ST2334

Probability and Statistics

Academic Year 2025/2026
Semester I

David Chew
Department of Statistics and Data Science
email: david.chew@nus.edu.sg

Typesetted using the MiKTEX sytem.

Contents

One

Basic Concepts of Probability

1 Probability Concepts and Definitions

In this section we introduce the basic terminology of probability theory: experiment, outcomes, sample space, events.

Definition 1 (Experiment, Sample space, Event)
*A **statistical experiment** is any procedure that produces data or observations.*

*The **sample space**, denoted by S, is the set of all possible outcomes of a statistical experiment.* The sample space depends on the problem of interest!

*A **sample point** is an outcome (element) in the sample space.*

*An **event** is a subset of the sample space.*

Example 1.1
Consider the experiment of **rolling a die**.

(i) If the problem of interest is "the number that shows on the top face", then
- Sample space: $S = \{1,2,3,4,5,6\}$.
- Sample point: 1 or 2 or 3 or 4 or 5 or 6.
- Some possible events are:
  - an event where an odd number occurs = $\{1,3,5\}$;
  - an event where a number greater than 4 occurs = $\{5,6\}$.

(ii) If the problem of interest is "whether the number is even or odd", then
- Sample space: $S = \{\text{even, odd}\}$.
- Sample point: "even" or "odd".
- A possible event is:
  - an event where an odd number occurs = $\{\text{odd}\}$.

REMARK
The sample space is itself an event and is called a **sure event**.

An event that contains no element is the empty set, denoted by $\emptyset$, and is called a **null event**.

2 EVENT OPERATIONS & RELATIONSHIPS

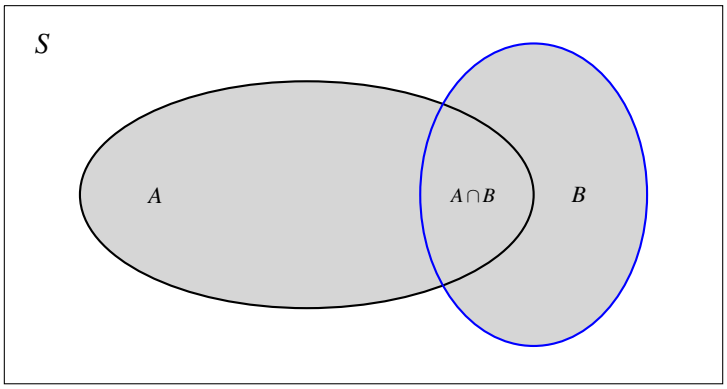
Let A and B be two events in the sample space S. We shall go through some event operations and relationships involving A and B.

- Event operations:
  (i) Union; (ii) Intersection; (iii) Complement.
- Event relationships:
  (i) Contained; (ii) Equivalent; (iii) Mutually exclusive.

Union

The **union** of events A and B, denoted by $A \cup B$, is the event containing all elements that belong to A or B or both. That is

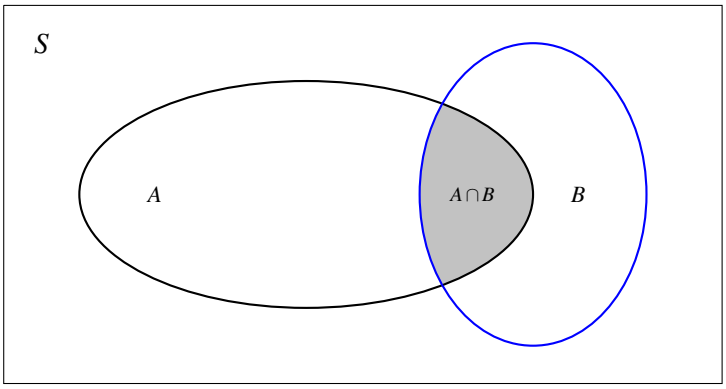
$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$



Intersection

The **intersection** of events A and B, denoted by $A \cap B$ or simply AB, is the event containing elements that belong to both A and B. That is

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



We can also consider the **union** and **intersection** of n events: $A_1, A_2, \ldots, A_n$.

- **Union**:

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \ldots \cup A_n = \{x : x \in A_1 \text{ or } x \in A_2 \text{ or } \ldots \text{ or } x \in A_n\},$$

  comprises of elements that belong to one or more of $A_1, \ldots, A_n$.

- **Intersection**:

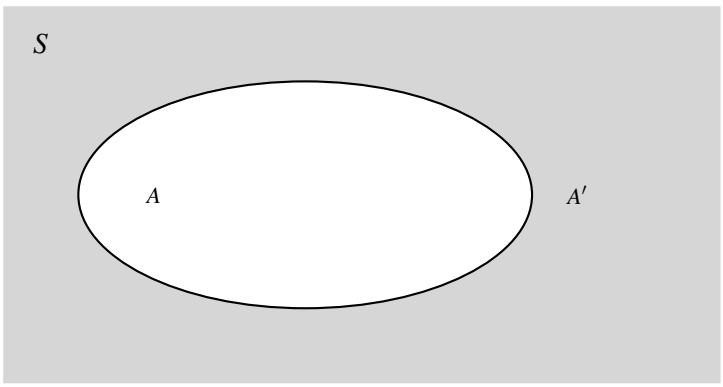
$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \ldots \cap A_n = \{x : x \in A_1 \text{ and } x \in A_2 \text{ and } \ldots \text{ and } x \in A_n\},$$

  comprises of elements that belong to every $A_1, \ldots, A_n$.

Complement

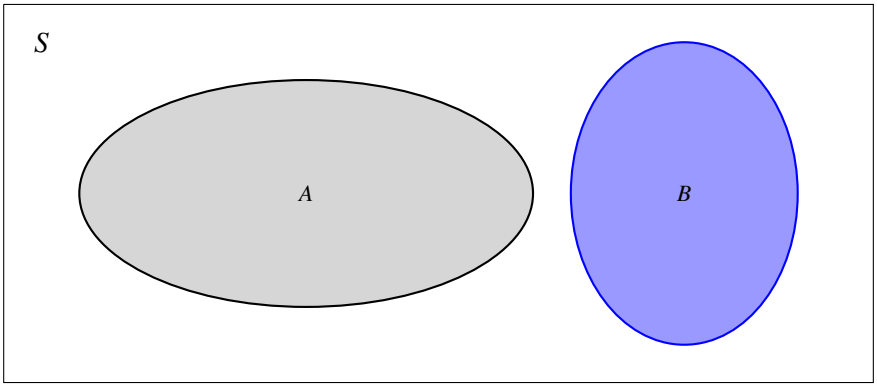
The **complement** of the event A with respect to S, denoted by A', is the event with elements in S, which are not in A. That is

$$A' = \{x : x \in S \text{ but } x \notin A\}.$$



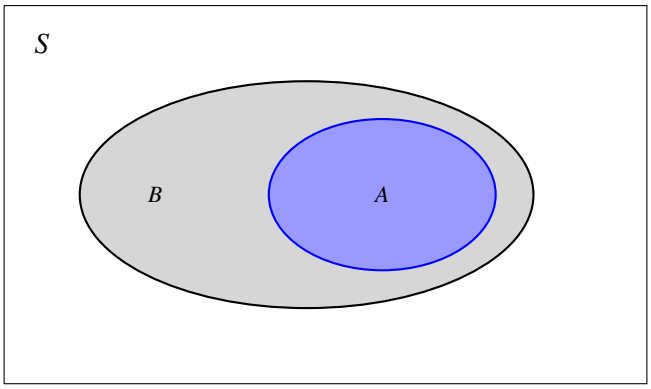
Mutually Exclusive

Events A and B are said to be **mutually exclusive** or **disjoint**, if $A \cap B = \emptyset$. That is, A and B have no element in common.



Contained and Equivalent

If all elements in A are also elements in B, then we say A is **contained** in B, denoted by $A \subset B$, or equivalently $B \supset A$.



If $A \subset B$ and $B \subset A$, then $A = B$. That is, set A and B are **equivalent**.

EXAMPLE 1.2
Consider the sample space and events:

$$S = \{1,2,3,4,5,6\}, \quad A = \{1,2,3\}, \quad B = \{1,3,5\}, \quad C = \{2,4,6\}.$$

Then

(i) $A \cup B = \{1,2,3,5\}$; $A \cup C = \{1,2,3,4,6\}$; $B \cup C = S$.

(ii) $A \cap B = \{1,3\}$; $A \cap C = \{2\}$; $B \cap C = \emptyset$.

(iii) $A \cup B \cup C = S$; $A \cap B \cap C = \emptyset$.

(iv) $A' = \{4,5,6\}$; $B' = \{2,4,6\} = C$.

Note that B and C are mutually exclusive, since $B \cap C = \emptyset$. On the other hand, A and B are not mutually exclusive as $A \cap B = \{1,3\} \neq \emptyset$.

MORE EVENT OPERATIONS

(a) $A \cap A' = \emptyset$ (b) $A \cap \emptyset = \emptyset$

(c) $A \cup A' = S$ (d) $(A')' = A$

(e) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(f) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(g) $A \cup B = A \cup (B \cap A')$

(h) $A = (A \cap B) \cup (A \cap B')$

DE MORGAN'S LAW

For any n events $A_1, A_2, \ldots, A_n$,

(i) $(A_1 \cup A_2 \cup \ldots \cup A_n)' = A_1' \cap A_2' \cap \ldots \cap A_n'$.

A special case: $(A \cup B)' = A' \cap B'$.

(j) $(A_1 \cap A_2 \cap \ldots \cap A_n)' = A_1' \cup A_2' \cup \ldots \cup A_n'$.

A special case: $(A \cap B)' = A' \cup B'$.

EXAMPLE 1.3
We return to Example 1.2 where

$$S=\{1,2,3,4,5,6\},\quad A=\{1,2,3\},\quad B=\{1,3,5\},\quad C=\{2,4,6\}.$$

We have

$$A'=\{4,5,6\},\quad B'=\{2,4,6\},\quad C'=\{1,3,5\}.$$

We check that

$$(A\cup B)'=\{1,2,3,5\}'=\{4,6\};\quad A'\cap B'=\{4,5,6\}\cap\{2,4,6\}=\{4,6\}.$$

This agrees with $(A\cup B)'=A'\cap B'$.
Also,

$$(A\cap B)'=\{1,3\}'=\{2,4,5,6\};\quad A'\cup B'=\{4,5,6\}\cap\{2,4,6\}=\{2,4,5,6\}.$$

This agrees with $(A\cap B)'=A'\cup B'$.

Similarly, we can check that

$$(A\cup B\cup C)'=\emptyset=A'\cap B'\cap C'\quad\text{and}\quad(A\cap B\cap C)'=S=A'\cup B'\cup C'.$$

3 COUNTING METHODS

In many instances, we need to count the number of ways that some operations can be carried out or that certain situations can happen.

There are two fundamental principles in counting:

Multiplication principle

Addition principle

They can be applied to derive some important counting methods: **permutation** and **combination**.

MULTIPLICATION PRINCIPLE
Suppose that r different experiments are to be performed sequentially. Suppose

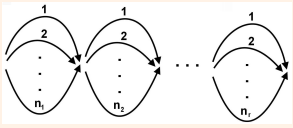
experiment 1 results in n_1 possible outcomes;

for each outcome above, experiment 2 results in n_2 possible outcomes;

… … … … …

for each outcome above, experiment r results in n_r possible outcomes.

Then there are $n_1n_2\cdots n_r$ possible outcomes for the r experiments.



EXAMPLE 1.4
How many possible outcomes are there when a die and a coin are thrown together?

Solution:
Note that for

- experiment 1: throwing a die, there are 6 possible outcomes: {1,2,3,4,5,6}.
- experiment 2: throwing a coin, with each outcome of experiment 1, there are 2 possible outcomes: $\{H, T\}$.

So altogether there are $6 \times 2 = 12$ possible outcomes.

In fact, the sample space is given by

$$S = \{(x, y) : x = 1, \ldots, 6; y = H \text{ or } T\}.$$

ADDITION PRINCIPLE
Suppose that an experiment can be performed by k different procedures.

Procedure 1 can be carried out in n_1 ways;

Procedure 2 can be carried out in n_2 ways;

… … … … …

Procedure k can be carried out in n_k ways.

Suppose that the "ways" under different procedures *do not overlap*. Then the total number of ways we can perform the experiment is

$$n_1 + n_2 + \ldots + n_k.$$

EXAMPLE 1.5 (ORCHARD ROAD)
We can take the MRT or bus from home to Orchard road. Suppose there are three bus routes and two MRT routes. How many ways can we go from home to Orchard road?

Solution:
Consider the trip from home to Orchard road as an experiment. Two procedures can used to complete the experiment:

Procedure 1: take MRT – 2 ways.

Procedure 2: take bus – 3 ways.

These ways do not overlap. So the total number of ways we can go from home to Orchard road is $2 + 3 = 5$.

PERMUTATION
A **permutation** is a selection and arrangement of r objects out of n. In this case, *order* is taken into consideration.

The number of ways to choose and arrange r objects out of n, where $r \leq n$, is denoted by P_r^n, where

$$P_r^n = \frac{n!}{(n-r)!} = n(n-1)(n-2)\dots(n-(r-1)).$$

obj 1	obj 2	obj 3	$\cdots$	obj r
n ways	$(n-1)$ ways	$(n-2)$ ways	$\cdots$	$(n-(r-1))$ ways

REMARK
When $r = n$, $P_n^n = n!$.

Essentially, it is the number of ways to arrange n objects in order.

EXAMPLE 1.6
Find the number of possible four-letter code words in which all letters are different.

Solution:
Note that there are $n = 26$ alphabets, and $r = 4$ in our case.

So the number of possible four-letter code words is

$$P_4^{26} = (26)(25)(24)(23) = 358800.$$

COMBINATION
A **combination** is a selection of r objects out of n, *without regard to the order*.

The number of combinations of choosing r objects out of n, denoted by C_r^n or $\binom{n}{r}$, is given by as

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

Note that this formula immediately implies $\binom{n}{r} = \binom{n}{n-r}$.

The derivation is as follows.

(A) Thinking in terms of permutation, *the number of ways to choose and arrange r objects out of n* is P_r^n.

(B) On the other hand, the same permutation task can be achieved by conducting the following two experiments sequentially:

(1) Select r objects out of n without regard to the order: $\binom{n}{r}$ ways.

(2) For each such combination, permute its r objects: P_r^r ways.

(C) Therefore, by the multiplication rule, *the number of ways to choose and arrange r objects out of n* is $\binom{n}{r} \times P_r^r$.

(D) As a consequence, $\binom{n}{r} \times P_r^r = P_r^n$, and so we obtain

$$\binom{n}{r} = \frac{P_r^n}{P_r^r} = \frac{n!/(n-r)!}{r!} = \frac{n!}{r!(n-r)!}.$$

EXAMPLE 1.7
From 4 women and 3 men, find the number of committees of size 3 that can be formed with 2 women and 1 man.

Solution:
The number of ways to select 2 women from 4 is $\binom{4}{2} = 6$.

The number of ways to select 1 man from 3 is $\binom{3}{1} = 3$.

By the multiplication rule, the number of committees formed with 2 women and 1 man is

$$\binom{4}{2} \times \binom{3}{1} = 6 \times 3 = 18.$$

4 PROBABILITY

Intuitively, the term probability is understood as the chance or how likely a certain event may occur.

