is

$$I''(\theta) = \frac{9\theta^2 - 16\theta + 9}{\theta^4(1-\theta)^3}.$$

   The discriminant of the quadratic expression on the numerator is negative, which shows that $I''(\theta) > 0$ for all θ, and hence $I(\theta)$ has a minimum at $\theta = 2/3$.

   Note that for the Bernoulli(θ) distribution, $I(\theta)$ reaches its minimum at $\theta = 1/2$:

$$I(\theta) = \frac{1}{\theta(1-\theta)}, \qquad I'(\theta) = \frac{2\theta-1}{\theta^2(1-\theta)^2}, \qquad I''(\theta) = \frac{2\big[\theta(1-\theta) + (2\theta-1)^2\big]}{\theta^3(1-\theta)^3}.$$

8.4 The Cramér-Rao lower bound

We now show that the variance of any unbiased estimator cannot be smaller than a certain fixed value which depends only on the Fisher information of the associated distribution. First we need the following lemma.

Lemma 8.21
Let $T(X)$ be an unbiased estimator of θ and let $U(X)$ be the score function of X. Then

$$\text{Cov}(T, U) = 1$$

Proof: We prove the lemma for continuous random variables.

By Lemma 8.11, $\mathbb{E}(U) = 0$ so

$$\begin{aligned}
\mathrm{Cov}(T, U) &= \mathbb{E}(TU) - \mathbb{E}(T)\mathbb{E}(U) = \mathbb{E}(TU) \\
&= \mathbb{E}\left[T(X)\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)\right] \\
&= \mathbb{E}\left[T(X)\left(\frac{1}{f(X;\theta)}\frac{\partial}{\partial\theta}f(X;\theta)\right)\right] \\
&= \int T(x)\left(\frac{1}{f(x;\theta)}\frac{\partial}{\partial\theta}f(x;\theta)\right)f(x;\theta)\,dx \\
&= \int T(x)\left(\frac{\partial}{\partial\theta}f(x;\theta)\right)\,dx \\
&= \frac{\partial}{\partial\theta}\int T(x)f(x;\theta)\,dx \quad \text{(by the regularity conditions)} \\
&= \frac{\partial}{\partial\theta}\mathbb{E}(T) = \frac{\partial}{\partial\theta}\theta = 1 \qquad \text{(because T is unbiased).}
\end{aligned}$$

Theorem 8.22 (CRLB)
Let X be a random variable and let $f(x;\theta)$ denote its PDF where θ is an unknown scalar. If $T(X)$ is an unbiased estimator of θ then

$$\mathrm{Var}(T) \geq \frac{1}{I(\theta)}.$$

Proof: By the Cauchy-Schwarz inequality applied to $T - \mathbb{E}(T)$ and $U - \mathbb{E}(U)$,

$$\begin{aligned}
1 = \mathrm{Cov}(T, U)^2 &= \mathbb{E}\big([T - \mathbb{E}(T)][U - \mathbb{E}(U)]\big)^2 \\
&\leq \mathbb{E}\big([T - \mathbb{E}(T)]^2\big)\mathbb{E}\big([U - \mathbb{E}(U)]^2\big) \\
&= \mathrm{Var}(T)\mathrm{Var}(U)
\end{aligned}$$

Hence

$$\mathrm{Var}(T) \geq \frac{1}{\mathrm{Var}(U)}.$$

Thus, because $I(\theta) = \mathrm{Var}(U)$, i.e. the variance of the score function, we conclude that

$$\mathrm{Var}(T) \geq \frac{1}{I(\theta)}.$$

Corollary 8.23
Let $X_1, X_2, \ldots, X_n$ be a random sample and let $f(x, \theta)$ denote their common PDF, where θ is an unknown scalar. If T_n is an unbiased estimator of θ then

$$\mathrm{Var}(T_n) \geq \frac{1}{nI(\theta)}$$

where $I(\theta)$ is the Fisher information of a single observation.

Remark 8.24

- The CRLB provides a **lower limit** on the variance of an unbiased estimator.
- Similar results hold for biased estimators, vectors of parameters, non-independent samples, and so on.
- The CRLB has deep connections with the **Heisenberg Uncertainty Principle**.

8.5 Efficiency

Definition 8.25
Let T be an unbiased estimator of the parameter θ.

1. The **efficiency** of T as an estimator of θ is defined by $e(T) = \dfrac{1/I(\theta)}{\mathrm{Var}(T)}$.
2. If $e(T) = 1$ then T is called an **efficient** estimator of θ.

An efficient estimator is thus an unbiased estimator whose variance achieves the Cramér-Rao lower bound (for all θ) and which therefore has the smallest variance among all possible unbiased estimators.

Example 8.26
Let $X_1, X_2, \dots, X_n$ be a random sample from the Bernoulli(θ) distribution, where $0 < \theta < 1$ is unknown. Show that the sample mean $\bar{X}$ is an efficient estimator of θ.

Solution:

Example 8.27
Let $X_1, X_2, \dots, X_n$ be a random sample from the $N(\theta, \sigma^2)$ distribution, whose mean θ is unknown but whose variance σ^2 is known. Show that $\bar{X}$ is an efficient estimator of θ.

Solution:

Example 8.28
Let X be a continuous random variable with the following PDF:

$$f(x) = \begin{cases} \dfrac{3\theta^3}{(x+\theta)^4} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

1. Show that $\mathbb{E}(X) = \theta/2$ and $\mathrm{Var}(X) = 3\theta^2/4$.
2. Show that $T = 2\bar{X}$ is an unbiased esimator of θ, and find its variance.
3. Show that the efficiency of T as an estimator of θ is $5/9$.

Solution:

Exercise 8.29

1. Let $X_1, X_2, \dots, X_n$ be a random sample from the Poisson(θ) distribution, where $\theta > 0$ is unknown. Show that the sample mean $\bar{X}$ is an efficient estimator of θ.

   Answer: The PMF of the Poisson(θ) distribution is

   $$f(x;\theta) = \frac{\theta^x \exp(-\theta)}{x!} \quad \text{for } x = 0, 1, 2, 3, \dots \text{ (zero otherwise)}$$

   Let $X \sim \text{Poisson}(\theta)$. The score function of X is

   $$u(\theta; X) = \frac{\partial}{\partial\theta}\log f(X,\theta) = \frac{\partial}{\partial\theta}(X\log\theta - \theta - \log x!) = \frac{X-\theta}{\theta}.$$

   Since $\mathbb{E}(X) = \theta$ and $\text{Var}(X) = \theta$, the Fisher information of a single observation is therefore

   $$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x,\theta)\right)^2\right] = \frac{\mathbb{E}\big[(X-\theta)^2\big]}{\theta^2} = \frac{\text{Var}(X)}{\theta^2} = \frac{1}{\theta}.$$

   Hence for a random sample of size n, the CRLB is equal to

   $$\frac{1}{nI(\theta)} = \frac{\theta}{n}.$$

   This is equal to the variance of the sample mean $\bar{X}$, so $\bar{X}$ is an efficient estimator of θ.

2. Let $X_1, X_2, \dots, X_n$ be a random sample from the $N(\mu, \theta)$ distribution, whose mean μ is known but whose variance θ is unknown. Using the fact that $\mathbb{E}\big((X-\mu)^4\big) = 3\theta^2$ for the normal distribution, show that the statistic

   $$T = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$$

   is an efficient estimator of the variance θ.

   Answer: Let $X \sim N(\mu, \theta)$. First we note that T is unbiased, because

   $$\mathbb{E}(T) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[(X_i-\mu)^2\big] = \text{Var}(X) = \theta.$$

   Furthermore, because the X_i are independent, the variance of T is

   $$\begin{aligned}\text{Var}(T) = \frac{1}{n}\text{Var}\big[(X-\mu)^2\big] &= \frac{1}{n}\Big(\mathbb{E}\big[(X-\mu)^4\big] - \mathbb{E}\big[(X-\mu)^2\big]^2\Big) \\ &= \frac{1}{n}\big(3\theta^2 - \theta^2\big) \\ &= \frac{2\theta^2}{n}\end{aligned}$$

Let $f(x;\theta)$ denote the PDF of the $N(\mu,\theta)$ distribution:

$$\begin{aligned}
f(x;\theta) &= \frac{1}{\sqrt{2\pi\theta}}\exp\left(-\frac{(x-\mu)^2}{2\theta}\right),\\
\log f(x;\theta) &= \log\left(\frac{1}{\sqrt{2\pi}}\right)+\log\left(\frac{1}{\sqrt{\theta}}\right)-\frac{(x-\mu)^2}{2\theta},\\
\frac{\partial}{\partial\theta}\log f(x;\theta) &= -\frac{1}{2\theta}+\frac{(x-\mu)^2}{2\theta^2},\\
\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) &= \frac{1}{2\theta^2}-\frac{(x-\mu)^2}{\theta^3}.
\end{aligned}$$

The Fisher information of a single observation is thus given by

$$\begin{aligned}
I(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right) &= -\mathbb{E}\left(\frac{1}{2\theta^2}-\frac{(X-\mu)^2}{\theta^3}\right)\\
&= -\frac{1}{2\theta^2}+\frac{\mathbb{E}(X-\mu)^2}{\theta^3}\\
&= -\frac{1}{2\theta^2}+\frac{\theta}{\theta^3}=\frac{1}{2\theta^2}.
\end{aligned}$$

By the Cramer-Rao theorem,

$$\mathrm{Var}(T)\geq\frac{1}{nI(\theta)}=\frac{2\theta^2}{n}$$

- The variance of T attains the CRLB for all values of $\theta>0$.
- T is therefore an efficient estimator of the variance of the Normal distribution.

3. Let $X\sim\text{Rayleigh}(\theta)$ whose CDF is given by

$$F(x)=1-e^{-x^2/2\theta}\quad\text{for } x\geq 0 \text{ and zero otherwise.}$$

   1. Show that $X^2\sim\text{Exponential}(2\theta)$ where 2θ is a scale paremter.
   2. Show that the MLE of θ is $T=\dfrac{1}{2n}\displaystyle\sum_{i=1}^n X_i^2$.
   3. Show that the T is an unbiased estimator of θ.
   4. Show that the T is an efficient estimator of θ.

Answer:

1. Let $Y=g(X)$ where $g(x)=x^2$. This is one-to-one and increasing over $\text{supp}(f_X)=[0,\infty)$. The inverse transformation is $g^{-1}(y)=\sqrt{y}$ and $\text{supp}(f_Y)=[0\infty)$. Hence
$$F_Y(y)=F_X\left[g^{-1}(y)\right]=1-e^{-y/2\theta}\quad\text{for } y\geq 0\text{, and zero otherwise.}$$
This is the CDF of the Exponential(2θ) distribution, where 2θ is a scale parameter.

2.
- $f(x,\theta)=\dfrac{x}{\theta}e^{-x^2/2\theta}$.
- $L(\theta)=\prod_{i=1}^n\dfrac{x_i}{\theta}e^{-x_i^2/2\theta}$.
- $\ell(\theta)=-n\log\theta+\sum_{i=1}^n\log x_i-\dfrac{1}{2\theta}\displaystyle\sum_{i=1}^n x_i^2$.
- $\ell'(\theta)=-\dfrac{n}{\theta}+\dfrac{1}{2\theta^2}\displaystyle\sum_{i=1}^n x_i^2$.

- $\ell''(\theta) = \dfrac{n}{\theta^2} - \dfrac{1}{\theta^3}\displaystyle\sum_{i=1}^n x_i^2.$

   Setting $\ell'(\theta) = 0$, the MLE of θ is $T = \dfrac{1}{2n}\displaystyle\sum_{i=1}^n X_i^2$.

3. Because $X^2 \sim \text{Exponential}(2\theta)$ we have $\mathbb{E}(X^2) = 2\theta$, so

$$\mathbb{E}(T) = \frac{1}{2n}\sum_{i=1}^n \mathbb{E}(X_i^2) = \theta.$$

   so T is unbiased.

4. Again using the fact that $\mathbb{E}(X^2) = 2\theta$,

$$I_n(\theta) = -\mathbb{E}\big[\ell''(\theta;\mathbf{X})\big] = -\frac{n}{\theta^2} + \frac{1}{\theta^3}\sum_{i=1}^n \mathbb{E}(X_i^2) = -\frac{n}{\theta^2} + \frac{n}{\theta^3}2\theta = \frac{n}{\theta^2}.$$

   Because $X^2 \sim \text{Exponential}(2\theta)$ we have $\mathbb{E}(X^2) = 2\theta$ and $\text{Var}(X^2) = 4\theta^2$. By indepedence,

$$\text{Var}(T) = \frac{1}{4n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{n}{4n^2}4\theta^2 = \frac{\theta^2}{n}$$

   Thus $\text{Var}(T)$ achieves the CRLB, so T is an efficient estimator for θ.

4. Let $X \sim \text{Exponential}(\lambda)$ where $\lambda > 0$ is an unknown rate parameter and let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X. The PDF of X is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let $S_n = \sum_{i=1}^n X_i$. This is the sum of n independent Exponential(λ) random variables, and has the so-called **Erlang** distribution with parameters $n \in \mathbb{N}$ and $\lambda > 0$, whose PDF is given by

$$f_S(s) = \frac{\lambda^n s^{n-1} e^{-\lambda s}}{\Gamma(k)} \quad \text{for } s > 0 \text{ and zero otherwise,}$$

where $\Gamma(k)$ is the so-called **gamma function**,

$$\Gamma(k) = \int_0^\infty t^{k-1}e^{-t}\,dt.$$

$\Gamma(k)$ is an extension of the factorial function and has the property $\Gamma(k+1) = k\Gamma(k)$.

1. Show that $\mathbb{E}(S_n^{-1}) = \dfrac{\lambda}{n-1}$ and $\text{Var}(S_n^{-1}) = \dfrac{\lambda^2}{(n-1)^2(n-2)}$.
2. Show that the MLE of λ is given by $T_n = \dfrac{n}{\sum_{i=1}^n X_i}$.
3. Show that T_n is an asymptotically unbiased estimator for λ as $n \to \infty$.
4. Show that T_n is an asymptotically efficient estimator of λ, in the sense that its variance converges to the CRLB as $n \to \infty$.

Answer:

1.
$$\mathbb{E}(S_n^{-1}) = \frac{\lambda^n}{\Gamma(n)}\int_0^\infty s^{n-2}e^{-\lambda s}\,ds = \frac{\lambda}{\Gamma(n)}\int_0^\infty u^{n-2}e^{-u}\,du = \frac{\lambda\Gamma(n-1)}{\Gamma(n)} = \frac{\lambda}{n-1}.$$
$$\mathbb{E}(S_n^{-2}) = \frac{\lambda^n}{\Gamma(n)}\int_0^\infty s^{n-3}e^{-\lambda s}\,ds = \frac{\lambda^2}{\Gamma(n)}\int_0^\infty u^{n-3}e^{-u}\,du = \frac{\lambda^2\Gamma(n-2)}{\Gamma(n)} = \frac{\lambda^2}{(n-1)(n-2)}.$$
$$\text{Var}(S_n^{-1}) = \mathbb{E}(S_n^{-2}) - \mathbb{E}(S_n^{-1})^2 = \frac{\lambda^2}{(n-1)^2(n-2)}.$$

2. - $L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$.
   - $\ell(\lambda) = n \log \lambda + (\theta - 1) \sum_{i=1}^{n} x_i$.
   - $\ell'(\lambda) = n/\theta + \sum_{i=1}^{n} x_i$.
   - $\ell''(\lambda) = -n/\theta^2 < 0$.

   Setting $\ell'(\lambda) = 0$ we obtain $T = n/\sum_{i=1}^{n} X_i$ as the MLE of λ.

3. T_n is an asymptotically unbiased estimator for λ because

$$\mathbb{E}(T_n) = n\mathbb{E}\left(\frac{1}{S_n}\right) = \left(\frac{n}{n-1}\right)\lambda. = \left(\frac{1}{1-1/n}\right)\lambda \to \lambda \text{ as } n \to \infty.$$

4. The Fisher information of the sample is

$$I_n(\lambda) = -\mathbb{E}\big[\ell''(\lambda)\big] = \frac{n}{\lambda^2}$$

so the CRLB is λ^2/n. Now,

$$\text{Var}(T_n) = n^2\text{Var}\left(\frac{1}{S_n}\right) = \frac{\lambda^2}{n}\left(\frac{n^3}{(n-1)^2(n-2)}\right) = \frac{\lambda^2}{n}\left(\frac{1}{(1-1/n)^2(1-2/n)}\right) \to \frac{\lambda^2}{n} \text{ as } n \to \infty.$$

Thus $\text{Var}(T_n)$ converges to the CRLB as $n \to \infty$.

Chapter 9 Hypothesis Testing

9.1 Null hypothesis siginficance testing

Let X be a random variable, let $\mathcal{M} = \{F(x;\theta) : \theta \in \Theta\}$ be a statistical model for its distribution and let $\{\Theta_0, \Theta_1\}$ be a **partition** of the parameter space:

$$\Theta_0 \cup \Theta_1 = \Theta \qquad \text{and} \qquad \Theta_0 \cap \Theta_1 = \emptyset.$$

This partition defines two possible statistical models for the distribution of X.

The **null** model:	$\mathcal{M}_0 = \{F(x,\theta) : \theta \in \Theta_0\}$.
The **alternative** model:	$\mathcal{M}_1 = \{F(x,\theta) : \theta \in \Theta_1\}$.

To decide which model is the 'correct' one, we need to test the claim that the true parameter belongs to the set Θ_0 against the alternative claim that it belongs to Θ_1. We denote these two hypotheses by H_0 and H_1 respectively and refer to them as follows.

The **null** hypothesis:	$H_0 : \theta \in \Theta_0$.
The **alternative** hypothesis:	$H_1 : \theta \in \Theta_1$.

Definition 9.1

- A hypothesis which specifies a particular value of θ is called a **simple hypothesis**.
- A hypothesis which specifies a set of values for θ is called a **composite hypothesis**.

For example, $H_0 : \theta = \theta_0$ is a simple hypothesis while $H_1 : \theta \neq \theta_0$ is a composite hypothesis.

9.1.1 Type I and Type II errors

In the absence of any evidence to the contrary, we assume that H_0 is correct. Suppose we now obtain a random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ from the distribution of X, and compute an estimate $T(\mathbf{x})$ of the true parameter value θ. We decide which hypothesis is correct based on the computed value of $T(\mathbf{x})$:

- if $T(\mathbf{x}) \in \Theta_0$ we retain the null hypothesis $H_0 : \theta \in \Theta_0$;
- if $T(\mathbf{x}) \in \Theta_1$ we reject H_0 in favour of the alternative hypothesis $H_1 : \theta \in \Theta_1$.

Definition 9.2

- A **Type I error** occurs when $\theta \in \Theta_0$ but $T(\mathbf{x}) \in \Theta_1$, which leads us to incorrectly reject H_0.
- A **Type II error** occurs when $\theta \in \Theta_1$ but $T(\mathbf{x}) \in \Theta_0$, which leads us to incorrectly retain H_0.

Hypothesis tests can be respresented by **decision tables**:

	Reality	
Decision	H_0 true ($\theta \in \Theta_0$)	H_0 false ($\theta \in \Theta_1$)
Retain H_0 ($T \in \Theta_0$)	Correct decision	Type II error
Reject H_0 ($T \in \Theta_1$)	Type I error	Correct decision

Different applications use different terminology. A decision table for a radar system, where the null hypothesis asserts the absence of a target, might be as follows:

	Reality		
Decision	Target Absent	Target Present	Action
Target Absent	Clear	Miss	Stay silent
Target Present	False alarm	Hit	Sound alarm

Here is a decision table for a medical diagnosis, where the null hypothesis asserts the absence of a disease:

	Reality		
Decision	Disease Absent	Disease Present	Action
Disease Absent	True negative	False negative	Do nothing
Disease Present	False positive	True positive	Prescribe

9.1.2 Critical regions

Definition 9.3
Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from the distribution of X. The set of all possible realisations of $\mathbf{X}$ is called the **sample space** which we denote by $D \subseteq \mathbb{R}^n$.

Model ($\mathcal{M}$)	Sample space (D)
Bernoulli	binary vectors (of length n)
Poisson	vectors of non-negative integers
Normal	vectors of real numbers

Definition 9.4
A **hypothesis test** of a null hypothesis H_0 against an alternative hypothesis H_1 is defined by a subset of the sample space called the **critical region** of the test: we **reject** H_0 if $\mathbf{X} \in C$ but **retain** H_0 if $\mathbf{X} \notin C$.

Critical regions can be specified in terms of a **test statistic** say $T : D \to \mathbb{R}$, in which case the critical region is specified by one or more **critical values**. For example, we might define

$$C = \{\mathbf{x} \in D : c_1 \leq T(\mathbf{x}) \leq c_2\}$$

in which case we reject H_0 if $T(\mathbf{x})$ falls between the critial values c_1 and c_2.

9.1.3 The size (or significance level) of a test

We would like to choose a critical region C that minimises the probability of making both Type I and Type II errors. These are conflicting objectives, as illustrated by the following extreme cases.

- If we choose $C = \emptyset$ we will never reject H_0 (because the random sample never falls into C), so we never make Type I errors when $\theta \in \Theta_0$ but always make Type II errors when $\theta \in \Theta_1$.
- If we choose $C = D$ we will always reject H_0 (because the random sample always falls into C), so we never make Type II errors when $\theta \in \Theta_1$ but always make Type I errors when $\theta \in \Theta_0$.

Remark 9.5 (Conservatie testing)
The null hypothesis represents the *status quo*. From a conservative standpoint, rejecting the status quo incorrectly (Type I error) is worse than retaining the status quo incorrectly (Type II error). As a result, hypothesis tests usually proceed in two stages:

1. find a set of tests for which $\mathbb{P}$(Type I error) is bounded above by some acceptable value, then
2. choose one of these tests so that $\mathbb{P}$(Type II error) is as small as possible.

Definition 9.6
Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample and let $F_\theta(x)$ be the common CDF of the component variables X_i. It can be shown that for every $\theta \in \Theta$ there exits a unique probability measure $\mathbb{P}_\theta$ on subsets of $\mathbb{R}^n$ for which the component variables X_i are independent and identically distribted according to $F_\theta(x)$. This is called the **probability distribution induced by X** on the sample space and is denoted by $\mathbb{P}_\theta(A) = \mathbb{P}_\theta(\mathbf{X} \in A)$, which is the probability that the random sample falls into the set $A \subseteq \mathbb{R}^n$ when the parameter value is θ.

Definition 9.7
The **size** of a critical region (also called the **significance level** of the test) is the maximum probability of making a Type I error. This is usually denoted by α,

$$\alpha = \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\mathbf{X} \in C).$$

If H_0 is a simple hypothesis, say $H_0 : \theta = \theta_0$, this reduces to $\alpha = \mathbb{P}_{\theta_0}(\mathbf{X} \in C)$.

Remark 9.8 (p-values)
Let $T : D \to \mathbb{R}$ be a test statistic, let $C = \{\mathbf{x} : T(\mathbf{x}) \leq c\}$ be a critical region and suppose that $\mathbf{x}_{\text{obs}}$ is the observed sample realisation. Then the **empirical size** (or p-value) of the test is defined to be

$$p = \max_{\theta \in \Theta_0} \mathbb{P}_\theta\big[T(\mathbf{X}) \leq T(\mathbf{x}_{\text{obs}})\big].$$

If H_0 is a simple hypothesis, say $H_0 : \theta = \theta_0$, this reduces to

$$p = \mathbb{P}_{\theta_0}\big[T(\mathbf{X}) \leq T(\mathbf{x}_{\text{obs}})\big].$$

For a test of size α, we reject H_0 if the empirical size satisfies $p \leq \alpha$.

9.1.4 The power of a test

Among all critical regions of size α we would like to choose one that minimises the probability $\mathbb{P}_\theta(\mathbf{X} \notin C)$ of making Type II errors for every $\theta \in \Theta_1$, or equivalently a critical region that maximises $\mathbb{P}_\theta(\mathbf{X} \in C)$ for every $\theta \in \Theta_1$.

Definition 9.9
Let C be a critical region for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. The **power function** of the associated test is

$$\gamma(\theta) = \mathbb{P}_\theta(\mathbf{X} \in C) \quad \text{which is defined for all } \theta \in \Theta_1.$$

The value $\gamma(\theta)$ is called the **power of the test to detect the alternative hypothesis at** $\theta \in \Theta_1$.

Remark 9.10
The probability of making Type II errors is often denoted by β:

$$\begin{aligned}\beta(\theta) &= \mathbb{P}_\theta(\mathbf{X} \notin C)\\ &= 1-\gamma(\theta).\end{aligned}$$

To maximise the power $\gamma(\theta)$ is to minimise the probability $\beta(\theta)$ of making a Type II error.

Example 9.11
Let $X_1, X_2, \ldots, X_8$ be a random sample from the Poisson(θ) distribution, where $\theta > 0$ is unknown. We reject the simple null hypothesis $H_0 : \theta = 0.5$ in favour of the alternative $H_1 : \theta > 0.5$ whenever the observed sum satisfies $\sum_{i=1}^{8} X_i \geq 8$.

