Cardiff University School of Mathematics

MA2500

Foundations of Probability and Statistics

Reading Material

2014-15

Version: 30/1/2015

Contents

1 Set Theory **1**
1.1 Elementary set theory . . . 1
1.2 Sample space, outcomes and events . . . 1
1.3 Countable unions and intersections . . . 2
1.4 Collections of sets . . . 3
1.5 Borel sets . . . 4
1.6 Exercises . . . 5

2 Probability Spaces **6**
2.1 Probability measures . . . 6
2.2 Null and almost-certain events . . . 6
2.3 Properties of probability measures . . . 7
2.4 Continuity of probability measures . . . 8
2.5 Exercises . . . 9

3 Conditional Probability **10**
3.1 Conditional probability . . . 10
3.2 Bayes' theorem . . . 10
3.3 Independence . . . 10
3.4 Conditional probability spaces . . . 11
3.5 Exercises . . . 13

4 Random Variables **14**
4.1 Random variables . . . 14
4.2 Indicator variables . . . 15
4.3 Simple random variables . . . 15
4.4 Probability on $\mathbb{R}$ . . . 15
4.5 Exercises . . . 16

5 Distributions **17**
5.1 Probability on the real line . . . 17
5.2 Cumulative distribution functions (CDFs) . . . 17
5.3 Properties of CDFs . . . 18
5.4 Discrete distributions and PMFs . . . 19
5.5 Continuous distributions and PDFs . . . 20
5.6 Exercises . . . 20

6 Transformations **22**
6.1 Transformations of random variables . . . 22
6.2 Support . . . 23
6.3 Transformations of CDFs . . . 23
6.4 Transformations of PMFs and PDFs . . . 24
6.5 The probability integral transform . . . 24

6.6 Exercises 25

7 Examples of transformations 27

7.1 Standard normal CDF $\longrightarrow$ Normal CDF 27
7.2 Standard normal CDF $\longrightarrow$ Chi-squared CDF 27
7.3 Standard uniform CDF $\longrightarrow$ Exponential CDF 27
7.4 Exponential CDF $\longrightarrow$ Pareto CDF 28
7.5 Normal PDF $\longrightarrow$ Standard normal PDF 28
7.6 Pareto PDF $\longrightarrow$ Standard uniform PDF 29
7.7 Normal PDF $\longrightarrow$ Lognormal PDF 29
7.8 Lomax PDF $\longrightarrow$ Logistic CDF 30
7.9 Exercises 30

8 Series and Integrals 32

8.1 Motivation 32
8.2 Series 32
8.3 Integrals 34

9 Expectation 37

9.1 Indicator variables 37
9.2 Simple random variables 38
9.3 Non-negative random variables 40
9.4 Signed random variables 41
9.5 Exercises 41

10 Computation of Expectation 42

10.1 Expectation with respect to CDFs 42
10.2 Discrete distributions 43
10.3 Continuous distributions 44
10.4 Transformed variables 45
10.5 Exercises 46

11 Concentration Inequalities 47

11.1 Markov's inequality 47
11.2 Chebyshev's inequality 48
11.3 Bernstein's inequality 49
11.4 Exercises 50

12 Probability Generating Functions 51

12.1 Generating functions 51
12.2 Probability generating functions 52
12.3 Exercises 54

13 Moment Generating Functions 55

13.1 Moment generating functions 55
13.2 Characteristic functions 57
13.3 Exercises 58

14 The Law of Large Numbers 60

14.1 Convergence 60
14.2 The law of large numbers 60
14.3 Bernoulli's law of large numbers* 61
14.4 Exercises 62

15 The Central Limit Theorem **63**
15.1 Poisson limit theorem . . . 63
15.2 Law of large numbers . . . 64
15.3 Central limit theorem . . . 65
15.4 Exercises . . . 66

16 Joint Distributions **68**
16.1 Joint distributions . . . 68
16.2 Properties of Joint CDFs . . . 68
16.3 Independent random variables . . . 69
16.4 Identically distributed random variables . . . 69
16.5 Jointly discrete distributions . . . 69
16.6 Jointly continuous distributions . . . 70
16.7 Exercises . . . 72

17 Covariance and Correlation **74**
17.1 Bivariate Distributions . . . 74
17.2 Covariance . . . 74
17.3 Correlation . . . 75
17.4 The Cauchy-Schwarz Inequality . . . 76
17.5 Exercises . . . 77

18 Conditional Distributions **78**
18.1 Conditional distributions . . . 78
18.2 Conditional expectation . . . 79
18.3 Law of total expectation . . . 80
18.4 Law of total variance . . . 80
18.5 Example . . . 81
18.6 Exercises . . . 83

19 The Bivariate Normal Distribution **84**
19.1 Bivariate transformations . . . 84
19.2 The bivariate normal distribution . . . 85
19.3 Properties of the bivariate normal distribution . . . 87
19.4 Conditional distributions . . . 88
19.5 Exercises . . . 89

Lecture 1 Set Theory

1.1 Elementary set theory

A set is a collection of distinct *elements*.

- If a is an element of the set A, we denote this by $a \in A$.
- If a is *not* an element of A, we denote this by $a \notin A$.
- The *cardinality* of a set is the number of elements it contains.
- The *empty set* contains no elements, and is denoted by $\emptyset$.

1.1.1 Set relations

Let A, B be sets.

- If $a \in B$ for every $a \in A$, we say that A is a *subset* of B, denoted by $A \subseteq B$.
- If $A \subseteq B$ and $B \subseteq A$, we say that A is *equal* to B, denoted by $A = B$,
- If $A \subseteq B$ and $A \neq B$, we say that A is a *proper subset* of B, denoted by $A \subset B$.

1.1.2 Set operations

Let A, B and Ω be sets, with $A, B \subseteq \Omega$.

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

1.1.3 Set algebra

Commutative property:
$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associative property:
$$(A \cup B) \cup C = A \cup (B \cup C)$$
$$(A \cap B) \cap C = A \cap (B \cap C)$$

Distributive property:
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

1.2 Sample space, outcomes and events

Definition 1.1

(1) Any process of observation or measurement whose outcome is uncertain is called a *random experiment*.

(2) A random experiment has a number of possible *outcomes*.

(3) Each time a random experiment is performed, *exactly one* of its outcomes will occur.

(4) The set of all possible outcomes is called the *sample space*, denoted by Ω.

(5) Outcomes are also called *elementary events*, and denoted by $\omega \in \Omega$.

Example 1.2

- $\{1, 2, \ldots, n\}$ is a finite sample space,
- $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ is a countably infinite sample space,
- $[0, 1]$ is an uncountable sample space.

Definition 1.3

(1) An *event* A is a subset of the sample space, Ω.

(2) If outcome ω occurs, we say that event A *occurs* if and only if $\omega \in A$.

(3) Two events A and B with $A \cap B = \emptyset$ are called *disjoint* or *mutually exclusive*.

(4) The empty set $\emptyset$ is called the *impossible event*.

(5) The sample space itself is called the *certain event*.

Remark 1.4

- If A occurs and $A \subseteq B$, then B occurs.
- If A occurs and $A \cap B = \emptyset$, then B does not occur.

1.3 Countable unions and intersections

Definition 1.5
Let Ω be any set. The set of all subsets Ω is called its *power set*.

- If Ω is a finite set, its power set is also finite.
- If Ω is a countably infinite set, its power set is uncountable set (Cantor's Theorem).
- If Ω is an uncountable set, its power set is also uncountable.

Definition 1.6
Let $A_1, A_2, \ldots$ be subsets of Ω.

(1) The (countable) *union* of $A_1, A_2, \ldots$ is the set

$$\bigcup_{i=1}^{\infty} A_i = \{\omega : \omega \in A_i \text{ for some } A_i\}.$$

(2) The (countable) *intersection* of $A_1, A_2, \ldots$ is the set

$$\bigcap_{i=1}^{\infty} A_i = \{\omega : \omega \in A_i \text{ for all } A_i\}.$$

Theorem 1.7 (De Morgan's laws)
For a countable collection of sets $\{A_1, A_2, \ldots\}$,

(1) $\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c$,

(2) $\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c$.

Proof:

1.4 Collections of sets

Definition 1.8
Let Ω be any set. Any subset of its power set is called a *collection of sets over* Ω.

Let Ω be the sample space of some random experiment. If we are interested whether the events A and B occur, we must also be interested in

- the event $A \cup B$: whether event A occurs *or* event B occurs;
- the event $A \cap B$: whether event A occurs *and* event B occurs;
- the event A^c: whether the event A does *not* occur.

Thus we can not use arbitrary collections of sets over Ω as the basis for investigating random experiments. Instead, we allow only collections which are *closed* under certain set operations.

Definition 1.9
A collection of sets $\mathcal{C}$ over Ω is said to be

(1) *closed under complementation* if $A^c \in \mathcal{C}$ for every $A \in \mathcal{C}$,

(2) *closed under pairwise unions* if $A \cup B \in \mathcal{C}$ for every $A, B \in \mathcal{C}$,

(3) *closed under finite unions* if $\bigcup_{i=1}^{n} A_i \in \mathcal{C}$ for every $A_1, A_2, \ldots A_n \in \mathcal{C}$,

(4) *closed under countable unions* if $\bigcup_{i=1}^{\infty} A_i \in \mathcal{C}$ for every $A_1, A_2, \ldots \in \mathcal{C}$.

Definition 1.10
A collection of sets $\mathcal{F}$ over Ω is called a *field* over Ω if

(1) $\Omega \in \mathcal{F}$,

(2) $\mathcal{F}$ is closed under complementation, and

(3) $\mathcal{F}$ is closed under pairwise unions.

Theorem 1.11 (Properties of fields)
Let $\mathcal{F}$ be a field over Ω. Then

(1) $\emptyset \in \mathcal{F}$,

(2) $\mathcal{F}$ is closed under set differences,

(3) $\mathcal{F}$ is closed under finite unions,

(4) $\mathcal{F}$ is closed under finite intersections.

Proof:

Definition 1.12
A collection of sets $\mathcal{F}$ over Ω is called a *σ-field* ("sigma-field") over Ω if

(1) $\Omega \in \mathcal{F}$,

(2) $\mathcal{F}$ is closed under complementation, and

(3) $\mathcal{F}$ is closed under countable unions.

Theorem 1.13 (Properties of σ-fields)
Let $\mathcal{F}$ be a σ-field over Ω. Then

(1) $\emptyset \in \mathcal{F}$,

(2) $\mathcal{F}$ is closed under set differences,

(3) $\mathcal{F}$ is closed under finite unions,

(4) $\mathcal{F}$ is closed under finite intersections,

(5) $\mathcal{F}$ is closed under countable intersections.

Proof:

1.5 Borel sets

In many situations of interest, random experiments yield outcomes that are *real numbers*.

Definition 1.14

- The *open interval* (a, b) is the set $\{x \in \mathbb{R} : a < x < b\}$.
- The *closed interval* $[a, b]$ is the set $\{x \in \mathbb{R} : a \leq x \leq b\}$.

Definition 1.15
The *Borel* σ-field over $\mathbb{R}$ is defined to be the smallest σ-field over $\mathbb{R}$ that contains all open intervals.

Remark 1.16

- The Borel σ-field is usually denoted by $\mathcal{B}$, and includes all closed interval, all half-open intervals, all finite sets and all countable sets.
- The elements of $\mathcal{B}$ are called *Borel sets* over $\mathbb{R}$.
- Borel sets can be thought of as the "nice" subsets of $\mathbb{R}$.

Proposition 1.17
The Borel σ-field over $\mathbb{R}$ contains all closed intervals.

Proof:

1.6 Exercises

Exercise 1.1

1. Let $\mathcal{F}$ be a field over Ω. Show that
   (a) $\emptyset \in \mathcal{F}$,
   (b) $\mathcal{F}$ is closed under set differences,
   (c) $\mathcal{F}$ is closed under pairwise intersections,
   (d) $\mathcal{F}$ is closed under finite unions,
   (e) $\mathcal{F}$ is closed under finite intersections.
2. Let $\mathcal{F}$ be a σ-field over Ω. Show that
   (a) $\mathcal{F}$ is closed under finite unions,
   (b) $\mathcal{F}$ is closed under finite intersections.
   (c) $\mathcal{F}$ is closed under countable intersections.

Exercise 1.2

1. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$.
   (a) What is the smallest σ-field containing the event $A = \{1, 2\}$?
   (b) What is the smallest σ-field containing the events $A = \{1, 2\}$, $B = \{3, 4\}$ and $C = \{5, 6\}$?
2. Let $\mathcal{F}$ and $\mathcal{G}$ be σ-fields over Ω.
   (a) Show that $\mathcal{H} = \mathcal{F} \cap \mathcal{G}$ is a σ-field over Ω.
   (b) Find a counterexample to show that $\mathcal{H} = \mathcal{F} \cup \mathcal{G}$ is not necessarily a σ-field over Ω.

Lecture 2 Probability Spaces

2.1 Probability measures

Definition 2.1
Let Ω be a sample space, and let $\mathcal{F}$ be a σ-field over Ω. A *probability measure* on $(\Omega, \mathcal{F})$ is a function

$$\begin{array}{rccc} \mathbb{P}: & \mathcal{F} & \to & [0,1] \\ & A & \mapsto & \mathbb{P}(A) \end{array}$$

such that $\mathbb{P}(\Omega) = 1$, and for any countable collection of pairwise disjoint events $\{A_1, A_2, \ldots\}$,

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

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

Remark 2.2
- The second property is called *countable additivity*.

Remark 2.3
In the more general setting of measure theory:

- The elements of $\mathcal{F}$ are called *measurable sets*.
- The pair $(\Omega, \mathcal{F})$ is called a *measurable space*.
- The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a *measure space*.

Example 2.4
A fair six-sided die is rolled once. A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for the experiment is given by

- $\Omega = \{1, 2, 3, 4, 5, 6\}$,
- $\mathcal{F} = \mathcal{P}(\Omega)$, where $\mathcal{P}(\Omega)$ denotes the power set of Ω,
- $\mathbb{P}(A) = |A|/|\Omega|$ for every $A \in \mathcal{F}$ (where $|A|$ denotes the cardinality of A).

If we are only interested in odd and even numbers, we can instead take

- $\Omega = \{1, 2, 3, 4, 5, 6\}$,
- $\mathcal{F} = \big\{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}\big\}$
- $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\{1, 3, 5\}) = 1/2$, $\mathbb{P}(\{2, 4, 6\}) = 1/2$, $\mathbb{P}(\{1, 2, 3, 4, 5, 6\}) = 1$.

2.2 Null and almost-certain events

Definition 2.5
(1) If $\mathbb{P}(A) = 0$, we say that A is a *null event*.

(2) If $\mathbb{P}(A) = 1$, we say that A occurs *almost surely* (or "*with probability 1*").

Remark 2.6

- A null event is not the same as the impossible event ($\emptyset$).
- An event that occurs almost surely is not the same as the certain event (Ω).

Example 2.7
A dart is thrown at a dartboard.

- The probability that the dart hits a given point of the dartboard is 0.
- The probability that the dart does not hit a given point of the dartboard is 1.

2.3 Properties of probability measures

Theorem 2.8 (Properties of probability measures)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A, B \in \mathcal{F}$.

