Raymond W. Yeung

Probability for Engineers
Lecture Notes for ENGG 2470A
(Spring 2020)

May 9, 2020

Contents

1

Preliminaries

1.1 Logic

Logic is a *language*. In logic, a *statement* can only be *True* (T) or *False* (F).

Example 1.1.

1. '$3 > 2$' — True
2. '$3 < 2$' — False
3. '$\forall x, x^2 \geq 0$' — True
4. 'Tom is a student of CUHK' — True/False
5. '$x > 1$, where $x \in \mathbb{R}$' — True/False
6. '$x > 1$, where $x \in \mathbb{C}$' — ??
7. 'Tom is handsome' — ??
8. 'The car is red' — ??

Remark In the *real world*, most statements are neither True or False. In other words, *logic may not be applicable!*

A *compound statement* is a statement obtained by combining more than one statement.

AND (Conjunction), $\wedge$

X	Y	$X \wedge Y$
T	T	T
T	F	F
F	T	F
F	F	F

Such a table is called a *truth table*. The only condition for '$X \wedge Y$' to be *True* is both X and Y are *True*. Equivalently, the condition for '$X \wedge Y$' to be *False* is at least one of X and Y is *False*.

OR (Disjuction), $\vee$

X	Y	$X \vee Y$
T	T	T
T	F	T
F	T	T
F	F	F

The only condition for '$X \vee Y$' to be *False* is both X and Y are *False*. Equivalently, the condition for '$X \vee Y$' to be *True* is at least one of X and Y is *True*.

Note OR is not the same as XOR in circuit theory.

NOT (Negation), $\sim$

X	$\sim X$
T	F
F	T

Note $\sim (\sim X) = X$.

Implication (Conditional), $\rightarrow$, $\Rightarrow$

X	Y	$X \rightarrow Y$
T	T	T
T	F	F
F	T	T
F	F	T

Note '$X \rightarrow Y$' can be read as one of the following:

1. 'if X then Y'
2. 'X implies Y
3. 'Y if X'
4. X is a sufficient condition of Y
5. 'X only if Y'
6. Y is a necessary condition of X

The only condition for '$X \rightarrow Y$' to be *False* is when X is *True* and Y is *False*. As we will see, this is the right formalism for Implication.

Example 1.2.

Tom: Will you betray me?
Jack: Only if you do.

What Jack says means

'Jack betrays Tom' $\rightarrow$ 'Tom betrays Jack'

Example 1.3. Consider the following database:

Student	Gender	College	Sports
Tom	M	CC	strong
Jack	M	NA	weak
Lillian	F	CC	strong
Jessica	F	Shaw	weak

According to the truth table for implication, the following can be verified:

'Jack belong to NA'	$\rightarrow$	'Jack is strong in sports'	*False*
'Lillian is male'	$\rightarrow$	'Lillian belongs to CC'	*True*
'Lillian is male'	$\rightarrow$	'Lillian is weak in sports'	*True*

These implications by themselves may not make much sense. Now consider the statement

'Student belongs to CC' $\rightarrow$ 'Student is male'.

Here, since the student is not specified, we interpret that the statement is for all students. Then in order to establish that the statement is *True* or *False*, we have to check it for every student.

Student	Student belongs to CC	$\rightarrow$	Student is male
Tom	T	T	T
Jack	F	T	T
Lillian	*T*	*F*	*F*
Jessica	F	T	T

From the above, we see that statement is *False* for the case that the student is Lillian (and *True* for all other cases), and so the statement is *False*. The case that makes the statement *False* is called a counterexample.

In fact, we only need to check the cases for Tom and Lillian, as these are the only 2 students belonging to CC.

Example 1.4. Consider the statement '$x \geq 2 \Rightarrow x \geq 0$'. Here the value of x is not specified, and we interpret that the statement is for all real number x. We can establish that this statement is *True* as follows.

for	$x \geq 2$	$\rightarrow$	$x \geq 0$
$x < 0$	F	T	F
$0 \leq x < 2$	F	T	T
$x \geq 2$	T	T	T

Again, we only need to check the case $x \geq 2$ because this is the only two case that '$x \geq 2$' is *True*.

Example 1.5. Consider the statement '$x \geq 0 \Rightarrow x \geq 2$'. We can prove that this statement is *False* by considering the following counterexample:

for	$x \geq 0$	$\rightarrow$	$x \geq 2$
$x = 1$	T	F	F

Equivalence, $\leftrightarrow$, $\Leftrightarrow$

$$\text{`}X \leftrightarrow Y\text{'} \text{ means } \underbrace{\begin{array}{l}\text{`}X \rightarrow Y\text{'} \\ (X \text{ only if } Y)\end{array} \quad \wedge \quad \begin{array}{l}\text{`}X \leftarrow Y\text{'} \\ (X \text{ if } Y)\end{array}}_{\text{`}X \text{ if and only if } Y\text{'}}$$

X	Y	$X \rightarrow Y$	$X \leftarrow Y$	$X \leftrightarrow Y$
T	T	T	T	T
T	F	F	T	F
F	T	T	F	F
F	F	T	T	T

Note '$X \leftrightarrow Y$ is *True* if either both X and Y are *True* or both X and Y are *False*.

Contrapositive and Converse

Consider '$X \to Y$'.
'$\sim Y \to \sim X$' is called the *contrapositive* of '$X \to Y$'.
'$Y \to X$' is called the *converse* of '$X \to Y$'.

Proposition 1.6.

1. '$X \to Y$' $\leftrightarrow$ '$\sim Y \to \sim X$' *(contrapositive)*
2. '$X \to Y$' $\not\leftrightarrow$ '$Y \to X$' *(converse)*

Proof

X	Y	$X \to Y$	$Y \to X$	$\sim Y$	$\to$	$\sim X$
T	T	T	T	F	T	F
T	F	F	T	T	F	F
F	T	T	F	F	T	T
F	F	T	T	T	T	T

Example 1.7.

1. 'If a car is electric, then the car has no exhaustion.' (original statement)
   $\leftrightarrow$
   'If a car has exhaustion, then the car is not electric.' (contrapositive)
2. 'If a car is electric, then the car has no exhaustion.' (original statement)
   $\not\leftrightarrow$
   'If a car has no exhaustion, then the car is electric.' (converse)

De Morgan's Law

1. '$\sim (X \vee Y)$' $\leftrightarrow$ '$(\sim X) \wedge (\sim Y)$'
2. '$\sim (X \wedge Y)$' $\leftrightarrow$ '$(\sim X) \vee (\sim Y)$'

Remark #1 and #2 are in fact equivalent. We now show that #2 follows from #1. In #1, replace X by $\sim X$ and Y by $\sim Y$. Then we have

$$\sim ((\sim X) \vee (\sim Y)) \;\leftrightarrow\; (\sim (\sim X)) \wedge (\sim (\sim Y))$$

or

$$\sim ((\sim X) \vee (\sim Y)) \;\leftrightarrow\; X \wedge Y.$$

Taking negation on both sides, we have

$$\sim\sim ((\sim X) \vee (\sim Y)) \;\leftrightarrow \sim (X \wedge Y)$$

or

$$(\sim X) \vee (\sim Y) \;\leftrightarrow \sim (X \wedge Y)$$

i.e., #2.

De Morgan's Law can readily be proved by a truth table. We leave the details as an exercise.

For All, There Exists, and Negation

'For all' ($\forall$) and 'there exists' ($\exists$) are two commonly used quantifiers in logic. We first look at an example.

Example 1.8. In Example 1.4, we have discussed the statement

$$\forall \text{ real number } x, \text{`}x \geq 2 \Rightarrow x \geq 0\text{'} \tag{1.1}$$

Here the (sub-)statement '$x \geq 2 \Rightarrow x \geq 0$' depends on x. Denote it by $Y(x)$. Formally, this is called a *statement function* with parameter x. Since x is not specified, we cannot determine whether $Y(x)$ is *True* or not. But if it is *True* for all real number $x \in \mathbb{R}$, we say that the statement (1.1) is *True*.

There is a close relation between 'for all' and 'there exists'. This is illustrated by the following day-to-day example.

Example 1.9. Consider the statement

$$\text{'All students in ENGG 2460A are female.'} \tag{1.2}$$

Evidently, the above statement is *False* if and only if the following statement is *True*:

$$\text{'At least one student in ENGG 2460A is not female.'} \tag{1.3}$$

In other words, (1.3) is the negation of (1.2). Expressing (1.2) in terms of 'for all', we have

'For all student x in ENGG 2460A, x is female.'

Likewise, expressing (1.3) in terms of 'there exists', we have

'There exists a student x in ENGG 2460A such that x is not female.'

The next proposition gives a precise relation between 'for all' and 'there exists'. To facilitate our discussion, we consider a statement of the form

$$\forall x, Y(x) \tag{1.4}$$

with the understanding that we know implicitly what x can be. For example, x can be any real number, any integer, any positive integer, or any student in ENGG 2460A as in the last example.

Proposition 1.10. $\sim (\forall x, Y(x)) \leftrightarrow \exists x \text{ s.t. } \sim Y(x)$ *('s.t.' is read as 'such that').*

Proof

LHS $\rightarrow$ RHS

We prove this part by proving the contrapositive, i.e., $\sim$ RHS $\rightarrow \sim$ LHS. Assume that RHS is *False*, i.e., there does not exist x such that $\sim Y(x)$ is *True*. Then for all x, $Y(x)$ is *True*, or LHS is *False*.

RHS $\rightarrow$ LHS

Assume that RHS is *True*. Then '$\forall x, Y(x)$' is *False*, or LHS is *True*.

From this proposition, we see that in order to prove that '$\forall x, Y(x)$' is *False*, we only need to come up with an x such that $Y(x)$ is *False*. We say that such an x gives a *counterexample*.

Proposition 1.10 in fact implies *De Morgan's Law*. Let an index i takes the values $1, 2, \ldots, n$. Consider the statement '$\forall i, X_i$'. By Proposition 1.10,

$$\sim (\forall i, X_i) \;\leftrightarrow\; \exists i \; s.t. \sim X_i \tag{1.5}$$

In the above, it is readily seen that $(\forall i, X_i)$ is equivalent to $X_1 \wedge X_2 \wedge \cdots \wedge X_n$, because the latter is *True* if and only if X_i is *True* for all i. On the other hand, RHS is equivalent to $(\sim X_1) \vee (\sim X_2) \vee \cdots \vee (\sim X_n)$, because the latter is *True* if and only if $(\sim X_i)$ is *True* for at least one i. It then follows that

$$\sim (X_1 \wedge X_2 \wedge \cdots \wedge X_n) \;\leftrightarrow\; (\sim X_1) \vee (\sim X_2) \vee \cdots \vee (\sim X_n),$$

which is a general form of De Morgan's Law.

1.2 Set Theory

- Set theory is a mathematical language *built on logic*.
- A set is a collection of objects, called *elements*.
- The notion of a set is "abstract," it is only a conceptual construction.

Examples of sets:

1. $A = \{\text{Tom, Jack, Lillian, Jessica}\}$
2. $B = \{1, 2, 3, 4, 5\}$
3. $C = \{x : x \geq 0\}$

- The size of a set A, called the *cardinality* of A, is the number of elements in A, denoted by $|A|$. Thus in the above, $|A| = 4$, $|B| = 5$, and $|C| = \infty$.

Two Important Sets

Universal set Ω the set that contains all objects of interest in a problem/discussion
Empty set $\emptyset$ the set that contains no element.

Three Basic Set Operations

Let A and B be sets.

1. Union
$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$
2. Intersection
$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$
3. Complement
$$A^c = \{x : x \notin A\} = \{x : \sim \text{`}x \in A'\}$$

Other Notions in Set Theory

Set Difference $A - B = \{x : x \in A \text{ and } x \notin B\}$

Remark

1. $A - B = A \cap B^c$
2. $A^c = \Omega - A = \{x : x \in \Omega \text{ and } x \notin A\}$.

Set Inclusion $A \subset B$ means '$x \in A$' $\Rightarrow$ '$x \in B$'.

Proposition 1.11. $A \subset B \Leftrightarrow B^c \subset A^c$.

Proof '$B^c \subset A^c$' means '$x \in B^c$' $\Rightarrow$ '$x \in A^c$', or '$x \notin B$' $\Rightarrow$ '$x \notin A$'. Take contrapositive to get

$$\sim \text{`}x \notin A\text{'} \Rightarrow \sim \text{`}x \notin B\text{'}$$
$$\text{`}x \in A\text{'} \Rightarrow \text{`}x \in B\text{'}$$

i.e., $A \subset B$.

Set Equivalence $A = B$ means one of the following:

1. $\{x : x \in A\} = \{x : x \in B\}$
2. '$x \in A$' $\Leftrightarrow$ '$x \in B$'
3. '$x \in A \Rightarrow x \in B$' $\wedge$ '$x \in B \Rightarrow x \in A$'
4. $A \subset B$ and $B \subset A$

Disjoint Sets

- A and B are said to be disjoint if $A \cap B = \emptyset$.
- Sets $A_i, i = 1, 2, \ldots, n$ are disjoint if $\forall 1 \leq i < j \leq n$,

$$A_i \cap A_j = \emptyset.$$

Venn Diagrams

These are diagrams for illustrating the relations among sets.

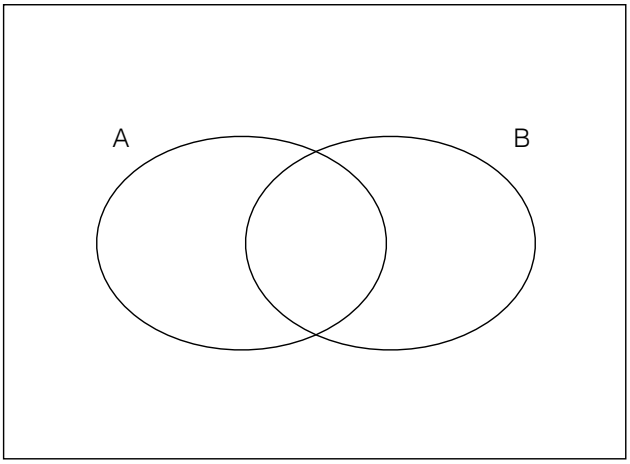


Fig. 1.1: Venn diagram for 2 sets

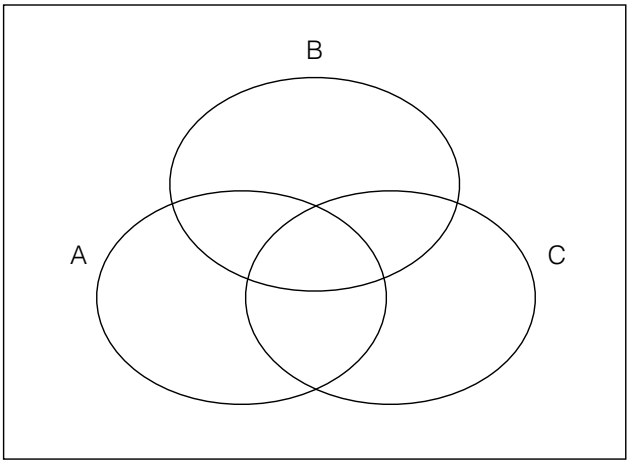


Fig. 1.2: Venn diagram for 3 sets

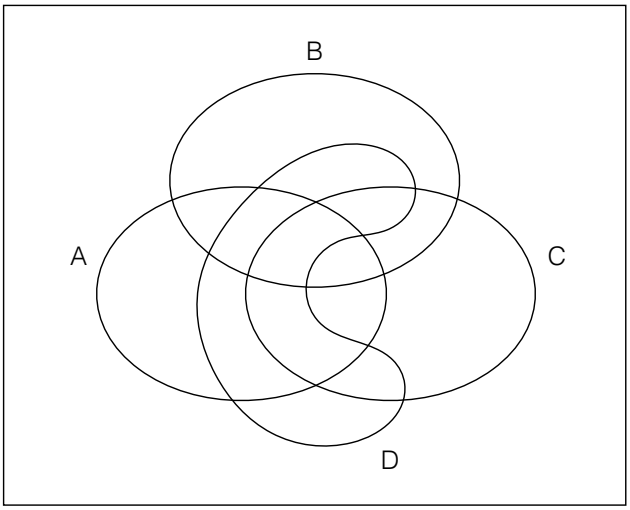


Fig. 1.3: Venn diagram for 4 sets

Venn diagram Illustration of various set-theoretic notions

Check out "Venn diagram" on Wikipedia.

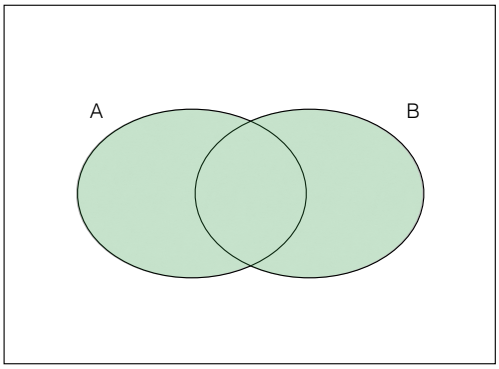


Fig. 1.4: $A \cup B$

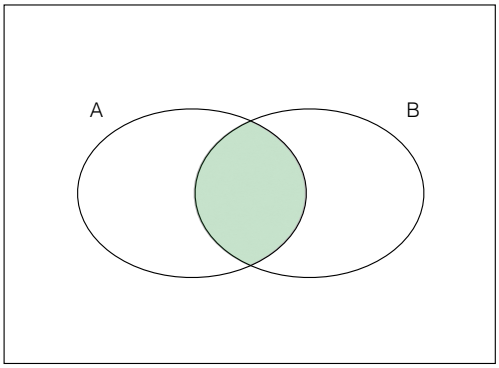


Fig. 1.5: $A \cap B$

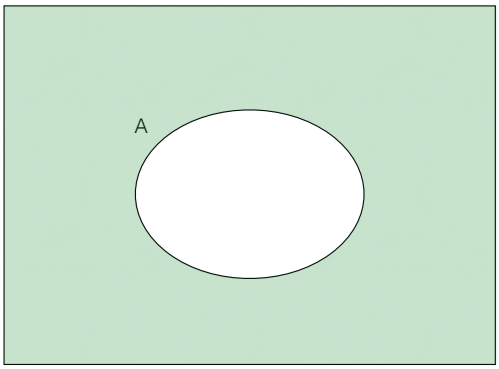


Fig. 1.6: A^c

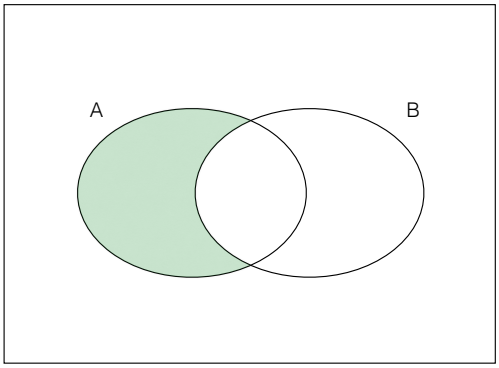


Fig. 1.7: $A - B$

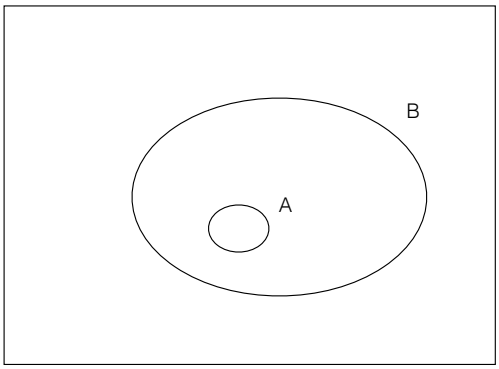


Fig. 1.8: $A \subset B$. Verify that '$A \subset B$' $\leftrightarrow$ '$B^c \subset A^c$'.

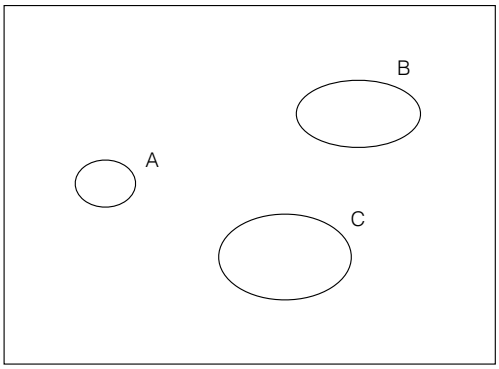


Fig. 1.9: A, B, and C are disjoint

Power Set

As discussed, a set is a conceptual construction. Thus we are free to construct sets as we want. For example, for a set A, we can construct a set

$$B = \{A\} = \{\{a\}\},$$

i.e., B is the set that contains the set A. Note that $A \neq B$, $A \in B$, and A is not a subset of B. For the latter, we normally would not write $A \not\subset B$ because the two sets A and B are "not at the same level."

There is a set we can constructed from any set A this is of special importance. For a set A, the *power set* of A, denoted by $\wp(A)$ (also commonly denoted by 2^A), is the set of all subsets of A. For example, if $A = \{a, b, c\}$, then

$$\wp(A) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

Note that $|\wp(A)| = 2^{|A|}$.

Proposition 1.12.

1. $(A^c)^c = A$
2. $\Omega^c = \emptyset$
3. $\emptyset^c = \Omega$
4. $A \cup \Omega = \Omega$
5. $A \cap \Omega = A$
6. $A \cap \emptyset = \emptyset$
7. $A \cup \emptyset = A$
8. $A \cup A^c = \Omega$
9. $A \cap A^c = \emptyset$

Proof

1.

$$\begin{aligned} A^c &= \{x :\sim \text{`}x \in A\text{`}\} \\ (A^c)^c &= \{x :\sim\sim \text{`}x \in A\text{`}\} \\ &= \{x : x \in A\} \\ &= A. \end{aligned}$$

2. $\Omega^c = \{x : x \notin \Omega\} = \emptyset$.
3. Take the complement of both sides of $\Omega^c = \emptyset$ (#2) to get $(\Omega^c)^c = \emptyset^c$, or $\Omega = \emptyset^c$.

The proofs of the remaining parts are left as an exercise.

De Morgan's Law

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

Remark #1 and #2 are in fact equivalent. We now show that #2 follows from #1. In #1, replace A by A^c ad replace B by B^c. Then we have

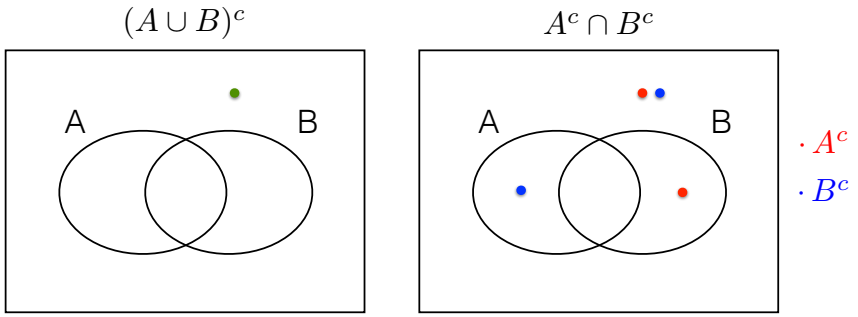
$$\begin{aligned} (A^c \cup B^c)^c &= (A^c)^c \cap (B^c)^c \\ &= A \cap B \end{aligned}$$

Take complement to get

$$\begin{aligned} [(A^c \cup B^c)^c]^c &= (A \cap B)^c \\ A^c \cup B^c &= (A \cap B)^c \end{aligned}$$

i.e., #2. We leave it as an exercise to show that #1 also follows from #2.

Verification of De Morgan's Law #1



De Morgan's law can formally be proved by logic:

$$(A \cup B)^c = \{x : \sim(x \in A \vee x \in B)\}$$
$$A^c \cap B^c = \{x : x \notin A \wedge x \notin B\}$$

The equivalence of the two underlined statements above corresponds to *De Morgan's law for logic* in Section 1.1, which asserts that

1. '$\sim (X \vee Y)$' $\leftrightarrow$ '$(\sim X) \wedge (\sim Y)$'
2. '$\sim (X \wedge Y)$' $\leftrightarrow$ '$(\sim X) \vee (\sim Y)$'

Distribution Laws

1. $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
2. $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

Note For real and complex number arithmetic,

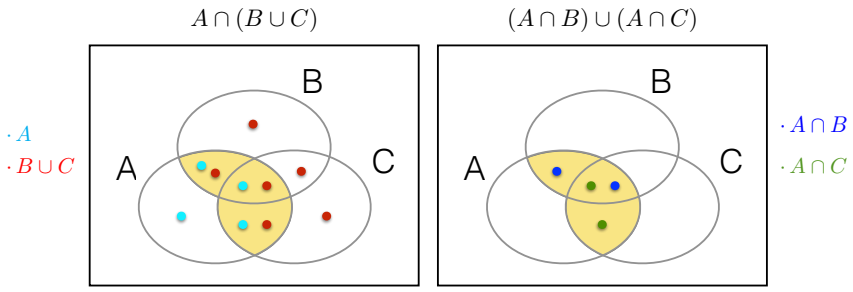
$$a \cdot (b + c) \quad = \quad (a \cdot b) + (a \cdot c)$$

but

$$a + (b \cdot c) \quad \neq \quad (a + b) \cdot (a + c)$$

In other words, for real and imaginary number arithmetic, multiplication distributes over addition but addition does not distribute over multiplication.

Verification of Distribution Law #1



Proposition 1.13. *$A \subset B$ if and only if $A - B = \emptyset$.*

Proof Consider

$$\begin{aligned} A &= A \cap \Omega \\ &= A \cap (B \cup B^c) \\ &= (A \cap B) \cup (A \cap B^c) \\ &= (A \cap B) \cup (A - B) \end{aligned}$$

Note that $A \cap B$ and $A - B$ are disjoint, and $A \subset B$ is equivalent to $A = A \cap B$. Therefore, in order for $A \subset B$, $A \cap B^c$ must be empty, or $A - B = \emptyset$ (cf. Fig. 1.7).

An Application in Philosophy

"White horse is not horse" is a famous paradox in Chinese philosophy. It is a paradox because "white horse is horse" appears to be wrong because there are horses that are not white. Therefore, "horse" means more than "white horse" and so the two cannot be the same. On the other hand, "white horse is not horse" appears to be wrong, too, because a white horse is obviously a horse. Hence, we cannot decide whether 'white horse is horse" or "white horse is not horse."

To resolve the paradox, we need the following set-theoretic setup. Let

$$\begin{aligned} W &= \text{the set of all white horses} \\ &= \{\text{horse} : \text{the horse is white}\} \\ H &= \text{the set of all horses} \end{aligned}$$

When we say "white horse *is* horse", it means

$$x \in W \Rightarrow x \in H, \text{ or } W \subset H.$$

When we say "white horse *is not* horse", it means

$$W \neq H.$$

Remark In natural language, there is an ambiguity in the use of 'is' and 'is not', because 'is' can mean "*a subset of*" or "*equivalent to*". Set theory, being a precise mathematical language, avoids such ambiguity.

