Cardiff University School of Mathematics

# MA1500

# Introduction to Probability Theory

## Lecture Notes

## 2014-15

Version: 30/1/2015

# Contents


**1 Set Theory** **1**
- 1.1 Elementary set theory . . . 1
- 1.2 De Morgan's laws . . . 2
- 1.3 Set difference . . . 3
- 1.4 Exercises . . . 3

**2 Events** **5**
- 2.1 Sample spaces . . . 5
- 2.2 Events . . . 5
- 2.3 Families of events . . . 6
- 2.4 Terminology . . . 7
- 2.5 Exercises . . . 7

**3 Probability** **9**
- 3.1 Probability measures . . . 9
- 3.2 Properties of probability measures . . . 10
- 3.3 Exercises . . . 10

**4 Conditional Probability** **13**
- 4.1 Conditional probability . . . 13
- 4.2 The partition theorem . . . 14
- 4.3 Bayes' theorem . . . 15
- 4.4 Exercises . . . 15

**5 Independence** **19**
- 5.1 Independence . . . 19
- 5.2 Pairwise independence and total indepencence . . . 19
- 5.3 Exercises . . . 20

**6 Classical Probability** **23**
- 6.1 The principle of indifference . . . 23
- 6.2 Sampling with replacement . . . 23
- 6.3 Sampling without replacement . . . 24
- 6.4 Exercises . . . 26

**7 The Frequentist Model** **29**
- 7.1 Finite probability spaces . . . 29
- 7.2 Relative frequency . . . 30
- 7.3 Conditional probability . . . 31
- 7.4 Infinite sample spaces . . . 32

**8 Random Variables** **33**
- 8.1 Random variables . . . 33

8.2 Cumulative distribution functions (CDFs) . . . 35
8.3 Probability mass functions (PMFs) . . . 36
8.4 Finite probability spaces . . . 36
8.5 Exercises . . . 38

**9 Expectation** **42**
9.1 Expectation . . . 42
9.2 Properties of expectation . . . 44
9.3 The sample mean . . . 45
9.4 Exercises . . . 45

**10 Moments** **48**
10.1 Transformations of random variables . . . 48
10.2 Moments . . . 49
10.3 Variance . . . 49
10.4 Location, scale and shape* . . . 51
10.5 Skewness and kurtosis* . . . 52
10.6 Exercises . . . 52

**11 The Uniform, Bernoulli and Binomial Distributions** **56**
11.1 Uniform distribution . . . 56
11.2 Bernoulli distribution . . . 57
11.3 Binomial distribution . . . 58
11.4 Exercises . . . 59

**12 Joint Distributions** **63**
12.1 Joint distributions . . . 63
12.2 Independent random variables . . . 64
12.3 Exercises . . . 66

**13 Covariance and Correlation** **69**
13.1 Product moments . . . 69
13.2 Correlation . . . 70
13.3 Covariance . . . 70
13.4 The correlation coefficient . . . 71
13.5 Exercises . . . 72

**14 Conditional Distributions** **74**
14.1 Conditional distributions . . . 74
14.2 Conditional expectation . . . 75
14.3 Law of total expectation . . . 75
14.4 Exercises . . . 76

**15 Discrete Probability** **82**
15.1 Cardinality . . . 82
15.2 Discrete probability spaces . . . 83
15.3 Discrete random variables . . . 83
15.4 Convergent Series . . . 84
15.5 Expectation . . . 85
15.6 Exercises . . . 86

**16 The Geometric and Poisson Distributions** **88**
16.1 Geometric distribution . . . 88
16.2 Poisson distribution . . . 90

16.3 The law of rare events . . . 91
16.4 Exercises . . . 93

**17 Continuous Probability** **97**
17.1 General probability spaces . . . 97
17.2 Continuous random variables . . . 98
17.3 Independence . . . 99
17.4 Expectation . . . 99
17.5 Exercises . . . 101

**18 The Uniform, Exponential and Normal Distributions** **106**
18.1 The uniform distribution . . . 106
18.2 The negative exponential distribution . . . 107
18.3 The normal distribution . . . 108
18.4 The chi-squared distribution* . . . 111
18.5 Exercises . . . 112

**19 Quantiles** **116**
19.1 Median . . . 116
19.2 Quantiles . . . 116
19.3 The quantile function . . . 118
19.4 Exercises . . . 119

**20 Normal Approximation** **121**
20.1 Approximation of discrete distributions by continuous distributions . . . 121
20.2 Normal approximation of the binomial distribution . . . 122
20.3 Normal approximation of the Poisson distribution . . . 122
20.4 Exercises . . . 123

# 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 $\emptyset$.

Algebra is the study of *operations* and *relations*.

- The basic relations of set algebra are *set inclusion* and *set equality*.
- The basic operations of set algebra are *complementation*, *union* and *intersection*.

### 1.1.1 Set relations

**Definition 1.1**
Let $A$ and $B$ be sets.

(1) If every element of $A$ is also an element of $B$, we say that $A$ is a *subset* of $B$.
This is denoted by $A \subseteq B$.

(2) If every element of $A$ is an element of $B$, and every element of $B$ is an element of $A$, we say that $A$ and $B$ are *equal*.
This is denoted by $A = B$.

(3) If $A$ is a subset of $B$, but $A$ is not equal to $B$, we say that $A$ is a *proper subset* of $B$.
This is denoted by $A \subset B$.

**Example 1.2**
Let $A = \{a, b\}$, $B = \{a, b\}$ and $C = \{a, b, c\}$.

- $A$ is a subset of $B$: $A \subseteq B$,
- $A$ is also equal to $B$: $A = B$, and
- $A$ is a proper subset of $C$: $A \subset C$.

### 1.1.2 Set operations

**Definition 1.3**
Let $A$, $B$ and $\Omega$ be sets, with $A, B \subseteq \Omega$.

(1) The *union* of $A$ and $B$ is the set

$$A \cup B = \{a \in \Omega : a \in A \text{ or } a \in B\}.$$

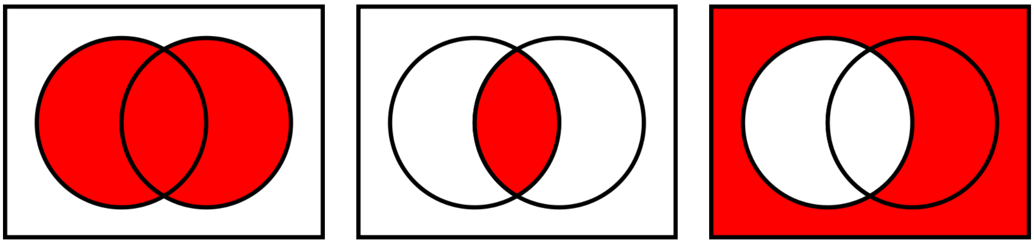
(2) The *intersection* of $A$ and $B$ is the set

$$A \cap B = \{a \in \Omega : a \in A \text{ and } a \in B\}.$$

(3) The *complement* of $A$ is the set

$$A^c = \{a \in \Omega : a \notin A\}.$$

**Example 1.4**
Let $A = \{a, b\}$, $B = \{b, c\}$ and $\Omega = \{a, b, c, d\}$.
Then $A \cup B = \{a, b, c\}$, $A \cap B = \{b\}$ and $A^c = \{c, d\}$.



Union Intersection Complement

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

### 1.1.3 Set algebra

**Definition 1.5**

(1) Commutative property.

- $A \cup B = B \cup A$,
- $A \cap B = B \cap A$.

(2) Associative property.

- $(A \cup B) \cup C = A \cup (B \cup C)$,
- $(A \cap B) \cap C = A \cap (B \cap C)$.

(3) Distributive property.

- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$,
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.

**Remark 1.6**
A statement such as $A \cup B \cap C$ is ambiguous.

## 1.2 De Morgan's laws

Union and intersection swap roles under complementation.

**Theorem 1.7**

(1) $(A \cup B)^c = A^c \cap B^c$.

(2) $(A \cap B)^c = A^c \cup B^c$.

**Proof:**

(1) Let $a \in (A \cup B)^c$. Then $a \notin A$ and $a \notin B$, so $a \in A^c \cap B^c$. Hence $(A \cup B)^c \subseteq A^c \cap B^c$.
Let $a \in A^c \cap B^c$. Then $a \notin A$ and $a \notin B$, so $a \notin A \cup B$. Hence $A^c \cap B^c \subseteq (A \cup B)^c$.
Thus it follows that $(A \cup B)^c = A^c \cap B^c$.

(2) Apply part (1) to the sets $A^c$ and $B^c$: $(A^c \cup B^c)^c = A \cap B$.
Then take the complement of both sides: $(A \cap B)^c = A^c \cup B^c$.

## 1.3 Set difference

**Definition 1.8**
Let $A$, $B$ and $\Omega$ be sets, with $A, B \subseteq \Omega$.

(1) The *set difference* between $A$ and $B$ is the set

$$A \setminus B = \{a \in \Omega : a \in A \text{ and } a \notin B\}.$$

(2) The *symmetric difference* between $A$ and $B$ is the set

$$A \triangle B = (A \setminus B) \cup (B \setminus A).$$

- $A \setminus B$ is the set of points that are in $A$ but not in $B$.
- $A \triangle B$ is the set of points that are in either $A$ or $B$, but not both.

**Example 1.9**
Let $A = \{a, b\}$ and $B = \{b, c\}$. Then

- $A \setminus B = \{a\}$
- $A \triangle B = \{a, c\}$.

## 1.4 Exercises

**Exercise 1.1**

1. Illustrate the basic set operations using Venn diagrams.
2. State and prove De Morgan's laws.

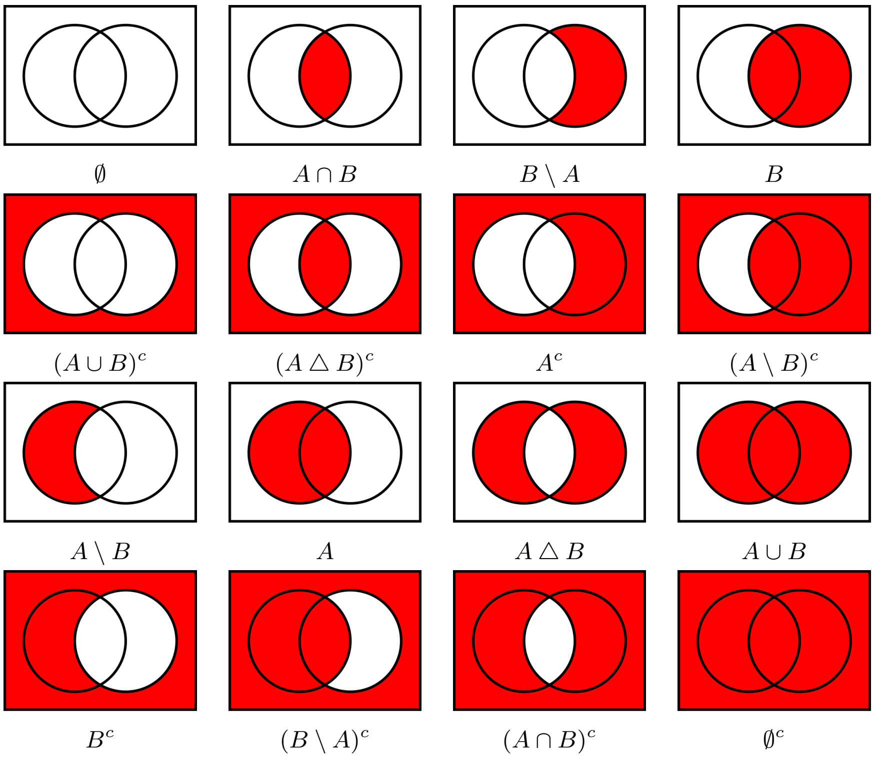



Figure 1.2: Set Operations ($A$ on the left, $B$ on the right).

# Lecture 2 Events

## A brief history of probability

Games of chance have been played since antiquity, but the mathematical principles of chance and uncertainty were first established only in the 17th century.

| | | |
|---|---|---|
| 1654 | Classical principles | Blaise Pascal (1623–1662) |
| | | Pierre de Fermat (1601–1665) |
| 1657 | *De Ratiociniis in Ludo Aleae* | Christiaan Huygens (1629–1695) |
| 1713 | *Ars Conjectandi* | Jakob Bernoulli (1654–1705) |
| 1718 | *The Doctrine of Chances* | Abraham de Moivre (1667–1754) |
| 1812 | *Theorie Analytique des Probabilites* | Pierre de Laplace (1749-1827) |
| 1919 | Relative frequency | Richard von Mises (1883–1953) |
| 1933 | Modern axiomatic theory | Andrey Kolmogorov (1903–1987) |

## 2.1 Sample spaces

**Definition 2.1**

(1) Any process of observation or measurement will be called an *experiment* or *trial*.

(2) Any experiment whose outcome is uncertain is called a *random experiment*.

(3) A random experiment has a set of possible *outcomes*.

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

(5) The set of all possible outcomes is called the *sample space* of the experiment, denoted by $\Omega$.

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

**Example 2.2**
For any random experiment, the sample space is the set of all possible outcomes:

| Experiment | Sample space |
|---|---|
| A coin is tossed once. | $\Omega = \{H, T\}$ |
| A six-sided die is rolled once. | $\Omega = \{1, 2, 3, 4, 5, 6\}$ |
| A coin is tossed repeatedly until a head occurs. | $\Omega = \{1, 2, 3, \ldots\}$ |
| The height of a randomly chosen student is measured: | $\Omega = [0, \infty)$ |

## 2.2 Events

**Definition 2.3**

- An *event* $A$ is a subset of the sample space, $\Omega$.
- If outcome $\omega$ occurs, we say that event $A$ *occurs* if and only if $\omega \in A$.
- Two events $A$ and $B$ with $A \cap B = \emptyset$ are called *disjoint* or *mutually exclusive*.
- The empty set $\emptyset$ is called the *impossible event*.

- The sample space $\Omega$ is called the *certain event*.

**Remark 2.4**

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

**Example 2.5**
A die is rolled once. The sample space can be represented by $\Omega = \{1, 2, 3, 4, 5, 6\}$.
We may be interested in whether or not the following events occur:

| Event | Subset |
|---|---|
| The outcome is the number 1. | $A = \{1\}$ |
| The outcome is an even number. | $A = \{2, 4, 6\}$ |
| The outcome is even but does not exceed 3. | $A = \{2, 4, 6\} \cap \{1, 2, 3\}$ |
| The outcome is not even | $A = \Omega \setminus \{2, 4, 6\}$ |

## 2.3 Families of events

**Definition 2.6**
Let $\Omega$ be any set.

(1) The set of all subsets of $\Omega$ is called its *power set*, which we denote by $\mathcal{P}(\Omega)$.

(2) Any subset of $\mathcal{P}(\Omega)$ is called a *family of sets over* $\Omega$.

Let $\Omega$ be the sample space of some random experiment. If we are interested in events $A$ and $B$, we must also be interested in whether:

- event $A$ occurs *or* event $B$ occurs – this is the event $A \cup B$,
- event $A$ occurs *and* event $B$ occurs – this is the event $A \cap B$,
- event $A$ does *not* occur – this is the event $A^c$.

We cannot therefore use arbitrary families of sets over $\Omega$ as the basis for investigating random experiments. Instead, we allow only families that are *closed* under certain set operations.

**Definition 2.7**
A family of sets $\mathcal{F}$ over $\Omega$ is said to be

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

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

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

**Definition 2.8**
A family of sets $\mathcal{F}$ over $\Omega$ is called a *field of sets* over $\Omega$ if

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

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

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

**Example 2.9**
A six-sided die is rolled once, and the score is observed. A suitable sample space for this experiment is the set $\Omega = \{1, 2, 3, 4, 5, 6\}$. The power set of $\Omega$ will always provide a field of sets to work with. However, suppose we are only interested in whether or not the outcome is an even number. In this case, we need only consider the following family of events:

$$\mathcal{F} = \big\{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \{1, 2, 3, 4, 5, 6\}\big\}.$$

We can see that $\mathcal{F}$ is a field of sets over $\Omega$, because

(1) it contains the sample space $\{1, 2, 3, 4, 5, 6\}$,

(2) the complement of every set in $\mathcal{F}$ is also contained in $\mathcal{F}$, and

(3) the union of any two sets in $\mathcal{F}$ is also contained in $\mathcal{F}$.

**Theorem 2.10 (Properties of fields)**
Let $\mathcal{F}$ be a field over $\Omega$. Then

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

(2) $\mathcal{F}$ is closed under pairwise intersections,

(3) $\mathcal{F}$ is closed under set differences.

**Proof:**

(1) We know that $\emptyset = \Omega^c$, and that $\Omega \in \mathcal{F}$. Because $\mathcal{F}$ is closed under complementation, it thus follows that $\emptyset \in \mathcal{F}$.

(2) Let $A, B \in \mathcal{F}$. By De Morgan's laws, we have that $A \cap B = (A^c \cup B^c)^c$. Because $\mathcal{F}$ is closed under complementation and pairwise unions, it thus follows that $A \cap B \in \mathcal{F}$.

(3) Let $A, B \in \mathcal{F}$. Set difference can be written as $A \setminus B = A \cap B^c$. Furthermore, by De Morgan's laws we see that $A \cap B^c = (A^c \cup B)^c$. Because $\mathcal{F}$ is closed under complementation and pairwise unions, it thus follows that $A \setminus B \in \mathcal{F}$.

## 2.4 Terminology

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

## 2.5 Exercises

**Exercise 2.1**

1. Identify a sample space, and the subset corresponding to event $A$, in each of the following scenarios:
   (a) A coin is tossed three times. $A$ is the event that at least two heads are obtained.
   (b) A game of football is played. $A$ is the event that the match ends in a draw.
   (c) A couple have two children. $A$ is the event that both are girls.
   (d) A shot hits a circular target of radius 10cm. $A$ is the event that the shot hits within 3cm of the centre.

2. A family of sets $\mathcal{F}$ over $\Omega$ is said to be
   - *closed under finite unions* if $A_1 \cup A_2 \cup \ldots \cup A_n \in \mathcal{F}$ whenever $A_1, A_2, \ldots A_n \in \mathcal{F}$,
   - *closed under finite intersections* if $A_1 \cap A_2 \cap \ldots \cap A_n \in \mathcal{F}$ whenever $A_1, A_2, \ldots A_n \in \mathcal{F}$.

Suppose that $\mathcal{F}$ is a field of sets over $\Omega$. Show that

(a) $\mathcal{F}$ is closed under finite unions, and that

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

# Lecture 3 Probability

## 3.1 Probability measures

Probability is defined to be a *function* that assigns numerical value to random events.

**Definition 3.1**
Let $\Omega$ be the sample space of some random experiment, and let $\mathcal{F}$ be a field of sets over $\Omega$. A *probability measure* on $(\Omega, \mathcal{F})$ is a function

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

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 3.2**
- The second property is called *countable additivity.*
- The number $\mathbb{P}(A)$ is called the *probability* of event $A \in \mathcal{F}$.

**Example 3.3**
Consider a random experiment in which a fair six-sided die is rolled once.

- A suitable sample space for the experiment is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- A suitable field of events for the experiment is the power set, $\mathcal{F} = \mathcal{P}(\Omega)$.
- Because the die is fair, a suitable probability measure is given by the function

$$\begin{aligned} \mathbb{P}: \mathcal{F} &\to [0,1] \\ A &\mapsto \tfrac{1}{6}|A|, \end{aligned} \qquad \text{where } |A| \text{ denotes the cardinality of } A.$$

| Event | Event | Probability |
|---|---|---|
| The outcome is the number 1. | $A = \{1\}$ | $\mathbb{P}(A) = 1/6$ |
| The outcome is an even number. | $A = \{2, 4, 6\}$ | $\mathbb{P}(A) = 3/6$ |
| The outcome is even but does not exceed 3. | $A = \{2, 4, 6\} \cap \{1, 2, 3\}$ | $\mathbb{P}(A) = 1/6$ |
| The outcome is not even | $A = \Omega \setminus \{2, 4, 6\}$ | $\mathbb{P}(A) = 3/6$ |

**Example 3.4**
A fair six-sided die is rolled once. If we are only interested in whether the outcome is an odd or even number, we can take

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

## 3.2 Properties of probability measures

**Theorem 3.5 (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:**

(1) Since $A \cup A^c = \Omega$ is a disjoint union and $\mathbb{P}(\Omega) = 1$, it follows by additivity that

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

(2) Since $\emptyset = \Omega^c$ and $\mathbb{P}(\Omega) = 1$, it follows by complemenarity that

$$\mathbb{P}(\emptyset) = \mathbb{P}(\Omega^c) = 1 - \mathbb{P}(\Omega) = 1 - 1 = 0.$$

(3) Let $A \subseteq B$ and let us write $B = A \cup (B \setminus A)$.
Since $A$ and $B \setminus A$ are disjoint sets, it follows by additivity that

$$\mathbb{P}(B) = \mathbb{P}\big[A \cup (B \setminus A)\big] = \mathbb{P}(A) + \mathbb{P}(B \setminus A).$$

Hence, because $\mathbb{P}(B \setminus A) \geq 0$, it follows that $\mathbb{P}(B) \geq \mathbb{P}(A)$.

(4) Let us write:

- $A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$
- $A = (A \setminus B) + (A \cap B)$
- $B = (B \setminus A) + (A \cap B)$

These are disjoint unions, so by additivity,

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

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

## 3.3 Exercises

**Exercise 3.1**

1. What does it mean to say that $\mathbb{P}$ is a probability measure over $(\Omega, \mathcal{F})$?
2. Show that $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ for any two events $A$ and $B$.
3. Let $A$ and $B$ be events such that $\mathbb{P}(A) = 0.4$, $\mathbb{P}(B) = 0.5$ and $\mathbb{P}(A \cup B) = 0.8$. Compute the following probabilities:
   (a) $\mathbb{P}(A \cap B)$.
   (b) $\mathbb{P}(A \cup B^c)$.
4. Let $A$ and $B$ be random events, with probabilities $\mathbb{P}(A) = 1/2$ and $\mathbb{P}(B) = 3/4$.
   (a) Show that $\frac{1}{4} \leq \mathbb{P}(A \cap B) \leq \frac{1}{2}$.
   (b) Show that $\frac{3}{4} \leq \mathbb{P}(A \cup B) \leq 1$.

# Lecture 4 Conditional Probability

## 4.1 Conditional probability

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

- If $B$ occurs and $A \cap B = \emptyset$, then $A$ cannot occur.
- If $B$ occurs and $B \subseteq A$, then $A$ is certain to occur.
- If $B$ occurs, then $A$ will also occur *if and only if* the event $A \cap B$ occurs.