More specifically, let A be an event in an experiment. We typically associate a number, called probability, to quantify how likely the event A occurs. This is denoted as $P(A)$.

Let us now investigate how we can obtain such a number.

You will discover that the fundamental concept of probability is extended from an idea based on intuition to a rigorous, abstract, and advanced mathematical theory. It has also wide applications in various scientific disciplines.

INTERPRETATION OF PROBABILITY: RELATIVE FREQUENCY
Suppose that we repeat an experiment E for a total of n times.

Let n_A be the number of times that the event A occurs.

Then $f_A = n_A/n$ is called the **relative frequency** of the event A in the n repetitions of E.

Clearly, f_A may not equal to $P(A)$ exactly. However when n grows large, we expect f_A to be close to $P(A)$; in the sense that $f_A \approx P(A)$. Or mathematically,

$$f_A \to P(A), \quad \text{as } n \to \infty.$$

Thus f_A "mimics" $P(A)$, and has the following properties:

(a) $0 \leq f_A \leq 1$.

(b) $f_A = 1$ if A occurs in every repetition.

(c) If A and B are mutually exclusive events, $f_{A\cup B} = f_A + f_B$.

Extending this idea, we can define **probability on a sample space** mathematically.

AXIOMS OF PROBABILITY
Probability, denoted by $P(\cdot)$, is a **function** on the collection of events of the sample space S, satisfying:

Axiom 1. For any event A,
$$0 \leq P(A) \leq 1.$$

Axiom 2. For the sample space,
$$P(S) = 1.$$

Axiom 3. For any two mutually exclusive events A and B, that is, $A \cap B = \emptyset$,
$$P(A \cup B) = P(A) + P(B).$$

EXAMPLE 1.8
Let H denote the event of getting a head when a coin is tossed. Find $P(H)$, if

(i) the coin is fair;

(ii) the coin is biased and a head is twice as likely to appear as a tail.

Solution:
The sample space is $S = \{H, T\}$.

(i) "The coin is fair" means that $P(H) = P(T)$.

The events $\{H\}$ and $\{T\}$ are mutually exclusive. Thus based on Axioms 2 and 3, we have

$$1 = P(S) = P(\{H\} \cup \{T\}) = P(H) + P(T) = 2P(H).$$

This gives $P(H) = 1/2$.

(ii) "A head is twice likely to appear as a tail" means $P(H) = 2P(T)$; therefore

$$1 = P(S) = P(\{H\} \cup \{T\}) = P(H) + P(T) = 3P(T).$$

This gives $P(T) = 1/3$ and $P(H) = 2/3$.

Basic Properties of Probability

Using the axioms, we can derive the following propositions.

PROPOSITION 2
The probability of the empty set $\emptyset$ is $P(\emptyset) = 0$.

Proof Since $\emptyset \cap \emptyset = \emptyset$ and $\emptyset = \emptyset \cup \emptyset$, applying Axiom 3 leads to

$$P(\emptyset) = P(\emptyset \cup \emptyset) = P(\emptyset) + P(\emptyset) = 2P(\emptyset).$$

This implies that $P(\emptyset) = 0$. ✠

PROPOSITION 3
If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, that is $A_i \cap A_j = \emptyset$ for any $i \neq j$, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$$

Proof This proposition can be established easily using induction and Axiom 3. ✠

PROPOSITION 4
For any event A, we have

$$P(A') = 1 - P(A).$$

Proof Since $S = A \cup A'$ and $A \cap A' = \emptyset$, based on Axioms 2 and 3, we have

$$1 = P(S) = P(A \cup A') = P(A) + P(A').$$

The result follows. ✠

PROPOSITION 5
For any two events A and B,

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

Proof Based on the properties

$$A = (A \cap B) \cup (A \cap B') \quad \text{and} \quad (A \cap B) \cap (A \cap B') = \emptyset,$$

we have

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

✠

PROPOSITION 6
For any two events A and B,

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

Proof Based on the event operations

$$A \cup B = B \cup (A \cap B') \quad \text{and} \quad B \cap (A \cap B') = \emptyset,$$

and Proposition 5 which states

$$P(A \cap B') = P(A) - P(A \cap B),$$

we have

$$P(A \cup B) = P(B) + P(A \cap B') = P(B) + P(A) - P(A \cap B).$$

✠

PROPOSITION 7
If $A \subset B$, then $P(A) \leq P(B)$.

Proof Since $A \subset B$, we have $A \cup B = B$. Also, we have

$$A \cup B = A \cup (B \cap A') \quad \text{and} \quad A \cap (B \cap A') = \emptyset.$$

Thus we obtain

$$P(B) = P(A \cup B) = P(A \cup (B \cap A')) = P(A) + P(B \cap A') \geq P(A).$$

✠

EXAMPLE 1.9
A retail establishment accepts either the American Express or the VISA credit card.

A total of 24% of its customers carry an American Express card, 61% carry a VISA card, and 11% carry both.

What is the probability that a customer carries a credit card that the establishment will accept?

Solution:
Let

$$A = \{\text{the customer carries an American Express Card}\}$$

and

$$V = \{\text{the customer carries an VISA Card}\}.$$

Then

$$P(A) = 0.24, \quad P(V) = 0.61, \quad P(A \cap V) = 0.11.$$

The question asked for $P(A \cup V)$, which is given as

$$P(A \cup V) = P(A) + P(V) - P(A \cap V) = 0.24 + 0.61 - 0.11 = 0.74.$$

FINITE SAMPLE SPACE WITH EQUALLY LIKELY OUTCOMES
Consider a sample space $S = \{a_1, a_2, \ldots, a_k\}$.

Assume that all outcomes in the sample space are **equally likely** to occur, i.e.,

$$P(a_1) = P(a_2) = \cdots = P(a_k).$$

Then for any event $A \subset S$,

$$P(A) = \frac{\text{number of sample points in } A}{\text{number of sample points in } S}.$$

EXAMPLE 1.10
A box contains 50 bolts and 150 nuts. Half of the bolts and half of the nuts are rusted.

If one item is chosen at random, what is the probability that it is rusted or is a bolt?

Solution:
We define the following events

$$A = \{\text{the item is rusted}\}, \quad B = \{\text{the item is a bolt}\}, \quad S = \{\text{all the items}\}.$$

Since the item is selected at random, each of the 200 elements in S is equally likely to be chosen.

- A consists of $25+75=100$ elements;
- B consists of 50 elements; and
- $A \cap B$ consists of 25 elements.

These give

$$P(A) = 100/200 \quad , \quad P(B) = 50/200 \quad , \quad P(A \cap B) = 25/200.$$

Therefore the required probability is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 5/8.$$

5 CONDITIONAL PROBABILITY

Sometimes we need to compute the probability of some events when some **partial information** is available.

Specifically, we might need to compute the probability of an event B, given that we have the information "an event A has occurred".

Mathematically, we denote

$$P(B|A)$$

as the conditional probability of the event B, given that event A has occurred.

DEFINITION 8 (CONDITIONAL PROBABILITY)
*For any two events A and B with $P(A) > 0$, the **conditional probability** of B given that A has occurred is defined by*

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

EXAMPLE 1.11
A fair die is rolled twice.

(i) What is the probability that the sum of the 2 rolls is even?

(ii) Given that the first roll is a 5, what is the (conditional) probability that the sum of the 2 rolls is even?

Solution:

We define the following events:

$$\begin{aligned} B &= \{\text{the sum of the 2 rolls is even}\}, \\ A &= \{\text{the first roll is a 5}\}. \end{aligned}$$

(i) The sample space is given by

		2nd roll					
		1	2	3	4	5	6
	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
1st roll	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

It is easy to see that $P(B) = 18/36$.

(ii) Since we know that A has already happened, we can just look at the fifth row:

		2nd roll					
		1	2	3	4	5	6
1st roll	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)

We are interested to look for instances along this row that gives an even sum. So $P(B|A) = 3/6$.

Alternatively, we can use the formula:

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{\frac{3}{36}}{\frac{6}{36}}.$$

REMARK (REDUCED SAMPLE SPACE)
$P(B|A)$ is read as:

"the conditional probability that B occurs given that A has occurred."

Since we know that A has occurred, regard A as our new, or **reduced sample space**.

The conditional probability that the event B given A will equal the probability of $A \cap B$ relative to the probability of A.

MULTIPLICATION RULE
Starting from the definition of conditional probability, and rearranging the terms, we have

$$\begin{aligned} & P(A \cap B) = P(A)P(B|A), \quad \text{if } P(A) \neq 0 \\ \text{or} \quad & P(A \cap B) = P(B)P(A|B), \quad \text{if } P(B) \neq 0. \end{aligned}$$

This is known as the **Multiplication Rule**.

INVERSE PROBABILITY FORMULA
The multiplication rule together with the definition of the conditional probability gives us:

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

This is known as the **Inverse Probability Formula**.

EXAMPLE 1.12
Deal 2 cards from a regular playing deck without replacement. What is the probability that both cards are aces?

Solution:

$$\begin{aligned} P(\text{both aces}) &= P(\text{1st card is ace and 2nd card is ace}) \\ &= P(\text{1st card ace}) \cdot P(\text{2nd card ace}|\text{1st card ace}) \\ &= \frac{4}{52} \cdot \frac{3}{51} = \frac{1}{221}. \end{aligned}$$

6 INDEPENDENCE

We saw several examples where conditioning on one event changes our beliefs about the probability of another event.

In this section, we discuss the important concept of independence, where learning that the event B occurred gives us no information that would change our probabilities for another event A occurring.

DEFINITION 9 (INDEPENDENCE)
*Two events A and B are **independent** if and only if*

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

We denote this by $A \perp B$.

*If A and B are not independent, they are said to be **dependent**, denoted by $A \not\perp B$.*

REMARK
If $P(A) \neq 0$, $A \perp B$ if and only if $P(B|A) = P(B)$.

This follows from the definition of conditional probability –

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

Intuitively, this is the same as saying that A and B are independent if the knowledge of A does not change the probability of B.

Likewise, if $P(B) \neq 0$, $A \perp B$ if and only if $P(A|B) = P(A)$.

EXAMPLE 1.13
Suppose we roll 2 fair dice.

(i) Let

$$A_6 = \{\text{the sum of two dice is 6}\}, \quad B = \{\text{the first die equals 4}\}.$$

Thus

$$P(A_6) = 5/36, \quad P(B) = 6/36 = 1/6 \quad \text{and} \quad P(A_6 \cap B) = 1/36.$$

As $P(A_6 \cap B) \neq P(A_6)P(B)$, we say that A_6 and B are **dependent**.

(ii) Let

$$A_7 = \{\text{the sum of two dice is 7}\}.$$

Then

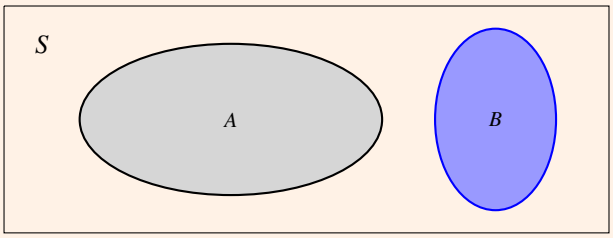
$$P(A_7 \cap B) = 1/36, \quad P(A_7) = 1/6 \quad \text{and} \quad P(B) = 1/6.$$

As $P(A_7 \cap B) = P(A_7)P(B)$, we say that A_7 and B are **independent**.

INDEPENDENT VS MUTUALLY EXCLUSIVE
Independence and **mutually exclusivity** are totally different concepts:

$$A, B \text{ independent} \Leftrightarrow P(A \cap B) = P(A)P(B)$$
$$A, B \text{ mutually exclusive} \Leftrightarrow A \cap B = \emptyset$$

"Mutually exclusivity" can be illustrated by a Venn diagram (like below), but we can not do that for "independence".



7 THE LAW OF TOTAL PROBABILITY

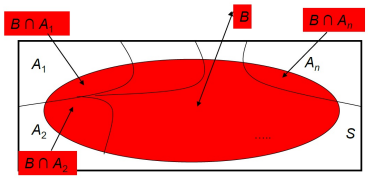
The definition of conditional probability has far-reaching consequences.

In this section we look at the Law of Total Probability (LOTP), which relates conditional probability to unconditional probability.

DEFINITION 10 (PARTITION)
*If $A_1, A_2, \ldots, A_n$ are mutually exclusive events and $\cup_{i=1}^{n} A_i = S$, we call $A_1, A_2, \ldots, A_n$ a **partition** of S.*

THEOREM 11 (LAW OF TOTAL PROBABILITY)
Suppose $A_1, A_2, \ldots, A_n$ is a partition of S. Then for any event B, we have

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



SPECIAL CASE: LAW OF TOTAL PROBABILITY
For any events A and B, we have

$$P(B) = P(A)P(B|A) + P(A')P(B|A').$$

EXAMPLE 1.14 (FRYING FISH)
At a nasi lemak stall, the chef and his assistant take turns to fry fish.

The chef burns his fish with probability 0.1, his assistant burns his fish with probability 0.23.

If the chef is frying fish 80% of the time, what is the probability that the fish you order is burnt?

Solution:
Let

$$\begin{aligned} B &= \{\text{the fish is burnt}\}, \\ C &= \{\text{the fish is fried by the chef}\}. \end{aligned}$$

We then need to compute $P(B)$. Using the Law of Total Probability,

$$P(B) = P(C)P(B|C) + P(C')P(B|C') = 0.8 \times 0.1 + 0.2 \times 0.23.$$

8 BAYES' THEOREM

We now discuss Bayes' Theorem (or Bayes' Rule), which will allow us to relate $P(A|B)$ to $P(B|A)$ and compute conditional probabilities in a wide range of problems.

THEOREM 12 (BAYES' THEOREM)
Let $A_1, A_2, \ldots, A_n$ be a partition of S, then for any event B and $k = 1, 2, \ldots, n$,

$$P(A_k|B) = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}.$$

Proof Bayes' Theorem can be derived based on the definition of conditional probability, the Multiplication Rule, and the Law of Total Probability.

In particular,

$$\begin{aligned} P(A_k|B) &= \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(B \cap A_i)} \\ &= \frac{P(A_k)P(B|A_k)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}. \end{aligned}$$

✠

SPECIAL CASE: BAYES' THEOREM
Let us consider a special case of Bayes' Theorem when $n = 2$.

$\{A, A'\}$ becomes a partition of S, and we have

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

EXAMPLE 1.15
The previous formula is practically meaningful.

For example, consider the events

$$A = \text{disease status of a person}, \quad B = \text{symptom observed}.$$

Then

- $P(A)$: probability of a disease in general;
- $P(B|A)$: if diseased, probability of observing symptom;
- $P(A|B)$: if symptom observed, probability of diseased.

EXAMPLE 1.16
Historically, we observe the collapse of some newly constructed house.

The chance that the design of the house is faulty is 1%. If the design is faulty, the chance that the house collapses is 75%; otherwise, the chance is 0.01%.

We observe that a newly constructed house collapsed, what is the probability that the design is faulty?