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

(2) $\mathbb{P}(\emptyset) = 0$,

(3) Monotonicity: if $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

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

Proof:

2.4 Continuity of probability measures

Theorem 2.9 (Continuity of probability measures)
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.

(1) For an increasing sequence of events $A_1 \subseteq A_2 \subseteq \ldots$ in $\mathcal{F}$,

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n).$$

(2) For a decreasing sequence of events $B_1 \supseteq B_2 \supseteq \ldots$ in $\mathcal{F}$,

$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} B_n\right) = \lim_{n\to\infty} \mathbb{P}(B_n).$$

Proof:

2.5 Exercises

Exercise 2.1

1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $A, B, C \in \mathcal{F}$. Show that

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

   This is called the *inclusion-exclusion principle.*

2. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.
   (a) Show that $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ for all $A, B \in \mathcal{F}$. This is called *subadditivity.*
   (b) Show that for any sequence $A_1, A_2, \ldots$ of events in $\mathcal{F}$,

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

   This is called *countable subadditivity.*

Exercise 2.2

1. Let A and B be events with probabilities $\mathbb{P}(A) = 3/4$ and $\mathbb{P}(B) = 1/3$.
   (a) Show that $\frac{1}{12} \leq \mathbb{P}(A \cap B) \leq \frac{1}{3}$, and construct examples to show that both extremes are possible.
   (b) Find corresponding bounds for $\mathbb{P}(A \cup B)$.

2. A roulette wheel consists of 37 slots of equal size. The slots are numbered from 0 to 36, with odd-numbered slots coloured red, even-numbered slots coloured black, and the slot labelled 0 coloured green. The wheel is spun in one direction and a ball is rolled in the opposite direction along a track running around the circumference of the wheel. The ball eventually falls on to the wheel and into one of the 37 slots. A player bets on the event that the ball stops in a red slot, and another player bets on the event that the ball stops in a black slot.
   (a) Define a suitable sample space Ω for this random experiment, and identify the events of interest.
   (b) Find the smallest field $\mathcal{F}$ over Ω that contains the events of interest.
   (c) Define a suitable probability measure $(\Omega, \mathcal{F})$ to represent the game.

Exercise 2.3

1. A biased coin has probability p of showing heads. The coin is tossed repeatedly until a head occurs. Describe a suitable probability space for this experiment.

2. A fair coin is tossed repeatedly.
   (a) Show that a head eventually occurs with probability one.
   (b) Show that a sequence of 10 consecutive tails eventually occurs with probability one.
   (c) Show that any finite sequence of heads and tails eventually occurs with probability one.

Lecture 3 Conditional Probability

3.1 Conditional probability

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $B \in \mathcal{F}$.

Definition 3.1
If $\mathbb{P}(B) > 0$, the *conditional probability of A given B* is defined to be

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

3.2 Bayes' theorem

Definition 3.2
A countable collection of sets $\{A_1, A_2, \ldots\}$ is said to form a *partition* of a set B if

(1) $A_i \cap A_j = \emptyset$ for all $i \neq j$, and

(2) $B \subseteq \bigcup_{i=1}^{\infty} A_i$.

Theorem 3.3 (The Law of Total Probability)
If $\{A_1, A_2, \ldots\}$ is a partition of B, then

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

Theorem 3.4 (Bayes' Theorem)
If $\{A_1, A_2, \ldots\}$ is a partition of B where $\mathbb{P}(B) > 0$, then

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

3.3 Independence

Definition 3.5
Two events A and B are said to be *independent* if $\mathbb{P}(A|B) = \mathbb{P}(A)$, or equivalently,

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

Definition 3.6
A collection of events $\{A_1, A_2, \ldots\}$ is said to be

(1) *pairwise independent* if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$.

(2) *totally independent* if, for every finite subset $\{B_1, B_2, \ldots, B_m\} \subset \{A_1, A_2, \ldots\}$,

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

This can also be written as $\mathbb{P}\left(\bigcap_{j=1}^{m} B_j\right) = \prod_{j=1}^{m} \mathbb{P}(B_j)$.

Remark 3.7
Total independence implies pairwise independence, but not vice versa.

3.4 Conditional probability spaces

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $B \in \mathcal{F}$.

Theorem 3.8
The family of sets $\mathcal{G} = \{A \cap B : A \in \mathcal{F}\}$ is a σ-field over B.

Remark 3.9
$\mathcal{G}$ contains all sets of the form $A \cap B$, where A is some element of $\mathcal{F}$. This means that $A' \in \mathcal{G}$ if and only if there is some $A \in \mathcal{F}$ for which $A' = A \cap B$.

Proof:

Theorem 3.10
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $B \in \mathcal{F}$, and let $\mathcal{G} = \{A \cap B : A \in \mathcal{F}\}$.
If $\mathbb{P}(B) > 0$, then

$$\begin{array}{rccc} \mathbb{Q}: & \mathcal{G} & \to & [0,1] \\ & A' & \mapsto & \mathbb{P}(A'|B) \end{array}$$

is a probability measure on $(B, \mathcal{G})$.

Remark 3.11
$(B, \mathcal{G}, \mathbb{Q})$ is called a *conditional probability space.*.

Proof:

Remark 3.12
We have shown that $\mathbb{Q}$ is a probability measure on $(B, \mathcal{G})$. Using an almost identical argument, it can be shown that $\mathbb{Q}$ is also a probability measure on $(\Omega, \mathcal{F})$.

- In the probability space $(B, \mathcal{G}, \mathbb{Q})$, outcomes $\omega \notin B$ are excluded from consideration.
- In the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$, outcomes $\omega \notin B$ are assigned probability zero.

3.5 Exercises

Exercise 3.1 [Revision]

1. Let Ω be a sample space, and let $A_1, A_2, \ldots$ be a partition of Ω with the property that $\mathbb{P}(A_i) > 0$ for all i.

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

   (b) Show that $\mathbb{P}(A_i|B) = \dfrac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$.

Exercise 3.2

1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$, and consider the function $\mathbb{Q} : \mathcal{F} \to [0,1]$ defined by $\mathbb{Q}(A) = \mathbb{P}(A|B)$.
   (a) Show that $(\Omega, \mathcal{F}, \mathbb{Q})$ is a probability space.
   (b) If $C \in \mathcal{F}$ and $\mathbb{Q}(C) > 0$, show that $\mathbb{Q}(A|C) = \mathbb{P}(A|B \cap C)$.

2. A random number N of dice are rolled. Let A_k be the event that $N = k$, and suppose that $\mathbb{P}(A_k) = 2^{-k}$ for $k \in \{1, 2, \ldots\}$ (and zero otherwise). Let S be the sum of the scores shown on the dice. Find the probability that:
   (a) $N = 2$ given that $S = 4$,
   (b) $S = 4$ given that N is even,
   (c) $N = 2$ given that $S = 4$ and the first die shows 1,
   (d) the largest number shown by any dice is r (where S is unknown).

3. Let $\Omega = \{1, 2, \ldots, p\}$ where p is a prime number. Let $\mathcal{F}$ be the power set of Ω, and let $\mathbb{P} : \mathcal{F} \to [0,1]$ be the probability measure on $(\Omega, \mathcal{F})$ defined by $\mathbb{P}(A) = |A|/p$, where $|A|$ denotes the cardinality of A. Show that if A and B are independent events, then at least one of A and B is either $\emptyset$ or Ω.

Lecture 4 Random Variables

4.1 Random variables

Random variables are functions that transform abstract sample spaces to the real numbers.

Definition 4.1
Let Ω be the sample space of some random experiment, and let $\mathcal{F}$ be a σ-field of events over Ω. A *random variable* on $(\Omega, \mathcal{F})$ is a function

$$\begin{aligned} X: \quad \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

with the property that $\{\omega : X(\omega) \in B\} \in \mathcal{F}$ for every $B \in \mathcal{B}$, where $\mathcal{B}$ is the Borel σ-field over $\mathbb{R}$.

Remark 4.2

- The set $\{\omega : X(\omega) \in B\}$ contains precisely those outcomes that are mapped by X into the set B.
- X is a random variable only if every set of this form is an element of the σ-field $\mathcal{F}$.
- This condition means that, for any Borel set B, the probability that X takes a value in B is well-defined.

Let us define the following notation:

$$\{X \in B\} = \{\omega : X(\omega) \in B\}$$

- The expression $\{X \in B\}$ should not be taken literally: X is a function, while B is a subset of the real numbers.
- Instead, think of $\{X \in B\}$ as the event that X takes a value in B.
- The condition $\{X \in B\} \in \mathcal{F}$ ensures that the probability of this event is well-defined.

We denote the probability of $\{X \in B\}$ by $\mathbb{P}(X \in B)$, by which we mean

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

Proposition 4.3
A function $X : \Omega \to \mathbb{R}$ is a random variable if and only if $\{X \leq x\} \in \mathcal{F}$ for every $x \in \mathbb{R}$.

[*Proof omitted.*]

Remark 4.4
To check whether or not a function $X : \Omega \to \mathbb{R}$ is a random variable, by the proposition we need not verify that $\{X \in B\} \in \mathcal{F}$ for all Borel sets $B \in \mathcal{B}$. Instead, it is enough to verify only that the sets $\{\omega : X(\omega) \leq x\}$ are included in $\mathcal{F}$ (for every $x \in \mathbb{R}$).

4.2 Indicator variables

The elementary random variable is the *indicator variable* of an event A.

Definition 4.5
The *indicator variable* of an event A is the random variable $I_A : \Omega \to \mathbb{R}$ defined by

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

Theorem 4.6
Let A and B be any two events. Then

(1) $I_{A^c} = 1 - I_A$

(2) $I_{A\cap B} = I_A I_B$

(3) $I_{A\cup B} = I_A + I_B - I_{A\cap B}$

Proof:

4.3 Simple random variables

Definition 4.7
A *simple random variable* is one that takes only finitely many values.

If $X : \Omega \to \mathbb{R}$ is a simple random variable, it can be represented as:

$$X(\omega) = \sum_{i=1}^{n} a_i I_{A_i}(\omega)$$

where

- $\{a_1, a_2, \ldots, a_n\} \subset \mathbb{R}$ is the range of X, and
- $\{A_1, A_2, \ldots, A_n\}$ is a partition of the sample space, Ω.

4.4 Probability on $\mathbb{R}$

Definition 4.8
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a random variable on $(\Omega, \mathcal{F})$. The function

$$\begin{aligned} \mathbb{P}_X : \quad \mathcal{B} &\to [0, 1] \\ B &\mapsto \mathbb{P}(X \in B). \end{aligned}$$

is called the *distribution* of X.

Theorem 4.9
$\mathbb{P}_X$ is a probability measure on $(\mathbb{R}, \mathcal{B})$.

Proof:

Remark 4.10
A random variable X transforms an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ into a more tractable probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$, where we can apply the methods of *real analysis*.

4.5 Exercises

Exercise 4.1

1. Let Ω be the sample space of some random experiment, and let $\mathcal{F}$ be a σ-field over Ω.
   (a) For any $A \in \mathcal{F}$, show that the function $X : \Omega \to \mathbb{R}$, defined by

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

   is a random variable on $(\Omega, \mathcal{F})$.
   (b) Let $A_1, A_2, \ldots, A_n \in \mathcal{F}$ be a partition of Ω and let $a_1, a_2, \ldots, a_n \in \mathbb{R}$. Show that the function $X : \Omega \to \mathbb{R}$, defined by

   $$X(\omega) = \sum_{i=1}^{n} a_i I_{A_i}(\omega) \qquad \text{where} \qquad I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A, \end{cases}$$

   is a random variable on $(\Omega, \mathcal{F})$.

Lecture 5 Distributions

5.1 Probability on the real line

Let $X : \Omega \to \mathbb{R}$ be a random variable, and recall the probability measure on $(\mathbb{R}, \mathcal{B})$, defined by

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

where $\mathcal{B}$ is the Borel σ-field over $\mathbb{R}$.

Definition 5.1

(1) The *distribution* of X is the probability measure $\mathbb{P}_X(B) = \mathbb{P}(X \in B)$.

(2) The *cumulative distribution function* (CDF) of X is the function $F(x) = \mathbb{P}(X \leq x)$.

(3) The *survival function* (SF) of X is the function $S(t) = \mathbb{P}(X > t)$.

Remark 5.2
The survival function is also called the *complementary* distribution function. If X represents the *lifetime* of some random system, then $S(t) = \mathbb{P}(X > t)$ is the probability that the system survives beyond time t. In this context, $F(t) = 1 - S(t)$ is called the *lifetime distribution function*.

5.2 Cumulative distribution functions (CDFs)

Proposition 4.3 states that $X : \Omega \to \mathbb{R}$ is a random variable if and only if the sets $\{X \leq x\}$ are *events* over Ω:

$$\{X \leq x\} = \big\{\omega : X(\omega) \leq x\big\} \in \mathcal{F} \quad \text{for all} \quad x \in \mathbb{R}.$$

It can be shown that the probability measure

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

is uniquely defined by the values it takes on the events $\{X \leq x\}$ for $x \in \mathbb{R}$. Consequently, the distribution of a random variable is uniquely determined by its *cumulative distribution function* (CDF):

Definition 5.3
The *cumulative distribution function* (CDF) of a random variable $X : \Omega \to \mathbb{R}$ is the function

$$\begin{aligned} F : \quad \mathbb{R} &\longrightarrow [0, 1] \\ x &\mapsto \mathbb{P}(X \leq x). \end{aligned}$$

Theorem 5.4
Let $F : \mathbb{R} \to [0, 1]$ be a CDF. Then there is a unique probability measure $\mathbb{P}_F : \mathcal{B} \to [0, 1]$ on the real line with the property that

$$\mathbb{P}_F\big((a, b]\big) = F(b) - F(a)$$

for every such half-open interval $(a, b] \in \mathcal{B}$.

[*Proof omitted.*]

- The triple $(\mathbb{R}, \mathcal{B}, \mathbb{P}_F)$ is sometimes called the *probability space induced by F*.

Remark 5.5
Compare the probability measure $\mathbb{P}_F$ of the interval $(a, b] \subset \mathbb{R}$ to the usual measure of its *length*:

- Length: $\mathbb{L}\big((a, b]\big) = b - a$
- Probability measure: $\mathbb{P}_F\big((a, b]\big) = F(b) - F(a)$.

Thus $\mathbb{P}_F\big((a, b]\big)$ quantifies the "amount of probability" in any given interval $(a, b]$.

5.3 Properties of CDFs

Theorem 5.6
A cumulative distribution function $F : \mathbb{R} \to [0, 1]$ has the following properties:

(1) if $x < y$ then $F(x) \leq F(y)$,

(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:

Theorem 5.7
Let $F : \mathbb{R} \to [0, 1]$ be a function with properties (i)-(iv) of Theorem 5.6. Then F is a cumulative distribution function.

[*Proof omitted.*]

Remark 5.8
The last two theorems make no explicit reference to random variables:

- many different random variables can have the same distribution function;
- a distribution function can represent many different random variables.