Given that $B$ occurs, the probability that $A$ also occurs is $\mathbb{P}(A \cap B)$ expressed as a proportion of $\mathbb{P}(B)$.

**Definition 4.1**
If $\mathbb{P}(B) > 0$, the *conditional probability of $A$ given $B$* is defined to be

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

**Remark 4.2**
- $\mathbb{P}(A|B) = 0$ whenever $A \cap B = \emptyset$, and
- $\mathbb{P}(A|B) = 1$ whenever $B \subseteq A$.

**Example 4.3**
Let $A$ and $B$ be two events, with probabilities $\mathbb{P}(A) = 0.3$, $\mathbb{P}(B) = 0.8$ and $\mathbb{P}(A \cap B) = 0.2$.
Find the probabilities $\mathbb{P}(A \cup B)$, $\mathbb{P}(A \cap B^c)$, $\mathbb{P}(A|B)$ and $\mathbb{P}(A|B^c)$.

**Solution:**

(1) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = 0.3 + 0.8 - 0.2 = 0.9$

(2) $\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = 0.3 - 0.2 = 0.1$

(3) $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B) = 0.2/0.8 = 0.25$

(4) $\mathbb{P}(A|B^c) = \mathbb{P}(A \cap B^c))/\mathbb{P}(B^c) = 0.1/0.2 = 0.5$

**Example 4.4 (The Second Child Paradox)**
If we know that a man has two children, and that one of them is a boy, what is the probability that he has two boys?

**Solution:** Let $\Omega = \{BB, BG, GB, GG\}$ denote the sample space, and let $A = \{BB, BG, GB\}$ be the event that the man has at least one boy. Then

$$\mathbb{P}(\{BB\}|A) = \frac{\mathbb{P}(\{BB\} \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(\{BB\})}{\mathbb{P}(\{BB, BG, GB\})} = \frac{1/4}{3/4} = \frac{1}{3}.$$

## 4.2 The partition theorem

**Definition 4.5**
A *partition* of a set $B$ is a collection of non-empty sets $\{A_1, A_2, \ldots\}$ such that every element of $B$ lies in exactly one of these sets, or equivalently,

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

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

**Theorem 4.6 (The Partition Theorem)**
If $\{A_1, A_2, \ldots\}$ is a partition of $B$, then

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

**Proof:** First we write $B$ as a disjoint union

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

By the countable additivity of probability measures,

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty}(B \cap A_i)\right) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(B \cap A_i) \\
&= \sum_{i=1}^{\infty} \mathbb{P}(B|A_i)\mathbb{P}(A_i).
\end{aligned}$$

## 4.3 Bayes' theorem

**Lemma 4.7**
For any two events $A$ and $B$ such that $\mathbb{P}(B) > 0$,

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

**Proof:** Set intersection is a commutative operation, so

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

**Theorem 4.8 (Bayes' Theorem)**
Let $\{A_1, A_2, \dots\}$ be a partition of an event $B$ and suppose that $\mathbb{P}(B) > 0$. Then

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

**Proof:** By Lemma 4.7,

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

where the last equality follows by the partition theorem.

**Example 4.9**
Bob tries to buy a newspaper every day. He tries in the morning with probability 1/3, in the evening with probability 1/2 and forgets completely with probability 1/6. The probability of successfully buying a newspaper in the morning is 9/10 (plenty of copies left), and in the evening is 2/10 (often sold out). If Bob buys a newspaper, what is the probability that he bought it in the morning?

**Solution:** Let $M$ be the event that Bob tries to buy a newspaper in the morning, $E$ the event that he tries in the evening, and $F$ the event that he forgets completely. Then

$$\mathbb{P}(M) = 1/3, \qquad \mathbb{P}(E) = 1/2, \qquad \mathbb{P}(F) = 1/6.$$

Let $N$ denote the event that Bob buys a newspaper. Then

$$\mathbb{P}(N|M) = 9/10, \qquad \mathbb{P}(N|E) = 2/10, \qquad \mathbb{P}(N|F) = 0.$$

By Bayes' Theorem,

$$\begin{aligned}
\mathbb{P}(M|N) = \frac{\mathbb{P}(N|M)\mathbb{P}(M)}{\mathbb{P}(N)} &= \frac{\mathbb{P}(N|M)\mathbb{P}(M)}{\mathbb{P}(N|M)\mathbb{P}(M) + \mathbb{P}(N|E)\mathbb{P}(E) + \mathbb{P}(N|F)\mathbb{P}(F)} \\
&= \frac{9/10 \times 1/3}{(9/10 \times 1/3) + (2/10 \times 1/2) + (0 \times 1/6)} \\
&= 3/4
\end{aligned}$$

If Bob buys a newspaper, the probability that he bought it in the morning is 0.75.

## 4.4 Exercises

**Exercise 4.1**

1. Let $A$ and $B$ be events such that $\mathbb{P}(A) = 0.4$, $\mathbb{P}(B) = 0.5$ and $\mathbb{P}(A \cup B) = 0.8$.

   Compute the following probabilities:

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

   (b) $\mathbb{P}(A \cup B^c)$.

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

   (d) $\mathbb{P}(A \mid A \cup B)$.

2. Let $A$, $B$ and $C$ be events such that $\mathbb{P}(A) = 0.7$, $\mathbb{P}(B) = 0.6$, $\mathbb{P}(C) = 0.5$, $\mathbb{P}(A \cap B) = 0.4$, $\mathbb{P}(A \cap C) = 0.3$, $\mathbb{P}(B \cap C) = 0.2$ and $\mathbb{P}(A \cap B \cap C) = 0.1$.

   Compute the following probabilities:

   (a) $\mathbb{P}(A \cup B)$.

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

   (c) $\mathbb{P}(A \mid A \cup B)$.

   (d) $\mathbb{P}(A \cup B \cup C)$.

   (e) $\mathbb{P}(A^c \cap B^c \cap C)$.

   (f) $\mathbb{P}(A^c \cap B^c \cap C|A \cup B)$.

3. A student has three opportunities to pass an exam. The probability of failing the first attempt is 0.6; the probability of failing the second attempt, given that they have failed the first is 0.75, and the probability of failing the third attempt, given that they have failed the first and second is 0.4.

   (a) What is the probability that the student eventually passes the exam.

   (b) What are the respective probabilities of passing at the first, second and third attempts.

4. An insurance company divides its customers into three categories: 60% of customers are classed as low-risk, 30% as moderate-risk and 10% as high-risk. The probabilities that low-risk customers, moderate-risk customers and high-risk customers make a claim in any given year are 0.01, 0.1 and 0.5 respectively. Given that a customer makes a claim this year, what is the probability that the customer is in the high-risk category?

5. A horse has three opportunities to clear a fence. The probability that it fails at the first attempt is 0.4. The probability that it fails at the second attempt, given that it has failed at the first attempt, is 0.3. The probability that it fails at the third attempt, given that it has failed at the first and second attempts, is 0.8.

   (a) What is the probability that the horse eventually clears the fence?

   (b) What are the respective probabilities that the horse clears the fence at the first, second and third attempts.

# Lecture 5 Independence

## 5.1 Independence

If the probability that event $A$ occurs is *not* affected by whether or not event $B$ occurs, then $\mathbb{P}(A|B) = \mathbb{P}(A)$. In such cases, we say that events $A$ and $B$ are *independent*:

**Definition 5.1**
Two events $A$ and $B$ are said to be *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

**Lemma 5.2**
If $A$ and $B$ are independent, then $A$ and $B^c$ are also independent.

**Proof:**

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

**Example 5.3**
A fair die is rolled once. Let $A$ be the event that the outcome is an even number, and let $B$ be the event that the outcome is divisible by 3. Are $A$ and $B$ independent?

**Solution:**

- $\mathbb{P}(A) = \mathbb{P}(\{2, 4, 6\}) = 1/2$, $\mathbb{P}(B) = \mathbb{P}(\{3, 6\}) = 1/3$, $\mathbb{P}(A \cap B) = \mathbb{P}(\{6\}) = 1/6$.
- Thus we see that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, so $A$ and $B$ are independent.

## 5.2 Pairwise independence and total indepencence

**Definition 5.4**
A family 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 $\{B_1, B_2, \ldots, B_m\} \subseteq \{A_1, A_2, \ldots\}$, we have

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

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

**Example 5.5 (de Méré's Paradox)**
Show that you are more likely to obtain a six in 4 rolls of a single fair die, than to obtain a double-six in 24 rolls of two fair dice.

**Solution:** Assume that the rolls are totally independent of each other.

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

**Example 5.6**
Consider a sample space $\Omega = \{1,2,3,4\}$ where each outcome is equally likely. Let $A = \{1,2\}$, $B = \{1,3\}$ and $C = \{1,4\}$. Show that $\{A,B,C\}$ is pairwise independent, but not totally independent.

**Solution:**

- $\mathbb{P}(A) = 1/2$ and $\mathbb{P}(B) = 1/2$
- $\mathbb{P}(A \cap B) = \mathbb{P}(\{1\}) = 1/4$.
- Hence $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, so $A$ and $B$ are independent.
- Similarly, $\mathbb{P}(A \cap C) = \mathbb{P}(A)\mathbb{P}(C)$ and $\mathbb{P}(B \cap C) = \mathbb{P}(B)\mathbb{P}(C)$.

Thus the set $\{A,B,C\}$ is pairwise independent.

- $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\{1\}) = 1/4$
- However, $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = 1/8$.
- Hence $\mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

Thus the set $\{A,B,C\}$ is not totally independent.

## 5.3 Exercises

**Exercise 5.1**

1. A fair six-sided die is rolled repeatedly. How many times should it be rolled to ensure that the probability of getting a six is at least 0.8?
2. A multiple choice test has five questions, with each question having four alternative choices. At least three questions must be answered correctly to pass the test. If a candidate chooses her answers at random, what is the probability that she passes the test? State any assumptions you make.
3. Two fair dice are rolled. Show that the event that their sum is 7 is independent of the score shown on the first die.
4. A fair die is rolled twice, each roll being independent of the other. Let $A$ be the event that the first roll shows 3, let $B$ be the event that the second roll shows 4, and let $C$ be the event that the total of the two rolls is 7.
   (a) Define a suitable sample space, and identify the subsets corresponding to events $A$, $B$ and $C$.
   (b) Show that $\{A,B,C\}$ is pairwise independent but not totally independent.
5. A coin has probability $p$ of showing heads. Let $q_n$ be the probability that in $n$ independent tosses, a head is observed an even number of times (for this question, take 0 to be an even number). Using the partition theorem, show that
$$q_n = p(1 - q_{n-1}) + (1-p)q_{n-1} \quad \text{for any} \quad n \geq 1.$$

# Lecture 6 Classical Probability

## 6.1 The principle of indifference

Let $\Omega$ be a finite sample space. If the outcomes are indistinguishable (except for their names), the *principle of indifference* states that each outcome should be assigned equal probability.

- We say that the outcomes are *equally likely* to occur.
- This is often an implicit assumption (e.g. "a card is chosen at random").

The probability of an event $A \subseteq \Omega$ is proportional to its *cardinality*:

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

- Classical examples involve *coins*, *dice* and *cards*.
- Problems can be solved by counting the number of ways that different events can occur.

## 6.2 Sampling with replacement

In many cases, once an element has been chosen it is *replaced* in the set, and can be chosen again in subsequent selections. If there are $n$ possible elements, there are $n^k$ distinct choices of $k$ elements under this selection model.

**Example 6.1**
There are exactly $2^k$ binary sequences of length $k$, $3^k$ ternary sequences of length $k$, etc.

**Example 6.2**
UK registration plates are of the form `AB 12 CDE`: the first two positions can be any letters except `I`, `Q` and `Z`, the third and fourth positions can be any single digit, and the last three positions can be any letters except `I` and `Q`. How many different UK vehicle registration plates are possible?

**Solution:** The total number of possible registration plates is

$$23 \times 23 \times 10 \times 10 \times 24 \times 24 \times 24 = 731\,289\,600.$$

## 6.3 Sampling without replacement

It also happens that once an element has been chosen, it cannot be chosen again in subsequent selections. There are two cases here, depending on whether or not we take the *order* in which elements are chosen into account.

**Definition 6.3**

(1) A *k-permutation* of a set of $n$ elements is an ordered sequence of $k$ elements, taken (without replacement) from the set. The number of $k$-permutations is

$$^nP_k = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

(2) A *k-combination* of a set of $n$ elements is an un-ordered subset of $k$ elements, taken (without replacement) from the set. The number of $k$-combinations is

$$^nC_k = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k(k-1)(k-2)\cdots 1} = \frac{n!}{(n-k)!k!}$$

This is called the *binomial coefficient*, and is usually denoted by $\binom{n}{k}$.

**Example 6.4**
An urn contains 3 white balls and 5 black balls. Two balls are drawn at random from the urn. What is the probability that they are both white?

**Solution:**

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

Hence, $\mathbb{P}(A) = |A|/|\Omega| = 3/28$.

**Example 6.5**
Find the probability that a hand of five cards contains (a) four cards of the same kind, and (b) two distinct pairs.

**Solution:** Let $\Omega$ be the set of all possible hands: $|\Omega| = \binom{52}{5} = 2598960$.

(1) Let $A$ be the event that the hand contains four of a kind.

- There are 13 different kinds of card (A,2,3,4,5,6,7,8,9,10,J,Q,K).
- For each kind, there are 48 different hands that contain all four cards of that kind.
- Hence there are $|A| = 13 \times 48 = 624$ hands containing four cards of the same kind.
- Thus $\mathbb{P}(\text{four-of-a-kind}) = |A|/|\Omega| = 624/2598960 = 0.00024$.

(2) Let $B$ be the event that the hand contains two pairs.

- Two (distinct) pairs can be chosen in $\binom{13}{2} = 78$ different ways.
- Each pair can be chosen in $\binom{4}{2} = 6$ different ways.
- The fifth card can be chosen in 44 different ways.
- Hence there are $|B| = 78 \times 6 \times 6 \times 44 = 123552$ hands containing two pairs.
- Thus $\mathbb{P}(\text{two-pairs}) = |B|/|\Omega| = 123552/2598960 = 0.0475$.

**Example 6.6 (The Division Paradox)**
A *fair game* is a game in which the probability of winning is equal to the probability of losing. Two players $A$ and $B$ decide to play a sequence of fair games until one of the players wins 6 games, but they stop when the score is 5:3 in favour of player $A$. How should the prize money be fairly divided?

**Solution:** Assume that the players carried on playing the sequence of games,

- The maximum number of additional games is 3.
- The possible outcomes are

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

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

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

**Example 6.7 (The Birthday Paradox)**
Find the smallest number of people for which the probability that least two share a birthday exceeds 1/2.

**Solution:**

- Suppose we have a set of $k$ randomly chosen people.
- Let $n = 365$ be the number of days in a year.
- Assume that each birthday is equally likely to occur.

Let $p(n, k)$ be the probability that at least two people share a birthday.

- We want to find the value of $k$ for which $p(n, k) > 0.5$.
- There are $n$ choices of birthday for each of the $k$ people, so the total number of outcomes is $n^k$.
- The number of outcomes in which all birthdays are different is the number of $k$-permutations of $n$ elements, which is equal to

$$^nP_k = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

Hence,

$$\begin{aligned} p(n,k) &= \mathbb{P}(\text{At least one pair share a birthday}) \\ &= 1 - \mathbb{P}(\text{No pairs share a birthday}) \\ &= 1 - \frac{\text{Number of outcomes with all birthdays different}}{\text{Total number of outcomes}} \\ &= 1 - \frac{n!}{n^k(n-k)!} \end{aligned}$$

The probabilities $p(365, k)$ for various values of $k$ are shown below:

| $k$ | 2 | 10 | 15 | 22 | 23 | 50 | 100 | 366 |
|---|---|---|---|---|---|---|---|---|
| $p(k, 365)$ | 0.003 | 0.12 | 0.25 | 0.48 | 0.51 | 0.97 | 0.99997 | 1 |

## 6.4 Exercises

**Exercise 6.1**

1. A student has two classical CDs, four jazz CDs, three rock CDs and three pop CDs, and wants to arrange them so that all CDs of the same genre are located next to each other. In how many distinct ways can this be done? If the CDs are arranged at random, what is the probability that this occurs?
2. (Tuesday's Child Paradox) A man tells you that he has two children, and that one is a boy born on a Tuesday. What is the probability that he has two boys?
3. (Division Paradox) Two players $A$ and $B$ play a series of games. The players agree beforehand that the winner of the series will be the person who first wins 5 games. For some reason, the players have to stop at the point when player $A$ has won 3 games and player $B$ has won 2 games. Based on past experience, the probability that player $A$ wins a game is $p$, and the probability that $B$ wins a game is $q$ (where $p + q = 1$). In view of this, how should should the players divide the prize money? State any assumptions you make.
4. (The Monty Hall Problem) You are a contestant in a game show. A prize is concealed behind one of three doors; the other two doors each conceals a goat. You win the prize if you choose the door behind which the prize is hidden. After you have chosen a door, but before the door is opened, the host opens one of the other two doors to reveal a goat, and then asks if you would like to switch from your current selection to the remaining unopened door. Should you switch, or should you stick with your original selection?

# Lecture 7 The Frequentist Model

## 7.1 Finite probability spaces

For the next few lectures, we will consider random experiments which have only finitely many outcomes.

- Let $\Omega = \{\omega_1, \omega_2, \ldots . \omega_n\}$ denote the sample space.

The power set $\mathcal{P}(\Omega)$ contains all events of interest. To define a probability measure on $\mathcal{P}(\Omega)$, we first define a *probability mass function* on the individual outcomes:

**Definition 7.1**
A *probability mass function* on $\Omega$ is a function

$$\begin{array}{rccc} p: & \Omega & \to & [0,1] \\ & \omega & \mapsto & p(\omega) \end{array}$$

with the property that $\displaystyle\sum_{\omega \in \Omega} p(\omega) = 1$.

**Remark 7.2**
In classical probability, we choose a sample space in which all outcomes are equally likely. The probability mass function is then taken to be $p(\omega) = 1/n$, where $n$ is the cardinality of the sample space.

The probability mass function on $\Omega$ can be extended to a probability measure on $\mathcal{P}(\Omega)$ by defining the probability of an event to be the sum of the probabilities of the outcomes it contains:

**Theorem 7.3**
Let $\Omega$ be a finite sample space, and let $p$ be a probability mass function on $\Omega$. Then the function

$$\begin{array}{rccc} \mathbb{P}: & \mathcal{P}(\Omega) & \to & [0,1] \\ & A & \mapsto & \displaystyle\sum_{\omega \in A} p(\omega) \end{array}$$

is a probability measure on subsets of $\Omega$.

**Proof:** First, because $p$ is a probability mass function, we see that

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

To prove additivity, let $A_1, A_2, \ldots, A_n$ be pairwise disjoint subsets of $\Omega$. (Note that because $\Omega$ is finite, it has only finitely many subsets.) Then

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{\omega \in A_1 \cup A_2 \cup \ldots \cup A_n} p(\omega) \\
&= \sum_{\omega \in A_1} p(\omega) + \sum_{\omega \in A_2} p(\omega) + \ldots + \sum_{\omega \in A_n} p(\omega) \\
&= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_n) \\
&= \sum_{i=1}^{n} \mathbb{P}(A_i).
\end{aligned}$$

Thus $\mathbb{P}$ satisfies the conditions required for it to be a probability measure.

## 7.2 Relative frequency

We have seen that probability is a measure of how likely an event is to occur.

- The *symmetry* that exists in many random experiments (e.g. those involving dice and cards) allows us to choose sensible values for the probability of various events.
- How can we define probability in more general situations?

If a random experiment can be repeated many times under the same conditions, it is natural to think of probability as the number of times an event occurs expressed as a proportion of the total number times the experiment is repeated.

- Let $N$ be the number of times the experiment is repeated.
- Let $N(A)$ be the number of times event $A$ occurs during these $N$ repetitions.

**Definition 7.4**
The ratio $N(A)/N$ is called the *relative frequency* of event $A$.

**Definition 7.5**
Under the *frequentist model*, the probability of an event $A$ is defined to be the limit of its relative frequency as the number of trials increases to infinity:

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

The frequentist model dominates in many areas of science (e.g. medical trials).

- In practical applications, the experiment is repeated many times, then the relative frequency $N(A)/N$ of an event $A$ is taken as an approximation of its "true" probability.

Frequentist probability has nice properties:

- $0 \leq \mathbb{P}(A) \leq 1$ for every event $A \in \mathcal{P}(\Omega)$.
- $\mathbb{P}(\emptyset) = 0$, because $N(\emptyset) = 0$ for any number of repetitions $N$.
- $\mathbb{P}(\Omega) = 1$, because $N(\Omega) = N$ for any number of repetitions $N$.

**Theorem 7.6**
Let $\Omega$ be a finite sample space. Then the function

$$\begin{aligned} \mathbb{P}: \ \mathcal{P}(\Omega) &\to [0,1] \\ A &\mapsto \lim_{N\to\infty} \frac{N(A)}{N} \end{aligned}$$

is a probability measure on subsets of $\Omega$.

**Proof:** Clearly, $\mathbb{P}(\Omega) = 1$, because $N(\Omega) = N$ for any number of repetitions $N$.
To prove additivity, let $A_1, A_2, \ldots, A_n$ be pairwise disjoint subsets of $\Omega$. (Note that because $\Omega$ is finite, it has only finitely many subsets.) Then

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \lim_{N\to\infty} \frac{N(A_1 \cup A_2 \cup \ldots \cup A_n)}{N} \\ &= \lim_{N\to\infty} \frac{N(A_1) + N(A_2) + \ldots + N(A_n)}{N} \qquad \text{(because the } A_i \text{ are disjoint)} \\ &= \lim_{N\to\infty} \frac{N(A_1)}{N} + \lim_{N\to\infty} \frac{N(A_2)}{N} + \ldots + \lim_{N\to\infty} \frac{N(A_n)}{N} \\ &= \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_n) \\ &= \sum_{i=1}^{n} \mathbb{P}(A_i). \end{aligned}$$

Thus $\mathbb{P}$ is a probability measure.

Because $\mathbb{P}$ is a probability measure, it has the following properties:

**Corollary 7.7**
(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: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

**Remark 7.8**
These properties can be proved directly. For example:

$$\mathbb{P}(A^c) = \lim_{N\to\infty} \frac{N(A^c)}{N} = \lim_{N\to\infty} \frac{N - N(A)}{N} = 1 - \lim_{N\to\infty} \frac{N(A)}{N} = 1 - \mathbb{P}(A).$$

## 7.3 Conditional probability

Under the frequentist model, a natural definition of $\mathbb{P}(A|B)$ is the number of trials in which $A$ and $B$ both occur, expressed as a proportion of the number of trials in which $B$ occurs.