1. Compute the size of the test.
2. Compute the power of the test at $\theta = 0.75$, $\theta = 1.0$ and $\theta = 1.25$.

Use the fact that if $X \sim \text{Poisson}(\theta_1)$ and $Y \sim \text{Poisson}(\theta_2)$ then $X + Y \sim \text{Poisson}(\theta_1 + \theta_2)$.

Solution:

9.2 One-sample tests

Example 9.12 (Binomial test)
Let $X_1, X_2, \dots, X_n$ be a random sample from the Bernoulli(θ) distribution, where θ is unknown.

1. Find a critical region of size α to test $H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$.
2. Find the power function of the test.

Solution:

Example 9.13 (z-test)
Let $X \sim N(\mu, \sigma^2)$ where μ is unknown but σ^2 is known.

1. Find a critical region for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$
2. Find the power function $\gamma(\mu)$ of the test.
3. Show that the power function is strictly increasing for $\mu > \mu_0$.

Solution:

Example 9.14 (*t*-test)
Let $X \sim N(\mu, \sigma^2)$ where μ and σ^2 are both unknown. Find a critical region of size α for testing

$H_0 : \mu = \mu_0$ against $H_1 : \mu < \mu_0$

Solution:

Example 9.15 (Normal approximation)
Let $X_1, X_2, \ldots, X_n$ be a random sample from the Bernoulli(θ) distribution, where θ is unknown. We reject the null hypothesis $H_0 : \theta = 1/2$ in favour of $H_1 : \theta > 1/2$ if the observed number of successes exceeds some constant value $c > 0$. Using a normal approximation, find values of n and c for which the size of the test is 0.1 and whose power at $\theta = 2/3$ is 0.95.

Solution:

Exercise 9.16

1. Let $X \sim \text{Binomial}(10, \theta)$ where θ is either equal to 0.25 or 0.5. The simple mull hypothesis $H_0 : \theta = 0.5$ is rejected in favour of the simple alternative $H_1 : \theta = 0.25$ if the observed value of X is at most equal to 3. Find the size and power of the test.

   Answer:

   - The critical region is $C = \{X \leq 3\}$. Under the null hypothesis, $X \sim \text{Binomial}(10, 0.5)$. The significance level of the test is therefore

   $$\alpha = \mathbb{P}_{0.5}(X \leq 3) = \mathbb{P}\big[X \leq 3 \text{ when } X \sim \text{Binomial}(10, 0.5)\big] = 0.1719 \quad \text{(from tables)}.$$

   - Under the alternative hypothesis $H_1 : \theta = 0.25$ we have $X \sim \text{Binomial}(10, 0.25)$. The power of the test to detect H_1 is therefore

   $$\gamma(0.25) = \mathbb{P}_{0.25}(X \leq 3) = \mathbb{P}\big[X \leq 3 \text{ when } X \sim \text{Binomial}(10, 0.25)\big] = 0.7759 \quad \text{(from tables)}.$$

2. Adult males diagnosed with lung cancer have a mortality rate of 70% within one year of the initial diagnosis. A research laboratory claims that a new treatment reduces this rate. Based on a random sample of 20 patients, find a critical region of size $\alpha = 0.15$ to test the claim, and compute the power of the test to detect a 20% reduction in the mortality rate.

   Answer: Let θ be the probability that a patient dies within one year of the initial diagnosis.

   - We wish to test the null hypothesis $H_0 : \theta = 0.7$ against the alternative $H_1 : \theta < 0.7$.

   Let $X_1, X_2, \ldots, X_{20}$ be a random sample from the Bernoulli(θ) distribution

   - In this context, 'success' corresponds to death within one year of the initial diagnosis!

Let $S = \sum_{i=1}^{20} X_i$ be the total number of deaths within one year of the initial diagnosis.

- Under the null hypothesis, $S \sim \text{Binomial}(20, 0.7)$.

A critical region for the test is $C = \{\mathbf{x} : S(\mathbf{x}) \leq k\}$, where k is chosen so that

$$\mathbb{P}_{0.7}(S \leq k) = 0.15.$$

Tabulated values of the Binomial$(20, 0.7)$ distribution yield

$$\mathbb{P}_{0.7}(S \leq 11) = 0.1133 \quad \text{and} \quad \mathbb{P}_{0.7}(S \leq 12) = 0.2277.$$

- It is not possible to find a critical region of size $\alpha = 0.15$ exactly.

The conservative approach would be to take $k = 11$ and $\alpha = 0.1133$.

- A 20% reduction in the mortality rate corresponds to $H_1 : \theta = 0.5$.

For the test of size $\alpha = 0.1133$, its power to detect $H_1 : \theta = 0.5$ is

$$\begin{aligned}\mathbb{P}_{0.5}(S \leq 11) &= \mathbb{P}(S \leq 11) \text{ where } S \sim \text{Binomial}(20, 0.5) \\ &= 0.7483 \quad \text{(from tables)}.\end{aligned}$$

For the test of size $\alpha = 0.2277$, its power to detect $H_1 : \theta = 0.5$ is

$$\begin{aligned}\mathbb{P}_{0.5}(S \leq 12) &= \mathbb{P}(S \leq 12) \text{ where } S \sim \text{Binomial}(20, 0.5) \\ &= 0.8684 \quad \text{(from tables)}.\end{aligned}$$

3. Let $X_1, X_2, \ldots, X_5$ be a random sample from the Bernoulli(θ) distribution. We wish to test the null hypothesis $H_0 : \theta \leq 0.5$ against the alternative $H_1 : \theta > 0.5$. H_0 is rejected by test A only if all five trials result in 'success', and rejected by test B if at least at least three trials result in 'success'. Find the size and power function of each test.

   Answer: The sample space is $D = \{0, 1\}^5$, the set of binary vectors of length 5. Let $Y = \sum_{i=1}^{5} X_i$ be the number of successes: $Y \sim \text{Binomial}(5, \theta)$.

   1. For test A,

   $$\alpha = \max_{0 \leq \theta \leq 0.5} \mathbb{P}_\theta(Y = 5) = 0.5^5 = 0.0312.$$

   $$\gamma(\theta) = \mathbb{P}_\theta(Y = 5) = \theta^5.$$

   2. For test B,

   $$\alpha = \max_{0.5 \leq \theta \leq 1} \mathbb{P}(Y \geq 5) = 10(0.5)^3(0.5)^2 + 5(0.5)^4(0.5) + (0.5)^5 = 16(0.5)^5 = 0.5,$$

   $$\gamma(\theta) = \mathbb{P}(Y \geq 3; \theta) = 10\theta^3(1-\theta)^2 + 5\theta^4(1-\theta) + \theta^5.$$

   Thus Test B is more powerful than Test A, but its size is greater that of Test A.

4. Let X be a random sample of size 1 from the Exponential(θ) distribution, where $\theta > 0$ is a rate parameter. The null hypothesis $H_0 : \theta = 1/2$ is rejected in favour of the simple alternative $H_1 : \theta = 1$ if the observed value x is such that

   $$\frac{f(x; 1/2)}{f(x; 1)} \leq \frac{3}{4}.$$

   where $f(x; \theta)$ is the PDF of X. (This is defined to be $f(x; \theta) = \theta e^{-\theta x}$ for $x > 0$, and zero otherwise.)

1. Show that the size of the test is $\alpha = 1/3$.
2. Find the power of the test at $\theta = 1$.

Answer: Let T denote the test statistic:

$$T(X) = \frac{f(X;1/2)}{f(X;1)} = \frac{1}{2}e^{X/2}$$

1. The size of the test is

$$\begin{aligned}\alpha = \mathbb{P}_{H_0}(T \leq 3/4) &= \mathbb{P}_{H_0}(e^{X/2} \leq 3/2) \\ &= \mathbb{P}_{H_0}\big[X \leq 2\log(3/2)\big] \\ &= 1 - e^{-\log 3/2} \\ &= 1 - 2/3 = 1/3.\end{aligned}$$

2. The power of the test when $\theta = 1$ is

$$\begin{aligned}\gamma(1) = \mathbb{P}_{H_1}(T \leq 3/4) &= \mathbb{P}_{H_1}\big[X \leq 2\log(3/2)\big] \\ &= 1 - e^{-2\log 3/2} \\ &= 1 - (2/3)^2 = 5/9.\end{aligned}$$

9.3 The likelihood ratio test

Definition 9.17
The **likelihood ratio** statistic $\lambda : D \to \mathbb{R}$ for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is defined by

$$\lambda(\mathbf{x}) = \frac{\max\{L(\theta;\mathbf{x}) : \theta \in \Theta_0\}}{\max\{L(\theta;\mathbf{x}) : \theta \in \Theta_1\}}$$

If H_0 and H_1 are both simple hypotheses, say $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, the likelihood ratio reduces to

$$\lambda(\mathbf{x}) = \frac{L(\theta_0;\mathbf{x})}{L(\theta_1;\mathbf{x})}$$

If $\lambda(\mathbf{x})$ is small then $L(\theta_1;\mathbf{x})$ is large relative to $L(\theta_0;\mathbf{x})$ which indicates that θ_1 is more likely than θ_0 of being the true parameter value, and as a result we might be inclined to reject H_0.

Definition 9.18
The **likelihood ratio test** (LRT) of $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ is defined by the critical region

$$C = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\}$$

where k is chosen according to the required size of the test.

If H_0 and H_1 are both simple hypotheses, this is sometimes called the **simple likelihood ratio test** (SLRT).

Example 9.19
Let $X_1, \ldots, X_n$ be a random sample from the Exponential(θ) distribution, where $\theta > 0$ is an unknown rate parameter. Derive an explicit form for the critical region of the likelihood ratio test for testing the simple hypothesis $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$.

Solution:

Example 9.20

Let $X_1,\ldots,X_n$ be a random sample from the distribution of X, whose PDF is given by

$$f(x;\theta)=\begin{cases}\theta x^{\theta-1} & 0\leq x\leq 1\\ 0 & \text{otherwise,}\end{cases}$$

where $\theta\geq 1$ is an unknown scalar parameter. Construct a likelihood ratio test of the null hypothesis $H_0:\theta=1$ against the alternative $H_1:\theta>1$.

Solution:

9.4 Most powerful tests

Definition 9.21
Let C_1 and C_2 be two critical regions of size α for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$ and consider the power functions $\gamma_1(\theta)$ and $\gamma_2(\theta)$ of the associated tests,

$$\gamma_1(\theta) = \mathbb{P}_\theta(X \in C_1) \quad \text{and} \quad \gamma_2(\theta) = \mathbb{P}_\theta(X \in C_2) \quad \text{for } \theta \in \Theta_1.$$

If $\gamma_1(\theta) > \gamma_2(\theta)$ for all $\theta \in \Theta_1$, we say that C_1 is a **more powerful test** than C_2.

Definition 9.22
Let C be a critical region of size α for testing $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$. Then the associated test is called a **most powerful test** of size α if for every subset $A \subset D$ of size α,

$$\mathbb{P}_\theta(\mathbf{X} \in C) \geq \mathbb{P}_\theta(\mathbf{X} \in A) \quad \text{for all } \theta \in \Theta_1.$$

This means that the test is at least as powerful as any other test of size α.

9.4.1 The Neyman-Pearson lemma

Theorem 9.23 (The Neyman-Pearson lemma)
The likelihood ratio test is a most powerful test of a simple null hypothesis against a simple alternative.

Proof: Let $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. The SLRT is given by the critical region

$$C = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\} \quad \text{where} \quad \lambda(\mathbf{x}) = \frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})}$$

Let A be another critical region of size α. We need to show that $\mathbb{P}_{\theta_1}(C) \geq \mathbb{P}_{\theta_1}(A)$.

- If $\mathbf{x} \in C \setminus A$, then $\lambda(\mathbf{x}) \leq k$, so $L(\theta_1; \mathbf{x}) \geq \frac{1}{k} L(\theta_0; \mathbf{x})$, or equivalently $f(\mathbf{x}; \theta_1) \geq \frac{1}{k} f(\mathbf{x}; \theta_0)$
- If $\mathbf{x} \in A \setminus C$, then $\lambda(\mathbf{x}) > k$, so $L(\theta_1; \mathbf{x}) \leq \frac{1}{k} L(\theta_0; \mathbf{x})$, or equivalently $f(\mathbf{x}; \theta_1) \leq \frac{1}{k} f(\mathbf{x}; \theta_0)$

Hence, for continuous distributions (the discrete case is similar),

$$\begin{aligned}
\mathbb{P}_{\theta_1}(C) - \mathbb{P}_{\theta_1}(A) &= \int_C L(\theta_1; \mathbf{x})\, d\mathbf{x} - \int_A L(\theta_1; \mathbf{x})\, d\mathbf{x} \\
&= \int_{C\setminus A} L(\theta_1; \mathbf{x})\, d\mathbf{x} - \int_{A\setminus C} L(\theta_1; \mathbf{x})\, d\mathbf{x} \\
&\geq \frac{1}{k} \left(\int_{C\setminus A} L(\theta_0; \mathbf{x})\, dx - \int_{A\setminus C} L(\theta_0; \mathbf{x})\, dx \right) \\
&= \frac{1}{k} \left(\int_C L(\theta_0; \mathbf{x})\, dx - \int_A L(\theta_0; \mathbf{x})\, dx \right) \\
&= \frac{1}{k}(\alpha - \alpha) \\
&= 0.
\end{aligned}$$

Thus $\mathbb{P}_{\theta_1}(C) \geq \mathbb{P}_{\theta_1}(A)$ and because this holds for any critical region A of size α, we conclude that the SLRT is a most powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

Example 9.24
Let $X_1, \ldots, X_n$ be a random sample from the Poisson(θ) distribution. Find a most powerful test of the simple hypothesis $H_0 : \theta = 2$ against the simple alternative $H_1 : \theta = 1/2$.

Solution:

Example 9.25
Let $X_1, X_2, \ldots, X_{10}$ be a random sample from the Bernoulli(θ) distribution, where θ is unknown. The null hypothesis $H_0 : \theta = 1/2$ is rejected in favour of the alternative $H_1 : \theta < 1/2$ whenever the sum of the observations satisfies $\sum_{i=1}^{10} X_i \leq 2$.

1. Find the size of the test.
2. Find the power of the test at $\theta = 1/4$ and the power of the test at $\theta = 1/5$.
3. Show that this is a most powerful test of $H_0 : \theta = 1/2$ against the simple alternative $H_1 : \theta = 1/4$.

Solution:

Exercise 9.26

1. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from a distribution whose PDF is $f(x;\theta) = \theta x^{\theta-1}$ for $0 < x < 1$, and zero otherwise. Show that a most powerful test of $H_0 : \theta = 1$ against $H_1 : \theta = 2$ is given by the critical region

$$C = \left\{\mathbf{x} : \prod_{i=1}^{n} x_i \geq c\right\}.$$

Answer: The likelihood function is

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta x^{\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{\theta-1}.$$

In particular, $L(1|\mathbf{x}) = 1$ and $L(2|\mathbf{x}) = 2^n \prod_{i=1}^{n} x_i$, so the likelihood ratio for testing H_0 against H_1 is given by

$$\lambda(\mathbf{x}) = \frac{L(1|\mathbf{x})}{L(2|\mathbf{x})} = \frac{1}{2^n \prod_{i=1}^{n} x_i}.$$

By the Neymann-Pearson lemma, a most powerful test is given by

$$C = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\} = \left\{\mathbf{x} : \frac{1}{2^n \prod_{i=1}^{n} x_i} \leq k\right\} = \left\{\mathbf{x} : \prod_{i=1}^{n} x_i \geq \frac{1}{k2^n}\right\}.$$

Hence a most powerful test is given by a set of the form $C = \big\{\mathbf{x} : \prod_{i=1}^{n} x_i \geq c\big\}$, where c is chosen to fix the size of the test.

2. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random sample from the $N(\theta, 100)$ distribution, and let

$$C = \left\{\mathbf{x} : \frac{1}{n}\sum_{i=1}^{n} x_i \geq k\right\}$$

where k is a constant.

1. Show that C defines a most powerful test of $H_0 : \theta = 75$ against $H_1 : \theta = 78$.
2. Find values for n and k such that $\mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq k\right) = 0.05$ and $\mathbb{P}_{H_1}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq k\right) = 0.90$, approximately.

Answer:

1. The density function is $f(x;\theta,\sigma^2) = \dfrac{1}{\sigma\sqrt{2\pi}}\exp\left(-\dfrac{1}{2\sigma^2}(x-\theta)^2\right)$.

   The likelihood ratio for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ is therefore

$$\begin{aligned}\lambda(\mathbf{x}) &= \frac{L(\theta_0;\mathbf{x})}{L(\theta_1;\mathbf{x})} \\ &= \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_i(x_i-\theta_0)^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_i(x_i-\theta_1)^2\right)} \\ &= \exp\left(-\frac{1}{2\sigma^2}\sum_i\left[(x_i-\theta_0)^2-(x_i-\theta_1)^2\right]\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left[-2(\theta_0-\theta_1)\sum_i x_i + n(\theta_0^2-\theta_1^2)\right]\right)\end{aligned}$$

   By the Neyman-Pearson lemma, a most powerful test is given by

$$\begin{aligned}\{\mathbf{x}:\lambda(\mathbf{x})\leq k\} &= \left\{\mathbf{x}:\exp\left(-\frac{1}{2\sigma^2}\left[-2(\theta_0-\theta_1)\sum_i x_i + n(\theta_0^2-\theta_1^2)\right]\right)\leq k\right\} \\ &= \left\{\mathbf{x}\,:\,\frac{1}{n}\sum_i x_i \geq \frac{1}{2}(\theta_0+\theta_1) - \frac{\sigma^2\log k}{n(\theta_0-\theta_1)}\right\}\end{aligned}$$

   where we have used the fact that $\theta_0 < \theta_1$.

   Thus if $\theta_0 < \theta_1$, the set $C = \left\{\mathbf{x}\,:\,\frac{1}{n}\sum_i x_i \geq k\right\}$ defines a most powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$.

2. Under $H_0 : \theta = 75$, each $X_i \sim N(75, 100)$ so the sample mean is $\bar{X} \sim N(75, 100/n)$. For a test of size $\alpha = 0.05$, the critical value c must be the 95th percentile of the $N(75, 100/n)$ distribution, so $c = 75 + 1.645(10/\sqrt{n})$ where 1.645 is the 95th percentile of the standard normal distribution $N(0, 1)$.

   Under $H_1 : \theta = 78$, each $X_i \sim N(78, 100)$ so the sample mean is $\bar{X} \sim N(78, 100/n)$. For a test of $\gamma = 0.9$, the critical value c must be the 10th percentile of the $N(78, 100/n)$ distribution, so $c = 78 - 1.280(10/\sqrt{n})$ where -1.280 is the 10th percentile of the standard normal distribution $N(0, 1)$.

   Equating these expressions for c, we obtain $\sqrt{n} = 10(1.645 + 1.280)/(78 - 75) = 9.75$ and therefore $n = 95.0625$, which means that we need a sample size $n = 96$ to ensure that the significance level does not exceed 0.05. Substituting for $\sqrt{n}$ in one of the above expression then yields the critical value $c = 76.6872$, which defines the critical region

$$C = \left\{\mathbf{x}\,:\,\frac{1}{n}\sum_i x_i \geq 76.6872\right\}.$$

   As shown above, this defines a most powerful test of $H_0 : \theta = 75$ against $H_1 : \theta = 78$.

3. Let $X_1, X_2, \ldots, X_{10}$ be a random sample from the $N(0, \sigma^2)$ distribution where σ^2 is unknown.

1. Find a most powerful test of size $\alpha = 0.05$ for testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 2$.
2. Is this also a most powerful test of $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 4$?
3. Is this a most powerful test of $H_0 : \sigma^2 = 1$ against the composite alternative $H_1 : \sigma^2 > 1$?

Answer:

1. The density function is $f(x, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$, the likelihood function is

$$L(\sigma^2; \mathbf{x}) = \prod_{i=1}^{10} f(x_i, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^5 \exp\left(-\frac{1}{2\sigma^2}\sum_i x_i^2\right)$$

and the likelihood ratio for $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 2$ is therefore

$$\lambda(\mathbf{x}) = \frac{L(1; \mathbf{x})}{L(2; \mathbf{x})} = 2^5 \exp\left(-\frac{1}{4}\sum_i x_i^2\right).$$

Let $\mathbf{X} = (X_1, X_2, \ldots, X_{10})$ denote the random sample. By the Neymann-Pearson lemma, a best critical region for the test is

$$\begin{aligned} C = \{\mathbf{x} : \lambda(\mathbf{x}) \leq k\} &= \left\{\mathbf{x} : 2^5 \exp\left(-\frac{1}{4}\sum_{i=1}^{10} x_i^2\right) \leq k\right\} \\ &= \left\{\mathbf{x} : \sum_{i=1}^{10} x_i^2 \geq 4\log\left(\frac{k}{2^5}\right)\right\} \\ &= \left\{\mathbf{x} : \sum_{i=1}^{10} x_i^2 \geq k'\right\} \end{aligned}$$

where k' is chosen to ensure that $\mathbb{P}(\mathbf{X} \in C; H_0) = 0.05$.

2. Under the null hypothesis we have $X_i \sim N(0, 1)$, so the sum-of-squares $\sum_{i=1}^{10} X_i^2$ has chi-squared distribution with 10 degrees-of-freedom. From tables, the critical value at $\alpha = 0.05$ for this distribution is 18.307, so the critical region is given by

$$C = \left\{\mathbf{x} : \sum_{i=1}^{10} x_i^2 \geq 18.307\right\}$$

This is also a most powerful test of $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 4$.

3. The argument of part (b) holds for any simple alternative hypthesis $H_1 : \sigma^2 = \sigma_1^2$ provided $\sigma_1^2 > 1$. This is therefore a **uniformly most powerful test** of size α for testing $H_0 : \sigma^2 = 1$ against every simple alternative in the composite hypothesis $H_1 : \sigma^2 > 1$.

Chapter 10 Non-parametric methods

So far we have assumed that the parametric form of the distribution is known, for example $X \sim \text{Bernoulli}(\theta)$ where $\theta \in [0,1]$ is unknown or $X \sim \text{Poisson}(\theta)$ where $\theta > 0$ is unknown. We now look at methods for estimating the distribution of a random variable where the parametric form of the distribution is not known. Such methods are called **non-parametric** or **distribution-free** methods.

We consider only continuous distributions, so we can assume that the inverse CDF $F^{-1}(u)$ exists for all $u \in [0,1]$ and that the probability of different observations taking the same value is zero.

10.1 Order statistics

Definition 10.1
Let X be a continuous random variable and let $F(x)$ denote its CDF. For $p \in [0,1]$, the p**th quantile** of the distribution is the value $x_p = F^{-1}(p)$, i.e. the value $x_p \in \mathbb{R}$ for which

$$\mathbb{P}(X \leq x_p) = p \qquad \text{for } p \in [0,1].$$

In particular,

- $x_{0.5}$ is the **median** of the distribution,
- $x_{0.25}$ is the **lower quartile**,
- $x_{0.75}$ is the **upper quartile**,
- $x_{0.75} - x_{0.25}$ is the **inter-quartile range**.

Remark 10.2
As we shall see, the median and inter-quartile range represent **location** and **scale** respectively.

Example 10.3
Find the median and inter-quartile range of the Exponential(λ) distribution, where $\lambda > 0$ is a rate parameter.

Solution:

10.1.1 Order statistics

The quantiles of a distribution can be estimated using the **order statistics** of a random sample.

Definition 10.4
Let $X_1, X_2, \dots, X_n$ be a random sample from an unknown distribution. The **order statistic of rank** k is the kth smallest observation in the sample and denoted by $X_{(k)}$.

Certain functions of the order statistics $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ are important statistics in their own right:

- $X_{(1)}$ is the **sample minimum**.
- $X_{(n)}$ is the **sample maximum**.
- $X_{(n)} - X_{(1)}$ is the **sample range**.
- The **sample median** is $\begin{cases} X_{(n/2+1/2)} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\big[X_{(n/2)} + X_{(n/2+1)}\big] & \text{if } n \text{ is even.} \end{cases}$
- The **lower quartile** is the median of $X_{(1)}, \dots, X_{(n/2)}$ if n is even or $X_{(1)}, \dots, X_{(n/2-1/2)}$ if n is odd.
- The **upper quartile** is the median of $X_{(n/2+1)}, \dots, X_{(n)}$ if n is even or $X_{(n/2+3/2)}, \dots, X_{(n)}$ if n is odd.

Remark 10.5
The **five number summary** is a commonly used set of descriptive statistics consisting of the five most important sample quantiles:

[sample minimum, lower quartile, median, upper quartile, sample maximum]

These are sometimes illustrated using **box plots**.

The distribution of $X_{(k)}$ can be expressed in terms of the common CDF of the sample points.