5.4 Discrete distributions and PMFs

The *range* of a random variable $X : \Omega \to \mathbb{R}$ is the set of all possible values it can take:

$$\text{Range}(X) = \{x \in \mathbb{R} : X(\omega) = x \text{ for some } \omega \in \Omega\}.$$

Definition 5.9

- $X : \Omega \to \mathbb{R}$ is called a *discrete random variable* if its range is a countable subset of $\mathbb{R}$.
- A discrete random variable is described by its *probability mass function* (PMF),

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

  which must have the property that $\sum_k f(k) = 1$.
- A probability mass function defines a *discrete probability measure* on $\mathbb{R}$,

$$\begin{array}{rccl} \mathbb{P}_X: & \mathcal{B} & \to & [0,1] \\ & B & \mapsto & \displaystyle\sum_{k\in B} \mathbb{P}(X=k), \end{array}$$

- The triple $(\mathbb{R}, \mathcal{B}, \mathbb{P}_X)$ is called a *discrete probability space* over $\mathbb{R}$.

5.5 Continuous distributions and PDFs

Definition 5.10

- A cumulative distribution function $F : \mathbb{R} \to [0,1]$ is said to be *absolutely continuous* if there exists an integrable function $f : \mathbb{R} \to [0,\infty)$ such that

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

- The function $f : \mathbb{R} \to [0,\infty)$ is called the *probability density function* (PDF) of F.
- The triple $(\mathbb{R}, \mathcal{B}, \mathbb{P}_F)$ is called a *continuous probability space* over $\mathbb{R}$.

Definition 5.11

A *continuous random variable* is one whose distribution function is absolutely continuous.

If $X : \Omega \to \mathbb{R}$ is a continuous random variable, then

- $f(x) = F'(x)$ for all $x \in \mathbb{R}$.
- Probabilities correspond to areas under the curve $f(x)$:

$$\mathbb{P}_X\big((a,b]\big) = \mathbb{P}(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)\,dx.$$

- Note that $\mathbb{P}(X = x) = 0$ for all $x \in \mathbb{R}$.

Remark 5.12

The continuity of a random variable $X : \Omega \to \mathbb{R}$ refers to the continuity of its distribution function, and *not* to the continuity (or otherwise) of itself as a function on Ω.

5.6 Exercises

Exercise 5.1

1. Let F and G be CDFs, and let $0 < \lambda < 1$ be a constant. Show that $H = \lambda F + (1-\lambda)G$ is also a CDF.
2. Let X_1 and X_2 be the numbers observed in two independent rolls of a fair die. Find the PMF of each of the following random variables:
   (a) $Y = 7 - X_1$,

(b) $U = \max(X_1, X_2)$,

(c) $V = X_1 - X_2$.

(d) $W = |X_1 - X_2|$.

3. 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.

   (b) Find the value of $P(X > 3/2)$.

   (c) Find the CDF of X.

4. The PDF of a continuous random variable X is given by $f(x) = \begin{cases} cx^{-d} & \text{for } x > 1, \\ 0 & \text{otherwise.} \end{cases}$

   (a) Find the range of values of d for which $f(x)$ is a probability density function.

   (b) If $f(x)$ is a density function, find the value of c, and the corresponding CDF.

5. Let $f(x) = \dfrac{ce^x}{(1+e^x)^2}$ be a PDF, where c is a constant. Find the value of c, and the corresponding CDF.

6. Let $X_1, X_2, \ldots$ be independent and identically distributed observations, and let F denote their common CDF. If F is unknown, describe and justify a way of estimating F, based on the observations. [Hint: consider the indicator variables of the events $\{X_j \leq x\}$.]

Lecture 6 Transformations

6.1 Transformations of random variables

Let $\mathcal{B}$ denote the Borel σ-field over $\mathbb{R}$.

Definition 6.1
A function $g : \mathbb{R} \to \mathbb{R}$ is said to be a *measurable function* if $g^{-1}(B) \in \mathcal{B}$ for all $B \in \mathcal{B}$.

Theorem 6.2
Let $X : \Omega \to \mathbb{R}$ be a random variable on $(\Omega, \mathcal{F})$, and let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function. Then the function $Y = g(X)$, defined by

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

is also a random variable on $(\Omega, \mathcal{F})$.

Proof:

We say that $Y = g(X)$ is a *transformation* of X.

- If we know the distribution of X, how can we deduce the distribution of Y?

In fact, the distribution of $Y = g(X)$ is completely determined by the distribution of X:

$$\begin{aligned} \mathbb{P}_Y(B) = \mathbb{P}(Y \in B) &= \mathbb{P}\big[\{\omega : Y(\omega) \in B\}\big] \\ &= \mathbb{P}\big[\{\omega : g[X(\omega)] \in B\}\big] \\ &= \mathbb{P}\big[\{\omega : X(\omega) \in g^{-1}(B)\}\big] \\ &= \mathbb{P}\big[X \in g^{-1}(B)\big] \\ &= \mathbb{P}_X\left[g^{-1}(B)\right] \end{aligned}$$

Remark 6.3

- Theorem 6.2 shows that $g(X) : \Omega \to \mathbb{R}$ is a random variable over $(\Omega, \mathcal{F})$.
- We can also think of $g : \mathbb{R} \to \mathbb{R}$ as a random variable over $(\mathbb{R}, \mathcal{B})$, whose distribution is given by

  $$\mathbb{P}(g \in B) = \mathbb{P}\big[g(X) \in B\big] = \mathbb{P}\big[X \in g^{-1}(B)\big] = \mathbb{P}_X\left[g^{-1}(B)\right],$$

  where $\mathbb{P}_X : \mathcal{B} \to [0, 1]$ is the distribution of X.
- The distribution of g over $(\mathbb{R}, \mathcal{B})$ is well-defined, because $g^{-1}(B) \in \mathcal{B}$ for all $B \in \mathcal{B}$.

6.2 Support

PMFs and many PDFs are defined to be zero over certain subsets of $\mathbb{R}$. We must ensure that the PMF or PDF of a transformed variable is defined correctly, over appropriate subsets of $\mathbb{R}$.

Definition 6.4

(1) A set $A \subset \mathbb{R}$ is said to be *closed* if it contains all its limit points.

(2) 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 $x \notin \text{supp}(h)$.

(3) The *support* of a random variable $X : \Omega \to \mathbb{R}$ is defined to be the support of its PMF (discrete case) or PDF (continuous case), denoted by $\text{supp}(f_X)$. This is the smallest closed set that contains the *range* of X.

Remark 6.5

Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$. The support of $Y = g(X)$ is the set

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

Example 6.6

The PDF of the continuous uniform distribution on $[0, 1]$ is

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

- The support of X is $\text{supp}(f_X) = [0, 1]$
- For the transformation $g(x) = x^2 + 2x + 3$, the support of $Y = g(X)$ is

$$\text{supp}(f_Y) = \{x^2 + 2x + 3 : x \in [0, 1]\} = [3, 6].$$

6.3 Transformations of CDFs

Theorem 6.7

Let X be a continuous random variable, and let f_X denote its PDF. Let $g : \mathbb{R} \to \mathbb{R}$ be an injective transformation over $\text{supp}(f_X)$ and let $Y = g(X)$. Finally, let $F_X(x)$ and $F_Y(y)$ respectively denote the CDFs of X and Y.

(1) If g is an increasing function, $F_Y(y) = F_X\big[g^{-1}(y)\big]$.

(2) If g is a decreasing function, $F_Y(y) = 1 - F_X\big[g^{-1}(y)\big]$.

Proof:

6.4 Transformations of PMFs and PDFs

Theorem 6.8 (Transformations of PMFs)
Let X be a discrete random variable and let f_X denote its PMF. Let $g:\mathbb{R}\to\mathbb{R}$ be an injective transformation over $\text{supp}(f_X)$, and let $Y=g(X)$. Then the PMF of Y is given by

$$f_Y(y)=f_X\left[g^{-1}(y)\right] \quad \text{for all} \quad y\in\text{supp}(f_Y).$$

Proof:

Theorem 6.9 (Transformations of PDFs)
Let X be a continuous random variable and let f_X denote its PDF. Let $g:\mathbb{R}\to\mathbb{R}$ be an injective transformation over $\text{supp}(f_X)$, and let $Y=g(X)$. Then, if the derivative of $g^{-1}(y)$ is continuous and non-zero over $\text{supp}(f_Y)$, the PDF of Y is given by

$$f_Y(y)=f_X\left[g^{-1}(y)\right]\left|\frac{d}{dy}g^{-1}(y)\right| \quad \text{for all} \quad y\in\text{supp}(f_Y).$$

Remark 6.10
- The transformation is equivalent to making a *change of variable* in an integral.
- The scale factor $\left|\frac{d}{dy}g^{-1}(y)\right|$ ensures that $f_Y(y)$ integrates to one.

Proof:

6.5 The probability integral transform

Theorem 6.11 (The Probability Integral Transform)
Let X be a continuous random variable, let $F(x)$ denote its CDF, and suppose that the inverse F^{-1} of the CDF exists for all $x\in\mathbb{R}$. Then the random variable $Y=F(X)$ has the continuous uniform distribution on $[0,1]$.

Proof:

Corollary 6.12
Let $F(x)$ be a CDF whose inverse exists for all $x \in \mathbb{R}$, and let $Y \sim \text{Uniform}(0, 1)$. Then F is the CDF of the random variable $X = F^{-1}(Y)$.

- Uniformly distributed pseudo-random numbers in $[0, 1]$ can be generated using sophisticated algorithms.
- Using the probability integral transform, we can convert uniformly distributed pseudo-random samples to pseudo-random samples from other (continuous) distributions:

(1) Generate uniformly distributed pseudo-random numbers $u_1, u_2, \ldots, u_n$ in $[0, 1]$.

(2) Compute $x_i = F^{-1}(u_i)$ for $i = 1, 2, \ldots, n$.

The set $\{x_1, x_2, \ldots, x_n\}$ is a pseudo-random sample from the distribution whose CDF is $F(x)$.

Example 6.13
Given an algorithm that generates uniformly distributed pseudo-random numbers in the range $[0, 1]$, show how to generate a pseudo-random sample from the exponential distribution with scale parameter $1/2$.

Solution:

6.6 Exercises

Exercise 6.1

1. Let X be a discrete random variable, with PMF $f_X(-2) = 1/3$, $f_X(0) = 1/3$, $f_X(2) = 1/3$, and zero otherwise. Find the distribution of $Y = X + 3$.
2. Let $X \sim \text{Binomial}(n, p)$ and define $g(x) = n - x$. Show that $g(X) \sim \text{Binomial}(n, 1 - p)$.

3. Let X be a random variable, and let F_X denote its CDF. Find the CDF of $Y = X^2$ in terms of F_X.

4. Let X be a random variable with the following CDF:

$$F_X(x) = \begin{cases} 1 - \dfrac{1}{x^3} & \text{for } x \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

   Find the CDF of the random variable $Y = 1/X$, and describe how a pseudo-random sample from the distribution of Y can be obtained using an algorithm that generates uniformly distributed pseudo-random numbers in the range $[0, 1]$.

Lecture 7 Examples of transformations

7.1 Standard normal CDF $\longrightarrow$ Normal CDF

Example 7.1
Let $Z \sim N(0,1)$. Find the CDF of $X = \mu + \sigma Z$ in terms of the CDF of Z.

Solution:

7.2 Standard normal CDF $\longrightarrow$ Chi-squared CDF

Example 7.2 (The chi-squared distribution)
Let $X \sim N(0,1)$. Find the CDF of $Y = X^2$.

Solution:

7.3 Standard uniform CDF $\longrightarrow$ Exponential CDF

Example 7.3
Let $X \sim \text{Uniform}[0,1]$, and let $Y = -\theta \log X$ where $\theta > 0$. Show that $Y \sim \text{Exponential}(\theta)$, where θ is a scale parameter.

Solution:

7.4 Exponential CDF $\longrightarrow$ Pareto CDF

Example 7.4
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:

Remark 7.5
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. However, we can see 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 *heavy-tailed* distributions.

7.5 Normal PDF $\longrightarrow$ Standard normal PDF

Example 7.6 (The standard normal distribution)
Let $X \sim N(\mu, \sigma^2)$, and define $Z = (X - \mu)/\sigma$. Find the PDF of Z.

Solution:

7.6 Pareto PDF $\longrightarrow$ Standard uniform PDF

Example 7.7 (The Pareto distribution)
The Pareto(θ, α) distribution is a continuous distribution with PDF

$$f_X(x) = \begin{cases} \dfrac{\alpha}{\theta}\left(\dfrac{\theta}{x}\right)^{\alpha+1} & \text{for } x > \theta, \\ 0 & \text{otherwise.} \end{cases}$$

Let $X \sim \text{Pareto}(1, 1)$. Find the PDF of $Y = 1/X$.

Solution:

7.7 Normal PDF $\longrightarrow$ Lognormal PDF

Example 7.8 (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:

7.8 Lomax PDF $\longrightarrow$ Logistic CDF

Example 7.9 (The logistic distribution)
The Lomax(θ, α) distribution[1] 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:

7.9 Exercises

Exercise 7.1

[1]The Lomax distribution is also known as the *Pareto Type II distribution* and the *shifted Pareto distribution*

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

2. Let X have exponential distribution with rate parameter $\lambda > 0$. The PDF of X is

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

   Find the PDFs of $Y = X^2$ and $Z = e^X$.

3. Let $X \sim \text{Pareto}(1,2)$. Find the PDF of $Y = 1/X$.

4. A continuous random variable U has PDF

$$f(u) = \begin{cases} 12u^2(1-u) & \text{for} \quad 0 < u < 1, \\ 0 & \text{otherwise.} \end{cases}$$

   Find the PDF of $V = (1-U)^2$.

5. The continuous random variable U has PDF

$$f_U(u) = \begin{cases} 1+u & -1 < u \leq 0, \\ 1-u & 0 < u \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

   Find the PDF of $V = U^2$. (Note that the transformation is not injective over $\text{supp}(f_U)$, so you should first compute the CDF of V, then derive its PDF by differentiation.)

6. Let X have exponential distribution with scale parameter $\theta > 0$. This has PDF

$$f(x) = \begin{cases} \frac{1}{\theta} \exp(-x/\theta) & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

   Find the PDF of $Y = X^{1/\gamma}$ where $\gamma > 0$.

7. Suppose that X has the *Beta Type I* distribution, with parameters $\alpha, \beta > 0$. This has PDF

$$f_X(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} & \text{for } 0 \leq x \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

   where $B(a,b) = \displaystyle\int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ is the so-called *beta function.* Show that the random variable $Y = \dfrac{X}{1-X}$ has the *Beta Type II* distribution, which has PDF

$$f_Y(y) = \begin{cases} \dfrac{1}{B(\alpha,\beta)} \dfrac{y^{\alpha-1}}{(1+y)^{\alpha+\beta}} & \text{for } y > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Lecture 8 Series and Integrals

8.1 Motivation

Discrete random variables

A discrete random variable can be represented by a sequence of real numbers.

- Let X take values in the set $\mathbb{N} = \{1, 2, 3, \ldots\}$, and let $p_k = \mathbb{P}(X = k)$.
- The PMF of X is the sequence $p_1, p_2, \ldots$.
- The only constraints on the sequence are that its terms p_k are all non-negative, and $\sum_{k=1}^{\infty} p_k = 1$.
- The expectation of X is given by the series $\mathbb{E}(X) = \sum_{k=1}^{\infty} k p_k$.
  - This series does not necessarily converge (to a finite value).
  - It may not even be well-defined.