- Let $N$ be the number of times the experiment is repeated.
- Let $N(A)$ be the number of times that event $A$ occurs.
- Let $N(B)$ be the number of times that event $B$ occurs.
- Let $N(A, B)$ be the number of times that events $A$ and $B$ both occur.

Then

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

which agrees with the definition of conditional probability given previously.

## 7.4 Infinite sample spaces

The frequentist model does not extend to random experiments that have infinitely many outcomes.

**Example 7.9**
Consider a random experiment in which a coin is tossed repeatedly until the first head occurs. The sample space for this experiment is an infinite set. Two possible representations are:

- $\Omega = \{H, TH, TTH, TTTH, TTTTH, TTTTTH, \ldots\}$, or alternatively,
- $\Omega = \{1, 2, 3, 4, 5, \ldots\}$.

Let $N$ be the number of times that the random experiment is repeated.

- For finite sample spaces, we can choose $N$ sufficiently large to ensure that every possible outcome can occur at least once.
- For infinite sample spaces, there are plausible outcomes that will not occur, regardless of how large we choose $N$.

Thus we cannot define probability on infinite sample spaces in terms of relative frequency.

# Lecture 8 Random Variables

## 8.1 Random variables

We are not always directly interested in the outcome of a random experiment, but rather in some *consequence* of the outcome. For example, a gambler is often more interested in his losses than in the outcomes of the individual games which led to them.

**Example 8.1**
A fair coin is tossed twice. Let $\Omega = \{HH, HT, TH, TT\}$ denote the sample space.

(1) The number of heads observed is a function $X : \Omega \to \mathbb{R}$, defined by

$$X(HH) = 2, \quad X(HT) = X(TH) = 1, \quad X(TT) = 0.$$

(2) Suppose we bet £1 on the outcome of this random experiment, and stand to win £4 if two heads occur, but otherwise lose our stake. Our winnings can be represented by a function $Y : \Omega \to \mathbb{R}$, defined by

$$Y(HH) = 4, \quad Y(HT) = Y(TH) = Y(TT) = -1.$$

- A random variable $X$ associates a real number $X(\omega)$ with each outcome $\omega \in \Omega$.
- Random variables are often used to pick out particular features of an experiment that are of interest.

**Definition 8.2**
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *random variable* on $(\Omega, \mathcal{F})$ is a function,

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

with the property that $\{\omega : X(\omega) \le x\} \in \mathcal{F}$ for every $x \in \mathbb{R}$.

**Discussion**

- The set $\{\omega : X(\omega) \le x\}$ contains those outcomes that are mapped by $X$ into the interval $(-\infty, x]$.
- Why do we insist on the condition that these sets are included in the field of events $\mathcal{F}$?
- Recall the definition of a probability measure:

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

- The condition ensures that the events $\{\omega : X(\omega) \le x\}$ have a well-defined probability.
- But why do we need this to be the case?

### 8.1.1 Notational conventions

- Random variables are denoted by upper-case letters such as $X$ and $Y$.
- Values taken by random variables are denoted by the corresponding lower-case letters, such as $x$ and $y$.

Let $B$ be any subset of the real numbers. We define the following abbreviations:

(1) $\{X \in B\}$ for the event $\{\omega : X(\omega) \in B\}$.

- This is the event that $X$ takes a value in the set $B$.

(2) $\mathbb{P}(X \in B)$ for the probability $\mathbb{P}\big(\{\omega : X(\omega) \in B\}$.

- This is the probability that $X$ takes a value in the set $B$.

(3) $\{X = x\}$ for the event $\{\omega : X(\omega) = x\}$.

- This is the event that $X$ takes the value $x$.

(4) $\mathbb{P}(X = x)$ for the probability $\mathbb{P}\big(\{\omega : X(\omega) = x\}$.

- This is the probability that $X$ takes the value $x$.

(5) $\{X \leq x\}$ for the event $\{\omega : X(\omega) \leq x\}$

- This is the event that $X$ takes a value at most equal to $x$.

(6) $\mathbb{P}(X \leq x)$ for the probability $\mathbb{P}\big(\{\omega : X(\omega) \leq x\}$.

- This is the probability that $X$ takes a value at most equal to $x$.

### 8.1.2 Indicator variables

**Definition 8.3**
The *indicator variable* $I_A : \Omega \to \mathbb{R}$ of an event $A$ is defined by

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

**Theorem 8.4**
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:** Exercise.
Note that for two functions to be equal, they must have the same domain, and be equal at every point of that domain, so for the first part we need to show that $I_{A^c}(\omega) = 1 - I_A(\omega)$ for every $\omega \in \Omega$, and similarly for parts (2) and (3).

### 8.1.3 Simple random variables

**Definition 8.5**
A *simple random variable* is one that takes only finitely many values.

Let $X : \Omega \to \mathbb{R}$ be a simple random variable, let $\{x_1, x_2, \ldots, x_n\}$ be the range of values that $X$ can take, and consider the finite partition $\{A_1, A_2, \ldots, A_n\}$ of the sample space, defined by

$$A_i = \{\omega : X(\omega) = x_i\}.$$

Then $X$ can be written as

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

where $I_{A_i}$ is the indicator variable of the event $A_i$.

**Example 8.6**
The random variables $X$ and $Y$ of example 8.1 are both simple random variables:

(1) Let $A_1 = \{TT\}$, $A_2 = \{TH, HT\}$, $A_3 = \{HH\}$, and let $x_1 = 0$, $x_2 = 1$, $x_3 = 2$. Then

$$X(\omega) = \sum_{i=1}^{3} x_i I_{A_i}(\omega) = I_{A_2}(\omega) + 2I_{A_3}(\omega).$$

(2) Let $B_1 = \{TT, TH, HT\}$, $B_2 = \{HH\}$, and let $y_1 = -1$, $y_1 = 4$. Then

$$Y(\omega) = \sum_{i=1}^{2} x_i I_{A_i}(\omega) = -I_{B_1}(\omega) + 4I_{B_2}(\omega).$$

## 8.2 Cumulative distribution functions (CDFs)

A random variable is completely described by its CDF:

**Definition 8.7**
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 *cumulative distribution function* (CDF) of $X$ is defined to be the function

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

**Remark 8.8**
- Recall that $\mathbb{P}(X \leq x)$ is shorthand notation for $\mathbb{P}\big(\{\omega : X(\omega) \leq x\}\big)$.
- To ensure that this probability is well-defined, we need that $\{\omega : X(\omega) \leq x\} \in \mathcal{F}$.
- The fact that $X$ is a random variable on $(\Omega, \mathcal{F})$ guarantees that this is the case.

**Example 8.9**
The CDFs of the random variables $X$ and $Y$ of example 8.1 are given by

$$F_X(x) = \begin{cases} 0 & x < 0, \\ 1/4 & 0 \leq x < 1, \\ 3/4 & 1 \leq x < 2, \\ 1 & x \geq 2, \end{cases} \qquad \text{and} \qquad F_Y(Y) = \begin{cases} 0 & y < -1, \\ 3/4 & -1 \leq y < 4, \\ 1 & y \geq 4, \end{cases}$$

respectively.

## 8.3 Probability mass functions (PMFs)

**Definition 8.10**
Let $X : \Omega \to \mathbb{R}$ be a simple random variable. The *probability mass function* (PMF) of $X$ is the function

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

**Example 8.11**
In example 8.1, because the coin is *fair* the PMFs of $X$ and $Y$ are, respectively,

$$f_X(x) = \begin{cases} 1/4 & \text{if } x = 0, \\ 1/2 & \text{if } x = 1, \\ 1/4 & \text{if } x = 2, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \qquad f_Y(y) = \begin{cases} 3/4 & \text{if } y = -1, \\ 1/4 & \text{if } y = 4. \\ 0 & \text{otherwise.} \end{cases}$$

## 8.4 Finite probability spaces

Let $\Omega$ be a finite sample space, let $p : \Omega \to [0,1]$ be a probability mass function on $\Omega$, and let $\mathbb{P}$ be the associated probability measure on the power set of $\Omega$. We often refer to the pair $(\Omega, \mathbb{P})$ as a *finite probability space*, without explicit reference to the fact that we take our field of events to be the power set $\mathcal{P}(\Omega)$.

**Lemma 8.12**
If $\Omega$ is a finite sample space, any function $X : \Omega \to \mathbb{R}$ is a simple random variable on $\Omega$.

**Proof:**

- $X$ is a random variable because for any $x \in \mathbb{R}$, the condition $\{\omega : X(\omega) \leq x\} \in \mathcal{P}(\Omega)$ is automatically satisfied, because the power set contains *all* subsets of $\Omega$.
- $X$ is a simple random variable because its domain is finite, so it can take at most finitely many values.

Let $(\Omega, \mathbb{P})$ be a finite probability space, let $p : \Omega \to [0,1]$ be the associated probability mass function, let $X : \Omega \to \mathbb{R}$ be a random variable on $\Omega$, and let $\{x_1, x_2, \dots, x_n\}$ be the range of values taken by $X$. Then the PMF of $X$ can be written as

$$\mathbb{P}(X = x_i) = \sum_{\omega \in A_i} p(\omega) \quad \text{where} \quad A_i = \{\omega : X(\omega) = x_i\} \quad \text{for } i = 1, 2, \dots, n.$$

Note that the events $A_1, A_2, \dots, A_n$ form a partition of $\Omega$.

**Example 8.13**
A biased coin is independently tossed three times, each time having probability $p$ of showing heads and $q = 1 - p$ of showing tails. Let $X$ be the number of heads that occur. Find the PMF of $X$.

**Solution:** The sample space is $\Omega = \{TTT, HTT, THT, TTH, THH, HTH, HHT, HHH\}$.
By independence, the associated probability mass function $p(\omega)$ is

| $\omega$ | $TTT$ | $HTT$ | $THT$ | $TTH$ | $THH$ | $HTH$ | $HHT$ | $HHH$ |
|---|---|---|---|---|---|---|---|---|
| $p(\omega)$ | $q^3$ | $pq^2$ | $pq^2$ | $pq^2$ | $p^2q$ | $p^2q$ | $p^2q$ | $p^3$ |

The number of heads is a random variable $X : \Omega \to \mathbb{R}$, given by

| $\omega$ | $TTT$ | $HTT$ | $THT$ | $TTH$ | $THH$ | $HTH$ | $HHT$ | $HHH$ |
|---|---|---|---|---|---|---|---|---|
| $X(\omega)$ | 0 | 1 | 1 | 1 | 2 | 2 | 2 | 3 |

$X$ takes values in the range $\{0, 1, 2, 3\}$, with probabilities

- $\mathbb{P}(X = 0) = \mathbb{P}(\{TTT\}) = q^3$,
- $\mathbb{P}(X = 1) = \mathbb{P}(\{HTT, THT, TTH\}) = 3pq^2$,
- $\mathbb{P}(X = 2) = \mathbb{P}(\{THH, HTH, HHT\}) = 3p^2q$,
- $\mathbb{P}(X = 3) = \mathbb{P}(\{HHH\}) = p^3$.

The PMF of $X$ can be represented in tabular form:

| $x$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| $f(x)$ | $q^3$ | $3pq^2$ | $3p^2q$ | $p^3$ |

## 8.5 Exercises

**Exercise 8.1**

1. Let $I_A$ denote the indicator function of an event $A$. Show that
   (a) $I_{A^c} = 1 - I_A$.
   (b) $I_{A\cap B} = I_A I_B$.
   (c) $I_{A\cup B} = I_A + I_B - I_{A\cap B}$.
2. Let $\Omega$ be the sample space of some random experiment, $\mathcal{F}$ be a field of events over $\Omega$, and let $A \in \mathcal{F}$. Show that the indicator variable $I_A : \Omega \to \mathbb{R}$ of event $A$, defined by
$$I(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{if } \omega \notin A, \end{cases}$$
is indeed a random variable on $(\Omega, \mathcal{F})$.
3. Let $X_1$ and $X_2$ be the numbers obtained in two independent throws of a fair die. Find the PMF of each of the following random variables:
   (a) $X_1$,
   (b) $Y = 7 - X_1$,
   (c) $U = \max(X_1, X_2)$,
   (d) $V = X_1 - X_2$.
   (e) $W = |X_1 - X_2|$.
4. An urn contains 7 red balls and 3 blue balls. If 5 balls are selected at random without replacement, find the PMF of the number of red balls selected.
5. Let $X$ be a random variable with the following PMF:

| $x$ | $-2$ | $0$ | $1$ | $4$ |
|---|---|---|---|---|
| $f(x)$ | 0.4 | 0.1 | 0.3 | 0.2 |

   Sketch the PMF and CDF of $X$.
6. Let $X$ be a random variable with PMF:
$$f(x) = \begin{cases} c & \text{if } x = 0, \\ 3c & \text{if } x = 1, \\ 6c & \text{if } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$
where $c$ is an unknown constant.
   (a) Find the value of $c$.
   (b) Find the probabilities $\mathbb{P}(X < 2)$, $\mathbb{P}(X \leq 2)$ and $\mathbb{P}(X > 1)$.
   (c) What is the smallest value of $k$ for which $\mathbb{P}(X \leq k) > 0.25$?
7. A pair of fair dice is rolled six times. What is the probability of getting a total of seven
   (1) twice,
   (2) at least once,
   (3) more than three times?

# Lecture 9 Expectation

### Classical mechanics

Consider a system of $n$ point particles with masses $m_1, m_2, \ldots, m_n$ placed at locations $x_1, x_2, \ldots, x_n$ along a thin rod. The point at which the rod balances is called the *centre of mass* of the system.

The centre of mass is the point $\mu$ for which

$$\sum_{i=1}^{n}(x_i - \mu)m_i = 0.$$

Solving this equation for $\mu$, we obtain

$$\mu = \frac{1}{M}\sum_{i=1}^{n} x_i m_i \qquad \text{where} \qquad M = \sum_{i=1}^{n} m_i \text{ is the total mass.}$$

- $\mu$ is a single number that describes the *location* of the system.
- $\mu$ says nothing about the *size* or *shape* of the system.

### Probability theory

Let $(\Omega, \mathbb{P})$ be a finite probability space, let $X : \Omega \to \mathbb{R}$ be a random variable on $\Omega$, and let $\{x_1, x_2, \ldots, x_n\}$ be the range of $X$. The PMF of $X$ is the function

$$f(x_i) = \mathbb{P}(X = x_i).$$

The "centre of mass" of the random variable $X$ is the value $\mu$ for which

$$\sum_{i=1}^{n}(x_i - \mu)f(x_i) = 0.$$

Solving for $\mu$ and using the fact that $\sum_{i=1}^{n} f(x_i) = 1$, we obtain

$$\mu = \sum_{i=1}^{n} x_i f(x_i).$$

- $\mu$ is called *expected value* or *expectation* of $X$.
- $\mu$ describes the *centre* or *location* of the system.

## 9.1 Expectation

Let $(\Omega, \mathbb{P})$ be a finite probability space, let $p : \Omega \to [0, 1]$ be the associated probability mass function, and let $X : \Omega \to \mathbb{R}$ be a random variable on $\Omega$.

**Definition 9.1**
The *expectation* of $X$ is defined to be

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

**Theorem 9.2 (Expectation of Indicator Variables)**
Let $I_A : \Omega \to \mathbb{R}$ be the indicator variable of event $A$. Then $\mathbb{E}(I_A) = \mathbb{P}(A)$.

**Proof:** $I_A(\omega) = 1$ if $\omega \in A$, and $I_A(\omega) = 0$ if $\omega \notin A$, so

$$\mathbb{E}(I_A) = \sum_{\omega\in\Omega} I(\omega)p(\omega) = \sum_{\omega\in A} 1 \times p(\omega) + \sum_{\omega\notin A} 0 \times p(\omega) = \sum_{\omega\in A} p(\omega) = \mathbb{P}(A)$$

- In definition 9.1, expectation is defined as a sum over the *domain* of $X$.
- Expectation can also be defined as a sum over the *range* of $X$:

**Theorem 9.3**
Let $\{x_1, x_2, \ldots, x_n\}$ be the range of $X$, and let $f(x)$ denote its PMF. The expectation of $X$ can be written as

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

**Proof:**

- Let $A_i = \{\omega : X(\omega) = x_i\}$ for $i = 1, 2, \ldots, n$.
- Each outcome $\omega \in \Omega$ belongs to exactly one of the sets $A_1, A_2, \ldots, A_n$.
- The sets $A_1, \ldots, A_n$ thus form a partition of the sample space $\Omega$.

Consequently,

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{\omega\in\Omega} X(\omega)p(\omega) \\
&= \sum_{\omega\in A_1} X(\omega)p(\omega) + \sum_{\omega\in A_2} X(\omega)p(\omega) + \ldots + \sum_{\omega\in A_n} X(\omega)p(\omega) \\
&= x_1 \sum_{\omega\in A_1} p(\omega) + x_2 \sum_{\omega\in A_2} p(\omega) + \ldots + x_n \sum_{\omega\in A_n} p(\omega) \\
&= x_1\mathbb{P}(A_1) + x_2\mathbb{P}(A_2) + \ldots + x_n\mathbb{P}(A_n) \\
&= x_1\mathbb{P}(X = x_1) + x_2\mathbb{P}(X = x_2) + \ldots + x_n\mathbb{P}(X = x_n) \\
&= \sum_{i=1}^{n} x_i\mathbb{P}(X = x_i) = \sum_{i=1}^{n} x_i f(x_i).
\end{aligned}$$

**Example 9.4**
Three fair dice are rolled independently. In return for a £1 stake, a player wins £1 if two of the dice show the same number and £5 if all three show the same number; otherwise the player loses the stake. Find the expected amount won by the player.

**Solution:** Let $X$ denote the amount won. The range of $X$ is the set $\{-1, 1, 5\}$.

- There are $6^3 = 216$ possible outcomes, each of equal probability $1/216$.
- $\{X = 5\}$ is the event that all dice show the same number.
  This can occur in six different ways so $\mathbb{P}(X = 5) = 6/216 = 1/36$.
- $\{X = 1\}$ is the event that exactly two of the dice show the same number.
  This can occur in 90 different ways so $\mathbb{P}(X = 1) = 90/216 = 15/36$.

- $\{X = -1\}$ is the event that the dice all show different numbers.
  This can occur in 120 different ways so $\mathbb{P}(X = 1) = 120/216 = 20/36$.

The expected amount won is therefore

$$\mathbb{E}(X) = \sum_{x\in\{-1,1,5\}} x\mathbb{P}(X = x) = \left(-1 \times \frac{20}{36}\right) + \left(1 \times \frac{15}{36}\right) + \left(5 \times \frac{1}{36}\right) = 0.$$

Games in which the expected amount won (or lost) is zero are called *fair* games.

## 9.2 Properties of expectation

**Theorem 9.5 (Properties of expectation)**
Let $\Omega$ be a finite sample space, and let $X, Y : \Omega \to \mathbb{R}$ be random variables on $\Omega$.

(1) **Positivity**. If $X(\omega) \geq 0$ for all $\omega \in \Omega$, then $\mathbb{E}(X) \geq 0$.

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

(3) **Monotonicity**. If $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

**Proof:**

(1) **Positivity**
If $X(\omega) \geq 0$ for all $\omega$, then $\mathbb{E}(X) = \sum_{\omega\in\Omega} X(\omega)p(\omega)$ is a sum of positive terms, so $\mathbb{E}(X) \geq 0$.

(2) **Linearity**
Let $a, b \in \mathbb{R}$. Because $\Omega$ is finite, the composite function $aX + bY$ is also a random variable on $\Omega$. We need to show that

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

- Let $X$ take values in the range $\{x_1, x_2, \ldots, x_m\}$, and let $A_i = \{\omega : X(\omega) = x_i\}$.
- Let $Y$ take values in the range $\{y_1, y_2, \ldots, y_n\}$, and let $B_j = \{\omega : Y(\omega) = y_j\}$.

The sets $\{A_i \cap B_j : i = 1, 2, \ldots, m; j = 1, 2, \ldots, n\}$ form a partition of $\Omega$.

- The random variable $aX + bY$ takes the value $ax_i + by_j$ for all $\omega \in A_i \cap B_j$.

By Theorem 9.3,

$$\begin{aligned}\mathbb{E}(aX + bY) &= \sum_{i=1}^{m}\sum_{j=1}^{n}(ax_i + by_j)\mathbb{P}(A_i \cap B_j)\\ &= a\sum_{i=1}^{m} x_i \sum_{j=1}^{n}\mathbb{P}(A_i \cap B_j) + b\sum_{j=1}^{n} y_j \sum_{i=1}^{m}\mathbb{P}(A_i \cap B_j)\end{aligned}$$

Furthermore,

- $\{A_i \cap B_j\}_{j=1}^{n}$ is a partition of $A_i$, so $\sum_{j=1}^{n}\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)$.
- $\{A_i \cap B_j\}_{i=1}^{m}$ is a partition of $B_j$, so $\sum_{i=1}^{m}\mathbb{P}(A_i \cap B_j) = \mathbb{P}(B_j)$.

Hence

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

as required.

(3) **Monotonicity**

$X \leq Y$ if and only if $Y - X \geq 0$.

- By positivity, $\mathbb{E}(Y - X) \geq 0$.
- By linearity, $\mathbb{E}(Y) - \mathbb{E}(X) \geq 0$, so $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

## 9.3 The sample mean

Consider a random experiment with a finite sample space $\Omega$. Let $X$ be a random variable on $\Omega$, and let $\{x_1, x_2, \ldots, x_n\}$ be the range of $X$.

- Suppose the random experiment is repeated $N$ times.
- Let $z_1, z_2, \ldots, z_N$ be the sequence of observed values.
- Let $N_i$ be the number of trials in which the observed value is $x_i$ (for $i = 1, 2, \ldots, n\}$

In common parlance, the "average value" of the observations is

$$\begin{aligned}
\bar{z} &= \frac{1}{N}\sum_{j=1}^{N} z_j \\
&= \frac{1}{N}\sum_{i=1}^{n} x_i N_i \quad \text{because exactly } N_i \text{ of the observations are equal to } x_i, \\
&= \sum_{i=1}^{n} x_i \left(\frac{N_i}{N}\right).
\end{aligned}$$

Under the frequentist model, the relative frequencies $N_i/N$ tend to the corresponding "true" probabilities $\mathbb{P}(X = x_i)$ as the number of repetitions $N \to \infty$.

- The "average value" $\bar{z}$ is called the *sample mean*.
- The sample mean is an empirical estimate of the expected value, $\mathbb{E}(X)$.
- The estimate becomes more accurate as the number of repetitions increases.