Set-Theoretic Interpretation of Mathematical Statements

Some examples of one-to-one correspondence between mathematical statements and set-theoretic statements:

Mathematical Statement	**Set-theoretic Statement**
$x \geq 6 \Rightarrow x \geq 3$	$\{x : x \geq 6\} \subset \{x : x \geq 3\}$
$(x-2)(x-3) = 0$ iff $x = 2$ *or* $x = 3$	$\{x : (x-2)(x+3) = 0\}$ $= \{2\} \cup \{3\} = \{2, 3\}$

What is a Meaningful Mathematical Result?

In mathematics, we are very often interested in a certain set A. For example,

$$A = \{x : x^2 - 5x + 6 = 0\},$$

the solution set of the quadratic equation $x^2 - 5x + 6 = 0$. Upon solving this quadratic equation, we find that the solution set is $\{2, 3\}$, and we call this set B. In other words, we have established that $A = B$.

We say that set B is a *characterization* of set A. The statement $A = B$ is a meaningful mathematical result because set B has a simpler description compared with set A.

Now consider another set

$$B' = \{x : x^3 - 4x^2 + x + 6 = 0 \text{ and } x \geq 0\}.$$

It can readily be checked that the roots of the cubic equation

$$x^3 - 4x^2 + x + 6 = 0$$

are $-1, 2$, and 3. Thus $B' = \{2, 3\}$, and so we also have $A = B'$. However, the statement $A = B'$ is usually not regarded as a meaningful result because B' has a more complicated description compared wth A.

In mathematics, instead of $A = B$, there are also results of the form $A \subset B$ or $B \subset A$. Here B is called a *partial characterization* of A. Again, the statements $A \subset B$ and $B \subset A$ are meaningful only if B has a simpler description compared with A.

1.3 Relation and Function

Relation

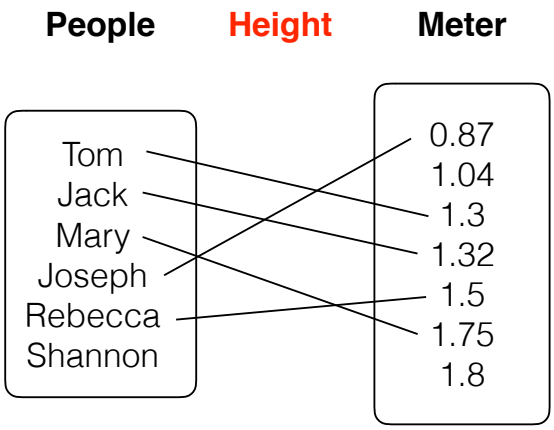
A relation is a subset of $A \times B$, where A and B are sets and $A \times B$ is the Cartesian product of A and B, defined by

$$A \times B = \{(x, y) : x \in A \text{ and } y \in B\}.$$

Example 1.14.

$$\text{People} = \{\text{Tom, Jack, Mary, Joseph, Rebecca, Shannon}\}$$
$$\text{Meter} = \{1.5, 1.8, 1.75, 1.3, 0.87, 1.32, 1.04\}$$
$$\text{Height} = \{(\text{Tom}, 1.3), (\text{Jack}, 1.32), (\text{Mary}, 1.75), (\text{Joseph}, 0.87), (\text{Rebecca}, 1.5)\}$$

A pair such as $(\text{Tom}, 1.3)$ is called an ordered pair.[1] Obviously, Height $\subset$ People $\times$ Meter. Therefore, Height is a relation. Note that each person has at most one height.

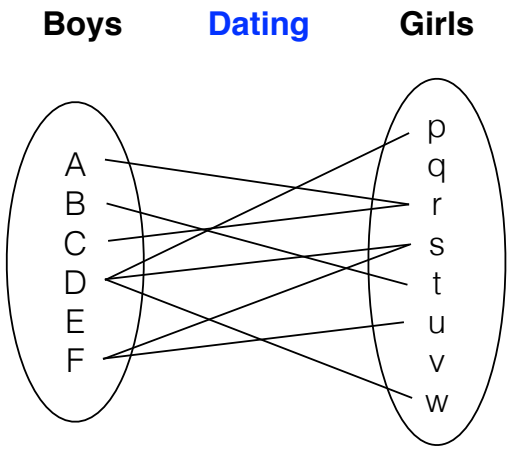


[1] An ordered pair means (x, y) is not regarded as the same as (y, x). On the other hand, the elements in a set are not ordered, i.e., $\{a, b, c\}$ and $\{a, c, b\}$ are regarded as the same set.

Example 1.15.

$$\text{Boys} = \{A, B, C, D, E, F\}$$
$$\text{Girls} = \{p, q, r, s, t, u, v, w\}$$

A relation called 'dating' is specified as below. Note that unlike the last example, there is no constraint on the subset 'dating' can take (in the last example, each person has only one height).



Function

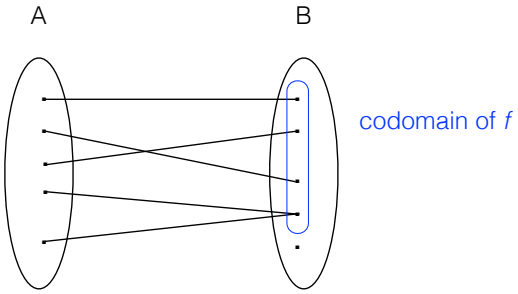
A *function* $f : A \to B$ is a *relation* such that *every* $x \in A$ is associated with *a unique element in* B, denoted by $f(x)$.

- The sets A and B are called the *domain* and the *range* of f, respectively.
- A generic element in A (or B) is called a *variable*.
- When f is a function, we call x the *independent variable*.
- The subset of B,

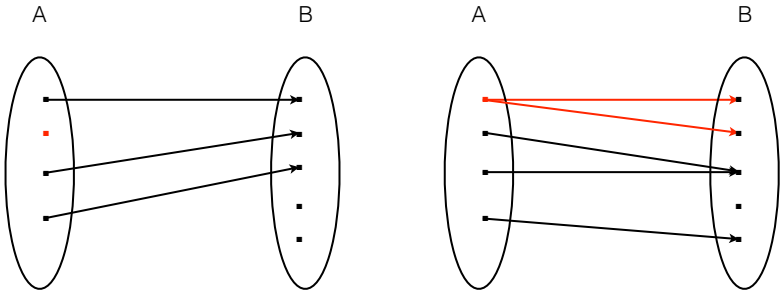
$$\{y \in B : y = f(x) \text{ for some } x \in A\}$$

  is called the codomain of f.

The following is an illustration of a function.



The following are not functions:



For the relation on the left, the red element in A is not assigned to any element in B. For the relation on the right, the red element in A is not assigned to a unique element in B.

Some terminologies:

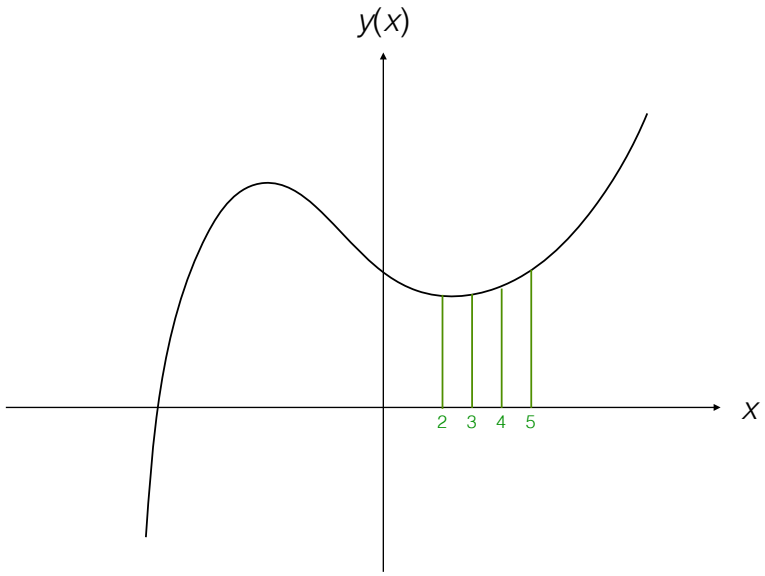
- For a given relation f defined on $A \times B$, let
  - x denotes *a generic element of A*
  - y denotes *a generic element of B*

  Here both x and y are referred to as variables. When we say

  $$\text{“}y \text{ is a function of } x\text{”},$$

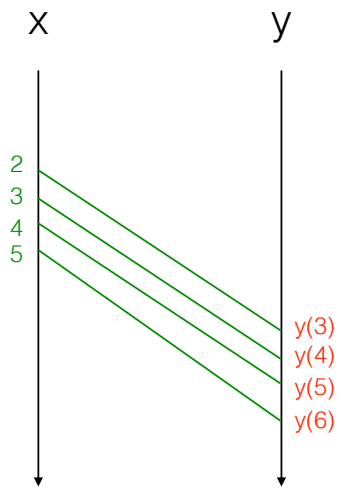
  we mean that the relation *f is a function from A to B*.
- Strictly speaking, for $x \in A$, $f(x)$ denotes the unique element in B that f associates x with.
- However, f is very often written as $f(x)$ to emphasize that *x is the independent variable of f*. An abuse of language!
- “function” and “mapping” are synonyms.

When both $A, B \subset \mathbb{R}$, a relation $f \subset A \times B$ may be represented by a graph. Below is an example.



In the above relation, y is a function of x, but x is not a function of y because for some y values, there is more than one corresponding x value.

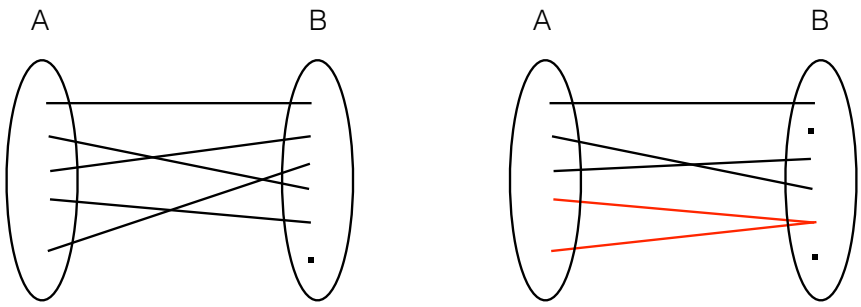
Alternatively, the relation f can be represented as follows:



Injection, Surjection, and Bijection

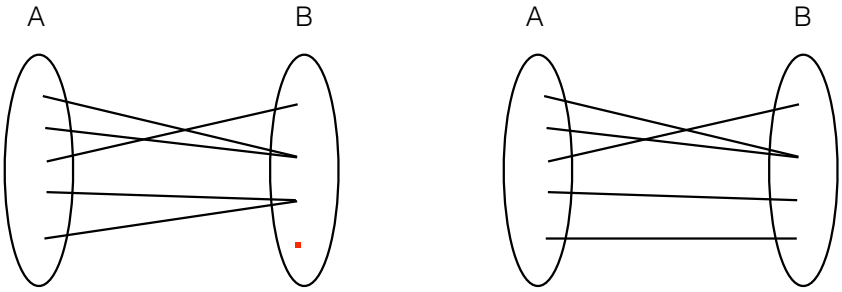
Let f be a function from A to B.

Injection f is called an injection, or an injective function, if for all y in the codomain of f, there is a unique $x \in A$ such that $f(x) = y$. For this reason, an injection is also called a *one-to-one* function (as opposed to a *many-to-one* function).



The function on the left is an injection, or a one-to-one function. The function on the right is not an injection. It is a many-to-one function because there exist $x_1 \neq x_2$ in A such that $f(x_1) = f(x_2)$.

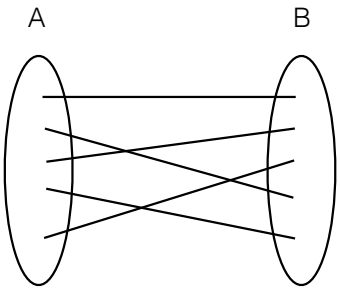
Surjection f is called a surjection, or a surjective function, if its codomain is equal to the range, i.e., for every $y \in B$, there exists $x \in A$ such that $y = f(x)$. A surjection is also called an onto function.



The function on the left is not a surjection because there is a $y \in B$ (the red dot) that does not correspond to any $x \in A$. The function on the right is a surjection.

Bijection f is called a bijection, or a bijective function, if it is *both* an *injection* and a *surjection*. Thus f is a bijection if it is one-to-one and the codomain is equal to the range. In other words, for every $y \in B$, there is a unique $x \in A$ such that $f(x) = y$. As such, there is a one-to-one correspondence (not the same meaning as a one-to-one function) between the elements in A and the elements in B, so that $|A| = |B|$. A bijection is also called a one-one onto function.

Here is an example of a bijection:



1.4 Partition

A partition of a set A is a set of nonempty subsets of A, $\{A_1, A_2, \ldots, A_k\}$ such that

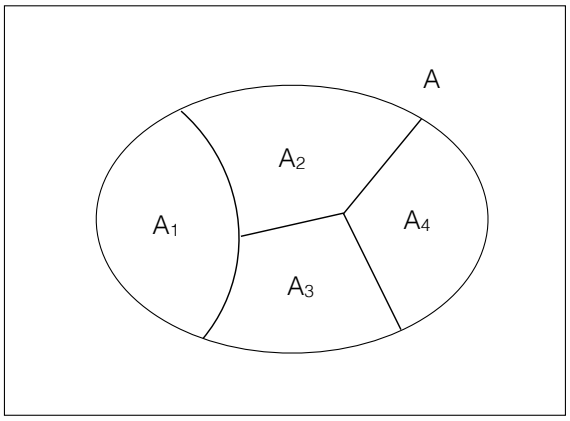
$$\bigcup_{i=1}^{k} A_i = A$$

and $A_i \cap A_j = \emptyset$ for all $i \neq j$. A_i are called the blocks of the partition.

Example 1.16. Let

$$\begin{aligned} A &= \{1, 2, \ldots, 10\} \\ A_1 &= \{1, 3, 7\} \\ A_2 &= \{2\} \\ A_3 &= \{4, 5, 6\} \\ A_4 &= \{8, 9, 10\} \end{aligned}$$

Then $\{A_1, A_2, A_3, A_4\}$ is a partition of A.

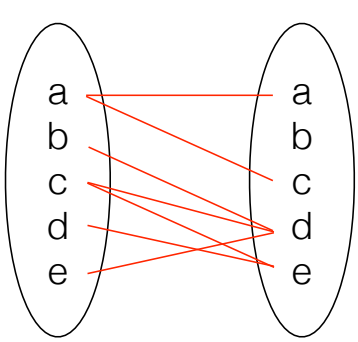


Recall that a relation is a subset of $A \times B$ for some sets A and B. Consider a relation R defined on $A \times A$ for some set A. If $(a, b) \in R$, we write '$a \sim b$' (reads a is related to b). One should not confuse the notation '$\sim$' here with the symbol for NOT in logic.

Example 1.17. Let $A = \{a, b, c, d, e\}$ and

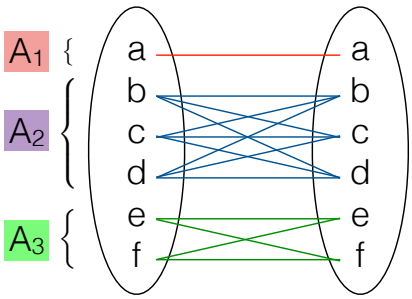
$$R = \{(a, a), (a, c), (b, d), (c, d), (c, e), (d, e), (e, d)\}$$

be a relation defined on $A \times A$.



	a	b	c	d	e
a	✔		✔		
b				✔	
c				✔	✔
d					✔
a				✔	

Example 1.18.



	a	b	c	d	e	f
a	✔					
b		✔	✔	✔		
c		✔	✔	✔		
d		✔	✔	✔		
e					✔	✔
f					✔	✔

Let

$$A_1 = \{a\},\ A_2 = \{b, c, d\},\ A_3 = \{e, f\}.$$

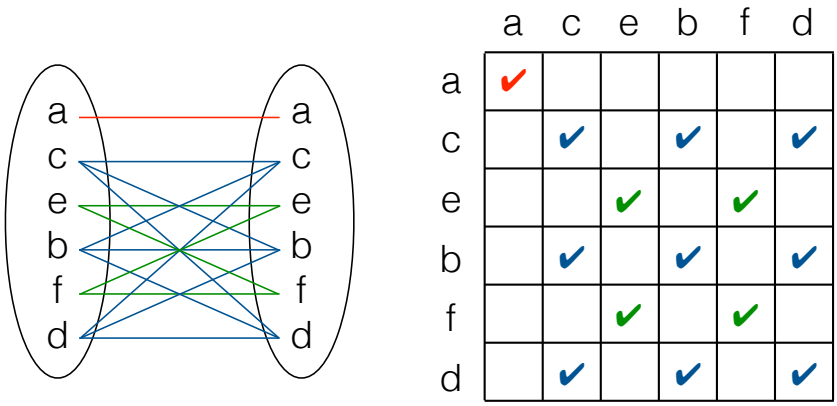
Here R induces the partition $\{A_1, A_2, A_3\}$. Precisely,

$$(x, y) \in R \text{ iff } x, y \in A_i \text{ for some } i = 1, 2, 3,$$

i.e.,

$$x \sim y \text{ iff } x \text{ and } y \text{ belong to the same block.} \tag{1.6}$$

Below is an alternative representation of R, from which it is less easy to see the partition induced by R.



Example 1.19. A is the set of all students living on campus. There are a total of k dormitories. Let A_i be the set of all students living in Dormitory i. Then

$$A_1 \cup A_2 \cup \cdots \cup A_k = A$$
$$A_i \cap A_j = \emptyset \quad \forall i \neq j$$

i.e., $\{A_1, A_2, \ldots, A_k\}$ is a partition of A.

1.5 Equivalence Relation

Consider relations defined on $A \times A$ for some set A. It is trivial to see that a partition of A induces a relation on $A \times A$.

Example 1.20. Consider the partition $\{A_1, A_2, A_3\}$ in Example 1.18. Then obviously it induces the partition as discussed therein, namely the relation specified by (1.6).

We have seen in Examples 1.17 and 1.18 that some relations on $A \times A$ induce a partition of A, while some others do not. For a given such a relation, we want to find a necessary and sufficient condition for the relation to induce a partition.

Toward finding such a condition, we observe from (1.6) that a relation R that induces a partition of A must satisfy the following properties:

1. reflexive for all a, $a \sim a$

2. symmetric if $a \sim b$, then $b \sim a$
3. transitive if $a \sim b \ \wedge \ b \sim c$, then $a \sim c$.

A relation satisfying all the above is called an equivalence relation. In other words, if R induces a partition (of A), then R is an equivalence relation (on $A \times A$). It turns out that the converse is also true.

This fundamental result will be fully developed in Theorem 1.22. Toward this end, we first prove the following lemma.

Lemma 1.21. *Let R be an equivalence relation on $A \times A$. For all $a \in A$, let*

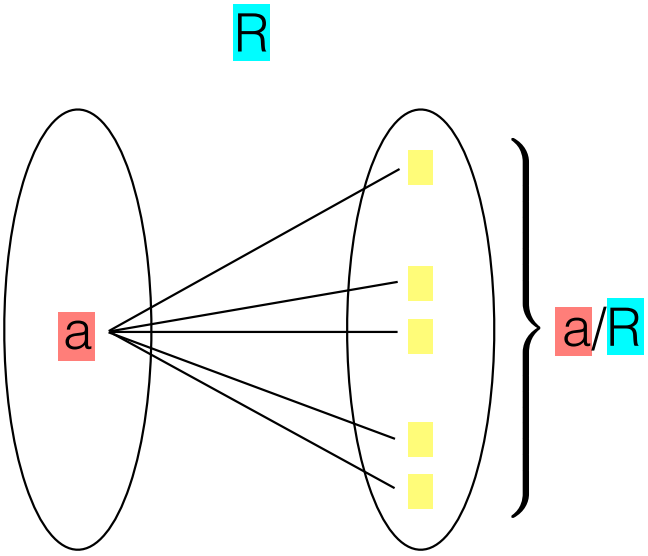
$$a/R = \{b \in A : a \sim b\}.$$

Then

1) For all $a \in A$, $a \in a/R$, and so a/R is nonempty.
2) For all $a, b \in A$, $a/R = b/R$ if and only if $a \sim b$.

Notes

- a/R is a set associated with the element a.
- Its definition depends on the equivalence relation R, as suggested by the notation.
- a/R is called the equivalence class of a (under the relation R).



Proof of Lemma 1.21

1) Since R is reflexive, $a \sim a$. Therefore , $a \in a/R$ by definition.
2) 'Only if': $a/R = b/R \Rightarrow a \sim b$

Assume that $a/R = b/R$, which is nonempty by 1). Let $c \in a/R = b/R$, i.e., $a \sim c$ and $b \sim c$. Then

$$b \sim c \overset{symmetric}{\Longrightarrow} \left.\begin{array}{l} c \sim b \\ a \sim c \end{array}\right\} \overset{transitive}{\Longrightarrow} a \sim b.$$

'If': $a \sim b \Rightarrow a/R = b/R$

Assume $a \sim b$. We need to prove that $a/R \subset b/R$ and $b/R \subset a/R$. First prove that $a/R \subset b/R$. Since a/R is nonempty, consider any $c \in a/R$, i.e., $a \sim c$. Then

$$a \sim b \;\Rightarrow\; \left.\begin{array}{l} a \sim c \\ b \sim a \end{array}\right\} \;\Rightarrow\; b \sim c.$$

Therefore, under the assumption that $a \sim b$,

$$\begin{array}{rcl} a \sim c & \Rightarrow & b \sim c \\ c \in a/R & \Rightarrow & c \in b/R \\ a/R & \subset & b/R. \end{array}$$

By means of a symmetrical argument, we can prove that $b/R \subset a/R$. Hence $a/R = b/R$.

The following is the main theorem of this section.

Theorem 1.22. *If R is an equivalence relation on A, then R induces a partition of A. More precisely, define*

$$A/R = \{a'/R : a' \in A\}, \tag{1.7}$$

called the quotient set of A by R. Then A/R is a partition of A. That is, R induces the partition A/R of A.

Proof

1. Note that a/R is nonempty.
2. For any $a \in A$, $a \in a/R \in A/R$. In other words, for any $a \in A$, there exists at least one element in A/R that contains a. Therefore,

$$\bigcup_{a'} a'/R = A.$$

3. Consider $a, b \in A$ such that $a/R \neq b/R$. We want to prove that

$$a/R \neq b/R \;\Rightarrow\; a/R \cap b/R = \emptyset.$$

Equivalently, we will prove the contrapositive of the above statement:

$$a/R \cap b/R \neq \emptyset \;\Rightarrow\; a/R = b/R.$$

Assume that $a/R \cap b/R \neq \emptyset$ and let $c \in a/R \cap b/R$. Then

$$\left.\begin{array}{l} c \in a/R \Rightarrow a \sim c \\ c \in b/R \Rightarrow b \sim c \Rightarrow c \sim b \end{array}\right\} \;\Rightarrow\; a \sim b \;\Rightarrow\; a/R = b/R,$$

where the last step is by Lemma 1.21. Hence, we have proved that A/R is indeed a partition of A.

Remarks

1. If $a \sim b$, then $a/R = b/R$ (by Lemma 1.21).
2. If $a \not\sim b$, then $a/R \neq b/R$ (by Lemma 1.21) and in fact $a/R \cap b/R = \emptyset$ (by Theorem 1.22).

Corollary 1.23. *$a \sim b$ if and only if a and b are in the same equivalence class.*

Proof

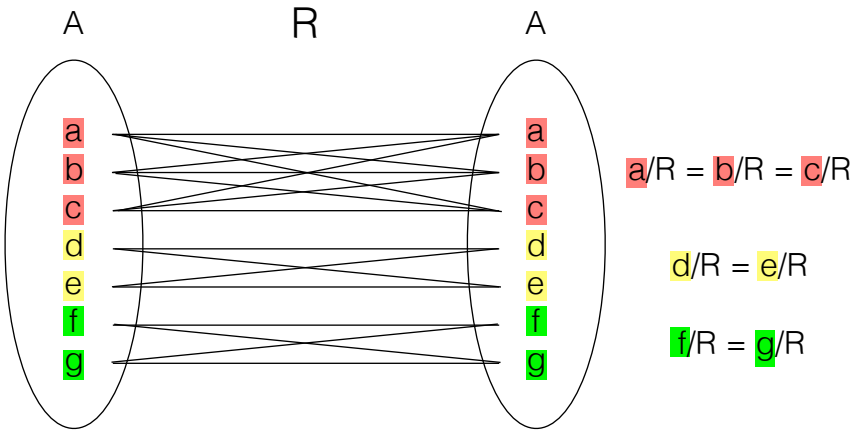
'Only if': $a \sim b \Rightarrow a$ and b are in the same equivalence class

We have $a \in a/R$ and $b \in b/R$. By Lemma 1.21, $a \sim b$ implies $a/R = b/R$. Therefore, a and b are in the same equivalence class.

'If': a and b are in the same equivalence class $\Rightarrow a \sim b$

Assume that a and b are in c/R for some $c \in A$. Then

$$\left.\begin{array}{l} c \sim a \Rightarrow a \sim c \\ c \sim b \end{array}\right\} \;\Rightarrow\; a \sim b.$$



Example 1.24. For the relation given in the figure above, we have

$$a/R = b/R = c/R = \{a, b, c\}$$
$$d/R = e/R = \{d, e\}$$
$$f/R = g/R = \{f, g\}.$$

The quotation set as specified by (1.7) is

$$A/R = \{\ a/R,\ b/R,\ c/R,\ d/R,\ e/R,\ f/R,\ g/R\ \}.$$

But since $a/R = b/R = c/R$, $d/R = e/R$, and $f/R = g/R$, we can write

$$A/R = \{\ a/R,\ d/R,\ f/R\ \}$$

because for a set, repeated elements do not count.

1.6 Different Proof Methods

We will illustrate different proof methods by means of Proposition 1.26, an elementary result. First, we state a few facts about prime factorization of integers.

Prime Factorization of Integers

- Every positive integer has a unique prime factorization. For example,

$$60 = 2^2 \cdot 3 \cdot 5.$$

- If a positive integer is even, its prime factorization must contain 2 as a factor, i.e., the power of 2 in the prime factorization must be at least 1.
- If a positive integer is the square of an integer, then in its prime factorization, all the powers must be even.
- In particular, if a positive integer is the square of an even integer, then its prime factorization must contain an even power of 2.

Example 1.25. Consider 36, the square of 6, an even integer. We have

$$36 = 2^2 \cdot 3^2.$$

Note that all the powers are even. Then

$$\sqrt{36} = (2^2 \cdot 3^2)^{1/2} = 2 \cdot 3 = 6.$$

Proposition 1.26. *Let x be a positive integer. If x^2 is even, then x is even. In terms of logic, this means*

$$\forall \text{ positive integer } x, \text{ '}x^2 \text{ is even} \Rightarrow x \text{ is even'}$$

In terms of set theory, this means

$$\{x \in \mathbb{Z}^+ : x^2 \text{ is even}\} \subset \{x \in \mathbb{Z}^+ : x \text{ is even}\},$$

where $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$.

We now prove Proposition 1.26 in different ways.