Continuous random variables

A continuous random variables can be represented by a function $f : \mathbb{R} \to \mathbb{R}$.

- Let X be a continuous random variable, and let $F(x) = \mathbb{P}(X \leq x)$ be its CDF.
- The PDF of X is the function $f(x) = F'(x)$.
- The only constraints on f are that $f(x) \geq 0$ for all $x \in \mathbb{R}$, and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.
- The expectation of X is given by the integral $\int_{-\infty}^{\infty} x\, f(x)\, dx$.
  - This integral does not necessarily converge (to a finite value).
  - It may not even be well-defined.

8.2 Series

8.2.1 Convergent sequences

Definition 8.1
An infinite sequence of real numbers $a_1, a_2, \ldots$ is said to *converge* if there exists some $a \in \mathbb{R}$ such that for all $\epsilon > 0$, there exists some $N \in \mathbb{N}$ with

$$|a_n - a| < \epsilon \quad \text{for all} \quad n > N.$$

Remark 8.2
(1) The number a is called the *limit* of the sequence, written as $a = \lim_{n\to\infty} a_n$.

(2) A sequence that is not convergent is said to be *divergent*.

8.2.2 Convergent series

Definition 8.3
Let $a_1, a_2, \dots$ be a sequence of real numbers. The infinite series $\sum_{n=1}^{\infty} a_n$ is said to be

(1) *convergent* if the sequence of partial sums $\sum_{n=1}^{m} a_n$ converges as $m \to \infty$,

(2) *absolutely convergent* if $\sum_{n=1}^{\infty} |a_n|$ is convergent,

(3) *conditionally convergent* if it is convergent, but is not absolutely convergent,

(4) *divergent* if the sequence of partial sums $\sum_{n=1}^{m} a_n$ diverges as $m \to \infty$.

Example 8.4

- $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$ converges.
- $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. (This is the *harmonic series.*)
- $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = \log 2$ is converges conditionally. (This is the *alternating harmonic series.*)

The alternating harmonic series is convergent,

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots = \log 2,$$

but not absolutely convergent, because

$$\sum_{n=1}^{\infty} \left| \frac{(-1)^{n+1}}{n} \right| = \sum_{n=1}^{\infty} \frac{1}{n}.$$

8.2.3 Positive and negative parts

If a series $\sum_n a_n$ has both positive and negative terms, we can write it as the difference of two series of non-negative terms:

$$\sum_{n=1}^{\infty} a_n = \sum_{n=1}^{\infty} a_n^+ - \sum_{n=1}^{\infty} a_n^-,$$

where

$$\begin{aligned} a_n^+ &= \max\{a_n, 0\} &&= \begin{cases} a_n & \text{if } a_n \geq 0, \\ 0 & \text{otherwise.} \end{cases} \\ a_n^- &= \max\{-a_n, 0\} &&= \begin{cases} -a_n & \text{if } a_n < 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

- $\displaystyle\sum_{n=1}^{\infty} a_n^+$ is called the *positive part* of the series.
- $\displaystyle\sum_{n=1}^{\infty} a_n^-$ is called the *negative part* of the series.

Since

$$\sum_{n=1}^{\infty} |a_n| = \sum_{n=1}^{\infty} a_n^+ + \sum_{n=1}^{\infty} a_n^-,$$

we see that

(1) If $\sum_n a_n$ is absolutely convergent, its positive and negative parts both converge.

(2) If $\sum_n a_n$ is conditionally convergent, its positive and negative parts both diverge.

Consider the alternating harmonic series:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \ldots = \log 2,$$

- The positive and negative parts of the alternating harmonic series both diverge.
- There is sufficient cancellation between its terms to ensure that the series itself converges.

8.2.4 The Riemann rearrangement theorem

Definition 8.5

(1) A bijection $\phi : \mathbb{N} \to \mathbb{N}$ is called a *permutation* of the labels $\{1, 2, \ldots\}$.

(2) The *rearrangement* of the series $\sum_n a_n$ by the permutation ϕ is the series $\sum_n a_{\phi(n)}$.

Theorem 8.6 (The Riemann rearrangement theorem)
Let $\sum_n a_n$ be a convergent series.

(1) If $\sum_n a_n$ is absolutely convergent, then every rearrangement $\sum_n a_{\phi(n)}$ is absolutely convergent to the same limit.

(2) If $\sum_n a_n$ is conditionally convergent, then for every $a \in \mathbb{R} \cup \{\pm\infty\}$ there exists a permutation $\phi : \mathbb{N} \to \mathbb{N}$ such that $\sum_n a_{\phi(n)} = a$.

[*Proof omitted.*]

- The limit of conditionally convergent series depends on the *order* in which the terms of the series are added together.
- In probability theory, we cannot deal with sums that are conditionally convergent.
- Expectation is only defined if the series $\sum_{i=1}^{\infty} kp_k$ is absolutely convergent.

8.3 Integrals

8.3.1 The Riemann integral

Let $g : [a, b] \to \mathbb{R}$ be a bounded function. A *partition* of $[a, b]$ is a set of intervals

$$\mathcal{P} = \{[x_0, x_1], [x_1, x_2], \ldots, [x_{n-1}, x_n]\}$$

where $a = x_0 < x_1 < x_2 < \ldots < x_n = b$.

The upper and lower *Riemann sums* of g with respect to $\mathcal{P}$ are, respectively,

$$\begin{aligned} U(\mathcal{P}, g) &= \sum_{i=1}^{n} M_i \Delta_i \quad \text{where} \quad M_i = \sup\{g(x) : x \in [x_{i-1}, x_i]\}, \\ L(\mathcal{P}, g) &= \sum_{i=1}^{n} m_i \Delta_i \quad \text{where} \quad m_i = \inf\{g(x) : x \in [x_{i-1}, x_i]\}, \end{aligned}$$

where $\Delta_i = x_i - x_{i-1}$ is the length of the interval $[x_{i-1}, x_i]$.

The upper and lower *Riemann integrals* of g on $[a, b]$ are, respectively,

$$\underline{\int_a^b} g(x)\,dx = \sup_{\mathcal{P}} L(\mathcal{P}, g) \quad \text{and} \quad \overline{\int_a^b} g(x)\,dx = \inf_{\mathcal{P}} U(\mathcal{P}, g).$$

where the supremum and infimum are taken over all possible partitions of $[a,b]$.

If the upper and lower Riemann integrals coincide, we say that g is *Riemann integrable*, in which case their common value is called the *Riemann integral* of g, denoted by

$$\int_a^b g(x)\,dx.$$

To extend the definition to (1) integrals over unbounded intervals, and (2) integrals of unbounded functions, we use limits to define *improper* integrals:

$$\int_{-\infty}^{\infty} e^{-x^2}\,dx = \lim_{n\to\infty}\int_{-n}^{n} e^{-x^2}\,dx,$$

$$\int_0^1 \frac{1}{\sqrt{x}}\,dx = \lim_{\epsilon\to 0}\int_{\epsilon}^{1}\frac{1}{\sqrt{x}}\,dx.$$

8.3.2 Integrable functions

A function $g:\mathbb{R}\to\mathbb{R}$ is said to be *integrable* (in the Riemann sense) if the area between the curve $g(x)$ and the horizontal axis is finite:

$$\int_{-\infty}^{\infty}|g(x)|\,dx < \infty.$$

If a function is not integrable, we say that its integral is *undefined* or *does not exist*.

Let $g:\mathbb{R}\to\mathbb{R}$ be an integrable function. The *integral* of g is the difference between the area above the horizontal axis and the area below the horizontal axis:

$$\int_{-\infty}^{\infty} g(x)\,dx = \int_{-\infty}^{\infty} g^+(x)\,dx - \int_{-\infty}^{\infty} g^-(x)\,dx,$$

where g^+ and g^- are respectively the *positive part* and *negative part* of g:

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

Note that:

- $g^+(x)\geq 0$ and $g^-(x)\geq 0$ for all $x\in\mathbb{R}$ (i.e. both are non-negative functions).
- $g(x) = g^+(x) - g^-(x)$ for all $x\in\mathbb{R}$.
- $|g(x)| = g^+(x) + g^-(x)$ for all $x\in\mathbb{R}$.

If g is integrable, then $\displaystyle\int g^+(x)\,dx$ and $\displaystyle\int g^-(x)\,dx$ are both finite:

$$\int g^+(x)\,dx + \int g^-(x)\,dx = \int g^+(x) + g^-(x)\,dx = \int |g(x)|\,dx < \infty.$$

If g is not integrable, one or both of $\displaystyle\int g^+(x)\,dx$ and $\displaystyle\int g^-(x)\,dx$ must be infinite:

- if $\displaystyle\int g^+(x)\,dx = \infty$ and $\displaystyle\int g^-(x)\,dx < \infty$, then $\displaystyle\int g(x)\,dx = +\infty$,
- if $\displaystyle\int g^+(x)\,dx < \infty$ and $\displaystyle\int g^-(x)\,dx = \infty$, then $\displaystyle\int g(x)\,dx = -\infty$,
- if $\displaystyle\int g^+(x)\,dx = \infty$ and $\displaystyle\int g^-(x)\,dx = \infty$, we say that $\displaystyle\int g(x)\,dx$ is *undefined*.

Example 8.7
Let $g:\mathbb{R}\to\mathbb{R}$ be the function $g(x)=\begin{cases}\sin x & \text{if } \ 0\leq x\leq 2\pi\\ 0 & \text{otherwise.}\end{cases}$

- Positive part: $g^+(x)=\begin{cases}\sin x & \text{if } 0\leq x\leq \pi,\\ 0 & \text{otherwise.}\end{cases}$.
- Negative part: $g^-(x)=\begin{cases}-\sin x & \text{if } \pi\leq x\leq 2\pi,\\ 0 & \text{otherwise.}\end{cases}$.

The area above and below the horizontal axis are respectively

$$\begin{aligned}A^+ &= \int_{-\infty}^{\infty} g^+(x)\,dx = \int_0^\pi \sin x\,dx = \big[-\cos x\big]_0^\pi = 2,\\ A^- &= \int_{-\infty}^{\infty} g^-(x)\,dx = \int_\pi^{2\pi} (-\sin x)\,dx = \big[\cos x\big]_\pi^{2\pi} = 2.\end{aligned}$$

Hence the integral of g over $\mathbb{R}$ is

$$\int_{-\infty}^{\infty} g(x)\,dx = \int_{-\infty}^{\infty} g^+(x)\,dx - \int_{-\infty}^{\infty} g^-(x)\,dx = A^+ - A^- = 0.$$

8.3.3 The Riemann-Stieltjes integral

Let $g:[a,b]\to\mathbb{R}$ be a bounded function, let $\mathcal{P}$ be a partition of $[a,b]$ and let $F:\mathbb{R}\to[0,1]$ be a CDF.

The upper and lower *Riemann-Stieltjes sums* of g with respect to $\mathcal{P}$ and F are, respectively,

$$\begin{aligned}U(\mathcal{P},g,F) &= \sum_{i=1}^n M_i\Delta_i \quad \text{where} \quad M_i=\sup\{g(x):x\in[x_{i-1},x_i]\},\\ L(\mathcal{P},g,F) &= \sum_{i=1}^n m_i\Delta_i \quad \text{where} \quad m_i=\inf\{g(x):x\in[x_{i-1},x_i]\},\end{aligned}$$

where $\Delta_i=F(x_i)-F(x_{i-1})$ is the probability measure induced by F of the interval $[x_{i-1},x_i]$.

The upper and lower *Riemann-Stieltjes integrals* of g on $[a,b]$ are, respectively,

$$\underline{\int_a^b} g(x)\,dF(x) = \sup_{\mathcal{P}} L(\mathcal{P},g,F) \quad \text{and} \quad \overline{\int_a^b} g(x)\,dF(x) = \inf_{\mathcal{P}} U(\mathcal{P},g,F).$$

where the supremum and infimum are taken over all possible partitions of $[a,b]$.

- If the upper and lower Riemann-Stieltjes integrals coincide, we say that g is *Riemann-Stieltjes integrable.*
- In this case, their common value is called the *Riemann-Stieltjes integral* of g, denoted by

$$\int_a^b g(x)\,dF(x).$$

Remark 8.8
Let F be the CDF of the *uniform* distribution on $[a,b]$:

$$F(x)=\begin{cases}0 & x<a,\\ \dfrac{x-a}{b-a} & a\leq x\leq b,\\ 1 & x>b.\end{cases}$$

In this case, for any interval $[x_{i-1},x_i]\subseteq[a,b]$ the probability measure induced by F is equal to its length, and the Riemann-Stieltjes integral reduces to the ordinary Riemann integral.

Lecture 9 Expectation

Expectation is to random variables what probability is to events.

- Random events are *sets*, and are studied using *set algebra*.
- Random variables are *functions*, and are studied using *mathematical analysis*.

Elementary probability theory provides the following computational formulae for the expectation of a random variable $X : \Omega \to \mathbb{R}$.

(1) If Ω is a finite sample space with probability mass function $p : \Omega \to \mathbb{R}$, the expectation of X is

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

(2) If X is a discrete random variable with PMF $f(x)$ and range $\{x_1, x_2, \ldots\}$, the expectation of X is

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

(3) if X is a continuous random variable with PDF $f(x)$, the expectation of X is

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

- The convergence of such sums and integrals is not guaranteed under all circumstances.
- If X takes only non-negative values, we can accept that $\mathbb{E}(X) = \infty$.
- If X can take both positive and negative values, we need that $\mathbb{E}(X) < \infty$.

9.1 Indicator variables

Consider the indicator variable of an event A,

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

To be consistent with the probability measure $\mathbb{P}(A)$, the only reasonable definition of expectation for I_A is the following:

Definition 9.1 (Expectation of indicator variables)
The expectation of an indicator variable $I_A : \Omega \to \mathbb{R}$ is defined by

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

9.2 Simple random variables

Let X be a simple random variable, let $\{x_1, x_2, \dots, x_n$ denote its range, and let $\{A_1, A_2, \dots, A_n\}$ be a partition of Ω such that $X(\omega) = x_i$ for all $\omega \in A_i$. Then X can be expressed as a finite linear combination of indicator variables,

$$X(\omega) = \sum_{i=1}^{n} x_i I_{A_i}(\omega) \qquad \text{where} \quad I_{A_i}(\omega) = \begin{cases} 1 & \text{if } \omega \in A_i, \\ 0 & \text{if } \omega \notin A_i. \end{cases}$$

Definition 9.2 (Expectation of simple random variables)
The expectation of a simple random variable $X = \sum_{i=1}^{n} x_i I_{A_i}$ is defined by

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

Remark 9.3
It can be shown all representations of X as finite linear combinations of inidcator variables yield the same value for $\mathbb{E}(X)$, so the expectation of a simple random variable is well-defined.