Solution:

Let

$$B = \{\text{The design is faulty}\}, \quad A = \{\text{The house collapses}\}.$$

We then have

$$P(B) = 0.01, \quad P(A|B) = 0.75, \quad \text{and} \quad P(A|B') = 0.0001.$$

The question asked for $P(B|A)$. We will compute it using Bayes' Theorem.

The denominator can be computed using the Law of Total Probability:

$$\begin{aligned} P(A) &= P(B)P(A|B) + P(B')P(A|B') \\ &= (0.01)(0.75) + (0.99)(0.0001) = 0.007599. \end{aligned}$$

The numerator is

$$P(A|B)P(B) = 0.75(0.01) = 0.0075.$$

As a consequence,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = 0.9870.$$

Two

Random Variables

1 Definition of a Random Variable

Often, when an experiment is performed, we are interested in some function (numerical characteristic) of the outcome, rather than the actual outcome itself.

For example,

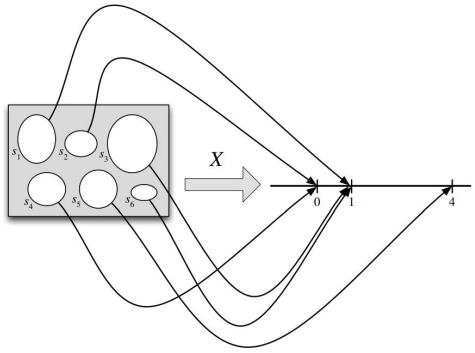
- in an experiment involving the examination of 100 electronic components, our interest is in the number of defective components.
- in an experiment of flipping a coin 100 times, our interest is in the number of heads obtained, instead of the "H" and "T" sequence.

This motivates us to assign numerical values to (possible) outcomes of an experiment.

Definition 1 (Random Variable)
Let S be the sample space of an experiment. A function X, which assigns a real number to every $s \in S$ is called a ***random variable****.*

Remark
So a random variable X is a function from S to $\mathbb{R}$:

$$X : S \mapsto \mathbb{R}.$$



A random variable maps the sample space into the real line. The random variable X depicted here is defined on a sample space with 6 elements $\{s_1, s_2, \ldots, s_6\}$, and has possible values 0, 1, and 4.

DEFINITION 2 (RANGE SPACE)
*The **range space** of X is the set of real numbers*

$$R_X = \{x | x = X(s), s \in S\}.$$

Each possible value x of X corresponds to an event that is a subset or element of the sample space S.

EXAMPLE 2.1
Let $S = \{HH, HT, TH, TT\}$ be the sample space associated with the experiment of flipping two fair coins.

Define the random variable:

$$X = \text{number of heads obtained.}$$

Note that X is a *function* from S to $\mathbb{R}$, the set of real numbers:

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

The range of X is $R_X = \{0, 1, 2\}$.

NOTATIONS

(i) We use upper case letters $X, Y, Z, X_1, X_2, \ldots$ to denote **random variables**.

(ii) We use lower case letters $x, y, z, x_1, x_2, \ldots$ to denote their **observed values** in the experiment.

(iii) The set $\{X = x\} = \{s \in S : X(s) = x\}$ is a subset of S.

(iv) If A is a subset of $\mathbb{R}$, the set $\{X \in A\} = \{s \in S : X(s) \in A\}$ is a subset of S.

(v) With the above expressions, we define $P(X = x)$ and $P(X \in A)$ as

$$\begin{aligned} P(X = x) &= P(\{s \in S : X(s) = x\}); \\ P(X \in A) &= P(\{s \in S : X(s) \in A\}). \end{aligned}$$

EXAMPLE 2.2

We revisit Example 2.1, where $S = \{HH, HT, TH, TT\}$ is the sample space associated with flipping two fair coins, and X is the number of heads obtained.

We then have

$$\begin{aligned} &\{X = 0\} = \{TT\}; && \{X = 1\} = \{HT, TH\}; \\ &\{X = 2\} = \{HH\}; && \{X \geq 1\} = \{HT, TH, HH\}. \end{aligned}$$

Thus

$$\begin{aligned} &P(X = 0) = P(TT) = 1/4; && P(X = 1) = P(\{HT, TH\}) = 2/4; \\ &P(X = 2) = P(HH) = 1/4; && P(X \geq 1) = P(\{HT, TH, HH\}) = 3/4. \end{aligned}$$

We can then summarize the probabilities of the random variable X using a table:

x	0	1	2
$P(X = x)$	1/4	1/2	1/4

2 Probability Distribution

There are two main types of random variables used in practice: discrete and continuous.

Let us denote by X the random variable, and it's range by R_X. For a

- **discrete random variable**, the number of values in R_X is **finite** or **countable**. That is, we can write $R_X = \{x_1, x_2, x_3, \ldots\}$.
- **continuous random variable**, R_X is an **interval** or a **collection of intervals**.

Discrete random variable

Consider a discrete random variable X with $R_X = \{x_1, x_2, x_3, \ldots\}$.

For each $x \in R_X$, let $P(X = x)$ be the probability that X takes the value x.

Definition 3 (Probability mass function)
For a discrete random variable X, define

$$f(x) = \begin{cases} P(X = x), & \text{for } x \in R_X; \\ 0, & \text{for } x \notin R_X. \end{cases}$$

*Then $f(x)$ is known as the **probability function (pf)**, or **probability mass function (pmf)** of X.*

*The collection of pairs $(x_i, f(x_i)), i = 1, 2, 3, \ldots$, is called the **probability distribution** of X.*

Properties of the Probability mass function
The probability mass function $f(x)$ of a discrete random variable **must** satisfy:

(1) $f(x_i) \geq 0$ for all $x_i \in R_X$;

(2) $f(x) = 0$ for all $x \notin R_X$;

(3) $\sum_{i=1}^{\infty} f(x_i) = 1$, or $\sum_{x_i \in R_X} f(x_i) = 1$.

For any set $B \subset \mathbb{R}$, we have

$$P(X \in B) = \sum_{x_i \in B \cap R_X} f(x_i).$$

EXAMPLE 2.3
We revisit Examples 2.1 and 2.2.

Recall that random variable X is the number of heads observed when we flip two fair coins.

The probability function of X is given below

x	0	1	2
$f(x)$	1/4	1/2	1/4

Note that $f(x)$ satisfies

(1) $f(x_i) \geq 0$ for $x_i = 0, 1$, or 2;

(2) $f(x) = 0$ for $x \notin R_X$;

(3) $f(0) + f(1) + f(2) = 1$.

When $B = [1, \infty)$,

$$P(X \in B) = f(1) + f(2) = 3/4.$$

Continuous random variable

For a continuous random variable X, R_X is an interval or a collection of intervals.

We next define the **probability function (pf)**, or **probability density function (pdf)**, to quantify the probability that X is in a certain range.

DEFINITION 4 (PROBABILITY DENSITY FUNCTION)
*The **probability density function** of a continuous random variable X, denoted by $f(x)$, is a function that satisfies:*

(1) $f(x) \geq 0$ for all $x \in R_X$; and $f(x) = 0$ for $x \notin R_X$;

(2) $\int_{R_X} f(x)\,\mathrm{dx} = 1$;

(3) For any a and b such that $a \leq b$,

$$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{dx}.$$

REMARK

- Note that Condition (2) is equivalent to

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

since $f(x) = 0$ for $x \notin R_X$.

- For any specific value x_0, we have

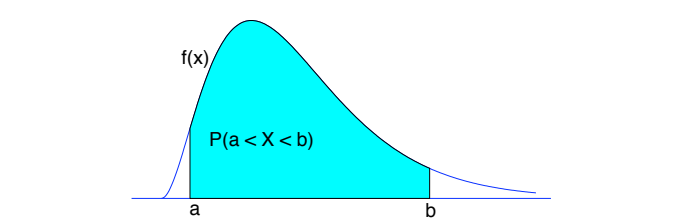
$$P(X = x_0) = \int_{x_0}^{x_0} f(x)\,\mathrm{dx} = 0.$$

The gives an example of "$P(A) = 0$, but A is not necessarily $\emptyset$."

- Furthermore, we have

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{dx}\,.$$

They all represent the area under the graph of $f(x)$ between $x = a$ and $x = b$.



- To check that a function $f(x)$ is a probability density function, it suffices to check Conditions (1) and (2). Namely,

(1) $f(x) \geq 0$ for all $x \in R_X$; and $f(x) = 0$ for $x \notin R_X$.

(2) $\int_{R_X} f(x)\,\mathrm{dx} = 1.$

EXAMPLE 2.4
Let X be a continuous random variable with probability density function given by

$$f(x) = \begin{cases} cx, & \text{for } 0 < x < 1; \\ 0, & \text{otherwise .} \end{cases}$$

(i) Find the value of c;

(ii) Find $P(X \leq 1/2)$.

Solution:

(i) Since

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{dx} = \int_0^1 cx\,\mathrm{dx} = c \cdot \frac{x^2}{2}\Big|_0^1 = c/2,$$

we set $c/2 = 1$. This results in $c = 2$.

(ii)

$$P(X \leq 1/2) = \int_{-\infty}^{1/2} f(x)\,\mathrm{dx} = \int_0^{1/2} 2x\,\mathrm{dx} = 1/4.$$

3 CUMULATIVE DISTRIBUTION FUNCTION

Another function that describes the distribution of a random variable is the **cumulative distribution function (cdf)**.

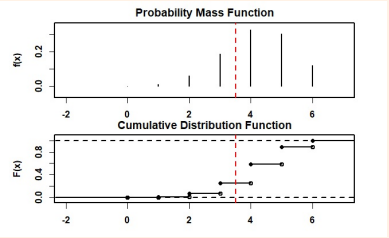
DEFINITION 5 (CUMULATIVE DISTRIBUTION FUNCTION)
*For any random variable X, we define its **cumulative distribution function (cdf)** by*

$$F(x) = P(X \leq x).$$

REMARK
This definition is applicable whether X is a discrete or continuous random variable.

CDF: DISCRETE RANDOM VARIABLE
If X is a discrete random variable, we have

$$\begin{aligned} F(x) &= \sum_{t \in R_X; t \le x} f(t) \\ &= \sum_{t \in R_X; t \le x} P(X = t) \end{aligned}$$



The cumulative distribution function of a discrete random variable is a step function.

For any two numbers $a < b$, we have

$$P(a \le X \le b) = P(X \le b) - P(X < a) = F(b) - F(a-),$$

where "$a-$" represents the "largest value in R_X that is smaller than a". Mathematically,

$$F(a-) = \lim_{x \uparrow a} F(x).$$

EXAMPLE 2.5
We revisit Examples 2.1 and 2.2. The random variable X is the number of heads observed when we flip two fair coins, and has the probability function

x	0	1	2
$f(x)$	1/4	1/2	1/4

We have

$$F(0) = f(0) = 1/4,\ F(1) = f(0) + f(1) = 3/4,\ F(2) = f(0) + f(1) + f(2) = 1.$$

We can therefore obtain the cumulative distribution function:

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/4, & 0 \le x < 1 \\ 3/4, & 1 \le x < 2 \\ 1, & 2 \le x \end{cases}.$$

EXAMPLE 2.6
Consider the cumulative distribution function from Example 2.5:

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/4, & 0 \le x < 1 \\ 3/4, & 1 \le x < 2 \\ 1, & 2 \le x \end{cases}.$$

Derive the corresponding probability function. [1]

Solution:
As $F(\cdot)$ only has four possible values, the distribution is a discrete distribution.

We obtain $R_X = \{0,1,2\}$, which are the jumping points of $F(\cdot)$. It is also the set where $f(x)$ is non-zero.

We have

$$\begin{aligned} f(0) &= P(X=0) = F(0) - F(0-) = 1/4 - 0 = 1/4; \\ f(1) &= P(X=1) = F(1) - F(1-) = 3/4 - 1/4 = 1/2; \\ f(2) &= P(X=2) = F(2) - F(2-) = 1 - 3/4 = 1/4. \end{aligned}$$

CDF: CONTINUOUS RANDOM VARIABLE
If X is a continuous random variable,

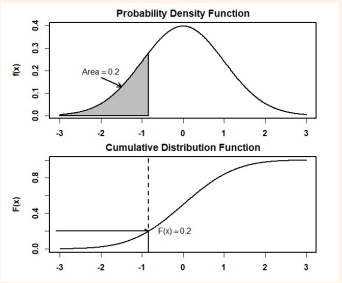
$$F(x) = \int_{-\infty}^{x} f(t)dt,$$

and

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{dx}}.$$

Further

$$P(a \le X \le b) = P(a < X < b) = F(b) - F(a).$$



EXAMPLE 2.7
Suppose the probability density function of a random variable X is

$$f(x) = \begin{cases} 2x, & 0 \le x < 1 \\ 0, & \text{elsewhere} \end{cases}.$$

The cumulative distribution function of X is then

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{dt} = \begin{cases} 0, & x < 0 \\ x^2, & 0 \le x < 1 \\ 1, & 1 \le x \end{cases}.$$

[1] Let us pretend for a while that the cumulative distribution function is the only information available for this distribution.

EXAMPLE 2.8
Consider the cumulative distribution function from Example 2.7:

$$F(x) = \begin{cases} 0, & x < 0 \\ x^2, & 0 \le x < 1 \\ 1, & 1 \le x \end{cases}.$$

Derive the corresponding probability function. [2]

Solution:
$F(x)$ is a cumulative distribution function for a continuous distribution, because when it is not equal to 0 and 1, it assumes different values in the interval $x \in [0,1)$.

$f(x) = 0$ when $x \notin [0,1)$ because $\frac{\mathrm{d}}{\mathrm{dx}}(0) = \frac{\mathrm{d}}{\mathrm{dx}}(1) = 0$.

$f(x) = \frac{\mathrm{d}}{\mathrm{dx}}(x^2) = 2x$ when $x \in [0,1)$.

REMARK

(i) No matter if X is discrete or continuous, $F(x)$ is non-decreasing. In the sense that for any $x_1 < x_2$, $F(x_1) \le F(x_2)$.

(ii) The probability function and cumulative distribution function have a one-to-one correspondence. That is, for any probability function given, the cumulative distribution function is uniquely determined; and vice versa.

(iii) The ranges of $F(x)$ and $f(x)$ satisfy:

- $0 \le F(x) \le 1$;
- for discrete distributions, $0 \le f(x) \le 1$;
- for continuous distributions, $f(x) \ge 0$, but **not necessary** that $f(x) \le 1$.

4 EXPECTATION AND VARIANCE

For a random variable X, one natural question to ask is: what is the **average value** of X, if the corresponding experiment is repeated many times?

For example, suppose X is the number obtained when we roll a die. We may want to know the average value obtained if we roll the die continuously.

Such an average, over the long run, is called the **mean** or **expectation** of X.

[2] Let us pretend for a while that the cumulative distribution function is the only information available for this distribution.

DEFINITION 6 (EXPECTATION: DISCRETE RANDOM VARIABLE)
*Let X be a discrete random variable with $R_X = \{x_1, x_2, x_3, \ldots\}$ and probability function $f(x)$. The **expectation** or **mean** of X is defined by*

$$E(X) = \sum_{x_i \in R_X} x_i f(x_i).$$

By convention, we also denote $\mu_X = E(X)$.