Proof 1 (Direct Proof)
Assume that x^2 is a positive even integer and let $x^2 = 2n$, where n is a positive integer. Then x^2 must contain 2 as a prime factor. Let

$$n = 2 \cdot \prod_{i=1}^{m} u_i^{2\,k_i} \tag{1.8}$$

be the unique prime factorization, where u_i is a prime and k_i is a positive integer. Note that the prime factorization of n must contain an odd power of 2 because x^2 is the square of a positive integer (whose prime factorization must contain an even power of 2), so (1.8) is the general form for n. It then follows that

$$x^2 = 2 \cdot n = 2\left(2 \cdot \prod_{i=1}^{m} u_i^{2k_i}\right) = 2^2\left(\prod_{i=1}^{m} u_i^{2k_i}\right)$$

and

$$x = 2 \cdot \prod_{i=1}^{m} u_i^{k_i}$$

Therefore, x is even.

Proof 2 (Proof by Contrapositive)
We instead prove the contrapositive of Proposition 1.26:

$$\forall \text{ positive integer } x, \text{ '}x \text{ is not even} \Rightarrow x^2 \text{ is not even'}$$

Let $x = 2n + 1$, where n is a nonnegative integer. Then $x^2 = (2n + 1)^2$ is odd because

$$\begin{aligned} x^2 &= (2n+1)^2 \\ &= 4n^2 + 4n + 1 \\ &= 4n(n+1) + 1. \end{aligned}$$

For Proposition 1.26, the direct proof is somewhat more difficult than the proof by contrapositive. In general, one of the proofs may be more difficult than the other.

We now look at another commonly proof method, called proof by contraction.

Proof 3 (Proof by Contradiction)
We want to prove that

$$\forall \text{ positive integer } x, \text{ '}x^2 \text{ is even} \Rightarrow x \text{ is even'}$$

Let x^2 is even and x is odd (i.e., the claimed conclusion is false). Since x is odd, let $x = 2n + 1$. Then $x^2 = (2n + 1)^2$ is odd. This is a contraction to the assumption that x^2 is even.

A Note on Proof by Contradiction

Direct Proof: $X \Rightarrow Y$
Proof by contrapositive: $\sim Y \Rightarrow \sim X$

Proof by contradiction:

1. Assume X and $\sim Y$.
2. Show that $\sim Y$ implies $\sim X$. This is a contradiction to the assumption X.
3. Therefore, under the assumption X, $\sim Y$ cannot be true. So Y is true.

As we see, proof by contradiction is just another form of proof by contrapositive.

1.7 Mathematical Induction

Example 1.27. Consider the hypothesis:

$$1+2+\cdots+n=\frac{n(n+1)}{2} \quad \text{for all } n\geq 1$$

Denote this hypothesis by $P(n)$.

Direct Proof

Let

$$\begin{array}{rcccccccc} x & = & 1 & +2 & +3 & + & \ldots & +(n-1) & +n \\ x & = & n & +(n-1) & +(n-2) & + & \ldots & +2 & +1 \\ \hline 2x & = & (n+1) & +(n+1) & +(n+1) & + & \ldots & +(n+1) & +(n+1) \end{array}$$

There are n terms on the RHS. Therefore, we have

$$2x=n(n+1)$$

or

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

Proof by Mathematical Induction

Let

$$P(n): 1+2+\cdots+n=\frac{n(n+1)}{2} \quad \text{for all } n\geq 1 \tag{1.9}$$

Base Case $n=1$

$$\begin{aligned} \text{LHS} &= 1 \\ \text{RHS} &= \frac{1\cdot(1+1)}{2}=1=\text{RHS} \end{aligned}$$

Therefore, $P(1)$ is true.

Induction Step Assume $P(n)$ is true for some $n\geq 1$. Consider

$$P(n+1): 1+2\cdots+n+(n+1)=\frac{(n+1)[(n+1)+1]}{2}=\frac{(n+1)(n+2)}{2},$$

which is obtained from (1.9) by replacing n by $n+1$. Now

$$\begin{aligned}
\text{LHS} &= 1+2+3+\cdots+n+(n+1) \\
&= (1+2+3+\cdots+n)+(n+1) \\
&= \frac{n(n+1)}{2}+(n+1) \\
&= (n+1)\left[\frac{n}{2}+1\right] \\
&= \frac{(n+1)(n+2)}{2} \\
&= \text{RHS}.
\end{aligned}$$

Hence, $P(n+1)$ is true.

Idea

- By verifying the base case, we have $P(1)$.
- Using the induction step, we have

$$\begin{aligned}
P(1) &\Rightarrow P(2) \\
P(2) &\Rightarrow P(3) \\
&\vdots
\end{aligned}$$

This proves $P(n)$ for all $n \geq 1$.

The Inclusion-Exclusion Formula

Recall that $|A|$ denotes the cardinality of a set A.

Example 1.28. For 2 sets A_1 and A_2, we have

$$|A_1 \cup A_2| = |A_1| + |A_2| - |A_1 \cap A_2|$$

This can easily be verified by a Venn diagram.

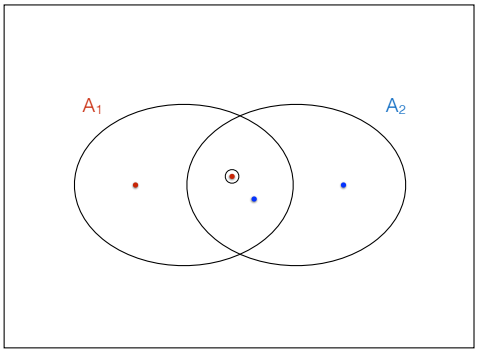
Example 1.29. For 3 sets A_1, A_2, and A_3, we have

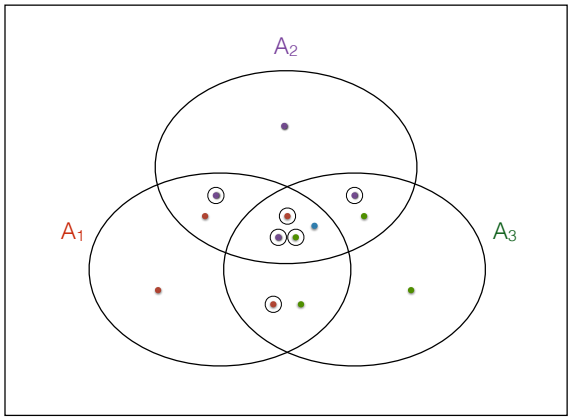
$$\begin{aligned}
|A_1 \cup A_2 \cup A_3| = {} & |A_1| + |A_2| + |A_3| \\
& -|A_1 \cap A_2| - |A_1 \cap A_3| - |A_2 \cap A_3| \\
& +|A_1 \cap A_2 \cap A_3|
\end{aligned}$$

Based on these two examples, we can hypothesize the following formula for n sets, known as the inclusion-exclusion formula.

$$\begin{aligned}
|A_1 \cup A_2 \cup \cdots \cup A_n| = {} & \sum_{1 \leq i \leq n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j| + \ldots \\
& +(-1)^{n+1} |A_1 \cap A_2 \cap \cdots \cap A_n|
\end{aligned}$$

The inclusion-exclusion formula has the following assertion for $n = 1, 2, 3$:





$\underline{n=1}$ $|A_1| = |A_1|$, trivial.
$\underline{n=2}$ Already verified in Example 1.28.
$\underline{n=3}$ Already verified in Example 1.29.

For $n \geq 4$, it is very difficult to verify. Instead will prove the formula by induction.

Base Case The hypothesis is verified for $n = 1$ and $\underline{n=2}$. We will see in the induction step that it does not suffice to verify only for $n = 1$ for the base case.

Induction Step Assume that the hypothesis is true for some $n \geq 2$. Prove that it is true for $n+1$. Then

$$
\begin{aligned}
&\left| \bigcup_{1 \leq i \leq n+1} A_i \right| \\
&= \left| \left(\bigcup_{1 \leq i \leq n} A_i \right) \cup A_{n+1} \right| \\
&\overset{i)}{=} \left| \bigcup_{1 \leq i \leq n} A_i \right| + |A_{n+1}| - \left| \left(\bigcup_{1 \leq i \leq n} A_i \right) \cap A_{n+1} \right| \\
&\overset{ii}{=} \left| \bigcup_{1 \leq i \leq n} A_i \right| + |A_{n+1}| - \left| \bigcup_{1 \leq i \leq n} (A_i \cap A_{n+1}) \right| \\
&\overset{iii)}{=} \left\{ \sum_{1 \leq i \leq n} |A_i| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j| \right. \\
&\quad \left. + \sum_{1 \leq i < j < k \leq n} |A_i \cap A_j \cap A_k| \ldots + (-1)^{n+1} |A_1 \cap A_2 \cap \ldots \cap A_n| \right\} \\
&\quad + |A_{n+1}| \\
&\quad - \left\{ \sum_{1 \leq i \leq n} |A_i \cap A_{n+1}| - \sum_{1 \leq i < j \leq n} |A_i \cap A_j \cap A_{n+1}| + \ldots \right. \\
&\quad \left. + (-1)^n \sum_{l=1}^{n} \left| \left(\bigcap_{\substack{1 \leq m \leq n \\ m \neq l}} A_m \right) \cap A_{n+1} \right| + (-1)^{n+1} |A_1 \cap A_2 \cap \ldots \cap A_n \cap A_{n+1}| \right\} \\
&= \sum_{1 \leq i \leq n+1} |A_i| - \sum_{1 \leq i < j \leq n+1} |A_i \cap A_j| + \sum_{1 \leq i < j < k \leq n+1} |A_i \cap A_j \cap A_k| - \ldots \\
&\quad + (-1)^{n+1} \sum_{1 \leq l \leq n+1} \left| \bigcap_{\substack{1 \leq m \leq n+1 \\ m \neq l}} A_m \right| + (-1)^{n+2} |A_1 \cap A_2 \cap \ldots \cap A_n \cap A_{n+1}|
\end{aligned}
$$

where

- in *i)*, we have applied the hypothesis for $n = 2$;
- *ii)* follows from the distribution law;

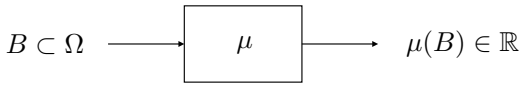
- in *iii*), we have applied the induction hypothesis for n to $\left|\bigcup_{1\le i\le n} A_i\right|$ and $\left|\bigcup_{1\le i\le n}(A_i \cap A_{n+1})\right|$, each being the cardinality of the union of n sets.

This proves the hypothesis for $n+1$.

Note In the base case, we need to verify the hypothesis for $n = 2$ because we need it in *iii*) above.

Set-Additive Functions

A function μ, which takes *a set* (a subset of Ω) to a real number, is called a *set function*.



The function μ is called a *set-additive function* if for all $B, B' \subset \Omega$ such that $B \cap B' = \emptyset$,

$$\mu(B \cup B') = \mu(B) + \mu(B').$$

Example 1.30 (Cardinality of Sets). Let $\mu(B) = |B|$ for $B \subset \Omega$. Obvioulsy, if B and B' are disjoint, then

$$\mu(B \cup B') = |B \cup B'| = |B| + |B'| = \mu(B) + \mu(B').$$

Remark If μ is always nonnegative (i.e., $\mu(B) \ge 0$ for all B) and μ is set-additive, then μ is called a measure, and $\mu(B)$ can be interpreted as the "weight" of B.

Example 1.31 (A Cookie). Let

Ω – a piece of cookie
B – a part of the cookie
$\mu(B)$ – the weight of B

If $B \cap B' = \emptyset$, then evidently

$$\mu(B \cup B) = \mu(B) + \mu(B').$$

This set-additivity of μ captures the notion of a "*weight distribution*".

Proposition 1.32. *The inclusion-exclusion formula can be extended to any set-additive function μ, i.e.,*

$$\mu(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{1\le i\le n} \mu(A_i) - \sum_{1\le i<j\le n} \mu(A_i \cap A_j) + \ldots + (-1)^{n+1}\mu(A_1 \cap A_2 \cap \cdots \cap A_n).$$

Proof The proof is exactly the same as the proof of the inclusion-exclusion formula.

Example 1.33 (Probability).

Ω – called the sample space of an experiment
$B \subset \Omega$ – called an event

For all $B \subset \Omega$, $\mu(B)$, denoted by $\Pr\{B\}$, is called the probability of event B. Then for all $B, B' \subset \Omega$ such that $B \cap B' = \emptyset$,

$$\Pr\{B \cup B'\} = \Pr\{B\} + \Pr\{B'\}.$$

Example 1.34 (Tossing 2 Fair Coins). Let

$$\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$$

be the set of all possible outcomes of tossing 2 fair coins. Then

$$\begin{aligned}\Pr\{\{(H,H),(T,T)\}\} &= \Pr\{\{(H,H)\} \cup \{(T,T)\}\} \\ &= \Pr\{\{(H,H)\}\} + \Pr\{\{(T,T)\}\} \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{1}{2}\end{aligned}$$

A Variation of the Inclusion-Exclusion Formula

By replacing '$\cup$' by '$\cap$' and '$\cap$' by '$\cup$' in the inclusion-exclusion formula, we obtain

$$\begin{aligned}\mu(A_1 \cap A_2 \cap \cdots \cap A_n) = &\sum_{1 \le i \le n} \mu(A_i) - \sum_{1 \le i < j \le n} \mu(A_i \cup A_j) + \ldots \\ &+ (-1)^{n+1} \mu(A_1 \cup A_2 \cup \cdots \cup A_n).\end{aligned}$$

The proof of this formula is very similar to the proof of the inclusion-exclusion formula and is left as an exercise.

1.8 Combinatorics

Combinatorics is about how to count. The most important concepts in combinatorics are factorial, permutation, and combination.

Factorial

Consider n balls numbered from 1 to n. How many ways are there to order these n balls? To tackle this problem, it helps to think of it as filling an array of n boxes by these n balls.

To fill the first box in the array, there are n choices. To fill the second box, there are $n-1$ choices because 1 ball has already been taken by the first box. Keep going this way, we see that the total number of choices of filling in the n boxes is

$$n \cdot (n-1) \cdot (n-2) \ \ldots \ 2 \cdot 1.$$

This number, denoted by $n!$, is called n *factorial*.

Permutation

Again consider n balls numbered from 1 to n. What is the number of ways to pick k balls out of these n balls in an ordered manner? It helps to think of it as filling an array of k boxes by these n balls.

Obviously, when $k = n$, the problem reduces to the one in the last subsection. Using the same argument, we see that the total number of ways is

$$\underbrace{n(n-1) \ \cdots \ (n-k+1)}_{k \text{ terms}},$$

denoted by $P(n,k)$. We can also write

$$P(n,k) = \frac{n!}{(n-k)!}.$$

Combination

Again consider n balls numbered from 1 to n. What is the number of ways to pick k balls out of these n balls in an unordered manner? We continue to think of it as filling an array of k boxes by these n balls, but the order is immaterial.

As the order is immaterial, for each way of picking k balls with the order considered, there are $k!$ ways that are considered as equivalent. As such, the total number of ways (unordered) is given by

$$\frac{P(n,k)}{k!} = \frac{n!}{k!(n-k)!},$$

denoted by $C(n,k)$, or

$$\binom{n}{k}$$

(read as "n choose k"). Evidently, $C(n,k)$ is also the number of subsets of $\{1,2,\ldots,n\}$ of size k.

The Binomial Formula

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^r b^{n-r}$$

Direct Proof

Consider

$$(a+b)^n = \underbrace{(a+b)(a+b)\cdots(a+b)}_{n \text{ terms}}$$

Expanding the RHS, the coefficient of $a^r b^{n-r}$, where $r = 0, 1, 2, \ldots, n$, is equal to

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

Proof by Induction Exercise.

For simple claims, a direct proof may be easier than mathematical induction. However, it may not be the same for more complicated claims.

2

Probability and Events

Probability theory is a mathematical theory for modeling the *outcome* of a *random experiment*. It has vast applications in science, engineering, finance, etc. In particular, it is an essential tool for understanding communication systems and networks.

2.1 Elements of a Probability Model

Probability is a *state of knowledge* about something unknown. This knowledge may change due to availability of certain information. These will be discussed in the next two sections.

A probability model for a random experiment consists of the following elements:

1. a *sample space* Ω containing all possible outcomes of the random experiment;
2. a *set function* P, called a *probability measure*, such that
   A1. $0 \leq P(E) \leq 1$ for all $E \subset \Omega$
   A2. $P(\Omega) = 1$
   A3. $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$.

Note: A set function is a function that maps a set to a real number.

The outcome of the random experiment, an element of Ω, is denoted by ω. A subset E of Ω is called an *event*, and we say that event E occurs if $\omega \in E$. The quantity $P(E)$ is interpreted as the *probability that event E occurs*. The probability measure $P(\cdot)$ represents our knowledge about the random experiement. A1 – A3 are called the *axioms of probability* due to the great mathematician A. Kolmogorov,
http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Kolmogorov.html.

Corollary 2.1. $P(\emptyset) = 0$.

Proof

$$\begin{aligned} P(\Omega) &= P(\Omega \cup \emptyset) \\ &= P(\Omega) + P(\emptyset) \quad \text{by A3} \end{aligned}$$

implies $P(\emptyset) = 0$.

Corollary 2.2. $P(A^c) = 1 - P(A)$.

Proof Consider

$$\begin{aligned} 1 &= P(\Omega) \\ &= P(A \cup A^c) \\ &= P(A) + P(A^c) \quad \text{by A3.} \end{aligned}$$

Therefore, $P(A^c) = 1 - P(A)$.

Corollary 2.3. $P(A) \leq P(B)$ *if* $A \subset B$.

Proof The fact that $B = A \cup (B - A)$ can easily be seen from a Venn diagram, so that

$$\begin{aligned} P(B) &= P(A \cup (B - A)) \\ &= P(A) + P(B - A) \quad \text{by A3} \\ &\geq P(A) \quad \text{by A1.} \end{aligned}$$

One can think of Ω as a cookie of weight 1, and P characterizes the weight distribution of the cookie. Obviously, $0 \leq P(E) \leq 1$ for all $E \subset \Omega$, and $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$ means that weight is additive. Note that the cookie can have *point masses*.

Example 2.4 (Tossing a fair die twice). Let

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\} = \{1, 2, \cdots, 6\}^2$$

and take

$$P(E) = \frac{1}{36}|E|$$

(to model that the die is *fair* and the two tosses are *independent*[1]). For example,

$$\begin{aligned} E &= \text{the set of all outcomes s.t. the sum is 6} \\ &= \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}. \end{aligned}$$

We now check that P is a valid probability measure by showing that it satisfies all the axioms of probability.

[1] To be discussed in the next chapter.

A1: Since $0 \leq |E| \leq 36$, A1 is obviously satisfied.
A2: Trivial.
A3: Since $|A \cup B| = |A| + |B|$ if $A \cap B = \emptyset$, we can readily verify A3.

Example 2.5. Following Example 2.4, we let X be the outcome of the first toss. Formally, $X : \Omega \to \{1, 2, \cdots, 6\}$, and $X((i, j)) = i$.

Example 2.6. Following Example 2.4, we define the random variable

$$Y = \begin{cases} 1 \text{ if } 2^{\text{nd}} \text{ toss} = 2 \times 1^{\text{st}} \text{ toss} \\ 0 \text{ otherwise.} \end{cases}$$

Formally, $Y : \Omega \to \{0, 1\}$, where

$$Y((1, 2)) = Y((2, 4)) = Y((3, 6)) = 1,$$

and $Y(\omega) = 0$ for $\omega \notin \{(1, 2), (2, 4), (3, 6)\}$.

2.2 Probability as a State of Knowledge

Think of probability as our state of knowledge about something. For example, our state of knowledge about the outcome of tossing a fair coin before the toss is

$$H : 0.5 \qquad T : 0.5$$

After the tossing and upon seeing the coin, our state of knowledge becomes

$$H : 1 \qquad T : 0$$

with probability $\frac{1}{2}$ (a HEAD is obtained), and becomes

$$H : 0 \qquad T : 1$$

with probability $\frac{1}{2}$ (a TAIL is obtained).

2.3 Conditional Probability

Definition 2.7. *The probability of event A conditioning on event B is defined by*

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

if $P(B) > 0$ (undefined if $P(B) = 0$).

The above formula specifies how we update our knowledge about event A from $P(A)$ to $P(A|B)$ upon knowing that the event B has occurred. If we fix the event B, then $P(\cdot|B)$ is a set function. Our knowledge of the random experiment before we know that the event B has occurred is given by $P(\cdot)$, and it becomes $P(\cdot|B)$ after we know that the event B has occurred. We will show later in Theorem 2.11 that $P(\cdot|B)$ is indeed a valid probability measure.

Example 2.8. Following Example 2.4,

$$P(\{(1,2)\}|\{(1,2),(2,4),(3,6)\}) = \frac{P(\{(1,2)\})}{P(\{(1,2),(2,4),(3,6)\})}.$$

Now $P(\{(1,2)\}) = \frac{1}{36}$, and

$$P(\{(1,2),(2,4),(3,6)\}) = \frac{3}{36}.$$

Therefore,

$$P(\{(1,2)\}|\{(1,2),(2,4),(3,6)\}) = \frac{1/36}{3/36} = \frac{1}{3}.$$

Thus, the probability that $(1,2)$ has occurred is updated from $\frac{1}{36}$ to $\frac{1}{3}$, upon knowing that one of $(1,2)$, $(2,4)$, and $(3,6)$ has occurred. Similarly,

$$P(\{(2,4)\}|\{(1,2),(2,4),(3,6)\}) = \frac{1}{3}$$

and

$$P(\{(3,6)\}|\{(1,2),(2,4),(3,6)\}) = \frac{1}{3}.$$

On the other hand,

$$\begin{aligned} &P(\{(1,3)\}|\{(1,2),(2,4),(3,6)\}) \\ &= \frac{P(\emptyset)}{P(\{(1,2),(2,4),(3,6)\})} \\ &= \frac{0}{\frac{1}{36}} \\ &= 0. \end{aligned}$$

Similarly,

$$P(\{(i,j)\}|\{(1,2),(2,4),(3,6)\}) = 0$$

for all $(i,j) \notin \{(1,2),(2,4),(3,6)\}$.

Example 2.9. Let $\Omega = \{1,2,3,4,5,6\}$, with

i	1	2	3	4	5	6
$P(\{i\})$	0.1	0.2	0.1	0.1	0.3	0.2

Let

$$B = \{\omega : \omega \geq 3\}.$$

Then it can easily be shown that

i	1	2	3	4	5	6
$P(\{i\}\|B)$	0	0	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{3}{7}$	$\frac{2}{7}$

Example 2.10. Following the last example, further define the event

$$A = \{\omega : \omega \text{ is odd } \}$$

Then

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\{3,5\})}{P(\{3,4,5,6\})} \\ &= \frac{0.1 + 0.3}{0.1 + 0.1 + 0.3 + 0.2} \\ &= \frac{4}{7}. \end{aligned}$$

We say that the set function P is a probability measure because it satisfies the axioms A1 – A3. The same can be said about $P(\cdot|B)$, the set function derived from P by conditioning on event B. This is proved in the next theorem.

Theorem 2.11. *Let B be any event such that $P(B) > 0$. Then the set function $P(\cdot|B)$ is a probability measure.*

Proof First, consider any event E. Then

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

since $P(E \cap B) \geq 0$ by A1. On the other hand,

$$\begin{aligned} P(E|B) &= \frac{P(E \cap B)}{P(B)} \\ &= \frac{P(E \cap B)}{P((E \cap B) \cup (E^c \cap B))} \\ &= \frac{P(E \cap B)}{P(E \cap B) + P(E^c \cap B)} \quad \text{by A3} \\ &\leq \frac{P(E \cap B)}{P(E \cap B)} \quad \text{by Corollary 2.3} \\ &= 1. \end{aligned}$$

Therefore, $P(\cdot|B)$ satisfies A1.

Second, consider

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

Therefore, $P(\cdot|B)$ satisfies A2.

Third, for any events E_1 and E_2 such that $E_1 \cap E_2 = \emptyset$, we have

$$\begin{aligned}
&P(E_1 \cup E_2|B) \\
&= \frac{P((E_1 \cup E_2) \cap B)}{P(B)} \\
&= \frac{P(E_1 \cap B) \cup (E_2 \cap B))}{P(B)} \\
&= \frac{P(E_1 \cap B) + P(E_2 \cap B)}{P(B)} \\
&= \frac{P(E_1 \cap B)}{P(B)} + \frac{P(E_2 \cap B)}{P(B)} \\
&= P(E_1|B) + P(E_2|B).
\end{aligned}$$

Therefore, $P(\cdot|B)$ satisfies A3, and we conclude that $P(\cdot|B)$ is a probability measure.

In fact, the set function $P(\cdot)$ can formally be viewed as $P(\cdot|\Omega)$. To see this, we only have to observe that for any event E,

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

Example 2.12. Following the setup in Example 2.10, since $P(\cdot|B)$ is a probability measure, we have

$$\begin{aligned}
P(A|B) &= P(\{3,5\}|B) \\
&= P(\{3\} \cup \{5\}|B) \\
&= P(\{3\}|B) + P(\{5\}|B) \quad \text{by A3} \\
&= \frac{1}{7} + \frac{3}{7} \quad \text{from Example 2.9} \\
&= \frac{4}{7},
\end{aligned}$$

which is consistent with what we obtained in Example 2.10.

Since $P(\cdot|B)$ is a probability measure, Corollaries 2.1, 2.2, and 2.3 can also be applied to $P(\cdot|B)$. Thus we have

Corollary 2.13.

1. $P(\emptyset|B) = 0$.
2. $P(A^c|B) = 1 - P(A|B)$.
3. $P(A|B) \leq P(A'|B)$ *if* $A \subset A'$.

2.4 The Law of Total Probability and the Bayes Theorem

Definition 2.14. *A collection of sets* $\{B_i\}$ *is a partition of* Ω *if*

1. $\cup_i B_i = \Omega$
2. $B_i \cap B_j = \emptyset$ *if* $i \neq j$.

Theorem 2.15 (The Law of Total Probability). *Let* $\{B_i\}$ *be a partition of* Ω. *Then for any event* A,

$$P(A) = \sum_i P(A|B_i)P(B_i).$$

Proof The theorem can be proved by considering

$$\begin{aligned} P(A) &= \sum_i P(A \cap B_i) \\ &= \sum_i P(A|B_i)P(B_i). \end{aligned}$$

In the law of total probability, the event A can be interpreted as an observation (or a consequence), and $\{B_i\}$ can be interpreted as a collection of mutually exclusive events which can possibly cause the observation A. Figure 2.1 illustrates the relation between A and $\{B_i\}$.