Theorem 10.6
Let $X_1, X_2, \dots, X_n$ be a random sample from an unknown distribution and let F denote their common CDF. Then the CDF of the kth order statistic $X_{(k)}$ is

$$\mathbb{P}(X_{(k)} \leq x) = \mathbb{P}(W \geq k) \quad \text{where} \quad W \sim \text{Binomial}\big(n, F(x)\big).$$

Proof: By independence,

$$\begin{aligned} \mathbb{P}(X_{(k)} \leq x) &= \mathbb{P}(\text{at least } k \text{ observations are} \leq x) \\ &= \mathbb{P}\big(\text{at least } k \text{ successes in } n \text{ Bernoulli trials where P(success)} = F(x)\big) \\ &= \mathbb{P}(W \geq k) \quad \text{where} \quad W \sim \text{Binomial}\big(n, F(x)\big). \end{aligned}$$

10.1.2 Empirical distribution functions

Let X be a random variable whose distribution is unknown, and let $X_1, X_2, \dots, X_n$ be a random sample from the distribution of X.

Definition 10.7
The **empirical cumulative distribution function** (ECDF) of X is

$$\hat{F}_X(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

where $I(X_i \leq x)$ is the indicator variable of the event $\{\omega : X_i(\omega) \leq x\}$.

In terms of order statistics, the ECDF can be written as

$$\hat{F}_X(x) = \begin{cases} 0 & \text{for } x < X_{(1)}, \\ i/n & \text{for } X_{(i)} < x \leq X_{(i+1)}, \\ 1 & \text{for } x \geq X_{(n)}. \end{cases}$$

Remark 10.8
$\hat{F}_X(x)$ is the proportion of observations that are less than or equal to x. By the law of large numbers applied to the indicator variables $I(X_i \leq x)$,

$$\hat{F}_X(x) \to \mathbb{P}(X \leq x) \quad \text{in probability as the sample size } n \to \infty \text{ for all } x \in \mathbb{R}.$$

Remark 10.9 (Goodness-of-fit)
Let F be an estimate for the CDF of X. We can quanify the so-called **goodness of fit** using a number of test statistics based on the empirical CDF of a random sample taken from the distribution of X, some of which are shown in Table 10.1.

Kolmogorov-Smirnov	$\max \lvert\hat{F}(x) - F(x)\rvert$
Cramer-von Mises	$\int_{-\infty}^{\infty} \big[\hat{F}(x) - F(x)\big]^2 f(x)\,dx$
Anderson-Darling	$\int_{-\infty}^{\infty} \frac{\big[\hat{F}(x) - F(x)\big]^2}{F(x)\big[1 - F(x)\big]} f(x)\,dx$

Table 10.1: Test statistics for goodness-of-fit.

10.2 Location-scale models

We seek to identify classes of parameters, in particular **location parameters** and **scale parameters**.

We can think of a parameter as a **function** of the CDF (or PDF/PMF) of a random variable: we call these **functionals**[1]. For example,

$$\begin{aligned} T(F_X) &= \int_{-\infty}^{\infty} x f_X(x)\,dx && \text{is the } \textbf{mean functional}, \\ T(F_X) &= F_X^{-1}(1/2) && \text{is the } \textbf{median functional}. \end{aligned}$$

Remark 10.10
The empirical CDF $\hat{F}_X$ is itself a CDF so we can apply the functional T to $\hat{F}_X$:

- $T(\hat{F}_X)$ is called the **induced estimator** of $T(F_X)$.

[1] The term *functional* is a generic term used for a function that maps functions to scalar values.

For example,

- if T is the mean functional, $T(\hat{F}_X)$ is the sample mean;
- if T is the median functional, $T(\hat{F}_X)$ is the sample median.

Definition 10.11

1. T is said to be a **location functional** if
$$T(F_{a+bX}) = a + bT(F_X) \quad \text{for all } a, b \in \mathbb{R}.$$
2. T is said to be a **scale functional** if
$$T(F_{a+bX}) = bT(F_X) \quad \text{for all } b > 0.$$

Example 10.12

1. Show that the mean is a location functional.
2. Show that the standard deviation is a scale functional.

Solution:

Definition 10.13
A statistical model $\mathcal{M}$ is called a

1. **location model** if $F(a + x) \in \mathcal{M}$ whenever $F \in \mathcal{M}$ and $a \in \mathbb{R}$.
2. **location-scale model** if $F(a + bx) \in \mathcal{M}$ whenever $F \in \mathcal{M}$ and $a, b \in \mathbb{R}$ with $b > 0$.

Example 10.14
Show that the family of uniform distributions $\mathcal{M} = \left\{\frac{x-L}{R-L} : L, R \in \mathbb{R}, L < R\right\}$ is a location-scale model.

Solution:

10.2.1 Q-Q plots

Let $\mathcal{M}$ be a location-scale model and let Z be a random variable whose CDF $F_Z \in \mathcal{M}$ is known. Suppose we have another random variable X whose CDF F_X is unknown, and that we wish to test whether $F_X \in \mathcal{M}$.

Because $\mathcal{M}$ is a location-scale model, if $F_X \in \mathcal{M}$ then

$$X = a + bZ$$

where a and $b > 0$ are unknown parameters. Since $b > 0$ the CDF of X satisfies

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(a + bZ \leq y) = \mathbb{P}\left(Z \leq \frac{x-a}{b}\right) = F_Z\left(\frac{x-a}{b}\right).$$

Let x_p and z_p denote the pth quantiles of X and Z respectively (where $0 < p < 1$). Then

- $z_p = F_Z^{-1}(p)$ is known
- $x_p = a + bF_Z^{-1}(p)$ is unknown (because a and b are unknown).

If $F_X \in \mathcal{M}$ we have the linear relationship

$$x_p = a + bz_p.$$

The quantiles x_p of F_X are unknown, but they can be estimated by order statistics.

- Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics of a random sample from the distribution of X.
- $X_{(k)}$ is a point estimator of the quantile x_{p_k} where $p_k = k/(n+1)$.

Definition 10.15
A plot of the order statistics $X_{(k)}$ against the quantiles z_{p_k} is called a **Quantile-Quantile plot** (Q-Q plot).

- If $F_X \in \mathcal{M}$, the plot should be approximately linear.
- The parameters a and b can be estimated by the intercept and gradient respectively.

Example 10.16
To test whether a random sample $X_1, X_2, \ldots, X_n$ is from a Normal distribution, we plot the order statistics $X_{(k)}$ against the values $z_{p_k} = \Phi^{-1}\big[k/(n+1)\big]$ for $k = 1, 2, \ldots, n$ where Φ is the CDF of $N(0,1)$. If the points lie on (or near) a straight line, we might conclude that the X_i are normally distributed (see Figure 10.1).

Exercise 10.17

1. Let X be a random variable, let F_X be its CDF and suppose that its distribution is symmetric about a. Using the fact that $X - a$ and $-(X - a)$ have the same distribution, show that any location functional satisfies $T(F_X) = a$.

   Answer: Because T is a location functional,

   $$T(F_{X-a}) = T(F_X) - a \qquad \text{and} \qquad T(F_{-(X-a)}) = -T(F_X) + a.$$

   Because $X - a$ and $-(X - a)$ have the same distribution, these are equal so $T(F_X) = a$, as required.

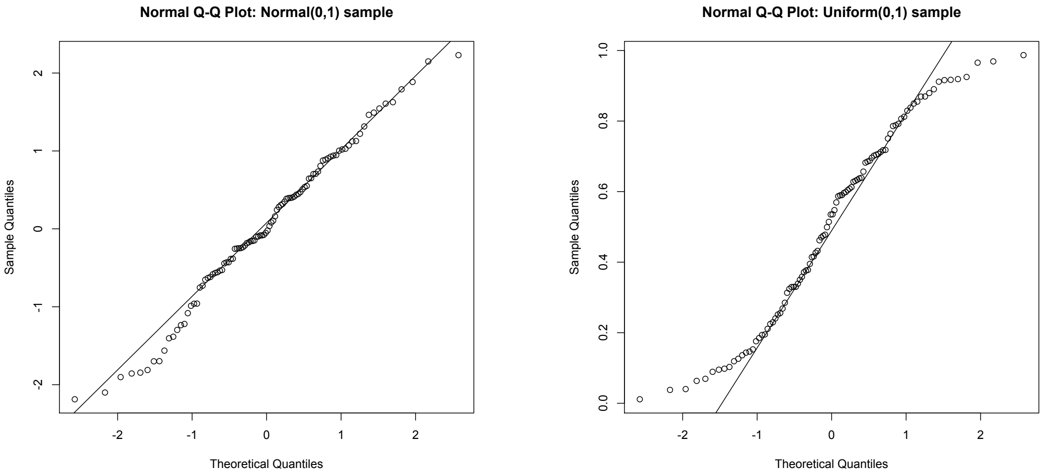

Figure 10.1: Q-Q plots for testing normality.

2. (a) Show that the median is a location functional.

Answer: Let $Y = a + bX$. Then

$$F_{a+bX}(y) = \begin{cases} F_X\left(\frac{y-a}{b}\right) & \text{if } b > 0, \\ 1 - F_X\left(\frac{y-a}{b}\right) & \text{if } b < 0. \end{cases}$$

Let $T(F_X) = F_X^{-1}(1/2)$ be the median functional. Then $F_X\big[T(F_X)\big] = 1/2$ so for $b > 0$,

$$F_{a+bX}\big[a + bT(F_X)\big] = F_X\left[\frac{a + bT(F_X) - a}{b}\right] = F_X\big[T(F_X)\big] = 1/2,$$

and for $b < 0$,

$$F_{a+bX}\big[a + bT(F_X)\big] = 1 - F_X\left[\frac{a + bT(F_X) - a}{b}\right] = 1 - F_X\big[T(F_X)\big] = 1 - 1/2 = 1/2.$$

In either case we have

$$T(F_{a+bX}) = F_{a+bX}^{-1}(1/2) = a + bT(F_X),$$

so the median is a location functional, as required.

(b) Show that the inter-quartile range is a scale functional.

Answer: Let $L(F_X) = F_X^{-1}(1/4)$ and $U(F_X) = F_X^{-1}(3/4)$ be the lower-quartile and upper-quartile functionals respectively. Then

$$F_X\big[L(F_X)\big] = 1/4 \quad \text{and} \quad F_X\big[U(F_X)\big] = 3/4.$$

For $b > 0$,

$$\begin{aligned} F_{a+bX}\big[a + bL(F_X)\big] &= F_X\left[\frac{a + bL(F_X) - a}{b}\right] = F_X\big[L(F_X)\big] = 1/4, \\ F_{a+bX}\big[a + bU(F_X)\big] &= F_X\left[\frac{a + bU(F_X) - a}{b}\right] = F_X\big[U(F_X)\big] = 3/4, \end{aligned}$$

Thus $L(F_{a+bX}) = F^{-1}_{a+bX}(1/4) = a + bL(F_X)$ and
$U(F_{a+bX}) = F^{-1}_{a+bX}(3/4) = a + bU(F_X)$.

The inter-quartile range functional is $T(F_X) = U(F_X) - L(F_X)$, so

$$T(F_{a+bX}) = F^{-1}_{a+bX}(3/4)-F^{-1}_{a+bX}(1/4) = \big[a+bU(F_X)\big]-\big[a+bL(F_X)\big] = b\big[U(F_X)-L(F_X)\big] = bT(F_X).$$

Thus the inter-quartile range is a scale functional, as required.

10.3 One-sample tests

10.3.1 The sign test

Let X be a continuous random variable with an unknown median η. The sign test is a non-parametric method for testing hypotheses about η.

The test statistic

Let $X_1, X_2, \dots, X_n$ be a random sample from the distribution of X and consider the null hypothesis $H_0 : \eta = \eta_0$ against a suitable alternative. If H_0 is correct then approximately half of the observations should be smaller than η_0 and approximately half should be larger than η_0.

Definition 10.18
The sign test statistic S^+_n is the number of observations larger than η_0:

$$S^+_n = \sum_{i=1}^n Z_i \quad \text{where} \quad Z_i = I(X_i > \eta_0).$$

We also define the complementary statistic $S^-_n = \sum_{i=1}^n (1 - Z_i)$ which is the number of observations smaller than η_0. Note that $S^+_n + S^-_n = n$.

Because the observations are independent, the distribution of our test statistic under H_0 is

$$S^+_n \sim \text{Binomial}(n, 0.5).$$

- Small values of S^+_n support the alternative $H_1 : \eta < \eta_0$.
- Large values of S^+_n support the alternative $H_1 : \eta > \eta_0$.

Example 10.19
The following are measurements of the breaking strength of a certain kind of two-inch cotton ribbon.

163	165	158	189	161	171	158	151	169	162
163	139	172	165	148	166	172	163	187	173

Conduct a sign test to decide between $H_0 : \eta = 160$ and $H_1 : \eta > 160$ at significance level $\alpha = 0.05$.

Solution:

Example 10.20 (Sign test for paired samples)
To evaluate a new traffic-control system, the number of accidents that occurred at 12 dangerous junctions were recorded during the four weeks prior to the installation of the new system, and for the four weeks after its installation. The following data were obtained.

Junction	1	2	3	4	5	6	7	8	9	10	11	12
Before	3	5	2	3	3	3	0	4	1	6	4	1
After	1	2	2	2	2	0	2	3	3	4	1	0

Use a sign test to evaluate the claim that the new system is more effective than the old system.

Solution:

Normal approximation

By the central limit theorem, if $X \sim \text{Binomial}(n, \theta)$ then for large n,

$$X \sim N\big(n\theta, n\theta(1-\theta)\big) \quad \text{approx.}$$

If $H_0 : \theta = 0.5$ is correct, the distribution of the test statistic is $S_n^+ \sim N(n/2, n/4)$ approx.

Definition 10.21 (The continuity correction)
Let X be a discrete random variable, taking values in the set $\{0, \pm 1, \pm 2, \ldots\}$. If the distribution of a continuous random variable Y is taken as an approximation of the distribution of X, we set

$$\mathbb{P}(X = k) = \mathbb{P}\left(k - \frac{1}{2} < Y < k + \frac{1}{2}\right).$$

This means that

- $\mathbb{P}(X < k) = \mathbb{P}(Y \leq k - 1/2)$ and $\mathbb{P}(X \leq k) = \mathbb{P}(Y \leq k + 1/2)$,
- $\mathbb{P}(X \geq k) = \mathbb{P}(Y \geq k - 1/2)$ and $\mathbb{P}(X > k) = \mathbb{P}(Y \geq k + 1/2)$

Example 10.22 (Sign test for large samples)
The following data are the amounts of sulphur oxide (in tonnes) emitted by a large industrial plant over a period of 40 days.

17	15	20	29	19	18	22	25	27	9
24	20	17	6	24	14	15	23	24	26
19	23	28	19	16	22	24	17	20	13
19	10	23	18	31	13	20	17	24	14

Construct a sign test of size $\alpha = 0.01$ to evaluate $H_0 : \eta = 21.5$ against $H_1 : \eta < 21.5$.

Solution:

Exercise 10.23

1. The biting rate of a particular species of fly was investigated. The biting rate is defined as the number of flies biting a volunteer during 15 minutes of exposure. The species is known to have a median biting rate of 5 bites per 15 minutes. It is hypothesized that the median biting range is higher in bright, sunny weather. To test the hypothesis, a total of 122 volunteers were exposed to flies on a sunny day, of which 95 experienced biting rates greater than 5. State the null and alternative hypotheses for the test, and state your conclusion for $\alpha = 0.01$.

   Answer: Let η deonte the (true) median biting rate.

   The hypothesis test is $H_0 : \eta = 5$ against, $H_1 : \eta > 5$.

   The test statistic is $S_n^+ = 95$ where $n = 122$.

   Under the null hypothesis, $\mathbb{P}(S_{122}^+ \geq 95) = \mathbb{P}\big[\text{Binomial}(122, 0.5) \geq 95\big]$.

   Since n is large, we use the normal approximation: under the null hypothesis,

   $$Z = \frac{(S_n^+ - 1/2) - n/2}{\sqrt{n/4}} \sim N(0,1) \quad \text{approx.}$$

   In this case, the test statistic is

   $$z = \frac{94.5 - 61}{\sqrt{30.5}} = 6.0659.$$

   An approximate p-value is $\mathbb{P}(Z > 6.6059) < 0.001$, so we reject H_0 at $\alpha = 0.01$.

2. Let $X \sim N(\mu, 1)$ where μ is unknown and suppose we wish to test the simple null hypothesis $H_0 : \mu = 0$ against the simple alternative $H_1 : \mu = 0.5$. A random sample of 9 observations is taken from the distribution of X and the number S^+ of positive values is counted.

   1. A sign test rejects the null hypothesis if S^+ exceeds 6. Find the size and power of the test.
   2. Construct a test based on the sample mean of the observations which has the same significance level as the sign test described above. Find the power of the test, and explain why this is higher than the power of the sign test.

   Answer:

   1. Under the null hypothesis, $S^+ \sim \text{Binomial}(9, 0.5)$. The size of the test is

      $$\alpha = \mathbb{P}_{\mu_0}(S^+ > 6) = \mathbb{P}\big(\text{Binomial}(9, 0.5) > 6\big) \approx 0.0898 \text{ (from tables).}$$

      Under $H_1 : \mu = 0.5$ we have $X \sim N(0.5, 1)$, so the probability that an observataion takes a positive value under H_1 is

      $$\mathbb{P}_{\mu_1}(X > 0) = \mathbb{P}\big(X > 0 \text{ where } X \sim N(0.5, 1)\big) = \mathbb{P}\big(Z > -0.5 \text{ where } Z \sim N(0,1)\big) \approx 0.69146 \quad \text{(from tables).}$$

      The power of the test to detect the alternative $H_1 : \mu = 0.5$ is therefore

      $$\gamma(0.5) = \mathbb{P}_{\mu_1}(S^+ > 6) = \mathbb{P}\big[\text{Binomial}(9, 0.6915) > 6\big] = \mathbb{P}\big[\text{Binomial}(9, 0.0.3085) < 3\big]$$

where the second equality follows by the fact that

$$\mathbb{P}\big[\text{Binomial}(n,p) > k\big] = \mathbb{P}\big[\text{Binomial}(n, 1-p) < n-k\big].$$

From tables, we find that

$$\mathbb{P}[\text{Binomial}(9, 0.30) \leq 2] \approx 0.4628 \quad \text{and} \quad \mathbb{P}[\text{Binomial}(9, 0.35) \leq 2] \approx 0.3373.$$

To find the required probability, we interpolate between these values:

$$\begin{aligned}\mathbb{P}\big[\text{Binomial}(9, 0.3085) \leq 2\big] &\approx \mathbb{P}\big[\text{Binomial}(9, 0.30) \leq 2\big] \\ &\quad + \left(\frac{0.3085 - 0.30}{0.35 - 0.30}\right)\Big(\mathbb{P}\big[\text{Bino}(9, 0.35) \leq 2\big] - \mathbb{P}\big[\text{Bino}(9, 0.30) \leq 2\big]\Big) \\ &= 0.4628 - (0.1708 \times 0.1255) \\ &= 0.4414.\end{aligned}$$

Alternatively, we can use the normal approximation: if $Y \sim \text{Binomial}(9, 0.6915)$ then $\mathbb{E}(Y) = 6.2231$ and $\text{Var}(Y) = 1.9201$. Using the continuity correction,

$$\mathbb{P}\big[\text{Binomial}(9, 0.6915) > 6\big] \approx \mathbb{P}\left[N(0,1) > \frac{6.5 - 6.2231}{\sqrt{1.9201}}\right] = \mathbb{P}\big(N(0,1) > 0.1998\big) \approx 0.4207.$$

2. Let $\bar{X}$ denote the sample mean, and consider the z-test, where $H_0 : \mu = 0$ is rejected in favour of $H_1 : \mu = 0.5$ if $\bar{X} > c$, with c chosen to give the required significance level. Here we require $\alpha = 0.0898$, so we need

$$\mathbb{P}_{\mu_0}(\bar{X} > c) = 0.0898.$$

Under H_0 we have $X_i \sim N(0,1)$, so $\mathbb{E}(\bar{X}) = 0$ and $\text{Var}(\bar{X}) = 1/9$. Thus $\bar{X} \sim N(0, 1/9)$, so the critical value c satisfies

$$\mathbb{P}\left(N(0,1) > \frac{c}{\sqrt{1/9}}\right) = 0.0898,$$

i.e. $1 - \Phi(3c) = 0.0898$, or $\Phi(3c) = 0.9102$. From tables, $\Phi(1.34076) = 0.91$ and $\Phi(1.40507) = 0.92$. Interpolating linearly between these values gives

$$3c = 1.34076 + 0.02(1.40507 - 1.34076) = 1.342,$$

so $c \approx 0.447$. The statistical power is

$$\begin{aligned}\gamma(0.5) = \mathbb{P}_{H_1}(\bar{X} > 0.447) &= \mathbb{P}\big(N(0.5, 1/9) > 0.447\big) \\ &= \mathbb{P}\big(N(0,1) > 3(0.447 - 0.5)\big) \\ &= \mathbb{P}\big(N(0,1) > -0.159)\big) \\ &\approx 0.564 \quad \text{(from tables)}.\end{aligned}$$

This is higher than the power of the sign test (which is approximately 0.43), because the z-test takes account of the magnitude of the observations, as well as their signs. If the data really do come from a normal distribution, the z-test is more powerful than the sign test for detecting $H_1 : \mu = 0.5$ against $H_0 : \mu = 0$.

10.3.2 The Wilcoxon signed-rank test

The Wilcoxon signed rank test is an extension of the sign test.

- The sign test counts the **number** of observations which are larger than some fixed value.

- The WSR test also considers the **relative size** of these observations.

Definition 10.24
For a random sample $X_1, \ldots, X_n$, the **rank** of observation X_i is its position in the sequence of observations sorted in ascending order:

$$R(X_i) = \sum_{j=1}^{n} I(X_j \leq X_i)$$

Note that the sum of the ranks is always equal to the sum of the first n positive integers:

$$\sum_{i=1}^{n} R(X_i) = \sum_{i=1}^{n} i = \frac{1}{2}n(n+1).$$

Remark 10.25 (Ties)
If two or more observations have the same value, they are said to be **tied**. If the distribution is continuous, ties occur with probability zero. Ties often occur in practical applications however, due to the limited precision of measurements. To deal with ties, the usual approach is to assign an **average rank** to each of the tied observations. For example, if there are $m-1$ observations strictly smaller than $X_i = X_j$, we set

$$R(X_i) = R(X_j) = \frac{m + (m+1)}{2} = m + \frac{1}{2}.$$

The test statistic

Let X be a continuous random variable whose distribution is **symmetric** and whose median η is unknown, and let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X. Without loss of generality, we consider the null hypothesis $H_0 : \eta = 0$:

- for a single sample $X_1, X_2, \ldots, X_n$, the null hypothesis $H_0 : \eta = \eta_0$ reduces to $H_0 : \eta = 0$ by looking at the differences $D_i = X_i - \eta_0$;
- For paired samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$, the null hypothesis $H_0 : \eta_1 = \eta_2$ reduces to $H_0 : \eta = 0$ by looking at the differences $D_i = X_i - Y_i$.

Definition 10.26
For the null hypothesis $H_0 : \eta = 0$, the **Wilcoxon signed rank** (WSR) statistic is

$$W_n^+ = \sum_{i=1}^{n} R_i Z_i \quad \text{where} \quad R_i = \sum_{j=1}^{n} I(|X_j| \leq |X_i|) \quad \text{and} \quad Z_i = I(X_i > 0).$$

- R_i is the rank of $|X_i|$ in the sample of absolute values $|X_1|, |X_2|, \ldots, |X_n|$;
- $Z_i = 1$ if $X_i > 0$, otherwise $Z_i = 0$.

We also define the complementary statistic $W_n^- = \sum_{i=1}^{n} R_i(1 - Z_i)$. Note that

$$W_n^+ + W_n^- = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} i = \frac{1}{2}n(n+1).$$

Example 10.27
A study of the effects of smoking by mothers on the birthweight of their children involved 12 pairs of mothers. The pairs of mothers were selected so that they were as similar as possible,

except that one mother of each pair was a smoker and the other was not. The birthweights (in kilograms) of the babies were as follows.

Pair	1	2	3	4	5	6	7	8	9	10	11	12
Non-smoker	3.22	4.48	3.90	3.47	3.07	3.23	4.25	3.31	3.33	3.78	3.18	4.60
Smoker	3.00	4.27	3.95	3.32	2.51	2.77	4.02	3.41	3.39	3.88	3.18	4.37

Do these data support the claim that mothers who smoke tend to have babies with smaller birthweights than those who do not?