Example 9.4
A fair coin is tossed three times. Let $X : \Omega \to \mathbb{R}$ be the random variable

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \in A_1 = \{TTT\} \\ 2 & \text{if } \omega \in A_2 = \{TTH, THT, HTT\} \\ 3 & \text{if } \omega \in A_3 = \{THH, HTH, HHT\} \\ 4 & \text{if } \omega \in A_4 = \{HHH\} \end{cases}$$

Compute the expected value of X.

Solution:

Definition 9.5
Let X and Y be random variables on Ω.

(1) If $X(\omega) \geq 0$ for all $\omega \in \Omega$, we say that X is *non-negative*. This is denoted by $X \geq 0$.

(2) If $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, we say that X is *dominated* by Y. This is denoted by $X \leq Y$.

Theorem 9.6 (Properties of expectation for simple random variables)
Let $X, Y : \Omega \to \mathbb{R}$ be simple random variables.

(1) **Positivity**. If $X \geq 0$ then $\mathbb{E}(X) \geq 0$.

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

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

Proof:

Example 9.7
Extending example 9.4, let $Y : \Omega \to \mathbb{R}$ be the random variable

$$Y(\omega) = \begin{cases} 2 & \text{if} \;\; \omega \in A'_1 = \{TTT, TTH\} \\ 3 & \text{if} \;\; \omega \in A'_2 = \{THT, THH\} \\ 4 & \text{if} \;\; \omega \in A'_3 = \{HTT, HTH\} \\ 5 & \text{if} \;\; \omega \in A'_4 = \{HHT, HHH\} \end{cases}$$

Compute the expected value of (i) Y and (ii) $3X + 2Y$.

Solution:

9.3 Non-negative random variables

Theorem 9.8
For every non-negative random variable $X \geq 0$, there exists an increasing sequence of simple non-negative random variables

$$0 \leq X_1 \leq X_2 \leq \dots$$

with the property that $X_n(\omega) \uparrow X(\omega)$ for each $\omega \in \Omega$ as $n \to \infty$.

[*Proof omitted.*]

Definition 9.9 (Expectation of non-negative random variables)
The *expectation* of a non-negative random variable X is defined to be

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X_n)$$

where the X_n are simple non-negative random variables with $X_n \uparrow X$ as $n \to \infty$.

Remark 9.10
It can be shown that all approximating sequences yield the same value for $\mathbb{E}(X)$, so the expectation of non-negative random variables is well-defined

Remark 9.11 (Infinite expectation)
The expectation of non-negative random variables can be infinite:

(1) If $\mathbb{E}(X) < \infty$ we say that X has *finite* expectation.

(2) If $\mathbb{E}(X) = \infty$ we say that X has *infinite* expectation.

Theorem 9.12 (Properties of expectation for non-negative random variables)
For non-negative random variables $X, Y \geq 0$,

(1) **Positivity**. If $X \geq 0$ then $\mathbb{E}(X) \geq 0$.

(2) **Linearity**. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ for every $a, b \in \mathbb{R}_+$.

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

(4) **Continuity**. If $X_n \to X$ as $n \to \infty$, where the X_n are non-negative, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$ as $n \to \infty$.

[*Proof omitted.*]

9.4 Signed random variables

We can extend the definition of expectation to random variables that take both positive and negative values, but only if the random variables are *integrable*:

Definition 9.13 (Integrable random variables)
A random variable X is said to be *integrable* if $\mathbb{E}(|X|) < \infty$.

Definition 9.14 (The positive and negative parts)
The positive and negative parts of a random variable X, denoted by X^+ and X^- respectively, are defined to be

$$\begin{aligned} X^+(\omega) &= \max\{0, X(\omega)\} &&= \begin{cases} X(\omega) & \text{if } X(\omega) \geq 0, \\ 0 & \text{if } X(\omega) < 0; \end{cases} \\ X^-(\omega) &= \max\{0, -X(\omega)\} &&= \begin{cases} -X(\omega) & \text{if } X(\omega) \leq 0, \\ 0 & \text{if } X(\omega) > 0. \end{cases} \end{aligned}$$

Note that X^+ and X^- are both non-negative random variables, with $X = X^+ - X^-$.

Definition 9.15 (Expectation of signed random variables)
The *expectation* of an integrable random variable $X : \Omega \to \mathbb{R}$ is

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

where X^+ and X^- are respectively the positive part and negative part of X:

Remark 9.16 (Undefined expectation)
Because $|X| = X^+ + X^-$, it follows by the linearity of expectation for non-negative random variables that

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

(1) If X is integrable, then $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ are both finite.

(2) If X is not integrable, then one or both of $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ must be infinite:

- if $\mathbb{E}(X^+) = \infty$ and $\mathbb{E}(X^-) < \infty$, we write $\mathbb{E}(X) = +\infty$;
- if $\mathbb{E}(X^+) < \infty$ and $\mathbb{E}(X^-) = \infty$, we write $\mathbb{E}(X) = -\infty$;
- if $\mathbb{E}(X^+) = \infty$ and $\mathbb{E}(X^-) = \infty$, we say that $\mathbb{E}(X)$ *does not exist*.

Theorem 9.17 (Properties of expectation for signed random variables)
Let X and Y be integrable random variables.

(1) **Monotonicity**. If $X \leq Y$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

(2) **Linearity**. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ for all $a, b \in \mathbb{R}$.

[*Proof omitted.*]

9.5 Exercises

Exercise 9.1

1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $0 \leq X_1 \leq X_2 \leq \ldots$ be an increasing sequence of non-negative random variables over $(\Omega, \mathcal{F})$ such that $X_n(\omega) \uparrow X(\omega)$ ans $n \to \infty$ for all $\omega \in \Omega$. Show that X is a random variable on $(\Omega, \mathcal{F})$.
2. Let X be an integrable random variable. Show that $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.
3. Let X and Y be integrable random variables. Show that $aX + bY$ is integrable.

Lecture 10 Computation of Expectation

10.1 Expectation with respect to CDFs

The natural definition of expectation for indicator variables (9.1) was extended to the expectation of simple random variables (9.2), then to non-negative variables (9.3) and finally to signed variables (9.4).

According to definition 9.9, to compute the expectation of a non-negative random variable X, we need to find an increasing sequence $0 \leq X_1 \leq X_2 \leq ...$ of simple random variables for which $X_n \uparrow X$ as $n \to \infty$, then compute the limit of $\mathbb{E}(X_n)$ as $n \to \infty$. This is not feasible in practical applications.

It turns out that the expectation of a random variable can be conventiently expressed as a Riemann-Stieltjes integral with respect to its CDF:

Theorem 10.1
Let $X : \Omega \to \mathbb{R}$ be a non-negative random variable, and let $F : \mathbb{R} \to [0, 1]$ denote its CDF. The expectation of X can be written as

$$\mathbb{E}(X) = \int_0^\infty x \, dF(x).$$

where the right-hand side is the Riemann-Stieltjes integral of x with respect to F.

[*Proof omitted.*]

The following theorem yields a computational formula for expectation, in terms of an ordinary Riemann integral:

Theorem 10.2
If X is non-negative,

$$\mathbb{E}(X) = \int_0^\infty 1 - F(x) \, dx$$

[*Proof omitted.*]

For signed random variables, we first compute $\mathbb{E}(X^+)$ and $\mathbb{E}(X^-)$ using to this formula, and then set $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$ as before. To do this, we must first find the CDFs of X^+ and X^-.

10.2 Discrete distributions

If X is discrete, the Riemann-Stieltjes integral of Theorem 10.1 reduces to a sum:

Theorem 10.3
Let X be a non-negative discrete random variable, let $\{x_1, x_2, \ldots\}$ be its range, and let $f(x)$ denote its PMF. Then

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

provided the sum is absolutely convergent.

[*Proof omitted.*]

Remark 10.4
This expression also holds for signed discrete random variables. Can you prove this?

The following is a special case of Theorem 10.2:

Theorem 10.5
Let X be a discrete non-negative random variable, taking values in the range $\{0, 1, 2, \ldots\}$. Then

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} \mathbb{P}(X > k)$$

Proof:

Example 10.6 (Geometric distribution)
Suppose X has the geometric distribution on $\{0, 1, 2, \ldots\}$, with probability-of-success parameter p. Given that the CDF of X is $\mathbb{P}(X \leq k) = 1 - (1-p)^{k+1}$, show that its expected value is equal to $(1-p)/p$.

Solution:

10.3 Continuous distributions

If X is continuous, the Riemann-Stieltjes integral of Theorem 10.1 reduces to an ordinary Riemann integral:

Theorem 10.7
Let X be a non-negative continuous random variable, and let $f(x)$ denote its PDF. Then

$$\mathbb{E}(X) = \int_0^\infty xf(x)\,dx,$$

provided the integral is absolutely convergent.

[*Proof omitted.*]

Remark 10.8
For signed continuous random variables, $\mathbb{E}(X) = \displaystyle\int_{-\infty}^{\infty} xf(x)\,dx$ provided the integral is absolutely convergent.

Theorem 10.9
Let X be a non-negative continuous random variable, and let F denote its CDF. Then

$$\mathbb{E}(X) = \int_0^\infty 1 - F(x)\,dx$$

[*Proof omitted.*]

Example 10.10 (Rayleigh distribution)
Let X be a continuous random variable having the *Rayleigh* distribution with parameter $\sigma > 0$. This has the following CDF:

$$F(x) = \begin{cases} 1 - e^{-x^2/2\sigma^2} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Show that $\mathbb{E}(X) = \sigma\sqrt{\dfrac{\pi}{2}}$.

Solution:

10.4 Transformed variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X : \Omega \to \mathbb{R}$ be a random variable, let $F : \mathbb{R} \to [0,1]$ be its CDF, and let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function.

- By Theorem 6.2, the transformed variable $g(X)$ is a random variable on $(\Omega, \mathcal{F})$.

(1) If $g(X)$ is a non-negative random variable,

$$\mathbb{E}\big[g(X)\big] = \int_0^\infty g(x)\, dF(x)$$

(2) If $g(X)$ is an integrable random variable,

$$\mathbb{E}\big[g(X)\big] = \int_0^\infty g^+(x)\, dF(x) - \int_0^\infty g^-(x)\, dF(x).$$

The latter expression reduces to

$$\mathbb{E}\big[g(X)\big] = \begin{cases} \displaystyle\sum_{i=1}^\infty g^+(x_i) f(x_i) - \sum_{i=1}^\infty g^-(x_i) f(x_i) & \text{when } X \text{ is discrete, and} \\ \displaystyle\int_0^\infty g^+(x) f(x)\, dx - \int_0^\infty g^-(x) f(x)\, dx & \text{when } X \text{ is continuous.} \end{cases}$$

Example 10.11
Let $X \sim \text{Uniform}[-1,1]$ be a continuous random variable. Find $\mathbb{E}(1/X^2)$ and $\mathbb{E}(1/X)$.

Solution:

10.5 Exercises

Exercise 10.1

1. Let X be the score on a fair die, and let $g(x) = 3x - x^2$. Find the expected value and variance of the random variable $Y = g(X)$.

2. A long line of athletes $k = 0, 1, 2, \ldots$ make throws of a javelin to distances $X_0, X_1, X_2, \ldots$ respectively. The distances are independent and identically distributed random variables, and the probability that any two throws are exactly the same distance is equal to zero. Let Y be the index of the first athlete in the sequence who throws further than distance X_0. Show that the expected value of Y is infinite.

3. Consider the following game. A random number X is chosen uniformly from $[0, 1]$, then a sequence $Y_1, Y_2, \ldots$ of random numbers are chosen independently and uniformly from $[0, 1]$. Let Y_n be the first number in the sequence for which $Y_n > X$. When this occurs, the game ends and the player is paid $(n - 1)$ pounds. Show that the expected win is infinite.

4. Let X be a discrete random variable with PMF
$$f(k) = \begin{cases} \dfrac{3}{\pi^2 k^2} & \text{if } k \in \{\pm 1, \pm 2, \ldots\} \\ 0 & \text{otherwise.} \end{cases}$$
Show that $\mathbb{E}(X)$ is undefined.

5. Let X be a continuous random variable having the Cauchy distribution, defined by the PDF
$$f(x) = \frac{1}{\pi(1 + x^2)} \qquad x \in \mathbb{R}$$
Show that $\mathbb{E}(X)$ is undefined.

6. A coin is tossed until the first time a head is observed. If this occurs on the nth toss and n is odd, you win $2^n/n$ pounds, but if n is even then you lose $2^n/n$ pounds. Show that the expected win is undefined.

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

8. Let X be a random variable with the following CDF:
$$F(x) = \begin{cases} 0 & \text{for } x \leq 1 \\ 1 - 1/x^2 & \text{for } x \geq 1 \end{cases}$$
Compute $\mathbb{E}(X)$, $\mathbb{E}(X^2)$, $\mathbb{E}(1/X)$ and $\mathbb{E}(1/X^2)$.

9. Let X be a continuous random variable with the following PDF:
$$f(x) = \begin{cases} 1 - |x| & \text{if } x \in [-1, 1] \\ 0 & \text{otherwise.} \end{cases}$$
Find the range of integer values $\alpha \in \mathbb{Z}$ for which $\mathbb{E}(X^\alpha)$ exists.

Lecture 11 Concentration Inequalities

11.1 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 probability of a non-negative random variable is provided by *Markov's inequality*.

Theorem 11.1 (Markov's inequality)
Let $X \geq 0$ be any non-negative random varible with finite mean. Then for every $a > 0$,

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

Proof:

Example 11.2
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:

Theorem 11.3 (Markov's inequality (General form))
Let X be any random varible with finite mean, and let $g : \mathbb{R} \to [0, \infty)$ be a non-negative function. Then for every $a > 0$,

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

Proof:

11.2 Chebyshev's inequality

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

Corollary 11.4 (Chebyshev's inequality)
Let X be any random varible with finite mean. Then for all $\epsilon > 0$,

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

Proof:

Example 11.5
Suppose that $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) = 1$. Find an integer value k such that $\mathbb{P}(|X| \geq k) \leq 0.01$.

Solution:

Example 11.6
Let X be a continuous random variable with expected value 3.6 and standard deviation 1.2. Find a lower bound for the probability $\mathbb{P}(1.2 \leq X \leq 6.0)$.

Solution:

11.3 Bernstein's inequality

Theorem 11.7 (Bernstein's inequality)
Let X be a random variable. Then for all $t > 0$,

$$\mathbb{P}(X > a) \leq e^{-ta}\mathbb{E}(e^{tX}).$$

Proof:

11.4 Exercises

Exercise 11.1

1. Let $X \sim \text{Uniform}[0, 20]$ be a continuous random variable.
   (1) Use Chebyshev's inequality to find an upper bound on the probability $\mathbb{P}(|X - 10| \geq z)$.
   (2) Find the range of z for which Chebyshev's inequality gives a non-trivial bound.
   (3) Find the value of z for which $\mathbb{P}(|X - 10| \geq z) \leq 3/4$.

2. Let X be a discrete random variable, taking values in the range $\{1, 2, \ldots, n\}$, and suppose that $\mathbb{E}(X) = \text{Var}(X) = 1$. Show that $\mathbb{P}(X \geq k + 1) \leq k^2$ for any integer k.