## 9.4 Exercises

### Exercise 9.1

1. Two fair dice are rolled independently. Let $X$ denote the larger of the two scores. Find the PMF of $X$, and its expected value $\mathbb{E}(X)$.
2. A thousand tickets are sold in a lottery, in which there is one prize of £200, four prizes of £50 and ten prizes of £10. If a ticket costs £1, calculate the expected net gain from one ticket.
3. A gambler chooses a number between 1 and 6. Three fair dice are rolled, and if the gambler's number appears $k$ times ($k = 1, 2, 3$), then she wins £$k$, but if her number fails to appear she loses £1. What is the gambler's expected winnings?
4. Let $X$ be a random variable with the following PMF, where $0 \leq \alpha \leq 1/2$ is a constant:

$$f(x) = \begin{cases} \alpha & \text{if } x = 1, \\ 1 - 2\alpha & \text{if } x = 2, \\ \alpha & \text{if } x = 3, \\ 0 & \text{otherwise,} \end{cases}$$

(a) Find $\mathbb{E}(X)$ and $\mathbb{E}\left(\frac{1}{X}\right)$.

(b) Find a condition on $\alpha$ under which $\mathbb{E}\left(\frac{1}{X}\right) = \frac{1}{\mathbb{E}(X)}$.

(c) Use your result to show that $\mathbb{E}\left(\frac{1}{X}\right) \neq \frac{1}{\mathbb{E}(X)}$ in general.

# Lecture 10 Moments

## 10.1 Transformations of random variables

Let $(\Omega, \mathbb{P})$ be a finite probability space, let $X$ be a random variable on $\Omega$, and let $\{x_1, \ldots, x_n\}$ be its range.

For any function $g : \mathbb{R} \to \mathbb{R}$, the *composition* of $g$ with $X$ is defined to be the function

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

The function $g(X)$ is also a random variable on $\Omega$. Let $\{y_1, y_2, \ldots, y_m\}$ denote its range.

The PMF of $g(X)$ is completely determined by the PMF of $X$:

$$\mathbb{P}\big[g(X) = y_j\big] = \sum_{x_i \in B_j} \mathbb{P}(X = x_i) \quad \text{where} \quad B_j = \{x_i : g(x_i) = y_j\}$$

Note that the sets $B_1, B_2, \ldots, B_m$ form a partition of $\{x_1, x_2, \ldots, x_n\}$.

To compute the expected value of the transformed variable $g(X)$, the following theorem shows that we need not compute its PMF explicitly. The result is sometimes called the *law of the unconscious statistician.*

**Theorem 10.1**
Let $f(x)$ denote the PMF of $X$. Then for any function $g : \mathbb{R} \to \mathbb{R}$, the expected value of $g(X)$ is

$$\mathbb{E}\big[g(X)\big] = \sum_{i=1}^{n} g(x_i) f(x_i)$$

**Proof:** Let $\{y_1, \ldots, y_m\}$ be the range of $g(X)$, and let

$$B_j = \{x_i : g(x_i) = y_j\}.$$

Then $\mathbb{P}\big[g(X) = y_j\big] = \sum_{x_i \in B_j} \mathbb{P}(X = x_i)$, so

$$\begin{aligned}
\mathbb{E}\big[g(X)\big] &= \sum_{j=1}^{m} y_j\, \mathbb{P}\big[g(X) = y_j\big] \\
&= \sum_{j=1}^{m} y_j \sum_{x_i \in B_j} \mathbb{P}(X = x_i) \\
&= \sum_{j=1}^{m} \sum_{x_i \in B_j} y_j \mathbb{P}(X = x_i) \\
&= \sum_{j=1}^{m} \sum_{x_i \in B_j} g(x_i) \mathbb{P}(X = x_i) \quad \text{because } y_j = g(x_i) \text{ whenever } x_i \in B_j, \\
&= \sum_{i=1}^{n} g(x_i) \mathbb{P}(X = x_i),
\end{aligned}$$

where the last equality follows because $\{B_1, B_2, \ldots, B_m\}$ is a partition of $\{x_1, \ldots, x_n\}$.

**Example 10.2**
Let $X$ be a random variable with the following PMF:

| $x$ | $-2$ | $-1$ | $1$ | $3$ |
|---|---|---|---|---|
| $\mathbb{P}(X = x)$ | $1/4$ | $1/8$ | $1/4$ | $3/8$ |

Find the expected value of $Y = X^2$.

**Solution:** The random variable $Y = X^2$ has the following PMF:

| $y$ | $1$ | $4$ | $9$ |
|---|---|---|---|
| $\mathbb{P}(Y = y)$ | $3/8$ | $1/4$ | $3/8$ |

The expected value of $Y$ is therefore

$$\mathbb{E}(Y) = \sum_{y\in\{1,4,9\}} y\,\mathbb{P}(Y = y) = \left(1 \times \frac{3}{8}\right) + \left(4 \times \frac{1}{4}\right) + \left(9 \times \frac{3}{8}\right) = \frac{19}{4}.$$

Alternatively,

$$\mathbb{E}(X^2) = \sum_{x\in\{-2,-1,1,3\}} x^2\,\mathbb{P}(X = x) = \left(4 \times \frac{1}{4}\right) + \left(1 \times \frac{1}{8}\right) + \left(1 \times \frac{1}{4}\right) + \left(9 \times \frac{3}{8}\right) = \frac{19}{4}.$$

## 10.2 Moments

**Definition 10.3**
Let $X : \Omega \to \mathbb{R}$ be a random variable, and let $k \in \{0, 1, 2, \ldots\}$ be a non-negative integer.

(1) $\mathbb{E}(X^k)$ is called the *$k$th moment of $X$ about the origin*, and denoted by $\mu'_k$. In particular,

- $\mu'_0 = \mathbb{E}(X^0) = 1$,
- $\mu'_1 = \mathbb{E}(X)$ is called the *mean* of $X$. This is usually denoted by $\mu$.
- $\mu'_2 = \mathbb{E}(X^2)$ is called the *mean-square* of $X$.

(2) $\mathbb{E}\big[(X-\mu)^k\big]$ is called the *$k$th moment of $X$ about the mean*, and denoted by $\mu_k$. In particular,

- $\mu_0 = 1$,
- $\mu_1 = 0$,
- $\mu_2 = \mathbb{E}\big[(X-\mu)^2\big]$ is called the *variance* of $X$. This is usually denoted by $\sigma^2$.

**Remark 10.4**
- Moments about the origin are also called *raw moments*.
- Moments about the mean are also called *central moments*.

## 10.3 Variance

Let $X$ be a random variable with range $\{x_1, x_2, \ldots, x_n\}$, let $f(x_i) = \mathbb{P}(X = x_i)$ denote its PMF.
The expected value and variance of $X$ are

$$\mathbb{E}(X) = \sum_{i=1}^n x_i f(x_i) \equiv \mu. \qquad \text{and} \qquad \mathrm{Var}(X) = \sum_{i=1}^n (x_i - \mu)^2 f(x_i) \equiv \sigma^2.$$

- The expected value $\mathbb{E}(X)$ represents the *centre* or *location* of a distribution..
- The variance $\mathrm{Var}(X)$ quantifies its *spread* or *dispersion* around the expected value.

Expectation is a linear operator (Theorem 9.5):

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

In contrast, variance does *not* have the linearity property:

**Theorem 10.5**
If $a, b \in \mathbb{R}$, then $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

**Proof:** By the linearity of expectation, $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ and

$$\begin{aligned}
\mathrm{Var}(aX + b) &= \mathbb{E}\left(\big[(aX + b) - \mathbb{E}(aX + b)\big]^2\right) \\
&= \mathbb{E}\left(\big[aX - \mathbb{E}(aX)\big]^2\right) \\
&= \mathbb{E}\left(a^2\big[X - \mathbb{E}(X)\big]^2\right) \\
&= a^2\mathbb{E}\left(\big[X - \mathbb{E}(X)\big]^2\right) \\
&= a^2\mathrm{Var}(X)
\end{aligned}$$

**Remark 10.6**
Note that the variance operator is *translation invariant*:

$$\mathrm{Var}(X + b) = \mathrm{Var}(X) \quad \text{for all} \quad b \in \mathbb{R}.$$

Central moments can be expressed in terms of raw moments. In particular, we have the following expression for variance, which is useful when performing computations.

**Lemma 10.7**
$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

**Proof:** Let $\{x_1, x_2, \ldots, x_n\}$ be the range of $X$, and let $\mu = \mathbb{E}(X)$. Taking $g(x) = (x - \mu)^2$ in Theorem 10.1,

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}\big[(X - \mu)^2\big] &= \sum_{i=1}^{n}(x_i - \mu)^2 f(x_i) \\
&= \sum_{i=1}^{n}(x_i^2 - 2x_i\mu + \mu^2) f(x_i) \\
&= \sum_{i=1}^{n} x_i^2 f(x_i) - 2\mu\sum_{i=1}^{n} x_i f(x_i) + \mu^2\sum_{i=1}^{n} f(x_i) \\
&= \mathbb{E}(X^2) - \mu^2
\end{aligned}$$

where the last equality follows because $\sum_{i=1}^{n} x_i f(x_i) = \mu$ and $\sum_{i=1}^{n} f(x_i) = 1$.

**Example 10.8**
Let $X$ be a random variable taking values in the range $\{1, 2, 3, 4, 5, 6\}$, and let $f(x)$ denote its PMF. Find the variance of $X$ when

(1) $f(x) = 1/6$ for all $x \in \{1, 2, 3, 4, 5, 6\}$;

(2) $f(x) = 1/4$ for $x \in \{3, 4\}$ and $f(x) = 1/8$ for $x \in \{1, 2, 5, 6\}$.

**Solution:**

- $\mathbb{E}(X) = \sum_{x=1}^{6} xf(x) = \frac{1}{6}(1+2+3+4+5+6) = \frac{7}{2}$.
- $\mathbb{E}(X^2) = \sum_{x=1}^{6} x^2 f(x) = \frac{1}{6}(1+4+9+16+25+36) = \frac{91}{6}$.
- $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12} = 2.92$.

- $\mathbb{E}(X) = \sum_{x=1}^{6} xf(x) = \frac{1}{8}(1+2+5+6) + \frac{1}{4}(3+4) = \frac{7}{2}$.
- $\mathbb{E}(X^2) = \sum_{x=1}^{6} x^2 f(x) = \frac{1}{8}(1+4+25+36) + \frac{1}{4}(9+16) = \frac{58}{4}$.
- $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{58}{4} - \frac{49}{4} = \frac{9}{4} = 2.25$.

## 10.4 Location, scale and shape*

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

For any random variable $X : \Omega \to \mathbb{R}$,

(1) the first moment of $X$ about the origin ($\mu$) describes its location,

(2) the second moment of $X$ about the mean ($\sigma^2$) describes its scale, and

(3) the higher moments of $X$ describe the shape of its distribution.

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

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

To find the value of $\mu$ that minimises the expected squared deviation, we can differentiate the right-hand side with respect to $\mu$ and set the resulting expression to zero. This yields $\mu = \mathbb{E}(X)$.

- The location of a distribution is described by its *expected value*, $\mu$.

**Scale** The size of $X$ should not depend on its location. Thus we consider the *centred* variable $Y = X - \mu$, which has the property that $\mathbb{E}(Y) = 0$. The expected squared deviation of $X$ around $\mu$ is equal to its *variance*, $\sigma^2 = \mathbb{E}\big[(X-\mu)^2\big]$.

- The scale of a distribution is quantified by its *standard deviation*, $\sigma$.

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

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

**Lemma 10.9**
$\mathbb{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$.

**Proof:**

- By the linearity of expectation, $\mathbb{E}(Z) = \mathbb{E}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\big[\mathbb{E}(X) - \mu\big] = 0$.

- By the properties of variance, $\text{Var}(Z) = \text{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X) = 1.$

## 10.5 Skewness and kurtosis*

Higher moments of the standardised variable $Z = \dfrac{X-\mu}{\sigma}$ quantify the *shape* of a distribution.

**Definition 10.10**
The *skewness* of a random variable $X$ is

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}$$

where $\mu_3 = \mathbb{E}\big[(X-\mu)^3\big]$ is the third central moment of $X$.

Skewness is a measure of *asymmetry*:

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

**Definition 10.11**
The *kurtosis* of a random variable $X$ is

$$\gamma_2 = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}$$

where $\mu_4 = \mathbb{E}\big[(X-\mu)^4\big]$ is the fourth central moment of $X$.

The word *kurtosis* comes from the Greek ***kurtos- 'bulge'***.

- Large kurtosis (leptokurtic: $\gamma_2 > 3$). Tall and narrow peak, with heavy tails.
- Small kurtosis (platykurtic: $\gamma_2 < 3$). Low and wide peak, with thin tails.

## 10.6 Exercises

**Exercise 10.1**

1. Let $X$ be a random variable with the following PMF, where $c$ is a constant:

| $x$ | $-2$ | $0$ | $2$ |
|---|---|---|---|
| $\mathbb{P}(X=x)$ | $c/6$ | $c/2$ | $c/3$ |

   (a) Find the value of $c$.
   (b) Show that $\mathbb{E}(X) = 1/3$ and $\text{Var}(X) = 17/9$.
   (c) Find $\mathbb{E}(Y)$ and $\text{Var}(Y)$ when $Y = 3X + 4$.

2. Let $X$ be a random variable with the following PMF, where $c$ is a constant:

| $x$ | $1$ | $2$ | $3$ |
|---|---|---|---|
| $\mathbb{P}(X=x)$ | $c/8$ | $c/8$ | $c/4$ |

(a) Find the value of $c$.

(b) Find $\mathbb{E}(X)$ and $\mathrm{Var}(X)$.

(c) Let $Y = 2X + 3$. Show that $\mathbb{E}(Y) = 15/2$ and $\mathrm{Var}(Y) = 11/4$.

3. Let $X$ be a random variable with the following PMF, where $c$ is a constant:

$$f(x) = \begin{cases} \frac{x}{8} & \text{for} \quad x = 1, 2, \\ c & \text{for} \quad x = 3, 4, \\ \frac{1}{8} & \text{for} \quad x = 5, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that $c = 1/4$.

(b) Find the mean and variance of $X$.

(c) Find the mean and variance of $Y = 2X - 1$.

4. Let $X$ be a random variable with mean $\mu$ and variance $\sigma^2$. If $c$ is any real number, show that

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

and deduce that $\mathbb{E}\big[(X - c)^2\big]$ is minimised when $c = \mu$.

# Lecture 11 The Uniform, Bernoulli and Binomial Distributions

In this lecture we look at some well-known distributions which take values in a finite subset of the non-negative integers.

## 11.1 Uniform distribution

The uniform distribution on $\{1, 2, \ldots, n\}$ assigns an equal probability to each value, and as such underpins the classical model of probability (Lecture 6).

| | |
|---|---|
| Notation | $X \sim \text{Uniform}\{1, 2, \ldots, n\}$ |
| Parameter(s) | $n \in \mathbb{N}$ |
| Range | $\{1, 2, \ldots, n\}$ |
| PMF | $f(k) = 1/n$ for all $k = 1, 2, \ldots, n$ |

**Lemma 11.1**
If $X \sim \text{Uniform}\{1, 2, \ldots, n\}$, then

$$\mathbb{E}(X) = \frac{n+1}{2} \quad \text{and} \quad \text{Var}(X) = \frac{n^2-1}{12}.$$

**Proof:**

$$\mathbb{E}(X) = \sum_{k=1}^{n} k\, f(k) = \frac{1}{n}\sum_{k=1}^{n} k = \frac{1}{n}\left(\frac{n(n+1)}{2}\right) = \frac{n+1}{2}.$$

$$\mathbb{E}(X^2) = \sum_{k=1}^{n} k^2\, f(k) = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{1}{n}\left(\frac{n(n+1)(2n+1)}{6}\right) = \frac{(n+1)(2n+1)}{6}.$$

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{(n-1)(n+1)}{12} = \frac{n^2-1}{12}.$$

**Example 11.2 (Entropy)**
Entropy is a way of quantifying the uncertainty associated with a random variable.

Let $X$ be a random variable taking values in $\{1, 2, \ldots, n\}$, and let $f(k) = \mathbb{P}(X = k)$ be its PMF. The *entropy* of $X$ is defined by

$$H(X) = -\sum_{k=1}^{n} f(k) \log f(k)$$

If $X$ is a non-random, say $f(j) = 1$ and $f(k) = 0$ for all $k \neq j$, then

$$H(X) = -\sum_{k=1}^{n} f(k) \log f(k) = -f(j) \log f(j) = -\log 1 = 0.$$

If $X$ is uniformly distributed, so that $f(k) = 1/n$ for all $k \in \{1, 2, \ldots, n\}$, then

$$H(X) = -\sum_{k=1}^{n} \frac{1}{n} \log\left(\frac{1}{n}\right) = -\log\left(\frac{1}{n}\right) = \log n.$$

In general, for any random variable $X$ taking values in $\{1, 2, \ldots, n\}$, it can be shown that

$$0 \leq H(X) \leq \log n.$$

Among all probability distributions on $\{1, 2, \ldots, n\}$, the uniform distribution has *maximum entropy.*

## 11.2 Bernoulli distribution

The Bernoulli distribution is the distribution of an indicator variable.

| | |
|---|---|
| Notation | $X \sim \text{Bernoulli}(p)$ |
| Parameter(s) | $p \in [0, 1]$ (probability of success) |
| Range | $\{0, 1\}$ |
| PMF | $f(0) = 1 - p$ and $f(1) = p$ |

**Lemma 11.3**
If $X \sim \text{Bernoulli}(p)$, then

$$\mathbb{E}(X) = p \quad \text{and} \quad \text{Var}(X) = p(1-p).$$

**Proof:**

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{k=0}^{1} k\, f(k) = \big[0 \times (1-p)\big] + \big[1 \times p\big] = p. \\
\mathbb{E}(X^2) &= \sum_{k=0}^{1} k^2\, f(k) = \big[0^2 \times (1-p)\big] + \big[1^2 \times p\big] = p. \\
\text{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p - p^2 = p(1-p).
\end{aligned}$$

**Example 11.4**
Find the entropy of the Bernoulli($p$) distribution, and show that it reaches its maximum value when $p = 1/2$.

**Solution:** Let $X \sim \text{Bernoulli}(p)$. Then $f(0) = 1 - p$ and $f(1) = p$, so

$$H(X) = -\sum_{k=1}^{n} f(k) \log f(k) = -(1-p)\log(1-p) - p\log p.$$

To find the maximum, we first differentiate with respect to p:

$$\begin{aligned}\frac{dH}{dp} &= \log(1-p) - (1-p)\frac{1}{1-p}(-1) - \log p - p\,\frac{1}{p} \\ &= \log(1-p) - \log p = \log\left(\frac{1-p}{p}\right).\end{aligned}$$

Setting this equal to zero,

$$\log\left(\frac{1-p}{p}\right) = 0 \quad \Rightarrow \quad \frac{1-p}{p} = 1 \quad \Rightarrow \quad p = 1/2, \quad \text{as required.}$$

## 11.3 Binomial distribution

The binomial distribution is the distribution of the number of successes in a sequence of independent Bernoulli trials.

| | | |
|---|---|---|
| Notation | $X \sim \text{Binomial}(n, p)$ | |
| Parameter(s) | $n \in \mathbb{N}$ | (number of trials) |
| | $p \in [0, 1]$ | (probability of success) |
| Range | $\{0, 1, 2, \ldots, n\}$ | |
| PMF | $f(k) = \binom{n}{k} p^k (1-p)^{n-k}$ | |

**Lemma 11.5**
If $X \sim \text{Binomial}(n, p)$, then

$$\mathbb{E}(X) = np \quad \text{and} \quad \text{Var}(X) = np(1-p).$$

**Proof:** Consider a sequence of independent Bernoulli trials in which the probability of success is $p$. Let $X_i$ be the indicator variable of the event that success occurs on the $i$th trial, and let $X$ be the total number of successes in $n$ trials:

$$X = \sum_{i=1}^{n} X_i \qquad \text{where} \quad X_i = \begin{cases} 1 & \text{if success on the } i\text{th trial,} \\ 0 & \text{otherwise.} \end{cases}$$

Then $X_i \sim \text{Bernoulli}(p)$ for each $i = 1, 2, 3, \ldots$, and $X \sim \text{Binomial}(n, p)$. Now,

$$\mathbb{E}(X_i) = p \quad \text{and} \quad \text{Var}(X_i) = p(1-p),$$

so by the linearity of expectation,

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n) = np,$$

and because the trials are independent,

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n) = np(1-p).$$

**Example 11.6**
A multiple choice test consists of ten questions, each with a choice of three different answers. A student decides to choose the answers at random. Find the mean and variance of the number of correct answers.

**Solution:** Let $X$ be the number of correct answers. Then $X \sim \text{Binomial}(n, p)$ where

- $n = 10$ (the total number of questions), and
- $p = 1/3$ (the probability of a correct answer).

Thus

$$\mathbb{E}(X) = np = \frac{10}{3} \quad \text{and} \quad \text{Var}(X) = np(1-p) = \frac{20}{9}.$$

## 11.4 Exercises

**Exercise 11.1**

1. A pair of fair dice is rolled six times. What is the probability of getting a total of seven
   (1) twice,
   (2) at least once,
   (3) more than three times?

2. A biased coin, for which the probability of getting a head is 1/4, is tossed 10 times. What are the probabilities of observing
   (1) exactly two heads,
   (2) fewer than two heads,
   (3) more than two heads,
   (4) at most two heads,
   (5) at least two heads?

3. The probability that a production line produces a faulty item is 0.1. Are you more likely to find at most one faulty item in a sample of 10 items, or at most two faulty items in a sample of 20 items?

4. Airlines find that customers who reserve a seat fail to turn up with probability 0.1. To avoid having empty seats, EasyJet always sell 10 tickets for their 9-seater aeroplane, while Ryanair always sell 20 tickets for their 18-seater aeroplane. Which of the two airlines is most often overbooked?

5. A random variable $X$ has binomial distribution with mean 1.5 and variance 1.275. Find the probability that $X$ is at most 2.

6. Let $X_1, X_2, \ldots, X_n$ be independent random variables, with $X_i \sim \text{Bernoulli}(p_i)$ for $i = 1, 2, \ldots, n$. Show that the mean and variance of their sum $X = X_1 + X_2 + \ldots + X_n$ are given by

$$\mathbb{E}(X) = \sum_{i=1}^{n} p_i \quad \text{and} \quad \text{Var}(X) = \sum_{i=1}^{n} p_i(1-p_i) \quad \text{respectively.}$$

7. Suppose that $n$ independent Bernoulli trials are carried out, each having probability of success $p$. Let the number of successes and failures obtained in these trials be denoted by $X$ and $Y$ respectively. Find the PMF of $Z = X - Y$, and show that $\mathbb{E}(Z) = n(2p - 1)$.

   [Hint: use the fact that $Y = n - X$.]