Solution:

Mean and variance of W_n^+

Theorem 10.28
Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous distribution whose density function is symmetric about its mean. Under the null hypothesis $H_0 : \eta = 0$,

$$\mathbb{E}(W_n^+) = \frac{n(n+1)}{4} \qquad \text{and} \qquad \mathrm{Var}(W_n^+) = \frac{n(n+1)(2n+1)}{24}.$$

Proof: Recall that

$$W_n^+ = \sum_{i=1}^n R_i Z_i \quad \text{where} \quad R_i = \sum_{j=1}^n I(|X_j| \leq |X_i|) \quad \text{and} \quad Z_i = I(X_i > 0).$$

Under $H_0 : \eta = 0$ we have $Z_i \sim \text{Bernoulli}(1/2)$ and hence $\mathbb{E}(Z_i) = 1/2$ and $\mathrm{Var}(Z_i) = 1/4$.

$$\begin{aligned}
\mathbb{E}(W_n^+) &= \sum_{i=1}^n R_i \mathbb{E}(Z_i) = \frac{1}{2}\sum_{i=1}^n R_i = \frac{1}{2}\sum_{i=1}^n i = \frac{1}{4}n(n+1). \\
\mathrm{Var}(W_n^+) &= \sum_{i=1}^n R_i^2 \mathrm{Var}(Z_i) = \frac{1}{4}\sum_{i=1}^n R_i^2 = \frac{1}{4}\sum_{i=1}^n i^2 = \frac{1}{24}n(n+1)(2n+1).
\end{aligned}$$

Normal approximation

W_n^+ is the sum of random variables and (although the random variables are not independent) it satisfies the central limit theorem. Under $H_0 : \eta = 0$, and provided n is sufficiently large, the distribution of W_n^+ is

$$W_n^+ \sim N\left(\frac{1}{4}n(n+1), \frac{1}{24}n(n+1)(2n+1)\right) \quad \text{approx.}$$

The continuity correction should be applied when defining the test statistic.

- Lower-talied test:

$$Z = \frac{(W_n^+ + \frac{1}{2}) - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \sim N(0,1) \quad \text{approx.}$$

- Upper-talied test:

$$Z = \frac{(W_n^+ - \frac{1}{2}) - \frac{1}{4}n(n+1)}{\sqrt{\frac{1}{24}n(n+1)(2n+1)}} \sim N(0,1) \quad \text{approx.}$$

Exact distribution of W_n^+ under H_0

The exact distribution of W_n^+ under H_0 can be obtained for small n by a recurrence relation. The base case is $n = 1$ (a single observation), for which $W_1^+ \in \{0, 1\}$ and under $H_0 : \eta = 0$,

$$\mathbb{P}(W_1^+ = 0) = \frac{1}{2} \quad \text{and} \quad \mathbb{P}(W_1^+ = 1) = \frac{1}{2}.$$

Theorem 10.29
Under $H_0 : \eta = 0$,

$$\mathbb{P}\big(W_{n+1}^+ = k\big) = \frac{1}{2}\mathbb{P}\big(W_n^+ = k\big) + \frac{1}{2}\mathbb{P}\big(W_n^+ = k - (n+1)\big).$$

for $k = 0, 1, \ldots, \frac{1}{2}n(n+1)$, and zero otherwise.

Proof:

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous and symmetric distribution.
- Let R_i be the rank of $|X_i|$ among the absolute values $|X_1|, |X_2|, \ldots, |X_n|$:

Then W_n^+ can be written as

$$W_n^+ = \sum_{r=1}^{n} rU_r \qquad \text{where} \qquad U_{R_i} = \begin{cases} 1 & X_i > 0, \\ 0 & \text{otherwise.} \end{cases}$$

- U_r indicates that the observation associated with rank r has a positive sign.
- Under the null hypothesis, $U_1, U_2, \ldots, U_n$ is a random sample from the Bernoulli(1/2) distribution.

Let $G_n(t)$ denote the probability generating function of W_n^+:

$$\begin{aligned} G_n(t) = \mathbb{E}(t^{W_n^+}) = \mathbb{E}(t^{\sum_{r=1}^n rU_r}) &= \mathbb{E}(t^{U_1}t^{2U_2}\cdots t^{nU_n}) \\ &= \prod_{r=1}^{n} \mathbb{E}(t^{rU_r}) \qquad \text{(by independence)} \\ &= \prod_{r=1}^{n} \Big[t^0\,\mathbb{P}(U_r = 0) + t^r\,\mathbb{P}(U_r = 1)\Big] \\ &= \prod_{r=1}^{n} \frac{1}{2}(1 + t^r). \end{aligned}$$

Hence,

$$G_{n+1}(t) = \frac{1}{2}(1 + t^{n+1})G_n(t).$$

By definition,

$$G_n(t) = \sum_{k=0}^{\frac{1}{2}n(n+1)} \mathbb{P}(W_n^+ = k)t^k \qquad \text{and} \qquad G_{n+1}(t) = \sum_{k=0}^{\frac{1}{2}(n+1)(n+2)} \mathbb{P}(W_{n+1}^+ = k)t^k.$$

Thus,

$$\sum_{k=0}^{\frac{1}{2}(n+1)(n+2)} \mathbb{P}(W_{n+1}^+ = k)t^k = \frac{1}{2}(1 + t^{n+1}) \sum_{k=0}^{\frac{1}{2}n(n+1)} \mathbb{P}(W_n^+ = k)t^k,$$

Comparing the coefficients of t^k yields the required recurrence relation:

$$\mathbb{P}\big(W_{n+1}^+ = k\big) = \frac{1}{2}\mathbb{P}\big(W_n^+ = k\big) + \frac{1}{2}\mathbb{P}\big(W_n^+ = k - (n+1)\big).$$

This recurrence relation is used to compute the quantiles of W_n^+ listed in statistical tables.

Exercise 10.30

1. We wish to test the hypothesis that two treatments A and B are equivalent, against the alternative hypothesis that the responses to treatment A tend to be larger than the responses to treatment B. We perform a paired difference experiment, and analyse the resulting data using the Wilcoxon signed rank test. The data is shown in the following table.

Pair	1	2	3	4	5	6	7	8	9	10
A	54	60	98	43	82	77	74	29	63	80
B	45	45	87	31	71	75	63	30	59	82

State the null and alternative hypotheses for the test, and conduct the test at significance level $\alpha = 0.05$.

Answer:

1. Let $D_i = A_i - B_i$ denote the differences, and let η denote the (true) median difference. We test the null hypothesis $H_0 : \eta = 0$ against the one-sided alternative $H_1 : \eta > 0$.
2. The signed ranks are computed as follows:

Pair	1	2	3	4	5	6	7	8	9	10
Difference	9	15	11	12	11	2	11	-1	4	-2
Signed Rank	5	10	7	9	7	2.5	7	-1	4	-2.5

- The Wilcoxon signed rank statistics are $W^+ = 51.5$ and $W^- = 3.5$.
- Check: $\frac{1}{2}n(n+1) = 55 = W^+ + W^-$.
- From tables, for a one-tailed test at $\alpha = 0.05$ and $n = 10$, the critical value is $w_c^+ = 44$.
- Thus we reject $H_0 : \eta = 0$ in favour of $H_1 : \eta > 0$, and conclude that responses to treatment A tend to be larger than responses to treatment B.

2. In a comparison of two populations A and B, a paired difference experiment with $n = 30$ pairs yields the Wilcoxon signed-rank statistic $w^+ = 354$.

1. Construct a test to determine whether or not population A is located to the right of population B.
2. Conduct the test at $\alpha = 0.05$.
3. Repeat part (2) using a normal approximation to the distribution of W^+.

Answer:

1. Let η_A and η_B denote the (true) medians of populations A and B respectively. To determine whether population A is located to the right of population B, we test the null hypothesis $H_0 : \eta_A = \eta_B$ against the alternative $H_1 : \eta_A > \eta_B$. We could also define $\eta = \eta_A - \eta_B$ to be the difference between the two medians, in which case we would test $H_0 : \eta = 0$ agaiaunst $H_1 : \eta > 0$.
2. From tables, for a one-tailed test at $\alpha = 0.05$ and $n = 30$, the critical value is $w_c^+ = 313$, so we reject H_0 in favour of H_1.

   To compute an approximate p-value for the test, from tables (for $n = 30$) we see that $w_c^+ = 345$ at $\alpha = 0.01$ and $w_c^+ = 356$ at $\alpha = 0.005$. Interpolating between these values,

$$\mathbb{P}_{H_0}(W^+ \geq 354) \approx 0.01 + \left(\frac{354 - 345}{356 - 345}\right)(0.005 - 0.01) = 0.0059.$$

3. The normal approximation is $\mathbb{P}(W^+ \geq w_c^+) \approx \mathbb{P}(Z \geq z_c)$ where

$$Z = \frac{(W^+ - 1/2) - \mathbb{E}(W^+)}{\sqrt{\text{Var}(W^+)}} \sim N(0,1) \quad \text{approx.},$$

Here, $\mathbb{E}(W^+)$ and $\text{Var}(W^+)$ are respectively the mean and variance of W^+ under H_0. In this case,

$$\mathbb{E}(W^+) = \frac{1}{4}n(n+1) = 232 \quad \text{and} \quad \text{Var}(W^+) = \frac{1}{24}n(n+1)(2n+1) = 2363,$$

so the test statistic is

$$z = \frac{353.5 - 232}{\sqrt{2363}} = 2.4994.$$

- At $\alpha = 0.05$ we have $z_c = 1.645$, so we reject H_0.
- From tables, an approximate p-value for the test is $1 - 0.99379 = 0.0062$.

3. To test whether or not a new diet actually results in weight loss, a researcher recruited nine subjects, measured their weight (in kilograms) before starting the diet, and then again after two months on the diet. The results are shown in the following table.

Case	1	2	3	4	5	6	7	8	9
Weight before	55	50	54	43	62	61	56	48	53
Weight after	47	40	52	43	51	62	50	47	56

(a) Use a sign test with $\alpha = 0.05$ to decide whether or not the diet results in weight loss.
Answer: Let η denote the (true) median weight loss.
We evaluate $H_0 : \eta = 0$ against the alternative $H_0 : \eta > 0$.

Case (i)	1	2	3	4	5	6	7	8	9
Weight before	55	50	54	43	62	61	56	48	53
Weight after	47	40	52	43	51	62	50	47	56
Difference (D_i)	8	10	2	0	11	-1	6	1	-3
Sign	$+$	$+$	$+$		$+$	$-$	$+$	$+$	$-$

- Exclude the zero difference, and take the sample size to be $n = 8$.
- $S_n^+ = \sum_i I(D_i > 0) = 6$.
- $S_n^- = \sum_i I(D_i < 0) = 2$.
- Check: $S_n^+ + S_n^- = n$.

Under H_0, $S_n^+ \sim \text{Binomial}(n, 0.5)$. From tables,

$$\mathbb{P}_{H_0}(S_n^+ \geq 6) = 1 - \mathbb{P}_{H_0}(S_n^+ \leq 5) = 1 - 0.8555 = 0.1445.$$

At $\alpha = 0.05$, we would decide to retain the null hypothesis $H_0 : \eta = 0$, and conclude that the diet results in no weight loss.

(b) Repeat the analysis using the Wilcoxon signed rank test.
Answer:

Difference (D_i)	8	10	2	0	11	-1	6	1	-3
Sign	$+$	$+$	$+$		$+$	$-$	$+$	$+$	$-$
Rank (R_i)	6	7	3		8	1.5	5	1.5	4

- $W^+ = \sum_i R_i I(D_i > 0) = 6 + 7 + 3 + 8 + 5 + 1.5 = 30.5$.
- $W^- = \sum_i R_i I(D_i < 0) = 1.5 + 4 = 5.5$.
- Check: $W^+ + W^- = \frac{1}{2}n(n+1)$.

From tables, critical values for an upper-tail test are

- $w_c^+ = 30$ at $\alpha = 0.05$, and
- $w_c^+ = 32$ at $\alpha = 0.025$.

Thus at $\alpha = 0.05$ we would reject the null hypothesis $H_0 : \eta = 0$ in favour of the alternative $H_0 : \eta > 0$, and conclude that the diet does indeed result in weight loss.

(c) Briefly discuss the reasons why the tests lead to different conclusions.

Answer: The sign test takes no account of the fact that the average amount of weight lost by those whose weight decreased over the study period, is significantly greater than the average amount of weight gained by those whose weight increased over the study period.

Chapter 11 Bivariate analysis

11.1 Two-sample tests

So far we have looked at one-sample tests, where hypotheses about an unknown parameter are tested based on a single sample $X_1, X_2, \ldots, X_n$ from the distribution in question. For paired samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ we apply one-sample tests to the differences $D_i = X_i - Y_i$. We now turn our attention to the question of whether two **independent** random samples are drawn from the same distribution.

- **Welch's t-test** is a parametric test to detect a difference between the means of two independent samples.
- The **Mann-Whitney test** is a non-parametric test to detect a difference between the medians of two independent samples.

11.1.1 Welch's t-test

Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ be independent random variables, and suppose we with to test the null hypothesis $H_0 : \mu_1 = \mu_2$ against a suitable alternative. Let $X_1, X_2, \ldots, X_m$ be a random sample from the distribution of X and let $Y_1, Y_2, \ldots, Y_n$ be a random sample from the distribution of Y.

Using characteristic functions it can be shown that a linear combination of normal variables is a normal variable: if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ then

$$aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

In particular,

$$X - Y \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2).$$

Because the samples are independent,

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}\right).$$

To estimate the variances σ_1^2 and σ_2^2 we use the sample variances of X and Y:

$$S_1^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \quad \text{and} \quad S_2^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2.$$

Our test statistic is

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{S_1^2/m + S_2^2/n}}$$

- Under $H_0 : \mu_1 = \mu_2$ we have that $T \sim t_\nu$ approximately, where

$$\nu = \frac{(s_1^2/m + s_2^2/n)^2}{(s_1^2/m)^2(m-1) + (s_2^2/n)^2(n-1)} \qquad \text{(Welch-Satterthwaite equation).}$$

- An approximate $100(1-\alpha)\%$ large-sample confidence interval for the difference $\mu_1 - \mu_2$ is given by
$$(\bar{X}-\bar{Y}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{m}+\frac{S_2^2}{n}}.$$

Equal variances

Suppose that X and Y have equal variances: $X \sim N(\mu_1, \sigma^2)$ and $Y \sim N(\mu_2, \sigma^2)$ so that their distributions belong to the location model $\mathcal{M} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}\}$. In this case we define the so-called **pooled estimator** of variance,

$$S_p^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}.$$

It is easy to show that S_p^2 is an unbiased estimator of σ^2, and that

$$T = \frac{(\bar{X}-\bar{Y}) - (\mu_1-\mu_2)}{S_p\sqrt{1/m+1/n}} \sim t_{m+n-2} \quad \text{under } H_0 : \mu_1 = \mu_2.$$

Example 11.1
Let $X_1, X_2, \ldots, X_{10}$ be a random sample from the $N(\mu_1, \sigma^2)$ distribution, and let $Y_1, Y_2, \ldots, Y_7$ be an independent random sample from the $N(\mu_2, \sigma^2)$ distribution. Realisations of the samples yield the sample means $\bar{x} = 4.2$ and $\bar{y} = 3.4$ and the sample variances $s_1^2 = 49$ and $s_2^2 = 32$.

1. Test the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$.
2. Find a 90% confidence interval for the difference $\mu_1 - \mu_2$.

Solution:

11.1.2 The Mann-Whitney test

Definition 11.2
Let $X_1, X_2, \ldots, X_m$ be a random sample from a continuous and symmetric distribution with median η_1, and let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a possibly shifted version the same distribution, with median η_2. To test the null hypothesis $H_0 : \eta_1 = \eta_2$ against a suitable alternative, the **Mann-Whitney** test statistic is

$$U_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij} \qquad \text{where} \qquad Z_{ij} = \begin{cases} 1 & X_i < Y_j, \\ 0.5 & X_i = Y_j, \\ 0 & X_i > Y_j. \end{cases}$$

We also define the complementary statistic

$$U'_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n}(1 - Z_{ij}).$$

Note that $\min(U_{m,n}) = 0$, $\max(U_{m,n}) = mn$, and

$$U_{m,n} + U'_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n}\big[Z_{ij} + (1 - Z_{ij})\big] = \sum_{i=1}^{m}\sum_{j=1}^{n} 1 = mn,$$

The direction of the alternative hypothesis determines how the test statistic should be used:

- $H_1 : \eta_1 < \eta_2$: large values of U support the alternative hypothesis.
- $H_1 : \eta_1 > \eta_2$: small values of U support the alternative hypothesis.
- $H_1 : \eta_1 \neq \eta_2$: small and large values of U support the alternative hypothesis.

Example 11.3
Suppose we have the random sample $\{14, 5, 8\}$ from the distribution of X, and the random sample $\{7, 12, 18, 11\}$ from the distribution of Y. Compute the Mann-Whitney test statistic for these data.

Solution:

Example 11.4
Two different models of car, with engines of a similar size, were compared for their fuel consumption. Five cars of each model were evaluated in independent tests. The observation made was the number of miles travelled using 10 litres of petrol. Use a suitable non-parametric test to assess the claim that Model 2 is more economical than Model 1.

Model 1	125.8	126.7	128.3	130.5	126.2
Model 2	127.8	131.4	129.6	130.2	128.1

Solution:

Mean and variance of $U_{m,n}$

Theorem 11.5
Under the null hypothesis $H_0 : \eta_1 = \eta_2$,

$$\mathbb{E}(U_{m,n}) = \frac{mn}{2} \quad \text{and} \quad \mathrm{Var}(U_{m,n}) = \frac{mn}{12}(m+n+1).$$

Proof: Under $H_0 : \eta_1 = \eta_2$ we have $\mathbb{P}(X_i < Y_j) = 1/2$, so $Z_{ij} \sim \text{Bernoulli}(1/2)$.

Hence $\mathcal{E}(Z_{ij}) = 1/2$ and therefore

$$\mathbb{E}(U_{m,n}) = \mathcal{E}\left(\sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathcal{E}(Z_{ij}) = \frac{mn}{2}.$$

The calculation to find the variance of U involves some tedious algebra:

$$\begin{aligned}
\mathbb{E}(U_{m,n}^2) &= \mathbb{E}\left[\left(\sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij}\right)\left(\sum_{k=1}^{m}\sum_{\ell=1}^{n} Z_{k\ell}\right)\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{m}\sum_{\ell=1}^{n}\mathbb{E}(Z_{ij}Z_{k\ell}) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}(Z_{ij}^2) + \sum_{i=1}^{m}\sum_{\substack{k=1\\k\neq i}}^{m}\sum_{j=1}^{n}\mathbb{E}(Z_{ij}Z_{kj}) \\
&\quad + \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{\substack{\ell=1\\ \ell\neq j}}^{n}\mathbb{E}(Z_{ij}Z_{i\ell}) + \sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq i}}^{m}\sum_{\substack{\ell=1\\ \ell\neq j}}^{n}\mathbb{E}(Z_{ij}Z_{k\ell})
\end{aligned}$$

The four sums have mn, $mn(m-1)$, $mn(n-1)$ and $mn(m-1)(n-1)$ terms, respectively.

1. If $k = i$ and $\ell = j$, the summand is $Z_{ij}Z_{k\ell} = Z_{ij}^2$, and $Z_{ij}^2 = 1$ only if $X_i < Y_j$. Under the null hypothesis, this occurs with probability 1/2.
2. If $k \neq i$ but $\ell = j$, Z_{ij} and $Z_{k\ell}$ are not independent. In this case, $Z_{ij}Z_{k\ell} = 1$ if and only if both $X_i < Y_j$ and $X_k < Y_j$. There are six possible arrangements of X_i, X_k and Y_j, of which two are such that $X_i < Y_j$ and $X_k < Y_j$. Under the null hypothesis, this occurs with probability 1/3.
3. If $k = i$ but $\ell \neq j$, by a similar argument we have $Z_{ij}Z_{k\ell} = 1$ if and only if both $X_i < Y_j$ and $X_i < Y_\ell$. Under the null hypothesis, this occurs with probability 1/3.
4. If $k \neq i$ and $\ell \neq j$ then Z_{ij} and $Z_{k\ell}$ are independent. The summand $Z_{ij}Z_{k\ell} = 1$ if and only if both $X_i < Y_j$ and $X_k < Y_\ell$. Under the null hypothesis, this occurs with probability $1/2 \times 1/2 = 1/4$.

Hence,

$$\begin{aligned}
\mathbb{E}(U_{m,n}^2) &= \left[mn \times \frac{1}{2}\right] + \left[m(m-1)n \times \frac{1}{3}\right] + \left[mn(n-1) \times \frac{1}{3}\right] + \left[m(m-1)n(n-1) \times \frac{1}{4}\right] \\
&= \frac{mn}{12}(1 + n + m + 3mn),
\end{aligned}$$

and

$$\mathrm{Var}(U_{m,n}) = \mathbb{E}(U_{m,n}^2) - \mathbb{E}(U_{m,n})^2 = \frac{mn}{12}(m+n+1).$$

Normal approximation

The Mann-Whitney statistic $U_{m,n}$ is a sum of random variables (namely the Z_{ij}). By the central limit theorem, the distribution of $U_{m,n}$ is approximately normal for m and n sufficiently large.

Lower-tail test:

$$Z = \frac{(U_{m,n} + \frac{1}{2}) - \frac{mn}{2}}{\sqrt{\frac{mn}{12}(m+n+1)}} \sim N(0,1) \quad \text{approx. for } m \text{ and } n \text{ sufficiently large.}$$

Upper-tail test:

$$Z = \frac{(U_{m,n} - \frac{1}{2}) - \frac{mn}{2}}{\sqrt{\frac{mn}{12}(m+n+1)}} \sim N(0,1) \quad \text{approx. for } m \text{ and } n \text{ sufficiently large.}$$

11.1.3 Exact distribution of $U_{m,n}$ under H_0

The exact distribution of $U_{m,n}$ under H_0 can be obtained for small m, n by a recurrence relation. First we define the base case: if $m = 0$ or $n = 0$, we set $U = 0$, so

- $\mathbb{P}(U_{m,0} = 0) = 1$ and $\mathbb{P}(U_{0,n} = 0) = 1$;
- $\mathbb{P}(U_{m,0} = u) = 0$ and $\mathbb{P}(U_{0,n} = u) = 0$ for $u \neq 0$.

Theorem 11.6
Under $H_0 : \eta_1 = \eta_2$, the PMF of $U_{m,n}$ satisfies

$$\mathbb{P}(U_{m,n} = u) = \left(\frac{m}{m+n}\right)\mathbb{P}(U_{m-1,n} = u) + \left(\frac{n}{m+n}\right)\mathbb{P}(U_{m,n-1} = u - m)$$

for $u = 0, 1, \ldots mn$ (and zero otherwise).

Proof: Let m and n be fixed and assume that $m > 0$ and $n > 0$. Under the null hypothesis,

$$\begin{aligned}\mathbb{P}(\text{The largest observation is one of the } x\text{-values}) &= \frac{m}{m+n},\\ \mathbb{P}(\text{The largest observation is one of the } y\text{-values}) &= \frac{n}{m+n}.\end{aligned}$$

Let $U = u$ and suppose that one of the x-values is the largest observation.

- The remaining $(m-1)$ x-values and n y-values constitute a random sample, with one fewer x-value, for which $U = u$.

Let $U = u$ and suppose that one of the y-values is the largest observation.

- The remaining m x-values and $(n-1)$ y-values constitute a random sample, with one fewer y-value, for which $U = u - m$. (The largest y-value adds m to the value U for the complete set of m x-values and n y-values.)

Thus we have that

$$\mathbb{P}(U_{m,n} = u) = \left(\frac{m}{m+n}\right)\mathbb{P}(U_{m-1,n} = u) + \left(\frac{n}{m+n}\right)\mathbb{P}(U_{m,n-1} = u - m)$$

This recurrence relation can be used to find the PMF of U for any m and n.

Example 11.7
In an experiment on the effects of exposure to ozone, 10 rats were exposed to the gas for a period. A control group of 10 rats were kept in an ozone-free atmosphere, but otherwise in similar conditions. The lung volumes in millilitres for the two groups of rats after the conclusion of the experiment are tabulated below. Perform a test of size $\alpha = 0.05$ to determine whether there is a statistically significant difference in the average lung volumes of the two groups of rats.

Exposed (X)	9.2	8.4	8.6	9.2	9.5	9.1	9.9	9.6	9.0	9.6
Not Exposed (Y)	8.8	8.6	8.7	8.4	9.1	9.2	8.3	8.5	8.8	8.2

Solution:

Exercise 11.8

1. We wish to determine whether the distribution of population B is located to the right of population A.

Sample A	37	40	33	29	42	33	35	28	34
Sample B	65	35	47	52					

   1. State the null and alternative hypotheses for the test.
   2. Perform the hypothesis test using the Mann-Whitney test at $\alpha = 0.05$.

   Answer:

   1. $H_0 : \eta_A = \eta_B$, $H_1 : \eta_A < \eta_B$.
   2. $U = 3 + 3 + 4 + 4 + 3 + 4 + 3.5 + 4 + 4 = 32.5$.
   From tables, with $m = 9$ and $n = 4$, a one-tailed test at $\alpha = 0.05$ has critical value $U_c = 29$.
   Since $U > U_c$, we reject H_0.