DEFINITION 7 (EXPECTATION: CONTINUOUS RANDOM VARIABLE)
*Let X be a continuous random variable with probability function $f(x)$. The **expectation** or **mean** of X is defined by*

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{dx} = \int_{x \in R_X} x f(x)\,\mathrm{dx}.$$

REMARK
The mean of X is *not necessarily* a possible value of the random variable X.

EXAMPLE 2.9
Suppose we toss a fair die and the upper face is recorded as X. We have

$$P(X = k) = 1/6, \quad \text{for } k = 1, 2, \ldots, 6,$$

and

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5.$$

Here we have a random variable whose mean is not a value that X assumes.

EXAMPLE 2.10
The probability density function of weekly gravel sales X is

$$f(x) = \begin{cases} \frac{3}{2}(1 - x^2), & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}.$$

We then have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{dx} = \int_0^1 \tfrac{3x}{2}(1 - x^2)\,\mathrm{dx} \\ &= \tfrac{3}{2}\int_0^1 (x - x^3)\,\mathrm{dx} = \tfrac{3}{2}\left(\tfrac{x^2}{2} - \tfrac{x^4}{4}\right)\Big|_0^1 = 3/8. \end{aligned}$$

PROPERTIES OF EXPECTATION

(1) Let X be a random variable, and let a and b be any real numbers. Then

$$E(aX+b) = aE(X)+b.$$

(2) Let X and Y be two random variables. We have

$$E(X+Y) = E(X)+E(Y).$$

(3) Let $g(\cdot)$ be an arbitrary function.

- If X is a **discrete** random variable with probability mass function $f(x)$ and range R_X,
$$E[g(X)] = \sum_{x\in R_X} g(x)f(x).$$
- If X is a **continuous** random variable with probability density function $f(x)$ and range R_X,
$$E[g(X)] = \int_{R_X} g(x)f(x)\,\mathrm{d}x\,.$$

Variance

Let $g(x) = (x-\mu_X)^2$, then $E[g(X)]$ is defined as the **variance** for X.

DEFINITION 8 (VARIANCE)
*Let X be a random variable. The **variance** of X is defined as*

$$\sigma_X^2 = V(X) = E(X-\mu_X)^2.$$

REMARK

- This definition is applicable whether X is discrete or continuous.
- If X is a **discrete** random variable with probability mass function $f(x)$ and range R_X,
$$V(X) = \sum_{x\in R_X} (x-\mu_X)^2 f(x).$$
- If X is a **continuous** random variable with probability density function $f(x)$,
$$V(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)\,\mathrm{d}x\,.$$
- $V(X) \geq 0$ for any X. Equality holds if and only $P(X = E(X)) = 1$, that is, when X is a **constant**.

- Let a and b be any real numbers, then $V(aX+b) = a^2V(X)$.
- The variance can also be computed by an alternative formula:

$$V(X) = E(X^2) - [E(X)]^2.$$

- The positive square root of the variance is defined as the **standard deviation** of X:

$$\sigma_X = \sqrt{V(X)}.$$

EXAMPLE 2.11
Let the probability function of a random variable X be given by

x	-1	0	1	2
$f(x)$	$1/8$	$2/8$	$1/8$	$4/8$

Find $E(X)$ and $V(X)$.

Solution:
The mean is given as

$$\begin{aligned} E(X) &= \sum_{x\in R_X} xf(x) \\ &= (-1)\left(\tfrac{1}{8}\right) + 0\left(\tfrac{2}{8}\right) + 1\left(\tfrac{1}{8}\right) + 2\left(\tfrac{4}{8}\right) = 1. \end{aligned}$$

The variance is given as

$$\begin{aligned} V(X) &= \sum_{x\in R_X} [x-E(X)]^2 f(x) = \sum_{x\in R_X} [x-1]^2 f(x) \\ &= (-1-1)^2\left(\tfrac{1}{8}\right) + (0-1)^2\left(\tfrac{2}{8}\right) \\ &\quad +(1-1)^2\left(\tfrac{1}{8}\right) + (2-1)^2\left(\tfrac{4}{8}\right) = \tfrac{5}{4}. \end{aligned}$$

EXAMPLE 2.12
Denote by X the amount of time that a book on reserve at the library is checked out by a randomly selected student. Suppose X has the probability density function

$$f(x) = \begin{cases} x/2, & 0 \le x < 2, \\ 0, & \text{otherwise.} \end{cases}$$

Compute $E(X)$, $V(X)$, and σ_X.

Solution:
We can compute

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{dx} = \int_0^2 x \cdot x/2\,\mathrm{dx} = \left.\frac{x^3}{6}\right|_0^2 = 4/3;$$
$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{dx} = \int_0^2 x^2 \cdot x/2\,\mathrm{dx} = \left.\frac{x^4}{8}\right|_0^2 = 2.$$

Using $V(X) = E(X^2) - [E(X)]^2$, we obtain

$$V(X) = 2 - (4/3)^2 = 2/9 \quad \text{and} \quad \sigma_X = \sqrt{V(X)} = \sqrt{2}/3.$$

Three

Joint Distributions

1 Joint Distributions for Multiple Random Variables

Very often, we are interested in more than one random variables *simultaneously*.

- For example, an investigator might be interested in both the height (H) and the weight (W) of individuals from a certain population.
- Another investigator could be interested in both the hardness (H) and the tensile strength (T) of a piece of cold-drawn copper.

Definition 1 (Two-dimensional Random Vector)
Let E be an experiment and S be a corresponding sample space. Suppose X and Y are two functions each assigning a real number to each $s \in S$.

*We call (X,Y) a **two-dimensional random vector**, or a **two-dimensional random variable**.*

Definition 2 (Range Space)
*Similar to the one-dimensional situation, we can denote the **range space** of (X,Y) by*

$$R_{X,Y} = \{(x,y) \mid x = X(s), y = Y(s), s \in S\}.$$

The definitions above can be extended to more than two random variables.

Definition 3 (n-dimensional Random Vector)
*Let $X_1, X_2, \ldots, X_n$ be n functions each assigning a real number to **every outcome** $s \in S$.*

*We call $(X_1, X_2, \ldots, X_n)$ a **n-dimensional random vector**, or a **n-dimensional random variable**.*

We define the discrete and continuous two-dimensional random variables as follows.

DEFINITION 4
(X,Y) *is a* ***discrete two-dimensional random variable*** *if the number of possible values of* $(X(s),Y(s))$ *are finite or countable. That is, the possible values of* $(X(s),Y(s))$ *may be represented by*

$$(x_i,y_j), \quad i=1,2,3,\ldots;j=1,2,3,\ldots$$

(X,Y) *is a* ***continuous two-dimensional random variable*** *if the possible values of* $(X(s),Y(s))$ *can assume any value in some region of the Euclidean space* $\mathbb{R}^2$.

REMARK
We can view X and Y separately to judge whether (X,Y) is discrete or continuous.

- If both X and Y are discrete random variables, then (X,Y) is discrete.
- If both X and Y are continuous random variables, then (X,Y) is continuous.
- Clearly, there are other cases. For example, X is discrete, but Y is continuous. These are not the focus of this course.

EXAMPLE 3.1
Consider a TV set that needs to be serviced.

Let X be the age of the set, rounded to the nearest year, and Y be the numbers of defective components in the set.

Then (X,Y) is a discrete 2-dimensional random variable and its range space is given as

$$R_{X,Y}=\{(x,y)\mid x=0,1,2,\ldots;y=0,1,2,\ldots,n\},$$

where n is the total number of components in the TV.

For example, $(X,Y)=(5,3)$ means that the TV is 5 years old and has 3 defective components.

Joint Probability Function

We will now introduce the probability functions for discrete and continuous random vectors.

For the discrete random vector, similar to the one-dimensional case, we define its probability function by associating a number with each possible value of the random variable.

DEFINITION 5 (DISCRETE JOINT PROBABILITY FUNCTION)
Let (X,Y) be a 2-dimensional ***discrete*** *random variable. Its* ***joint probability (mass) function*** *is defined by*

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

for $(x,y) \in R_{X,Y}$.

PROPERTIES OF THE DISCRETE JOINT PROBABILITY FUNCTION
The joint probability mass function has the following properties:

(1) $f_{X,Y}(x,y) \geq 0$ for any $(x,y) \in R_{X,Y}$.

(2) $f_{X,Y}(x,y) = 0$ for any $(x,y) \notin R_{X,Y}$.

(3) $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} f_{X,Y}(x_i,y_j) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = 1.$
Equivalently, $\sum\sum_{(x,y)\in R_{X,Y}} f(x,y) = 1.$

(4) Let A be any subset of $R_{X,Y}$, then

$$P((X,Y) \in A) = \sum\sum_{(x,y)\in A} f_{X,Y}(x,y).$$

EXAMPLE 3.2
Find the value of k such that

$$f(x,y) = kxy, \quad \text{for } x = 1,2,3 \text{ and } y = 1,2,3,$$

can serve as a joint probability function.

Solution:
Note that $R_{X,Y} = \{(x,y) \mid x = 1,2,3; y = 1,2,3\}$, and

$$\begin{aligned} f(1,1) &= k, & f(1,2) &= 2k, & f(1,3) &= 3k, \\ f(2,1) &= 2k, & f(2,2) &= 4k, & f(2,3) &= 6k, \\ f(3,1) &= 3k, & f(3,2) &= 6k, & f(3,3) &= 9k. \end{aligned}$$

Using Property (3), we have

$$\begin{aligned} 1 &= \sum\sum_{(x,y)\in R_{X,Y}} f(x,y) \\ &= 1k+2k+3k+2k+4k+6k+3k+6k+9k. \end{aligned}$$

This results in $k = 1/36$.

DEFINITION 6 (CONTINUOUS JOINT PROBABILITY FUNCTION)
*Let (X,Y) be a 2-dimensional **continuous** random variable. Its **joint probability (density) function** is a function $f_{X,Y}(x,y)$ such that*

$$P((X,Y)\in D) = \iint_{(x,y)\in D} f_{X,Y}(x,y)\,\mathrm{dy}\,\mathrm{dx},$$

for any $D \subset \mathbb{R}^2$. More specifically,

$$P(a\le X\le b, c\le Y\le d) = \int_a^b \int_c^d f_{X,Y}(x,y)\,\mathrm{dy}\,\mathrm{dx}\,.$$

PROPERTIES OF THE CONTINUOUS JOINT PROBABILITY FUNCTION
The joint probability density function has the following properties:

(1) $f_{X,Y}(x,y)\ge 0$, for any $(x,y)\in R_{X,Y}$.

(2) $f_{X,Y}(x,y) = 0$, for any $(x,y)\notin R_{X,Y}$.

(3) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy} = 1$.

Equivalently, $\iint_{(x,y)\in R_{X,Y}} f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy} = 1$.

EXAMPLE 3.3
Find the value c such that $f(x,y)$ below can serve as a joint probability density function for a random variable (X,Y):

$$f(x,y) = \begin{cases} cx(x+y), & 0 \le x \le 1; 1 \le y \le 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Solution:
In order for $f(x,y)$ to be a probability density function, we need

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,\mathrm{dy}\,\mathrm{dx} = \int_0^1\int_1^2 cx(x+y)\,\mathrm{dy}\,\mathrm{dx} \\ &= c\int_0^1 x\left[x+\frac{1}{2}y^2\right]_1^2 \mathrm{dx} = c\int_0^1 x(x+1.5)\,\mathrm{dx} \\ &= c\left[\frac{1}{3}x^3 + 1.5\cdot\frac{1}{2}x^2\right]_0^1 = c\cdot\frac{13}{12}. \end{aligned}$$

This implies that $c = 12/13$.

2 MARGINAL AND CONDITIONAL DISTRIBUTIONS

We now consider the marginal distributions.

Put simply, the marginal distribution of X is the individual distribution of X, ignoring the value of Y.

DEFINITION 7 (MARGINAL PROBABILITY DISTRIBUTION)
Let (X,Y) be a two-dimensional random variable with joint probability function $f_{X,Y}(x,y)$. We define the ***marginal distribution*** *of X as follows.*

If Y is a discrete random variable, then for any x,

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

If Y is a continuous random variable, then for any x,

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

REMARK

- $f_Y(y)$ for Y is defined in the same way as that of X.
- We can view the marginal distribution as the "projection" of the 2D function $f_{X,Y}(x,y)$ to the 1D function.
- Intuitively, it is the distribution of X by ignoring the presence of Y.

  For example, consider a person from a certain community.

  - Suppose $X =$ body weight, $Y =$ height, and (X,Y) has joint distribution $f_{X,Y}(x,y)$.
  - The marginal distribution $f_X(x)$ of X is the **distribution of body weights for all people in the community**.
- $f_X(x)$ should not involve the variable y. This can be viewed from its definition: y is either summed out or integrated over.
- $f_X(x)$ is a **probability function**; so it satisfies all the properties of the probability function.

EXAMPLE 3.4

We revisit Example 3.2. The joint probability function is given by

$$f(x,y) = \frac{1}{36}xy, \quad \text{for } x = 1,2,3 \text{ and } y = 1,2,3.$$

Note that X has three possible values: 1, 2, and 3. The marginal distribution for X is given by

- for $x = 1$, $f_X(1) = f(1,1) + f(1,2) + f(1,3) = 6/36 = 1/6$.
- for $x = 2$, $f_X(2) = f(2,1) + f(2,2) + f(2,3) = 12/36 = 1/3$.
- for $x = 3$, $f_X(3) = f(3,1) + f(3,2) + f(3,3) = 18/36 = 1/2$.

For other values of x, $f_X(x) = 0$.
Alternatively, for each $x \in \{1,2,3\}$,

$$f_X(x) = \sum_y f(x,y) = \sum_{y=1}^{3} \frac{1}{36}xy = \frac{1}{36}x\sum_{y=1}^{3} y = \frac{1}{6}x.$$

DEFINITION 8 (CONDITIONAL DISTRIBUTION)

Let (X,Y) be a random variable with joint probability function $f_{X,Y}(x,y)$. Let $f_X(x)$ be the marginal probability function for X. Then for any x such that $f_X(x) > 0$, the ***conditional probability function of Y given $X = x$*** *is defined to be*

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

REMARK

- For any y such that $f_Y(y) > 0$, we can similarly define the **conditional distribution of X given $Y = y$** as

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

- $f_{Y|X}(y|x)$ is defined only for x such that $f_X(x) > 0$; likewise $f_{X|Y}(x|y)$ is defined only for y such that $f_Y(y) > 0$.
- The intuitive meaning of $f_{Y|X}(y|x)$: the distribution of Y given that the random variable X is observed to take the value x.
- Considering y as the variable (and x as a fixed value), $f_{Y|X}(y|x)$ is a probability function, so it must satisfy all the properties of a probability function.
- However, $f_{Y|X}(y|x)$ is not a probability function for x. This means that there is **NO** requirement that
  - $\int_{-\infty}^{\infty} f_{Y|X}(y|x)\,\mathrm{dx} = 1$, for X continuous; or
  - $\sum_x f_{Y|X}(y|x) = 1$, for X discrete.
- With this definition, we immediately have
  - If $f_X(x) > 0$, $f_{X,Y}(x,y) = f_X(x) f_{Y|X}(y|x)$.
  - If $f_Y(y) > 0$, $f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y)$.
- One immediate application of the conditional distribution is to compute, for continuous random variable,

$$\begin{aligned} P(Y \le y|X = x) &= \int_{-\infty}^{y} f_{Y|X}(y|x)\,\mathrm{dy}; \\ E(Y|X = x) &= \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,\mathrm{dy}. \end{aligned}$$

  Their interpretations are clear: the former is the probability that $Y \le y$, given $X = x$; the latter is the average value of Y given $X = x$.