# Long proof of Lemma 11.5*

To show: if $X \sim \text{Binomial}(n,p)$, then $\mathbb{E}(X) = np$ and $\text{Var}(X) = np(1-p)$.

Let $q = 1-p$. The PMF of $X$ can be written as

$$f(k) = \binom{n}{k} p^k q^k, \qquad k = 0,1,\ldots,n.$$

The expected value is

$$\begin{aligned}\mathbb{E}(X) = \sum_{k=0}^{n} kf(k) = \sum_{k=0}^{n} \binom{n}{k} kp^k q^{n-k} &= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k} \\ &= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} q^{n-k}.\end{aligned}$$

Letting $j = k-1$,

$$\begin{aligned}\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} q^{n-k} = \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-j-1)!} p^j q^{n-j-1} &= \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-j-1} \\ &= (p+q)^{n-1} = 1 \qquad \text{(by the binomial theorem).}\end{aligned}$$

Hence $\mathbb{E}(X) = np$. To compute the variance, we use the identity

$$\mathbb{E}(X^2) = \mathbb{E}\big[X(X-1)\big] + \mathbb{E}(X).$$

First,

$$\begin{aligned}\mathbb{E}[X(X-1)] = \sum_{k=0}^{n} k(k-1)f(k) &= \sum_{k=0}^{n} \binom{n}{k} k(k-1) p^k q^{n-k} \\ &= \sum_{k=2}^{n} \frac{n!}{(k-2)!(n-k)!} p^k q^{n-k} \\ &= n(n-1)p^2 \sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} q^{n-k}.\end{aligned}$$

Letting $j = k-2$,

$$\begin{aligned}\sum_{k=2}^{n} \frac{(n-2)!}{(k-2)!(n-k)!} p^{k-2} q^{n-k} = \sum_{j=0}^{n-2} \frac{(n-2)!}{j!(n-2-j)!} p^j q^{n-2-j} &= \sum_{j=0}^{n-2} \binom{n-2}{j} p^j q^{n-2-j} \\ &= (p+q)^{n-2} = 1 \qquad \text{(by the binomial theorem).}\end{aligned}$$

Hence $\mathbb{E}[X(X-1)] = n(n-1)p^2$. This yields

$$\mathbb{E}(X^2) = \mathbb{E}[X(X-1)] + \mathbb{E}(X) = n(n-1)p^2 + np,$$

so the variance of $X$ is

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= np(1-p).\end{aligned}$$

as required.

# Lecture 12 Joint Distributions

Let $X$ and $Y$ be simple random variables, let $\{x_1, x_2, \dots, x_m\}$ be the range of $X$, and let $\{y_1, y_2, \dots, y_n\}$ be the range of $Y$.

## 12.1 Joint distributions

**Definition 12.1**

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

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

(2) The *marginal CDF* of $X$ is the function

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

and the marginal CDF of $Y$ is

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

**Remark 12.2**

$\mathbb{P}(X \leq x, Y \leq y)$ is the probability of the event $\{\omega : X(\omega) \leq x \text{ and } Y(\omega) \leq y\}$,

For simple random variables, it is often easier to work with joint PMFs:

**Definition 12.3**

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

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

(2) The *marginal PMF* of $X$ is the function

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

and the marginal PMF of $Y$ is

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

**Lemma 12.4**

The marginal PMFs of $X$ and $Y$ satisfy

$$f_X(x_i) = \sum_{j=1}^{n} f_{X,Y}(x_i, y_j) \quad \text{and} \quad f_Y(y_j) = \sum_{i=1}^{m} f_{X,Y}(x_i, y_j).$$

**Proof:** By definition, the joint PMF of $X$ and $Y$ can be written as $f_{X,Y}(x_i, y_j) = \mathbb{P}(A_{i,j})$, where

$$A_{i,j} = \{X = x_i, Y = y_j\} \equiv \{\omega : X(\omega) = x_i \text{ and } Y(\omega) = y_j\}.$$

The sets $A_{i,1}, A_{i,2}, \ldots, A_{i,n}$ form a partition of the event $\{X = x_i\}$, so by the partition theorem,

$$f_X(x_i) = \mathbb{P}(X = x_i) = \sum_{j=1}^{n} \mathbb{P}(A_{i,j}) = \sum_{j=1}^{n} f_{X,Y}(x_i, y_j).$$

Similarly, $A_{1,j}, A_{2,j}, \ldots, A_{m,j}$ form a partition of the event $\{Y = y_j\}$, so by the partition theorem,

$$f_Y(y_j) = \mathbb{P}(Y = y_j) = \sum_{i=1}^{m} \mathbb{P}(A_{i,j}) = \sum_{i=1}^{m} f_{X,Y}(x_i, y_j).$$

**Example 12.5**
A fair die is rolled once. Let $\omega$ denote the outcome, and consider the random variables

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

Find the joint PMF of $X$ and $Y$.

**Solution:**

| $\omega$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X(\omega)$ | 1 | 2 | 1 | 2 | 1 | 2 |
| $Y(\omega)$ | 1 | 1 | 1 | 2 | 2 | 2 |

The joint PMF of $X$ and $Y$ is shown in the following table.

| | $Y = 1$ | $Y = 2$ |
|---|---|---|
| $X = 1$ | 1/3 | 1/6 |
| $X = 2$ | 1/6 | 1/3 |

The marginal PMFs are recovered by summing the rows and columns of the table.

## 12.2 Independent random variables

Two random variables are said to be *independent* if the value taken by one does not affect the distribution of the other.

**Definition 12.6**
$X$ and $Y$ are said to be *independent* if the events $\{X = x\}$ and $\{Y = y\}$ are independent for all $x, y \in \mathbb{R}$.

The joint PMF of two independent random variables is equal to the product of their marginal PMFs:

**Lemma 12.7**
$X$ and $Y$ are independent if and only if

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

**Proof:** Recall that two events $A$ and $B$ are said to be independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Thus $X$ and $Y$ are independent if and only if $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$ for all $x, y \in \mathbb{R}$, or equivalently

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

Note that if $x \notin \{x_1, x_2, \dots, x_m\}$ or $y \notin \{y_1, y_2, \dots, y_n\}$, both sides are equal to zero.

**Example 12.8**
Let $X$ and $Y$ be random variables with a joint PMF shown in the following table. Find the marginal PMFs of $X$ and $Y$, and decide whether or not $X$ and $Y$ are independent.

| | | $y$ | | |
|---|---|---|---|---|
| | | 2 | 3 | 4 |
| | 1 | 1/12 | 1/6 | 0 |
| $x$ | 2 | 1/6 | 0 | 1/3 |
| | 3 | 1/12 | 1/6 | 0 |

**Solution:** The marginal distributions are obtained by summing the rows and columns of the table:

| $x$ | 1 | 2 | 3 |
|---|---|---|---|
| $f_X(x)$ | 1/4 | 1/2 | 1/4 |

| $y$ | 2 | 3 | 4 |
|---|---|---|---|
| $f_Y(y)$ | 1/3 | 1/3 | 1/3 |

We have (for example) that $\mathbb{P}(X = 2, Y = 3) = 0$ but $\mathbb{P}(X = 2)\mathbb{P}(Y = 3) = 1/6$, so $X$ and $Y$ are not independent.

**Theorem 12.9**
Let $X$ and $Y$ be independent, and let $g, h : \mathbb{R} \to \mathbb{R}$ be any two functions. Then $g(X)$ and $h(Y)$ are also independent.

**Proof:**
Let $\{c_1, c_2, \dots, c_M\}$ be the range of $g(X)$, let $\{d_1, d_2, \dots, d_N\}$ be the range of $h(Y)$, and consider the following partitions.

- The sets $C_k = \{x_i : g(x_i) = c_k\}$ for $k = 1, 2, \dots, M$.
  These partition the set $\{x_i : i = 1, 2, \dots, m\}$, which is the range of $X$.
- The sets $D_l = \{y_j : h(y_j) = d_l\}$ for $l = 1, 2, \dots, N$.
  These partition the set $\{y_j : j = 1, 2, \dots, n\}$, which is the range of $Y$.
- The sets $E_{k,l} = \{(x_i, y_j) : g(x_i) = c_k, h(y_j) = d_l\}$ for $k = 1, 2, \dots, M$ and $l = 1, 2, \dots, N$.
  These partition the set of pairs $\{(x_i, y_j) : i = 1 : 2, \dots, m; j = 1, 2, \dots, n\}$.

By construction, $g(x) = c_k$ and $h(y) = d_l$ if and only if $(x, y) \in E_{k,l}$, so by the partition theorem,

$$\begin{aligned}
\mathbb{P}\big[g(X) = c_k, h(Y) = d_l\big] &= \sum_{(x,y)\in E_{k,l}} \mathbb{P}(X = x, Y = y) \\
&= \sum_{(x,y)\in E_{k,l}} \mathbb{P}(X = x)\mathbb{P}(Y = y) \qquad \text{(by independence)}, \\
&= \sum_{x\in C_k} \mathbb{P}(X = x) \sum_{y\in D_l} \mathbb{P}(Y = y) \\
&= \mathbb{P}\big[g(X) = c_k\big]\, \mathbb{P}\big[h(Y) = d_l\big].
\end{aligned}$$

## 12.3 Exercises

**Exercise 12.1**

1. A fair coin is tossed twice. Let $X$ be the number of heads, and let $Y$ be the indicator variable of the event $\{X = 2\}$. Find the joint PMF of $X$ and $Y$.

2. 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 $Z = X(1 - X)$. Find the joint PMF of $X$ and $Z$.

3. Let $X$ and $Y$ be two independent random variables with PMFs

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

and

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

respectively.

   (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.

4. The random variables $X$ and $Y$ have the joint PMF

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

   where $c$ is a constant.
   (a) Find the value of $c$.
   (b) Find $\mathbb{P}(X = 0, Y = -2)$.
   (c) Find $\mathbb{P}(X = 2)$.
   (d) Find $\mathbb{P}(|X - Y| \leq 1)$.

# Lecture 13 Covariance and Correlation

Let $X$ and $Y$ be simple random variables. Let $\{x_1, x_2, \ldots, x_m\}$ be the range of $X$, and let $\{y_1, y_2, \ldots, y_n\}$ be the range of $Y$. Recall that $X$ and $Y$ together are described by their joint PMF,

$$f(x, y) = \mathbb{P}(X = x, Y = y).$$

where $f(x, y) \neq 0$ only if $x \in \{x_1, x_2, \ldots, x_m\}$ and $y \in \{y_1, y_2, \ldots, y_n\}$.

## 13.1 Product moments

Let $g : \mathbb{R}^2 \to \mathbb{R}$ be any function of two variables:

$$\begin{array}{rccc} g: & \mathbb{R}^2 & \to & \mathbb{R} \\ & (x,y) & \mapsto & g(x,y) \end{array}$$

$X$ and $Y$ can be combined to create a new random variable $g(X, Y)$, defined by

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

As in Theorem 10.1, the expected value of $g(X, Y)$ can be computed without having to compute its distribution explicitly:

$$\mathbb{E}\big[g(X,Y)\big] = \sum_{i=1}^{m}\sum_{j=1}^{n} g(x_i, y_j) f(x_i, y_j).$$

The special case $g(x, y) = xy$ yields the following definition.

**Definition 13.1**
$\mathbb{E}(XY)$ is called the *product moment* of $X$ and $Y$.

To compute $\mathbb{E}(XY)$, we can either compute the PMF of $XY$ explicitly, or use the joint PMF of $X$ and $Y$ as above:

$$\mathbb{E}(XY) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_i\, y_j\, f(x_i, y_j).$$

**Lemma 13.2**
If $X$ and $Y$ are independent then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

**Proof:**

$$\begin{aligned}
\mathbb{E}(XY) &= \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j\, f(x_i, y_j) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} x_i y_j\, f_X(x_i) f_Y(y_j) \qquad \text{by independence.} \\
&= \left(\sum_{i=1}^{m} x_i\, f_X(x_i)\right)\left(\sum_{j=1}^{n} y_j\, f_Y(y_j)\right) \\
&= \mathbb{E}(X)\mathbb{E}(Y).
\end{aligned}$$

## 13.2 Correlation

**Definition 13.3**
Two random variables $X$ and $Y$ are said to be *correlated* if $\mathbb{E}(XY) \neq \mathbb{E}(X)\mathbb{E}(Y)$, otherwise they are said to be *uncorrelated*.

**Theorem 13.4**
If $X$ and $Y$ are uncorrelated then $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

**Proof:** If $X$ and $Y$ are uncorrelated, we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$, so

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

**Remark 13.5**
If $X$ and $Y$ are independent then they are uncorrelated, but the converse is not necessarily true.

## 13.3 Covariance

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

**Definition 13.6**
The *covariance* of $X$ and $Y$ is

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

**Remark 13.7**

(1) Variance is a special case of covariance: $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.

(2) $X$ and $Y$ are uncorrelated if and only if $\mathrm{Cov}(X,Y) = 0$.

**Example 13.8**
Let $X$ and $Y$ be random variables with joint PMF shown in the following table.

| | | $y$ = 2 | 3 | 4 |
|---|---|---|---|---|
| | 1 | 1/12 | 1/6 | 0 |
| $x$ | 2 | 1/6 | 0 | 1/3 |
| | 3 | 1/12 | 1/6 | 0 |

Find the covariance of $X$ and $Y$.

**Solution:** To find $\mathbb{E}(X)$ and $\mathbb{E}(Y)$, we need the marginal PMFs of $X$ and $Y$:

| $x$ | 1 | 2 | 3 |
|---|---|---|---|
| $P(X=x)$ | 1/4 | 1/2 | 1/4 |

| $y$ | 2 | 3 | 4 |
|---|---|---|---|
| $P(Y=y)$ | 1/3 | 1/3 | 1/3 |

- $\mathbb{E}(X) = (1 \times 1/4) + (2 \times 1/2) + (3 \times 1/4) = 2.$
- $\mathbb{E}(Y) = (2 \times 1/3) + (3 \times 1/3) + (4 \times 1/3) = 3.$

To find $\mathbb{E}(XY)$, we compute the PMF of the random variable $Z = XY$:

| $z$ | 2 | 3 | 4 | 6 | 8 | 9 |
|---|---|---|---|---|---|---|
| $P(XY=z)$ | 1/12 | 1/6 | 1/6 | 1/12 | 1/3 | 1/6 |

The expectation of $XY$ is therefore equal to

$$\mathbb{E}(XY) = \left(2 \times \frac{1}{12}\right) + \left(2 \times \frac{1}{6}\right) + \left(4 \times \frac{1}{6}\right) + \left(6 \times \frac{1}{12}\right) + \left(8 \times \frac{1}{3}\right) + \left(9 \times \frac{1}{6}\right) = 6.$$

Hence the covariance of $X$ and $Y$ is

$$\operatorname{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 6 - (2 \times 3) = 0.$$

Thus $X$ and $Y$ are uncorrelated (although they are not independent).

## 13.4 The correlation coefficient

The correlation coefficient of $X$ and $Y$ is the product moment of the standardised variables

$$\frac{X - \mathbb{E}(X)}{\sqrt{\operatorname{Var}(X)}} \quad \text{and} \quad \frac{Y - \mathbb{E}(Y)}{\sqrt{\operatorname{Var}(Y)}}.$$

**Definition 13.9**
The *correlation coefficient* of $X$ and $Y$ is

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

**Remark 13.10**
- We will show that $|\rho(X,Y)| \leq 1$ for any two random variables $X$ and $Y$.
- $\rho(X,Y)$ provides a *standardised* measure of (linear) dependence between $X$ and $Y$.
- Note that $X$ and $Y$ are uncorrelated if and only if $\rho(X,Y) = 0$.

**Theorem 13.11 (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 $Y = aX$ for some $a \in \mathbb{R}$.

**Proof:** Let $a \in \mathbb{R}$ consider the random variable $Z = aX - Y$.
By the properties of expectation,

$$\begin{aligned} Z^2 \geq 0 &\Rightarrow \mathbb{E}(Z^2) \geq 0 \\ &\Rightarrow \mathbb{E}(a^2X^2 - 2aXY + Y^2) \geq 0 \\ &\Rightarrow a^2\mathbb{E}(X^2) - 2a\mathbb{E}(XY) + \mathbb{E}(Y^2) \geq 0. \end{aligned}$$

- Let $h(a) = a^2\mathbb{E}(X^2) - 2a\mathbb{E}(XY) + \mathbb{E}(Y^2)$. This is a quadratic expression in $a$.

Since $h(a) \geq 0$ for all $a \in \mathbb{R}$, the roots of the quadratic equation $h(a) = 0$, given by

$$a = \frac{\mathbb{E}(XY) \pm \sqrt{\mathbb{E}(XY)^2 - \mathbb{E}(X^2)\mathbb{E}(Y^2)}}{\mathbb{E}(X^2)}$$

are either both complex (discriminant is negative) or co-incide (discriminant is zero).

- Hence $\mathbb{E}(XY)^2 - \mathbb{E}(X^2)\mathbb{E}(Y^2) \leq 0$, or equivalently, $\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$.

Finally, the discriminant is zero if and only if $Z = 0$.

- Hence $\mathbb{E}(XY)^2 = \mathbb{E}(X^2)\mathbb{E}(Y^2)$ if and only if $Y = aX$ for some $a \in \mathbb{R}$.

**Theorem 13.12**
For any two random variables $X$ and $Y$,

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

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

**Proof:** By the Cauchy-Schwarz inequality,

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

with equality if and only if $Y = aX$ for some $a \in \mathbb{R}$. Hence,

$$|\text{Cov}(X,Y)| \leq \sqrt{\text{Var}(X)\text{Var}(Y)}, \text{ which implies that } |\rho(X,Y)| \leq 1,$$

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

## 13.5 Exercises

**Exercise 13.1**

1. Let $X$ and $Y$ be two Bernoulli random variables with parameter $\frac{1}{2}$. Show that the random variables $U = X + Y$ and $V = |X - Y|$ are uncorrelated but not independent.

# Lecture 14 Conditional Distributions

Let $X$ and $Y$ be simple random variables, let $\{x_1, x_2, \dots, x_m\}$ be the range of $X$, and let $\{y_1, y_2, \dots, y_n\}$ be the range of $Y$.

## 14.1 Conditional distributions

Let $A$ and $B$ be two events. Recall that the conditional probability of $A$ given $B$ is

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

This idea extends to random variables.

**Definition 14.1**
Let $x \in \mathbb{R}$ be a fixed number, and suppose that $\mathbb{P}(X = x) > 0$.

(1) The *conditional CDF* of $Y$ given $X = x$ is the function

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

(2) The *conditional PMF* of $Y$ given $X = x$ is the function

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

**Lemma 14.2**
The conditional PMF of $Y$ given $X = x$ can be written as

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

where $f_{X,Y}$ is the joint PMF of $X$ and $Y$, and $f_X$ is the marginal PMF of $X$.

**Proof:**

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

**Remark 14.3 (Independence)**
Recall that $X$ and $Y$ are independent if and only if

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

Hence by Lemma 14.2, $X$ and $Y$ are independent if and only if

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

where $f_Y$ is the marginal PMF of $Y$. Thus if $X$ and $Y$ are independent, the value taken by $X$ does not affect the distribution of $Y$.

**Example 14.4**
Let $X$ and $Y$ be random variables with joint PMF shown in the following table.

| | | $y$ | | |
|---|---|---|---|---|
| | | 2 | 3 | 4 |
| | 1 | 1/12 | 1/6 | 0 |
| $x$ | 2 | 1/6 | 0 | 1/3 |
| | 3 | 1/12 | 1/6 | 0 |

Find the conditional distribution of $Y$ given that (i) $X = 1$, (ii) $X = 2$, (iii) $X = 3$.

**Solution:** The conditional distributions are obtained by re-scaling the rows of the table.

| | $Y = 2$ | $Y = 3$ | $Y = 4$ |
|---|---|---|---|
| $f_{Y\mid X}(y\,\vert\,0)$ | 1/3 | 2/3 | 0 |
| $f_{Y\mid X}(y\,\vert\,1)$ | 1/3 | 0 | 2/3 |
| $f_{Y\mid X}(y\,\vert\,2)$ | 1/3 | 2/3 | 0 |

## 14.2 Conditional expectation

Let $x$ be a value such that $\mathbb{P}(X = x) > 0$.

**Definition 14.5**
(1) The *conditional expectation of* $Y$ *given* $X = x$ is a number,

$$\mathbb{E}(Y|X = x) = \sum_{j=1}^{n} y_j\, f_{Y|X}(y_j|x).$$

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

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

**Remark 14.6**
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 $X$, and by Theorem 10.1,

$$\mathbb{E}\big[\mathbb{E}(Y|X)\big] \equiv \mathbb{E}\big[g(X)\big] = \sum_{i=1}^{m} g(x_i) f_X(x_i) \equiv \sum_{i=1}^{m} \mathbb{E}(Y|X = x_i) f_X(x_i)$$

where $f_X$ is the marginal distribution of $X$.

## 14.3 Law of total expectation

**Theorem 14.7 (The law of total expectation)**
For any two simple random variables $X$ and $Y$,

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

**Proof:**

$$\mathbb{E}\big[\mathbb{E}(Y|X)\big] = \sum_{i=1}^{m} \mathbb{E}(Y|X=x_i) f_X(x_i) = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} y_j f_{Y|X}(y_j|x_i) \right) f_X(x_i)$$
$$= \sum_{j=1}^{n} y_j \left( \sum_{i=1}^{m} f_{Y|X}(y_j|x_i) f_X(x_i) \right)$$
$$= \sum_{j=1}^{n} y_j \left( \sum_{i=1}^{m} f_{X,Y}(x_i, y_j) \right) = \sum_{j=1}^{n} y_j f_Y(y_j) = \mathbb{E}(Y).$$

**Remark 14.8**
The law of total expectation provides a useful way of computing $\mathbb{E}(Y)$:

$$\mathbb{E}(Y) = \sum_{i=1}^{m} \mathbb{E}(Y|X=x_i)\mathbb{P}(X=x_i).$$

This is analogous to the *partition theorem*.

**Example 14.9**
Consider the random variables $X$ and $Y$ of Example 14.4.

(1) Find the conditional expectation $Y$ given that (i) $X=1$, (ii) $X=2$ and (iii) $X=3$.

(2) Find the distribution of $\mathbb{E}(Y|X)$, and verify that the law of total expectation holds.

**Solution:** The conditional expectation of $Y$ given that $X=1$, $X=2$ and $X=3$ are, respectively,

- $\mathbb{E}(Y|X=1) = \big(2 \times \frac{1}{3}\big) + \big(3 \times \frac{2}{3}\big) + (4 \times 0) = \frac{8}{3}$,
- $\mathbb{E}(Y|X=2) = \big(2 \times \frac{1}{3}\big) + (3 \times 0) + \big(4 \times \frac{2}{3}\big) = \frac{10}{3}$,
- $\mathbb{E}(Y|X=3) = \big(2 \times \frac{1}{3}\big) + \big(3 \times \frac{2}{3}\big) + (4 \times 0) = \frac{8}{3}$.