2. Independent random samples are selected from two populations. The data is shown in the following table.

Sample 1	15	10	12	16	13	8		
Sample 2	5	12	9	9	8	4	5	10

   1. Use the Mann-Whitney to determine whether the data provide sufficient evidence to indicate a shift in the locations of the probability distributions of the sampled populations. Test using $\alpha = 0.05$.
   2. Do the data provide sufficient evidence to indicate that the probability distribution of the first propulation is shifted to the right of the second population? Use the Mann-Whitney test with $\alpha = 0.05$

   Answer: Recall that

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij} \quad \text{and} \quad U' = \sum_{i=1}^{m}\sum_{j=1}^{n}(1 - Z_{ij}) \quad \text{where} \quad Z_{ij} = \begin{cases} 1 & X_i < Y_j, \\ 0.5 & X_i = Y_j, \\ 0 & X_i > Y_j. \end{cases}$$

From the table,

$$\begin{aligned} U &= 0 + 1.5 + 0.5 + 0 + 0 + 4.5 = 6.5, \\ U' &= 6 + 3.5 + 5 + 5 + 5.5 + 6 + 6 + 4.5 = 41.5. \end{aligned}$$

1. $H_0 : \eta_1 = \eta_2$ and $H_1 : \eta_1 \neq \eta_2$.
   From tables, for a two-tailed test at $\alpha = 0.05$ with $m = 6$ and $n = 8$, the upper-tail critical value is $U_c = 40$ and the lower-tail critical value is $mn - U_c = 48 - 40 = 8$. The rejection region is therefore $\{U : U \leq 8 \text{ or } U \geq 40\}$. Since the observed value of U lies in the critical region, we reject H_0, and conclude that the medians of the two populations are different.
2. $H_0 : \eta_1 = \eta_2$ and $H_1 : \eta_1 > \eta_2$.
   Only small values of U support the alternative hypothesis $H_1 : \eta_1 > \eta_2$.
   From tables, for a one-tailed test at $\alpha = 0.05$ with $m = 6$ and $n = 8$, the upper-tail critical value is $U_c = 37$, so the lower-tail critical value is $mn - U_c = 48 - 37 = 11$. The rejection region is therefore $\{U : U \leq 11\}$. Since the observed value of U lies in the critical region, we reject H_0, and conclude that population 1 lies to the right of population 2.

3. The percentage of carbon in iron samples taken from two different furnaces was measured. The results obtained are as follows

Furnace 1	2.28	2.34	2.37	2.39	2.40	2.41
Furnace 2	2.36	2.40	2.42	2.44	2.44	2.48

Perform a suitable non parametric test to determine if there is a significant difference between the median percentage carbon in the two furnaces. Discuss how you have dealt with ties and the effect that these might have had on your conclusions.

Answer: We assume that the furnaces are independent of each other, and use the Mann-Whintey test.

- Using the counting method: $U = 6 + 6 + 5 + 5 + 4.5 + 4 = 30.5$
- Alternatively, using rank sums:

Furnace 1	2.28	2.34	2.37	2.39	2.40	2.41	
Rank	1	2	4	5	6.5	8	$R_X = 26.5$

Furnace 2	2.36	2.40	2.42	2.44	2.44	2.48	
Rank	3	6.5	9	10.5	10.5	12	$R_Y = 51.5$

Hence $U = R_Y - \frac{1}{2}n(n+1) = 51.5 - 21 = 30.5$.

From tables, the critical value for a 2-tailed test at significance level $\alpha = 0.05$ is $U_c = 31$, and at $\alpha = 0.1$ is $U_c = 29$. Thus the test statistic is outside the critical region $\{U : U \geq 31\}$ at $\alpha = 0.05$, but inside the critical region $\{U : U \geq 29\}$ at the $\alpha = 0.1$ significance level. Thus we would retain the null hypothesis at $\alpha = 0.05$, but it is a close decision.

When computing the test statistic, ties were handled by assigning the average of the ranks that would have been assigned had there not been any ties. If the two observations recorded as 2.40 had been recorded more accurately, and the value for Furnace 2 had turned out to be bigger than the value for Furnace 1, the Mann-Whitney statistic would be $U = 31$. This would have led us to reject the null hypothesis at $\alpha = 0.05$, which illustrates that the decision is indeed a close call.

11.2 Sums of squares

Let $X \sim N(\mu, \sigma^2)$ where μ is unknown, and let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution of X. The usual test statistic for deciding between $H_0 : \mu = \mu_0$ and a suitable alternative is the standardised sum

$$Z = \sum_{i=1}^{n} \left(\frac{X_i - \mu_0}{\sigma} \right) \sim N(0,1) \text{ under } H_0.$$

Another test statistic is provided the standardized **sum-of-squares**,

$$T = \sum_{i=1}^{n} \left(\frac{X_i - \mu_0}{\sigma} \right)^2 \sim \chi^2_n \text{ under } H_0.$$

where χ^2_n is the **chi-squared distribution** with n degrees of freedom.

Remark 11.9
If σ^2 is unknown we replace it by the sample variance s^2, in which case $T \sim \chi^2_{n-1}$.

11.2.1 The χ^2 distribution

Definition 11.10
The χ^2_n distribution is defined by the PDF

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(n/2)2^{n/2}} \, x^{n/2-1} e^{-x/2} & \text{for } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where the parameter n is called the **degrees of freedom**.

The χ^2_n distribution is a special case of the $\Gamma(k, \theta)$ distribution, where $k = n/2$ and $\theta = 2$ is a scale parameter. In particular, $\mathbb{E}(X) = n$ and $\text{Var}(X) = 2n$. The following theorem (which we shall not prove) asserts that the sum-of-squares of n independent standard normal variables has the χ^2_n-distribution.

Theorem 11.11
If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal variables then $\displaystyle\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$.

Example 11.12
A quality control supervisor at a paint factory knows that the exact amount each tin contains will vary due to certain uncontrollable factors that affect the amount of fill. The mean fill is important, but equally important is the variation of each fill. If the variance σ^2 of the fill is large, some tins will contain too much paint, and others too little. A regulatory agency specifies that the variance of the amount of fill in $250ml$ tins should be less than $3ml$. To determine whether or not the process is meeting this specification, the supervisor randomly selects 10 tins and measures the contents of each tin. The mean fill over the sample is found to be $250.78ml$, and the sample variance is $s^2 = 1.03$. Do the data indicate that the factory is operating within the regulatory limits?

Solution:

11.2.2 The F distribution

Definition 11.13
Let T_1 and T_2 be independent random variables with $T_1 \sim \chi^2_m$ and $T_2 \sim \chi^2_n$. The distribution of the ratio

$$F = \frac{T_1/m}{T_2/n}.$$

is called the **F-distribution with m and n degrees of freedom**, and denoted by $F \sim F_{m,n}$.

Example 11.14
A researcher wants to compare the metabolic rates of mice subjected to different drugs. The weight of the mice may affect their metabolic rates, so the researcher wishes to obtain mice that are relatively homogeneous with respect to weight. Five hundred mice will be needed to complete the study. Currently, 16 mice from supplier 1 and another 13 mice from supplier 2 are available for comparison. The researcher weighs these mice and finds that the sample standard deviations are $s_1 = 0.2021$ and $s_2 = 0.0982$ respectively. Is there sufficient evidence to indicate a significant

difference in the variance of the weight of mice obtained from the two suppliers at the $\alpha = 0.1$ level?

Solution:

11.2.3 The non-central χ^2 distribution

Definition 11.15
Let $X_1, X_2, \dots, X_n$ be independent random variables with $X_i \sim N(\mu_i, 1)$. The distribution of the sum-of-squares

$$W = \sum_{i=1}^{n} X_i^2$$

is called the **non-central chi-squared distribution**, with n degrees of freedom and non-centrality parameter

$$\lambda = \sum_{i=1}^{n} \mu_i^2.$$

We write this as $W \sim \chi_n^2(\lambda)$, in which case

$$\mathbb{E}(W) = n + \lambda \quad \text{and} \quad \operatorname{Var}(W) = 2(n + 2\lambda).$$

When $\lambda = 0$, all μ_i must be zero and the $\chi_n^2(\lambda)$ distribution reduces to the ordinary χ_n^2 distribution. Any non-zero mean μ_i increases the value of λ and hence increases $\mathbb{E}(W)$ and $\operatorname{Var}(W)$ compared to those of the ordinary χ_n^2 distribution.

11.3 Analysis of variance

Analysis of variance is a method for testing hypotheses about means by looking at sample variances.

Let Y be a continuous variable, let $X \in \{1, 2, \ldots, k\}$ be a simple random variable representing **group membership**, and consider the location model

$$Y = \mathbb{E}(Y|X = i) + \epsilon \qquad \text{where } \epsilon \sim N(0, \sigma^2).$$

Let $\mu = \mathbb{E}(Y)$ and $\mu_i = \mathbb{E}(Y|X = i)$. We wish to test the null hypothesis that the conditional means μ_i are all equal against the alternative hypothesis that they are not.

$$\begin{aligned}&H_0 : \mu_1 = \mu_2 = \ldots = \mu_k\\ &H_1 : \mu_i \neq \mu_j \text{ for some } i \neq j.\end{aligned}$$

11.3.1 Partition of variance

Given that $X = i$ we have $Y \sim N(\mu_i, \sigma^2)$ so

$$\mathbb{E}(Y|X = i) = \mu_i \quad \text{and} \quad \mathrm{Var}(Y|X = i) = \sigma^2.$$

Hence

$$\mathrm{Var}\big[\mathbb{E}(Y|X)\big] = \sum_{i=1}^{k}(\mu_i - \mu)^2\mathbb{P}(X = i) \quad \text{and} \quad \mathbb{E}[\mathrm{Var}(Y|X)\big] = \sigma^2.$$

By the law of total variance,

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}\big[\mathrm{Var}(Y|X)\big] + \mathrm{Var}\big[\mathbb{E}(Y|X)\big]\\ &= \sigma^2 + \sum_{i=1}^{k}(\mu_i - \mu)^2\mathbb{P}(X = i).\end{aligned}$$

Thus we have divided the variance of Y into two components,

- an **explained** component $\sum_{i=1}^{k}(\mu_i - \mu)^2\mathbb{P}(X = i)$ due to variation **between** groups;
- an **unexplained** component σ^2 due to variation **within** groups.

11.3.2 Test statistics

Suppose we obtain independent random samples from each group.

- $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ where $Y_{1j} \sim N(\mu_1, \sigma^2)$,
- $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ where $Y_{2j} \sim N(\mu_2, \sigma^2)$,

  $\ldots$
- $Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$ where $Y_{kj} \sim N(\mu_k, \sigma^2)$.

Let $N = \sum_{i=1}^{k} n_i$ be the total number of observations. We estimate the overall mean μ and the individual group means μ_i by the sample means

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij} \quad \text{and} \quad \hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad \text{respectively.}$$

Consider the sum of the squared differences between the Y_{ij} and the overall sample mean $\hat{\mu}$,

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i+\hat{\mu}_i-\hat{\mu})^2 \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left[(Y_{ij}-\hat{\mu}_i)^2+2(Y_{ij}-\hat{\mu}_i)(\hat{\mu}_i-\hat{\mu})+(\hat{\mu}_i-\hat{\mu})^2\right] \\ &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i)^2+\sum_{i=1}^{k}n_i(\hat{\mu}_i-\hat{\mu})^2\end{aligned}$$

We write this as $SST = SSE + SSG$ where

$$\begin{aligned}SST &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu})^2 && \text{is the } \textbf{total sum-of-squares,} \\ SSE &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i)^2 && \text{is the } \textbf{error sum-of-squares} \text{ (due to variation within groups),} \\ SSG &= \sum_{i=1}^{k}n_i(\hat{\mu}_i-\hat{\mu})^2 && \text{is the } \textbf{groups sum-of-squares} \text{ (due to variation between groups).}\end{aligned}$$

Lemma 11.16
The error sum-of-squares and groups sum-of-squares both have χ^2-distribution:

$$\begin{aligned}&\frac{1}{\sigma}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i)^2 && \sim \chi^2_{N-k} \\ &\frac{1}{\sigma}\sum_{i=1}^{k}n_i(\hat{\mu}_i-\hat{\mu})^2 && \sim \chi^2_{k-1}(\lambda) \text{ where } \lambda=\sum_{i=1}^{k}n_i(\mu_i-\mu)^2.\end{aligned}$$

Proof:

1. By independence, because $Y_{ij} \sim N(\mu_i, \sigma^2)$ we have

$$\sum_{j=1}^{n_i}\left(\frac{Y_{ij}-\mu_i}{\sigma}\right)^2 \sim \chi^2_{n_i}.$$

   Replacing the unknown expectation μ_i by the sample mean $\hat{\mu}_i$, we obtain

$$\frac{1}{\sigma^2}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_i)^2 \sim \chi^2_{n_i-1}.$$

   If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ then $U+V \sim \chi^2_{m+n}$, so

$$\frac{1}{\sigma^2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu})^2 \sim \chi^2_{N-k}.$$

2. Because $\hat{\mu}_i$ is a sample mean, we have $\hat{\mu}_i \sim N(\mu_i, \sigma^2/n_i)$, so

$$\sqrt{n_i}\left(\frac{\hat{\mu}_i-\mu}{\sigma}\right) \sim N(\mu_i-\mu, 1)$$

Because all observations are independent, the sample means $\hat{\mu}_i$ are also independent so

$$\sum_{i=1}^{k} n_i \left(\frac{\bar{Y}_i - \mu}{\sigma}\right)^2 \sim \chi^2_k(\lambda) \quad \text{where} \quad \lambda = \sum_{i=1}^{k} (\mu_i - \mu)^2.$$

Finally, replacing the unknown expectation μ by the sample mean $\hat{\mu}$, we obtain

$$\frac{1}{\sigma^2} \sum_{i=1}^{k} n_i (\hat{\mu}_i - \hat{\mu})^2 \sim \chi^2_{k-1}(\lambda).$$

Theorem 11.17 (Test Statistic for ANOVA)
Let $F = s_G^2 / s_E^2$ where

$$s_G^2 = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\hat{\mu}_i - \hat{\mu})^2 \quad \text{and} \quad s_E^2 = \frac{1}{N-k} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^2.$$

Then $F \sim F_{k-1,N-k}$ under $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$.

Remark 11.18
Under the alternative hypothesis we have $\lambda > 0$, in which case the s_G^2 is likely to be larger than it would be under the null hypothesis. Thus we require an upper-tail test: H_0 is rejected whenever $F > F_\alpha$ where F_α is the upper-tail critical value of the $F_{k-1,N-k}$ distribution at significance level α.

The various statistics computed during a one-way analysis of variance are usually reported in tabular form:

Source	df	SS	MS	F
Groups	$k-1$	SSG	s_G^2	$F = s_G^2/s_E^2$
Error	$N-k$	SSE	s_E^2	
Total	$N-1$	SST		

Example 11.19
The data below are the yields (per hectare) of eight types of wheat, recorded over four independent trials.

Type	Yield			
1	182	214	216	231
2	196	202	208	224
3	203	212	221	242
4	198	203	207	222
5	171	192	197	204
6	194	218	223	232
7	208	216	218	239
8	183	188	193	198

Perform a one-way analysis-of-variance to determine whether there are significant differences among the mean yields of the eight types.

Solution:

11.4 Simple linear regression

Let X and Y be continuous random variables. We wish to investigate how X influences the behaviour of Y.

- X is called the **explanatory variable** or the **independent variable**.
- Y is called the **response variable** or the **dependent variable**.

Suppose we observe that X takes the value x. Unless Y is completely determined by X, we cannot predict its value with certainty so instead we focus on the problem of estimating its conditional expectation $E(Y|X=x)$. This leads us to represent Y as the sum of two random variables:

$$Y = \mu(X) + \epsilon \qquad \text{where } \epsilon \sim N(0,\sigma^2).$$

- $\mu(x) = \mathbb{E}(Y|X=x)$ is called the **regression function**;
- $\epsilon = Y - \mathbb{E}(Y|X)$ is called the **error variable**.

Lemma 11.20
$\text{Var}(Y) = \text{Var}\big[\mu(X)\big] + \text{Var}(\epsilon)$.

Proof: $\epsilon = Y - \mathbb{E}(Y|X)$, so by the definition of conditional variance,

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

By the law of total variance,

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}\big[\mathbb{E}(Y|X)\big] + \mathbb{E}\big[\mathrm{Var}(Y|X)\big] \\ &= \mathrm{Var}\big[\mu(X)\big] + \mathbb{E}\big[\mathrm{Var}(\epsilon)\big] \\ &= \mathrm{Var}\big[\mu(X)\big] + \mathrm{Var}(\epsilon).\end{aligned}$$

Lemma 11.20 expresses the variance of Y as the sum of an **explained variance**, attributed to the variance of the explanatory variable X, and an **unexplained variance** attributed to the error variable ϵ.

11.4.1 Linear models

Definition 11.21
A **linear model** is a model which is linear in its parameters.

- $\mu(x) = \alpha + \beta x + \gamma x^2$ is a linear model.
- $\mu(x) = \alpha e^{\beta x}$ is not a linear model.

A **simple linear model** is a linear model of the form $\mu(x) = \alpha + \beta x$.

The simple linear regression model is

$$Y = \alpha + \beta X + \epsilon \quad \text{where} \quad \epsilon \sim N(0, \sigma^2).$$

Given that $X = x$, we have that $Y \sim N(\alpha + \beta x, \sigma^2)$ and in particular,

$$\mathbb{E}(Y|X = x) = \alpha + \beta x \quad \text{and} \quad \mathrm{Var}(Y|X = x) = \sigma^2.$$

11.4.2 Test statistics

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample from the joint distribution of X and Y. In Week 8 we saw that the maximum likelihood estimators of α, β and σ^2 are respectively

$$\hat\alpha = \bar Y - \hat\beta \bar X, \qquad \hat\beta = \frac{\sum_{i=1}^n (X_i - \bar X)(Y_i - \bar Y)}{\sum_{i=1}^n (X_i - \bar X)^2} \qquad \text{and} \qquad \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n \hat\epsilon_i^2$$

where $\hat\epsilon_i = Y_i - (\hat\alpha + \hat\beta X_i)$ is the so-called **residual variable** at X_i.

Let $x_1, x_2, \ldots, x_n$ be a fixed realisation of the marginal sample $X_1, X_2, \ldots, X_n$. Then $\hat\alpha$, $\hat\beta$ and $\hat\epsilon_i$ are all linear combinations of $Y_1, Y_2, \ldots, Y_n$. Because the Y_i are independent normal variables it thus follows that $\hat\alpha$, $\hat\beta$ and $\hat\epsilon_i$ are also normal variables.

Theorem 11.22
The MLEs of α, β and σ^2 respectively satisfy

$$\hat\alpha \sim N\left(\alpha, \frac{\sigma^2}{n}\right), \quad \hat\beta \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar x)^2}\right) \quad \text{and} \quad \frac{n\hat\sigma}{\sigma} \sim \chi^2_{n-2}.$$

Proof:

1. The expected value of $\hat{\alpha}$ is

$$\begin{aligned}\mathbb{E}(\hat{\alpha}) = \mathbb{E}(\bar{Y} - \beta\bar{x}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\beta}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) - \frac{\beta}{n}\sum_{i=1}^{n} x_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i) - \frac{\beta}{n}\sum_{i=1}^{n} x_i \\ &= \alpha.\end{aligned}$$

Because $\mathrm{Var}(Y_i) = \sigma^2$, the variance of $\hat{\alpha}$ is

$$\begin{aligned}\mathrm{Var}(\hat{\alpha}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\beta}{n}\sum_{i=1}^{n} x_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(Y_i) = \frac{\sigma^2}{n}.\end{aligned}$$

Hence $\hat{\alpha} \sim N(\alpha, \sigma^2/n)$, as required.

2. Since $\mathbb{E}(Y_i) = \alpha + \beta x_i$ and $\mathbb{E}(\bar{Y}) = \alpha + \beta\bar{x}$, the expected value of $\hat{\beta}$ is therefore

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= \mathbb{E}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}\beta(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta.\end{aligned}$$

Using the fact that $\sum_{i=1}^{n} x_i = n\bar{x}$, it is easy to see that $\hat{\beta}$ can be rewritten as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Because $\mathrm{Var}(Y_i) = \sigma^2$, the variance of $\hat{\beta}$ is

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) = \mathrm{Var}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] &= \frac{1}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\mathrm{Var}(Y_i) \\ &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\end{aligned}$$

Hence $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$, as required.

3. Recall that

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon_i}^2 \quad \text{where} \quad \hat{\epsilon}_i = Y_i - (\hat{\alpha} + \hat{\beta}x_i).$$

Consider

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{n}\big[Y_i - (\alpha + \beta x_i)\big]^2 &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\big[(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)x_i + (Y_i - (\hat{\alpha} + \hat{\beta}x_i))\big]^2 \\ &= \frac{n(\hat{\alpha} - \alpha)^2}{\sigma^2} + \frac{(\hat{\beta} - \beta)^2}{\sigma^2}\sum_{i=1}^{n} x_i^2 + \frac{n\hat{\sigma}^2}{\sigma^2}.\end{aligned}$$

The first three terms in this expression all have chi-squared distribution.

- Because $Y_i \sim N(\alpha+\beta x_i, \sigma^2)$ it follows that $\big[Y_i - (\alpha+\beta x_i)\big]/\sigma \sim N(0,1)$, so
$$\frac{1}{\sigma^2}\sum_{i=1}^n \big[Y_i - (\alpha+\beta x_i)\big]^2 \sim \chi^2_n.$$
- Because $\hat{\alpha} \sim N(\alpha, \sigma^2/n)$ it follows that $\sqrt{n}(\hat{\alpha}-\alpha)/\sigma \sim N(0,1)$, so
$$\frac{n(\hat{\alpha}-\alpha)^2}{\sigma^2} \sim \chi^2_1.$$
- Because $\hat{\beta} \sim N\left(\beta, \sigma^2/\sum_{i=1}^n (x_i-\bar{x})^2\right)$ it follows that $(\hat{\beta}-\beta)\sqrt{\sum_{i=1}^n x_i^2}/\sigma \sim N(0,1)$, so
$$\frac{(\hat{\beta}-\beta)^2}{\sigma^2}\sum_{i=1}^n x_i^2 \sim \chi^2_1.$$

It is easy to see that if $U \sim \chi^2_a$ and $V \sim \chi^2_b$ are independent, then $U+V \sim \chi^2_{a+b}$. It thus follows that $n\hat{\sigma}/\sigma \sim \chi^2_{n-2}$ as required.

We estimate the error variance using the following unbiased estimator (instead of the MLE),

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{\epsilon}_i^2 \quad \text{where} \quad \hat{\epsilon}_i = Y_i - (\hat{\alpha}+\hat{\beta}x_i).$$

This estimator for σ^2 yields the following test statistics:

$$T_1 = \frac{\hat{\alpha}-\alpha}{\sqrt{\hat{\sigma}^2/(n-2)}} \quad \text{and} \quad T_2 = \frac{\hat{\beta}-\beta}{\sqrt{n\hat{\sigma}^2/[(n-2)\sum_{i=1}^n (x_i-\bar{x})^2]}}.$$

- Under the null hypothesis $H_0 : \alpha = 0$,
$$T_1 = \frac{\hat{\alpha}-\alpha}{\sqrt{\hat{\sigma}^2/(n-2)}} \sim t_{n-2}.$$
- Uhder the null hypothesis $H_0 : \beta = 0$,
$$T_2 = \frac{\hat{\beta}-\beta}{\sqrt{n\hat{\sigma}^2/[(n-2)\sum_{i=1}^n (x_i-\bar{x})^2]}} \sim t_{n-2}.$$

T_2 can be used to test whether or not Y depends (linearly) on X:

$$\begin{aligned} H_0 &: Y = \alpha + \epsilon, \\ H_1 &: Y = \alpha + \beta X + \epsilon. \end{aligned}$$

11.4.3 ANOVA for regression

The **predicted value** of Y_i is the random variable

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i.$$

The total deviation of Y_i from the overall mean $\bar{Y}$ can be divided into two components,

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}),$$

from which it follows that

$$\sum_{i=1}^n (Y_i-\bar{Y})^2 = \sum_{i=1}^n (Y_i-\hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i-\bar{Y})^2.$$

As with ANOVA, we define the following sums-of-squares.

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 && \text{The \textbf{total} sum-of-squares.}\\
SSR &= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 && \text{The \textbf{regression} sum-of-squares.}\\
SSE &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 && \text{The \textbf{error} sum-of-squares.}
\end{aligned}
$$

The total sum-of-squares SST is determined by the marginal sample $(Y_1, Y_2, \dots, Y_n)$. Substituting for $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ we see that the regression sum-of-squares satisfies

$$SSR = \frac{\left[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The error sum-of-squares is then obtained via $SSE = SST - SSR$.

Under $H_0 : \beta = 0$,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \sim \chi^2_1 \quad \text{and} \quad \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \sim \chi^2_{n-2}.$$

Thus we have the test statistic

$$F = \frac{SSR}{(n-2)^{-1}SSE} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{(n-2)^{-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \sim F_{1,n-2} \quad \text{under } H_0 : \beta = 0.$$

which provides another means of testing whether or not Y depends linearly on X.

11.4.4 The coefficient of determination

Recall that the **correlation coefficient** of any pair of random variables X and Y is

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big]}{\sqrt{\mathbb{E}\big[(X - \mathbb{E}X)^2\big]\mathbb{E}\big[(Y - \mathbb{E}Y)^2\big]}}$$

For a bivariate random sample $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ the **sample correlation coefficient**, also known as the **Pearson correlation** is defined by

$$R = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

For the simple linear regression model $Y = \alpha + \beta X + \epsilon$, the MLE of β can be written as,

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = R\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}.$$

The **coefficient of determination** is the square of the sample correlation, and denoted by R^2:

$$R^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{SSR}{SST}.$$

where SSR and SST are the regression sum-of-squares and total sum-of-squares respectively.