  For the discrete case, the results can be similarly established, based on the definition of $f_{Y|X}(y|x)$.

EXAMPLE 3.5
We revisit Examples 3.2 and 3.4. The joint probability function for (X,Y) is given by

$$f(x,y) = {}^{xy}/_{36}, \quad \text{for } x = 1,2,3 \text{ and } y = 1,2,3.$$

The marginal probability function for X is

$$f_X(x) = {}^{x}/_{6}, \quad \text{for } x = 1,2,3.$$

Therefore $f_{Y|X}(y|x)$ is defined for any $x = 1,2,3$:

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{({}^{xy}/_{36})}{({}^{x}/_{6})} = {}^{y}/_{6}, \quad \text{for } y = 1,2,3.$$

We can also compute

$$\begin{aligned} P(Y = 2|X = 1) &= f_{Y|X}(2|1) = \frac{1}{6} \times 2 = 1/3; \\ P(Y \leq 2|X = 1) &= P(Y = 1|X = 1) + P(Y = 2|X = 1) \\ &= f_{Y|X}(1|1) + f_{Y|X}(2|1) = 1/6 + 1/3 = 1/2; \\ E(Y|X = 2) &= 1 \cdot f_{Y|X}(1|2) + 2 \cdot f_{Y|X}(2|2) + 3 \cdot f_{Y|X}(3|2) \\ &= 1 \cdot (1/6) + 2 \cdot (2/6) + 3 \cdot (3/6) = 7/3. \end{aligned}$$

3 INDEPENDENT RANDOM VARIABLES

We next discuss independence for random variables.

DEFINITION 9 (INDEPENDENT RANDOM VARIABLES)
Random variables X and Y are ***independent*** *if and only if for* ***any*** *x and y,*

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

Random variables $X_1, X_2, \ldots, X_n$ *are* ***independent*** *if and only if for* ***any*** $x_1, x_2, \ldots, x_n$,

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

REMARK

- The above definition is applicable whether (X,Y) is continuous or discrete.
- The "product feature" in the definition implies one necessary condition for independence: $R_{X,Y}$ needs to be a product space. In the sense that if X and Y are independent, for any $x \in R_X$ and any $y \in R_Y$, we have

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) > 0,$$

implying $R_{X,Y} = \{(x,y) | x \in R_X; y \in R_y\} = R_X \times R_Y$.

Conclusion:
If $R_{X,Y}$ is not a product space, then X and Y are not independent!

PROPERTIES OF INDEPENDENT RANDOM VARIABLES

Suppose X,Y are independent random variables.

(1) If A and B are arbitrary subsets of $\mathbb{R}$, the events $X \in A$ and $Y \in B$ are independent events in S. Thus

$$P(X \in A; Y \in B) = P(X \in A) P(Y \in B).$$

In particular, for any real numbers x,y,

$$P(X \leq x; Y \leq y) = P(X \leq x) P(Y \leq y).$$

(2) For arbitrary functions $g_1(\cdot)$ and $g_2(\cdot)$, $g_1(X)$ and $g_2(Y)$ are independent. For example,

- X^2 and Y are independent.
- $\sin(X)$ and $\cos(Y)$ are independent.
- e^X and $\log(Y)$ are independent.

(3) Independence is connected with conditional distribution.

- If $f_X(x) > 0$, then $f_{Y|X}(y|x) = f_Y(y)$.
- If $f_Y(y) > 0$, then $f_{X|Y}(x|y) = f_X(x)$.

EXAMPLE 3.6
The joint probability function of (X,Y) is given below.

x	y			$f_X(x)$
	1	3	5	
2	0.1	0.2	0.1	0.4
4	0.15	0.3	0.15	0.6
$f_Y(y)$	0.25	0.5	0.25	1

Are X and Y independent?

Solution:
We need to check that for every x and y combination, whether we have

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

For example, from the table, we have $f_{X,Y}(2,1) = 0.1$; $f_X(2) = 0.4$, $f_Y(1) = 0.25$. Therefore

$$f_{X,Y}(2,1) = 0.1 = 0.4 \times 0.25 = f_X(2)f_Y(1).$$

In fact, we can check for each $x \in \{2,4\}$ and $y \in \{1,3,5\}$ combination, the equality holds. Therefore X and Y are independent.

4 EXPECTATION AND COVARIANCE

Similar to one dimensional random variable, we can talk about the expectation of a random vector.

DEFINITION 10 (EXPECTATION)
Consider any two variable function $g(x,y)$.

If (X,Y) is a discrete random variable,

$$E(g(X,Y)) = \sum_x \sum_y g(x,y) f_{X,Y}(x,y).$$

If (X,Y) is a continuous random variable,

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

If we let

$$g(X,Y) = (X - E(X))(Y - E(Y)) = (X - \mu_X)(Y - \mu_y),$$

the expectation $E[g(X,Y)]$ leads to the covariance of X and Y.

DEFINITION 11 (COVARIANCE)
*The **covariance** of X and Y is defined to be*

$$\operatorname{cov}(X,Y) = E[(X - E(X))(Y - E(Y))].$$

REMARK
If X and Y are discrete random variables,

$$\operatorname{cov}(X,Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y).$$

If X and Y are continuous random variables,

$$\operatorname{cov}(X,Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x,y)\,\mathrm{dx}\,\mathrm{dy}.$$

PROPERTIES OF THE COVARIANCE

The covariance has the following properties.

(1) $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$.

This is true because

$$\begin{aligned}\text{cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] = E[XY - Y\mu_X - X\mu_Y + \mu_X\mu_Y] \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y = E(XY) - \mu_X\mu_Y.\end{aligned}$$

(2) If X and Y are independent, then $\text{cov}(X,Y) = 0$. However, $\text{cov}(X,Y) = 0$ does not imply that X and Y are independent.

Take note that the two statements can be summarised as:

(i) $X \perp Y \Rightarrow \text{cov}(X,Y) = 0$;

(ii) $X \perp Y \not\Leftarrow \text{cov}(X,Y) = 0$.

For (i), note that if X and Y are independent, then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. So

$$\begin{aligned}E(XY) &= \sum_i \sum_j x_i y_j f_{X,Y}(x_i,y_j) = \sum_i \sum_j x_i y_j f_X(x_i) f_Y(y_j) \\ &= \sum_i x_i f_X(x_i) \sum_j y_j f_Y(y_j) = E(X)E(Y).\end{aligned}$$

(3) $\text{cov}(aX+b, cY+d) = ac \cdot \text{cov}(X,Y)$.

This can be derived using the following 3 formulas:

(i) $\text{cov}(X,Y) = \text{cov}(Y,X)$;

(ii) $\text{cov}(X+b,Y) = \text{cov}(X,Y)$;

(iii) $\text{cov}(aX,Y) = a\,\text{cov}(X,Y)$.

(4) $V(aX+bY) = a^2V(X) + b^2V(Y) + 2ab \cdot \text{cov}(X,Y)$.

This can be derived using the following 2 formulas:

(i) $V(aX) = a^2V(X)$;

(ii) $V(X+Y) = V(X) + V(Y) + 2\,\text{cov}(X,Y)$.

EXAMPLE 3.7
We are given the joint distribution for (X,Y):

x	y				$f_X(x)$
	0	1	2	3	
0	1/8	1/4	1/8	0	1/2
1	0	1/8	1/4	1/8	1/2
$f_Y(y)$	1/8	3/8	3/8	1/8	1

(i) Find $E(Y-X)$.

(ii) Find $\text{cov}(X,Y)$.

Solution:

(i) Method 1:

$$\begin{aligned} E(Y-X) &= (0-0)(1/8)+(1-0)(1/4)+(2-0)(1/8) \\ &\quad +\ldots+(3-1)(1/8)=1. \end{aligned}$$

Method 2:

$$E(Y-X)=E(Y)-E(X)=1.5-0.5=1,$$

where

$$\begin{aligned} E(Y) &= 0\cdot(1/8)+1\cdot(3/8)+2\cdot(3/8)+3\cdot(1/8)=1.5 \\ E(X) &= 0\cdot(1/2)+1\cdot(1/2)=0.5. \end{aligned}$$

(ii) We use $\text{cov}(X,Y)=E(XY)-E(X)E(Y)$ to compute. Note that we have computed $E(X)$ and $E(Y)$ in Part (i).

$$\begin{aligned} E(XY) &= (0)(0)(1/8)+(0)(1)(1/4)+(0)(2)(1/8) \\ &\quad +\ldots+(1)(3)(1/8)=1. \end{aligned}$$

Therefore

$$\text{cov}(X,Y)=E(XY)-E(X)E(Y)=1-(0.5)(1.5)=0.25.$$

Four

Special Probability Distributions

1 Discrete Distributions

Situations involving uncertainty and probability fall into certain broad classes, and we can use the same set of rules and principles for all situations within a class.

So it is beneficial for us to study whole classes of discrete random variables that arise frequently in applications.

Remark
Recall that for a discrete random variable X, the number of possible values in the range space R_X is either **finite** or **countable**.

Then the elements of R_X can be listed as $x_1, x_2, x_3, \ldots$.

Discrete Uniform Distribution

One of the simplest class of a discrete random variable is the discrete uniform distribution.

Definition 1 (Discrete Uniform Distribution)
If a random variable X assumes the values $x_1, x_2, \ldots, x_k$ with equal probability, then X follows a ***discrete uniform distribution****.*

The probability mass function for X is given by

$$f_X(x) = \begin{cases} \dfrac{1}{k}, & x = x_1, x_2, \ldots, x_k; \\ 0, & \textit{otherwise}. \end{cases}$$

THEOREM 2
Suppose X follows the discrete uniform distribution with $R_X = \{x_1, x_2, \ldots, x_k\}$.

The expectation of X is given by

$$\mu_X = E(X) = \sum_{i=1}^{k} x_i f_X(x_i) = \frac{1}{k}\sum_{i=1}^{k} x_i.$$

The variance of X is given by

$$\sigma_X^2 = V(X) = E(X^2) - (E(X))^2 = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - \mu_X^2.$$

EXAMPLE 4.1
A bulb is selected at random from a box that contains a 40-watt bulb, a 60-watt bulb, a 80-watt bulb, and a 100-watt bulb.

Each bulb has 1/4 probability of being selected.

Let X be the wattage of the bulb being selected. Identify the distribution of X, and compute its mean and variance.

Solution:
X follows a uniform distribution and

$$R_X = \{40, 60, 80, 100\}.$$

Further, the probability mass function for X is given as

$$f_X(x) = \begin{cases} \frac{1}{4}, & x = 40, 60, 80, 100; \\ 0, & \text{otherwise.} \end{cases}$$

We can compute the expectation as

$$E(X) = \sum_i x_i f_X(x_i) = 40 \cdot (1/4) + 60 \cdot (1/4) + 80 \cdot (1/4) + 100 \cdot (1/4) = 70.$$

The variance is also found to be

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= 40^2 \cdot (1/4) + 60^2 \cdot (1/4) + 80^2 \cdot (1/4) + 100^2(1/4) - 70^2 \\ &= 500. \end{aligned}$$

Bernoulli Trial, Bernoulli Random Variable and Bernoulli Process

Numerous experiments have two possible outcomes.

If an item is selected from the assembly line and inspected, it is either defective or not defective. A piece of fruit is either damaged or not damaged.

Such experiments are called Bernoulli trials after the Swiss mathematician Jacob Bernoulli.

DEFINITION 3 (BERNOULLI TRIAL)
*A **Bernoulli trial** is a random experiment with only two possible outcomes.*

One is called a "success", and the other a "failure". We often code the two outcomes as "1" (success) and "0" (failure).

DEFINITION 4 (BERNOULLI RANDOM VARIABLE)
*Let X be the number of success in a Bernoulli trial. Then X has only two possible values: 1 or 0, and is called a **Bernoulli random variable**.*

Denote by p, where $0 \leq p \leq 1$, the probability of success for a Bernoulli trial. Then X has the probability mass function

$$f_X(x) = P(X = x) = \begin{cases} p, & x = 1; \\ 1-p, & x = 0. \end{cases}$$

This probability mass function can also be written as

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

REMARK
We denote a Bernoulli random variable by $X \sim \text{Bernoulli}(p)$, and write $q = 1 - p$.

Then the probability mass function becomes

$$f_X(1) = p, \quad f_X(0) = q.$$

THEOREM 5
For a Bernoulli random variable defined as above, we have

$$\begin{aligned} \mu_X &= E(X) = p, \\ \sigma_X^2 &= V(X) = p(1-p) = pq. \end{aligned}$$

PARAMETERS
In certain instances, $f_X(x)$ may rely on one or more unknown quantities: different values of the quantities lead to different probability distributions.

Such a quantity is called the **parameter** of the distribution.

For example, p is the parameter for the Bernoulli distribution.

The collection of distributions that are determined by one or more unknown parameters is called a **family of probability distributions**.

Thus the aforementioned Bernoulli distributions determined by the parameter p is a family of probability distributions.

EXAMPLE 4.2
The following are all examples of Bernoulli trials:

A coin toss
: Say we want heads. Then "heads" is a success, and "tails" is a failure.

Rolling a die
: Say we only care about rolling a 6. Then the outcome space is binarized to "success"$= \{6\}$ and "failure" $= \{1,2,3,4,5\}$.

Polls
: Choosing a voter at random to ascertain if he will vote "yes" in an upcoming referendum.

EXAMPLE 4.3
A box contains 4 blue and 6 red balls. Draw a ball from the box at random.

What is the probability that a blue ball is chosen?

Solution:
Let $X = 1$ if a blue ball is drawn; and $X = 0$ otherwise.

Then X is a Bernoulli random variable and

$$P(X = 1) = 4/10 = 0.4.$$

Furthermore, the probability mass function for X is given by

$$f_X(x) = \begin{cases} 0.4, & x = 1; \\ 0.6, & x = 0. \end{cases}$$

DEFINITION 6 (BERNOULLI PROCESS)
*A **Bernoulli process** consists of a sequence of repeatedly performed **independent and identical** Bernoulli trials.*

*Consequently, a Bernoulli process generates a sequence of **independent and identically distributed** Bernoulli random variables: $X_1, X_2, X_3, \ldots$.*

Several distributions useful in applications are based on the Bernoulli trial and Bernoulli process. We will look at them in the next subsections:

- **Binomial distribution**;
- **Negative Binomial distribution**; **Geometric distribution**;
- **Poisson distribution**.

Binomial Distribution

Suppose we have n independent and identically distributed Bernoulli trials. We can use the binomial distribution to address some interesting questions. For example,

- A student randomly guesses at 5 multiple-choice questions. What is the number of questions the student guessed correctly?
- Randomly pick a family with 4 kids. What is the number of girls amongst the kids?
- A urn has 4 black balls and 3 white balls. Draw 5 balls with replacement. How many black balls will there be?