3. Let $k \in \mathbb{N}$. Show that Markov's inequality is tight (i.e. cannot be improved) by finding a non-negative random variable X such that
$$\mathbb{P}\big[X \geq k\mathbb{E}(X)\big] = \frac{1}{k}.$$

4. What does the Chebyshev inequality tell us about the probability that the value taken by a random variable deviates from its expected value by six or more standard deviations?

5. Let S_n be the number of successes in n Bernoulli trials with probability p of success on each trial. Use Chebyshev's Inequality to show that, for any $\epsilon > 0$, the upper bound
$$\mathbb{P}\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{1}{4n\epsilon^2}$$
is valid for any p.

6. Let $X \sim N(0, 1)$.
   (1) Use Chebyshev's Inequality to find upper bounds for the probabilities $\mathbb{P}(|X| \geq 1)$, $\mathbb{P}(|X| \geq 2)$ and $\mathbb{P}(|X| \geq 3)$.
   (2) Use statistical tables to find the area under the standard normal curve over the intervals $[-1, 1]$, $[-2, 2]$ and $[-3, 3]$.
   (3) Compare the bounds computed in part (a) with the exact values found in part (b). How good is the Chebyshev inequality in this case?

7. Let X be a random variable with mean $\mu \neq 0$ and variance σ^2, and define the *relative deviation* of X from its mean by $D = \left|\dfrac{X - \mu}{\mu}\right|$. Show that
$$\mathbb{P}(D \geq a) \leq \left(\frac{\sigma}{\mu a}\right)^2.$$

Lecture 12 Probability Generating Functions

12.1 Generating functions

Generating functions, first introduced by de Moivre in 1730, are power series used to represent sequences of real numbers. It is often easier to work with generating functions than with the original sequences.

Definition 12.1
Let $a = (a_0, a_1, a_2, \ldots)$ be a sequence of real numbers. The *generating function* of the sequence is the function $G_a(t)$, defined for every $t \in \mathbb{R}$ for which the sum converges, by

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

The sequence can be reconstructed from $G_a(t)$ by setting

$$a_n = \frac{1}{n!} G_a^{(n)}(0),$$

where $G_a^{(n)}(t)$ is nth derivative of $G_a(t)$. In particular,

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

Example 12.2
The *convolution* of two sequences $a = (a_0, a_1, a_2, \ldots)$ and $b = (b_0, b_1, b_2, \ldots)$ is another sequence $c = (c_0, c_1, c_2, \ldots)$, whose kth term is

$$c_k = a_0 b_k + a_1 b_{k-1} + \ldots + a_k b_0 = \sum_{i=0}^{k} a_i b_{k-i}.$$

Convolutions can be difficult to handle. However, the generating function of a convolution is just the product of the generating functions of the original sequences:

$$\begin{aligned} G_c(t) = \sum_{k=0}^{\infty} c_k t^k &= \sum_{k=0}^{\infty} \left[\sum_{i=0}^{n} a_i b_{k-i} \right] t^k \\ &= \sum_{i=0}^{\infty} a_i t^i \sum_{k=i}^{\infty} b_{k-i} t^{k-i} = \sum_{i=0}^{\infty} a_i t^i \sum_{j=1}^{\infty} b_j t^j = G_a(t) G_b(t). \end{aligned}$$

A convolution of sequences is replaced by a product of generating functions.

12.1.1 Properties of generating functions*

A generating function $G_a(t)$ is a power series whose coefficients are the terms of the sequence a. All power series have the following properties:

- **Convergence**. There exists a *radius of convergence* $R \geq 0$ such that $G_a(t)$ is absolutely convergent when $|t| < R$, and divergent when $|t| > R$.
- **Differentiation**. $G_a(t)$ may be differentiated or integrated any number of times whenever $|t| < R$.
- **Uniqueness**. If $G_a(t) = G_b(t)$ for all $|t| < R'$, where $0 < R' \leq R$, then $a_n = b_n$ $\forall n$.

- **Abel's theorem**. If $a_k > 0$ for all k, and $G_a(t)$ converges for all $|t| < 1$, then

$$G_a(1) = \lim_{t\uparrow 1} G_a(t) = \lim_{t\uparrow 1} \sum_{k=0}^{\infty} a_k t_k = \sum_{k=0}^{\infty} a_k.$$

12.2 Probability generating functions

Definition 12.3
Let X be a discrete random variable taking values in the range $\{0, 1, 2, \ldots\}$, and let f denote its PMF. The *probability generating function* (PGF) of X is the generating function of its PMF:

$$G(t) = \mathbb{E}(t^X) = \sum_{k=0}^{\infty} f(k)t^k$$

Remark 12.4
- $G(t)$ converges for all $|t| \leq 1$.
- $G(0) = 0$.
- $G(1) = \sum_{k=0}^{\infty} f(k) = 1$.

Example 12.5
The PGFs of some notable discrete distributions on $\{0, 1, 2, \ldots\}$ are computed as follows:

(1) **Constant**: if $\mathbb{P}(X = c) = 1$,

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

(2) **Bernoulli**: if $X \sim \text{Bernoulli}(p)$, its PMF is

$$f(k) = \begin{cases} 1-p & \text{if } k = 0, \\ p & \text{if } k = 1, \end{cases}$$

and zero otherwise, so its PGF is

$$G(t) = \sum_{k=0}^{\infty} f(k)t^k = (1-p)t^0 + pt^1 = 1 - p + pt.$$

(3) **Poisson**: if X$\sim$ Poisson(λ), its PMF is

$$f(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \text{for } k = 0, 1, 2, \ldots,$$

and zero otherwise, so its PGF is

$$G(t) = \sum_{k=0}^{\infty} f(k)t^k = \sum_{k=0}^{\infty} \left(\frac{\lambda^k e^{-\lambda}}{k!}\right) t^k = e^{-\lambda} \sum_{i=1}^{\infty} \frac{(\lambda t)^k}{k!} = e^{-\lambda} e^{\lambda t} = e^{\lambda(t-1)}.$$

(4) **Geometric**: if $X \sim \text{Geometric}(p)$, its PMF is

$$f(k) = (1-p)^k p \text{ for } k = 0, 1, 2, \ldots,$$

and zero otherwise, so its PGF is

$$G(t) = \sum_{k=0}^{\infty} f(k)t^k = \sum_{k=0}^{\infty} (1-p)^k p t^k = p \sum_{k=0}^{\infty} \big[(1-p)t\big]^k = \frac{p}{1-(1-p)t} \qquad \text{for all } |t| < \frac{1}{1-p}.$$

Here, we have used the fact that $\displaystyle\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ for $|r| < 1$.

Theorem 12.6
Let X be a random variable, and let $G(t)$ denote its PGF. Then $\mathbb{E}(X) = G'(1)$, and more generally,

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

where $G^{(n)}(1)$ is the nth derivative of $G(t)$ evaluated at $t = 1$.

Proof:

Remark 12.7
$\mathbb{E}\big[X(X-1)\dots(X-n+1)\big]$ is called the nth *factorial moment* of X.

Example 12.8
The variance of X can be written in terms of $G(t)$ as follows:

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

12.2.1 Sums of random variables

Let X and Y be two independent discrete random variables, both taking values in $\{0, 1, 2, \dots\}$. The PMF of their sum $X + Y$ is given by the convolution of the individual PMFs,

$$\mathbb{P}(X + Y = k) = \sum_{j=0}^{\infty} \mathbb{P}(X = j)\mathbb{P}(Y = k - j).$$

The corresponding PGFs satisfy a more straightforward, multiplicative relationship:

Theorem 12.9
If X and Y are independent, then $G_{X+Y}(t) = G_X(t)G_Y(t)$.

Proof:

Corollary 12.10
If $S = X_1 + X_2 + \ldots + X_n$ is a sum of independent random variables taking values in the non-negative integers, its PGF is

$$G_S(t) = G_{X_1}(t)G_{X_2}(t)\cdots G_{X_n}(t)$$

Example 12.11
Show that the PGF of the Binomial(n, p) distribution is $G(t) = (1 - p + pt)^n$.

Solution:

Example 12.12
Let $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ) be independent. Show that $X + Y \sim$ Poisson$(\lambda + \mu)$.

Solution:

12.3 Exercises

Exercise 12.1

1. Let $X \sim$ Binomial(m, p) and $Y \sim$ Binomial(n, p). Show that $X + Y \sim$ Binomial$(m + n, p)$,
2. Show that a discrete distribution on the non-negative integers is uniquely determined by its PGF, in the sense that if two such random variables X and Y have PGFs $G_X(t)$ and $G_Y(t)$ respectively, then $G_X(t) = G_Y(t)$ if and only if $\mathbb{P}(X = k) = \mathbb{P}(Y = k)$ for all $k = 0, 1, 2, \ldots$.
3. The PGF of a random variable is given by $G(t) = 1/(2 - t)$. What is its PMF?
4. Let $X \sim$ Binomial(n, p). Using the PGF of X, show that

$$\mathbb{E}\left(\frac{1}{1 + X}\right) = \frac{1 - (1 - p)^{n+1}}{(n + 1)p}.$$

Lecture 13 Moment Generating Functions

13.1 Moment generating functions

- PGFs are defined only for discrete random variables taking non-negative integer values.
- MGFs are defined for any random variable.

Definition 13.1
The *moment generating function* (MGF) of a random variable X is a function $M : \mathbb{R} \to [0, \infty]$ given by

$$M(t) = \mathbb{E}(e^{tX}).$$

Remark 13.2
(1) e^{tX} is non-negative, so its expectation is well-defined, and $\mathbb{E}(e^{tX}) \geq 0$.

(2) For a discrete random variable X taking non-negative integer values,

$$M(t) = \mathbb{E}(e^{tX}) = \mathbb{E}\big[(e^t)^X\big] = G(e^t),$$

where G is the PGF of X.

(3) MGFs are related to *Laplace transforms.*

Example 13.3
The MGFs of some notable discrete distributions can be computed as follows:

$$\begin{array}{llll} X \sim \text{Bernoulli}(p): & G(t) = 1 - p + pt & \Rightarrow & M(t) = 1 - p + pe^t \\ X \sim \text{Binomial}(n, p): & G(t) = (1 - p + pt)^n & \Rightarrow & M(t) = (1 - p + pe^t)^n \\ X \sim \text{Poisson}(\lambda): & G(t) = e^{\lambda(t-1)} & \Rightarrow & M(t) = e^{\lambda(e^t - 1)} \end{array}$$

MGFs have properties similar to those of PGFs:

Theorem 13.4
(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:

Theorem 13.5
Let $M(t)$ be the MGF of the random variable X. If $M(t)$ converges on an open interval $(-R, R)$ centred at the origin, then

$$M(t) = \sum_{k=0}^{\infty} \frac{\mathbb{E}(X^k)}{k!} t^k$$

Proof:

Corollary 13.6
Let X be a random variable, and let $M(t)$ denote its MGF. Then

$$\mathbb{E}(X^n) = M^{(n)}(0),$$

where $M^{(n)}(0)$ is the nth derivative of $M(t)$ evaluated at $t = 0$. In particular,

$$M(0) = 1, \quad M'(0) = \mathbb{E}(X), \quad M''(0) = \mathbb{E}(X^2), \quad \text{and so on.}$$

Example 13.7 (Exponential distribution)
Let $X \sim \text{Exponential}(\lambda)$ where $\lambda > 0$ is a rate parameter.

(1) Show that the MGF of X is given by $M(t) = \dfrac{\lambda}{\lambda - t}$.

(2) Use $M(t)$ to find the mean and variance of X.

Solution:

Example 13.8 (Gamma distribution)
The PDF of the $\text{Gamma}(k, \theta)$ distribution is given by

$$f(x) = \begin{cases} \dfrac{x^{k-1}e^{-x}}{\Gamma(k)} & x > 0, \\ 0 & \text{othewise.} \end{cases}$$

Show that the MGF of the $\text{Gamma}(k, \theta)$ distribution is given by

$$M(t) = \frac{1}{(1 - \theta t)^k}$$

Solution:

Example 13.9 (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

$$M(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right).$$

Solution:

13.2 Characteristic functions

MGFs are useful, but the expectations that define them may not always be finite. Characteristic functions do not suffer this disadvantage.

Definition 13.10
The *characteristic function* of a random variable X is a function $\phi : \mathbb{R} \to \mathbb{C}$ given by

$$\phi(t) = \mathbb{E}(e^{itX}) \qquad \text{where} \qquad i = \sqrt{-1}.$$

Remark 13.11

- If $M(t)$ is the MGF of X, its characteristic function is given by $\phi(t) = M(it)$.
- $\phi : \mathbb{R} \to \mathbb{C}$ exists for all $t \in \mathbb{R}$.
- Characteristic functions are related to *Fourier transforms*.

Theorem 13.12

(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 omitted.*]

13.2.1 The inversion theorem

The *inversion theorem* asserts that a random variable is entirely specified by its characteristic function, meaning that X and Y have the same characteristic function if and only if they have the same distribution. We state the inversion theorem only for continuous distributions:

Theorem 13.13 (Fourier inversion theorem)
If X is continuous with density function f and characteristic function ϕ, then

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi(t)\,dt$$

at every point x at which f is differentiable.

[*Proof omitted.*]

13.2.2 The continuity theorem

For a sequence of random variables $X_1, X_2, \ldots$, the *continuity theorem* asserts that if the cumulative distribution functions $F_1, F_2, \ldots$ of the sequence approaches some limiting distribution F, then the charactaristic functions $\phi_1, \phi_2, \ldots$ of the sequence approaches the characteristic function of F.

Definition 13.14
A sequence of distribution functions $F_1, F_2, \ldots$ is said to *converge* to the distribution function F, denoted by $F_n \to F$, if $F_n(x) \to F(x)$ as $n \to \infty$ at each point x at which F is continuous.

Theorem 13.15 (Continuity theorem)
Let $F_1, F_2, \ldots$ and F be distribution functions, and let $\phi_1, \phi_2, \ldots$ and ϕ denote the corresponding characteristic functions.

(1) If $F_n \to F$ then $\phi_n(t) \to \phi(t)$ for all t.

(2) If $\phi_n(t) \to \phi(t)$ then $F_n \to F$ provided $\phi(t)$ exists and is continuous at $t = 0$.

[*Proof omitted.*]

13.3 Exercises

Exercise 13.1

1. Let X be a discrete random variable, taking values in the set $\{-3, -2, -1, 0, 1, 2, 3\}$ with uniform probability, and let $M(t)$ denote the MGF of X.
   (1) Show that $M(t) = \frac{1}{7}(e^{-3t} + e^{-2t} + e^{-t} + 1 + e^t + e^{2t} + e^{3t})$.
   (2) Use $M(t)$ to compute the mean and variance of X.

2. A continuous random variable X has MGF given by $M(t) = \exp(t^2 + 3t)$. Find the distribution of X.

3. Let X be a discrete random variable with probability mass function

$$\mathbb{P}(X = k) = \begin{cases} q^k p & k = 0, 1, 2, \dots \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < p < 1$ and $q = 1 - p$.

(1) Show that the MGF of X is given by $M(t) = \dfrac{p}{1 - qe^t}$ for $t < -\log q$.

(2) Find the PGF of X.

(3) Use the PGF of X to find the PMF of $Y = X + 1$.