Thus R^2 is the proportion of the total variation explained by the regression model: it quantifies how well the regression line fits the data points, and as such is an example of a **goodness-of-fit** statistic (the value $R^2 = 1$ indicates that the regression line perfectly fits the data). The corresponding quantity for one-way ANOVA is the so-called **eta-squared** effect size:

$$\eta^2 = \frac{SSG}{SST}.$$

Example 11.23
The table below shows the deaths due to bronchitis (x) and corresponding daily temperatures (y), averaged over a long period.

x	253	232	210	200	191	187	134	102	81	25
y	35	37	39	41	43	45	47	49	51	53

1. Use the simple linear model $y = \alpha + \beta x + \epsilon$ to perform a least-squares regression of y against x.
2. Test whether the slope of the regression line is significantly different from zero at the 5% significance level.

Solution:

Chapter A Random processes

Let Ω be the sample space of some random experiment and let $T \subseteq \mathbb{R}$ be a set of **times**. The canonical examples are $T = \{0, 1, 2, \ldots\}$ which defines a discrete-time process, and $T = [0, \infty)$ which defines a continuous-time process.

Definition 1.1
A **random process** on Ω is a collection of random variables $\{X_t : t \in T\}$, where each $X_t : \Omega \to \mathbb{R}$ is a random variable on Ω.

1. If T is countable, $\{X_t\}$ is called a **discrete-time** random process.
2. If T is uncountable, $\{X_t\}$ is called a **continuous-time** random process.

We can think of a random process $\{X_t\}$ as a mapping:

$$\begin{array}{rccc} \{X_t\} : & T \times \Omega & \to & \mathbb{R} \\ & (t, \omega) & \mapsto & X_t(\omega). \end{array}$$

Definition 1.2
For a fixed outcome $\omega \in \Omega$, the associated realisation $\{X_t(\omega) : t \in T\}$ of the random process $\{X_t : t \in T\}$ is called a **trajectory** or **sample path** of the process.

If T is a finite set then $\{X_t\}$ is random vector which is defined by its joint CDF. If T is an infinite set (either countable or uncountable) it is not easy to define a CDF to describe $\{X_t\}$. To do it, we have to deal with the joint distributions of $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ for all $n \in \mathbb{N}$ and all choices of $t_1, t_2, \ldots, t_n \in T$. These are called the **finite-dimensional distributions** of the random process.

A.0.1 The canonical probability space

Let $\Omega = [0, 1]^{\mathbb{N}}$ be the set of infinite sequences of real numbers from the unit interval $[0, 1]$, let $\omega \in \Omega$ and write this as

$$\omega = (\omega_0, \omega_1, \omega_2, \ldots).$$

The nth term of the sequence is extracted by the random variable

$$\begin{array}{rccc} \gamma_n : & \omega & \to & [0, 1] \\ & \omega & \mapsto & \omega_n. \end{array}$$

We state the following theorem without proof.

Theorem 1.3
There exists a probability measure $\mathbb{P}$ on Ω) such that the random variables $\gamma_0, \gamma_1, \ldots$ are independent and uniformly distributed on $[0, 1]$.

We think of the single outcome $\omega = (\omega_0, \omega_1, \omega_2, \ldots)$ as the source of all randomness in our experiment, and all other quantities are deterministic functions of this outcome (random variables).

A.0.2 The Bernoulli process

Definition 1.4
A random process $\{X_n\}$ consisting of independent and identically distributed Bernoulli(p) variables is called the Bernoulli(p) process.

In terms of the canonical probability space,

$$X_n = \begin{cases} 1 & \gamma_n \leq p, \\ 0 & \text{otherwise.} \end{cases}$$

Trajectories of the Bernoulli process are binary sequences (of infinite length).

Example 1.5
Let $\{X_n\}$ be a Bernoulli(p) process. If $p > 0$, show that a 'success' is eventually observed with probability 1.

Solution:

Example 1.6
Suppose have observed the first n terms of a Bernoulli(p) process. Let τ be the time until we next observe a 'success'. Show that $\tau \sim \text{Geometric}(p)$.

Solution:

Exercise 1.7
Show that the geometric distribution has the so-called memoryless property: if $\tau \sim \text{Geometric}(p)$, then

$$\mathbb{P}(\tau > m+n|\tau > n) = \mathbb{P}(\tau > m).$$

Answer:

$$\begin{aligned}\mathbb{P}(\tau > m+n|\tau > n) &= \mathbb{P}(\tau > m+n, \tau > n)/\mathbb{P}(\tau > n)\\ &= \mathbb{P}(\tau > m+n)/\mathbb{P}(\tau > n)\\ &= (1-p)^{m+n}/(1-p)^n\\ &= (1-p)^m\\ &= \mathbb{P}(\tau > m).\end{aligned}$$

A.1 Random walks

The **simple random walk** is the simplest model of a **diffusion process**. We consider a particle which inhabits the set of integer points $\mathbb{Z}$ and at each (discrete) time step the particle moves either one step to the right with probability p or one step to the left with probability $q = 1 - p$.

- For fixed ω the random variable X_n describes the trajectory of the particle over time.
- For fixed n the random variable X_n describes the (spatial) distribution of the particle at time n.

Definition 1.8
A discrete random process $\{X_n\}$ is called a **simple random walk** with parameter $p \in (0, 1)$ if

1. $X_{n+1} - X_n$ is independent of $X_0, X_1, \dots, X_n$,
2. $\mathbb{P}(X_{n+1} = X_n + 1) = p$ and $\mathbb{P}(X_{n+1} = X_n - 1) = q$ where $q = 1 - p$.

If $p = 1/2$ the random walk is called **symmetric**.

In terms of the random variables γ_n defined on the canonical probability space let us define a new sequence of random variables,

$$\xi_n = \begin{cases} 1 & \gamma_n \leq p, \\ -1 & \text{otherwise.} \end{cases}$$

Lemma 1.9
The random process $\{X_n\}$ where $X_n = X_0 + \sum_{k=1}^n \xi_k$ is a simple random walk.

Proof:

1. The ξ_n are independent (they are transforms of the γ_n, which are independent), so the increment $X_{n+1} - X_n$ is independent of $\xi_1, \xi_2, \dots, \xi_n$, and because $X_0, X_1, \dots, X_n$ are just linear combinations of $\xi_1, \xi_2, \dots, \xi_n$, it follows that $X_{n+1} - X_n$ is independent of $X_0, X_1, \dots, X_n$.
2. Because $\gamma_{n+1} \sim \text{Uniform}[0, 1]$,

   $$\mathbb{P}(X_{n+1} = X_n + 1) = \mathbb{P}(\xi_{n+1} = 1)\mathbb{P}(\gamma_{n+1} \leq p) = p.\mathbb{P}(X_{n+1} = X_n - 1) \quad = \mathbb{P}(\xi_{n+1} = -1)\mathbb{P}(\gamma_{n+1} > p) = 1 - p.$$

A.1.1 Properties

Definition 1.10
A **stationary** random process is one whose behaviour does not change when shifted in time and space.

Lemma 1.11
The simple random walk satisfies

$$\begin{array}{lcll}\mathbb{P}(X_n = x \,|\, X_0 = a) & = & \mathbb{P}(X_n = x + b \,|\, X_0 = a + b) & \text{(spatial homogeneity), and} \\ \mathbb{P}(X_n = x \,|\, X_0 = a) & = & \mathbb{P}(X_{m+n} = x \,|\, X_m = a) & \text{(temporal homgeneity)}\end{array}$$

and is therefore a stationary process.

Proof:

$$\begin{aligned}\mathbb{P}(X_n = x|X_0 = a) &= \mathbb{P}\left(a + \textstyle\sum_{k=1}^n \xi_k = x\right) \\ &= \mathbb{P}\left((a+b)\textstyle\sum_{k=1}^n \xi_k = x + b\right) = \mathbb{P}(X_n = x + b|X_0 = a + b) \\ \mathbb{P}(X_n = x|X_0 = a) &= \mathbb{P}\left(a + \textstyle\sum_{k=1}^n \xi_k = x\right) \\ &= \mathbb{P}\left(a + \textstyle\sum_{k=m+1}^{m+n} \xi_k = x\right) = \mathbb{P}(X_{m+n} = x|X_m = a).\end{aligned}$$

A.1.2 Sample paths

The motion of a particle can be represented by the sequence $\{(n, X_n) : n = 0, 1, 2, \ldots\}$., which is called the **trajectory** or **path** of the particle. By convention, sample paths are plotted with time on the horizontal axis, and displacement on the vertical axis.

Any event can be expressed in terms of an appropriate set of paths: the probability of the event is the probability that one of the associated paths is realized. The set of sample paths therefore serves as a **sample space** for analysing the simple random walk.

Let C_n be the set of all possible paths of length n,

$$C_n = \big\{(x_0, x_1, \ldots, x_n) \in \mathbb{Z}^{n+1} : x_{k+1} - x_k = \pm 1 \text{ for } k = 0, 1, 2, \ldots, n-1\big\}.$$

The probability that the first n steps of the random walk follows any particular path
$\mathbf{x} = (x_0, x_1, \ldots, x_n)$ is $p^c q^d$ where

- c is the number of positive steps (right/up), and
- d is the number of negative steps (left/down).

The distribution of the particle at time n

We can compute the PMF of X_n by examining the ensemble of all possible paths over the first n steps. Let $C_n(a, b)$ be the set of paths from $(0, a)$ to (n, b),

$$C_n(a, b) = \big\{\mathbf{x} \in C_n : x_0 = a, x_n = b\big\}.$$

Lemma 1.12
The number of possible paths from $(0, a)$ to (n, b) is

$$|C_n(a, b)| = \binom{n}{\frac{1}{2}(n + b - a)}$$

provided that $(n + b - a)/2$ belongs to the set $\{0, 1, 2 \ldots, n\}$.

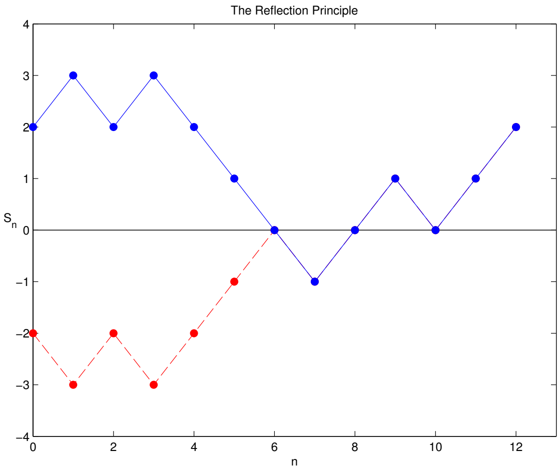

Figure A.1: The Reflection Principle

If $X_0 = 0$, the condition says that the particle can occupy only odd-numbered sites after an odd number of steps, and only even-numbered sites after an even number of steps.)

Proof: Choose a path from $(0, a)$ to (n, b). Let u be the number of positive steps, and d the number of negative steps.

- $u + d = n$ is the total number of steps;
- $u - d = b - a$ is the overall displacement to the right.

Solving for c and d, we get $c = \frac{1}{2}(n+b-a)$ and $d = \frac{1}{2}(n-b+a)$. The number of paths from $(0, a)$ to (n, b) is the number of ways of choosing exactly u positive steps from the n available steps, so

$$|C_n(a,b)| = \binom{n}{u} = \binom{n}{\frac{1}{2}(n+b-a)}$$

Corollary 1.13

$$\mathbb{P}(X_n = b | X_0 = a) = \binom{n}{\frac{1}{2}(n+b-a)} p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)}$$

Proof: Each path in $C_n(a, b)$ occurs with probability $p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)}$, so

$$\mathbb{P}(X_n = b | X_0 = a) = \binom{n}{\frac{1}{2}(n+b-a)} p^{\frac{1}{2}(n+b-a)} q^{\frac{1}{2}(n-b+a)}$$

A.1.3 The Reflection Principle

Counting sample paths is made easier by using the **reflection principle**.

Theorem 1.14
Let $a, b > 0$, and let $C_n^0(a, b)$ be the set of paths from $(0, a)$ to (n, b) which visit the (spatial) origin:

$$C_n^0(a,b) = \{\mathbf{x} \in C_n(a,b) : x_k = a \text{ for some } k = 1, 2, \ldots, n-1\}.$$

The number of such paths is equal to the total number of paths from $(0, -a)$ to (n, b),

$$|C_n^0(a,b)| = |C_n(-a,b)|.$$

Proof: Each path from $(0,a)$ to (n,b) intersects the horizontal axis at some earliest point $(k,0)$. If we reflect the segment of the path with $0 \leq t \leq k$ in the horizontal axis, we obtain a path from $(0,a)$ to (n,b) which intersects the horizontal axis. This operation puts the elements of $C_n^0(a,b)$ and $C_n(-a,b)$ in one-to-one correspondence, which proves the theorem.

Theorem 1.15 (The ballot theorem)
Let $b > a \geq 0$, and let $C_n^*(a,b)$ to be the set of paths from $(0,a)$ to (n,b) which do **not** revisit the initial location $X_0 = a$,

$$C_n^*(a,b) = \big\{\mathbf{x} \in C_n(a,b) : x_k = a \text{ for some } k = 1,2,\ldots,n-1\big\}.$$

The number of such paths is given by

$$|C_n^*(a,b)| = \frac{b-a}{n}|C_n(a,b)|,$$

where $C_n(a,b)$ is the set of all paths from $(0,a)$ to (n,b).

Proof: Without loss of generality, let $a = 0$. The first step of all such paths must be to $(1,1)$, so by the reflection principle the number of such paths is

$$\begin{aligned} |C_n^*(0,b)| &= |C_{n-1}(1,b)| - |C_{n-1}^0(1,b)| \\ &= |C_{n-1}(1,b)| - |C_{n-1}(-1,b)| \end{aligned}$$

The result then follows by the fact that the total number of paths $(0,0)$ to (n,b) is equal to

$$|C_n(0,b)| = \binom{n}{\frac{1}{2}(n+b)}$$

and

$$\begin{aligned} |C_{n-1}(1,b)| &= \tbinom{n-1}{\frac{1}{2}(n+b)} &= \tfrac{n+b}{2n}|C_n(0,b)| \\ |C_{n-1}(-1,b)| &= \tbinom{n-1}{\frac{1}{2}n+b-1} &= \tfrac{n-b}{2n}|C_n(0,b)| \end{aligned}$$

Example 1.16
In an election, candidate A scores α votes and candidate B scores β votes, where $\alpha > \beta$. What is the probability that candidate A was always ahead of candidate B during the election?

Solution:

A.1.4 Hitting times

Let $X_n = \sum_{k=1}^{n} \xi_k$ be a simple random walk with $X_0 = 0$. The **first hitting time** at level ℓ is the time at which the trajectory first reaches level $\ell \in \mathbb{N}$,

$$T_\ell = \min\{n : X_n = \ell\}$$

If $T_\ell = n$, we must have that (1) $X_n = \ell$ and (2) $X_k < \ell$ for all $k = 0,1,2,\ldots,n-1$.

Example 1.17 (Gambler's ruin)
A gambler starts with £x and plays a game in which a fair coin is tossed repeatedly. Each time, if the coin shows heads then he wins £1, but if the coin shows tails he loses £1. The gambler stops when either he goes bankrupt or otherwise reaches some pre-determined amount £a. Find the probability that the gambler goes bankrupt.

Solution:

Theorem 1.18
Let $\{X_n\}$ be a simple random walk with $X_0 = 0$, and let T be the first hitting time at level $\ell = 1$. The PMF of T is given by the following recursive formula,

$$p_n = q\sum_{k=1}^{n-2} p_k p_{n-k-1}, \qquad p_0 = 0, p_1 = p, q = 1 - p..$$

where $p_n = \mathbb{P}(T = n)$.

Proof:

- We cannot move from 0 to 1 in an even number of steps, so $p_{2n} = 0$ for $n \in \mathbb{Z}_{\geq 0}$.
- For $n = 1$, the first step must be upwards, so $p_1 = p$.
- For $n > 1$, the first step must be downwards (which occurs with probability q), and then we need to climb from -1 to 0, and then from 0 to 1.

$$\begin{aligned}\mathbb{P}(T = n) &= q\sum_{k=1}^{n-2} \mathbb{P}(k \text{ steps to first hit 0 from } -1\text{, and } n-k-1 \text{ steps to first hit 1 from 0}) \\ &= q\sum_{k=1}^{n-2} \mathbb{P}(k \text{ steps to first hit 0 from } -1)\mathbb{P}(n-k-1 \text{ steps to first hit 1 from 0}) \\ &= q\sum_{k=1}^{n-2} \mathbb{P}(k \text{ steps to first hit 1 from 0})\mathbb{P}(n-k-1 \text{ steps to first hit 1 from 0}) \\ &= q\sum_{k=1}^{n-2} p_k p_{n-k-1}\end{aligned}$$

Theorem 1.19
The PGF of T is given by $G(t) = \dfrac{1-\sqrt{1-4pqt^2}}{2qt}$.

Proof: Let $G(t) = \sum_{k=0}^{\infty} p_k t^k$ be the PGF of T. The square of the PGF can be written as

$$G(t)^2 = \sum_{k=0}^{\infty}\left(\sum_{i=0}^{k} p_i p_{k-i}\right) t^k$$

Because $p_0 = 0$ and $p_{k+1} = q\sum_{i=1}^{k-1} p_i p_{k-i}$, the inner sum can be written as

$$\sum_{i=0}^{k} p_i p_{k-i} = \sum_{i=1}^{k-1} p_i p_{k-i} = q^{-1} p_{k+1} \quad \text{for } k \geq 2.$$

(and zero for $k = 0$ and $k = 1$). Hence,

$$G(t)^2 = q^{-1}\sum_{k=2}^{\infty} p_{k+1} t^{k+1} = G(t) - pt$$

so

$$qtG(t)^2 = \sum_{k=2}^{\infty} p_{k+1} t^{k+1} = \sum_{k=0}^{\infty} p_k t^k - pt = G(t) - pt.$$

Thus we obtain a quadratic equation for $G(t)$,

$$qtG(t)^2 - G(t) + pt = 0$$

and solving for $G(t)$ we obtain

$$G(t) = \frac{1 \pm \sqrt{1-4pqt^2}}{2qt}$$

For any PGF we must have that $G(t) \leq 1$, so we conclude that

$$G(t) = \frac{1 - \sqrt{1-4pqt^2}}{2qt}$$

Remark 1.20
The probabilities p_n can be computed by taking successive derivatives of $G(t)$ and evaluating these at $t = 1$. An explicit expression is given by

$$p_{2n-1} = p^n q^{n-1} \frac{2}{n} \binom{2n-3}{n-2}.$$

This result can also be obtained by the reflection principle.

Remark 1.21
Since $G(1) = \sum_{k=0}^{\infty} p_k$ we might be inclined to think that $G(1) = 1$. In fact,

$$G(1) = \frac{1 - \sqrt{1-4pq}}{2q} = \frac{1 - |p-q|}{2q} = \begin{cases} 1 & \text{for } p \geq 1/2, \\ p/q & \text{for } p < 1/2. \end{cases}$$

Thus if $p < q$ we see that $G(1) < 1$, which show that the random walk might never reach level 1 when the probability of a positive step is smaller than the probability of a negative step. In this case, $G(1) = \sum_{k=0}^{\infty} p_k = \mathbb{P}(T < \infty)$, so $\mathbb{P}(T = \infty) > 0$. It is a remarkable fact that if $p = 1/2$, the random walk will **always** hit level 1 sooner or later, but this need not happen if $p < 1/2$. This behaviour is known as **criticality** - many systems exhibit qualitatively different behaviour when the value of a parameter p lies either side of some critical value p_c.

Remark 1.22
We can compute the **expected** time before the random walk hits 1 for the first time. If $p < 1/2$ then $\mathbb{P}(T = \infty) > 0$ so $\mathbb{E}(T) = \infty$. For the case $p \geq 1/2$, note that

$$G'(t) = \frac{2p}{\sqrt{1-4pqt^2}} - \frac{1 - \sqrt{1-4pqt^2}}{2qt^2}.$$

$$\begin{aligned} p = 1/2: &\quad \mathbb{E}(T) = \lim_{t \nearrow 1} G'(t) = \lim_{t \nearrow 1} \left(\tfrac{1}{\sqrt{1-t^2}} - \tfrac{1-\sqrt{1-t^2}}{t^2} \right) = +\infty. \\ p > 1/2: &\quad \mathbb{E}(T) = \lim_{t \nearrow 1} G'(t) = \tfrac{1}{p-q}. \end{aligned}$$

Exercise 1.23
Consider a simple random walk $X_0, X_1, X_2, \ldots$ with $X_0 = 0$. Let q be the probability that the random walk eventually returns to the starting position X_0. If $q = 1$, position X_0 is called **recurrent**; if $q < 1$, position X_0 is called **transient**. Show that X_0 is transient if and only if $\sum_{n=1}^{\infty} \mathbb{P}(X_n = 0) < \infty$. [*Hint*: find expressions for the expected number of times that X_0 is re-visited.]

Answer: Let

$$I_n = \begin{cases} 1 & \text{if } X_n = 0 \\ 0 & \text{otherwise.} \end{cases}$$

and let $N = \sum_{n=1}^{\infty} I_n$ be the number of times that state $X_0 = 0$ is revisited.

The expected value of N is given by

$$\mathbb{E}(N) = \mathbb{E}\left(\sum_{n=1}^{\infty} I_n \right) = \sum_{n=1}^{\infty} \mathbb{E}(I_n) = \sum_{n=1}^{\infty} \mathbb{P}(X_n = 0)$$

The expected value of N is also given by

$$\begin{aligned}\mathbb{E}(N) &= \sum_{k=1}^{\infty} k\mathbb{P}(N=k)\\ &= \sum_{k=1}^{\infty}\big[k\mathbb{P}(N\geq k) - k\mathbb{P}(N\geq k+1)\big]\\ &= \sum_{k=1}^{\infty} k\mathbb{P}(N\geq k) - \sum_{k=2}^{\infty}(k-1)\mathbb{P}(N\geq k)\\ &= \sum_{k=1}^{\infty}\mathbb{P}(N\geq k)\\ &= \sum_{k=1}^{\infty} q^k\end{aligned}$$

where the last equality follows by the fact that every return occurs independently with probability q. Combining these results, we get

$$\sum_{n=1}^{\infty}\mathbb{P}(X_n=0) = \sum_{k=1}^{\infty} q^k$$

which diverges if $q=1$, and converges if $q<1$. Thus the random walk is recurrent precisely when $\sum_{n=1}^{\infty}\mathbb{P}(X_n=0)$ is infinite.

A.2 Branching processes

In Victorian England, several aristocratic families realised that their family names could become extinct. In 1873, Sir Francis Galton posed the following question in the *Educational Times*:

> How many male children (on average) must each generation of a family have for the family name to continue in perpetuity?

The first complete answer was put forward by Reverend Henry Watson. Galton and Watson published a paper entitled *On the probability of extinction of families* in 1874. Their model is as follows.

1. A population starts with a single individual, $Z_0 = 1$
2. At time $n=1$, this individual gives birth to Z_1 offspring, where $Z_1\in\{0,1,2,\ldots\}$, then dies.
3. If $Z_1=0$, the population is extinct and $Z_n=0$ for all $n\geq 2$.
4. If $Z_1>0$, each of the Z_1 individuals in the first generation gives birth to a random number of offspring at time $n=2$: the first has $Z_{1,1}$ offspring, the second has $Z_{1,2}$ offspring, ..., and the has last Z_{1,Z_1} offspring.
5. Assume that every individual in every generation has the same offspring distribution, and the number of offpring born to any individual is independent of the number born to any other individual.
6. The total number of individuals in the second generation is

$$Z_2 = \sum_{k=1}^{Z_1} Z_{1,k}.$$

7. The third, fourth, fifth etc. generations are produced in the same way.
8. If it happens that $Z_m = 0$ for some m then $Z_n = 0$ for all $m \geq n$, and the population is **extinct**.

Definition 1.24
A random process $Z_0, Z_1, Z_2 \ldots$ with the properties described above is called a **simple branching process** (or Galton-Watson process).

The offspring distribution determines the evolution of a branching process. Galton's question is to find conditions on the offspring distribution under which

$$\mathbb{P}\big(Z_n \geq 1 \text{ for all } n = 0, 1, 2, \ldots\big) = 1.$$

Let $p_0, p_1, p_2, \ldots$ denote the offspring distribution, and let $G(t) = \sum_{k=1}^{\infty} p_k t^k$ be its PGF.