The marginal distribution of $X$, along with the associated values of $\mathbb{E}(Y|X=x)$, are shown in the following table.

| $x$ | 1 | 2 | 3 |
|---|---|---|---|
| $\mathbb{P}(X=x)$ | 1/4 | 1/2 | 1/4 |
| $\mathbb{E}(Y\|X=x)$ | 8/3 | 10/3 | 8/3 |

Hence the distribution of $\mathbb{E}(Y|X)$ is

| $z$ | 8/3 | 10/3 |
|---|---|---|
| $\mathbb{P}\big(\mathbb{E}(Y\|X)=z\big)$ | 1/2 | 1/2 |

and the expected value of $\mathbb{E}(Y|X)$ is therefore

$$\mathbb{E}\big[\mathbb{E}(Y\mid X)\big] = \left(\frac{8}{3} \times \frac{1}{2}\right) + \left(\frac{10}{3} \times \frac{1}{2}\right) = 3.$$

This agrees with the value of $\mathbb{E}(Y)$ computed from the marginal distribution of $Y$, which verifies that the law of total expectation holds in this case.

## 14.4 Exercises

**Exercise 14.1**

1. Two random variables $X$ and $Y$ have joint PMF shown in the following table:

| | | $y$ 0 | 1 | 2 |
|---|---|---|---|---|
| | 0 | 1/24 | 3/24 | 2/24 |
| $x$ | 1 | 2/24 | 4/24 | 6/24 |
| | 2 | 1/24 | 1/24 | 4/24 |

   (a) Find the covariance and correlation coefficient of $X$ and $Y$
   (b) Find conditional expectation of $X$ given that $Y = 1$.

2. Two random variables $X$ and $Y$ have joint PMF shown in the following table:

| | | $y$ $-1$ | 0 | 1 |
|---|---|---|---|---|
| | 0 | 2/28 | 2/28 | 3/28 |
| $x$ | 1 | 1/28 | 2/28 | 4/28 |
| | 2 | 1/28 | 4/28 | 9/28 |

   (a) Find the marginal distributions of $X$ and $Y$.
   (b) Are $X$ and $Y$ are independent? Justify your answer.
   (c) Find the expected values of $X$ and $Y$.
   (d) Find the covariance of $X$ and $Y$.
   (e) Find the conditional expectation of $Y$ given that
      (i) $X = 0$,
      (ii) $X = 1$,
      (iii) $X = 2$.
   (f) Find the distribution of the conditional expectation $\mathbb{E}(Y \mid X)$.
   (g) Check that the law of total expectation holds in this case.

3. A fair six-sided die is thrown once. Let $X$ be the score shown on the die, and let $A$ be the event that $X$ is an even number. Find the conditional PMF of $X$ given that $A$ occurs, and use this to find the conditional expectation of $X$ given that $A$ occurs.

4. Two fair coins are tossed. Let $X_1$ and $X_2$ be random variables, with $X_1 = 1$ if the first coin lands on heads and $X_1 = -1$ if it lands on tails, and $X_2 = 1$ if the second coin lands on heads and $X_2 = -1$ if it lands on tails. Now consider the random variables

$$X = \frac{X_1 + X_2}{2} \qquad \text{and} \qquad Y = \frac{X_1 - X_2}{2}.$$

   (a) Compute the correlation coefficient of $X$ and $Y$.
   (b) Compute the conditional PMF and conditional expectation of $Y$ given that
      (i) $X = -1$,
      (ii) $X = 0$,
      (iii) $X = 1$.
   (c) Verify that the law of total expectation holds in this case.

# Lecture 15 Discrete Probability

So far, we have considered random experiments that have only a finite number of outcomes.

- Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be a finite sample space.
- The power set of $\Omega$, which contains all possible events, is also a finite set.

To define a probability measure on $\mathcal{P}(\Omega)$, we first defined a *probability mass function* on $\Omega$:

$$\begin{array}{rccl} p: & \Omega & \to & [0,1] \\ & \omega_k & \mapsto & p(\omega_k) \end{array}$$

with the property that $\displaystyle\sum_{k=1}^{n} p(\omega_k) = 1$. This was extended to a probability measure on $\mathcal{P}(\Omega)$ by defining the probability of an event to be the sum of the probabilities of the outcomes it contains:

$$\begin{array}{rccl} \mathbb{P}: & \mathcal{P}(\Omega) & \to & [0,1] \\ & A & \mapsto & \displaystyle\sum_{\omega \in A} p(\omega) \end{array}$$

A theory of probability on countably infinite sample spaces can be developed in much the same way, the main difference being that finite sums are replaced by infinite sums.

## 15.1 Cardinality

**Definition 15.1**
Let $A$ be a set, let $\mathbb{N}$ denote the set of natural numbers, and let $\mathbb{R}$ denote the set of real numbers.

(1) If there exists a bijection $\phi : A \to \mathbb{N}$, we say that $A$ is *countably infinite*.

(2) If there exists a bijection $\phi : A \to \mathbb{R}$, we say that $A$ is *uncountable*.

A set is called *countable* if it is either finite or countably infinite.

**Example 15.2**
Consider a random experiment in which a coin is tossed repeatedly until the first head occurs. There is a countably infinite set of possible sequences:

$$\Omega = \{H, TH, TTH, TTTH, TTTTH, \ldots\}$$

- Let $\omega_k$ be the outcome consisting of $k-1$ tails followed by a head
- Let $0 < t < 1$ be the probability that the coin shows heads.

If we assume that the trials are independent, the probability mass function for this experiment is

$$p(\omega_k) = (1-t)^{k-1}t \qquad \text{for} \quad k \in \{1, 2, 3, \ldots\}.$$

The sum of these probabilities is equal to one:

$$\sum_{k=1}^{\infty} p(\omega_k) = t\sum_{k=1}^{\infty}(1-t)^{k-1} = t\sum_{k=0}^{\infty}(1-t)^k = \frac{t}{1-(1-t)} = 1.$$

Here, we have used the formula for the sum of a geometric progression with $a = t$ and $r = (1-t)$:

$$a + ar + ar^2 + \ldots = \frac{a}{1-r} \qquad \text{provided } |r| < 1,$$

Let $A$ be the event that the coin is tossed an even number of times: $A = \{\omega_2, \omega_4, \omega_6, \ldots\}$.
The probability of this event is:

$$\mathbb{P}(A) = \sum_{\omega \in A} p(\omega) = \sum_{j=1}^{\infty} p(\omega_{2j}) = \sum_{j=0}^{\infty} t(1-t)^{2j+1} = t(1-t)\sum_{j=0}^{\infty}(1-t)^{2j} = \frac{t(1-t)}{1-(1-t)^2} = \frac{1-t}{2-t}.$$

where we have again used the formula for the sum of a geometric progression, this time with $a = t(1-t)$ and $r = (1-t)^2$.

## 15.2 Discrete probability spaces

**Definition 15.3**
Let $\Omega$ be a countable sample space, and let $\mathcal{P}(\Omega)$ denote its power set.

(1) A *probability mass function* on $\Omega$ is a function

$$\begin{array}{cccc} p & : & \Omega & \to & [0,1] \\ & & \omega & \mapsto & p(\omega), \end{array}$$

with the property that $\displaystyle\sum_{\omega\in\Omega} p(\omega) = 1$.

(2) A *probability measure* on $\Omega$ is a function

$$\begin{array}{ccccc} \mathbb{P} & : & \mathcal{P}(\Omega) & \to & [0,1] \\ & & A & \mapsto & \displaystyle\sum_{\omega\in A} p(\omega) \end{array}$$

where $p(\omega)$ is a probability mass function on $\Omega$.

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

**Remark 15.4**
Finite sets are countable, so finite probability spaces are also discrete probabilty spaces.

## 15.3 Discrete random variables

**Definition 15.5**
Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an arbitrary probability space. A random variable $X : \Omega \to \mathbb{R}$ is called

(1) *simple* if it takes only a finite number of distinct values, and

(2) *discrete* if it takes at most a countably infinite number of distinct values.

**Remark 15.6**
- Finite sets are countable, so simple random variables are also discrete random variables.
- Any random variable on a discrete probability space must be a discrete random variable.

**Definition 15.7**
The *probability mass function* (PMF) of a discrete random variable $X$ is the function

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

**Remark 15.8**

(1) $f(x) = 0$ for all but a countable number of values $x \in \mathbb{R}$.

(2) If $\{x_1, x_2, x_3, \ldots\}$ is the range of $X$, then $\sum_{k=1}^{\infty} f(x_k) = 1$. Can you prove this?

## 15.4 Convergent Series

Let $a_1, a_1, a_2, \ldots$ be an infinite sequence of real numbers. The associated *series* is the (ordered) sum of its terms:

$$\sum_{k=0}^{\infty} a_k = a_0 + a_1 + a_2 + \ldots$$

**Definition 15.9**

A series $\displaystyle\sum_{k=0}^{\infty} a_k$ is said to *converge* if $\displaystyle\sum_{k=0}^{\infty} a_k < \infty$, otherwise it is said to *diverge*.

**Example 15.10 (Convergent series)**

(1) $\displaystyle\sum_{k=0}^{\infty} ar^k = a + ar + ar^2 + ar^3 + \ldots = \frac{a}{1-r}$ provided $|r| < 1$.

(2) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots = \frac{\pi^2}{6}$.

(3) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k^4} = 1 + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \ldots = \frac{\pi^4}{90}$.

(4) $\displaystyle\sum_{k=0}^{\infty} \frac{1}{2^k} = 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots = 2$.

(5) $\displaystyle\sum_{k=0}^{\infty} \frac{1}{k!} = 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{24} + \ldots = e$.

(6) $\displaystyle\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = 1 + \lambda + \frac{\lambda^2}{2} + \frac{\lambda^3}{6} + \frac{\lambda^4}{24} + \ldots = e^{\lambda}$ for all $\lambda \in \mathbb{R}$.

Not all series are convergent.

**Example 15.11 (Divergent series)**

(1) $\displaystyle\sum_{k=1}^{\infty} k = 1 + 2 + 3 + 4 + \ldots$

(2) $\displaystyle\sum_{k=1}^{\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \ldots$ (This is called the *harmonic series*.)

Let $a_0, a_1, a_2, \ldots$ be an infinite sequence of non-negative real numbers. If the associated series $\sum_{k=0}^{\infty} a_k$ converges, the sequence can be used to define a PMF on the the non-negative integers $\{0, 1, 2, \ldots\}$ as follows:

$$f(k) = \frac{a_k}{\sum_{k=0}^{\infty} a_k} \qquad \text{for } k = 0, 1, 2, \ldots.$$

Note that $\sum_{k=0}^{\infty} f(k) = 1$, so $f$ is indeed a PMF. For the series listed in Example 15.10,

- the first series yields the *geometric distribution* (next lecture)
- the last series yields the *Poisson distribution* (next lecture).

## 15.5 Expectation

The expectation of a discrete random variable is computed in much the same way that the expectation of a simple random variable is computed, the only difference again being that finite sums are replaced by infinite sums.

**Definition 15.12**
Let $X$ be a random variable on a discrete probability space $(\Omega, \mathbb{P})$. The *expectation* of $X$ is defined to be

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

We also have the following analogues of Theorem 9.3 and Theorem 10.1, which were proved for simple random variables. The proofs for the discrete case are similar (see exercises).

**Theorem 15.13**
Let $X$ be a discrete random variable, let $\{x_1, x_2, \ldots\}$ be the range of $X$, and let $f(x)$ be its PMF. Then

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

and for any function $g : \mathbb{R} \to \mathbb{R}$,

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

**Example 15.14**
Let $X$ be a discrete random variable, taking values in the set $\{1, 2, 3, \ldots\}$ with probabilities

$$\mathbb{P}(X = k) = \frac{1}{2^k}$$

Show that $\mathbb{E}(X) = 2$.

**Solution:**

$$\begin{aligned}
\mathbb{E}(X) = \sum_{k=1}^{\infty} k\mathbb{P}(X = k) &= \sum_{k=1}^{\infty} k\frac{1}{2^k} \\
&= \sum_{j=0}^{\infty}(j+1)\frac{1}{2^{j+1}} \qquad \text{(where we have set } j = k-1\text{)} \\
&= \frac{1}{2}\left(\sum_{j=0}^{\infty} j\frac{1}{2^j} + \sum_{j=0}^{\infty}\frac{1}{2^j}\right) \\
&= \frac{1}{2}\left(\sum_{j=1}^{\infty} j\frac{1}{2^j} + 2\right) \\
&= \frac{1}{2}\mathbb{E}(X) + 1.
\end{aligned}$$

so $\mathbb{E}(X) = 2$, as required.

**Example 15.15**
Let $X$ be a discrete random variable, taking values in the range $\{1, 2, \ldots\}$ with PMF given by

$$f(k) = \frac{6}{\pi^2 k^2}.$$

Show that this is indeed a PMF, and compute the expected value of $X$.

**Solution:** This is a PDF because $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$. The expected value of $X$ is

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} k f(k) = \frac{6}{\pi^2} \sum_{k=0}^{\infty} \frac{1}{k}.$$

The series $\sum_{k=0}^{\infty} \frac{1}{k}$ diverges, so $X$ has infinite expectation!

## 15.6 Exercises

**Exercise 15.1**

1. Consider the function

$$p(k) = \begin{cases} \dfrac{c}{k(k+1)} & \text{for } k = 1, 2, 3, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

   Find the value of $c$ for which this function is a probability mass function on the positive integers.

   [*Hint. Use partial fractions and the so-called "method of differences".*]

2. Let $X : \Omega \to \mathbb{R}$ be a discrete random variable, let $\{x_1, x_2, \ldots\}$ be the range of $X$, let $f(x)$ be the PMF of $X$, and let $g : \mathbb{R} \to \mathbb{R}$ be any function.

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

   (b) Show that $\mathbb{E}\big[g(X)\big] = \sum_{i=1}^{\infty} g(x_i) f(x_i)$.

# Lecture 16 The Geometric and Poisson Distributions

In this lecture we look at some well-known distributions which take values in the non-negative integers.

## 16.1 Geometric distribution

The geometric distribution on $\{1, 2, 3, \ldots\}$ is the distribution of the number of trials up to and including the first success in a sequence of independent Bernoulli trials.

| | |
|---|---|
| Notation | $X \sim \text{Geometric}(p)$ |
| Parameter(s) | $p \in [0, 1]$ (probability of success) |
| Range | $\{1, 2, \ldots\}$ |
| PMF | $fk) = (1-p)^{k-1}p$ |
| CDF | $F(k) = 1 - (1-p)^k$ |

**Lemma 16.1**
If $X \sim \text{Geometric}(p)$, then

$$\mathbb{E}(X) = \frac{1}{p} \quad \text{and} \quad \text{Var}(X) = \frac{1-p}{p^2}.$$

**Proof:** We make use of the geometric series

$$\sum_{k=0}^{\infty} r^k = 1 + r + r^2 + \ldots = \frac{1}{1-r} \qquad \text{for } |r| < 1$$

Let $q = 1 - p$. The PDF of $X \sim \text{Geometric}(p)$ is

$$f(k) = q^{k-1}p.$$

$$\begin{aligned}
\mathbb{E}(X) = \sum_{k=1}^{\infty} k\mathbb{P}(X = k) &= \sum_{k=1}^{\infty} kq^{k-1}p \\
&= \sum_{j=0}^{\infty} (j+1)q^j p \qquad \text{(where we have set } j = k-1) \\
&= q\sum_{j=0}^{\infty} jq^{j-1}p + p\sum_{j=0}^{\infty} q^j \\
&= q\sum_{j=1}^{\infty} j\,\mathbb{P}(X = j) + \frac{p}{1-q} \\
&= q\mathbb{E}(X) + 1.
\end{aligned}$$

Solving this equation, we get $\mathbb{E}(X) = \dfrac{1}{p}$.

$$
\begin{aligned}
\mathbb{E}(X^2) = \sum_{k=1}^{\infty} k^2 q^{k-1} p &= \sum_{j=0}^{\infty} (j+1)^2 q^j p \qquad\qquad \text{(where we have set } j = k-1\text{)} \\
&= \sum_{j=0}^{\infty} (j^2 + 2j + 1) q^j p \\
&= \sum_{j=0}^{\infty} j^2\, q^j p + 2 \sum_{j=0}^{\infty} j\, q^j p + \sum_{j=0}^{\infty} q^j p \\
&= q \sum_{j=1}^{\infty} j^2\, q^{j-1} p \ + \ 2q \sum_{j=1}^{\infty} j\, q^{j-1} p \ + \ p \sum_{j=0}^{\infty} q^j \\
&= q \sum_{j=1}^{\infty} j^2\, \mathbb{P}(X = j) \ + \ 2q \sum_{j=1}^{\infty} j\, \mathbb{P}(X = j) \ + \ \frac{p}{1-q} \\
&= q\mathbb{E}(X^2) + 2q\mathbb{E}(X) + 1.
\end{aligned}
$$

Substituting for $\mathbb{E}(X)$,

$$
\begin{aligned}
\mathbb{E}(X^2) &= \frac{2q\mathbb{E}(X) + 1}{1-q} = \frac{(2-p)}{p^2}, \text{ so} \\
\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 &= \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.
\end{aligned}
$$

as required.

**Example 16.2**
Let $p = 0.001$ be the probability that a certain type of lightbulb fails on any given day.

(1) What is the expected lifetime of this type of lightbulb?

(2) What is the probability that a lightbulb lasts for more than 30 days?

**Solution:** Let $X$ denote the lifetime of a bulb. Then $X \sim \text{Geometric}(p)$ where $p = 0.001$.

(1) $\mathbb{E}(X) = \dfrac{1}{p} = 1000$ days.

(2) The distribution function of $X$ is

$$\mathbb{P}(X \leq k) = 1 - (1-p)^k,$$

so $\mathbb{P}(X > k) = (1-p)^k$ and hence

$$\mathbb{P}(X > 30) = (1 - 0.001)^{30} = 0.999^{30} = 0.9704.$$

**Remark 16.3**
Let $X$ be the number of failures before the first success in a sequence of independent Bernoulli trials (i.e. the number of trials up to but *not* including the first success). Confusingly, the distribution of $X$ is also called the geometric distribution. In this case,

- $X$ takes values in the non-negative integers $\{0, 1, 2, \ldots\}$,
- its PMF is $f(k) = (1-p)^k p$, and
- $\mathbb{E}(X) = \dfrac{1-p}{p}$ and $\mathrm{Var}(X) = \dfrac{1-p}{p^2}$ (as before).

## 16.2 Poisson distribution

Consider an event that occurs repeatedly at random times. If the time interval between successive occurences are independent of each other, the number of occurences per unit time can be modelled by a Poisson distribution, named after Siméon Poisson (1781-1840). This distribution has a single parameter, called the *rate parameter*, which is the mean number of occurences per unit time.

| | |
|---|---|
| Notation | $X \sim \text{Poisson}(\lambda)$ |
| Parameter(s) | $\lambda > 0$ (rate parameter) |
| Range | $\{0, 1, 2, \ldots\ldots\}$ |
| PMF | $f(k) = \dfrac{\lambda^k}{k!} e^{-\lambda}$ |

**Lemma 16.4**
If $X \sim \text{Poisson}(\lambda)$, then

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

**Proof:**

$$\begin{aligned}
\mathbb{E}(X) = \sum_{k=0}^{\infty} k\mathbb{P}(X = k) &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{k\lambda^k}{k!} && \text{because the term for } k = 0 \text{ is zero,} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} && \text{because } k! = k(k-1)!, \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} && \text{where we have set } j = k - 1, \\
&= \lambda && \text{because } \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = e^{\lambda}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}(X^2) = \sum_{k=0}^{\infty} k^2\mathbb{P}(X = k) &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{k^2\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=1}^{\infty} \frac{k^2\lambda^k}{k!} && \text{because the term for } k = 0 \text{ is zero,} \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{k\lambda^{k-1}}{(k-1)!} && \text{because } k! = k(k-1)!, \\
&= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{(j+1)\lambda^j}{j!} && \text{where we have set } j = k - 1, \\
&= \lambda e^{-\lambda} \left[ \sum_{j=0}^{\infty} \frac{j\lambda^j}{j!} + \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \right] \\
&= \lambda e^{-\lambda} \left[ \lambda e^{\lambda} + e^{\lambda} \right] && \text{because } e^{-\lambda} \sum_{j=0}^{\infty} \frac{j\lambda^j}{j!} = \mathbb{E}(X) = \lambda, \\
&= \lambda(\lambda + 1).
\end{aligned}$$

Thus

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda.$$

**Example 16.5**
A call centre receives an average of five calls every three minutes. What is the probability that there will be

(1) no calls in the next minute, and

(2) at least two calls in the next minute?

**Solution:** Let $X$ be the number of calls received in any given minute. The average number of calls per minute is $\lambda = \frac{5}{3}$. If we model $X$ by a Poisson distribution rate parameter $\lambda$, then the probability that $k$ calls are received in any given minute is

$$\mathbb{P}(X = k) = \frac{1}{k!}\left(\frac{5}{3}\right)^k e^{-5/3}$$

(1) $\mathbb{P}(X = 0) = e^{-5/3} = 0.189$.

(2) $\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - e^{-5/3} - \frac{5}{3}e^{-5/3} = 0.496$.

## 16.3 The law of rare events

The *law of rare events*, also known as the *Poisson limit theorem*, shows that the binomial distribution can be approximated by a Poisson distribution when the number of trials ($n$) is large, and the probability of success ($p$) is small.

**Theorem 16.6 (The law of rare events)**
Let $X \sim \text{Binomial}(n, p)$, and suppose that $p \to 0$ as $n \to \infty$ in such a way that the product $\lambda = np$ remains constant. Then

$$\mathbb{P}(X = k) \to \frac{\lambda^k}{k!}e^{-\lambda} \quad \text{as} \quad n \to \infty.$$

**Proof:** To prove the theorem, we need the fact that for any $c \in \mathbb{R}$,

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

Let $\lambda = np$.

$$\begin{aligned}
\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^k &= \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\
&= \frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n!}{(n-k)!n^k}\left(\frac{\lambda^k}{k!}\right)\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}
\end{aligned}$$

Now,

$$\begin{aligned}
\frac{n!}{(n-k)!n^k} &= \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k} \\
&= 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right) \\
&\to 1 \quad \text{as} \quad n \to \infty
\end{aligned}$$

Furthermore,

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda} \quad \text{and} \quad \left(1 - \frac{\lambda}{n}\right)^{-k} \to 1 \quad \text{as} \quad n \to \infty.$$

Thus we have that

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^k \to \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{as} \quad n \to \infty.$$

which is the PDF of a Poisson($\lambda$) random variable.