DEFINITION 7 (BINOMIAL RANDOM VARIABLE)
*A **Binomial random variable** counts the number of successes in n trials of a Bernoulli Process. That is, suppose we have n trials where*

- *the probability of success for each trial is the same p,*
- *the trials are independent.*

Then the number of successes, denoted by X, in the n trials is a binomial random variable.

We say X has a binomial distribution and write it as $X \sim \text{Bin}(n, p)$.

The probability of getting exactly x successes is given as

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, 2, \ldots, n.$$

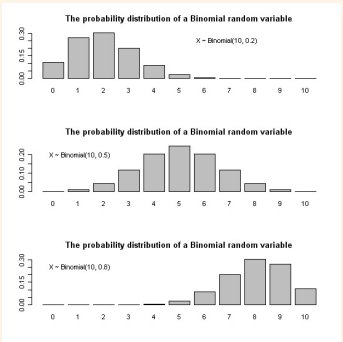
It can be shown that $E(X) = np$, *and* $V(X) = np(1-p)$.

REMARK
When $n = 1$, the probability mass function for the binomial random variable X is reduced to

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

This is the probability mass function for the Bernoulli distribution. Therefore the Bernoulli distribution is a special case of the binomial distribution.

BINOMIAL PROBABILITY MASS FUNCTION
The probability mass function for $\text{Bin}(10, 0.2)$, $\text{Bin}(10, 0.5)$, and $\text{Bin}(10, 0.8)$.



EXAMPLE 4.4
Flip a fair coin independently 10 times. What is the probability of observing exactly 6 heads?

Solution:
Let X be the number of heads in 10 flips of the coin.

Each flip of the coin can be observed as a Bernoulli trial, with the probability of getting head (success) $p = 0.5$. Then X is the number success out of 10 Bernoulli trials; so $X \sim \text{Bin}(10, 0.5)$.

We can compute

$$P(X = 6) = \binom{10}{6} (0.5)^6 (1 - 0.5)^{10-6} = 0.205.$$

Negative Binomial Distribution

Consider a Bernoulli process, where the Bernoulli trials can be repeated as many times as desired or necessary.

Suppose we are interested in the number of trials needed so that k number of successes occur.

DEFINITION 8 (NEGATIVE BINOMIAL DISTRIBUTION)
Let X be the number of independent and identically distributed Bernoulli(p) *trials needed until the kth success occurs. Then X follows a* ***Negative Binomial distribution****, denoted by* $X \sim \text{NB}(k, p)$.

The probability mass function of X is given by

$$f_X(x) = P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad \text{for } x = k, k+1, k+2, \ldots$$

It can be shown that $E(X) = \dfrac{k}{p}$ *and* $V(X) = \dfrac{(1-p)k}{p^2}$.

EXAMPLE 4.5
Keep rolling a fair die, until the 6th time we get the number 6. What is the probability that we need to roll the die 10 times?

Solution:
Let X be the number of rolls needed to get the 6th number 6. Then $X \sim \text{NB}(6, 1/6)$.

Using the probability mass function of the negative binomial distribution:

$$P(X=10) = \binom{10-1}{6-1}(1/6)^6(1-1/6)^4 = 0.001302.$$

Geometric Distribution

The **Geometric distribution** is a special case of the negative binomial distribution.

DEFINITION 9 (GEOMETRIC DISTRIBUTION)
Let X be the number of independent and identically distributed Bernoulli(p) *trials needed until the first success occurs. Then X follows a* ***Geometric distribution****, denote by* $X \sim \text{Geom}(p)$.

The probability mass function of X is given by

$$f_X(x) = P(X=x) = (1-p)^{x-1}p.$$

It can be shown that $E(X) = \dfrac{1}{p}$ *and* $V(X) = \dfrac{1-p}{p^2}$.

Poisson Distribution

A number of probability distributions come about through limiting arguments applied to other distributions. One useful distribution of this type is called the Poisson distribution.

DEFINITION 10 (POISSON RANDOM VARIABLE)
The ***Poisson random variable*** *X denotes the number of events occurring in a* ***fixed period of time or fixed region****.*

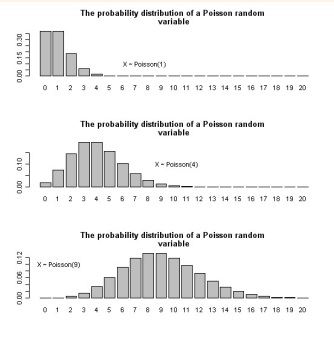
We denote $X \sim \text{Poisson}(\lambda)$*, where the parameter* $\lambda > 0$ *is the expected number of occurrences during the given period/region. Its probability mass function is given by*

$$f_X(k) = P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!},$$

where $k = 0, 1, \ldots$ *is the number of occurrences of such events.*

It can be shown that $E(X) = \lambda$ *and* $V(X) = \lambda$.

POISSON PROBABILITY MASS FUNCTION
The probability mass function for Poisson(1), Poisson(4), and Poisson(9).



EXAMPLE 4.6
The "fixed period of time" given in the definition can be a time period of any length: a minute, a day, a week, a month, etc. The "fixed region" can be of any size.

Here are some examples of events that may be modeled by the Poisson distribution:

(a) The number of spelling mistakes one makes while typing a single page.

(b) The number of times a web server is accessed per minute.

(c) The number of road kill (animals killed) found per unit length of road.

(d) The number of mutations in a given stretch of DNA after a certain amount of radiation exposure.

(e) The number of unstable atomic nuclei that decayed within a given period of time in a piece of radioactive substance.

(f) The distribution of visual receptor cells in the retina of the human eye.

(g) The number of light bulbs that burn out in a certain amount of time.

DEFINITION 11 (POISSON PROCESS)
*The **Poisson process** is a continuous time process. We count the number of occurrences within some interval of time. The defining properties of a Poisson process with rate parameter α are*

- *the expected number of occurrences in an interval of length T is αT;*
- *there are no simultaneous occurrences;*
- *the number of occurrences in disjoint time intervals are independent.*

The number of occurrences in any interval T of a Poisson process follows a Poisson(αT) *distribution.*

EXAMPLE 4.7
The average number of robberies in a day is four in a certain big city. What is the probability that six robberies occurring in two days?

Solution:
Let X_1 be the number of robberies in one day. Then $X_1 \sim \text{Poisson}(4)$ from the given conditions.

Let X be the number of robberies in two days. Then

$$X \sim \text{Poisson}(2 \times 4) = \text{Poisson}(8).$$

We then have

$$P(X = 6) = \frac{e^{-8}8^6}{6!} = 0.1222.$$

Poisson Approximation to Binomial

The Poisson random variable has a tremendous range of applications in diverse areas because it may be used as an approximation for a binomial random variable under certain conditions.

The following result shows us how.

PROPOSITION 12 (POISSON APPROXIMATION TO BINOMIAL)
Let $X \sim \text{Bin}(n, p)$. Suppose that $n \to \infty$ and $p \to 0$ in such a way that $\lambda = np$ remains a constant. Then approximately, $X \sim \text{Poisson}(np)$.
That is

$$\lim_{p \to 0; n \to \infty} P(X = x) = \frac{e^{-np}(np)^x}{x!}.$$

REMARK
The approximation is good when

- $n \geq 20$ and $p \leq 0.05$, or if
- $n \geq 100$ and $np \leq 10$.

EXAMPLE 4.8
The probability, p, of an individual car having an accident at a junction is 0.0001.

If there are 1000 cars passing through the junction during certain period of a day, what is the probability of two or more accidents occurring during that period?

Solution:
Let X be the number of accidents among the 1000 cars.

Then $X \sim \text{Bin}(1000, 0.0001)$. If we compute using the binomial distribution,

$$P(X \geq 2) = \sum_{x=2}^{1000} \binom{1000}{x} 0.0001^x 0.9999^{1000-x}.$$

We can also use the Poisson approximation.

We have $n = 1000$ and $p = 0.0001$. Hence $np = \lambda = 0.1$.

Thus

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - e^{-0.1} - e^{-0.1}(0.1)^1/1! \\ &= 0.0047. \end{aligned}$$

2 CONTINUOUS DISTRIBUTION

There are many "natural" random variables whose set of possible values is uncountable. For example, consider

- the lifetime of an electrical appliance; or
- the amount of rainfall we get in a month.

How then, can we model such variables?

To achieve this aim, we shall now study some classes of continuous random variables.

Continuous Uniform Distribution

Intuitively, a uniform random variable on the interval (a,b) is a completely random number between a and b. We formalize the notion of "completely random" on an interval by specifying that the probability density function should be constant over the interval.

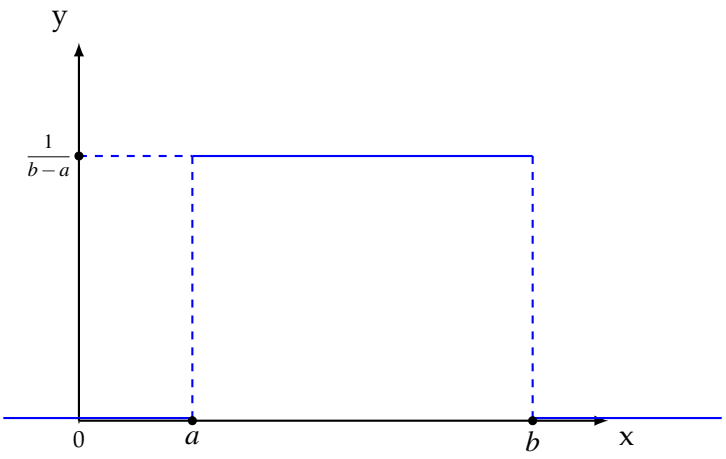
DEFINITION 13 (CONTINUOUS UNIFORM DISTRIBUTION)
A random variable X is said to follow a ***uniform distribution*** *over the interval (a,b) if its probability density function is given by*

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

We denote this by $X \sim \mathrm{U}(a,b)$.

It can be shown that $E(X) = \dfrac{a+b}{2}$ and $V(X) = \dfrac{(b-a)^2}{12}$.

The probability density function for the continuous uniform distribution can be drawn as below.



The cumulative distribution function for the continuous uniform distribution is given by

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

EXAMPLE 4.9
A point is chosen at random on the line segment $[0,2]$.

What is the probability that the chosen point lies between 1 and $\frac{3}{2}$?

Solution:
Let X be the position of the point. Then $X \sim \mathrm{U}(0,2)$, and we have

$$f_X(x) = \begin{cases} \frac{1}{2}, & 0 \leq x \leq 2; \\ 0, & \text{otherwise.} \end{cases}$$

Then the required probability is

$$P\left(1 \leq X \leq \tfrac{3}{2}\right) = \int_1^{3/2} \tfrac{1}{2}\,\mathrm{dx} = \tfrac{1}{2}\,[x]_1^{3/2} = \tfrac{1}{4}.$$

Exponential Distribution

The exponential distribution is the continuous counterpart to the geometric distribution. It is often used to model the waiting time to the first success in *continuous time*.

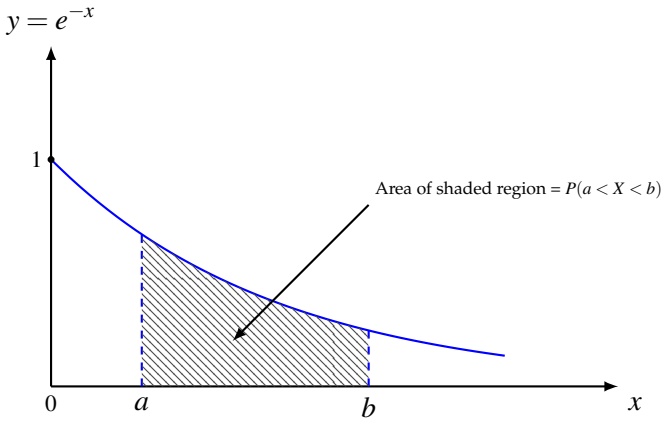
DEFINITION 14 (EXPONENTIAL DISTRIBUTION)
A continuous random variable X is said to follow an ***exponential distribution*** *with parameter $\lambda > 0$ if its probability density function is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0. \end{cases}$$

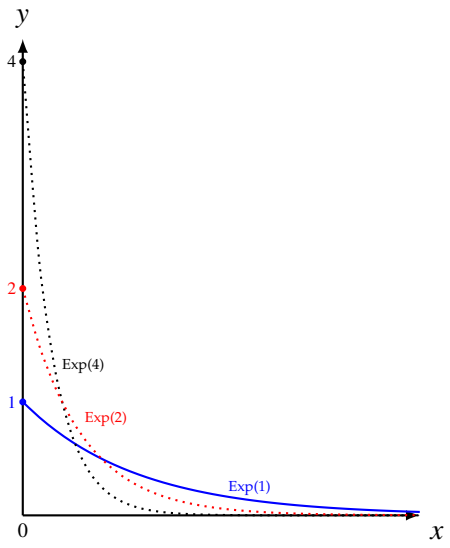
We denote $X \sim \mathrm{Exp}(\lambda)$.

It can be shown that $E(X) = \dfrac{1}{\lambda}$ and $V(X) = \dfrac{1}{\lambda^2}$.

The probability density function for $\text{Exp}(1)$.



The probability density function for $\text{Exp}(\lambda)$, where $\lambda = 1, 2, 4$.



The cumulative distribution function of $X \sim \text{Exp}(\lambda)$ is given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

ALTERNATIVE FORM OF THE EXPONENTIAL
The probability density function of the exponential distribution can be written in the following alternative form

$$f_X(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu}, & x \geq 0; \\ 0, & x < 0. \end{cases}$$

The parameters μ and λ have the relationship $\mu = 1/\lambda$.

We will then have

$$E(X) = \mu, \quad V(X) = \mu^2, \quad \text{and} \quad F_X(x) = 1 - e^{-x/\mu}, \quad \text{for } x \geq 0.$$

EXAMPLE 4.10
Suppose that the failure time, T, of a system is exponentially distributed, with a mean of 5 years.

What is the probability that at least two out of five of these systems are still functioning at the end of 8 years?

Solution:
Since $E(T) = 5$, therefore $\lambda = 1/5$.

We then have $T \sim \text{Exp}(1/5)$, and so

$$P(T > 8) = 1 - P(T \leq 8) = 1 - F_X(8) = e^{-(1/5)\times 8} = e^{-1.6} \approx 0.2.$$

Now let X be the number of systems out of 5 that are still functioning after 8 years. We see tha $X \sim \text{Bin}(5, 0.2)$. Hence,

$$P(X \geq 2) = 0.2627.$$

THEOREM 15
Suppose that X has an exponential distribution with parameter $\lambda > 0$. Then for any two positive numbers s and t, we have

$$P(X > s+t | X > s) = P(X > t).$$

REMARK

The above theorem states that the exponential distribution has "**no memory**" or is "**memoryless**".

To illustrate, suppose $X \sim \text{Exp}(\lambda)$ is the life length of a bulb. Then

$$P(X > s+t|X > s) = P(X > t).$$

This means that, if the bulb has lasted s time units (that is, $X > s$), the probability that it will last for another t units (that is, $X > s+t$), is the same as the probability that it will last for the first t units as a brand new bulb.

Normal Distribution

We next look at one of the most important class of continuous random variables.