(4) Use $M(t)$ to find the mean and variance of X.

4. Let $M(t)$ denote the MGF of the normal distribution $\mathrm{N}(0, \sigma^2)$. By exanding $M(t)$ as a power series in t, show that the moments μ_k of the $\mathrm{N}(0, \sigma^2)$ distribution are zero if k is odd, and equal to

$$\mu_{2m} = \frac{\sigma^{2m}(2m)!}{2^m m!} \qquad \text{if } k = 2m \text{ is even.}$$

5. Let $X \sim \text{Exponential}(\theta)$ where θ is a scale parameter.

(1) Show that the MGF of X is $M(t) = \dfrac{1}{1 - \theta t}$.

(2) By expanding this expression as a power series in t, find the first four non-central moments of X.

(3) Find the skewness γ_1 and the excess kurtosis γ_2 of X.

6. Let $X_1, X_2, \dots$ be independent and identically distributed random variables, with each $X_i \sim \mathrm{N}(\mu, \sigma^2)$.

(1) Find the MGF of the random variable $\bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.

(2) Show that $\bar{X}$ has a normal distribution, and find its mean and variance.

7. Let $X_1 \sim \text{Gamma}(k_1, \theta)$ and $X_2 \sim \text{Gamma}(k_2, \theta)$ be independent random variables. Use the MGFs of X_1 and X_2 to find the distribution of the random variable $Y = X_1 + X_2$.

8. A coin has probability p of showing heads. The coin is tossed repeatedly until exactly k heads occur. Let N be the number of times the coin is tossed. Using the continuity theoram for characteristic functions, show that the distribution of the random variable $X = 2pN$ converges to a gamma distribution as $p \to 0$.

9. Let X and Y be independent and identically distributed random variables, with means equal to 0 and variances equal to 1. Let $\phi(t)$ denote their common characteristic function, and suppose that the random variables $X + Y$ and $X - Y$ are independent. Show that $\phi(2t) = \phi(t)^3\phi(-t)$, and hence deduce that X and Y must be independent standard normal variables.

Lecture 14 The Law of Large Numbers

14.1 Convergence

We define the following notions of convergence for sequences of random variables.

Definition 14.1
Let $X_1, X_2, \ldots$ 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 mean* if $\mathbb{E}(|X_n - X|) \to 0$ as $n \to \infty$,

(4) $X_n \to X$ *in probability*, if for all $\epsilon > 0$, $\mathbb{P}(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$,

(5) $X_n \to X$ *in distribution* if $F_n(x) \to F(x)$ as $n \to \infty$ for every point x at which F is continuous.

Theorem 14.2
(1) Convergence almost surely implies convergence in probability.

(2) Convergence in mean square implies convergence in mean.

(3) Convergence in mean implies convergence in probability.

(4) Convergence in probability implies convergence in distribution.

[*Proof omitted.*]

14.2 The law of large numbers

Theorem 14.3 (The weak law of large numbers)
Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having 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:

Theorem 14.4 (The law of large numbers: convergence in mean square)
Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having finite mean μ, and finite variance. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{in mean square as } n \to \infty.$$

Proof:

Remark 14.5 (Frequentist probability)
A random 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. Then the sample mean of the X_i is the *relative frequency* of event A over these n repetitions, and by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{P}(A) \quad \text{as } n \to \infty.$$

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.

14.3 Bernoulli's law of large numbers*

In the proof of Theorem 14.3, we used Chebyshev's inequality to show that

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \qquad \forall\, \epsilon > 0.$$

We say that the *rate* at which $\bar{X}_n \to \mu$ is of order $O(1/n)$ as $n \to \infty$. In the proof of the following theorem, we use Bernstein's inequality to show that the sample mean of Bernoulli random variables satisfies

$$\mathbb{P}\left(|\bar{X}_n - \mu| > \epsilon\right) \leq e^{-\frac{1}{2}n\epsilon^2} \qquad \forall\, \epsilon > 0.$$

In this case, the rate at which $\bar{X}_n \to \mu$ as $n \to \infty$ is said to be *exponentially fast.*

Theorem 14.6 (Bernoulli's Law of Large Numbers)
Let $X_1, X_2, \ldots$ be independent, with each $X_i \sim \text{Bernoulli}(p)$, and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean of the first n variables in the sequence. Then for every $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - p| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty.$$

Proof:

14.4 Exercises

Exercise 14.1

1. Let c be a constant, and let $X_1, X_2, \ldots$ be a sequence of random variables with $\mathbb{E}(X_n) = c$ and $\mathrm{Var}(X_n) = 1/\sqrt{n}$ for each n. Show that the sequence converges to c in probability as $n \to \infty$.

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

Lecture 15 The Central Limit Theorem

We will need the following result from elementary analysis:

Lemma 15.1
For any constant $c \in \mathbb{R}$,

$$\left(1 + \frac{c}{n}\right)^n \to e^c \quad \text{as} \quad n \to \infty.$$

Proof:

We will also need the following analogue of Theorem 13.5, which is a consequence of Taylor's theorem for functions of a complex variable. Here, $o(t^k)$ denotes a quantity with the property that $o(t^k)/t^k \to 0$ in the limit as $t \to 0$, and represents an 'error' term that is asymptotically smaller than the other terms of the expression in the limit as $t \to 0$, and which can therefore be neglected when t is sufficiently small. (This is called *Landau notation.*)

Theorem 15.2
If $\mathbb{E}(|X^k|) < \infty$, then

$$\phi(t) = \sum_{j=0}^{k} \frac{\mathbb{E}(X^j)}{j!}(it)^j + o(t^k) \qquad \text{as} \quad t \to 0,$$

[*Proof omitted.*]

15.1 Poisson limit theorem

Theorem 15.3 (The Poisson limit theorem)
If $X_n \sim \text{Binomial}(n, \lambda/n)$ then the distribution of X_n converges to the Poisson(λ) distribution as $n \to \infty$.

Proof:

15.2 Law of large numbers

Theorem 15.4
Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with common mean $\mu < \infty$. Then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \quad \text{in distribution as} \quad n \to \infty.$$

- Unlike Theorem 14.3, this result does not require that the X_i have bounded variance.
- Convergence in distribution is however a weaker property than convergence in probability.

Proof:

15.3 Central limit theorem

Let $X_1, X_2, \ldots$ be i.i.d. random variables, and consider the partial sums

$$S_n = X_1 + X_2 + \ldots + X_n.$$

By independence, $\mathbb{E}(S_n) = n\mu$ and $\text{Var}(S_n) = n\sigma^2$.

The central limit theorem says that, *irrespective of the distribution of the* X_i, the distribution of the standardised variables

$$S_n^* = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{\text{Var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges to the standard normal distribution as $n \to \infty$.

Theorem 15.5 (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 normalised sums

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}} \qquad \text{where} \qquad S_n = X_1 + \ldots + X_n,$$

converges to the standard normal distribution $\text{N}(0, 1)$ as $n \to \infty$.

Proof:

Example 15.6 (Erlang Distribution)
The *Erlang distribution* with parameters $k \in \mathbb{N}$ and $\lambda > 0$ is defined to be the sum of k independent and identically distributed random variables $X_1, X_2, \ldots, X_k$, where each X_i is exponentially distributed with (rate) parameter λ. Show that if $Y \sim \text{Erlang}(k, \lambda)$, then the random variable

$$Z_k = \frac{\lambda Y - k}{\sqrt{k}}$$

has approximately the standard normal distribution when k is large.

Solution:

15.4 Exercises

Exercise 15.1

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.

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.

3. Let $X \sim \text{Binomial}(n_1, p_1)$ and $X_2 \sim \text{Binomial}(n_2, p_2)$ be independent random variables.

   (1) Use the central limit theorem to find the approximate distribution of $Y = X_1 - X_2$ when n_1 and n_2 are both large.

(2) Let $Y_1 = X_1/n_1$ and $Y_2 = X_2/n_2$. Show that $Y_1 - Y_2$ is approximately normally distributed with mean $p_1 - p_2$ and variance $\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$ when n_1 and n_2 are both large.

(3) Show that when n_1 and n_2 are both large,

$$\frac{(Y_1 - Y_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1) \qquad \text{approx.}$$

4. 5% of items produced by a factory production line are defective. Items are packed into boxes of 2000 items. As part of a quality control exercise, a box is chosen at random and found to contain 120 defective items. Use the central limit theorem to estimate the probability of finding at least this number of defective items when the production line is operating properly.

5. Use the central limit theorem to prove the law of large numbers.

6. 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.

Lecture 16 Joint Distributions

16.1 Joint distributions

Definition 16.1
Let $X, Y : \Omega \to \mathbb{R}$ be random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

(1) The *joint distribution* of X and Y is the function

$$\begin{array}{rccc} \mathbb{P}_{X,Y}: & \mathcal{B}^2 & \to & [0,1] \\ & (A,B) & \mapsto & \mathbb{P}(X \in A, Y \in B). \end{array}$$

(2) The *joint CDF* of X and Y is

$$\begin{array}{rccc} F_{X,Y}: & \mathbb{R}^2 & \to & [0,1] \\ & (x,y) & \mapsto & \mathbb{P}(X \le x, Y \le y). \end{array}$$

(3) The *marginal CDF* of X is the function

$$\begin{array}{rccc} F_X: & \mathbb{R} & \to & [0,1] \\ & x & \mapsto & \mathbb{P}(X \le x), \end{array}$$

and the marginal CDF of Y is

$$\begin{array}{rccc} F_Y: & \mathbb{R} & \to & [0,1] \\ & y & \mapsto & \mathbb{P}(Y \le y). \end{array}$$

16.2 Properties of Joint CDFs

Theorem 16.2
Let $F : \mathbb{R}^2 \to [0,1]$ be a joint CDF.

(1) Limiting behaviour:

$$\begin{array}{lll} \lim_{x\to-\infty} F(x,y) = 0, & \lim_{y\to-\infty} F(x,y) = 0, & \lim_{\substack{x\to-\infty \\ y\to-\infty}} F(x,y) = 0, \\ \lim_{x\to+\infty} F(x,y) = F_Y(y), & \lim_{y\to+\infty} F(x,y) = F_X(x), & \lim_{\substack{x\to+\infty \\ y\to+\infty}} F(x,y) = 1. \end{array}$$

(2) Monotonicity:

$$F(x,y) \le F(x+u, y+v) \quad \text{for all } u, v \ge 0.$$

(3) Inclusion-exclusion:

$$\mathbb{P}\big(a < X \le b,\ c < Y \le d\big) = F(b,d) - F(a,d) - F(b,c) + F(a,c).$$

(4) Upper continuity:

$$F(x+u, y+v) \longrightarrow F(x,y) \quad \text{as} \quad u \downarrow 0 \text{ and } v \downarrow 0,$$

where $u \downarrow 0$ means that u converges to zero through positive values (a.k.a. "from above").

[*Proof omitted.*]

16.3 Independent random variables

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

Definition 16.3
Two random variables $X, Y : \Omega \to \mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be *independent* if the events

$$\begin{aligned} \{X \leq x\} &\equiv \{\omega : X(\omega) \leq x\} \\ \{Y \leq y\} &\equiv \{\omega : Y(\omega) \leq y\} \end{aligned}$$

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

The following lemma is easily proved.

Lemma 16.4
Let X and Y be random variables with joint CDF $F_{X,Y}(x, y)$ and marginal CDFs $F_X(x)$ and $F_Y(y)$ respectively. Then X and Y are independent if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \text{for all} \quad x, y \in \mathbb{R}.$$

16.4 Identically distributed random variables

Definition 16.5
Two random variables $X, Y : \Omega \to \mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are said to be *identically distributed* if $\mathbb{P}_X = \mathbb{P}_Y$, or equivalently $F_X = F_Y$.

Thus X and Y are identically distributed if and only if

- $\mathbb{P}_X(B) \equiv \mathbb{P}(X \in B) = \mathbb{P}(Y \in B) \equiv \mathbb{P}_Y(B)$ for all $B \in \mathcal{B}$, or equivalently
- $F_X(t) \equiv \mathbb{P}(X \leq t) = \mathbb{P}(Y \leq t) \equiv F_Y(t)$ for all $t \in \mathbb{R}$.

16.5 Jointly discrete distributions

Definition 16.6

(1) Two random variables X and Y are called *jointly discrete* if the random vector (X, Y) only takes values in a countable subset of $\mathbb{R}^2$.

(2) Two jointly discrete random variables X and Y are described by their *joint PMF*:

$$\begin{aligned} f_{X,Y} : \quad \mathbb{R}^2 &\to [0, 1] \\ (x, y) &\mapsto \mathbb{P}(X = x, Y = y). \end{aligned}$$

(3) The *marginal PMF of* X is the function $f_X(x) = \mathbb{P}(X = x)$.

(4) The *marginal PMF of* Y is the function $f_Y(y) = \mathbb{P}(Y = y)$.

Example 16.7
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.

Solution:

The following lemma is easily proved.

Lemma 16.8
Two jointly discrete random variables X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \quad \text{for all} \quad x,y \in \mathbb{R}.$$

16.6 Jointly continuous distributions

Definition 16.9

(1) Two random variables X and Y are called *jointly continuous* if their joint CDF can be written as

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,du\,dv \qquad x,y \in \mathbb{R}$$

for some integrable function $f_{X,Y} : \mathbb{R}^2 \to [0,\infty)$ called the *joint PDF* of X and Y.

(2) The *marginal PDFs* of X and Y are defined by $f_X(x) = F'_X(x)$ and $f_Y(y) = F'_Y(y)$ respectively, where $F_X(x)$ and $F_Y(y)$ are the marginal CDFs of X and Y respectively.

Example 16.10
A dart is thrown at a circular dartboard of radius ρ. The point at which the dart hits the board determines a distance R from the centre, and an angle Θ with (say) the upward vertical. Assume that the dart does in fact hit the board, and that regions of equal area are equally likely to be hit. Show that R and Θ are jointly continuous random variables.

Solution:

16.6.1 Independence

Lemma 16.11
Two jointly continuous random variables X and Y are independent if and only if

(1) $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$, and

(2) the support of $f_{X,Y}$ is a rectangular region in $\mathbb{R}^2$.

Proof:

Remark 16.12
If the value taken by X affects the range of values taken by Y, then clearly Y dependes on X. Hence if X and Y are independent, we need that $\text{supp}(f_{X,Y})$ can be expressed as the Cartesian product of two sets in $\mathbb{R}$:

$$\text{supp}(f_{X,Y}) = \text{supp}(f_X) \times \text{supp}(f_Y) \qquad \text{where} \quad \text{supp}(f_X), \text{supp}(f_Y) \subseteq \mathbb{R}.$$

For example:

- The unit square is fine: $\text{supp}(f_{X,Y}) = \{(x,y) : 0 \leq x, y \leq 1\} = [0,1] \times [0,1]$.
- The unit disc is not: $\text{supp}(f_{X,Y}) = \{(x,y) : x^2 + y^2 \leq 1\}$.

Example 16.13
Two jointly continuous random variables X and Y have joint PDF

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

(1) Show that $c = 24$.

(2) Find the marginal PDFs of X and Y.

Solution:

Example 16.14
The continuous random variables X and Y have joint PDF

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

where c is a constant.