Theorem 1.25
The PGF of Z_n is the n-fold composition of $G(t)$ with itself,

$$G_{Z_n}(t) = \underbrace{G(G(\ldots G(t)\ldots))}_{n \text{ times}} \qquad (n \geq 1).$$

Proof: For $n = 1$, Z_1 has distribution $p_0, p_1, p_2, \ldots$ so $G_{Z_1}(t) = G(t)$. Suppose the statement holds for some $n \in \mathbb{N}$. Then

$$Z_{n+1} = \sum_{i=1}^{Z_n} Z_{n,i}.$$

is a random sum of Z_n independent variables with PMF $p_0, p_1, \ldots$, and where the number of summands Z_n is independent of the summands $Z_{n,1}, Z_{n,2}, \ldots, Z_{n,Z_n}$.

Hence, by Theorem 6.9,

$$G_{Z_{n+1}}(t) = G_{Z_n}\big[G(t)\big]$$

and by the inductive hypothesis,

$$G_{Z_{n+1}}(t) = \underbrace{G(G(\ldots G(t)\ldots))}_{n+1 \text{ times}}$$

as required.

Theorem 1.26
Let $Z_0, Z_1, Z_2, \ldots$ be a simple branching process, and let μ and σ^2 be the mean and variance of its offspring distribution. Then $\mathbb{E}(Z_n) = \mu^n$ and

$$\mathrm{Var}(Z_n) = \sigma^2\mu^n(1 + \mu + \mu^2 + \cdots + \mu^n) = \begin{cases} \sigma^2(n+1) & \text{if } \mu = 1, \\ \sigma^2\mu^n\left(\frac{1-\mu^{n+1}}{1-\mu}\right) & \text{if } \mu \neq 1. \end{cases}$$

Proof: For brevity, let $G_n(t)$ denote the PGF $G_{Z_n}(t)$ of Z_n

1. To find the mean, differentiate $G_n(t) = G\big(G_{n-1}(t)\big)$,

$$G'_n(t) = G'\big(G_{n-1}(t)\big)G'_{n-1}(t).$$

Evaluating this at $t = 1$, and using the fact that $G_{n-1}(1) = 1$,

$$\mathbb{E}(Z_n) = G'_n(1) = G'(1)G'_{n-1}(1) = \mu\mathbb{E}(Z_{n-1})$$

The result then follows by induction.

2. To find the variance, differentiate $G_n(s) = G\big(G_{n-1}(s)\big)$ twice to obtain
$$G_n''(1) = G''(1)G_{n-1}'(1)^2 + G'(1)G_{n-1}''(1)$$
and substitute in the expression $\mathrm{Var}(Z_n) = G_n''(1) + G_n'(1) - G_n'(1)^2$.

A.2.1 Extinction probability

The event that the population becomes extinct can be written as
$$E = \big\{\omega \in \Omega : Z_n(\omega) = 0 \text{ for some } n \in \mathbb{N}\big\}.$$
This can be written as the union of an expanding sequence of events $E_1 \subseteq E_2 \subseteq \ldots$,
$$E = \textstyle\bigcup_{n=1}^{\infty} E_n \quad \text{where} \quad E_n = \{\omega \in \Omega : Z_n(\omega) = 0\}.$$
By the continuity of probability measures, the **extinction probability** is
$$\mathbb{P}(E) = \lim_{n\to\infty} \mathbb{P}(E_n).$$
Using the fact that $\mathbb{P}(E_n) = \mathbb{P}(Z_n = 0) = G_{Z_n}(0)$,
$$\mathbb{P}(E) = \lim_{n\to\infty} G_{Z_n}(0) = \lim_{n\to\infty} \underbrace{G(G(\ldots G(0)\ldots))}_{n \text{ times}}$$
Remarkably, the extinction probability $\mathbb{P}(E)$ can be computed even when G_{Z_n} is not known explicitly.

Theorem 1.27
The extinction probability is the smallest non-negative solution of the so-called **extinction equation**,
$$x = G(x)$$
where G is the PGF of the offspring distribution.

Proof: Let $e = \mathbb{P}(E)$ be the extinction probability. First we show that e is a solution of $x = G(x)$. Let
$$x_n = \underbrace{G(G(\ldots G(0)\ldots))}_{n \text{ times}}$$
Then (1) $e = \lim_{n\to\infty} x_n$ and (2) $G(x_n) = x_{n+1}$, so
$$e = \lim_{n\to\infty} x_n = \lim_{n\to\infty} x_{n+1} = \lim_{n\to\infty} G(x_n) = G(\Big(\lim_{n\to\infty} x_n\Big) = G(e),$$
where we have used the fact that G is a continuous function.

To show that $e = \mathbb{P}(E)$ is the smallest solution, let e' be another solution of $x = G(x)$ in $[0, 1]$. Since $e' \geq 0$ and $G(t)$ is a non-decreasing function,
$$G(0) \leq G(e') = e'.$$
Applying G to both sides, since $G(t)$ is increasing,
$$G(G(0)) \leq G(G(e')) = G(e') = e'.$$
Repeating this procedure, we get
$$\mathbb{P}(E_n) = \underbrace{G(G(\ldots G(0)\ldots))}_{n \text{ times}} \leq e'.$$
Hence,
$$e = \mathbb{P}(E) = \lim_{n\to\infty} \mathbb{P}(E_n) \leq \lim_{n\to\infty} e' = e',$$
so e is not larger than any other solution e' of $x = G(x)$.

A.3 Martingales

Definition 1.28
A random process $X_0, X_1, \dots$ is called a **martingale** with respect to the sequence of random variables $\xi_1, \xi_2, \dots$ if

1. $\mathbb{E}(|X_n|) < \infty$ and
2. $\mathbb{E}(X_{n+1}|\xi_1, \dots, \xi_n) = X_n$.

- A **sub-martingale** has $\mathbb{E}(X_{n+1}|\xi_1, \dots, \xi_n) \geq X_n$ (the process tends to increase over time).
- A **super-martingale** has $\mathbb{E}(X_{n+1}|\xi_1, \dots, \xi_n) \leq X_n$ (the process tends to decrease over time).

Example 1.29 (Random walk)
Let $X_n = X_0 + \sum_{k=1}^{n} \xi_n$ be a symmetric simple random walk (so $\mathbb{P}(\xi_k = 1) = \mathbb{P}(\xi_k = -1) = 1/2$).

Show that $\{X_n\}$ is a martingale with respect to the increments $\xi_1, \xi_2, \dots$.

Solution:

Example 1.30 (Sub-martingale)
Let $\xi_1, \xi_2, \dots$ be independent random variables with zero means, finite variances and partial sums $X_n = \sum_{i=1}^{n} \xi_i$. Show that X_n^2 is a sub-martingale with respect to $\xi_1, X_2, \dots$.

Solution:

Theorem 1.31 (Martingale Convergence Theorem)
Let $X_0, X_1, \dots$ be a martingale such that $\mathbb{E}(|X_n|)$ is bounded for all $n = 0, 1, 2, \dots$. Then there exists a finite random variable X such that $X_n \to X$ with probability one as $n \to \infty$.

The MCT can be used to prove the following remarkable theorem.

Theorem 1.32
A symmetric simple random walk on $\mathbb{Z}$ will visit every point with probability 1.

Proof: Let $X_0 = 0$ and $b \in \mathbb{Z}$, and suppose (without loss of generality) that $b < 0$. Let T be the first time n for which $X_n = b$, and consider the corresponding **stopped process** $\tilde{X}_0, \tilde{X}_1, \tilde{X}_2, \dots$ defined by $\tilde{X}_n = X_{\min\{n,T\}}$.

- The stopped process remains in position b from time T onwards.

It is easy to show that $\tilde{X}_n$ is a martingale, and because $\tilde{X}_n - b$ is non-negative, $\mathbb{E}(|\tilde{X}_n - b|) = \mathbb{E}(\tilde{X}_n - b) < \infty$.

By the Martingale Convergence Theorem, the limit $\tilde{X} = \lim_{n\to\infty} \tilde{X}_n$ exists and is finite with probability one.

- In particular, $|\tilde{X}_{n+1} - \tilde{X}_n|$ converges to zero, and must therefore be less than 1 for large n.
- However $|\tilde{X}_{n+1} - \tilde{X}_n| = 1$ whenever $n < T$.
- Thus we have $T < \infty$, and hence $X_n = b$ for some n.

Exercise 1.33
Let $Z_0, Z_1, Z_2, \dots$ be a simple branching process with $Z_0 = 1$. Show that the sequence $W_1, W_2, \dots$ with $W_n = Z_n/\mathbb{E}(Z_n)$ is a martingale with respect to $Z_1, Z_2, \dots$.
Answer: Conditioned on $Z_n = z_n$, the number Z_{n+1} is the sum of z_n independent familiy sizes,

$$\mathbb{E}(Z_{n+1} : Z_n = z_n) = \mu z_n$$

where μ is the expected family size. By the Markov property,

$$\mathbb{E}(Z_{n+1} : Z_1, Z_2, \dots, Z_n) = \mu Z_n$$

Since $\mathbb{E}(Z_n) = \mu^n$, we conclude that

$$\mathbb{E}(W_{n+1} : Z_1, Z_2, \dots, Z_n) = W_n$$

as required.

Exercise 1.34
An urn contains one red ball and one green ball. At each time step, we choose one ball uniformly at random from the urn, and replace it along with another ball of the same colour. Let R_n and G_n respectively denote the number of red balls and green balls after n steps, and let M_n denote the fraction of green balls in the urn.

1. Show that M_n is a martingale.
2. Show that M_n converges to a finite limit with probability 1 as $n \to \infty$

Answer:

1. M_n is a martingale because

$$\begin{aligned}\mathbb{E}(M_{n+1}|R_0,G_0,\ldots,R_n,G_n) &= \left(\frac{R_n}{R_n+G_n}\right)\left(\frac{G_n}{R_n+G_n+1}\right)+\left(\frac{G_n}{R_n+G_n}\right)\left(\frac{G_n+1}{R_n+G_n+1}\right)\\ &= \frac{G_n}{R_n+G_n}\\ &= M_n\end{aligned}$$

2. Since $M_n \geq 0$ is bounded for all $n \in \mathbb{N}$, it follows by the martingale that there exists a finite random variable M such that $M_n \to M$ with probability one as $n \to \infty$. In fact, it can be shown that

$$\mathbb{P}(G_n = m+1) = \binom{m}{n}\frac{m!(n-m)!}{(n+1)!} = \frac{1}{n+1}$$

and hence

$$\mathbb{P}(M_n \leq x) = \frac{\lfloor x(n+2)-1\rfloor}{n+1} \to x \quad \text{as } n \to \infty$$

Thus the distribution of M_n approaches a uniform distribution on $[0,1]$ as $n \to \infty$.

Chapter B Bayesian Inference

So far we have looked at **frequentist inference**, which assumes that an unknown parameter θ has a fixed (but unknown) value.

- The PMF/PDF of an observation is written as $f(x;\theta)$.
- The likelihood function is written as $L(\theta;x)$.
- Estimators such as the MME and MLE claim to estimate the 'true' value of θ.

For **Bayesian inference**, we instead think of an unknown parameter θ as a **random variable**.

- The PMF/PDF of an observation is written as $f(x|\theta)$.
- The likelihood function of is written as $L(\theta|x)$.
- We seek to estimate the distribution of θ.

B.1 Bayes' theorem

If the events $A_1, A_2, \ldots$ form a partition of event B, Bayes' theorem states that

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_k \mathbb{P}(B|A_k)\mathbb{P}(A_k)}.$$

Let X and Y be discrete random variables taking values in the sets $\{x_1, x_2, \ldots\}$ and $\{y_1, y_2, \ldots\}$ respectively. Because the events $\{Y = y_1\}, \{Y = y_2\}, \ldots$ form a partition of the event $\{X = x_i\}$, we have

$$\mathbb{P}(Y = y_j|X = x_i) = \frac{\mathbb{P}(X = x_i|Y = y_j)\mathbb{P}(Y = y_j)}{\sum_k \mathbb{P}(X = x_i|Y = y_k)\mathbb{P}(Y = y_k)}.$$

To avoid cluttering the notation with subscripts, we write this as

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

where the sum in the denominator is taken over the range of Y. In terms of PMFs, this becomes

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\sum_y f_{X|Y}(x|y)f_Y(y)}.$$

This extends directly to the case of continuous random variables, the only difference being that the denominator (which is the marginal PDF of X) is expressed by an integral:

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

For Bayesian inference, the unknown parameter θ takes the role of Y in the above formulation. To simplify the notation we denote the PMF/PDF of θ by the symbol π:

$$\pi(\theta|\mathbf{x}) = \begin{cases} \dfrac{f(\mathbf{x}|\theta)\pi(\theta)}{\sum_\theta f(\mathbf{x}|\theta)\pi(\theta)} & (\theta \text{ discrete}), \\[2ex] \dfrac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)\,d\theta} & (\theta \text{ continuous}). \end{cases}$$

This can also be expressed in terms of likelihood functions,

$$\pi(\theta|\mathbf{x}) = \begin{cases} \dfrac{L(\theta|\mathbf{x})\pi(\theta)}{\sum_\theta L(\theta|\mathbf{x})\pi(\theta)} & (\theta \text{ discrete}), \\[2ex] \dfrac{L(\theta|\mathbf{x})\pi(\theta)}{\int L(\theta|\mathbf{x})\pi(\theta)\,d\theta} & (\theta \text{ continuous}). \end{cases}$$

B.2 The prior and posterior distributions

The prior distribution

Suppose we have an initial estimate for distribution of θ, perhaps obtained as a result of some preliminary experiments.

- This is called the **prior distribution** of θ, which we denote by $\pi_0(\theta)$.

An initial point estimate of θ can be computed from the prior distribution, for example

- by the **mean** of the prior distribution: $\hat{\theta} = \mathbb{E}\big[\pi_0(\theta)\big]$ or
- by the **mode** of the prior distribution: $\hat{\theta} = \operatorname{argmax}_\theta \big[\pi_0(\theta)\big]$.

If we have no prior knowledge, we should initially consider every value of θ to be equally likely. For example, if θ is continuous and all we know is that θ belongs to some interval $[a, b]$, we should adopt the **uniform** distribution over $[a, b]$ as the prior distribution of θ,

$$\pi_0(\theta) = \begin{cases} 1/(b-a) & \text{if } a \leq \theta \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

In this context, the uniform distribution is often called the **naïve** or **non-informative** prior.

The posterior distribution

Suppose we now obtain some sample data $\mathbf{x} = (x_1, \ldots, x_n)$.

- Bayes' theorem can be used to combine the prior distribution with the data.
- This yields an updated PMF/PDF $\pi_1(\theta)$, called the **posterior distribution** of θ.

By Bayes' theorem,

$$\pi_1(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi_0(\theta)}{\sum_\theta f(\mathbf{x}|\theta)\pi_0(\theta)} \qquad \text{or} \qquad \pi_1(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi_0(\theta)}{\int f(\mathbf{x}|\theta)\pi_0(\theta)\,d\theta}.$$

In terms of likelihood functions,

$$\pi_1(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})\pi_0(\theta)}{\sum_\theta L(\theta|\mathbf{x})\pi_0(\theta)} \qquad \text{or} \qquad \pi_1(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})\pi_0(\theta)}{\int L(\theta|\mathbf{x})\pi_0(\theta)\,d\theta}.$$

Having obtained the posterior distribution, we can compute an updated estimate of θ, for example

- by the mean of the posterior distribution: $\hat{\theta} = \mathbb{E}\big[\pi_1(\theta)\big]$, or
- by the mode of the posterior distribution: $\hat{\theta} = \mathrm{argmax}_\theta\,\big[\pi_1(\theta)\big]$.

Definition 2.1
The mode of the posterior is called the **maximum a-posteriori** or **MAP** estimator of θ.

Remark 2.2
The denominator of the posterior depends only on $\mathbf{x}$, so the posterior is proportional to the likelihood times the prior:

$$\pi_1(\theta|\mathbf{x}) \propto L(\theta|\mathbf{x})\pi_0(\theta),$$

The MAP estimator is the value of θ that maximizes the numerator $L(\theta|\mathbf{x})\pi_0(\theta)$ of the posterior.

- Note that when π_0 is the uniform distribution, the MAP estimator is just the MLE.

To compute the mean of $\pi_1(\theta|\mathbf{x})$ we also need to compute the denominator. This is not always easy, which is why the MAP estimator is more widely used in practical applications.

Example 2.3
Suppose we have three tins of biscuits. The first tin contains 30 chocolate and 10 plain biscuits, the second tin contains 20 chocolate and 20 plain biscuits, and the third tin contains 10 chocolate and 30 plain biscuits. A tin is selected at random, and a biscuit is chosen at random from the tin.

1. If a chocolate biscuit is chosen, estimate which tin was selected.

The biscuit is replaced, then a biscuit is again chosen from the tin.

2. If a chocolate biscuit is chosen, update your estimate regarding which tin was selected.
3. If a plain biscuit is chosen, update your estimate regarding which tin was selected.

Solution:

Remark 2.4
We can think of the biscuit tins in Example 2.3 as competing scientific hypotheses:

- The probability assigned to each hypothesis indicates its **relative plausibility**.
- We update the relative plausibility of each competing hypothesis based on **observation**.

In this way, Bayesian inference embodies the **scientific method**.

Exercise 2.5
Suppose we have three coins A, B and C which have probabilities 1/4, 1/2 and 3/4 respectively of showing heads. A coin is chosen at random, and tossed three times. If exactly two heads are obtained, use the maximum a-posteriori (MAP) estimator to estimate which coin was chosen.

Answer: First we define the parameter $\theta \in \{1, 2, 3\}$ such that $\{\theta = 1\}$ is the event that coin A is chosen, $\{\theta = 2\}$ is the event that coin B is chosen, and $\{\theta = 3\}$ is the event that coin C is chosen. We should initially assume that each coin is equally likely to be chosen, so we choose the uniform prior distribution:

$$\begin{aligned}
\pi_0(1) &= \mathbb{P}(\theta = 1) = 1/3 \\
\pi_0(2) &= \mathbb{P}(\theta = 2) = 1/3 \\
\pi_0(3) &= \mathbb{P}(\theta = 3) = 1/3
\end{aligned}$$

Let T be the event that exactly two heads are obtained. Then

$$\begin{aligned}
\mathbb{P}(T|\theta = 1) &= 3(1/4)^2(3/4) = 9/64 \\
\mathbb{P}(T|\theta = 2) &= 3(1/2)^2(1/2) = 3/8 \\
\mathbb{P}(T|\theta = 3) &= 3(1/4)(3/4)^2 = 27/64
\end{aligned}$$

The denominator of the posterior is the overall probability of obtaining exactly two heads:

$$\begin{aligned}\mathbb{P}(T) &= \mathbb{P}(T|\theta=1)\mathbb{P}(\theta=1)+\mathbb{P}(T|\theta=2)\mathbb{P}(\theta=2)+\mathbb{P}(T|\theta=3)\mathbb{P}(\theta=3)\\ &= 3(1/4)^2(3/4)(1/3)+3(1/2)^3(1/3)+3(3/4)^2(1/4)(1/3)\\ &= 3/64+8/64+9/64\\ &= 20/64\end{aligned}$$

Hence the posterior distribution is given by

$$\pi_1(\theta)=\frac{\mathbb{P}(T|\theta)\pi_0(\theta)}{\mathbb{P}(T)}=$$

from which we obtain

$$\begin{aligned}\pi_1(1)&=3/20,\\ \pi_1(2)&=8/20,\\ \pi_1(3)&=9/20.\end{aligned}$$

The MAP estimator (mode of the posterior) is $\theta=3$, which corresponds to coin C.

B.3 The binomial model

A suitable model for estimating the distribution of a parameter in the interval $[0,1]$ is provided by the **beta distribution**.

Definition 2.6
The beta distribution with parameters $\alpha,\beta>0$ is defined by the PDF

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} & \text{if } 0\leq x\leq 1,\\ 0 & \text{otherwise,}\end{cases}$$

where $B(\alpha,\beta)$ is the so-called **beta function**,

$$B(\alpha,\beta)=\int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,dt,$$

which is defined for all $\alpha,\beta>0$.

Lemma 2.7
Let $X\sim\text{Beta}(\alpha,\beta)$. Then $\mathbb{E}(X)=\dfrac{\alpha}{\alpha+\beta}$ and $\text{Mode}(X)=\dfrac{\alpha-1}{\alpha+\beta-2}$ provided that $\alpha,\beta>1$.

Proof: Exercise.

Example 2.8
Let $X\sim\text{Binomial}(n,\theta)$ where n is known, but $0<\theta<1$ is unknown.

1. We conduct a sequence of n independent trials and observe k successes. Find a suitable prior distribution for θ, compute the posterior distribution, and find its mean and mode.

2. We conduct a further sequence of n independent trials, this time observing k' successes. Compute an updated posterior distribution for θ, and find its mean and mode.

Solution:

B.4 The exponential model

A suitable model for estimating the distribution of a non-negative parameter $\theta \geq 0$ is provided by the **gamma distribution**.

Definition 2.9
The **gamma distribution** with parameter $\alpha, \beta > 0$ is defined by the PDF

$$f(x; \alpha, \beta) = \begin{cases} \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

where $\Gamma(\alpha)$ is the so-called **gamma function**,

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt, \quad \text{which is defined for all } \alpha \in \mathbb{R}.$$

Lemma 2.10
Let $X \sim \text{Gamma}(\alpha, \beta)$. Then $\mathbb{E}(X) = \dfrac{\alpha}{\beta}$ and $\text{Mode}(X) = \dfrac{\alpha - 1}{\beta}$ provided that $\alpha > 1$.

Proof: Exercise.

Example 2.11
Let $X \sim \text{Exponential}(\lambda)$, where $\lambda > 0$ is an unknown rate parameter. Let $X_1, X_2, \ldots, X_n$ be a random sample of observations from the distribution of X, and suppose that we adopt the $\text{Gamma}(\alpha, \beta)$ distribution as a prior distribution for λ, where $\alpha, \beta > 0$ are fixed values (perhaps estimated in some preliminary experiments).

1. Find the mean and mode of the prior distribution.
2. Show that the posterior of λ is the $\text{Gamma}(\alpha + n, \beta + \sum_{i=1}^n x_i)$ distribution.

3. Find the mean and mode of the posterior distribution.

Solution:

Exercise 2.12

1. Let $X \sim \text{Geometric}(\theta)$ where $0 < \theta < 1$ is unknown.

   (a) A single experiment yields the observation k. Find a suitable prior distribution for θ, compute the corresponding posterior distribution, and find the MAP estimator of θ for this posterior.

   Answer: Let $f(x|\theta)$ be the PMF of the Geometric(θ) distribution:

$$f(x|\theta) = \theta^x (1-\theta)^{n-x}$$

Without any information about θ, we should choose the **naïve** prior:

$$\pi_0(\theta) = \begin{cases} 1 & \text{if } 0 \leq \theta \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

For the observation $X = k$, the likelihood function is

$$L(\theta|k) = f(k|\theta) = \theta(1-\theta)^{k-1}$$

The posterior distribution is:

$$\pi_1(\theta|k) = \frac{L(\theta|k)\pi_0(\theta)}{\int L(\theta|k)\pi_0(\theta)\,d\theta} = \frac{\theta(1-\theta)^{k-1}}{\int_0^1 \theta(1-\theta)^{k-1}\,d\theta}.$$

We recognise this as the PDF of the Beta(α,β) distribution, with parameters $\alpha = 2$ and $\beta = k$. The mode of the Beta(α,β) distribution is $(\alpha-1)/(\alpha+\beta-2)$, so the MAP estimator is

$$\hat{\theta}_{MAP} = \frac{1}{k}.$$

(b) A second experiment yields the observation k'. Compute an updated posterior distribution for θ, and find a new MAP estimator for θ.

Answer: For the observation $X = k'$, the likelihood function is

$$L(\theta|k') = f(k'|\theta) = \theta(1-\theta)^{k'-1}$$

Using π_1 as the new prior distribution for θ, the new posterior distribution π_2 is

$$\pi_2(\theta|k,k') = \frac{L(\theta|k')\pi_1(\theta)}{\int_0^1 L(\theta|k')\pi_1(\theta)\,d\theta} = \frac{\theta^2(1-\theta)^{k+k'-2}}{\int_0^1 \theta^2(1-\theta)^{k+k'-2}\,d\theta}.$$

which we recognise as the PDF of the Beta(α,β) distribution, with parameters $\alpha = 3$ and $\beta = k + k' - 1$. Hence the new MAP estimator is

$$\hat{\theta}_{MAP} = \frac{2}{k+k'}.$$

2. Let $X \sim$ Poisson(λ), where $\lambda > 0$ is unknown, and let $X_1, X_2, \ldots, X_n$ be a random sample of observations from the distribution of X. Suppose we adopt the Gamma(α,β) distribution as a prior distribution for λ, where $\alpha, \beta > 0$ are fixed values.

(a) Show that the MAP estimator of λ is given by

$$\hat{\lambda}_{MAP} = \frac{\alpha - 1 + \sum_{i=1}^n X_i}{n+\beta}$$

Answer: Let $f(x|\lambda)$ be the PMF of the Poisson(λ) distribution:

$$f(x|\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = 0,1,2,\ldots, \\ 0 & \text{otherwise.} \end{cases}$$

The PDF of the prior distribution is

$$\pi_0(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\lambda^{\alpha-1}\exp(-\beta\lambda).$$

which has mean α/β and mode $(\alpha-1)/\beta$.