DEFINITION 16 (NORMAL DISTRIBUTION)

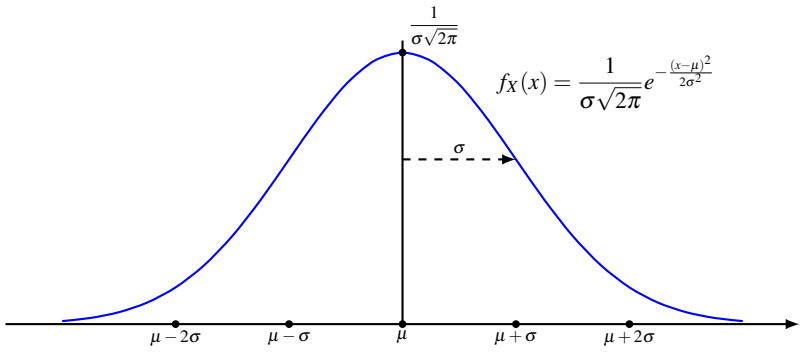
A random variable X is said to follow a ***normal distribution*** *with parameters μ and σ^2 if its probability density function is given by*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty.$$

We denote $X \sim N(\mu, \sigma^2)$.

It can be shown that $E(X) = \mu$ and $V(X) = \sigma^2$.

The probability density function of the normal distribution is positive over the whole real line, symmetrical about $x = \mu$, and bell-shaped.



PROPERTIES OF THE NORMAL DISTRIBUTION

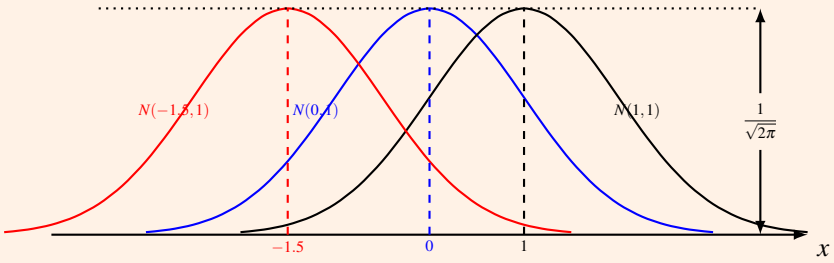
We go through some important properties of the normal distribution.

(1) The total area under the curve and above the horizontal axis is equal to 1.

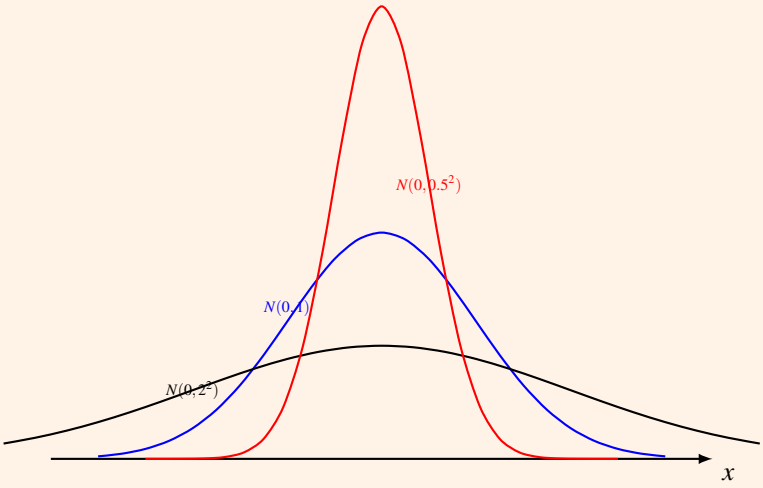
$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{dx} = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \mathrm{dx} = 1.$$

This validates that $f_X(\cdot)$ is a probability density function.

(2) Two normal curves are identical in shape if they have the same σ^2. But they are centered at different positions when their means are different.



(3) As σ increases, the curve flattens; and vice versa.



(4) Given $X \sim N(\mu, \sigma^2)$, let

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

Then Z follows the $N(0,1)$ distribution, with $E(Z) = 0$ and $V(Z) = 1$.

We say that Z has a **standardized normal** or **standard normal** distribution, and the probability density function of Z is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

REMARK

- Calculating normal probabilities is a challenge because
  - there is no close formula for the integration,
  - and the computation relies on numerical integration.
- Fortunately we can use Property 4 from above.

  Suppose $X \sim N(\mu, \sigma^2)$ and we seek $P(x_1 < X < x_2)$. Consider

  $$x_1 < X < x_2 \iff \tfrac{x_1-\mu}{\sigma} < \tfrac{X-\mu}{\sigma} < \tfrac{x_2-\mu}{\sigma}.$$

  Consider the transformation $Z = \frac{X-\mu}{\sigma}$, and let $z_1 = \frac{x_1-\mu}{\sigma}$ and $z_2 = \frac{x_2-\mu}{\sigma}$. Then

  $$P(x_1 < X < x_2) = P(z_1 < Z < z_2).$$

- By convention, we use $\phi(\cdot)$ and $\Phi(\cdot)$ to denote the probability density function and cumulative distribution function of the standard normal. That is,

  $$\begin{aligned} \phi(z) &= f_Z(z) = \tfrac{1}{\sqrt{2\pi}} e^{-z^2/2} \\ \Phi(z) &= \int_{-\infty}^{z} \phi(t)\,\mathrm{dt} = \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\,\mathrm{dt}. \end{aligned}$$

  Then for $X \sim N(\mu, \sigma^2)$ and any real numbers x_1, x_2,

  $$P(x_1 < X < x_2) = \Phi\left(\tfrac{x_2-\mu}{\sigma}\right) - \Phi\left(\tfrac{x_1-\mu}{\sigma}\right).$$

- Thus we can use the standard normal to calculate any normal probability.

  To do so, $\Phi(z)$ can be tabulated, or computed using statistical software.
- The standard normal distribution has the following properties:
  - $P(Z \geq 0) = P(Z \leq 0) = \Phi(0) = 0.5$;
  - For any z, $\Phi(z) = P(Z \leq z) = P(Z \geq -z) = 1 - \Phi(-z)$;
  - If $Z \sim N(0,1)$, then $-Z \sim N(0,1)$;
  - If $Z \sim N(0,1)$, then $\sigma Z + \mu \sim N(\mu, \sigma^2)$.

EXAMPLE 4.11
Given $X \sim N(50, 100)$, compute $P(45 < X < 62)$.

Solution:
We have $\mu = 50$, $\sigma = 10$. Then

$$\begin{aligned} P(45 < X < 62) &= P\left(\tfrac{45-50}{10} < \tfrac{X-50}{10} < \tfrac{62-50}{10}\right) \\ &= P(-0.5 < Z < 1.2) \\ &= P(Z < 1.2) - P(Z \leq -0.5) \\ &= \Phi(1.2) - \Phi(-0.5), \end{aligned}$$

where $\Phi(1.2)$ and $\Phi(-0.5)$ can either be computed using software or obtained from a statistical table.

DEFINITION 17 (QUANTILE)
The ***αth (upper) quantile****, where $0 \leq \alpha \leq 1$, of the random variable X is the number x_α that satisfies*

$$P(X \geq x_\alpha) = \alpha.$$

THE Z UPPER QUANTILE
Specifically, we denote by z_α the αth (upper) quantile, or the 100α percentage point, of $Z \sim N(0,1)$. That is

$$P(Z \geq z_\alpha) = \alpha.$$

Here are some common values of z_α:

$$z_{0.05} = 1.645, \quad z_{0.01} = 2.326.$$

Since $\phi(z)$, the probability density function of Z, is symmetrical about 0, then

$$P(Z \geq z_\alpha) = P(Z \leq -z_\alpha) = \alpha.$$

EXAMPLE 4.12
Find z such that

(a) $P(Z < z) = 0.95$;

(b) $P(|Z| \leq z) = 0.98$.

Solution:

(a) We need z such that

$$P(Z > z) = 1 - P(Z < z) = 0.05.$$

Therefore $z = z_{0.05} = 1.645$.

(b) We have

$$\begin{aligned} 0.98 &= P(|Z| \leq z) = 1 - P(|Z| > z) \\ &= 1 - P(Z > z) - P(Z < -z) = 1 - 2P(Z > z). \end{aligned}$$

This means that $P(Z > z) = 0.01$. Therefore $z = z_{0.01} = 2.326$.

Normal Approximation to Binomial

Recall that when $n \to \infty$, $p \to 0$, and np remains a constant, we can use the Poisson distribution to approximate the binomial distribution.

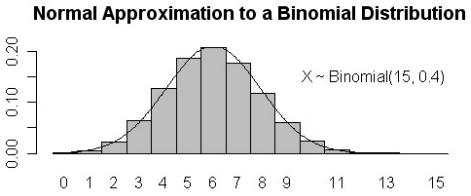
When $n \to \infty$, but p remains a constant (practically, p is not very close to 0 or 1), we can use the normal distribution to approximate the binomial distribution.

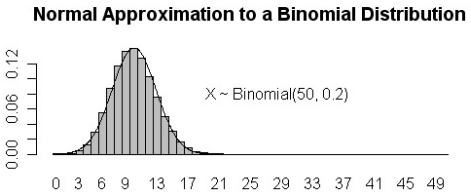
A good rule of thumb is to use the normal approximation when

$$np > 5 \quad \text{and} \quad n(1-p) > 5.$$

PROPOSITION 18 (NORMAL APPROXIMATION TO BINOMIAL)
Let $X \sim \text{Bin}(n,p)$, so that $E(X) = np$ and $V(X) = np(1-p)$. Then as $n \to \infty$,

$$Z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - np}{\sqrt{np(1-p)}} \text{ is approximately } \sim N(0,1).$$





Five

Sampling and Sampling Distributions

1 Population and Sample

The aim of *Statistical Inference* is to say something about the population based on a sample.

> **Definition 1 (Population & Sample)**
> *The totality of all possible outcomes or observations of a survey or experiment is called a* ***population****.*
>
> *A* ***sample*** *is any subset of a population.*

Every outcome or observation can be recorded as a numerical or categorical value.

So each member of a population can be regarded as a value of a random variable.

Note that a population can be finite or infinite.

Finite Population
A **finite population** consists of a finite number of elements.

For example, it can be

- the monthly income of Singaporeans;
- all the books in the Central Library; or
- the CAP scores of students in NUS.

INFINITE POPULATION
An **infinite population** is one that consists of an infinitely (countable and uncountable) large number of elements.

For example, it can be

- the results of all possible rolls of a pair of dice;
- the depths at all conceivable positions of a lake; or
- the PSI level in the air at various parts of Singapore.

REMARK
Some finite populations are so large that in theory we assume them to be infinite, since it may be impractical/uneconomical to observe all its values.

2 RANDOM SAMPLING

We often know that the population belongs to (or can be modeled using) a known (family of) distribution(s).

However, the values of parameters (for example, p, μ or σ) that specify the distribution(s) are unknown.

For example:

- A pollster is sure that the responses to his "agree/disagree" question will follow a binomial distribution, but p, the proportion of those who "agree" in the population, is unknown.
- An agronomist believes that the yield per acre of a variety of wheat is approximately normally distributed, but the mean μ and the standard deviation σ of the yields are unknown.

Thus we rely on a sample to learn about these parameters and study the properties of the population.

- The sample should be representative of the population. We have different types of sampling schemes attempting to do that.
- For the probability methods, it is possible to fully describe the quantitative properties of the sample.
- We will focus on the simple random sample. It is often known simply as a random sample.

DEFINITION 2 (SIMPLE RANDOM SAMPLE)
A set of n members taken from a given population is called a ***sample*** *of size n.*

A ***simple random sample (SRS)*** *of n members is a sample that is chosen such that every subset of n observations of the population has the same probability of being selected.*

REMARK
With simple random sampling, everyone has the same chance of inclusion in the sample, so it is fair.

It tends to yield a sample that resembles the population. This reduces the chance that the sample is seriously biased in some way, leading to inaccurate inferences about the population.

EXAMPLE 5.1 (DRUG EXPERIMENT)
Suppose that a researcher in a medical center plans to compare two drugs for some adverse condition. She has four patients with this condition, and she wants to randomly select two to use each drug. Denote the four patients by P_1, P_2, P_3, and P_4.

In selecting $n = 2$ subjects to use the first drug, the six possible samples are

$$(P_1, P_2), (P_1, P_3), (P_1, P_4), (P_2, P_3), (P_2, P_4), (P_3, P_4).$$

REMARK
More generally, let N denote the population size. The population has $\binom{N}{n}$ possible samples of size n.

For large values of N and n, one can use software easily to select the sample from a list of the population members using a random number generator.

Sampling from an Infinite Population

When lists are available and items are readily numbered, it is easy to draw random samples from finite populations.
Unfortunately, it is often impossible to proceed in the way we have just described for an infinite population.

DEFINITION 3 (SIMPLE RANDOM SAMPLE: INFINITE POPULATION)
Let X be a random variable with certain probability distribution $f_X(x)$.

Let $X_1, X_2, \ldots, X_n$ be n independent random variables each having the same distribution as X. Then $(X_1, X_2, \ldots, X_n)$ is called a ***random sample of size*** *n from a population with distribution $f_X(x)$.*

The joint probability function of $(X_1, X_2, \ldots X_n)$ is given by

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n),$$

where $f_X(x)$ is the probability function of the population.

3 SAMPLING DISTRIBUTION OF SAMPLE MEAN

Our main purpose in selecting random samples is to elicit information about the unknown population parameters.

For instance, we wish to know the proportion of people in Singapore who prefer a certain brand of coffee.

A large random sample is then selected from the population and the proportion of this sample favouring the brand of coffee in question is calculated.

This value is now used to make some inference concerning the true proportion in the population.

DEFINITION 4 (STATISTIC)
Suppose a random sample of n observations $(X_1, \ldots, X_n)$ has been taken. A function of $(X_1, \ldots, X_n)$ is called a ***statistic****.*

EXAMPLE 5.2 (SAMPLE MEAN)
The **sample mean**, defined as

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

is a statistic.

If the values in a random sample are observed and they are $(x_1, \ldots, x_n)$, then the realization of the statistic $\overline{X}$ is given by

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

EXAMPLE 5.3 (SAMPLE VARIANCE)
The **sample variance**, defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

is a statistic.

Similarly, if the values in a random sample are observed and they are $(x_1, \ldots, x_n)$, then the realization of the statistic S^2 is given by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

STATISTICS ARE RANDOM VARIABLES

- Note that X_1 is a random variable and so are $X_2, \ldots, X_n$.
- Thus $\overline{X}$ and S^2 are random variables as well.
- As many random samples are possible from the same population, we expect the statistic to vary somewhat from sample to sample.
- Hence a statistic is a random variable. It is meaningful to consider the probability distribution of a statistic.

DEFINITION 5 (SAMPLING DISTRIBUTION)
*The probability distribution of a statistic is called a **sampling distribution**.*

Two Results

We next present two key results about the sampling distribution of the sample mean.

- Theorem 6 provides formulas for the center and the spread of the sampling distribution.
- Theorem 9 describes the shape of the sampling distribution, showing that it is often approximately normal.