(1) Show that $c = 6$.

(2) Find the marginal PDFs of X and Y. Are X and Y independent?

(3) Show that $\mathbb{P}(X + Y \geq 1) = 9/10$.

Solution:

16.7 Exercises

Exercise 16.1

1. Let X be a Bernoulli random variable with parameter p.

(a) Let $Y = 1 - X$. Find the joint PMF of X and Y.

(b) Let $Y = 1 - X$ and $Z = XY$. Find the joint PMF of X and Z.

2. Let X and Y be two independent discrete random variables with the following PMFs:

x	1	2
$f_X(x)$	1/3	2/3

y	-1	0	1
$f_Y(y)$	1/4	1/2	1/4

(a) Compute the joint PMF of X and Y.

(b) Compute the joint PMF of the random variables $U = 1/X$ and $V = Y^2$.

(c) Show that U and V are independent.

3. Two discrete random variables X and Y have the following joint PMF:

$$f_{X,Y}(x,y) = \begin{cases} c|x+y| & \text{for} \quad x, y \in \{-2,-1,0,1,2\}, \\ 0 & \text{otherwise,} \end{cases}$$

where c is a constant.

(a) Show that $c = 1/40$.

(b) Find $\mathbb{P}(X = 0, Y = -2)$.

(c) Find $\mathbb{P}(X = 2)$.

(d) Find $\mathbb{P}(|X - Y| \leq 1)$.

4. Two continuous random variables X and Y have the joint PDF

$$f_{X,Y}(x,y) = \begin{cases} 2x & \text{if} \quad 0 \leq x, y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

(a) Find the conditional distribution of Y given that $X = x$.

(b) Find $\mathbb{P}(Y \leq 0.5 | X = 0.5)$ and $\mathbb{P}(Y \leq 0.5 | X = 0.75)$.

(c) Find the marginal distribution of Y and hence find $\mathbb{P}(Y \leq 0.5)$.

5. Two continuous random variables X and Y have the following joint PDF:

$$f_{X,Y}(x,y) = \begin{cases} c(x^2 + y) & \text{when } -1 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 - x^2, \\ 0 & \text{otherwise.} \end{cases}$$

where c is a constant.

(a) Show that $c = 5/4$.

(b) Find $\mathbb{P}(0 \leq X \leq 0.5)$.

(c) Find $\mathbb{P}(Y \leq X + 1)$.

(d) Find $\mathbb{P}(Y = X^2)$.

Lecture 17 Covariance and Correlation

17.1 Bivariate Distributions

Definition 17.1
Let $X, Y : \Omega \to \mathbb{R}$ be two random variables defined on the same probability space, let $F_{X,Y}$ denote their joint CDF, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a (Borel) measureable function on $\mathbb{R}^2$. Then

$$\mathbb{E}\big[g(X,Y)\big] = \iint g(x,y)\,dF(x,y)$$

whenever this integral exists. In particular,

(1) if X and Y are jointly discrete, with joint PMF $f_{X,Y}(x,y)$, then

$$\mathbb{E}\big[g(X,Y)\big] = \sum_{x,y} g(x,y) f_{X,Y}(x,y)$$

whenever this sum exists, and

(2) if X and Y are jointly continuous, with joint PDF $f_{X,Y}(x,y)$, then

$$\mathbb{E}\big[g(X,Y)\big] = \iint g(x,y) f_{X,Y}(x,y)\,dx\,dy$$

whenever this integral exists.

17.2 Covariance

Definition 17.2
The *product moment* of X and Y is defined to be

$$\mathbb{E}(XY) = \iint xy\,dF(x,y)$$

whenever this integral exists. In particular,

(1) if X and Y are jointly discrete, with joint PMF $f_{X,Y}(x,y)$, then

$$\mathbb{E}(XY) = \sum_{x,y} xy\, f_{X,Y}(x,y)$$

whenever this sum is absolutely convergent, and

(2) if X and Y are jointly continuous, with joint PDF $f_{X,Y}(x,y)$, then

$$\mathbb{E}(XY) = \iint xy\, f_{X,Y}(x,y)\,dx\,dy$$

whenever this integral is absolutely convergent.

Definition 17.3
(1) The *covariance* of X and Y is

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

(2) The *correlation coefficient* of X and Y is

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

Remark 17.4

- $\mathrm{Cov}(X,Y)$ is the product moment of the *centred* variables $X - \mathbb{E}(X)$ and $Y - \mathbb{E}(Y)$.
- $\rho(X,Y)$ is the product moment of the *standardized* variables $\dfrac{X-\mathbb{E}(X)}{\sqrt{\mathrm{Var}(X)}}$ and $\dfrac{Y-\mathbb{E}(Y)}{\sqrt{\mathrm{Var}(Y)}}$.

Remark 17.5 (Variance of sums of random variables)
For any random variables $X_1, X_2, \ldots, X_n$,

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{i=1}^n \sum_{j=i+1}^n \mathrm{Cov}(X_i, X_j).$$

17.3 Correlation

Correlation quantifies the (linear) dependence between random variables.

Definition 17.6
Two random variables X and Y are said to be *correlated* if $\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y)$.

Lemma 17.7
If X and Y are independent, they are uncorrelated.

We prove the lemma only for discrete random variables (the continuous case is similar).

Proof:

Theorem 17.8
If X and Y are uncorrelated, $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof:

Example 17.9
Let $Y_1, \ldots, Y_r$ be independent and identically distributed random variables, with each $Y_i \sim \mathrm{Geometric}(p)$. Find the mean and variance of their sum $X = \sum_{i=1}^r Y_i$.

Solution:

17.4 The Cauchy-Schwarz Inequality

Lemma 17.10
If $X \geq 0$ and $\mathbb{E}(X) = 0$ then $\mathbb{P}(X = 0) = 1$.

[*Proof omitted.*]

Theorem 17.11 (Cauchy-Schwarz inequality)
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 $a \in \mathbb{R}$.

Proof:

Corollary 17.12
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 $a \in \mathbb{R}$.

Proof:

17.5 Exercises

Exercise 17.1

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 common distribution has finite mean and variance.

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 $\mathrm{Cov}(X,Y)$.

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.

Lecture 18 Conditional Distributions

18.1 Conditional distributions

Let $X, Y : \Omega \to \mathbb{R}$ be random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 18.1
(1) The *conditional distribution* of Y given X is the function

$$\begin{array}{rccl} \mathbb{P}_{Y|X}: & \mathcal{B}^2 & \to & [0,1] \\ & (A,B) & \mapsto & \mathbb{P}(Y \in B \,|\, X \in A). \end{array}$$

(2) The *conditional CDF* of Y given X is the function

$$\begin{array}{rccl} F_{Y|X}: & \mathbb{R}^2 & \to & [0,1] \\ & (x,y) & \mapsto & \mathbb{P}(Y \leq y \,|\, X \leq x). \end{array}$$

The following lemma is easily proved.

Lemma 18.2
The conditional CDF of Y given 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.

18.1.1 Discrete case

Definition 18.3
Let X and Y be jointly discrete random variables, and let x be such that $\mathbb{P}(X = x) > 0$. The *conditional PMF* of Y given $X = x$ is the function

$$\begin{array}{rccc} f_{Y|X}: & \mathbb{R} & \to & [0,1] \\ & y & \mapsto & \mathbb{P}(Y = y \,|\, X = x). \end{array}$$

The following lemma is easily proved.

Lemma 18.4
The conditional PMF of Y given $X = x$ satisfies

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

Theorem 18.5
If X and Y are jointly discrete random variables, then

$$f_Y(y) = \sum_x f_{Y|X}(y|x) f_X(x).$$

where the sum is taken over the range of X.

Proof:

18.1.2 Continuous case

Let X and Y be jointly continuous random variables.

- Suppose we observe that X takes the value x.
- Since $\mathbb{P}(X = x) = 0$, we cannot condition on the event $\{X = x\}$.

Definition 18.6
Let X and Y be jointly continuous random variables. The *conditional PDF* of Y given $X = x$ is

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

Theorem 18.7
If X and Y are jointly continuous random variables, then

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

Proof:

18.2 Conditional expectation

Definition 18.8
(1) The *conditional expectation of* Y *given* $X = x$ is a number,

$$\mathbb{E}(Y|X = x) \;=\; \begin{cases} \displaystyle\sum_y y\, f_{Y|X}(y|x) & \text{(discrete case)}, \\ \displaystyle\int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy & \text{(continuous case)}. \end{cases}$$

(2) The *conditional expectation of* Y *given* X is a random variable,

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

Remark 18.9
Let $g(x) = \mathbb{E}(Y|X = x)$. The distribution of the random variable $g(X) = \mathbb{E}(Y|X)$ depends only on the distribution of X. Its expectation is given by

$$\mathbb{E}\big[\mathbb{E}(Y|X)\big] \;=\; \begin{cases} \displaystyle\sum_x \mathbb{E}(Y|X = x) f_X(x) & \text{(discrete case)}, \\ \displaystyle\int_{-\infty}^{\infty} \mathbb{E}(Y|X = x) f_X(x)\, dx & \text{(continuous case)}. \end{cases}$$

where f_X is the marginal PMF or PDF of X.

18.3 Law of total expectation

Theorem 18.10 (Law of total expectation)
Let X and Y be random variables on the same probability space. Then

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

We prove the theorem for continuous random variables (the discrete case follows similarly).

Proof:

18.4 Law of total variance

Theorem 18.11 (Law of total variance)
Let X and Y be random variables on the same probability space. Then

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

This is sometimes called the *variance decomposition formula*.

Proof:

18.4.1 Variance decomposition

$\mathbb{E}(Y|X)$ can be thought of as a *model* of Y in terms of X.

- $\text{Var}\big[\mathbb{E}(Y|X)\big]$ is the variance of the model. This is called the *explained variance*.
- $Y - \mathbb{E}(Y|X)$ is called the *residual*, representing that part of Y not explained by the model $\mathbb{E}(Y|X)$.
- $\text{Var}(Y|X) = \mathbb{E}\big([Y - \mathbb{E}(Y|X)]^2 \,\big|\, X\big)$ is called the *residual variance* at X.
- $\mathbb{E}\big[\text{Var}(Y|X)\big]$ is the expected residual variance. This is called the *unexplained variance*.

The law of total variance divides the variance into *unexplained* and *explained* components:

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

or

$$\frac{\mathbb{E}\big[\text{Var}(Y|X)\big]}{\text{Var}(Y)} + \frac{\text{Var}\big[\mathbb{E}(Y|X)\big]}{\text{Var}(Y)} = 1.$$

This idea is important in statistics.

18.4.2 Linear models

Suppose we adopt a *linear model* of Y against X:

$$\mathbb{E}(Y|X) = a + bX.$$

It can be shown that the residual variance is minimised when

$$a = \mathbb{E}(Y) - \left[\frac{\text{Cov}(X,Y)}{\text{Var}(X)}\right]\mathbb{E}(X) \qquad \text{and} \qquad b = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}.$$

The proportion of the total variance explained by the model is the square of the correlation coefficient:

$$\frac{\text{Var}\big[\mathbb{E}(Y|X)\big]}{\text{Var}(Y)} = \rho(X,Y)^2.$$

This is known as the *coefficient of determination*, usually denoted by R^2, which quantifies the extent to which a linear model $Y = a + bX$ captures the relationship (if any) between X and Y.

18.5 Example

Example 18.12
The jointly continuous random variables X and Y have following joint PDF:

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

(1) Find the marginal PDFs of X and Y.

(2) Find the mean and variance of Y.

(3) Find the conditional PDF of X given $Y = y$.

(4) Find the conditional PDF of Y given $X = x$.

(5) Are X and Y independent?

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

Solution:

18.6 Exercises

Exercise 18.1

1. A fair coin is tossed three times. Let I_j be the indicator variable of the event that a head occurs on the jth toss. Compute the conditional expectation $E(Y|X)$ and verify the identity $E(E(Y|X)) = E(Y)$ in each of the following cases:

   (1) $X = \max\{I_1, I_2, I_3\}$ and $Y = \min\{I_1, I_2, I_3\}$,

   (2) $X = I_1 + I_2$ and $Y = I_2 + I_3$.

2. Let X and Y be continuous random variables with joint density function

$$f(x,y) = \begin{cases} c(x+y) & 0 \leq x, y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

   (1) Show that $c = 1$.

   (2) Compute the conditional expectation $\mathbb{E}(Y|X)$.

   (3) Verify the identity $\mathbb{E}\big(\mathbb{E}(Y|X)\big) = \mathbb{E}(Y)$.

3. Let the joint density of random variables X and Y be

$$f(x,y) = \begin{cases} cxy & \text{for} \quad 0 \leq x, y \leq 1 \text{ where } \ x + y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

   (1) Compute the normalization constant c.

   (2) Compute the conditional expectation $\mathbb{E}(Y|X)$.

   (3) Verify the identity $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$.

Lecture 19 The Bivariate Normal Distribution

19.1 Bivariate transformations

Definition 19.1
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 19.2
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

$$f_{X,Y}(x, y) = |J| f_{U,V}\left[g^{-1}(x, y)\right] \quad \text{where} \quad J = \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{vmatrix} \quad \text{with} \quad (u, v) = g^{-1}(x, y).$$

Remark 19.3
The absolute value $|J|$ is a scale factor, which ensures that $f_{X,Y}(x, y)$ integrates to one.

Example 19.4
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 $U, V \sim \text{Exponential}(1)$ are independent, find the joint PDF of X and Y.

Solution:

19.2 The bivariate normal distribution

Theorem 19.5
if $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, then

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

[*Proof omitted.*]

Corollary 19.6
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 19.7
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 19.8
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 19.9
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:

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 19.10
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:

19.3 Properties of the bivariate normal distribution

Theorem 19.11
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:

19.4 Conditional distributions

Theorem 19.12
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

$$\begin{aligned}\mathbb{E}(Y|X=x) &= \mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(x-\mu_1),\\ \mathrm{Var}(Y|X=x) &= \sigma_2^2(1-\rho^2),\end{aligned}$$

and the conditional mean and variance of Y given X is

$$\begin{aligned}\mathbb{E}(Y|X) &= \mathbb{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\big[X - \mathbb{E}(X)\big],\\ \mathrm{Var}(Y|X) &= \mathrm{Var}(Y)(1-\rho^2).\end{aligned}$$

Proof:

19.5 Exercises

Exercise 19.1

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

   (b) Check that $\text{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)\,dxdy = \rho$.

   (c) Show that if X and Y are uncorrelated, then they are independent.

2. Let X and Y have standard bivariate normal distribution. Find the conditional distribution of Y given $X = x$, and hence show that $\mathbb{E}(Y|X) = \rho X$.

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.

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 $\mathbb{E}(Z^2) = 1$.

5. Let $U, V \sim N(0,1)$. Show that the random variables $X = U + V$ and $Y = U - V$ 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

$$N\left(\mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(x - \mu_1), \sigma_2^2(1-\rho^2)\right).$$

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

$$f_{X+Y}(t) = \int_{-\infty}^{\infty} f_{X,Y}(x, t-x)\,dx = \int_{-\infty}^{\infty} f_{X,Y}(t-y, y)\,dy.$$

   (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 19.5.)