Example 2.16. Let Ω be the set of all students in a class, M be the set of all males students, and F be the set of all female students. Obviously, $\{M, F\}$ is a partition of Ω. Let D be the set of students who wear dresses. Then the law of total probablity says that

$$P(D) = P(D|M)P(M) + P(D|F)P(F).$$

Theorem 2.17 (Bayes Theorem). *For any events* A *and* B *such that* $P(A), P(B) > 0$,

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

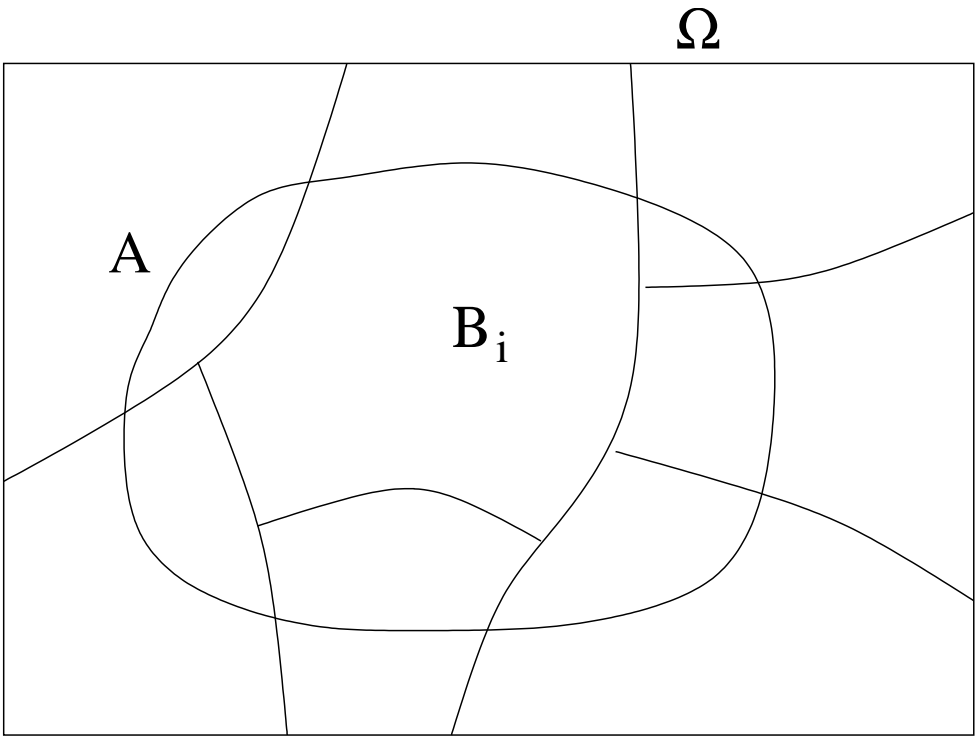


Fig. 2.1: A illustration of the relation between A and $\{B_i\}$.

Proof By definition, we have

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

or

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

Exchanging the roles of A and B above, we obtain

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

Since $B \cap A = A \cap B$ (and hence $P(B \cap A) = P(A \cap B)$), we have

$$P(B|A)P(A) = P(A|B)P(B),$$

or

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

Corollary 2.18. *Let $\{B_i\}$ be a partition of Ω. Then for any event A,*

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}.$$

Proof This follows immediately from the preceding two theorems.

The formula in Corollary 2.18 expresses $P(B_i|A)$ in terms of $\{P(A|B_j)\}$ and $\{P(B_j)\}$.

Example 2.19. The probability that a student being lazy (L) is 0.2. The probability that a student fails (F) in a certain examination is 0.8 given that (s)he is lazy, while the probability that a student fails in the examination is 0.15 given that (s)he is not lazy. That is,

$$\begin{aligned} P(L) &= 0.2 \\ P(F|L) &= 0.8 \\ P(F|L^c) &= 0.15. \end{aligned}$$

We are interested in the probabilities $P(L|F)$, $P(L^c|F)$, $P(L|F^c)$, and $P(L^c|F^c)$.

We proceed to determine $P(L|F)$. By Corollary 2.18, we have

$$\begin{aligned} P(L|F) &= \frac{P(F|L)P(L)}{P(F|L)P(L) + P(F|L^c)P(L^c)} \\ &= \frac{P(F|L)P(L)}{P(F|L)P(L) + P(F|L^c)(1 - P(L))} \\ &= \frac{(0.8)(0.2)}{(0.8)(0.2) + (0.15)(1 - 0.2)} \\ &= \frac{4}{7}. \end{aligned}$$

By the remark at the end of Section 2.4, we have

$$P(L^c|F) = 1 - P(L|F) = \frac{3}{7}.$$

The probabilities $P(L|F^c)$ and $P(L^c|F^c)$ can be determined likewise. As an exercise, the reader should identify the sets A and $\{B_i\}$ in Corollary 2.18.

2.5 Independent Events

Let A and B be two events, and assume for the time being that $P(A), P(B) > 0$. The events A and B are *independent* if the probability of one event is not changed by knowing the other event. Specifically,

$$P(A) = P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

and

$$P(B) = P(B|A) = \frac{P(A \cap B)}{P(A)}. \tag{2.2}$$

It is easy to see that both (2.1) and (2.2) are equivalent to

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

The above relation will be used as the definition for independence. In fact, it is more general then either (2.1) or (2.2) because it avoids the problem of division by zero when $P(A)$ or $P(B)$ is zero.

Definition 2.20. *Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.*

The reader should carefully distinguish between two events being disjoint and two events being independent. The latter has to do with the probability measure P while the former does not.

For three events A, B, and C, we have two notions of independence.

Definition 2.21 (Pairwise Independence). *Three events $A, B,$ and C are pairwise independent if*

$$P(A \cap B) = P(A)P(B), \tag{2.3}$$
$$P(A \cap C) = P(A)P(C), \tag{2.4}$$
$$P(B \cap C) = P(B)P(C). \tag{2.5}$$

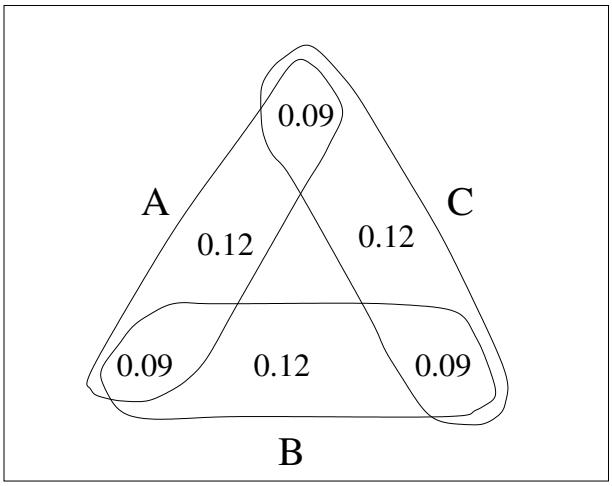


Fig. 2.2: The probability measure for Example 2.23.

Definition 2.22 (Mutual Independence). *Three events A, B, and C are mutually independent if (2.3)-(2.5) are satisfied and also*

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

Clearly, if events A, B, and C are mutually independent, then they are also pairwise independent. However, the converse is not true, as is shown in the next example.

Example 2.23. Consider the probability measure shown in Figure 2.2. First, we have

$$P(A) = 0.09 + 0.12 + 0.09 = 0.3.$$

By symmetry, we also have $P(B) = P(C) = 0.3$. Now,

$$P(A \cap B) = P(A \cap C) = P(B \cap C) = 0.09.$$

Therefore, it is readily seen that events A, B, and C are pairwise independent. However, since

$$P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C),$$

events A, B, and C are not mutually independent.

3

Random Variables

As we have discussed, a random variable X is a function of ω, the outcome of the random experiment in discourse. Formally, $X : \Omega \to \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers. The *alphabet* of X, denoted by $\mathcal{X}$, is a subset of $\mathbb{R}$ containing all possible values taken by X. If $\mathcal{X}$ is a finite set, we denote its cardinality (size) by $|\mathcal{X}|$.

Example 3.1. Let $\Omega = \{1, 2, 3, 4, 5, 6\}$, with

ω	1	2	3	4	5	6
$P(\{\omega\})$	0.1	0.2	0.1	0.1	0.3	0.2

as in Example 2.10. Define random variables X and Y as follows.

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

We can let $\mathcal{X} = \{0, 1, 2, 3, 4\}$, with $|\mathcal{X}| = 5$. Similarly, we can let $\mathcal{Y} = \{0, 1\}$.

Example 3.2. $X : \Omega \to \mathbb{R}$ such that $X(\omega) = \omega$ defines a random variable X.

Example 3.3. Suppose we choose a time in a day at random. To model this, we let

$$\Omega = \{(h, m) : 0 \leq h \leq 23, 0 \leq m \leq 59\}.$$

Then the two functions *HR* and *MIN* defined by

$$HR(\omega) = h$$

and

$$MIN(\omega) = m$$

are random variables representing respectively the hour and minute of the random time ω chosen. Likewise, the function

$$AM(\omega) = \begin{cases} 0 \text{ if } h \geq 12 \\ 1 \text{ otherwise} \end{cases}$$

is the random variable denoting "morning".

Example 3.4. Let Ω be the set of all possible states of the body of a person chosen at random. Then the height, the weight, the body temperature, the blood pressure, the eye color, etc, of the chosen person are all random variables and they are functions of ω. One can think of a random variable as a "measurement" of ω.

One can think of a random variable X as described by a piece of wire extending from $-\infty$ to ∞ with total weight 1 and a certain weight distribution, and the probability that X takes a value in a set $A \subset \mathbb{R}$ is the weight of the set A. So the problem is how to characterize a weight distribution on a piece of wire. In the rest of the chapter, we will see various such characterizations.

3.1 Probability Mass Function

A random variable X is called *discrete* if the set of all values taken by X is discrete (finite or countably infinite). Such a random variable is characterized by a probability mass function (pmf) which gives the probability of occurrence of each possible value of $X(\omega)$. When there is no ambiguity, we will write $P(X = i)$ as p_i. For example, in Example 3.1, we have

$$\begin{aligned} p_0 &= P(X = 0) \\ &= P(\{\omega : X(\omega) = 0\}) \\ &= P(\{\omega : \ \omega = 1 \text{ or } \omega = 3\}) \\ &= P(\{1, 3\}) \\ &= P(\{1\} \cup \{3\}) \\ &= P(\{1\}) + P(\{3\}) \quad \text{by A3} \\ &= 0.1 + 0.1 \\ &= 0.2. \end{aligned}$$

We usually do not go for such level of formality, but the above shows what

$$P(X = 0) = P(\omega = 1) + P(\omega = 3)$$

exactly means.

Theorem 3.5. *A pmf $\{p_i\}$ satisfies the following two properties:*

1. $p_i \geq 0$ *for all* i;
2. $\sum_i p_i = 1$.

Proof The proof is left as an exercise.

Example 3.6 (Binomial Distribution). The *binomial distribution* $\mathcal{B}(n, p)$ with parameters n and p takes integer values between 0 and n, with

$$p_i = \binom{n}{i} p^i q^{n-i}$$

for $0 \leq i \leq n$, and $p_i = 0$ otherwise, where $0 \leq p \leq 1$ and $q = 1 - p$. It is obvious that $p_i \geq 0$. Moreover,

$$\begin{aligned} \sum_i p_i &= \sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} \\ &= (p+q)^n \\ &= 1^n \\ &= 1, \end{aligned}$$

where we have invoked the binomial formula

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}.$$

It can be shown that p_i peaks at $i \approx np$.

Example 3.7 (Poisson Distribution). The *Poisson distribution* with parameter λ for $\lambda \geq 0$ is defined by

$$p_k = \begin{cases} \frac{e^{-\lambda}\lambda^k}{k!} & \text{if } k \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} \sum_k p_k &= \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} e^{\lambda} \\ &= 1. \end{aligned}$$

3.2 Cumulative Distribution Function

We already have seen the pmf characterization of a discrete random variable. In this section, we introduce a more powerful characterization called the *cumulative distribution function* (CDF), which can characterize any random variable. The CDF of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P(X \leq x)$$

for all $-\infty < x < \infty$. $F_X(x)$ will be written as $F(x)$ when there is no ambiguity.

Theorem 3.8. *A CDF $F(x)$ satisfies the following properties:*

1. *$F(x)$ is non-decreasing and right-continuous;*
2. *$F(-\infty) = 0$;*
3. *$F(\infty) = 1$.*

Proof The first property results from the definition of $F(x)$. To prove the second property, we consider

$$\begin{aligned} F(-\infty) &= P(X \leq -\infty) \\ &= P(\{\omega : X(\omega) \leq -\infty\}) \\ &= P(\emptyset) \\ &= 0. \end{aligned}$$

To prove the third property, we consider

$$\begin{aligned} F(\infty) &= P(X \leq \infty) \\ &= P(\{\omega : X(\omega) \leq \infty\}) \\ &= P(\Omega) \\ &= 1. \end{aligned}$$

Basically, $F(x)$ gives the weight of the left part of the piece of wire in discussion up to and including the point x. From $F(x)$, for any interval $(a, b]$, we can determine $P(X \in (a, b])$ from

$$P(X \in (a, b]) = F(b) - F(a)$$

and hence the probability that X takes a value in any union of intervals by A3. Thus the CDF completely characterizes the weight distribution of the piece of wire and hence the random variable X!

If X is a discrete random variable taking integer values, then

$$p_i = F(i) - F(i-1)$$

and

$$F(i) = \sum_{j \le i} p_j.$$

Thus the pmf $\{p_i\}$ and the CDF $F(x)$ completely specify each other.

According to the CDF, a random variable can be classified as follows:

1. *Discrete* – if the CDF has increments only at discrete jumps;
2. *Continuous* – if the CDF has only continuous increment;
3. *Mixed* – otherwise.

Example 3.9 (Exponential Distribution). For an exponential distribution with parameter λ, denoted by $\mathcal{E}(\lambda)$,

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

Figure 3.1 is a sketch of the CDF of an exponential distribution.

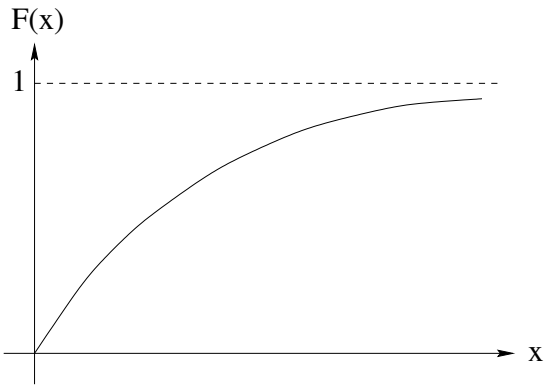


Fig. 3.1: The CDF of an exponential distribution.

Example 3.10 (Uniform Distribution). For the uniform distribution on $[0, 1]$,

$$F(x) = \begin{cases} 0 \text{ if } x < 0 \\ x \text{ if } 0 \le x \le 1 \\ 1 \text{ if } x > 1. \end{cases}$$

Figure 3.2 is a sketch of the CDF of this distribution.

Example 3.11 (Mixed Distribution). For the distribution uniform on $[0, 1]$ with a point probability mass $\frac{2}{3}$ at $x = 1$, the CDF is given by

$$F(x) = \begin{cases} 0 \text{ if } x < 0 \\ \frac{x}{3} \text{ if } 0 \le x < 1 \\ 1 \text{ if } x \ge 1. \end{cases}$$

Figure 3.3 is a sketch of the CDF of this distribution.

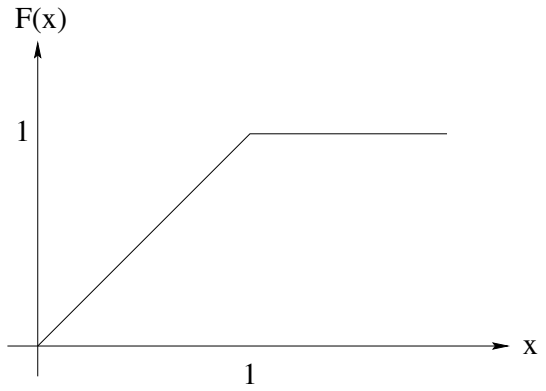


Fig. 3.2: The CDF of the uniform distribution on $[0, 1]$.

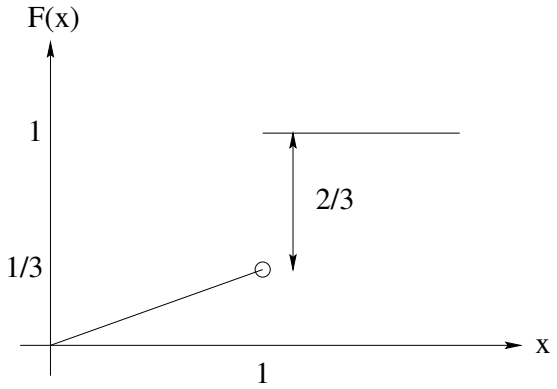


Fig. 3.3: The CDF of the mixed distribution in Example 3.11.

3.3 Probability Density Function

In the last section, we have seen how we can characterize the weight distribution on of a piece of wire by the CDF. In this section, we will see how we can characterize the same thing by means of a "density".

Suppose we say that the density of the piece of wire in a neighborhood of x is equal to a, with the unit kg/m. What it means is that for a segment of wire around x with length Δx, the weight is approximately $a \cdot \Delta x$. Let $f(x)$ denotes the density at x for a piece of wire whose weight distribution is described by a CDF $F(x)$. Then what we want for $f(x)$ is that

$$F(x) = \int_{-\infty}^{x} f(u)du \tag{3.1}$$

for all x (this is exactly what density means). Assuming that $F(x)$ is differentiable (with respective to x) and using the *fundamental theorem of calculus*[1], by differentiating both sides of the above, we have

$$\frac{dF(x)}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f(u)du = f(x),$$

giving

$$f(x) = \frac{dF(x)}{dx}. \tag{3.2}$$

The function $f(x)$ is called the *probability density function* (pdf).

If $F(x)$ is not differentiable at a certain x_0, then we cannot define $f(x)$ by (3.2). This can happen if $F(x)$ does not change smoothly at x_0, for example, at $x = 0$ in Figure 3.2. However, a careful examination of (3.1) reveals that the density $f(x)$ at any *isolated* point, as long as it is finite, does not affect the CDF $F(x)$. Therefore, in such situations, we can simply set $f(x)$ to be any finite value.

In fact, if $f(x)$ is finite, then

$$P(X = x) = 0.$$

This can be seen as follows. Let A_Δ be an interval with width Δ such that $x \in A_\Delta$. Then

$$0 \le P(X = x) \le P(X \in A_\Delta) \approx f(x)\Delta.$$

Letting $\Delta \to 0$, we see that $P(X = x) = 0$. However, although the event $\{\omega : X(\omega) = x\}$ has zero probability measure, it is *not* impossible.

Example 3.12. Suppose we throw a dart at a target and the point where the dart lands distributes uniformly on the target. Then the probability that the dart lands on any particular point is zero, but it is not impossible for the dart to land on that point.

---

[1] The fundamental theorem of calculus states that

$$\frac{d}{dx}\int_{c}^{x} f(u)du = f(x).$$

This can be seen by considering

$$\begin{aligned}\frac{d}{dx}\int_{c}^{x} f(u)du &= \lim_{h\to 0}\frac{1}{h}\left[\int_{c}^{x+h} f(u)du - \int_{c}^{x} f(u)du\right] \\ &= \lim_{h\to 0}\frac{1}{h}\int_{x}^{x+h} f(u)du \\ &= f(x)\end{aligned}$$

if f is sufficiently smooth at x.

Example 3.13 (Exponential Distribution). For an exponential distribution with parameter λ,

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3.4 is a sketch of the pdf of an exponential distribution.

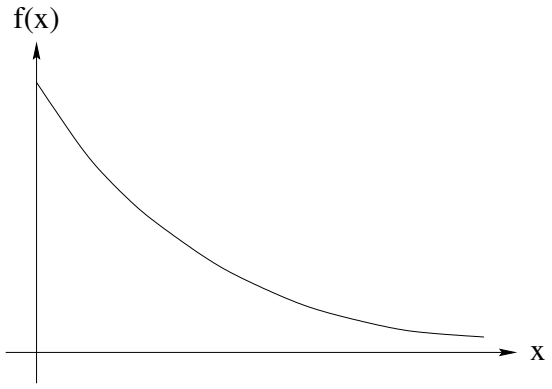


Fig. 3.4: The pdf of an exponential distribution.

Example 3.14 (Uniform Distribution). For the uniform distribution on $[0, 1]$,

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

Figure 3.5 is a sketch of the pdf of this distribution.

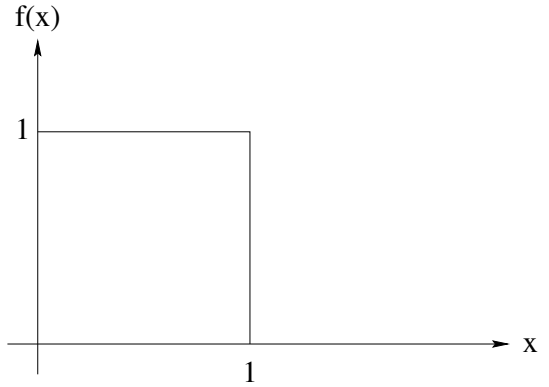


Fig. 3.5: The pdf of the uniform distribution on $[0, 1]$.

However, we do run into a serious technical difficulty in defining the density function $f(x)$ if $F(x)$ has a discrete jump at x_0, i.e., there is a point mass at x_0. The problem we face is how to define $f(x)$ in view of (3.1) so that $F(x)$ changes abruptly as x changes from $x_0 - \epsilon$ to x_0 for small $\epsilon > 0$. In fact, no such function $f(x)$ exists. Instead, we introduce the notion of a *Dirac delta function*.[2] A delta function is called a *generalized function*, which is not a true function. Generalized function theory is a very deep theory, but it suffices for us to understand what a delta function means.

The best way to think of a delta function, denoted by $\delta(x)$, is that it is function which takes the value 0 everywhere except around $x = 0$. Around $x = 0$, the function is a very sharp peak such that the total area under $\delta(x)$ is equal to 1. That is

$$\int_{-\infty}^{\infty} \delta(x)dx = \lim_{\epsilon \to 0} \int_{-\epsilon}^{\epsilon} \delta(x)dx = 1.$$

By means of the delta function, we will be able to describe a CDF $F(x)$ with discrete jumps. This will be illustrated in the next example.

Example 3.15 (Mixed Distribution). For the distribution uniform on $[0, 1]$ with a point probability mass $\frac{2}{3}$ at $x = 1$, the pdf is sketched in Figure 3.6. Here, $\frac{2}{3}\delta(x-1)$ denotes a delta scaled to two-third of $\delta(x)$ and relocated at $x = 1$.

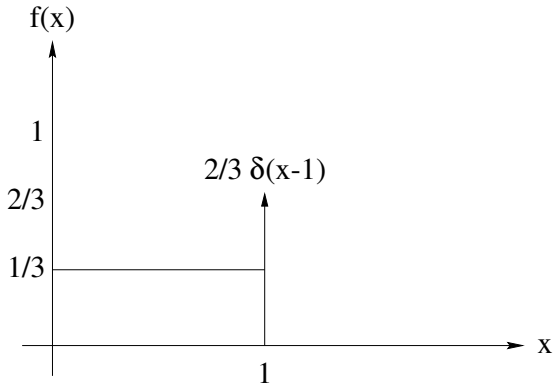


Fig. 3.6: The pdf of the mixed distribution in Example 3.11.

Theorem 3.16. *A pdf $f(x)$ satisfies the following properties:*

1. $f(x) \geq 0$;
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

[2] named after the great physicist Paul Dirac.

Proof First, $f(x)$ is nonnegative because $F(x)$ is non-decreasing. Second,

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

3.4 Function of a Random Variable

Suppose we have a random variable X specified by a given distribution (pdf or CDF), and we pass X through a function g to obtain a random variable $Y = g(X)$. What is the distribution of Y? Note that Y, as a random variable, is also a function of ω because $Y = g(X(\omega))$.

This problem is rather trivial if both X and Y are discrete. To be specific, for each $y \in \mathcal{Y}$, $P(Y = y)$ can be determined by

$$P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x),$$

where $g^{-1}(y)$ is the inverse image of y under the mapping g. For continuous and mixed random variables, the idea is similar but the technicality is a little bit more involved.

Let us consider the case that both X and Y are continuous, and we will determine $f_Y(y)$ in terms of $f_X(x)$. To start with, we further assume that g is monotone so that the inverse function g^{-1} is defined. Toward this end, in light of Figure 3.7, consider

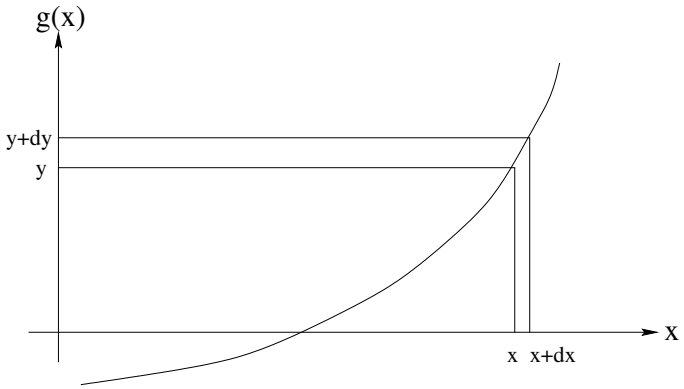


Fig. 3.7: A monotone function g.

$$P(y < Y < y + dy) = P(x < X < x + dx),$$

where $y = g(x)$, or $x = g^{-1}(y)$. Since

$$P(y < Y < y + dy) \approx f_Y(y)|dy|$$

and

$$P(x < X < x + dx) \approx f_X(x)|dx|,$$

we have

$$f_Y(y)|dy| \approx f_X(x)|dx|$$
$$f_Y(y) \approx \frac{f_X(x)}{\frac{|dy|}{|dx|}},$$

and in the limit, we have

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|} \Big|_{x=g^{-1}(y)} .$$

Example 3.17 (Linear Transformation). Let $g(x) = ax + b$, where $a \neq 0$. Then

$$g'(x) = a$$

and

$$x = g^{-1}(y) = \frac{y - b}{a}.$$

Therefore,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y - b}{a}\right).$$

Example 3.18. We give an alternative solution to the problem in the last example. We consider two cases for a. First, for $a > 0$, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P\left(X \leq \frac{y - b}{a}\right) \\ &= F_X\left(\frac{y - b}{a}\right). \end{aligned}$$

Then

$$f_Y(y) = \frac{d}{dy} F_X\left(\frac{y - b}{a}\right) = \frac{1}{a} f_X\left(\frac{y - b}{a}\right). \tag{3.3}$$

For $a < 0$, we have

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P\left(X \ge \frac{y-b}{a}\right) \\ &= 1 - P\left(X < \frac{y-b}{a}\right) \\ &= 1 - \left[P\left(X \le \frac{y-b}{a}\right) - P\left(X = \frac{y-b}{a}\right)\right] \\ &= 1 - \left[F_X\left(\frac{y-b}{a}\right) - 0\right] \\ &= 1 - F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

Upon differentiating with respect to y, we obtain

$$f_Y(y) = -\frac{1}{a} f_X\left(\frac{y-b}{a}\right). \tag{3.4}$$

Combining (3.3) and (3.4) for the two cases, we have

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right),$$

which is the same as what we have obtained before.