**Example 16.7**
On average, a typist makes one error in every 500 words. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

**Solution:** Assume that typing a single word incorrectly is a Bernoulli trial with probability of 'success' equal to $\frac{1}{500}$, and that whether any given word is typed incorrectly is independent of any other word being typed incorrectly.
Let $X$ be the number of errors in five pages, or 1500 words.
Then $X \sim \text{Binomial}\left(1500, \frac{1}{500}\right)$, so

$$P(X \leq 2) = \sum_{k=0}^{2} \binom{1500}{k} \left(\frac{1}{500}\right)^k \left(\frac{499}{500}\right)^{1500-k} = 0.4230.$$

Using the Poisson approximation with $\lambda = 1500 \times \frac{1}{500} = 3$,

$$\mathbb{P}(X \leq 2) \approx e^{-3} \left(1 + 3 + \frac{3^2}{2}\right) = 0.4232.$$

## 16.4 Exercises

**Exercise 16.1**

1. A machine produces washers which must satisfy certain size and quality constraints. It produces defective items with probability $p$. Every hour a sample of 10 washers is inspected, and if two or more washers are found to be defective, the machine is stopped and overhauled.
   (a) Write down an expression for the probability that the machine is stopped after a particular sample.
   (b) Find the expected time between successive stoppages.
2. The *negative binomial distribution* is the distribution of the number independent Bernoulli trials that are required to obtain a fixed number of successes. Let $X \sim \text{NegativeBinomial}(r, p)$, where $r$ is the desired number of successes, and $p$ is the probability of success on each trial.
   (a) Show that the PMF of $X$ is given by
   $$\mathbb{P}(X = k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r \qquad \text{for } k = r, r+1, r+2, \ldots \text{ (and zero otherwise).}$$
   (b) Using the fact that $X$ can be written as the sum of independent geometric random variables, show that $\mathbb{E}(X) = r/p$ and $\text{Var}(X) = r(1-p)/p^2$.
   (c) Biological populations are often sampled using a technique called *inverse binomial sampling*, where individuals are successively chosen and examined until the required number of individuals with a particular characteristic has been obtained. A biologist wishes to obtain a sample of 100 fruit flies having a certain genetic trait that occurs at a rate of one in every twenty fruit files in the population. Write down an expression for the probability that the biologist has to examine at least $k$ flies?
   (d) To encourage sales of a certain breakfast cereal, each packet contains a prize token with probability $p$. To secure a particular prize, a total of 10 tokens must be obtained. Find the PMF of the number of packets that must be purchased to obtain 10 tokens, then find the mean and variance of this number.
3. Let $X \sim \text{Poisson}(1)$ and define the random variable
   $$\begin{aligned} Y &= X \quad \text{if } X \leq 3, \\ Y &= 3 \quad \text{if } X > 3. \end{aligned}$$
   Find the PMF of $Y$, and compute its expected value.
4. A commercial insecticide is advertised as being 99.9% effective. Suppose that 4000 insects infest a rose garden where the insecticide has been applied. Let $X$ denote the number of surviving insects.
   (a) What probability distribution might be a reasonable model for this experiment?
   (b) Write down an expression for the probability that fewer than three insects survive.
   (c) Compute an approximation to the probability that fewer than three insects survive.
5. Let $Y \sim \text{Poisson}(\lambda)$.
   (a) Starting with the series expansion $e^x = \displaystyle\sum_{k=0}^{\infty} \frac{x^k}{k!}$ of the exponential function, show that
   $$x + \frac{x^3}{3!} + \frac{x^5}{5!} + \frac{x^7}{7!} + \ldots = \frac{1}{2}(e^x - e^{-x}) \qquad \text{(for every } x \in \mathbb{R}\text{).}$$
   (b) Using your answer to part (a), find the probability that the value taken by $Y$ is an odd number.
   (c) Show that this probability is approximately equal to $1/2$ when $\lambda$ is large.
6. Let $X$ and $Y$ be independent random variables, with $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$. Show that $X + Y \sim \text{Poisson}(\lambda + \mu)$.
   [*Hint. Find the probability* $\mathbb{P}(X + Y = k)$ *and use the binomial theorem:* $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$.]

# Lecture 17 Continuous Probability

## 17.1 General probability spaces

Let $\Omega = \{\omega_1, \omega_2, \ldots\}$ be a countably infinite sample space.

- There are infinitely many possible subsets of $\Omega$.
- We may wish to consider an infinite number of events.

For example, suppose that a coin is tossed repeatedly until the first head occurs. For every prime number $p$, let $A_p$ be the event that the number of times the coin is tossed is divisible by $p$. There are infinitely many prime numbers, so we have an infinite number of events to consider: $A_2, A_3, A_5, A_7, \ldots$. In such situations, to define probability measures properly, we must consider families of events that are not only closed under finite unions (Lecture 2), but which are also closed under *countable* unions.

**Definition 17.1**
Let $\{A_1, A_2, \ldots\}$ be a countable family of sets over $\Omega$.

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

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

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

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

**Definition 17.2**
Let $\Omega$ be any set. A family of sets $\mathcal{F}$ is called a *$\sigma$-field* over $\Omega$ if

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

(2) $\mathcal{F}$ is closed under complementation: if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$, and

(3) $\mathcal{F}$ is closed under countable unions: if $A_1, A_2, \ldots \in \mathcal{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

In general, a probability space consists of a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

(1) $\Omega$ is a sample space,

(2) $\mathcal{F}$ is a $\sigma$-field, and

(3) $\mathbb{P}$ a probability measure.

**Remark 17.3 (Cardinality)**

- If $\Omega$ is countably infinite, its power set is uncountable (Cantor's theorem).
  - In this case, we can still use the power set of $\Omega$ as a $\sigma$-field of events.
- If $\Omega$ is uncountable, its power set is "very uncountable".
  - In this case, the power set of $\Omega$ is simply too big, and we must define an explicit $\sigma$-field of events.

## 17.2 Continuous random variables

- Discrete random variables take values in a finite or countably infinite set.
- Continuous random variables take arbitrary values in an interval or collection of intervals.

**Definition 17.4**
A random variable $X$ is called *continuous* if its CDF can be written as

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

for some integrable function $f : \mathbb{R} \to [0, \infty)$ called the *probability density function* (PDF) of $X$.

**Remark 17.5**
(1) $f(x) = F'(x)$ for all $x \in \mathbb{R}$.

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

(3) Probabilities correspond to areas under the PDF curve: $\mathbb{P}(a < X \leq b) = \displaystyle\int_a^b f(x)\,dx = F(b) - F(a)$.

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

**Example 17.6**
Let $X$ be a random variable with the following CDF:

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

where $\lambda > 0$ is a constant. Show that $X$ is a continuous random variable.

**Solution:** $X$ is a continuous random variable, because there exists a function $f(x) = \lambda e^{-\lambda x}$ such that

$$\begin{aligned} \int_{-\infty}^{x} f(t)\,dt &= \lambda \int_{-\infty}^{x} e^{-\lambda t}\,dt \\ &= \lambda \left[ \frac{-e^{-\lambda t}}{\lambda} \right]_{-\infty}^{x} \\ &= 1 - e^{-\lambda x} \\ &= F(x). \end{aligned}$$

This is the *negative exponential distribution* with (rate) parameter $\lambda$ (see next lecture).

**Example 17.7**
A straight rod is thrown down at random onto a horizontal plane, and the angle between the rod and a certain fixed orientation is measured. In the absence of any further information, a natural probability measure on the sample space $\Omega = [0, 2\pi)$ is

$$\mathbb{P}\big(\{\omega \in (a, b)\}\big) = \frac{b - a}{2\pi} \quad \text{for} \quad 0 \leq a < b < 2\pi.$$

Let $F_X$ and $F_Y$ be the CDFs of the random variables $X(\omega) = \omega$ and $Y(\omega) = \omega^2$ respectively:

$$F_X(x) = \begin{cases} 0 & x \leq 0, \\ \dfrac{x}{2\pi} & 0 \leq x < 2\pi, \\ 1 & x \geq 2\pi, \end{cases} \quad \text{and} \quad F_Y(y) = \begin{cases} 0 & y \leq 0, \\ \dfrac{\sqrt{y}}{2\pi} & 0 \leq y < 4\pi^2, \\ 1 & y \geq 4\pi^2. \end{cases}$$

The PDFs of $X$ and $Y$ are, respectively,

$$f_X(x) = \begin{cases} \dfrac{1}{2\pi} & 0 \leq x \leq 2\pi, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad f_Y(y) = \begin{cases} \dfrac{1}{4\pi\sqrt{y}} & 0 \leq y \leq 4\pi^2, \\ 0 & \text{otherwise.} \end{cases}$$

**Remark 17.8**
Let $X$ be a continuous random variable, and let $f(x)$ denote its PDF.

- The numerical value $f(x)$ is *not* a probability.

However, the following heuristic interpretation is often useful. Let $\delta x > 0$ be a small positive number. If $f(x)$ is a continuous function, then

$$\begin{aligned} \mathbb{P}(x < X \leq x + \delta x) &= F(x + \delta x) - F(x) \\ &= \int_{-\infty}^{x+\delta x} f(x)\,dx - \int_{-\infty}^{x} f(x)\,dx \\ &= \int_{x}^{x+\delta x} f(x)\,dx \\ &\approx f(x)\delta x. \end{aligned}$$

Thus we can think of $f(x)\delta x$ as the 'amount' of probability in the interval $[x, x + \delta x]$. Note that

$$f(x) = \lim_{\delta x \to 0} \frac{F(x + \delta x) - F(x)}{\delta x} = F'(x).$$

## 17.3 Independence

Recall that the *joint CDF* of two random variables $X$ and $Y$ is defined by

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

and that the joint PMF of two discrete random variables $X$ and $Y$ is

**Definition 17.9**
Two arbitrary random variables $X$ and $Y$ are called *independent* if the events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for all $x, y \in \mathbb{R}$.

- If $X$ and $Y$ are independent, then $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$.

## 17.4 Expectation

Let $X$ be a discrete random variable, let $\{x_1, x_2, \ldots\}$ be its range, and let $f(x)$ be its PMF. The expectation of $X$ is defined by a *sum*:

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

In contrast, the expectation of a continuous random variable is defined by an *integral*:

**Definition 17.10**
Let $X$ be a continuous random variable, and let $f(x)$ be its PDF. The expectation of $X$ is

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

provided this integral exists.

**Theorem 17.11**
Let $X$ be a continuous random variable, let $f(x)$ be its PDF, and let $g : \mathbb{R} \to \mathbb{R}$ be any well-behaved function. Then

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

[*Proof omitted.*]

**Definition 17.12**
Variance, covariance and the correlation coefficient are again defined in terms of expectation:

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

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

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

The following properties of expectation and variance also hold for continuous variables:

**Theorem 17.13**
Let $X$ and $Y$ be continuous random variables, and let $a, b \in \mathbb{R}$.

(1) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.

(2) $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

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

**Proof:** See exercises.

**Example 17.14**
Let $X$ be a continuous random variable with the following CDF:

$$F(x) = \begin{cases} \dfrac{x-a}{b-a} & \text{if } \ a \le x \le b \\ 0 & \text{otherwise.} \end{cases}$$

Find the PDF of $X$, and compute its expected value and variance.

**Solution:** The PDF of $X$ is given by $F'(x)$:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } \ a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

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

$$\mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx = \frac{1}{b-a}\int_a^b x^2\, dx = \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{(b^2+ab+a^2)}{3}.$$

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{(b^2+ab+a^2)}{3} - \frac{(b+a)^2}{4} = \frac{(b^2-2ab+a^2)}{12} = \frac{(b-a)^2}{12}.$$

This is the (continuous) *uniform* distribution on $[a,b]$ (see next lecture).

## 17.5 Exercises

**Exercise 17.1**

1. Let $A_1, A_2, \dots$ be a countable family of sets. Prove the following versions of De Morgan's laws:
   (a) $\left(\bigcup_{i=1}^{\infty} A_i\right)^c = \bigcap_{i=1}^{\infty} A_i^c$.
   (b) $\left(\bigcap_{i=1}^{\infty} A_i\right)^c = \bigcup_{i=1}^{\infty} A_i^c$.
   Hence, or otherwise, show that $\sigma$-fields are closed under countable intersections.

2. A continuous random variable $X$ has PDF
$$f(x) = \begin{cases} cx(1-x) & \text{if } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$
   where $c$ is a constant.
   (a) Find the value of $c$.
   (b) Find the CDF of $X$.
   (c) Show that $\mathbb{P}\left(\frac{1}{3} \le X \le \frac{2}{3}\right) = \frac{13}{27}$.
   (d) Find the expected value and variance of $X$.

3. Let $X$ be a random variable with PDF $f(x)$ and range $[-1, 1]$ (meaning that $f(x) = 0$ for all $|x| > 1$). Find the mean and variance of $X$ in each of the following cases:
   (a) $f(x) = (3/4)(1 - x^2)$.
   (b) $f(x) = (\pi/4)\cos(\pi x/2)$.
   (c) $f(x) = (x + 1)/2$.
   (d) $f(x) = (3/8)(x + 1)^2$.

4. Let $X$ be the amount of petrol (in thousands of litres) sold per week in a certain garage. The PDF of $X$ is as follows:
$$f(x) = \begin{cases} cx^3(9 - x^2) & \text{if } 0 \le x \le 3, \\ 0 & \text{otherwise,} \end{cases}$$
   where $c$ is a constant.
   (a) Find the value of $c$.
   (b) Show that $\mathbb{E}(X) = 72/35$.
   (c) Compute the standard deviation of $X$.

5. Let $X$ be a random variable having the *Cauchy* distribution, whose PDF is as follows:
$$f(x) = \frac{1}{\pi(1 + x^2)} \qquad (x \in \mathbb{R}).$$
   (1) Show that $f(x)$ is indeed a PDF, and sketch the curve $y = f(x)$.
   (2) Find the CDF of $X$.
   (3) Find $\mathbb{P}(-1 \le X \le 1)$.

6. The operational lifetime (in days) of a battery-operated toy can be modelled by a continuous random variable $X$ with the following PDF:
$$f(x) = \begin{cases} \dfrac{cx(50 + x)}{5000} & \text{if } 0 \le x \le 50, \\ c & \text{if } 50 < x \le 100, \\ 0 & \text{otherwise.} \end{cases}$$
   (1) Evaluate $c$, and find the mean operational lifetime of the toy.
   (2) If the purchase price of the toy is £5 and battery costs are 2p per day, find the average cost-per-day.

7. Let $X$ be continuous random variable, and let $a, b \in \mathbb{R}$.
   (a) Show that $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.
   (b) Show that $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$.

# Lecture 18 The Uniform, Exponential and Normal Distributions

## 18.1 The uniform distribution

| | |
|---|---|
| Notation | $X \sim \text{Uniform}[a,b]$ |
| Parameter(s) | $a, b \in \mathbb{R}$, with $a < b$ |
| Range | $[a,b] \subset \mathbb{R}$ |
| PDF | $f(x) = \dfrac{1}{b-a}$ |
| CDF | $F(x) = \dfrac{x-a}{b-a}$ |

**Lemma 18.1**
If $X \sim \text{Uniform}[a,b]$, then

$$\mathbb{E}(X) = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(b-a)^2}{12}.$$

**Proof:** See Lecture 17.

**Example 18.2**
A string of length $L$ is cut at a point chosen uniformly at random along the string. Let $Y$ be the area of the rectangle, two of whose sides are formed by the two pieces of string. Find $\mathbb{E}(Y)$, and the probability that $Y$ exceeds $5L^2/36$.

**Solution:** Let $X$ and $L - X$ denote the lengths of the two pieces. The area of the rectangle is $Y = X(L - X)$. Since $X$ has the uniform distribution on $[0, L]$, it has the following PDF:

$$f(x) = \begin{cases} \dfrac{1}{L} & \text{if } 0 \leq x \leq L, \\ 0 & \text{otherwise.} \end{cases}$$

Let $g(x) = x(L - x)$. The expected value of $Y = g(X)$ is

$$\mathbb{E}(Y) = \int_{-\infty}^{\infty} g(x)f(x)\,dx = \int_0^L x(L-x)\frac{1}{L}\,dx = \frac{1}{L}\left[\frac{Lx^2}{2} - \frac{x^3}{3}\right]_0^L = \frac{L^2}{6}.$$

The probability that $Y = g(X)$ exceeds $5L^2/36$ is

$$\begin{aligned} \mathbb{P}\left(Y > \frac{5L^2}{36}\right) &= \mathbb{P}\left(LX - X^2 > \frac{5L^2}{36}\right) \\ &= \mathbb{P}\left(X^2 - LX + \frac{5L^2}{36} < 0\right) \\ &= \mathbb{P}\left[\left(X - \frac{5L}{6}\right)\left(X - \frac{L}{6}\right) < 0\right] \end{aligned}$$

- The expression $\left(X - \frac{5L}{6}\right)\left(X - \frac{L}{6}\right) < 0$ if and only if exactly one of the factors is negative.
- This occurs if and only if $\frac{L}{6} < X < \frac{5L}{6}$.

Hence,

$$\mathbb{P}\left(Y > \frac{5L^2}{36}\right) = \mathbb{P}\left(\frac{L}{6} < X < \frac{5L}{6}\right) = \int_{L/6}^{5L/6} \frac{1}{L}\,dx = \frac{2}{3}.$$

## 18.2 The negative exponential distribution

| | |
|---|---|
| Notation | $X \sim \text{Exponential}(\lambda)$ |
| Parameter(s) | $\lambda > 0$ (rate) |
| Range | $[0, \infty)$ |
| PDF | $f(x) = \lambda e^{-\lambda x}$ |
| CDF | $F(x) = 1 - e^{-\lambda x}$ |

**Lemma 18.3**
If $X \sim \text{Exponential}(\lambda)$, then

$$\mathbb{E}(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(X) = \frac{1}{\lambda^2}.$$

**Proof:**

$$\mathbb{E}(X) = \int x f(x)\,dx = \lambda \int_0^\infty x e^{-\lambda x}\,dx$$

Integrating by parts,

$$\int_0^\infty x e^{-\lambda x}\,dx = \left[-\frac{x e^{-\lambda x}}{\lambda}\right]_0^\infty + \int_0^\infty \frac{e^{\lambda x}}{\lambda}\,dx = 0 + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}\,dx = \frac{1}{\lambda}\left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty = \frac{1}{\lambda^2}.$$

Hence,

$$\mathbb{E}(X) = \lambda \int_0^\infty x e^{-\lambda x}\,dx = \frac{1}{\lambda}.$$

To compute the variance,

$$\mathbb{E}(X^2) = \int x^2 f(x)\,dx = \lambda \int_0^\infty x^2 e^{-\lambda x}\,dx$$

Integrating by parts,

$$\int_0^\infty x^2 e^{-\lambda x}\,dx = \left[x^2\left(-\frac{e^{-\lambda x}}{\lambda}\right)\right]_0^\infty + \int_0^\infty 2x\left(\frac{e^{-\lambda x}}{\lambda}\right)dx = \frac{2}{\lambda}\int_0^\infty x e^{-\lambda x}\,dx = \frac{2}{\lambda^3}.$$

Hence

$$\mathbb{E}(X^2) = \lambda \int_0^\infty x^2 e^{-\lambda x}\,dx = \frac{2}{\lambda^2},$$

and the variance is

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

**Example 18.4**
The lifetime of a type of electrical fuse has exponential distribution with a mean of 4 years.

(1) Find the probability that a fuse lasts for at least 2 years.

(2) Find the probability that a fuse lasts for at least 6 years, given that it lasts for at least 4 years.

**Solution:** Let $X$ be the lifetime of a fuse. Since the mean is 4, the (rate) parameter of the distribution is $\lambda = 1/4$, so $X$ has the following CDF:

$$F(x) = \begin{cases} 1 - e^{-x/4} & \text{if } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Thus $\mathbb{P}(X \geq 2) = 1 - \mathbb{P}(X \leq 2) = 1 - F(2) = 1 - (1 - e^{-2/4}) = e^{-1/2} = 0.6065$, and

$$\mathbb{P}(X \geq 6 \,|\, X \geq 4) = \frac{\mathbb{P}(X \geq 6 \text{ and } X \geq 4)}{\mathbb{P}(X \geq 4)} = \frac{\mathbb{P}(X \geq 6)}{\mathbb{P}(X \geq 4)} = \frac{e^{-6/4}}{e^{-4/4}} = e^{-1/2} = 0.6065 \quad \text{(again).}$$

This illustrates that the exponential distribution has the so-called 'memoryless' property.

The exponential distribution is related to the Poisson distribution: if the number of 'arrivals' per unit time has Poisson distribution with mean $\lambda$, the inter-arrival times are independent random variable, each having exponential distribution with rate parameter $\lambda$.

**Example 18.5**
The number of customers entering a supermarket per minute has the Poisson(10) distribution.

(1) What is the CDF of the time between the arrival of successive customers?

(2) What is the probability that a customer arrives between three and five seconds after the previous customer?

**Solution:**

(1) Let $\lambda = 10$ denote the mean number of customers arriving per minute, and let $T$ denote the time between successive arrivals. Then $T$ has exponential distribution with rate parameter $\lambda$, so its distribution function is

$$F(t) = \begin{cases} 1 - e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

(2) The probability that a customer arrives between 3 and 5 seconds after the previous customer is

$$\begin{aligned} \mathbb{P}(3/60 \leq T \leq 5/60) &= F(5/60) - F(3/60) \\ &= e^{-3\lambda/60} - e^{-5\lambda/60} \\ &= e^{-1/2} - e^{-5/6} = 0.6065 - 0.4346 = 0.1719. \end{aligned}$$

## 18.3 The normal distribution

| | |
|---|---|
| Notation | $X \sim \mathrm{N}(\mu, \sigma^2)$ |
| Parameter(s) | $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$ |
| Range | $\mathbb{R}$ |
| PDF | $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ |

The normal distribution is based on the *Gaussian integral*:

$$\int_{-\infty}^{\infty} e^{-t^2}\, dt = \sqrt{\pi}$$

For this reason, it is often called the *Gaussian distribution.*

**Definition 18.6**
The *standard normal distribution* is the normal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$.

Let $Z \sim \mathrm{N}(0,1)$ be a standard normal variable.

(1) The distribution function of $Z$ is usually denoted by $\Phi(z) = \mathbb{P}(Z \leq z)$.