Let $\boldsymbol{x}=(x_1,x_2,\ldots,x_n)$ be a realisation of the sample. The likelihood function is

$$L(\lambda|\mathbf{x})=\prod_{i=1}^n f(x_i|\lambda)=\prod_{i=1}^n \frac{\lambda^{x_i}e^{-\lambda}}{x_i!}=\lambda^{\sum x_i}e^{-n}\prod_{i=1}^n\frac{1}{x_i!}$$

The PDF of the posterior distribution is

$$\pi_1(\lambda|\mathbf{x})=\frac{L(\lambda|\mathbf{x})\pi_0(\lambda)}{\int_0^\infty L(\lambda|\mathbf{x})\pi_0(\lambda)\,d\lambda}.$$

To find the MAP estimator, we need to find the value of λ that maximises the numerator:

$$L(\lambda|\mathbf{x})\pi_0(\lambda)=c\lambda^{(\alpha-1+\sum x_i)}e^{-(n+\beta)\lambda}\quad\text{where}\quad c=\left(\prod_{i=1}^n\frac{1}{x_i!}\right)\frac{\beta^\alpha}{\Gamma(\alpha)}$$

Let $g(\lambda)=\lambda^{\alpha-1+\sum x_i}e^{-(n+\beta)\lambda}$. Then

$$g'(\lambda)=\lambda^{(\alpha-2+\sum x_i)}e^{-(n+\beta)\lambda}\left[(\alpha-1+\sum_{i=1}^n x_i)-\lambda(n+\beta)\right]$$

Setting $g'(\lambda)$ to zero and solving for λ, we obtain the MAP estimator

$$\hat\lambda_{MAP}=\frac{\alpha-1+\sum_{i=1}^n X_i}{n+\beta}$$

as required.

(b) Comment on the limiting cases (i) $n=0$ and (ii) $n\to\infty$.

Answer:

- When $n=0$, $\hat\lambda_{MAP}=(\alpha-1)/\beta$ is the mode of the prior distribution.
- When $n\to\infty$,

$$\hat\lambda_{MAP}=\frac{\frac{\alpha-1}{n}+\frac{1}{n}\sum_{i=1}^n X_i}{1+\frac{\beta}{n}}\to\frac{1}{n}\sum_{i=1}^n X_i$$

As the sample size increases, the effect of the prior decreases, and the MAP estimator approaches the sample mean in the limit as $n\to\infty$.

3. Let $X_1,\ldots,X_n$ be a random sample from the $N(\mu,\sigma^2)$ distribution, where the mean μ is unknown but the variance σ^2 is known. Suppose we adopt the $N(\mu_0,\sigma_0^2)$ distribution as a prior for the unknown mean μ (where μ_0 and σ_0^2 are known constants). Compute the maximum a-posteriori (MAP) estimator of μ.

Answer: Let $\pi_0(\mu)$ denote the prior density function of μ:

$$\pi_0(\mu)=\frac{1}{\sigma_0\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right].$$

For the observed sequence $\mathbf{x}=(x_1,x_2,\ldots,x_n)$, the likelihood function is

$$\begin{aligned}L(\mu\,|\,\mathbf{x})&=\prod_{i=1}^n\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\exp\left[-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right]\\&=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n/2}\exp\left[-\frac{1}{2}\sum_{i=1}^n\left(\frac{x_i-\mu}{\sigma}\right)^2\right].\end{aligned}$$

The posterior density function of μ combines the data and the prior:

$$\pi_1(\mu) = \frac{L(\mu\,|\,\mathbf{x})\pi_0(\mu)}{\int L(\mu|\mathbf{x})\pi_0(\mu)\,d\mu}.$$

The MAP estimator of μ is the value that maximises the posterior $\pi_1(\mu)$. Since the denominator in the above expression for π_1 is a constant, it is sufficient to find the value of μ that maximises the numerator,

$$L(\mu|\mathbf{x})\pi_0(\mu) = \left(\frac{1}{\sigma_0\sqrt{2\pi}}\right)\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right].$$

Let

$$g(\mu) = \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 - \frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right].$$

The value of μ that maximizes $L(\mu|\mathbf{x})\pi_0(\mu)$ also maximizes $g(\mu)$. The first derivative of g with respect to μ is

$$g'(\mu) = \left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) - \frac{1}{\sigma_0^2}(\mu-\mu_0)\right] g(\mu).$$

Setting this equal to zero,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = \frac{1}{\sigma_0^2}(\mu-\mu_0),$$

and solving for μ, we obtain the MAP estimator,

$$\hat{\mu}_{MAP} = \left(\frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}\right)\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i + \frac{\mu_0}{\sigma_0^2}\right).$$

- This expression can be rearranged to give

$$\hat{\mu}_{MAP}\left(1+\frac{n\sigma_0^2}{\sigma^2}\right) = \frac{\sigma_0^2}{\sigma^2}\sum_{i=1}^{n}x_i + \mu_0.$$

  This shows that $\hat{\mu} = \mu_0$ when $n = 0$, so the $\hat{\mu}_{MAP}$ is equal to the mean of the prior distribution when there is no data.

- The expression can also be rearranged to give

$$\hat{\mu}_{MAP}\left(1+\frac{\sigma^2}{n\sigma_0^2}\right) = \frac{1}{n}\sum_{i=1}^{n}x_i + \left(\frac{\sigma^2}{n\sigma_0^2}\right)\mu_0.$$

  This shows that $\hat{\mu}_{MAP} \to \bar{X}$ as $n \to \infty$ (which is independent of the prior).

Chapter C The Bivariate Normal Distribution

C.1 Bivariate transformations

Definition 3.1
Let $h : \mathbb{R}^2 \to \mathbb{R}^2$ and let $(u, v) = h(x, y)$. The **Jacobian determinant** of the transformation h is the determinant of its 2×2 matrix of partial derivatives:

$$J = \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{vmatrix}$$

Theorem 3.2
Let U and V be jointly continuous random variables, let $f_{U,V}$ be their joint PDF, let $g : \mathbb{R}^2 \to \mathbb{R}^2$ be an injective transform over the support of $f_{U,V}$ and let $(X, Y) = g(U, V)$. Then the joint PDF of X and Y is given by

$$f_{X,Y}(x, y) = |J| f_{U,V}\left[g^{-1}(x, y)\right]$$

where J is the Jacobian determinant of the transformation g^{-1},

$$J = \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{vmatrix}$$

where $(u, v) = g^{-1}(x, y)$.

Remark 3.3
The absolute value $|J|$ is a scale factor, which ensures that the transformed PDF $f_{X,Y}(x, y)$ integrates to one.

Example 3.4
Let U and V be continuous random variables, and let $X = U + V$ and $Y = U - V$.

1. Find the joint PDF of X and Y in terms of the joint PDF of U and V.
2. If $U, V \sim$ Exponential(1) are independent, find the joint PDF of X and Y.

Solution:

C.2 The bivariate normal distribution

Theorem 3.5
if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, then

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Corollary 3.6
If $U, V \sim N(0,1)$ are independent, then $aU + bV \sim N(0, a^2 + b^2)$ for all $a, b \in \mathbb{R}$.

Definition 3.7
A pair of random variables U and V have the **standard bivariate normal distribution** if their joint PDF $f : \mathbb{R}^2 \to [0, \infty)$ can be written as

$$f_{U,V}(u,v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\big(u^2 - 2\rho uv + v^2\big)\right)$$

where ρ is a constant satisfying $-1 < \rho < 1$.

Definition 3.8
A pair of random variables X and Y are said to have **bivariate normal distribution** with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 and correlation ρ, if their joint PDF can be written as

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right)$$

The following lemma can be used to derive many properties of the bivariate normal distribution.

Lemma 3.9
Let $U, V \sim N(0,1)$ be independent, let $\rho \in (-1, +1)$. Then the random variables

$$\begin{aligned} X &= \mu_1 + \sigma_1 U, \\ Y &= \mu_2 + \sigma_2\big(\rho U + \sqrt{1-\rho^2}V\big) \end{aligned}$$

have bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ.

Proof: To find the joint PDF of X and Y, let $g(u,v)$ denote the transformation:

$$g(u,v) = \big[\mu_1 + \sigma_1 u, \mu_2 + \sigma_2(\rho u + \sqrt{1-\rho^2}v)\big].$$

The inverse transformation is

$$g^{-1}(x,y) = \left(\frac{x-\mu_1}{\sigma_1}, \frac{1}{\sqrt{1-\rho^2}}\left[\left(\frac{y-\mu_2}{\sigma_2}\right) - \rho\left(\frac{x-\mu_1}{\sigma_1}\right)\right]\right)$$

The joint PDF of X and Y is $f_{X,Y}(x,y) = |J| f_{U,V}(u,v)$, where J is the Jacobian determinant of the inverse transformation:

$$J = \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{vmatrix} = \begin{vmatrix} \dfrac{1}{\sigma_1} & 0 \\ \dfrac{1}{\rho\sigma_1} & \dfrac{1}{\sigma_2\sqrt{1-\rho^2}} \end{vmatrix} = \frac{1}{\sigma_1\sigma_2\sqrt{1-\rho^2}}$$

Because U and V are independent,

$$f_{U,V}(u,v) = f_U(u) f_V(v) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(u^2+v^2)\right) \qquad u,v \in \mathbb{R}.$$

and since

$$\begin{aligned} u^2+v^2 &= \left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \frac{1}{1-\rho^2}\left[\left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \rho^2\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right] \\ &= \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right] \end{aligned}$$

it follows that

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_1\sigma_1\sqrt{1-\rho^2}}\exp\left(\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right)$$

as required.

The following theorem shows that if X and Y have bivariate normal distribution, then any linear combination of X and Y is normally distributed.

Theorem 3.10
Let X and Y have bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ. Then

$$aX + bY \sim N\big(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\sigma_1\sigma_2\rho + b^2\sigma_2^2\big)$$

Proof: Let $Z = aX + bY$, let U and V be independent standard normal random variables, and let

$$\begin{aligned} X' &= \mu_1 + \sigma_1 U \\ Y' &= \mu_2 + \sigma_2\big(\rho U + \sqrt{1-\rho^2}V\big) \end{aligned}$$

By Lemma 3.9, X and Y have the same joint distribution as X' and Y', so $Z = aX + bY$ has the same distribution as

$$Z' = aX' + bY' = (a\mu_1 + b\mu_2) + (a\sigma_1 + b\sigma_2\rho)U + b\sigma_2\sqrt{1-\rho^2}V$$

Because $U, V \sim N(0,1)$ are independent, it follows by Corollary 3.6 that

$$Z' \sim N\left(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\sigma_1\sigma_2\rho + b^2\sigma_2^2\right),$$

so $Z = aX + bY$ has normal distribution, as required.

C.3 Properties of the bivariate normal distribution

Theorem 3.11
Let X and Y have bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ. Then

1. $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$,
2. ρ is the correlation coefficient of X and Y, and
3. X and Y are independent if and only if $\rho = 0$.

Proof: Let $U, V \sim N(0,1)$ and define

$$\begin{aligned} X &= \mu_1 + \sigma_1 U \\ Y &= \mu_2 + \sigma_2\big(\rho U + \sqrt{1-\rho^2}V\big) \end{aligned}$$

1. In the proof of Theorem 3.10:
   - taking $a = 1$ and $b = 0$ yields $X \sim N(\mu_1, \sigma_1^2)$, and
   - taking $a = 0$ and $b = 1$ yields $Y \sim N(\mu_2, \sigma_2^2)$.
2. Using the fact that $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$ for all $a, b, c, d \in \mathbb{R}$,

$$\begin{aligned} \text{Cov}(X,Y) &= \text{Cov}\big[\mu_1 + \sigma_1 U, \mu_2 + \sigma_2(\rho U + \sqrt{1-\rho^2}V)\big] \\ &= \sigma_1\sigma_2 \text{Cov}(U, \rho U + \sqrt{1-\rho^2}V) \\ &= \sigma_1\sigma_2\big[\rho\mathbb{E}(U^2) + \sqrt{1-\rho^2}\mathbb{E}(UV)\big] \\ &= \sigma_1\sigma_2\rho. \end{aligned}$$

   Thus $\rho = \dfrac{\text{Cov}(X)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$ as required.

3. If X and Y are independent, they are uncorrelated. If X and Y are uncorrelated then $\rho = 0$, so the joint PDF of X and Y satisfies

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right) \times \frac{1}{\sqrt{2\pi}\sigma_2}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right) \\ &= f_X(x)f_Y(y). \end{aligned}$$

   Because this holds for all $x, y \in \mathbb{R}$, it follows that X and Y are independent.

C.4 Conditional distributions

Theorem 3.12
Let X and Y have bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ. Then the conditional distribution of Y given $X = x$ is also normal, with conditional mean and variance given by

$$\mathbb{E}(Y|X = x) = \mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(x - \mu_1),$$

$$\text{Var}(Y|X = x) = \sigma_2^2(1 - \rho^2),$$

and the conditional mean and variance of Y given X is

$$\mathbb{E}(Y|X) = \mathbb{E}(Y) + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}\big[X - \mathbb{E}(X)\big],$$

$$\text{Var}(Y|X) = \text{Var}(Y)(1 - \rho^2).$$

Proof: Let $U, V \sim N(0,1)$ be independent, and define the random variables

$$\begin{aligned} X &= \mu_1 + \sigma_1 U, \\ Y &= \mu_2 + \sigma_2\big[\rho U + \sqrt{1-\rho^2}V\big] \\ &= \mu_2 + \sigma_2\left[\rho\left(\frac{X-\mu_1}{\sigma_1}\right) + \sqrt{1-\rho^2}V\right]. \end{aligned}$$

If X is fixed at x, then Y is a linear transformation of V, so the conditional distribution of Y given that $X = x$ is a normal distribution. Furthermore, since $\mathbb{E}(V) = 0$ and $\mathrm{Var}(V) = 1$ we have

$$\mathbb{E}(Y|X=x) = \mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(x-\mu_1)$$

$$\mathrm{Var}(Y|X=x) = \sigma_2^2(1-\rho^2)$$

as required.

Remark 3.13

- Given $X = x$, the conditional mean is obtained by adjusting $\mathbb{E}(Y)$ by an amount proportional to the difference between $X = x$ and its mean $\mathbb{E}(X)$. The size of this adjustment is determined by (1) the size of Y relative to X, expressed by the ratio σ_2/σ_1, and (2) the correlation coefficient ρ, which quantifies the linear dependence between X and Y. Note that $\rho = 0$ implies that $\mathbb{E}(Y|X=x) = 0$ for all x (which is not surprising, given that uncorrelated normal variables are independent).

- The conditional variance $\mathrm{Var}(Y|X=x)$ quantifies the variability in Y that is **not** explained by the fact that X takes the value x. The squared correlation coefficient ρ^2 thus quantifies the proportion of the overall variance $\mathrm{Var}(Y)$ accounted for by the fact that $X = x$.

- This idea of 'explained' and 'unexplained' variance is apparent in the law of total variance:

$$\mathrm{Var}(Y) = \mathbb{E}\big[\mathrm{Var}(Y|X)\big] + \mathrm{Var}\big[\mathbb{E}(Y|X)\big]$$

  - $\mathrm{Var}(Y)$ is the total variance,
  - $\mathbb{E}\big[\mathrm{Var}(Y|X)\big]$ is the variance explained by X, and
  - $\mathrm{Var}\big[\mathbb{E}(Y|X)\big]$ is the variance not explained by X.

C.5 Mulivariate transformations

Definition 3.14
Let $h : \mathbb{R}^n \to \mathbb{R}^n$ be a transformation of n variables, and let $(x_1, x_2, \ldots, x_n) = h(y_1, y_2, \ldots, y_n)$. The **Jacobian** (or **Jacobian determinant**) of the transformation h is the determinant of its $n \times n$ matrix of partial derivatives:

$$J = \begin{vmatrix} \dfrac{\partial x_1}{\partial y_1} & \dfrac{\partial x_1}{\partial y_2} & \cdots & \dfrac{\partial x_1}{\partial y_n} \\ \dfrac{\partial x_2}{\partial y_1} & \dfrac{\partial x_2}{\partial y_2} & \cdots & \dfrac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial x_n}{\partial y_1} & \dfrac{\partial x_n}{\partial y_2} & \cdots & \dfrac{\partial x_n}{\partial y_n} \end{vmatrix}$$

Theorem 3.15
Let $X = (X_1, X_2, \ldots, X_n)$ be a vector of continuous random variables, let $f_X(x_1, x_2 \ldots, x_n)$ be their joint PDF, let $g : \mathbb{R}^n \to \mathbb{R}^n$ be an injective transformation, and let $Y = (Y_1, Y_2, \ldots, Y_n)$ be the continuous random vector defined by $Y = g(X)$, i.e.

$$(Y_1, Y_2, \ldots, Y_n) = g(X_1, X_2, \ldots, X_n)$$

If the Jacobian of the inverse transformation g^{-1} is continuous and non-zero over the range of the transformation, the joint PDF of Y is

$$f_Y(y_1, y_2, \ldots, y_n) = |J| f_X(x_1, x_2, \ldots, x_n)$$

where

$$(x_1, x_2, \ldots, x_n) = g^{-1}(y_1, y_2, \ldots, y_n)$$

is the solution of $(y_1, y_2, \ldots, y_n) = g(x_1, \ldots, x_n)$.

Remark 3.16
The scale factor $|J|$ ensures that $f_Y(y_1, y_2, \ldots, y_n)$ integrates to one.

C.6 The multivariate normal distribution

The bivariate normal distribution can be formulated using matrix notation. Let

$$\mathbf{Z} = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

The matrix Σ is called the **variance-covariance** matrix.

- The determinant of Σ is
$$|\Sigma| = (1-\rho^2)\sigma_1^2\sigma_2^2.$$
- The inverse of Σ is
$$\Sigma^{-1} = \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix}.$$

Let $\mathbf{z} = (x, y)^T$ and consider the quadratic form

$$\begin{aligned}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu) &= \frac{1}{(1-\rho^2)\sigma_1^2\sigma_2^2} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{pmatrix} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix} \\ &= \frac{1}{(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\end{aligned}$$

The PDF $f(x, y)$ of the bivariate normal distribution can therefore be written as

$$f(x, y) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right).$$

Using this notation, we define the **multivariate normal distribution** as follows.

Definition 3.17
A random vector $\mathbf{Z} = (X_1, X_2, \ldots, X_n)^T$ is said to have **multivariate normal distribution** with mean vector $\mu = (\mu_1, \mu_2, \ldots, \mu_n)^T$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \ldots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_1\sigma_2 & \sigma_2^2 & \ldots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1}\sigma_1\sigma_n & \rho_{n2}\sigma_n\sigma_2 & \ldots & \sigma_n^2 \end{pmatrix}$$

if its joint density function can be written as

$$f(\mathbf{z}) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right)$$

where $\mu_i = \mathbb{E}(X_i)$, $\sigma_i^2 = \mathrm{Var}(X_i)$ and $\rho_{ij} = \dfrac{\mathbb{E}(X_iX_j)}{\sigma_i\sigma_j}$.

Exercise 3.18

1. Let X and Y have standard bivariate normal distribution, with joint PDF given by

   $$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$$

   where ρ is a constant satisfying $-1 < \rho < 1$.

   (a) Check that $f(x,y)$ is indeed a joint PDF, by verifying that $f(x,y) \geq 0$ and $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dxdy = 1$.
   Answer: TODO

   (b) Check that $\mathrm{Cov}(X,Y) = \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dxdy = \rho$.
   Answer: TODO

   (c) Show that if X and Y are uncorrelated, then they are independent.
   Answer: TODO

2. Let X and Y have standard bivariate normal distribution. Find the conditional distribution of Y given $X = x$, and hence show that $\mathbb{E}(Y|X) = \rho X$.

   Answer: The conditional distribution of Y given $X = x$ is $N(\rho x, 1-\rho^2)$.

3. Let X and Y have standard bivariate normal distribution. Show that X and $Z = \dfrac{Y - \rho X}{\sqrt{1-\rho^2}}$ are independent standard normal random variables.

   Answer: TODO

4. Let X and Y have standard bivariate normal distribution, and let $Z = \max\{X, Y\}$. Show that $\mathbb{E}(Z) = \sqrt{(1-\rho)/\pi}$ and $\mathbb{E}(Z^2) = 1$.

   Answer: TODO

5. Let $U, V \sim N(0,1)$. Show that the random variables $X = U + V$ and $Y = U - V$ are independent.

   Answer:

   - The transformation is $g(u,v) = (u+v, u-v)$.

- To compute the inverse transformation, consider $x = u + v$ and $y = u - v$.
- Solving these, we obtain $u = \frac{1}{2}(x + y)$ and $v = \frac{1}{2}(x - y)$.
- Thus the inverse transformation is $(u, v) = g^{-1}(x, y) = \big(\frac{1}{2}(x + y), \frac{1}{2}(x - y)\big)$

The Jacobian determinant of $g^{-1}(x, y)$ is

$$J = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} = \begin{vmatrix} [r]\frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

The joint PDF of U and V is

$$f(u, v) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{u^2 + v^2 - 2\rho uv}{2(1 - \rho^2)}\right)$$

Now,

$$\begin{aligned} u^2 + v^2 - 2\rho uv &= \left(\frac{1}{2}(x + y)\right)^2 + \left(\frac{1}{2}(x - y)\right)^2 - 2\rho\left(\frac{1}{2}(x + y)\right)\left(\frac{1}{2}(x - y)\right) \\ &= \frac{1}{2}x^2(1 - \rho) + \frac{1}{2}y^2(1 + \rho) \end{aligned}$$

The joint PDF of X and Y is therefore

$$\begin{aligned} f(x, y) &= \frac{1}{2}\frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{x^2(1 - \rho)}{4(1 - \rho^2)} - \frac{y^2(1 + \rho)}{4(1 - \rho^2)}\right) \\ &= \frac{1}{\sqrt{4\pi(1 + \rho)}} \exp\left(-\frac{x^2}{4(1 + \rho)}\right) \times \frac{1}{\sqrt{4\pi(1 - \rho)}} \exp\left(-\frac{y^2}{4(1 - \rho)}\right) \end{aligned}$$

This is the product of the PDF of a $N\big(0, 2(1 + \rho)\big)$ variable and the PDF of a $N\big(0, 2(1 - \rho)\big)$ variable. Thus X and Y are independent.

6. Let X and Y have bivariate normal distribution with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and correlation ρ. Show that the conditional distribution of Y given $X = x$ is

$$N\left(\mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(x - \mu_1), \sigma_2^2(1 - \rho^2)\right).$$

Answer: TODO

7. (a) Let X and Y be jointly continuous random variables, and let $f_{X,Y}$ be their joint PDF. Show that the PDF of the random variable $X + Y$ can be written as

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_{X,Y}(x, t - x)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(t - y, y)\,dy.$$

Answer: Let $A = \{(x, y) : x + y \leq z\} \subset \mathbb{R}^2$. Then

$$\mathbb{P}(X + Y \leq z) = \iint_A f(x, y)\,dxdy = \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{z-x} f_{X,Y}(x, y)\,dy\,dx$$

We change the variable of integration (in the inner integral), making the substitution $y = t - x$:

$$\begin{aligned} F_{X+Y}(z) = \mathbb{P}(X + Y \leq z) &= \int_{x=-\infty}^{\infty}\int_{t=-\infty}^{z} f_{X,Y}(x, t - x)\,dt\,dx \\ &= \int_{t=-\infty}^{z}\int_{x=-\infty}^{\infty} f_{X,Y}(x, t - x)\,dx\,dt \end{aligned}$$

where the final equality follows by reversing the order of integration. Thus the PDF of $X+Y$ is

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_{X,Y}(x, t-x)\,dx \qquad \text{as required.}$$

(b) Hence, or otherwise, show that if $U, V \sim N(0,1)$ are independent, then $U+V \sim N(0,2)$. (This is a special case of Theorem 3.5.)

Answer: By part (a), if two random variables X and Y are independent, the PDF of $X+Y$ is the **convolution** of the marginal PDFs:

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_X(x) f_Y(t-x)\,dx = \int_{-\infty}^{\infty} f_X(t-y) f_Y(y)\,dy.$$

U and V are independent, so their joint PDF is

$$f(u,v) = f_U(u) f_V(v) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(u^2+v^2)\right) \qquad u, v \in \mathbb{R}.$$

Let $W = U+V$. Then because U and V are independent,

$$\begin{aligned} f_W(w) &= \int_{-\infty}^{\infty} f_U(u) f_V(w-u)\,du \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2}\left(u^2+(w-u)^2\right)\right] du \\ &= \frac{1}{2\pi} e^{-\frac{1}{4}w^2} \int_{-\infty}^{\infty} \exp\left[-\left(u-\frac{w}{2}\right)^2\right] du \end{aligned}$$

We change the variable of integration, by making the substitution $t = \sqrt{2}\left(u - \dfrac{w}{2}\right)$:

$$f_W(w) = \frac{1}{2\sqrt{\pi}} e^{-\frac{1}{4}w^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}w^2}\,dv = \frac{1}{2\sqrt{\pi}} e^{-\frac{w^2}{4}},$$

which is the PDF of the $N(0,2)$ distribution