We now turn to the case that g is not monotone, i.e., $y = g(x)$ has possibly more than one solution for some y. We will discuss the case that for a particular y, there exist x_i, $i = 1, 2$, such that $y = g(x_i)$ (cf. Figure 3.8). The general

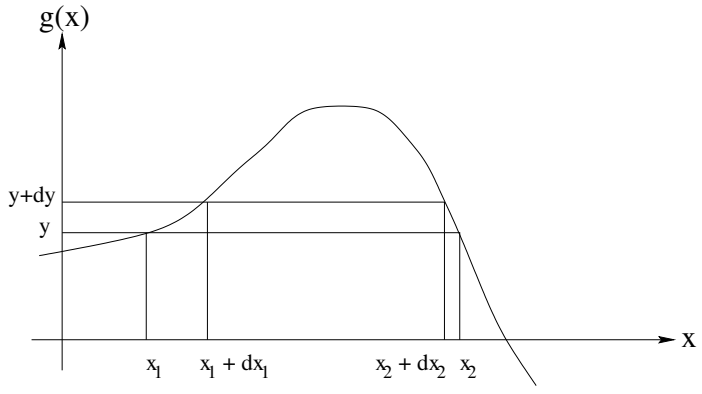


Fig. 3.8: A function g with two inverses x_1 and x_2 at y.

case is a straightforward extension. From Figure 3.8, we see that

$$\{y < Y < y + dy\} = \{x_1 < X < x_1 + dx_1\} \cup \{x_2 < X < x_2 + dx_2\}.$$

Then

$$P(y < Y < y + dy) = P(x_1 < X < x_1 + dx_1) + P(x_2 < X < x_2 + dx_2).$$

Following exactly the argument as we used for the case that g is monotone, we obtain

$$f_Y(y) = \frac{f_X(x_1)}{|g'(x_1)|} + \frac{f_X(x_2)}{|g'(x_2)|}.$$

Example 3.19. Let $g(x) = a(x-c)^2 + b$, $a > 0$. Now values of y less than b are impossible, so that $f_Y(y) = 0$ for $y \leq b$. For $y > b$, we have

$$x_1 = c - \sqrt{\frac{y-b}{a}}$$
$$x_2 = c + \sqrt{\frac{y-b}{a}}.$$

Since $g'(x) = 2a(x-c)$, we have

$$f_Y(y) = \frac{1}{2\sqrt{a(y-b)}} \left[f_X\left(c - \sqrt{\frac{y-b}{a}}\right) + f_X\left(c + \sqrt{\frac{y-b}{a}}\right) \right].$$

3.5 Expectation

Definition 3.20. *The expectation of X, written as $E[X]$ or simply EX, is defined as*

$$EX = \sum_{x \in \mathcal{X}} xp(x)$$

for X discrete, and

$$EX = \int_{-\infty}^{\infty} x f(x) dx$$

for X continuous. EX is sometimes written as $\bar{x}$.

The expectation of a random variable X is the value taken by X *on the average*. It is also called the *mean* and it can be regarded as the *center of mass* of the piece of wire describing the distribution of X. Specifically, the location of the center of mass of the piece of wire is given by

$$\frac{\sum_{x \in \mathcal{X}} xp(x)}{\sum_{x \in \mathcal{X}} p(x)} = \sum_{x \in \mathcal{X}} xp(x) = EX.$$

Proposition 3.21. *If $X = c$ with probability 1, then $EX = c$.*

Proof This can be proved by considering

$$\begin{aligned} EX &= cP(X = c) \\ &= c \cdot 1 \\ &= c. \end{aligned}$$

Example 3.22 (Binomial Distribution). Let $X \sim \mathcal{B}(n, p)$. Then

$$EX = \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k}.$$

This summation can be evaluated explicitly by considering the binomial formula:

$$(a+b)^n \equiv \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Here we fix b and regard each side as a function of a, so that the left hand side and the right hand side are identical functions of a. Differentiating with respect to a, we have

$$n(a+b)^{n-1} \equiv \sum_{k=0}^{n} \binom{n}{k} k a^{k-1} b^{n-k},$$

and multiplying by a gives

$$na(a+b)^{n-1} \equiv \sum_{k=0}^{n} \binom{n}{k} k a^k b^{n-k}.$$

Then let $a = p$ and $b = 1 - p$ to obtain

$$EX = np.$$

Example 3.23 (Exponential Distribution). Let $X \sim \mathcal{E}(\lambda)$. Then

$$EX = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx.$$

Let $u = \lambda x$ and $dv = e^{-\lambda x}$. Then $du = \lambda dx$ and $v = -\frac{1}{\lambda} e^{-\lambda x}$. Therefore,

$$\begin{aligned} EX &= -xe^{-\lambda x} \Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx \\ &= -(0-0) - \frac{e^{-\lambda x}}{\lambda} \Big|_0^\infty \\ &= -\frac{1}{\lambda}(0-1) \\ &= \frac{1}{\lambda}. \end{aligned}$$

In evaluating $\lim_{x\to\infty} xe^{-\lambda x}$, we use L'Hospital's Rule as follows.

$$\begin{aligned}\lim_{x\to\infty} xe^{-\lambda x} &= \lim_{x\to\infty} \frac{x}{e^{\lambda x}} \quad \left(= \frac{\infty}{\infty}\right) \\ &= \lim_{x\to\infty} \frac{\frac{d}{dx}x}{\frac{d}{dx}e^{\lambda x}} \\ &= \lim_{x\to\infty} \frac{1}{\lambda e^{\lambda x}} \\ &= 0.\end{aligned}$$

The expectation of a function of a random variable is defined in the natural way.

Definition 3.24. *The expectation of a function g of a random variable X is defined by*

$$Eg(X) = \sum_{x\in\mathcal{X}} g(x)p(x)$$

for X discrete, and

$$Eg(X) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

for X continuous.

Proposition 3.25. *For any constant c, $E[c] = c$.*

Proof This can be proved by considering

$$\begin{aligned}E[c] &= \int cf(x)dx \\ &= c\int f(x)dx \\ &= c\cdot 1 \\ &= c.\end{aligned}$$

3.6 Moments

In Definition 3.24, by letting $g(X) = X^n$ and $g(X) = (X - EX)^n$, we obtain the nth *moment* and the nth *central moment* of X, respectively, where $n \geq 1$. The nth moment and the nth central moment are denoted by m_n and μ_n, respectively, i.e.,

$$\begin{aligned}m_n &= EX^n \\ \mu_n &= E[(X - EX)^n].\end{aligned}$$

(Note that in the definition of μ_n, EX is just a constant.) The first moment, i.e., m_1, is called the mean (or expectation), while the second central moment, i.e., μ_2, is called the *variance*. The variance of X is also denoted by $\mathrm{var}X$.

Proposition 3.26. $\mu_1 = 0$.

Proof First,

$$\mu_1 = E[X - EX] = \int (x - EX) f(x) dx.$$

Regarding the distribution of X as represented by a piece of wire, the quantity $(x - EX)f(x)dx$ is the "moment" (as in mechanics) about the center of mass (at position EX) due to the weight $f(x)dx$ at position x. So this proposition says that for any weight distribution, the total moment about the center of mass is equal to 0.

For X continuous, it suffices to consider

$$\begin{aligned}
\mu_1 &= E[X - EX] \\
&= \int (x - EX) f(x) dx \\
&= \int x f(x) dx - EX \int f(x) dx \\
&= EX - EX \\
&= 0.
\end{aligned}$$

The proof for X discrete is similar.

For each function g, $Eg(X)$ give partial information about the distribution of X. For example, EX gives the average value taken by X. Likewise, each m_n and μ_n gives some partial information about the distribution of X. However, we point out that under many situations, the distribution of X is completely specified by $\{m_n\}$, i.e., the set of all moments of X gives complete information about the distribution of X. It can also be shown that $\{m_n\}$ is completely specified by μ_n, $n = 2, 3, \ldots$ together with m_1, and vice versa. The following theorem is a special case of this fact.

Theorem 3.27. $\mathrm{var}X = EX^2 - (EX)^2$.

Proof The theorem is proved by considering

$$\begin{aligned}
\mathrm{var}X &= \int (x - EX)^2 f(x) dx \\
&= \int (x^2 - 2x(EX) + (EX)^2) f(x) dx \\
&= \int x^2 f(x) dx - 2(EX) \int x f(x) dx + (EX)^2 \int f(x) dx \\
&= EX^2 - 2(EX)^2 + (EX)^2 \\
&= EX^2 - (EX)^2.
\end{aligned}$$

Corollary 3.28. *$EX^2 \geq (EX)^2$ with equality if and only if $X = EX$ with probability 1.*

Proof Since $(x - EX)^2 \geq 0$ for all x,

$$\text{var}X = \int (x - EX)^2 f(x) dx \geq 0. \tag{3.5}$$

Therefore,

$$EX^2 - (EX)^2 = \text{var}X \geq 0.$$

We now show that the above inequality is tight if and only if $P(X = EX) = 1$. If $P(X = EX) = 1$, then obviously $\text{var}X = 0$. The proof of the converse, however, is much more technical and is deferred to Appendix 3.7.

Example 3.29 (Gaussian Distribution). The Gaussian distribution $\mathcal{N}(m, \sigma^2)$ is a continuous distribution defined by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

for $-\infty < x < \infty$. It is one of the most important distributions in probability theory. It will be discussed in depth when we discuss the *central limit theorem*.

It can be shown that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = 1.$$

Furthermore, it can be shown that the mean and variance $\mathcal{N}(m, \sigma^2)$ are respectively m and σ^2.

3.7 Jensen's inequality

Definition 3.30. *Let $\{x_i\}$ be a set of vectors and $\{\alpha_i\}$ be a set of real numbers. Then*

$$y = \sum_i \alpha_i x_i$$

is called a linear combination *of $\{x_i\}$. If in addition $\{\alpha_i\}$ satisfies*

1. *$\alpha_i \geq 0$ for all i, and*
2. *$\sum_i \alpha_i = 1$,*

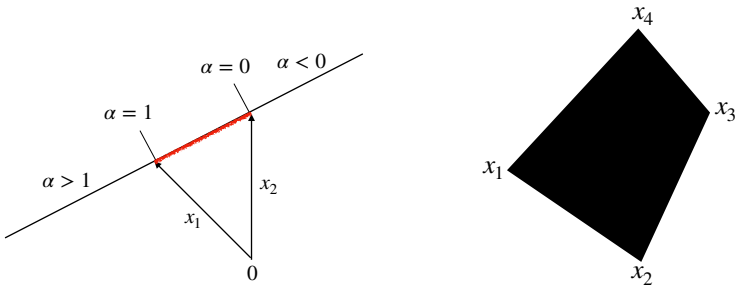
then y is called a convex combination *of $\{x_i\}$.*

Remark If $\{\alpha_i\}$ satisfies $\alpha_i \geq 0$ for all i and $\sum_i \alpha_i = 1$, then $\{\alpha_i\}$ is a probability mass function. In this course, the vectors x_i are usually scalars. In particular,

$$EX = \sum_x xp(x)$$

is a convex combination of all the elements of $\mathcal{X}$.

Figure 3.9(a) is an illustration of a linear (convex) combination $\alpha x_1 + (1-\alpha)x_2$ of two vectors x_1 and x_2, and Figure 3.9(b) is an illustration of the set of all convex combinations of the four vectors x_1, x_2, x_3, and x_4.



(a) (b)

Fig. 3.9: Linear combinations and convex combinations.

Definition 3.31 (convex set). *A set S is convex if for any $x_1, x_2 \in S$,*

$$(\alpha x_1 + (1-\alpha)x_2) \in S$$

for any $0 \leq \alpha \leq 1$.

Definition 3.32 (convex function). *A function g defined on a convex set S is called convex if for any $x_1, x_2 \in S$ and any $0 \leq \alpha \leq 1$,*

$$g(\alpha x_1 + (1-\alpha)x_2) \leq \alpha g(x_1) + (1-\alpha)g(x_2). \tag{3.6}$$

A function g is concave if and only if $-g$ is convex.

In the definition of a convex function, it is crucial that the set S is convex because otherwise for $x_1, x_2 \in S$ and $0 \leq \alpha \leq 1$, $\alpha x_1 + (1-\alpha)x_2$ may not be in S. If so, g is not defined at $\alpha x_1 + (1-\alpha)x_2$ and (3.6) cannot even be stated.

Figure 3.10 is an illustration of a convex function when S is some convex subset of the real line (i.e., an interval). In general, S is some convex subset of $\mathbb{R}^n$.

Definition 3.33 (convex hull). *For any set $S \subset \mathbb{R}^n$, its convex hull is defined as*

$$\begin{aligned}\text{con}(S) = \{&x \in \mathbb{R}^n : x = \alpha x_1 + (1-\alpha)x_2 \text{ for some } x_1, x_2 \in S \\ &\text{and } 0 \leq \alpha \leq 1\}.\end{aligned}$$

In words, $\text{con}(S)$ *consists of the convex combination of any pair of vectors in S.*

Example 3.34. The convex hull of a set S is convex. Figure 3.9(b) shows the convex hull of the set $\{x_1, x_2, x_3, x_4\}$.

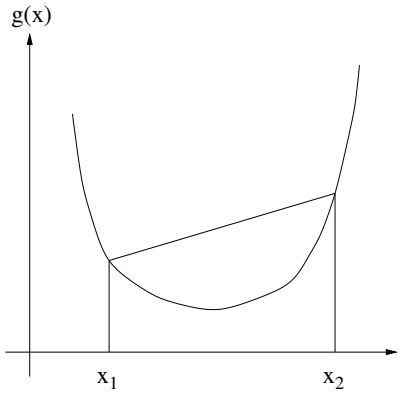


Fig. 3.10: A convex function.

Theorem 3.35 (Jensen's inequality). *Let g be a convex function. Then $Eg(X) \geq g(EX)$.*

Proof

We give a proof for the case when X is a continuous random variable and denote the pdf of X by $f(x)$. Consider Fig. 3.11. Let $L(x) = a + bx$ be a tangent to the graph $g(x)$ at $x = EX$, where a and b are some real constants. Evidently, we have $g(x) \geq L(x)$ for all x, with equality at $x = EX$, i.e., $g(EX) = L(EX)$.

Now consider

$$\begin{aligned} E[g(X)] &= \int g(x)f(x)dx \\ &\geq \int L(x)f(x)dx \\ &= \int (a+bx)f(x)dx \\ &= a\int f(x)dx + b\int xf(x)dx \\ &= a + b(EX) \\ &= L(EX) \\ &= g(EX). \end{aligned}$$

Remark The computation of $Eg(X)$ requires the knowledge of the distribution of X, while the computation of $g(EX)$ requires only the knowledge of EX. Therefore, Jensen's inequality provides a lower bound on $Eg(X)$ when only EX is known.

Corollary 3.36. *Let g be a concave function. Then $Eg(X) \leq g(EX)$.*

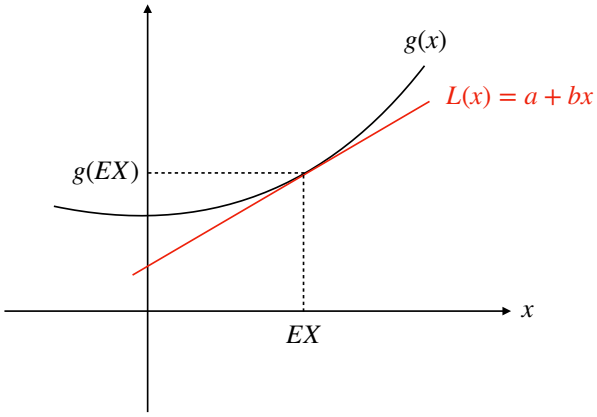


Fig. 3.11: A proof of Jensen's inequality.

Example 3.37. $g(x) = |x|$ is convex. By Jensen's inequality, $E|X| \geq |EX|$.

Example 3.38. $g(x) = x^2$ is convex. By Jensen's inequality, $EX^2 \geq (EX)^2$, i.e., $\text{var}X \geq 0$. This has been obtained in Corollary 3.28.

Appendix 3.A: Proof of Corollary 3.28

This appendix is for the more enthusiastic readers. Here, we complete the proof of Corollary 2.27 by showing that if $\text{var}X = 0$, then $P(X = EX) = 1$, or $P(X \neq X) = 0$. Assume that $\text{var}X = 0$. First, observe that

$$\{x : x \neq EX\} = \{x : (x - EX)^2 > 0\}.$$

Define the set

$$A_\epsilon = \{x : (x - EX)^2 > \epsilon^2\}.$$

Note that

$$A_0 = \{x : (x - EX)^2 > 0\} = \{x : x \neq EX\},$$

and that for $\epsilon \leq \epsilon'$, $A_{\epsilon'} \subset A_\epsilon$, so that $P(A_{\epsilon'}) \leq P(A_\epsilon)$. Our goal is to establish that $P(A_0) = 0$. We first show that for any $\epsilon > 0$, $P(A_\epsilon) = 0$. Consider

$$\begin{aligned}
0 &= \text{var}X \\
&= \int_{x\in A_\epsilon} + \int_{x\notin A_\epsilon} (x-EX)^2 f(x)dx \\
&\geq \int_{x\in A_\epsilon} (x-EX)^2 f(x)dx \\
&\geq \epsilon^2 \int_{x\in A_\epsilon} f(x)dx \\
&= \epsilon^2 P(A_\epsilon) \\
&\geq 0.
\end{aligned}$$

Thus

$$0 \leq \epsilon^2 P(A_\epsilon) \leq 0,$$

implying that $P(A_\epsilon) = 0$. Now since $P(A_0) \geq P(A_\epsilon) = 0$ for all $\epsilon > 0$, we still cannot conclude that $P(A_0) = 0$. Toward this end, by means of the *monotone convergence theorem* in real analysis (beyond the scope of it course), it can be shown that

$$\begin{aligned}
P(A_0) &= P\left(\lim_{\epsilon\to 0} A_\epsilon\right) \\
&= \lim_{\epsilon\to 0} P(A_\epsilon) \\
&= 0.
\end{aligned}$$

This completes the proof.

4

Multivariate Distributions

In Chapter 3, we have discussed a single random variable. In this and subsequent chapters, we will discuss a *finite* collection of jointly distributed random variables. In this chapter, we will discuss two jointly distributed random variables mostly, but generalization to more than two random variables is straightforward. Emphasis will be on continuous distributions.

A random variable is a function of ω. Two random variables are an ordered pair of functions of ω, for example, $(x(\omega), y(\omega))$. For one random variable, we think of its distribution as the weight distribution of a piece of wire whose total weight is 1. For two random variables, we think of their joint distribution as the weight distribution of a sheet whose total weight is 1.

As for a single random variable, we will have both CDF and pdf characterizations of the joint distribution.

4.1 Joint Cumulative Distribution Function

Let X and Y be two jointly distributed continuous random variables. The joint CDF of X and Y is defined by

$$F_{XY}(x, y) = P(X \leq x, Y \leq y).$$

When there is no ambiguity, $F_{XY}(x, y)$ will be abbreviated as $F(x, y)$.

Theorem 4.1. *A joint CDF $F(x, y)$ satisfies the following properties:*

1. $F(x, y) \geq 0$;
2. $F(-\infty, -\infty) = 0$, $F(x, -\infty) = 0$, *and* $F(-\infty, y) = 0$, *for any* $-\infty < x, y < \infty$;
3. $F(\infty, \infty) = 1$;
4. *For any* $x_1 < x_2$ *and* $y_1 < y_2$,

$$F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \geq 0.$$

Proof Property 1 is proved by noting

$$F(x, y) = P(X \leq x, Y \leq y) \geq 0.$$

Property 2 is proved by considering

$$\begin{aligned} F(-\infty, -\infty) &= P(X \leq -\infty, Y \leq -\infty) \\ &= P(\{\omega : X(\omega) \leq -\infty, Y(\omega) \leq -\infty\}) \\ &= P(\emptyset) \\ &= 0. \end{aligned}$$

That $F(x, -\infty) = 0$ and $F(-\infty, y)$ can be proved likewise. The proof for Property 3 is similar to that for $F(\infty) = 1$ for the single variable case, so it is omitted. Property 4 can be proved by considering

$$\begin{aligned} &F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) \\ &= [F(x_2, y_2) - F(x_1, y_2)] - [F(x_2, y_1) - F(x_1, y_1)] \\ &= P(x_1 < X \leq x_2, Y \leq y_2) - P(x_1 < X \leq x_2, Y \leq y_1) \\ &= P(x_1 < X \leq x_2, y_1 < Y \leq y_2) \qquad (4.1) \\ &\geq 0. \end{aligned}$$

Letting $y_1 = -\infty$ and $y_2 = y$ in Property 4, we have

$$F(x_2, y) - F(x_1, y) - F(x_2, -\infty) + F(x_1, -\infty) \geq 0,$$

which implies

$$\begin{aligned} F(x_2, y) - F(x_1, y) &\geq F(x_2, -\infty) - F(x_1, -\infty) \\ &= 0 - 0 \\ &= 0, \end{aligned}$$

where we have invoked Property 2. Similarly, we can show that for $x_1 < x_2$,

$$F(x_2, y) \geq F(x_1, y).$$

Therefore, $F(x, y)$ is non-decreasing in both x and y, but this is weaker than Property 4 as we will see in the next example.

Example 4.2. Let

$$\begin{aligned} F(x_1, y_1) &= 4 \\ F(x_1, y_2) &= 7 \\ F(x_2, y_1) &= 8 \\ F(x_2, y_2) &= 10. \end{aligned}$$

Then we see that

$$\begin{aligned} F(x_2, y_2) &> F(x_2, y_1) \\ F(x_2, y_2) &> F(x_1, y_2) \\ F(x_2, y_1) &> F(x_1, y_1) \\ F(x_1, y_2) &> F(x_1, y_1). \end{aligned}$$

However,

$$F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1) = 10 - 7 - 8 + 4 = -1.$$

The reader should compare Property 4 with the non-decreasing property of $F(x)$ for a single random variable. We can think of Property 4 as the "jointly non-decreasing" property.

So from (4.1), we can obtain the probability that (X, Y) is within any rectangle in $\mathbb{R}^2$ from the joint CDF F_{XY}. In principle, we can obtain $P((X, Y) \in A)$ for essentially any subset A of $\mathbb{R}^2$ by A3[1] through approximation and taking limit. Therefore, the joint CDF fully characterizes the joint distribution.

The *marginal* CDFs, namely F_X and F_Y, can readily be obtained from F_{XY}. Specifically,

$$F_X(x) = P(X \leq x) = P(X \leq x, Y \leq \infty) = F_{XY}(x, \infty).$$

Similarly,

$$F_Y(y) = F_{XY}(\infty, y).$$

4.2 Joint Probability Density Function

Our task is to find a joint pdf $f(x, y)$ (if exists) for a given joint CDF $F(x, y)$ such that

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x', y') dy' dx'. \tag{4.2}$$

Assume that $\frac{\partial^2}{\partial x \partial y} F(x, y)$ exists. Then by the fundamental theorem of calculus, we have

$$\begin{aligned} \frac{\partial^2}{\partial x \partial y} F(x, y) &= \frac{\partial}{\partial y} \frac{\partial}{\partial x} \int_{-\infty}^{x} \int_{-\infty}^{y} f(x', y') dy' dx' \\ &= \frac{\partial}{\partial y} \int_{-\infty}^{y} f(x, y') dy' \\ &= f(x, y). \end{aligned}$$

Therefore,

[1] A3 refers to the third axiom of probability.

$$f(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y). \tag{4.3}$$

Now $F(x,y)$ can be obtained from $f(x,y)$ by (4.2), and vice versa by (4.3). Hence, both $F(x,y)$ and $f(x,y)$ are complete characterizations of the joint distribution.

Theorem 4.3. *A joint pdf $f(x,y)$ satisfies the following properties:*

1. $f(x,y) \geq 0$;
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1$.

Proof First,

$$\begin{aligned}
&f(x,y)\\
&= \frac{\partial}{\partial x}\frac{\partial}{\partial y}F(x,y)\\
&= \frac{\partial}{\partial x}\left[\lim_{\Delta y\to 0}\frac{1}{\Delta y}(F(x,y+\Delta y) - F(x,y))\right]\\
&= \lim_{\Delta x\to 0}\frac{1}{\Delta x}\left[\lim_{\Delta y\to 0}\frac{1}{\Delta y}(F(x+\Delta x, y+\Delta y) - F(x+\Delta x, y))\right.\\
&\quad \left. - \lim_{\Delta y\to 0}\frac{1}{\Delta y}(F(x,y+\Delta y) - F(x,y))\right]\\
&= \lim_{\Delta x\to 0}\lim_{\Delta y\to 0}\frac{1}{\Delta x \Delta y}\left[F(x+\Delta x, y+\Delta y) - F(x+\Delta x, y)\right.\\
&\quad \left. -F(x,y+\Delta y) + F(x,y)\right].
\end{aligned}$$

By Property 4 of a joint CDF in Theorem 4.1, we see that the expression inside the square bracket above is nonnegative. It then follows that $f(x,y) \geq 0$. Second,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = F(\infty,\infty) = 1$$

by Property 2 in Theorem 4.1. The theorem is proved.

The marginal pdfs, i.e., f_x and f_Y, can also readily be obtained from f_{XY}. Specifically,

$$\begin{aligned}
f_X(x) &= \frac{d}{dx}F_X(x)\\
&= \frac{d}{dx}F_{XY}(x,\infty)\\
&= \frac{d}{dx}\int_{-\infty}^{x}\int_{-\infty}^{\infty} f_{XY}(x',y)dydx'\\
&= \int_{-\infty}^{\infty} f_{XY}(x,y)dy.
\end{aligned}$$

Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)dx. \tag{4.4}$$

Remark For discrete random variables X and Y, the analog of (4.4) is

$$P(Y = y) = \sum_x P(X = x, Y = y).$$

4.3 Independence of Random Variables

Two random variables X and Y are said to be *independent* if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \tag{4.5}$$

for all subsets A and B of $\mathbb{R}$. When both X and Y are discrete, (4.5) is equivalent to

$$P(X = i, Y = j) = P(X = i)P(Y = j) \tag{4.6}$$

for all $i \in \mathcal{X}$ and $j \in \mathcal{Y}$. However, when both X and Y are continuous, it is extremely difficult to check the condition (4.5), as it involves an uncountably many subsets of $\mathbb{R}$. Instead, we will discuss two simpler characterizations of independence of two random variables.

Theorem 4.4 (CDF Characterization). *Two random variables X and Y are independent if and only if*

$$F_{XY}(x, y) = F_X(x)F_Y(y) \tag{4.7}$$

for all x and y.

Proof If X and Y are independent, then (4.5) is satisfied for all subsets A and B of $\mathbb{R}$. Then for all x and y,

$$\begin{aligned} F_{XY}(x, y) &= P(X \le x, Y \le y) \\ &= P(X \le x)P(Y \le y) \\ &= F_X(x)F_Y(y). \end{aligned}$$

For the converse, we will only give the proof for a special case. Assume (4.7) is satisfied for all x and y, and suppose A and B are intervals, with $A = (x_1, x_2]$ and $B = (y_1, y_2]$. Then

$$\begin{aligned} &P(X \in A, Y \in B) \\ &= P(x_1 < X \le x_2, y_1 < Y \le y_2) \\ &= F_{XY}(x_2, y_2) - F_{XY}(x_1, y_2) - F_{XY}(x_2, y_1) + F_{XY}(x_1, y_1) \quad \text{by (4.1)} \\ &= F_X(x_2)F_Y(y_2) - F_X(x_1)F_Y(y_2) - F_X(x_2)F_Y(y_1) + F_X(x_1)F_Y(y_1) \\ &= (F_X(x_2) - F_X(x_1))(F_Y(y_2) - F_Y(y_1)) \\ &= P(X \in A)P(Y \in B), \end{aligned}$$

i.e., (4.5) is satisfied. Using this argument, one can prove the same result by induction for A and B being finite unions of intervals. The details are omitted.