(2) The density function of $Z$ is usually denoted by $\phi(z)$:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

**Lemma 18.7**
If $X \sim \text{Normal}(\mu, \sigma^2)$, then $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$.

**Proof:** First consider $Z \sim \mathrm{N}(0,1)$.

$$\begin{aligned}
\mathbb{E}(Z) = \int_{-\infty}^{\infty} z\phi(z)\,dz &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-\frac{1}{2}z^2}\,dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d}{dz}(-e^{-\frac{1}{2}z^2})\,dz \\
&= \frac{1}{\sqrt{2\pi}} \left[-e^{-\frac{1}{2}z^2}\right]_{-\infty}^{\infty} \\
&= 0.
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}(Z^2) = \int_{-\infty}^{\infty} z^2\phi(z)\,dz &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2}\,dz \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z\frac{d}{dz}(-e^{-z^2/2})\,dz \\
&= \frac{1}{\sqrt{2\pi}} \left[-ze^{-z^2/2}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\,dz \qquad \text{(integration by parts)} \\
&= 0 + \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2}\,dt \qquad \text{(using a change of variable: } t = z\sqrt{2}) \\
&= 1,
\end{aligned}$$

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

For the general case, let $X = \mu + \sigma Z$ where $Z \sim N(0,1)$. The CDF of $X$ is

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

and (using the chain rule), the density function of $X$ is

$$\begin{aligned}
f(x) &= \frac{d}{dx}F(x) \\
&= \frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right) \\
&= \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) \\
&= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]
\end{aligned}$$

so $X \sim \mathrm{N}(\mu, \sigma^2)$. By the linearity of expectation,

$$\mathbb{E}(X) = \mathbb{E}(\mu + \sigma Z) = \mu + \sigma\mathbb{E}(Z) = \mu,$$

and by the properties of variance

$$\text{Var}(X) = \text{Var}(\mu + \sigma Z) = \sigma^2\text{Var}(Z) = \sigma^2.$$

**Example 18.8**
The height $X$ in metres of a randomly chosen adult male has a distribution which is approximately normal with mean $\mu$ and variance $\sigma^2$. If 10% of males in the population are taller than 1.8 metres, and 5% are shorter than 1.6 metres, find $\mu$ and $\sigma$.

For $Z \sim \mathrm{N}(0,1)$, you may assume that $\mathbb{P}(Z \leq -1.645) \approx 0.05$ and $\mathbb{P}(Z \leq 1.280) \approx 0.90$.

- The values $z_{0.05} = -1.645$ and $z_{0.90} = 1.280$ are called the *critical points* of the standard normal distribution at levels 0.05 and 0.90 respectively.
- Critical values can be looked up in statistical tables, or computed using statistical software packages.

**Solution:**

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

From the question, we know that $\mathbb{P}(X \leq 1.6) = 0.05$ and $\mathbb{P}(X \leq 1.8) = 0.9$.

- Let $Z = \dfrac{X-\mu}{\sigma}$.

Then $Z \sim \mathrm{N}(0,1)$, with $\mathbb{P}\left(Z \leq \dfrac{1.6-\mu}{\sigma}\right) = 0.05$ and $\mathbb{P}\left(Z \leq \dfrac{1.8-\mu}{\sigma}\right) = 0.90$.

From the critical points given in the question,

$$\frac{1.6-\mu}{\sigma} = -1.645 \quad \text{and} \quad \frac{1.8-\mu}{\sigma} = +1.280.$$

Solving these equations, we obtain

$$\mu = 1.712 \quad \text{and} \quad \sigma = 0.0684.$$

The height of adult males therefore has mean 1.712m and standard deviation 0.068m.

## 18.4 The chi-squared distribution*

During an experiment, we make a sequence of independent observations $X_1, X_2, \ldots, X_n$. An expert tells us that our observations should be normally distributed with mean $\mu$ and variance $\sigma^2$. How can we test to see whether our data fits this model?

Under the proposed model, the standardised variables $\mathbb{Z}_i = (X_i - \mu)/\sigma$ should have the standard normal distribution $\mathrm{N}(0,1)$. To quantify the extent to which our data fits the model, we compute the total squared deviation between our standardised observations $Z_1, Z_2, \ldots, Z_n$ and their hypothetical mean, which is zero. This is called the *chi-squared* statistic:

$$\chi_n^2 = \sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$$

If the expert is correct, $\chi_n^2$ will have the so-called *chi-squared* distribution with $n$ degrees of freedom. Consequently, if the observed value of $\chi_n^2$ is far from the centre of this distribution (i.e. the observed value is in the upper or lower tail of the distribution), we might conclude that the observations are *not* normally distributed.

Let $Z_1, Z_2, \ldots, Z_k$ be independent standard normal random variables, and let

$$X = \sum_{i=1}^k Z_i^2.$$

Then $X$ is said to have *chi-squared distribution* with $k$ degrees of freedom.

| | |
|---|---|
| Notation | $X \sim \text{Chi-squared}(k)$ |
| Parameter(s) | $k \in \mathbb{N}$ (degrees of freedom) |
| Range | $\mathbb{R}^+$ |
| PDF | $f(x) = \dfrac{x^{k/2-1}e^{-x/2}}{2^{k/2}\Gamma(k/2)}$ |

**Lemma 18.9**
If $X \sim \text{Chi-squared}(k)$, then $\mathbb{E}(X) = k$ and $\mathrm{Var}(X) = 2k$.

**Proof:** If $Z_i \sim \mathrm{N}(0,1)$, then

$$\mathbb{E}(Z_i^2) = \mathrm{Var}(Z_i) + \mathbb{E}(Z_i)^2 = 1.$$

Thus by the linearity of expectation,

$$\mathbb{E}(X) = \mathbb{E}(Z_1^2) + \mathbb{E}(Z_2^2) + \ldots + \mathbb{E}(Z_k^2) = k$$

It is easy to show that $\mathbb{E}(Z_i^4) = 3$, so

$$\mathrm{Var}(Z_i^2) = \mathbb{E}(Z_i^4) - \mathbb{E}(Z_i^2)^2 = 2.$$

Hence, because the $Z_i$ are independent,

$$\mathrm{Var}(X) = \mathrm{Var}(Z_1^2) + \mathrm{Var}(Z_2^2) + \ldots + \mathrm{Var}(Z_k^2) = 2k,$$

as required.

## 18.5 Exercises

**Exercise 18.1**

1. The first bus of the day arrives at a certain stop at 7 a.m., and then at 10 minute intervals until the evening. A girl arrives at the stop at a time which is uniformly distributed between 7.15 a.m. and 7.45 a.m. Find the probability that she waits for a bus for
   (1) less than two minutes,
   (2) more than four minutes.

2. Let $X_1$, $X_2$, $X_3$ and $X_4$ be independent and identically distributed continuous random variables.
   (1) Show that $\mathbb{P}(X_1 < X_2 < X_3 < X_4) = \frac{1}{24}$.
   (2) Find $\mathbb{P}(X_1 > X_2 < X_3 < X_4)$.

3. The time $T$ between successive arrivals at a hospital has PDF
$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$
   Exactly 100 inter-arrival times were measured, and the total was found to be 500 hours.
   (1) Estimate the value of $\lambda$.
   (2) Use this estimate of $\lambda$ to estimate the probabilities $\mathbb{P}(T \leq 5)$ and $\mathbb{P}(T \leq 10 \,|\, T > 5)$.

4. A teacher set and marked an examination, and found that the distribution of marks were (approximately) normally distributed with mean 42 and standard deviation 14. The school's policy is to present scaled marks whose distribution is (approximately) normal with mean 50 and standard deviation 15. Find the (linear) transformation that the teacher should apply to the raw marks to accomplish this. What is the transformed mark corresponding to a raw mark of 40?

5. Let $X_1$ and $X_2$ be independent and identically distributed continuous random variables, let $F(x)$ denote their common CDF, and let $f(x)$ denote their common PDF.
   (1) Show that the CDF and PDF of $V = \max\{X_1, X_2\}$ are $F_V(v) = F(v)^2$ and $f_V(v) = 2F(v)f(v)$ respectively.
   (2) Find the CDF and PDF of $U = \min\{X_1, X_2\}$.

6. Let $X : \Omega \to \mathbb{R}$ be a continuous random variable with CDF $F(x)$, and suppose that $X$ is *non-negative* in the sense that $X(\omega) \geq 0$ for all $\omega \in \Omega$. Show that
$$\mathbb{E}(X) = \int_0^\infty \big(1 - F(x)\big)\, dx$$

7. Let $X$ have exponential distribution with (rate) parameter $\lambda$. The CDF of $X$ is as follows:
$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$
   Show that the exponential distribution has the so-called *memoryless* property: for any $t > 0$,
$$\mathbb{P}(X \leq x + t \,|\, X > t) = \mathbb{P}(X \leq x) \quad \text{for all} \quad x \geq 0.$$

# Lecture 19 Quantiles

Let $X$ be a continuous random variable, and let $F$ denote its CDF.

- Recall that CDFs are *increasing* functions: $x < y \;\Rightarrow\; F(x) \leq F(y)$.
- In this lecture, we will assume that $F$ is *strictly increasing*: $x < y \;\Rightarrow\; F(x) < F(y)$.
- This condition ensures that the inverse function $F^{-1} : [0,1] \to \mathbb{R}$ exists.
- The ideas presented here can be extended to discrete distributions, with some modifications.

## 19.1 Median

The mean of a distribution can be thought of as its *location*.

- The expected *squared deviation* of $X$ from the point $c$ is given by $\mathbb{E}\big[(X-c)^2\big]$.
- The mean of $X$ is the value of $c$ that minimises the expected squared deviaiton.

Another location parameter is provided by a *median* of a distribution.

- The expected *absolute deviation* of $X$ from the point $c$ is given by $\mathbb{E}\big[|X-c|\big]$.
- The median of $X$ is a value of $c$ that minimises the expected absolute deviaiton.

**Definition 19.1**
A median of a distribution is a value $\eta \in \mathbb{R}$ such that

$$F(\eta) = \frac{1}{2}.$$

**Remark 19.2**
If $F$ is continuous and strictly increasing, the median is uniquely defined.

## 19.2 Quantiles

The $q$-quantiles of a distribution divide the real line into $q$ 'equal' parts:

**Definition 19.3**
For $q \in \mathbb{N}$, the $q$-quantiles of a distribution are values $x_1 < x_2 < \ldots < x_{q-1}$ such that

$$F(x_k) = \frac{k}{q} \qquad \text{for} \quad k = 1, 2, \ldots, q-1$$

**Remark 19.4**
If $F$ is continuous and strictly increasing, $q$-quantiles are uniquely defined.

In particular,

- The 2-quantile is the *median* of $F$.
- The 4-quantiles are the called the *quartiles* of $F$.
- The 100-quantiles are called the *percentiles* (or *percentage points*) of $F$.

### 19.2.1 Quartiles

**Definition 19.5**

(1) The *lower quartile* of $F$ is a number $x_L$ such that $F(x_L) = \dfrac{1}{4}$.

(2) The *upper quartile* of $F$ is a number $x_U$ such that $F(x_U) = \dfrac{3}{4}$.

(3) The *inter-quartile range* of $F$ is the difference $x_U - x_L$ between its upper and lower quartiles.

**Remark 19.6**

- The median quantifies the *location* of a distribution.
- The inter-quartile range quantifies the *size* or *scale* of a distribution.
- The lower quartile is the median of the lower half of the distribution.
- The upper quartile is the median of the upper half of the distribution.

**Example 19.7**
A continuous random variable $X$ has the following PDF: $f(x) = \begin{cases} 3x^2/8 & \text{if } 0 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$

Find the median and the interquartile range of this distribution.

**Solution:**

(1) The median $\eta$ satisfies

$$F(\eta) = \frac{3}{8}\int_0^{\eta} x^2\,dx = \frac{1}{2} \quad\Rightarrow\quad \frac{3}{8}\times\frac{\eta^3}{3} = \frac{1}{2} \quad\Rightarrow\quad \eta = 4^{1/3} = 1.5874.$$

(2) The lower quartile $x_L$ satisfies

$$F(x_L) = \frac{3}{8}\int_0^{x_L} x^2\,dx = \frac{1}{4} \quad\Rightarrow\quad \frac{3}{8}\times\frac{x_L^3}{3} = \frac{1}{4} \quad\Rightarrow\quad x_U = 2^{1/3} = 1.2599.$$

The upper quartile $x_U$ satisfies

$$F(x_U) = \frac{3}{8}\int_0^{x_U} x^2\,dx = \frac{3}{4} \quad\Rightarrow\quad \frac{3}{8}\times\frac{x_U^3}{3} = \frac{3}{4} \quad\Rightarrow\quad x_L = 6^{1/3} = 1.8171.$$

Hence the inter-quartile range is equal to $x_U - x_L = 1.8171 - 1.2599 = 0.5572$.

### 19.2.2 Percentiles

**Definition 19.8**
The $k$th percentile of $F$ is a value $x_k$ such that

$$F(x_k) = \frac{k}{100} \qquad \text{for} \quad k = 1, 2, \ldots, 99$$

In particular,

- The 25th percentile is the lower quartile.
- The 50th percentile is the median.
- The 75th percentile is the upper quartile.

## 19.3 The quantile function

**Definition 19.9**
Let $F$ be a continuous and stricty increasing CDF. The *quantile function* is the inverse of $F$:

$$\begin{array}{rccc} Q: & [0,1] & \longrightarrow & \mathbb{R}[0,1] \\ & p & \mapsto & F^{-1}(p) \end{array}$$

For any $p \in [0,1]$, $Q(p)$ is the value of $x$ for which $F(x) = p$.

(1) $Q(p)$ called the *critical point* of the distribution at level $p$, and is often denoted by $x_p$.

For example, the critical points at levels $p = 0.05$ and $p = 0.95$ satisfy

$$\begin{aligned} F(x_{0.05}) &= \mathbb{P}(X \leq x_{0.05}) &= 0.05. \qquad &\text{(lower tail)} \\ 1 - F(x_{0.95}) &= \mathbb{P}(X \geq x_{0.95}) &= 0.05. \qquad &\text{(upper tail)} \end{aligned}$$

**Remark 19.10**

- The event $\{x_{0.05} < X < x_{0.95}\}$ occurs with probability 0.9: this is a *typical* event.
- The event $\{X \leq x_{0.05}\} \cup \{X \geq x_{0.95}\}$ occurs with probability 0.1" this is an *extreme* event.
- Confidence intervals and statistical hypothesis tests are constructed on this basis.

**Example 19.11**
Let $X$ have (negative) exponential distribution with rate parameter $\lambda$.

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

Derive an explicit expression for the quantile function of $F$.

**Solution:** Let $x_p = Q(p)$ where $p \in [0,1]$. Then

$$\begin{aligned} p = F(x_p) &\Rightarrow p = 1 - e^{-\lambda x_p} \\ &\Rightarrow e^{-\lambda x_p} = 1 - p \\ &\Rightarrow -\lambda x_p = \log(1-p) \\ &\Rightarrow x_p = -\frac{\log(1-p)}{\lambda}. \end{aligned}$$

In particular,

$$x_{0.05} = \frac{\log(20/19)}{\lambda}, \quad x_{0.25} = \frac{\log(4/3)}{\lambda}, \quad x_{0.50} = \frac{\log(2)}{\lambda}, \quad x_{0.75} = \frac{\log(4)}{\lambda}, \quad x_{0.95} = \frac{\log(20)}{\lambda}.$$

**Remark 19.12 (Statistical Tables)**

- For many important distributions, it is not possible to derive explicit (closed-form) expressions for their quantile functions.
- In such cases, critical points must be estimated using numerical approximation techniques.
- For a number of standard distributions, tables of such critical points are available.
- These tables list critical points for various values of $p \in [0, 1]$.
- In recent years, statistical tables have been supplanted by statistical software packages.

## 19.4 Exercises

**Exercise 19.1**

1. Let $X$ be a continuous random variable.

   - A *mode* of $X$ is a number $a \in \mathbb{R}$ such that $f(a) \geq f(x)$ for all $x \in \mathbb{R}$.

   Find a mode and a median of a random variable having the following PDF:

$$f(x) = \begin{cases} \dfrac{2}{3}\cos\left(x - \dfrac{\pi}{6}\right) & \text{if } \ 0 \leq x \leq \dfrac{2\pi}{3}, \\ 0 & \text{otherwise.} \end{cases}$$

2. A continuous random variable $X$ has PDF

$$f(x) = \begin{cases} ce^{-x} & \text{if } \ x \geq 0, \\ 0 & \text{otherwise,} \end{cases}$$

   where $c$ is a constant.

   (1) Find the value of $c$.
   (2) Find the CDF of $X$.
   (3) Find the median of the distribution.

3. Suppose that $X$ has the exponential distribution with (rate) parameter $\lambda > 0$. The PDF of $X$ is

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

   Find the median and the inter-quartile range of this distribution.

# Lecture 20 Normal Approximation

## 20.1 Approximation of discrete distributions by continuous distributions

Recall the PMFs of the binomial and Poisson distributions:

- If $X \sim \text{Binomial}(n, p)$ then

$$\mathbb{P}(X = k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n.$$

- If $X \sim \text{Poisson}(\lambda)$ then

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

If $k$ is large, computing $k!$ can take a long time. To avoid this, we make use of the fact that, under certain conditions, the binomial and Poisson distributions can both be approximated by the normal distribution.

Binomial distribution:

- If $X \sim \text{Binomial}(n, p)$ then $\mathbb{E}(X) = np$ and $\text{Var}(X) = np(1-p)$.
- If $np$ and $n(1-p)$ are both sufficiently large,

$$X \sim \text{N}\big(np, np(1-p)\big) \quad \text{approx.}$$

Poisson distribution:

- If $X \sim \text{Poisson}(\lambda)$ then $\mathbb{E}(X) = \lambda$ and $\text{Var}(X) = \lambda$.
- If $\lambda$ is sufficiently large,

$$X \sim \text{N}(\lambda, \lambda) \quad \text{approx.}$$

### 20.1.1 The continuity correction

**Definition 20.1 (The continuity correction)**
Let $X$ be a discrete random variable, taking values in the set $\{0, \pm 1, \pm 2, \ldots\}$. If the distribution of a continuous random variable $Y$ is taken as an approximation to the distribution of $X$, we set

$$\mathbb{P}(X = k) = \mathbb{P}\left(k - \frac{1}{2} < Y < k + \frac{1}{2}\right).$$

In particular,

- $\mathbb{P}(X < k) = \mathbb{P}(Y \leq k - 1/2)$ and $\mathbb{P}(X \leq k) = \mathbb{P}(Y \leq k + 1/2)$,
- $\mathbb{P}(X > k) = \mathbb{P}(Y \geq k + 1/2)$ and $\mathbb{P}(X \geq k) = \mathbb{P}(Y \geq k - 1/2)$,
- $\mathbb{P}(X = k) = \mathbb{P}(k - 1/2 < Y \leq k + 1/2)$.

## 20.2 Normal approximation of the binomial distribution

**Theorem 20.2**
If $X \sim \text{Binomial}(n,p)$, then

$$\mathbb{P}(X=k) \to \frac{1}{\sqrt{2\pi np(1-p)}} \int_{k-1/2}^{k+1/2} \exp\left[-\frac{1}{2}\left(\frac{x-np}{\sqrt{np(1-p)}}\right)^2\right] dx \quad \text{as} \quad n \to \infty.$$

[*Proof omitted.*]

**Example 20.3**
A fair coin is tossed 500 times. What is the probability that 270 heads are obtained?

**Solution:** Let $X$ be the number of heads obtained. Then $X \sim \text{Binomial}(500, 270)$, so

$$\mathbb{P}(X > 270) = \sum_{k=271}^{n} \binom{500}{k} \left(\frac{1}{2}\right)^{500}$$

This is difficult to compute (the exact value is 0.0333).
Using the normal approximation, the associated normal variable is $Y \sim \text{N}(250, 125)$, so

$$\begin{aligned}
\mathbb{P}(X \geq 271) &\approx \mathbb{P}(Y > 270.5) \\
&= 1 - \mathbb{P}(Y \leq 270.5) \\
&= 1 - \mathbb{P}\left(Z \leq \frac{270.5-250}{\sqrt{125}}\right) \qquad \text{where} \quad Z \sim \text{N}(0,1), \\
&= 1 - \Phi(1.8336) \\
&= 0.0334 \qquad \text{(from tables).}
\end{aligned}$$

## 20.3 Normal approximation of the Poisson distribution

**Theorem 20.4**
If $X \sim \text{Poisson}(\lambda)$, then

$$\mathbb{P}(X=k) \to \frac{1}{\sqrt{2\pi\lambda}} \int_{k-1/2}^{k+1/2} \exp\left[-\frac{1}{2}\left(\frac{x-\lambda}{\sqrt{\lambda}}\right)^2\right] dx \quad \text{as} \quad \lambda \to \infty.$$

[*Proof omitted.*]

**Example 20.5**
A newsagent knows from past experience that the weekly demand for a certain magazine has Poisson distribution with mean 20. How many copies of the magazine should the newsagent stock in order to satisfy the weekly demand with probability 0.95?

**Solution:** Suppose the newsagent stocks $N$ copies of the magazine. Let $X \sim \text{Poisson}(20)$ represent the weekly demand. We need that

$$\mathbb{P}(\text{demand satisfied}) = \mathbb{P}(X \leq N) = \sum_{k=0}^{N} \frac{20^k}{k!} e^{-20} = 0.95.$$

Using the normal approximation, the associated normal variable is $Y \sim \mathrm{N}(20,20)$, so

$$\begin{aligned}
\mathbb{P}(X \leq N) = 0.95 \quad &\Rightarrow \quad \mathbb{P}(Y \leq N + 0.5) = 0.95 \\
&\Rightarrow \quad \mathbb{P}\left(Z \leq \frac{N + 0.5 - 20}{\sqrt{20}}\right) = 0.95 \qquad \text{where} \quad Z \sim \mathrm{N}(0,1), \\
&\Rightarrow \quad \Phi\left(\frac{N - 19.5}{\sqrt{20}}\right) = 0.95 \\
&\Rightarrow \quad \frac{N - 19.5}{\sqrt{20}} = 1.645 \qquad \text{(from tables: } z_{0.95} = 1.645\text{)}, \\
&\Rightarrow \quad N = 27 \qquad \text{(to the nearest integer)}.
\end{aligned}$$

Thus the newsagent should stock 27 copies of the magazine.

## 20.4 Exercises

**Exercise 20.1**

1. 90% of all items produced by a manufacturing process are satisfactory. Find an approximation for the probability that a sample of 250 items contains exactly 25 defective items.
2. A casino buys a new die and rolls it 600 times. Let $N$ denote the number of times a six occurs.
   (1) Find the probability that $N$ is between 90 and 100, assuming that the die is fair.
   (2) Find the value $c$ for which $\mathbb{P}(100 - c \leq N \leq 100 + c) = 0.95$, assuming that the die is fair.
   (3) What might the casino conclude if a six occurred $N = 120$ times?