Theorem 4.5 (pdf Characterization). *If the pdf f_{XY} for random variables X and Y exists, then X and Y are independent if and only if*

$$f_{XY}(x,y) = f_X(x)f_Y(y) \tag{4.8}$$

for all x and y.

Proof

The following apply to all x and y. Assume that $f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)$ exists, and that

$$F_{XY}(x,y) = F_X(x)F_Y(y). \tag{4.9}$$

Then

$$\begin{aligned} f_{XY}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y) \\ &= \frac{\partial}{\partial x}\frac{\partial}{\partial y} F_X(x)F_Y(y) \\ &= \frac{\partial}{\partial x}\left[F_X(x)\left(\frac{\partial F_Y(y)}{\partial y}\right)\right] \\ &= \left(\frac{\partial F_X(x)}{\partial x}\right)\left(\frac{\partial F_Y(y)}{\partial y}\right) \\ &= f_X(x)f_Y(y). \end{aligned}$$

Conversely, assume (4.8) holds. Then

$$\begin{aligned} F_{XY}(x,y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{XY}(x',y')dy'dx' \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_X(x')f_Y(y')dy'dx' \\ &= \int_{-\infty}^{x} f_X(x')\left[\int_{-\infty}^{y} f_Y(y')dy'\right]dx' \\ &= \left[\int_{-\infty}^{x} f_X(x')dx'\right]\left[\int_{-\infty}^{y} f_Y(y')dy'\right] \\ &= F_X(x)F_Y(y). \end{aligned}$$

Therefore, (4.8) and (4.9) are equivalent. The theorem is proved.

Remark For two discrete random variables X and Y, the analog of Theorem 4.5 is that X and Y are independent if and only if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

for all x and y.

For three or more random variables, like events, we distinguish between *pairwise independence* and *mutual independence*.

Definition 4.6. *Random variables $X_1, X_2, \ldots, X_n$ are mutually independent if*

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \ldots P(X_n \in A_n)$$

where A_i is a subset of $\mathbb{R}$.

Definition 4.7. *Random variables $X_1, X_2, \ldots, X_n$ are pairwise independent if X_i and X_j are independent for any $1 \leq i < j \leq n$.*

Theorem 4.8. *Random variables $X_1, X_2, \ldots, X_n$ are mutually independent if and only if*

$$F_{X_1 \ldots X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n) \tag{4.10}$$

for all $x_i, 1 \leq i \leq n$, or equivalently,

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \ldots f_{X_n}(x_n)$$

for all $x_i, 1 \leq i \leq n$ if $f_{X_1 \ldots X_n}$ exists.

Proof Omitted.

Theorem 4.9. *If $X_1, X_2, \ldots, X_n$ are mutually independent, then they are pairwise independent.*

Proof If $X_1, X_2, \ldots, X_n$ are mutually independent, then (4.10) holds. For any fixed $1 \leq i < j \leq n$, by setting $x_k = \infty$ for all $k \neq i, j$ in (4.10), the left hand side becomes $F_{X_i X_j}(x_j, x_j)$, while the right hand sides becomes $F_{X_i}(x_i) F_{X_j}(x_j)$ since $F_{X_k}(\infty) = 1$ for all $k \neq i, j$. Therefore,

$$F_{X_i X_j}(x_j, x_j) = F_{X_i}(x_i) F_{X_j}(x_j),$$

or X_i and X_j are independent. Hence, we conclude that $X_1, X_2, \ldots, X_n$ are pairwise independent.

We note that, however, pairwise independence does not imply mutual independence.

Theorem 4.10. *If $X_1, X_2, \ldots, X_n$ are mutually independent, then so are $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ for any functions g_i, $1 \leq i \leq n$.*

Proof Consider any subsets $A_1, A_2, \ldots, A_n$ of $\mathbb{R}$. Then

$$\begin{aligned} P(g_i(X_i) \in A_i, 1 \le i \le n) &= P(X_i \in g_i^{-1}(A_i), 1 \le i \le n) \\ &= \prod_{i=1}^{n} P(X_i \in g_i^{-1}(A_i)) \\ &= \prod_{i=1}^{n} P(g_i(X_i) \in A_i), \end{aligned}$$

where we have invoked the mutual independence assumption to obtain the second equality. Thus $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ are also mutually independent.

Definition 4.11. *Random variables $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) if they are mutually independent and each X_i has the same marginal distribution.*

For i.i.d. random variables $X_1, X_2, \ldots, X_n$, we can think of each of them being an identical copy of a *generic* random variable X, where the random variables are mutually independent. Very often the notation '$X_i \sim X$' is used to mean that X_i and X have the same distribution.

Example 4.12 (Binomial Distribution). Let $X_1, X_2, \ldots, X_n$ be i.i.d. binary random variables with

$$P(X_i = 0) = 1 - p \quad \text{and} \quad P(X_i = 1) = p.$$

Then $N = \sum_{i=1}^{n} X_i$ has distribution $\mathcal{B}(n, p)$, i.e.,

$$P(N = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $0 \le k \le n$.

4.4 Conditional Distribution

The distribution of a random variable X represents our knowledge about the outcome of X. This is referred to as the *a priori* distribution of X. Upon knowing the occurrence of a certain event or the outcome of another jointly distributed random variable, our knowledge of X changes accordingly. Our new knowledge about X is represented by the *conditional distribution* of X.

4.4.1 Conditioning on an Event

We first consider the distribution of a single random variable X conditioning on an event $\{X \in A\}$, where A is a subset of $\mathbb{R}$ and $P(X \in A) > 0$. Let F be the CDF of X. The CDF of X conditioning on $\{X \in A\}$ is defined by

$$F_X(x|X \in A) = P(X \le x|X \in A).$$

Then

$$\begin{aligned} F_X(x|X \in A) &= \frac{P(X \le x, X \in A)}{P(X \in A)} \\ &= \frac{P(X \in (-\infty, x] \cap A)}{P(X \in A)}. \end{aligned} \tag{4.11}$$

Note that $F_X(\cdot|X \in A)$ satisfies the three properties of a CDF in Theorem 3.8.

The pdf of X conditioning on $\{X \in A\}$ is defined by

$$f_X(x|X \in A) = \frac{d}{dx} F_X(x|X \in A).$$

It is readily seen from (4.11) that

$$f_X(x|X \in A) = \begin{cases} \frac{f_X(x)}{P(X \in A)} & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Again, note that $f_X(\cdot|X \in A)$ satisfies the two properties of a pdf in Theorem 3.16.

Example 4.13. Consider the CDF and pdf in Figure 4.1(a) and (b) for random variable X, respectively, and let A be the subset of $\mathbb{R}$ as shown in Figure 4.1(a). Then it is readily seen that

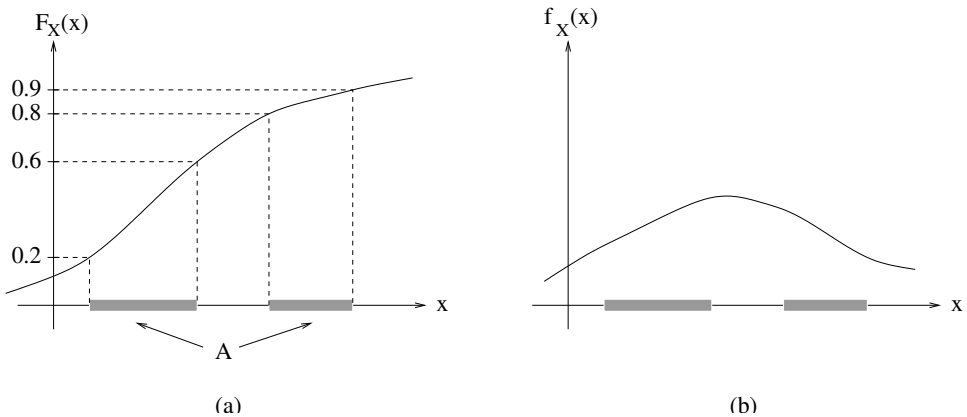


Fig. 4.1: The CDF and pdf of a distribution.

$$P(X \in A) = (0.6 - 0.2) + (0.9 - 0.8) = 0.5,$$

and $F_X(x|X \in A)$ and $f_X(x|X \in A)$ are illustrated in Figure 4.2 accordingly.

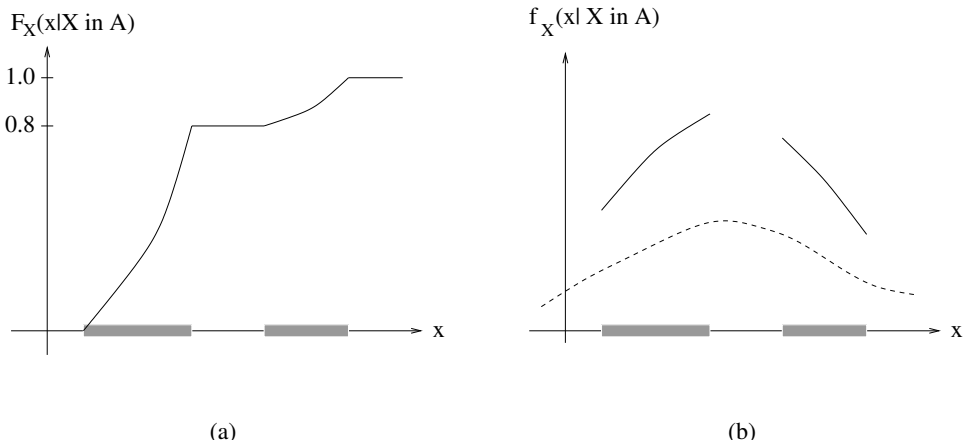


Fig. 4.2: The CDF and pdf conditioning on $\{X \in A\}$.

We now consider the joint distribution of two random variables X and Y conditioning on an event $\{(X,Y) \in A\}$, where A is a subset of $\mathbb{R}^2$. The CDF of (X,Y) conditioning on $\{(X,Y) \in A\}$ is defined by

$$F_{XY}(x,y|(X,Y) \in A) = P(X \le x, Y \le y|(X,Y) \in A). \tag{4.12}$$

This is illustrated in Figure 4.3, where

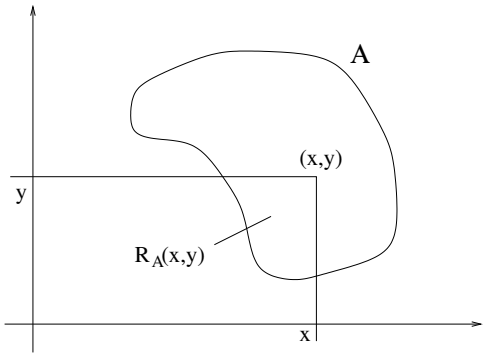


Fig. 4.3: The set $R_A(x,y)$.

$$R_A(x,y) = \{(x',y') : x' \le x \text{ and } y' \le y\} \cap A.$$

Then from (4.12), we have

$$F_{XY}(x,y|(X,Y)\in A) = \frac{P((X,Y)\in R_A(x,y))}{P((X,Y)\in A)} = \frac{\int\int_{R_A(x,y)} f_{XY}(x',y')dy'dx'}{P((X,Y)\in A)}. \tag{4.13}$$

Accordingly, the jointly pdf of (X,Y) conditioning on $\{(X,Y)\in A\}$ is defined as

$$f_{XY}(x,y|(X,Y)\in A) = \frac{\partial^2}{\partial x\partial y}F_{XY}(x,y|(X,Y)\in A). \tag{4.14}$$

It turns out that carrying out the above partial differentiation is considerably more complicated than the case when only one random variable is involved (see Appendix 4.A). So we instead will obtain $f_{XY}(x,y|(X,Y)\in A)$ by another approach. We observe that $f_{XY}(x,y|(X,Y)\in A)$ has to satisfy the following properties: For $(x,y)\notin A$,

$$f_{XY}(x,y|(X,Y)\in A) = 0,$$

and for any small subset S of A containing (x,y) with area Δ,

$$P((X,Y)\in S|(X,Y)\in A) \approx f_{XY}(x,y|(X,Y)\in A)\Delta. \tag{4.15}$$

The left hand side above can be written as

$$P((X,Y)\in S|(X,Y)\in A) = \frac{P((X,Y)\in S\cap A)}{P((X,Y)\in A)} = \frac{P((X,Y)\in S)}{P((X,Y)\in A)} \approx \frac{f_{XY}(x,y)\Delta}{P((X,Y)\in A)}. \tag{4.16}$$

Equating (4.16) and the right hand side of (4.15), we obtain

$$f_{XY}(x,y|(X,Y)\in A) = \frac{f_{XY}(x,y)}{P((X,Y)\in A)}.$$

Therefore,

$$f_{XY}(x,y|(X,Y)\in A) = \begin{cases} \frac{f_{XY}(x,y)}{P((X,Y)\in A)} & \text{if } (x,y)\in A \\ 0 & \text{if } (x,y)\notin A. \end{cases}$$

Again, we note that $f_{XY}(\cdot|(X,Y)\in A)$ satisfies the two properties of a joint pdf in Theorem 4.3.

4.4.2 Conditioning on Another Random Variable

In this subsection, we will discuss the distribution of a random variable Y conditioning on the event $\{X=x\}$, where X and Y are jointly distributed

random variables. For X continuous, $P(X = x) = 0$. Thus we are running into the problem of conditioning on an event with zero probability.

To get around this problem, instead of conditioning on $\{X = x\}$ which has zero probability, we condition on the event $\{x - \frac{\epsilon}{2} < X < x + \frac{\epsilon}{2}\}$, or equivalently the event $\{(X, Y) \in A(x, \epsilon)\}$, which has nonzero probability, where

$$A(x, \epsilon) = \left\{(x', y) : x - \frac{\epsilon}{2} < x' < x + \frac{\epsilon}{2}\right\}.$$

The set $A(x, \epsilon)$ is illustrated in Figure 4.4. Eventually, we will take the limit as $\epsilon \to 0$.

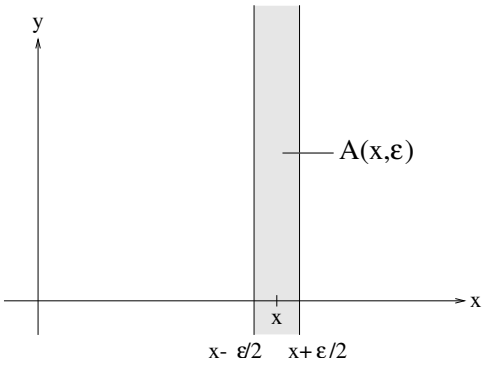


Fig. 4.4: The set $A(x, \epsilon)$.

The joint pdf of X and Y conditioning on $\{(X, Y) \in A(x, \epsilon)\}$, from the last subsection, is given by

$$f_{XY}(x', y|(X, Y) \in A(x, \epsilon)) = \begin{cases} \frac{f_{XY}(x', y)}{P((X,Y) \in A(x,\epsilon))} & \text{if } x - \frac{\epsilon}{2} < x' < x + \frac{\epsilon}{2} \\ 0 & \text{otherwise.} \end{cases}$$

By integrating over all x', we obtain the conditional marginal pdf of Y as follows.

$$\begin{aligned} & f_Y(y|(X, Y) \in A(x, \epsilon)) \\ &= \int_{-\infty}^{\infty} f_{XY}(x', y|(X, Y) \in A(x, \epsilon))dx' \\ &= \int_{x-\frac{\epsilon}{2}}^{x+\frac{\epsilon}{2}} \frac{f_{XY}(x', y)}{P((X, Y) \in A(x, \epsilon))}dx' \\ &\approx \frac{f_{XY}(x, y)\epsilon}{f_X(x)\epsilon} \\ &= \frac{f_{XY}(x, y)}{f_X(x)}. \end{aligned}$$

Assuming that $f_X(x) > 0$ and taking the limit as $\epsilon \to 0$, we define the pdf of Y conditioning on $\{X = x\}$ by

$$f_{Y|X}(y|x) = \lim_{\epsilon \to 0} f_Y(y|(X,Y) \in A(x,\epsilon)) = \frac{f_{XY}(x,y)}{f_X(x)}. \tag{4.17}$$

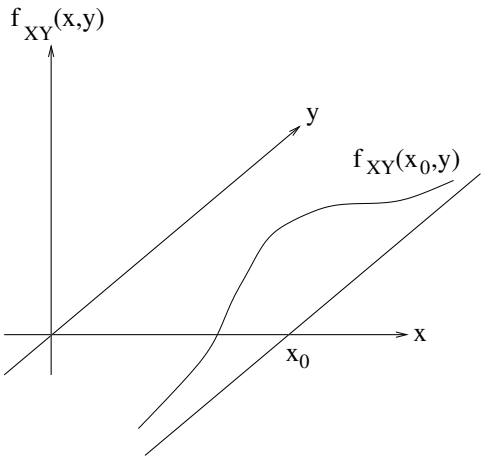


Fig. 4.5: An illustration of $f_{XY}(x_0, y)$.

Figure 4.5 is an illustration of $f_{XY}(x_0, y)$ for a fixed x_0. Thus we see from (4.17) that $f_{Y|X}(y|x)$ is just the normalized version of $f_{XY}(x, y)$ for fixed x. (For fixed x, $f_X(x)$ is a constant.) Similarly, we define

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}. \tag{4.18}$$

From (4.17) and (4.18), we have

$$f_{Y|X} = \frac{f_{X|Y} f_Y}{f_X},$$

which is a form of the Bayes Theorem.

Proposition 4.14. *If $f_{Y|X} = f_Y$, then X and Y are independent.*

Proof If $f_{Y|X} = f_Y$, then

$$\frac{f_{XY}}{f_X} = f_{Y|X} = f_Y.$$

This implies $f_{XY} = f_X f_Y$, i.e., X and Y are independent.

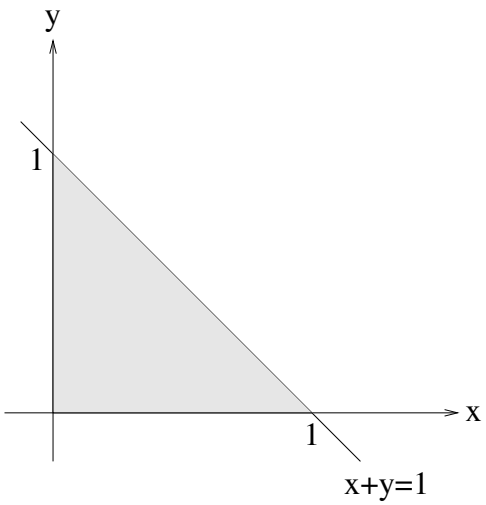


Fig. 4.6: The pdf in Example 4.15.

Example 4.15. Consider the joint pdf $f_{XY}(x, y) = cxy$ for $x, y > 0$ and $x+y < 1$ (see Figure 4.6).

We will first determine the normalizing constant c. Then we will determine the marginal pdfs f_X and f_Y, and show that X and Y are not independent. In fact, we can tell immediately that X and Y are not independent by inspecting Figure 4.6, because the possible range of Y depends on the value taken by X.

We use Property 2 of a joint pdf in Theorem 4.3 to determine c. Consider

$$\begin{aligned} 1 &= c \int_0^1 \int_0^{1-x} xy dy dx \\ &= c \int_0^1 x \frac{(1-x)^2}{2} dx \\ &= \frac{c}{2} \int_0^1 (x - 2x^2 + x^3) dx \\ &= \frac{c}{2} \left[\frac{x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{4} \right]_0^1 \\ &= \frac{c}{24}, \end{aligned}$$

which implies $c = 24$.

Now, for $0 < x < 1$,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)dy \\ &= \int_0^{1-x} cxydy \\ &= \frac{cx(1-x)^2}{2}. \end{aligned}$$

Thus

$$f_X(x) = \begin{cases} \frac{cx(1-x)^2}{2} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

By symmetry, we also have

$$f_Y(y) = \begin{cases} \frac{cy(1-y)^2}{2} & \text{if } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then for $(x,y) = (\frac{3}{4}, \frac{3}{4})$, $f_{XY}(x,y) = 0$, while $f_X(x)f_Y(y) > 0$. Therefore, X and Y are not independent.

4.5 Functions of Random Variables

Suppose we are given two random variables X and Y, and let $Z = g(X,Y)$. We are interested in the distribution of Z. Specifically, we want to determine the distribution of Z in terms of the joint distribution of X and Y.

4.5.1 The CDF Method

In this method, $F_Z(z)$ is obtained by considering

$$\begin{aligned} F_Z(z) &= P(Z \le z) \\ &= P(g(X,Y) \le z) \\ &= \int\int_{\{(x,y):g(x,y)\le z\}} f_{XY}(x,y)dydx. \end{aligned}$$

Example 4.16. Let $Z = \frac{X}{Y}$. Then

$$F_Z(z) = P(Z \le z) = P\left(\frac{X}{Y} \le z\right).$$

Figure 4.7 is an illustration of the set $\{(x,y) : \frac{x}{y} \le z\}$.

Then

$$F_Z(z) = \int_0^{\infty}\int_{-\infty}^{zy} f_{XY}(x,y)dxdy + \int_{-\infty}^{0}\int_{zy}^{\infty} f_{XY}(x,y)dxdy, \qquad (4.19)$$

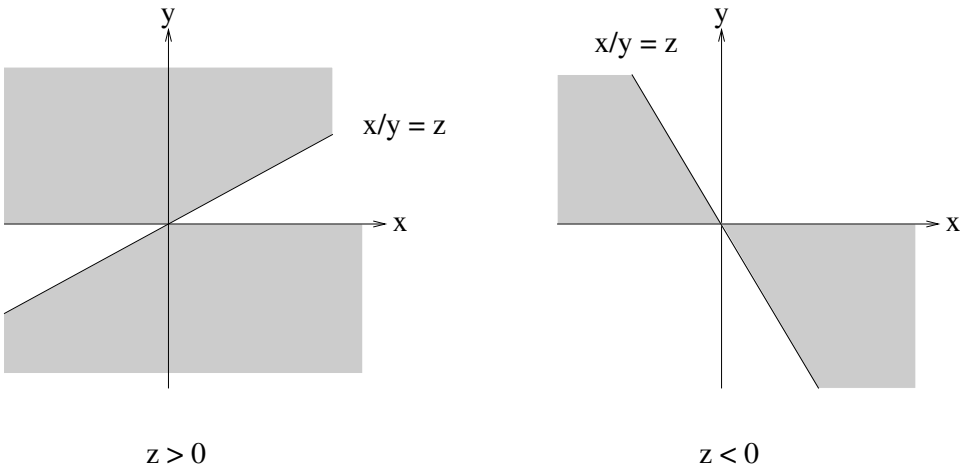


Fig. 4.7: An illustration of the set $\{(x, y) : \frac{x}{y} \leq z\}$.

and the pdf of Z is obtained by

$$f_Z(z) = \frac{d}{dz} F_Z(z).$$

In order to differentiate the double integrals in (4.19), we use the chain rule for differentiation. Specifically,

$$\begin{aligned}
&\frac{d}{dz} \int_0^\infty \int_{-\infty}^{zy} f_{XY}(x, y) dx dy \\
&= \int_0^\infty \frac{d}{dz} \int_{-\infty}^{zy} f_{XY}(x, y) dx dy \\
&= \int_0^\infty \frac{d(zy)}{dz} \frac{d}{d(zy)} \int_{-\infty}^{zy} f_{XY}(x, y) dx dy \\
&= \int_0^\infty y f_{XY}(zy, y) dy.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\frac{d}{dz} \int_{-\infty}^0 \int_{zy}^\infty f_{XY}(x, y) dx dy \\
&= \int_{-\infty}^0 \frac{d}{dz} \int_{zy}^\infty f_{XY}(x, y) dx dy \\
&= -\int_{-\infty}^0 \frac{d(zy)}{dz} \frac{d}{d(zy)} \int_{\infty}^{zy} f_{XY}(x, y) dx dy \\
&= -\int_{-\infty}^0 y f_{XY}(zy, y) dy.
\end{aligned}$$

Therefore,

$$\begin{aligned} f_Z(z) &= \int_0^\infty y f_{XY}(zy, y)dy - \int_{-\infty}^0 y f_{XY}(zy, y)dy \\ &= \int_0^\infty y f_{XY}(zy, y)dy + \int_{-\infty}^0 (-y) f_{XY}(zy, y)dy \\ &= \int_0^\infty |y| f_{XY}(zy, y)dy + \int_{-\infty}^0 |y| f_{XY}(zy, y)dy \\ &= \int_{-\infty}^\infty |y| f_{XY}(zy, y)dy. \end{aligned}$$

The next example is *extremely* important and should be studied very carefully.

Example 4.17 (Convolution). Let $Z = X+Y$, where X and Y are independent. Then

$$\begin{aligned} F_Z(z) &= P(Z \le z) \\ &= P(X + Y \le z) \\ &= \int_{-\infty}^\infty \int_{-\infty}^{z-y} f_{XY}(x, y)dxdy \\ &= \int_{-\infty}^\infty \int_{-\infty}^{z-y} f_X(x) f_Y(y)dxdy \\ &= \int_{-\infty}^\infty f_Y(y) \int_{-\infty}^{z-y} f_X(x)dxdy \\ &= \int_{-\infty}^\infty f_Y(y) F_X(z - y)dy. \end{aligned}$$

Differentiating with respect to z, we obtain

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_Z(z) \\ &= \frac{d}{dz} \int_{-\infty}^\infty f_Y(y) F_X(z - y)dy \\ &= \int_{-\infty}^\infty f_Y(y) \left[\frac{d}{dz} F_X(z - y) \right] dy \\ &= \int_{-\infty}^\infty f_Y(y) f_X(z - y)dy. \end{aligned}$$

The integral above is called the convolution of $f_X(z)$ and $f_Y(z)$, denoted by $f_X * f_Y(z)$. Since $X + Y = Y + X$, by exchanging the roles of X and Y in the above integral, we also have

$$f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx.$$

In summary, if Z is equal to $X + Y$ where X and Y are independent, then $f_Z(z) = f_X * f_Y(z)$.

4.5.2 The pdf Method

This method is to consider $f_Z(z)$ directly. Specifically,

$$\begin{aligned} f_Z(z)dz &= P(Z \in [z, z+dz)) \\ &= P(g(X,Y) \in [z, z+dz)) \\ &= \int\int_{\{(x,y): g(x,y) \in [z, z+dz)\}} f_{XY}(x,y) dxdy. \end{aligned}$$

Our task is to express the right hand side as an expression times dz, so that $F_Z(z)$ is obtained by cancelling dz on both sides. This may not always be easy to do, but we will illustrate how this can possibly be done by an example.

Example 4.18. Let X and Y be i.i.d. $\sim N(0, \sigma^2)$, so that

$$\begin{aligned} f_{XY}(x,y) &= f_X(x) f_Y(y) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} \\ &= \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \end{aligned}$$

for all x and y. Let $Z = \sqrt{X^2 + Y^2}$. Then

$$\begin{aligned} & f_Z(z)dz \\ &= \int\int_{\{(x,y): \sqrt{x^2+y^2} \in [z, z+dz)\}} f_{XY}(x,y) dxdy \\ &= \int_0^{2\pi} \int_z^{z+dz} f_{XY}(r\cos\theta, r\sin\theta) r dr d\theta \quad \text{(cf. Figure 4.8)} \\ &= \frac{1}{2\pi\sigma^2} \int_0^{2\pi} \int_z^{z+dz} e^{-\frac{r^2\cos^2\theta + r^2\sin^2\theta}{2\sigma^2}} r dr d\theta \\ &= \frac{1}{2\pi\sigma^2} \int_0^{2\pi} \int_z^{z+dz} e^{-\frac{r^2}{2\sigma^2}} r dr d\theta \\ &= \frac{1}{2\pi\sigma^2} \int_0^{2\pi} e^{-\frac{z^2}{2\sigma^2}} z dz d\theta \\ &= \frac{1}{2\pi\sigma^2} e^{-\frac{z^2}{2\sigma^2}} z dz \int_0^{2\pi} d\theta \\ &= \frac{1}{2\pi\sigma^2} e^{-\frac{z^2}{2\sigma^2}} z dz (2\pi) \\ &= \frac{1}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}} z dz. \end{aligned}$$

Cancelling dz on both sides, we have

$$f_Z(z) = \frac{z}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}}$$

for all $z \geq 0$. Obviously, $f_Z(z) = 0$ for all $z < 0$.

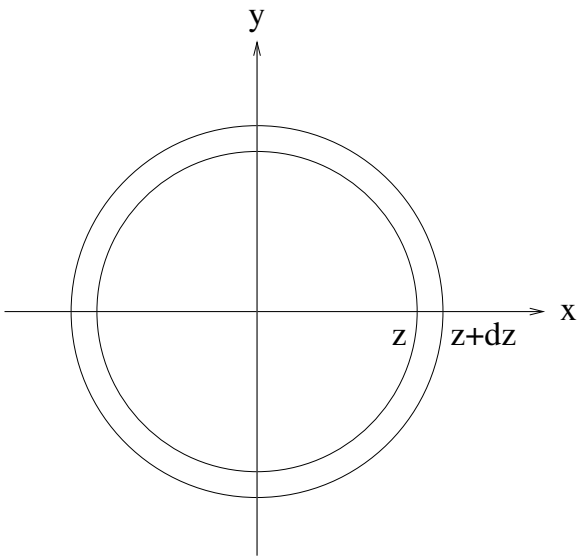


Fig. 4.8: An illustration for the range of integration in polar coordinates.

The reader is encouraged to repeat the above example by the CDF method.

4.6 Transformation of Random Variables

So far, we have discussed how we can obtained the distribution of a single random variable obtained as a function of some given random variables. In this section, we will discuss how we can obtain the joint distribution of a pair of random variables obtained as functions of a given pair of random variables.

Let $(X_1, X_2) \sim f_{X_1X_2}(x_1, x_2)$, and let (x_1, x_2) be the values taken by (X_1, X_2). Suppose for each (x_1, x_2), there is a unique pair (z_1, z_2) satisfying

$$x_1 = g_1(z_1, z_2)$$
$$x_2 = g_2(z_1, z_2)$$

for some functions g_1 and g_2. Write $\mathbf{x} = g(\mathbf{z})$, where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad \mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Then g is an *invertible* transformation and g^{-1} exists. See Figure 4.9 for an illustration of g which takes a point in the z_1-z_2 plane to the x_1-x_2 plane. Similarly, write

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}.$$

Our task is to determine $f_{Z_1 Z_2}(z_1, z_2)$, where (Z_1, Z_2) is the pair of jointly distributed random variables given by $\mathbf{Z} = g^{-1}(\mathbf{X})$.

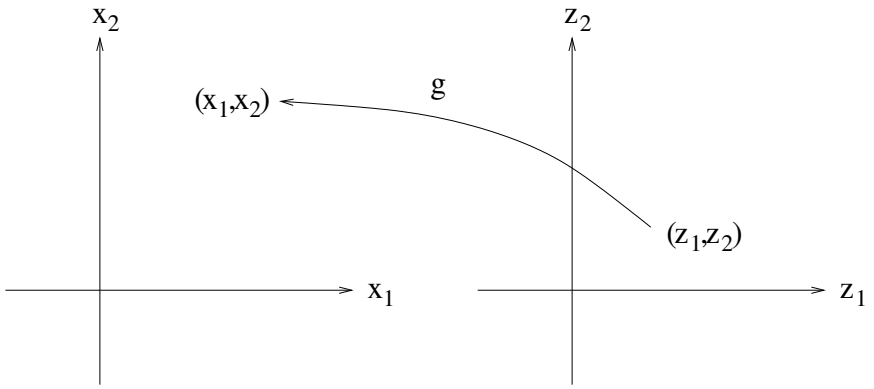


Fig. 4.9: An illustration of the function g.

Consider the rectangular element in the z_1-z_2 plane in Figure 4.10. We label the four corners of this element by $\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$, where $\mathbf{z}_0 = (z_1, z_2)$. When we trace the line from $\mathbf{z}_0$ to $\mathbf{z}_1$ in the z_1-z_2 plane, under the mapping g, the curve between $\mathbf{x}_0 = (x_1, x_2)$ and $\mathbf{x}_1$ is traced in the x_1-x_2 plane. The same applies when we trace the lines from $\mathbf{z}_0$ to $\mathbf{z}_2$, from $\mathbf{z}_1$ to $\mathbf{z}_3$, and from $\mathbf{z}_2$ to $\mathbf{z}_3$. When Δz_1 and Δz_2 are small, the region traced out in the x_1-x_2 plane can be approximated by a parallelogram. Specifically, the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are given by

$$\mathbf{v}_1 = \frac{\partial x_1}{\partial z_1}(\Delta z_1)\mathbf{i} + \frac{\partial x_2}{\partial z_1}(\Delta z_1)\mathbf{j}$$
$$\mathbf{v}_2 = \frac{\partial x_1}{\partial z_2}(\Delta z_2)\mathbf{i} + \frac{\partial x_2}{\partial z_2}(\Delta z_2)\mathbf{j}.$$

Let A_z denote the area of the rectangular element in the z_1-z_2 plane, and let A_x denote the area of the corresponding region in the x_1-x_2 plane. Now observe that (Z_1, Z_2) is in the rectangular element in the z_1-z_2 plane if and only if (X_1, X_2) is in the corresponding region in the x_1-x_2 plane. Equating

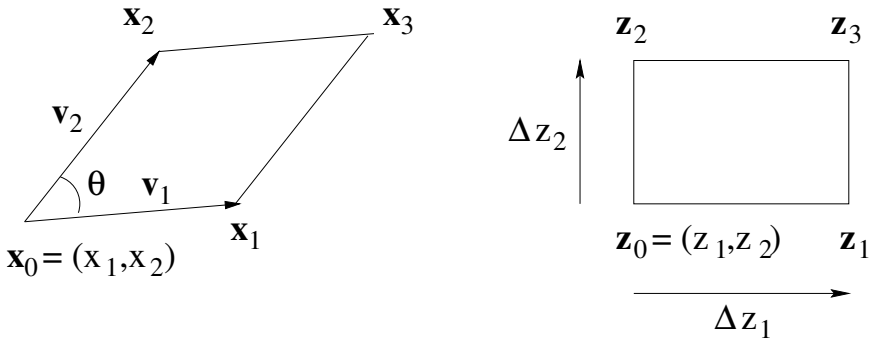


Fig. 4.10: Two corresponding regions in the x_1-x_2 plane and z_1-z_2 plane.

the probability of this event in terms of $f_{X_1X_2}(x_1, x_2)$ and $f_{Z_1Z_2}(z_1, z_2)$, we have,

$$f_{X_1X_2}(x_1, x_2)A_x \approx f_{Z_1Z_2}(z_1, z_2)A_z,$$

so that

$$f_{Z_1Z_2}(z_1, z_2) \approx f_{X_1X_2}(x_1, x_2)\frac{A_x}{A_z}, \tag{4.20}$$

where $A_z = |\Delta z_1||\Delta z_2|$. So, we need to determine the value of A_x as follows.

$$\begin{aligned}
A_x &= |\mathbf{v}_1||\mathbf{v}_2|\sin\theta \\
&= |\mathbf{v}_1 \times \mathbf{v}_2| \\
&= \left\| \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial x_1}{\partial z_1}(\Delta z_1) & \frac{\partial x_2}{\partial z_1}(\Delta z_1) & 0 \\ \frac{\partial x_1}{\partial z_2}(\Delta z_2) & \frac{\partial x_2}{\partial z_2}(\Delta z_2) & 0 \end{vmatrix} \right\| \\
&= \left| \left(\frac{\partial x_1}{\partial z_1}\frac{\partial x_2}{\partial z_2}(\Delta z_1 \Delta z_2) - \frac{\partial x_1}{\partial z_2}\frac{\partial x_2}{\partial z_1}(\Delta z_1 \Delta z_2) \right) \mathbf{k} \right| \\
&= \left\| \begin{vmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_2}{\partial z_1} \\ \frac{\partial x_1}{\partial z_2} & \frac{\partial x_2}{\partial z_2} \end{vmatrix} \right\| |\Delta z_1||\Delta z_2|.
\end{aligned}$$

Defining[2]

[2] It is easy to see that we can also write (4.21) as

$$J\begin{pmatrix} x_1 & x_2 \\ z_1 & z_2 \end{pmatrix} = \begin{vmatrix} \frac{\partial x_1}{\partial z_1} & \frac{\partial x_1}{\partial z_2} \\ \frac{\partial x_2}{\partial z_1} & \frac{\partial x_2}{\partial z_2} \end{vmatrix}.$$

$$J\begin{pmatrix} x_1 & x_2 \\ z_1 & z_2 \end{pmatrix} = \begin{vmatrix} \dfrac{\partial x_1}{\partial z_1} & \dfrac{\partial x_2}{\partial z_1} \\ \dfrac{\partial x_1}{\partial z_2} & \dfrac{\partial x_2}{\partial z_2} \end{vmatrix} \tag{4.21}$$

we have

$$A_x \approx \left| J\begin{pmatrix} x_1 & x_2 \\ z_1 & z_2 \end{pmatrix} \right| A_z.$$

Therefore, from (4.20), we have

$$f_{Z_1Z_2}(z_1, z_2) = \left| J\begin{pmatrix} x_1 & x_2 \\ z_1 & z_2 \end{pmatrix} \right| f_{X_1X_2}(x_1, x_2). \tag{4.22}$$

Exchanging the roles of (X_1, X_2) and (Z_1, Z_2) in the above derivation and accordingly in (4.22), we obtain

$$f_{X_1X_2}(x_1, x_2) = \left| J\begin{pmatrix} z_1 & z_2 \\ x_1 & x_2 \end{pmatrix} \right| f_{Z_1Z_2}(z_1, z_2). \tag{4.23}$$

Therefore, combining (4.22) and (4.23), we have

$$f_{Z_1Z_2}(z_1, z_2) = \left| J\begin{pmatrix} x_1 & x_2 \\ z_1 & z_2 \end{pmatrix} \right| \left| J\begin{pmatrix} z_1 & z_2 \\ x_1 & x_2 \end{pmatrix} \right| f_{Z_1Z_2}(z_1, z_2),$$

which implies

$$\left| J\begin{pmatrix} z_1 & z_2 \\ x_1 & x_2 \end{pmatrix} \right| = \left| J\begin{pmatrix} x_1 & x_2 \\ z_1 & z_2 \end{pmatrix} \right|^{-1}. \tag{4.24}$$

This is the two-dimensional generalization of the result

$$\frac{dy}{dx} = \left(\frac{dx}{dy}\right)^{-1}$$

in elementary calculus.

Example 4.19 (Polar Coordinates). Let

$$\begin{aligned} x &= r\cos\theta \\ y &= r\sin\theta. \end{aligned}$$

Then

$$\left| J\begin{pmatrix} x & y \\ r & \theta \end{pmatrix} \right| = \left\| \begin{matrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{matrix} \right\| = r.$$

Therefore,

$$f_{R\Theta}(r, \theta) = r f_{XY}(r\cos\theta, r\sin\theta).$$

Example 4.20. Following the last example, this time we want to determine f_{XY} in terms of $f_{R\Theta}$. From (4.24), we have

$$\left|J\begin{pmatrix} r & \theta \\ x & y \end{pmatrix}\right| = \left|J\begin{pmatrix} x & y \\ r & \theta \end{pmatrix}\right|^{-1} = \frac{1}{r} = \frac{1}{\sqrt{x^2+y^2}}.$$

Therefore,

$$f_{XY}(x,y) = \frac{1}{\sqrt{x^2+y^2}} f_{R\Theta}\left(\sqrt{x^2+y^2}, \tan^{-1}\frac{y}{x}\right).$$

Alternatively, we can write

$$\begin{aligned} r &= \sqrt{x^2+y^2} \\ \theta &= \tan^{-1}\frac{y}{x} \end{aligned}$$

and apply (4.21) directly to obtain $\left|J\begin{pmatrix} r & \theta \\ x & y \end{pmatrix}\right|$, but this would be considerably more difficult.

Appendix 4.A: Differentiating $F_{XY}(\cdot|A)$

From (4.13), we have

$$\begin{aligned} f_{XY}(x,y|(X,Y)\in A) &= \frac{\frac{\partial^2}{\partial x \partial y}\int\int_{R_A(x,y)} f_{XY}(x',y')dy'dx'}{P((X,Y)\in A)} \\ &= \frac{\frac{\partial^2}{\partial x \partial y} S(x,y)}{P((X,Y)\in A)}, \end{aligned} \tag{4.25}$$

where

$$S(x,y) = \int\int_{R_A(x,y)} f_{XY}(x',y')dy'dx'.$$

Following the steps toward proving Property 1 of Theorem 4.3, we have

$$\begin{aligned} \frac{\partial^2 S(x,y)}{\partial x \partial y} = \lim_{\Delta x\to 0}\lim_{\Delta y\to 0} \frac{1}{\Delta x \Delta y} &[S(x+\Delta x, y+\Delta y) - S(x+\Delta x, y) \\ &- S(x, y+\Delta y) + S(x,y)]. \end{aligned}$$

Let

$$B_\Delta(x,y) = \{(x',y') : x < x' \le x+\Delta x, y < y' \le y+\Delta y\}.$$

We leave it as an exercise for the reader to show that the expression in the square bracket above is equal to

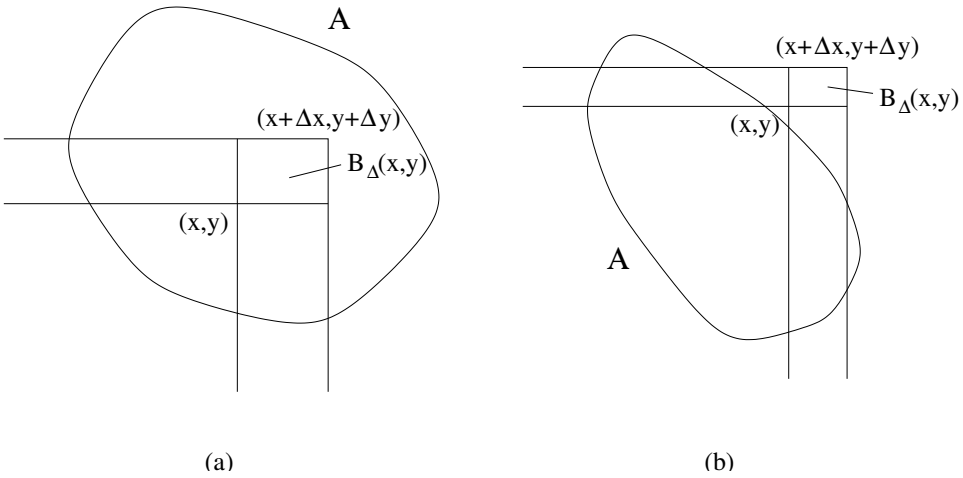


Fig. 4.11: The set $B_\Delta(x,y)$ for $(x,y) \in A$ and $(x,y) \notin A$.

$$\int\int_{B_\Delta(x,y)\cap A} f_{XY}(x',y')dy'dx' \tag{4.26}$$

(cf. Figure 4.11). If $(x,y) \in A$, then $B_\Delta(x,y) \cap A = B_\Delta(x,y)$, so that (4.26) is equal to

$$\int_x^{x+\Delta x}\int_y^{y+\Delta y} f_{XY}(x',y')dy'dx',$$

and hence

$$\frac{\partial^2 S(x,y)}{\partial x \partial y} = f_{XY}(x,y).$$

Otherwise, $B_\Delta(x,y) \cap A = \emptyset$, so that (4.26) vanishes, and

$$\frac{\partial^2 S(x,y)}{\partial x \partial y} = 0.$$

Therefore, from (4.25), we have

$$f_{XY}(x,y|(X,Y) \in A) = \begin{cases} \frac{f_{XY}(x,y)}{P((X,Y)\in A)} & \text{if } (x,y) \in A \\ 0 & \text{if } (x,y) \notin A. \end{cases}$$

5

Expectation and Moment Generating Functions

We have already discussed the expectation of a single random variable in Chapter 2. In this chapter, we will discuss expectation as an *operator* when more than one random variable is involved. We will also introduce *moment generating functions*, a fundamental tool in probability theory.

The results will be proved by assuming that the random variables are continuous. The proofs for discrete random variables are analogous.

5.1 Expectation as a Linear Operator

We have defined the expectation of a function of a random variable in Definition 3.24. When the function involves two random variables, we have the following definition.

Definition 5.1. *The expectation of a function g of random variables X and Y is defined by*

$$Eg(X,Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} g(x,y)p(x,y)$$

for X and Y discrete, and

$$Eg(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$$

for X and Y continuous.

One can regard expectation as an operator which inputs a random variable and outputs the expectation of that random variable. We will establish in the following theorem that expectation is a linear operator.

Theorem 5.2. *Expectation is a linear operator, i.e., for any real numbers a and b,*

$$E[aX + bY] = aEX + bEY. \tag{5.1}$$

Proof This can be proved by considering

$$\begin{aligned} E[aX+bY] &= \int\int (ax+by) f_{XY}(x,y)dxdy \\ &= a\int\int x f_{XY}(x,y)dxdy + b\int\int y f_{XY}(x,y)dxdy \\ &= a\int x\left[\int f_{XY}(x,y)dy\right]dx + b\int y\left[\int f_{XY}(x,y)dx\right]dy \\ &= a\int x f_X(x)dx + b\int y f_Y(y)dy \\ &= aEX + bEY. \end{aligned}$$

Corollary 5.3. $E[aX] = aEX$.

Proof This is a special case of Theorem 5.2 with b equals 0.

Example 5.4. Consider

$$\begin{aligned} \mathrm{var}X &= E[X-EX]^2 \\ &= E[X^2 - 2(EX)X + (EX)^2] \\ &= EX^2 - 2(EX)(EX) + E[(EX)^2] \\ &= EX^2 - 2(EX)^2 + (EX)^2 \\ &= EX^2 - (EX)^2. \end{aligned}$$

This is an alternative proof for Theorem 3.27 with simplified notation.

Proposition 5.5. $\mathrm{var}(aX) = a^2\mathrm{var}X$.

Proof This can be proved by considering

$$\begin{aligned} \mathrm{var}(aX) &= E(aX)^2 - (E(aX))^2 \\ &= a^2EX^2 - a^2(EX)^2 \\ &= a^2(EX^2 - (EX)^2) \\ &= a^2\mathrm{var}X. \end{aligned}$$

Proposition 5.6. *Let m and σ^2 be the mean and the variance of a random variable X. Then $Y = \frac{X-m}{\sigma}$ has zero mean and unit variance.*

Proof To prove that $EY = 0$, consider

$$\begin{aligned} EY &= E\left[\frac{X-m}{\sigma}\right] \\ &= \frac{1}{\sigma}(EX - m) \\ &= 0. \end{aligned}$$

We will prove that $\text{var}Y = 1$ in two steps. First,

$$\begin{aligned}\text{var}(X-m) &= E[X-m]^2 - (E[X-m])^2 \\ &= E[X-m]^2 - 0 \\ &= \text{var}X,\end{aligned}$$

which is intuitively clear. Second, by Proposition 5.5, we have

$$\begin{aligned}\text{var}Y &= \frac{1}{\sigma^2}\text{var}(X-m) \\ &= \frac{1}{\sigma^2}\text{var}X \\ &= \frac{\sigma^2}{\sigma^2} \\ &= 1.\end{aligned}$$

Theorem 5.7. *If X and Y are independent, then*

$$E[g_1(X)g_2(Y)] = (Eg_1(X))(Eg_2(Y)). \tag{5.2}$$

Proof If X and Y are independent, then so are $g_1(X)$ and $g_2(Y)$ by Theorem 4.10. Therefore,

$$\begin{aligned}E[g_1(X)g_2(Y)] &= \int\int g_1(x)g_2(y)f_{XY}(x,y)dxdy \\ &= \int\int g_1(x)g_2(y)f_X(x)f_Y(y)dxdy \\ &= \int g_2(y)f_Y(y)\left[\int g_1(x)f_X(x)dx\right]dy \\ &= \left[\int g_1(x)f_X(x)dx\right]\left[\int g_2(y)f_Y(y)dy\right] \\ &= (Eg_1(X))(Eg_2(Y)).\end{aligned}$$

Remark The equation (5.1) is valid for *all* joint distribution for X and Y. By contrast, (5.2) may not be valid if X and Y are not independent.

5.2 Conditional Expectation

Definition 5.8. *Let X and Y be jointly distribution random variables. The expectation of X conditioning on the event $\{Y = y\}$ is defined by*

$$E[X|Y=y] = \int_{\infty}^{\infty} x f_{X|Y}(x|y)dx.$$

We note that $E[X|Y=y]$ depends only on the value of y, and hence is a function of y. If the value of Y is not specified, then

$$E[X|Y] = \int_{\infty}^{\infty} x f_{X|Y}(x|Y)dx$$

is a random variable, and more specifically, a function of the random variable Y. We now prove an important theorem regarding conditional expectation.

Theorem 5.9. $EX = EE[X|Y]$.

Proof Consider

$$\begin{aligned}
EX &= \int x f_X(x)dx \\
&= \int x \left(\int f_{XY}(x,y)dy \right) dx \\
&= \int \int x f_{X|Y}(x|y) f_Y(y)dydx \\
&= \int \int x f_{X|Y}(x|y) f_Y(y)dxdy \\
&= \int \left(\int x f_{X|Y}(x|y)dx \right) f_Y(y)dy \\
&= E\left[\int x f_{X|Y}(x|Y)dx \right] \\
&= EE[X|Y].
\end{aligned}$$

Note that in the above, the inner expectation is a conditional expectation, while the outer expectation is an expectation on a function of Y.

Example 5.10. Let $X \sim \mathcal{E}(\lambda)$ and Y distributes uniformly on $[0,x]$ when $X = x$. We now apply Theorem 5.9 to determine EY.

$$\begin{aligned}
EY &= EE[Y|X] \\
&= \int_0^{\infty} E[Y|X=x] f_X(x)dx \\
&= \int_0^{\infty} \left(\frac{x}{2} \right) \lambda e^{-\lambda x} dx \\
&= \frac{1}{2} \int_0^{\infty} x \lambda e^{-\lambda x} dx.
\end{aligned}$$

The above integral is the expectation of $\mathcal{E}(\lambda)$, which was shown in Example 3.9 to be $\frac{1}{\lambda}$. Therefore, $EY = \frac{1}{2\lambda}$.

Alternatively, we can consider

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{x} \text{ if } 0 < y < x \\ 0 \text{ otherwise,} \end{cases}$$

so that

$$\begin{aligned} f_{XY}(x,y) &= f_{Y|X}(y|x)f_X(x) \\ &= \begin{cases} \frac{1}{x}\lambda e^{-\lambda x} \text{ if } 0 < y < x, \\ 0 \qquad\quad \text{otherwise.} \end{cases} \end{aligned}$$

Then

$$f_Y(y) = \int_y^\infty \frac{1}{x}\lambda e^{-\lambda x} dx$$

and

$$EY = \int_0^\infty y f_Y(y) dy.$$

However, evaluations of these integrals are quite difficult.

5.3 Covariance and Schwartz's Inequality

Let X and Y be random variables, and consider $Z = X + Y$. We are naturally interested in the relation between the variance of Z with the variances of X and Y. Toward this end, consider

$$\begin{aligned} \text{var}Z &= EZ^2 - (EZ)^2 \\ &= E(X+Y)^2 - (E[X+Y])^2 \\ &= E[X^2 + Y^2 + 2XY] - (EX + EY)^2 \\ &= EX^2 + EY^2 + 2E[XY] - [(EX)^2 + (EY)^2 + 2(EX)(EY)] \\ &= [EX^2 - (EX)^2] + [EY^2 - (EY)^2] + 2[E[XY] - (EX)(EY)] \\ &= \text{var}X + \text{var}Y + 2[E[XY] - (EX)(EY)]. \end{aligned}$$

Define

$$\text{cov}(X,Y) = E[XY] - (EX)(EY), \tag{5.3}$$

so that we can write

$$\text{var}Z = \text{var}X + \text{var}Y + 2\text{cov}(X,Y).$$

When X and Y are independent, by Theorem 5.7,

$$\begin{aligned} \text{cov}(X,Y) &= E[XY] - (EX)(EY) \\ &= (EX)(EY) - (EX)(EY) \\ &= 0, \end{aligned}$$

so that

$$\text{var}Z = \text{var}X + \text{var}Y. \tag{5.4}$$

When $\text{cov}(X,Y) = 0$, or equivalently

$$E[XY] = (EX)(EY),$$

or equivalently (5.4) holds, X and Y are said to be *uncorrelated*. As shown, X and Y are uncorrelated if X and Y are independent. However, the converse is not true.

Example 5.11. Consider the following joint pmf for random variables X and Y:

$p(x,y)$	$y=0$	$y=1$
$x=-1$	0	0.25
$x=0$	0.5	0
$x=1$	0	0.25

Then it is easy to X and Y are uncorrelated but they are not independent.

Proposition 5.12. $\mathrm{cov}(X,Y) = E[(X-EX)(Y-EY)]$.

Proof This can be proved by considering

$$\begin{aligned}
&E[(X-EX)(Y-EY)]\\
&= E[XY-(EY)X-(EX)Y+(EX)(EY)]\\
&= E[XY]-(EY)(EX)-(EX)(EY)+(EX)(EY)\\
&= E[XY]-(EY)(EX)\\
&= \mathrm{cov}(X,Y).
\end{aligned}$$

In fact, it is ready seen from (5.3) and Proposition 5.12 that $\mathrm{cov}(X,X) = \mathrm{var}X$. Thus the variance of a random variable X can be regarded as the covariance between X and itself.

We now prove an inequality known as *Schwartz's inequality*. This inequality gives an upper bound on the magnitude of $E[XY]$ in terms of the second moments of X and Y.

Theorem 5.13 (Schwartz's Inequality). $(E[XY])^2 \le (EX^2)(EY^2)$.

Proof Let X and Y be any fixed jointly distributed random variables. Let λ be any real number, and consider the random variable $(X-\lambda Y)^2$. Since this random variable can only take nonnegative values, we have

$$E[(X-\lambda Y)^2] \ge 0, \tag{5.5}$$

or

$$EX^2 - 2\lambda E[XY] + \lambda^2 EY^2 \ge 0.$$

Rearranging the terms, we have

$$(EY^2)\lambda^2 - 2(EXY)\lambda + EX^2 \ge 0. \tag{5.6}$$

Since X and Y are fixed random variables, EY^2, EXY, and EX^2 in the above are fixed real numbers. Therefore, the left hand side above is quadratic in λ, and it achieves its minimum when

$$\lambda = \lambda^* \stackrel{\text{def}}{=} -\frac{-2(EXY)}{2(EY^2)} = \frac{EXY}{EY^2}.$$

Since (5.5) holds for all λ, it holds for $\lambda = \lambda^*$. Substituting $\lambda = \lambda^*$ into (5.6), we have

$$EY^2\left(\frac{EXY}{EY^2}\right)^2 - 2(EXY)\left(\frac{EXY}{EY^2}\right) + EX^2 \geq 0$$

$$\frac{(EXY)^2}{EY^2} - 2\frac{(EXY)^2}{EY^2} + EX^2 \geq 0.$$

Hence,

$$(EXY)^2 \leq (EX^2)(EY^2),$$

proving the theorem.

Corollary 5.14. *Schwartz's inequality is tight, i.e.,*

$$(EXY)^2 = (EX^2)(EY^2),$$

if and if X is proportional to Y.

Proof From the proof of the last theorem, we see that Schwartz's inequality is simply the inequality in (5.5) with $\lambda = \lambda^*$. Therefore, Schwartz's inequality is tight if and only if

$$E(X - \lambda^* Y)^2 = 0, \tag{5.7}$$

which in turn holds if and only if with probability 1, $X - \lambda^* Y = 0$, or $X = \lambda^* Y$, implying that X is proportional to Y. On the other hand, if X is proportional to Y, i.e., $X = \lambda Y$ for some constant λ, then it is readily verified that Schwartz's inequality is tight.

Denote var$X$ and var$Y$ by σ_x^2 and σ_y^2, respectively. Now in Schwartz's inequality, if we replace X by $(X - EX)$ and Y by $(Y - EX)$[1], we immediately obtain

$$[E(X - EX)(Y - EY)]^2 \leq [E(X - EX)^2][E(Y - EY)^2],$$

or

$$\text{cov}(X, Y)^2 \leq \sigma_x^2 \sigma_y^2,$$

where we have invoked Proposition 5.12. Thus,

$$-\sigma_x \sigma_y \leq \text{cov}(X, Y) \leq \sigma_x \sigma_y. \tag{5.8}$$

[1] This is valid because X and Y can be any random variables.

The ratio

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y},$$

called the *coefficient of correlation*, is a measure of the degree of correlation between X and Y. It is readily seen from (5.8) that

$$-1 \leq r \leq 1.$$

When $r = 0$, we say that X and Y are uncorrelated. When $0 < r \leq 1$, we say that X and Y are positively correlated. When $-1 \leq r < 0$, we say that X and Y are negatively correlated.

When $r = 0$, $\text{var}(X+Y)$ is simply equal to $\text{var}X + \text{var}Y$. When $0 < r \leq 1$,

$$\begin{aligned} \sigma_x^2 + \sigma_y^2 &< \text{var}(X + Y) \\ &\leq \sigma_x^2 + \sigma_y^2 + 2\sigma_x \sigma_y \\ &= (\sigma_x + \sigma_y)^2. \end{aligned}$$

Thus

$$\sigma_x^2 + \sigma_y^2 < \text{var}(X + Y) \leq (\sigma_x + \sigma_y)^2.$$

When $-1 \leq r < 0$,

$$\begin{aligned} \sigma_x^2 + \sigma_y^2 &> \text{var}(X + Y) \\ &\geq \sigma_x^2 + \sigma_y^2 - 2\sigma_x \sigma_y \\ &= (\sigma_x - \sigma_y)^2. \end{aligned}$$

Thus

$$(\sigma_x - \sigma_y)^2 \leq \text{var}(X + Y) < \sigma_x^2 + \sigma_y^2.$$

Example 5.15. We already have seen that when X and Y are independent, then they are uncorrelated, or $r = 0$. When $X = Y$,

$$\text{cov}(X, Y) = \text{cov}(X, X) = \text{var}X = \sigma_x^2,$$

so that

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{\sigma_x^2}{\sigma_x^2} = 1,$$

and

$$\text{var}(X + Y) = \text{var}(2X) = 4\text{var}X,$$

by Proposition 5.5. When $X = -Y$,

$$\text{cov}(X, Y) = \text{cov}(X, -X) = -\text{var}X = -\sigma_x^2,$$

so that

$$r = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} = \frac{-\sigma_x^2}{\sigma_x^2} = -1,$$

and

$$\text{var}(X + Y) = \text{var}\ 0 = 0.$$

5.4 Moment Generating Functions

The moments of a random variable (or of a distribution) were introduced in Section 3.6. In this section, we introduce the *moment generating function* of a random variable X, denoted by $\Phi_X(s)$, where the parameter s is a real number.

Definition 5.16. *The moment generating function of a random variable X (or of the distribution of X) is defined by*

$$\Phi_X(s) = E[e^{sX}], \tag{5.9}$$

where s is a real number. For the discrete case,

$$\Phi_X(s) = \sum_x p(x)e^{sx}, \tag{5.10}$$

and for the continuous case,

$$\Phi_X(s) = \int_{-\infty}^{\infty} e^{sx} f(x)dx. \tag{5.11}$$

$\Phi_X(s)$ is said to exist if it exists (i.e., the summation in (5.10) or the integral in (5.11) converges) for some range of s .

Note that for a fixed s, e^{sX} is a function of X. So $\Phi_X(s)$ is simply the expectation of this particular function of X, and its value depends on s. Therefore, $\Phi_X(s)$ is a function of s.

Proposition 5.17. $\Phi_X(0) = 1$.

Proof By letting $s = 0$ in (5.9), we have

$$\Phi_X(0) = E[e^{0X}] = E[1] = 1,$$

proving the corollary. This equality can be interpreted by letting $s = 0$ in (5.10) and (5.11), which is nothing but the normalizing conditions for $\{p(x)\}$ and $f(x)$.

The moment generating function can be regarded as the "transform" of the distribution. In (5.10), if $\mathcal{X} = \{0, 1, 2, \ldots\}$, by letting $s = \ln z$ so that $e^s = z$, the right hand side becomes

$$\sum_{k=0}^{\infty} p(k)z^k,$$

which is the z-transform of the pmf $\{p(k)\}$ (with k being the time index). In (5.11), by letting $s = -j\omega$, $\Phi_X(s)$ becomes the Fourier transform of the pdf

$f(x)$ (with x being the index of time). (But of course, s in $\Phi_X(s)$ is actually a real number.)

It can be shown that there is a one-to-one correspondence between the distribution of X and the moment generating function of X. That is, for each moment generating function, there is a *unique* distribution corresponding to it. In other words, the distribution of a random variable X can (in principle) be recovered from its moment generating function.

The moments of X can be obtained by successive differentiation of $\Phi_X(s)$ with respect to s. Specifically,

$$\begin{aligned}\Phi'_X(s) &= \frac{d}{ds}E[e^{sX}] \\ &= E\left[\frac{d}{ds}e^{sX}\right] \\ &= E[Xe^{sX}],\end{aligned}$$

and in general

$$\Phi_X^{(n)}(s) = E[X^n e^{sX}], \tag{5.12}$$

where $\Phi_X^{(n)}(s) = \frac{d^n}{ds^n}E[e^{sX}]$. Then by putting $s = 0$ in (5.12), we obtain

$$\Phi_X^{(n)}(0) = E\left[X^n e^{0X}\right] = E[X^n].$$

That is why $\Phi_X(s)$ is called the moment generating function of X.

Theorem 5.18. *If X and Y are independent, then $\Phi_{X+Y}(s) = \Phi_X(s)\Phi_Y(s)$.*

Proof This can be proved by considering

$$\begin{aligned}\Phi_{X+Y}(s) &= E\left[e^{s(X+Y)}\right] \\ &= E\left[e^{sX}e^{sY}\right] \\ &= E\left[e^{sX}\right]E\left[e^{sY}\right] \\ &= \Phi_X(s)\Phi_Y(s).\end{aligned}$$

We know from Example 4.17 that if X and Y are independent, then $Z = X + Y$ has pdf given by $f_X * f_Y(z)$. The above theorem says that $\Phi_Z(s) = \Phi_X(s)\Phi_Y(s)$, so it is actually analogous to the *convolution theorem* in linear system theory.

Example 5.19. Let $X \sim \mathcal{E}(\lambda)$. Then for $s < \lambda$,

$$\begin{aligned}\Phi_X(s) &= E\left[e^{sX}\right] \\ &= \int_0^\infty \lambda e^{-\lambda x} e^{sx} dx \\ &= \lambda \int_0^\infty e^{-(\lambda - s)x} dx \\ &= \lambda \frac{e^{-(\lambda-s)x}}{-(\lambda - s)} \Big|_0^\infty \\ &= \frac{\lambda}{\lambda - s}(1 - 0) \\ &= \frac{\lambda}{\lambda - s}.\end{aligned}$$

Example 5.20. Let N be the geometric random variable with parameter p, where $0 < p < 1$, i.e.,

$$P(N = n) = \begin{cases} p(1-p)^n & n = 0, 1, 2, \ldots, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned}\Phi_N(s) &= E\left[e^{sN}\right] \\ &= \sum_{n=0}^{\infty} p(1-p)^n e^{sn} \\ &= p \sum_{n=0}^{\infty} [(1-p)e^s]^n \\ &= \frac{p}{1 - (1-p)e^s},\end{aligned}$$

where we require that $(1-p)e^s < 1$, or $s < -\ln(1-p)$.

Remark: We use the convention adopted by most books in probability in defining the geometric distribution in the above example. A random variable N' with

$$P(N' = n) = \begin{cases} p(1-p)^{n-1} & n = 1, 2, \ldots, \\ 0 & \text{otherwise} \end{cases}$$

is said to have a *truncated* geometric distribution with parameter p. If we toss a coin with the probability of obtaining a head equals p repeatedly, then it is easy to see that the first time a head is obtained has truncated geometric distribution with parameter p. Note that if N is geometric with parameter p, then $N' = N + 1$ is truncated geometric with parameter p.

In the next example, we give a powerful application of the moment generating function.

Example 5.21. Let $X_k, k = 1, 2, \ldots,$ be i.i.d. random variables each $\sim \mathcal{E}(\lambda)$, N be an independent truncated geometric random variable (independent of X_k's) with parameter p, and $Y = \sum_{k=1}^{N} X_k$. That is, Y is the sum of a truncated-geometric number of i.i.d. exponential random variables.

We now determine the distribution of Y via $\Phi_Y(s)$. Consider

$$\begin{aligned}
\Phi_Y(s) &= E\left[e^{sY}\right] \\
&= E\left[e^{s\sum_{k=1}^{N} X_k}\right] \\
&= EE\left[e^{s\sum_{k=1}^{N} X_k} \mid N\right] \\
&= EE\left[e^{s(X_1+\ldots+X_N)} \mid N\right] \\
&= EE\left[e^{sX_1}\ldots e^{sX_N} \mid N\right] \\
&\overset{a)}{=} E\left\{E\left[e^{sX_1} \mid N\right]\cdots E\left[e^{sX_N} \mid N\right]\right\} \\
&\overset{b)}{=} E\left\{E\left[e^{sX_1}\right]\cdots E\left[e^{sX_N}\right]\right\} \\
&\overset{c)}{=} E\left[\Phi_{X_1}(s)\right]^N \\
&\overset{d)}{=} \sum_{n=1}^{\infty}\left(\frac{\lambda}{\lambda-s}\right)^n p(1-p)^{n-1} \\
&= \sum_{n=0}^{\infty}\left(\frac{\lambda}{\lambda-s}\right)^{n+1} p(1-p)^n \\
&= \frac{\lambda p}{\lambda-s}\sum_{n=0}^{\infty}\left[\frac{\lambda(1-p)}{\lambda-s}\right]^n \\
&= \frac{\lambda p}{\lambda-s}\frac{1}{1-\frac{\lambda(1-p)}{\lambda-s}} \\
&= \frac{\lambda p}{\lambda-s-\lambda(1-p)} \\
&= \frac{\lambda p}{\lambda p-s},
\end{aligned}$$

where

a) follows because X_k's are independent, and they remain to be so when conditioning on N because N is independent of X_k's;
b) follows because N is independent of X_k's;
c) follows because X_k are i.i.d.;
d) follows because $\Phi_{X_1}(s) = \frac{\lambda}{\lambda-s}$.

From the form of $\Phi_Y(s)$, we see that $Y \sim \mathcal{E}(\lambda p)$.

6

Limit Theorems

Recall that in a random experiment, the outcome ω is drawn from the sample space Ω according to the probability measure P, and a random variable X is a function of ω. Now suppose we construct random variables $X_k(\omega)$, where $k = 1, 2, \ldots$. Then $\{X_k\}$ is called a *stochastic process* (or *random process*).

Since a stochastic process involves an infinite number of random variables, it cannot be described by a joint distribution. In fact, a stochastic process can be very complicated in general. But if X_k's are i.i.d., in that case we say that $\{X_k\}$ is an i.i.d. process, then it is relatively easy to describe. The process of tossing a coin repeatly can be modeled as an i.i.d. process.

Limit theorems deal with the asymptotic behaviors of stochastic processes, and they play an important role in probability theory. In this chapter, we will discuss the two most fundamental such theorems, namely the weak law of large number (WLLN) and the central limit theorem (CLT).

6.1 The Weak Law of Large Numbers

Consider tossing a fair coin n times. Intuitively, when n is large, the relative frequency of obtaining a head should be approximately equal to 0.5. The WLLN makes this idea precise, as we will see, and it gives a physical meaning to the probability of an event.

Theorem 6.1 (Chebyshev's Inequality). *For any random variable X and any $\epsilon \geq 0$,*

$$P(|X - EX| \geq \epsilon) \leq \frac{\mathrm{var}X}{\epsilon^2}.$$

Proof The inequality can be proved by considering

$$\begin{aligned}
\mathrm{var}X &= \int_{-\infty}^{\infty} (x - EX)^2 f(x)dx \\
&\geq \int_{-\infty}^{EX-\epsilon} (x - EX)^2 f(x)dx + \int_{EX+\epsilon}^{\infty} (x - EX)^2 f(x)dx \\
&\geq \epsilon^2 \left[\int_{-\infty}^{EX-\epsilon} f(x)dx + \int_{EX+\epsilon}^{\infty} f(x)dx \right] \\
&= \epsilon^2 P(|X - EX| \geq \epsilon).
\end{aligned}$$

The interpretation of Chebyshev's inequality is as follows. Here, $\{|X - EX| \geq \epsilon\}$ is the event that X deviates from its mean by at least ϵ. The upper bound on the probability of the event is such that:

- if $\mathrm{var}X$ is small, then it is small;
- if ϵ is large, then it is small.

In fact, sometimes this upper bound can be larger than 1, making it useless. In general, Chebyshev's inequality is rather loose.

Theorem 6.2 (WLLN). *Let $\{X_k\}$ be i.i.d. random variables with mean m and variance σ^2. Then*

$$P\left(\left| \frac{1}{n} \sum_{k=1}^{n} X_k - m \right| \geq \epsilon \right) \leq \frac{\sigma^2}{n\epsilon^2}. \tag{6.1}$$

Proof Let

$$Y_n = \frac{1}{n} \sum_{k=1}^{n} X_k. \tag{6.2}$$

The WLLN is actually an application of Chebyshev's inequality to Y_n. Now

$$EY_n = E\left[\frac{1}{n} \sum_{k=1}^{n} X_k \right] = \frac{1}{n} \sum_{k=1}^{n} EX_k = m,$$

and

$$\begin{aligned}
\mathrm{var}Y_n &\overset{a)}{=} \frac{1}{n^2} \mathrm{var} \left\{ \sum_{k=1}^{n} X_k \right\} \\
&\overset{b)}{=} \frac{1}{n^2} \sum_{k=1}^{n} \mathrm{var}X_k \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n},
\end{aligned}$$

where a) follows from Proposition 5.5 and b) follows because X_k are independent. Then by Chebyshev's inequality, we have

$$P(|Y_n - EY_n| \geq \epsilon) \leq \frac{\text{var}Y_n}{\epsilon^2},$$

or

$$P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - m\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

We note that in (6.1), the upper bound tends to zero for all finite σ and ϵ as $n \to \infty$. Thus the WLLN says that as long as n is sufficiently large, the probability that the average of $X_1, \ldots, X_n$, i.e., $\frac{1}{n}\sum_{k=1}^{n} X_k$, deviates from m by any prescribed amount is arbitrarily small. Technically, we say that

$$\frac{1}{n}\sum_{k} X_k \to m \quad \textit{in probability.}$$

We now give another interpretation of the WLLN. Let X be the generic random variable and $f(x)$ be its density function, i.e.,

$$X \sim f(x).$$

It can readily be shown that $aX \sim \frac{1}{a}f(\frac{x}{a})$ for any $a > 0$. Then by Theorem 5.18,

$$X_1 + \ldots + X_n \sim \underbrace{f * \ldots * f}_{n}(x) = f^{(n)}(x),$$

so that

$$\frac{1}{n}(X_1 + \ldots + X_n) \sim nf^{(n)}(nx).$$

Thus the WLLN says that

$$nf^{(n)}(nx) \to \delta(x - m)$$

regardless of the actual form of f.

Having proved the WLLN, we now show an important consequence of it. Consider any subset A of $\mathcal{X}$, the alphabet of the generic random variable X of the i.i.d. process $\{X_k\}$. We are interested in the relative frequency of occurrence of A in the first n trial. More precisely, we define the random variable (called an *indicator function*)

$$I_k(A) = \begin{cases} 1 \text{ if } X_k \in A \\ 0 \text{ if } X_k \notin A, \end{cases}$$

so that

$$\begin{aligned} EI_k(A) &= 1 \cdot P(X_k \in A) + 0 \cdot P(X_k \notin A) \\ &= P(X_k \in A) \\ &= P(X \in A). \end{aligned}$$

Since X_k are i.i.d., so are $I_k(A)$ because $I_k(A)$ is a function of X_k. The relative frequency of occurrence of A is given by

$$\frac{1}{n}\sum_{k=1}^{n} I_k(A).$$

Since X_k are i.i.d., so are $I_k(A)$ because $I_k(A)$ is a function of X_k. Then by the WLLN,

$$\frac{1}{n}\sum_{k}^{n} I_k(A) \to P(X \in A) \quad \textit{in probability.}$$

Therefore, the probability of an event can be interpreted as the long term relative frequency of occurrence of the event when the experiment is repeated indefinitely.

Example 6.3. If a fair dice is tossed a large number of times, the average outcome is close to

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

with probability almost 1.

Example 6.4. If a fair coin is tossed a large number of times, the relative frequency of occurrence of a head is close to 0.5 with probability almost 1.

Example 6.5. If a fair dice is tossed a large number of times, the relative frequency of occurrence of a multiple of 3 is close to

$$P(\{3,6\}) = p(3) + p(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

with probability almost 1.

6.2 The Central Limit Theorem

The Gaussian distribution was introduced in Example 3.29. It is one of the most important distributions in probability theory because of its wide applications. Noise in communication systems, the marks in an examination, etc, can be modeled by the Gaussian distribution. In this section, we will discuss the central limit theorem (CLT) that explains the origin of this important distribution. The theorem is first stated below.

Theorem 6.6 (Central Limit Theorem). *Let $\{X_k\}$ be i.i.d. with zero mean and unit variance, and let*

$$Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k. \tag{6.3}$$

Then

$$\lim_{n\to\infty} Z_n \sim \mathcal{N}(0,1).$$

The reader should compare the definition of Y_n in (6.2) for the WLLN with the definition of Z_n here, and observe that for Y_n the scaling factor is n while for Z_n it is $\sqrt{n}$. To understand the CLT properly, we note that

$$EZ_n = E\left[\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k\right] = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} E[X_k] = 0$$

since X_k has zero mean, and

$$\begin{aligned}
\mathrm{var} Z_n &= \mathrm{var}\left[\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X_k\right] \\
&\overset{a)}{=} \left(\frac{1}{\sqrt{n}}\right)^2 \mathrm{var}\left[\sum_{k=1}^{n} X_k\right] \\
&\overset{b)}{=} \frac{1}{n} \sum_{k=1}^{n} \mathrm{var} X_k \\
&\overset{c)}{=} \frac{1}{n} \cdot n \\
&= 1,
\end{aligned}$$

where

a) follows from Proposition 5.5;
b) follows because X_k's are mutually independent (cf. Section 5.3);
c) follows since X_k has unit variance.

So the only thing the CLT says about $\lim_{n\to\infty} Z_n$ is that it has a Gaussian distribution; the mean and the variance of Z_n are already fixed from the setup.

Before we prove the CLT, we first derive the moment generating function of a Gaussian distribuion.

Lemma 6.7. *Let $X \sim \mathcal{N}(m, \sigma^2)$. Then $\Phi_X(s) = e^{ms + \frac{\sigma^2 s^2}{2}}$.*

Proof Consider

$$\begin{aligned}
\Phi_X(s) &= E[e^{sX}] \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{sx} e^{-\frac{(x-m)^2}{2\sigma^2}} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[(x-m)^2 - 2\sigma^2 sx]} dx \\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}[(x-m)^2 - 2\sigma^2 s(x-m) + \sigma^4 s^2]} e^{\frac{1}{2\sigma^2}(\sigma^4 s^2 + 2\sigma^2 sm)} dx \\
&= e^{\frac{1}{2\sigma^2}(\sigma^4 s^2 + 2\sigma^2 sm)} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-(m+\sigma^2 s))^2} dx \\
&= e^{ms + \frac{\sigma^2 s^2}{2}},
\end{aligned}$$

where the last step follows because the density function of $\mathcal{N}(m + \sigma^2 s, \sigma^2)$ integrates to 1, proving the lemma.

Corollary 6.8. *Let $X \sim \mathcal{N}(0,1)$. Then $\Phi_X(s) = e^{\frac{s^2}{2}}$.*

Corollary 6.9. *Let $X_i \sim \mathcal{N}(m_i, \sigma_i^2)$, $i = 1, 2$, where X_1 and X_2 are independent, and let $Y = X_1 + X_2$. Then $Y \sim \mathcal{N}(m_1 + m_2, \sigma_1^2 + \sigma_2^2)$.*

Proof The proof is left as an exercise.

This corollary says that the sum of two independent Gaussian random variable is again Gaussian.

Proof of Theorem 6.6 We give an outline of the proof as follows. Denote the generic random variable of $\{X_k\}$ by X. Consider the second order approximation of $\Phi_X(s)$ via the Taylor expansion[1] for small s:

$$\Phi_X(s) \approx 1 + \Phi'_X(0)s + \frac{1}{2}\Phi''_X(0)s^2.$$

Since $\Phi'_X(0) = EX$ and $\Phi''_X(0) = EX^2$, by the assumption that X has zero mean and unit variance, we have

$$\Phi'_X(0) = 0$$

and

$$\Phi''_X(0) = EX^2 = \mathrm{var}X + (EX)^2 = 1.$$

Therefore,

[1] The Taylor expansion for a function $h(s)$ at $s = 0$ is given by

$$\sum_{n=0}^{\infty} \frac{h^{(n)}(0)}{n!} s^n = h(0) + h'(0) + \frac{1}{2}h''(0) + \ldots,$$

where $h^{(n)}$ denotes the nth derivative of h.

$$\Phi_X(s) \approx 1 + \frac{s^2}{2}$$

for small s. By Theorem 5.18, we have

$$\Phi_{\sum_{k=1}^n X_k}(s) = [\Phi_X(s)]^n \approx \left(1 + \frac{s^2}{2}\right)^n.$$

Then for large n,

$$\begin{aligned}
\Phi_{Z_n}(s) &= E\left[e^{sZ_n}\right] \\
&= E\left[e^{s\frac{1}{\sqrt{n}}\sum_k X_k}\right] \\
&= \Phi_{\sum_k X_k}\left(\frac{s}{\sqrt{n}}\right) \\
&\approx \left(1 + \frac{s^2}{2n}\right)^n,
\end{aligned}$$

where the above approximation is justified because $\frac{s}{\sqrt{n}}$ is small for large n. Then

$$\lim_{n\to\infty} \Phi_{Z_n}(s) = \lim_{n\to\infty} \left(1 + \frac{\frac{s^2}{2}}{n}\right)^n = e^{\frac{s^2}{2}},$$

where we have used the identity

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

Finally, we conclude by Corollary 6.8 that

$$\lim_{n\to\infty} Z_n \sim \mathcal{N}(0,1).$$

This proves the CLT.

The version of the CLT we have proved in Theorem 6.6 applies only when X_k's have zero mean and unit variance. In order to apply it to a general i.i.d. process, we need the following scaling property of Gaussian distributions.

Proposition 6.10. *Let $X \sim \mathcal{N}(m, \sigma^2)$, and $Y = aX + b$. Then $Y \sim \mathcal{N}(am + b, a^2\sigma^2)$. That is, Y again has a Gaussian distribution.*

Proof Consider

$$\begin{aligned}
\Phi_Y(s) &= E[e^{sY}] \\
&= E[e^{s(aX+b)}] \\
&= e^{bs}E[e^{(as)X}] \\
&= e^{bs}\Phi_X(as) \\
&= e^{bs}e^{m(as)+\frac{\sigma^2(as)^2}{2}} \\
&= e^{(am+b)s+\frac{(a^2\sigma^2)s^2}{2}}.
\end{aligned}$$

Thus $Y \sim \mathcal{N}(am + b, a^2\sigma^2)$.

Example 6.11 (Standard Normal CDF). In this example, we give a useful application of Proposition 6.10. Suppose $Y \sim \mathcal{N}(m, \sigma^2)$, and we are interested in $P(Y \leq c)$ for some constant c. Now

$$Y \leq c$$

is equivalent to

$$\frac{Y-m}{\sigma} \leq \frac{c-m}{\sigma}.$$

By defining $X = \frac{Y-m}{\sigma}$, the above is equivalent to

$$X \leq \frac{c-m}{\sigma}.$$

We see from Proposition 5.6 that X has zero mean and unit variance. Moreover, by Proposition 6.10, X is in fact a Gaussian random variable. Therefore,

$$P(Y \leq c) = P\left(X \leq \frac{c-m}{\sigma}\right) = \Phi\left(\frac{c-m}{\sigma}\right),$$

where

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-x^2}{2}} dx$$

is called the standard normal CDF.

Now suppose $\{X_k\}$ is i.i.d. with mean m and variance σ^2. Then by Proposition 5.6, $\{X'_k\}$, where

$$X'_k = \frac{X_k - m}{\sigma}, \tag{6.4}$$

is i.i.d. with zero mean and unit variance. From (6.4), we have

$$X_k = \sigma X'_k + m,$$

so that

$$\sum_{k=1}^{n} X_k = \sum_{k=1}^{n} (\sigma X'_k + m) = \sigma\sqrt{n}\left(\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X'_k\right) + nm.$$

When n is large, by the CLT,

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{n} X'_k \sim \mathcal{N}(0, 1).$$

Then by Proposition 6.10,

$$\sum_{k=1}^{n} X_k \sim \mathcal{N}(nm, n\sigma^2).$$

That is, the sum of a large number of i.i.d. random variables is approximately Gaussian. Again, the mean and the variance of the above Gaussian distribution are determined from the setup and has nothing to do with the CLT.