THEOREM 6 (MEAN AND VARIANCE OF $\overline{X}$)
For random samples of size n taken from an infinite population with mean μ_X and variance σ_X^2, the sampling distribution of the sample mean $\overline{X}$ has mean μ_X and variance $\frac{\sigma_X^2}{n}$. That is,

$$\mu_{\overline{X}} = E\left(\overline{X}\right) = \mu_X \quad \text{and} \quad \sigma_{\overline{X}}^2 = V\left(\overline{X}\right) = \frac{\sigma_X^2}{n}.$$

VALIDITY OF $\overline{X}$ AS AN ESTIMATOR FOR μ_X

- The expectation of $\overline{X}$ is equal to the population mean μ_X.
- In "the long run", $\overline{X}$ does not introduce any systematic bias as an estimator of μ_X. So $\overline{X}$ can serve as a valid estimator of μ_X.
- For an infinite population, when n gets larger and larger, σ_X^2/n, the variance of $\overline{X}$, becomes smaller and smaller, that is, the accuracy of $\overline{X}$ as an estimator of μ_X keeps improving.

DEFINITION 7 (STANDARD ERROR)
The spread of a sampling distribution is described by its standard deviation, which is called the ***standard error****.*

The standard deviation of the sampling distribution of $\overline{X}$ is called the standard error of $\overline{X}$. We denote it by $\sigma_{\overline{X}}$.

REMARK
The standard error of $\overline{X}$ describes how much $\bar{x}$ tends to vary from sample to sample of size n.

The symbol $\sigma_{\overline{X}}$ (instead of σ) and the terminology standard error (instead of standard deviation) distinguishes this measure from the standard deviation σ of the population.

Because σ_X^2/n decreases as n increases, $\overline{X}$ tends to be closer to μ_X as n increases. The result that $\overline{X}$ converges to μ_X as n grows indefinitely is called the **Law of Large Numbers**.

THEOREM 8 (LAW OF LARGE NUMBERS (LLN))
If $X_1, \ldots, X_n$ are independent random variables with the same mean μ and variance σ^2, then for any $\varepsilon \in \mathbb{R}$,

$$P(|\overline{X} - \mu| > \varepsilon) \to 0 \text{ as } n \to \infty.$$

REMARK
This says that as the sample size increases, the probability that the sample mean differs from the population mean goes to zero.

Another way of looking at this is that it is increasingly likely that $\overline{X}$ is close to μ_X, as n gets larger.

4 CENTRAL LIMIT THEOREM

The result that the sampling distribution of $\overline{X}$ is approximately normal is called the **Central Limit Theorem**.

THEOREM 9 (CENTRAL LIMIT THEOREM (CLT))
If $\overline{X}$ is the mean of a random sample of size n taken from a population having mean μ and finite variance σ^2, then, as $n \to \infty$,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to Z \sim N(0,1).$$

Equivalently, this means

$$\overline{X} \to N\left(\mu, \frac{\sigma^2}{n}\right).$$

WHAT IS THE BIG DEAL?
The Central Limit Theorem states that, under rather general conditions, for large n, sums and means of random samples drawn from a population follows the normal distribution closely.

Note that if the random sample comes from a normal population, $\overline{X}$ is normally distributed regardless of the value of n.

RULE OF THUMB
The Central Limit Theorem says that, if you take the mean of a large number of independent samples, then the distribution of that mean will be approximately normal.

- If the population you are sampling from is symmetric with no outliers, a good approximation to normality appears after as few as 15-20 samples.
- If the population is moderately skewed, such as exponential or χ^2, then it can take between 30-50 samples before getting a good approximation.
- Data with extreme skewness, such as some financial data where most entries are 0, a few are small, and even fewer are extremely large, may not be appropriate for the Central Limit Theorem even with 1000 samples.

EXAMPLE 5.4 (BOWLING LEAGUE)
In a bowling league season, bowlers bowl 50 games and the average score is ranked at the end of the season. Historically, John averages 175 a game with a standard deviation of 30. What is the probability that John will average more than 180 this season?

Solution:
We do not know the distribution of X, but we know that $\mu = 175$, $\sigma = 30$ and $n = 50$. Let $\overline{X}$ be the sample mean.

By CLT, we can approximate $\overline{X}$ by $N(\mu, \sigma^2/n)$. The question asks for the probability

$$\begin{aligned} P(\overline{X} > 180) &= P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} > \frac{180 - \mu}{\sigma/\sqrt{n}}\right) \\ &\approx P(Z > 1.18) = 0.119. \end{aligned}$$

5 OTHER SAMPLING DISTRIBUTIONS

We next describe the χ^2, t, and F distributions, which are examples of distributions that are derived from random samples from a normal distribution.

The emphasis is on understanding the relationships between the random variables and how they can be used to describe distributions related to the sample statistics $\overline{X}$ and S^2.

Your goal should be to get comfortable with the idea that sample statistics have known distributions.

DEFINITION 10 (THE χ^2 DISTRIBUTION)
Let Z be a standard normal random variable. A random variable with the same distribution as Z^2 is called a ***χ^2 random variable with one degree of freedom****.*

Let $Z_1, \ldots, Z_n$ be n independent and identically distributed standard normal random variables. A random variable with the same distribution as $Z_1^2 + \cdots + Z_n^2$ is called a ***χ^2 random variable with n degrees of freedom****.*

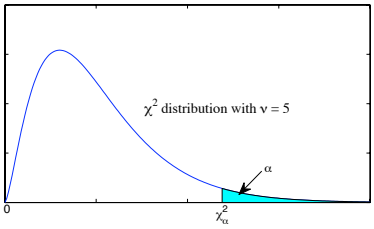
REMARK
We denote a χ^2 random variable with n degrees of freedom as $\chi^2(n)$.

PROPERTIES OF χ^2 DISTRIBUTIONS

1. If $Y \sim \chi^2(n)$, then $E(Y) = n$ and $V(Y) = 2n$.
2. For large n, $\chi^2(n)$ is approximately $N(n, 2n)$.
3. If Y_1 and Y_2 are independent χ^2 random variables with m and n degrees of freedom respectively, then $Y_1 + Y_2$ is a χ^2 random variable with $m+n$ degrees of freedom.
4. The χ^2 distribution is a family of curves, each determined by the degrees of freedom n. All the density functions have a long right tail.

DEFINITION 11
Define $\chi^2(n;\alpha)$ such that for $Y \sim \chi^2(n)$,

$$P(Y > \chi^2(n;\alpha)) = \alpha.$$



The sampling distribution of $(n-1)S^2/\sigma^2$

Recall that for $X_1, \ldots, X_n$ independent and identically distributed with $E(X) = \mu$ and $V(X) = \sigma^2$, the sample variance is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Though it can be shown that $E(S^2) = \sigma^2$, the sampling distribution of the random variable S^2 has little practical application in statistics.
We shall instead consider the sampling distribution of the random variable $\dfrac{(n-1)S^2}{\sigma^2}$ when $X_i \sim N(\mu, \sigma^2)$, for all i.

THEOREM 12
If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the random variable

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

has a χ^2 distribution with $n-1$ degrees of freedom.

DEFINITION 13 (THE t-DISTRIBUTION)
Suppose $Z \sim N(0,1)$ and $U \sim \chi^2(n)$. If Z and U are independent, then

$$T = \frac{Z}{\sqrt{U/n}}$$

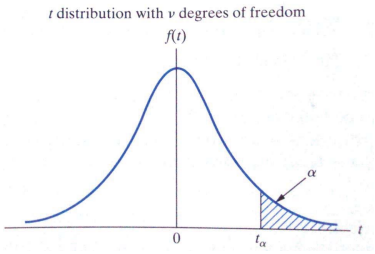
*follows the **t-distribution with n degrees of freedom**.*

PROPERTIES OF THE t-DISTRIBUTION

- The t-distribution with n degrees of freedom, also called the Student's t-distribution, is denoted by $t(n)$.
- The t-distribution approaches $N(0,1)$ as the parameter $n \to \infty$. When $n \geq 30$, we can replace it by $N(0,1)$.
- If $T \sim t(n)$, then $E(T) = 0$ and $V(T) = n/(n-2)$ for $n > 2$.
- The graph of the t-distribution is symmetric about the vertical axis and resembles the graph of the standard normal distribution.

DEFINITION 14
Define $t_{n;\alpha}$ such that for $T \sim t(n)$,

$$P(T > t_{n;\alpha}) = \alpha.$$



THE IMPORTANCE OF THE *t*-DISTRIBUTION
The t-distribution will play an important role in the later chapters, where it appears as the result of random sampling.

The following theorem establishes the connection between a random sample $X_1, \ldots, X_n$ and the t-distribution.

THEOREM 15
If $X_1, \ldots, X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2, then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}}$$

follows a t-distribution with $n-1$ degrees of freedom.

EXAMPLE 5.5 (MIDTERM SCORE)
The lecturer of a class announced that the mean score of the midterm is 16 out of 30. A student doubts it, so he randomly chose 5 classmates and asked them for their scores: 20, 19, 24, 22, 25.

Should the student believe that the mean score is 16? Assume the scores are approximately normally distributed.

Solution:
The student has $n = 5$ sampled data

$$x_1 = 20, x_2 = 19, x_3 = 24, x_4 = 22, x_5 = 25.$$

If $\mu = 16$,

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - 16}{S/\sqrt{5}}$$

should follow a t-distribution with $5 - 1 = 4$ degrees of freedom.

With the observed data $\bar{x} = 22$ and $s = 2.55$ so

$$t = \frac{22 - 16}{2.55/\sqrt{5}} = 5.26.$$

Using software, $P(t(4) > 5.26) = 0.003$. This says that there is only a 0.003 chance that T is 5.26 (or larger), provided the lecturer is telling the truth that $\mu = 16$.

So should the student believe him based on his findings?

DRINK BEER AND DO STATISTICS!
The t-distributions were discovered by William S. Gosset in 1908. Gosset was a statistician employed by the Guinness brewing company which had stipulated that he not publish under his own name. He therefore wrote under the pen name "Student".

For a biography of Gosset, browse to
http://www-history.mcs.st-andrews.ac.uk/Biographies/Gosset.html

DEFINITION 16 (THE F-DISTRIBUTION)
Suppose $U \sim \chi^2(m)$ and $V \sim \chi^2(n)$ are independent. Then the distribution of the random variable

$$F = \frac{U/m}{V/n}$$

*is called a **F-distribution with (m,n) degrees of freedom**.*

PROPERTIES OF THE F-DISTRIBUTION

- The F-distribution with (m,n) degrees of freedom is denoted by $F(m,n)$.
- If $X \sim F(m,n)$, then
$$E(X) = \frac{n}{n-2}, \quad \text{for } n > 2$$
and
$$V(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad \text{for } n > 4.$$
- If $F \sim F(n,m)$, then $1/F \sim F(m,n)$. This follows immediately from the definition of the F-distribution.
- Values of the F-distribution can be found in the statistical tables or software. The values of interests are $F(m,n;\alpha)$ such that
$$P(F > F(m,n;\alpha)) = \alpha,$$
where $F \sim F(m,n)$.
- It can be shown that
$$F(m,n;1-\alpha) = 1/F(n,m;\alpha).$$

EXAMPLE 5.6

For example,
$$F(4,5;0.05) = 5.19$$
means that $P(F > 5.19) = 0.05$, where $F \sim F(4,5)$.

Six

Estimation

We now learn about a powerful use of statistics:

STATISTICAL INFERENCE
about POPULATION PARAMETERS
using SAMPLE DATA.

In case you wonder about the relevance of learning about probability and sampling distribution, this is why:

- Statistical inference methods use probability calculations that assume that the data were gathered with a random sample or a randomized experiment.
- The probability calculations refer to a sampling distribution of a statistic, which is often approximately a normal distribution.

There are two types of statistical inference methods

- estimation of population parameters; and
- testing hypotheses about the parameter values.

This chapter discusses the first — estimating population parameters.

TWO TYPES OF ESTIMATIONS

Point estimation
Based on sample data, a single number is calculated to estimate the population parameter. The rule or formula that describes this calculation is called the point estimator. The resulting number is called a point estimate.

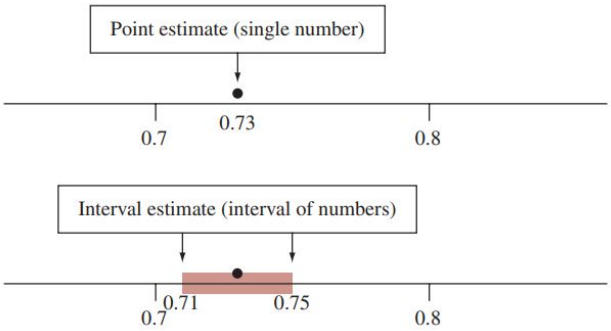
Interval estimation
Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie.

EXAMPLE 6.1
One survey asked, "Do you believe in hell?"

From sample data, the point estimate for the proportion of adult (in the population) who would respond "yes" is 0.73. The adjective "point" refers to using a single number as the parameter estimate.

An interval estimate predicts that the proportion of adult (in the population) who believe in hell falls between 0.71 and 0.75.

The next figure illustrates the difference between point estimate and interval estimate for the previous example.



1 POINT ESTIMATION

Suppose we are interested to estimate the parameter μ, the population mean. Assume that we have the following data, a random sample consisting

$$X_1, X_2, \ldots, X_n.$$

DEFINITION 1 (ESTIMATOR)
*An **estimator** is a rule, usually expressed as a formula, that tells us how to calculate an **estimate** based on information in the sample.*

EXAMPLE 6.2 (POINT ESTIMATOR)
We want to estimate the average waiting time for a bus (μ) for students attending ST2334. The lecturer asked 4 students their waiting times $X_1, \ldots, X_4$ for a bus. The (observed) results are

$$x_1 = 6, x_2 = 1, x_3 = 4, x_4 = 9.$$

We can use $\overline{X} = \frac{1}{4}(X_1 + \cdots + X_4)$ to estimate μ. In this case, $\overline{X}$ is the estimator (for μ), and the computed value $\bar{x} = 5$ is the estimate.

QUESTIONS

- *How good is the estimator?*
- *What would be a criteria for a "good" estimator?*

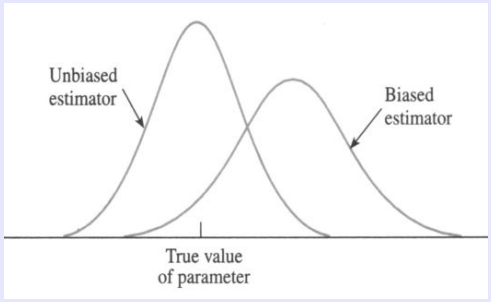
Unbiased Estimator

One of the reasons we think $\overline{X}$ is a good estimator of μ is because $E(\overline{X}) = \mu$. That is, "on average", the estimator is right.

In general, we represent the parameter of interest by θ. For example, θ can be p, μ, or σ.

DEFINITION 2 (UNBIASED ESTIMATOR)
*Let $\hat{\Theta}$ be an estimator of θ. Then $\hat{\Theta}$ is a random variable based on the sample. If $E(\hat{\Theta}) = \theta$, we call $\hat{\Theta}$ an **unbiased estimator** of θ.*

REMARK
An unbiased estimator has mean value equals to the true value of the parameter.



EXAMPLE 6.3 (UNBIASED ESTIMATOR)
Let $X_1, X_2, \ldots, X_n$ be a random sample from the same population with mean μ and variance σ^2. Then

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

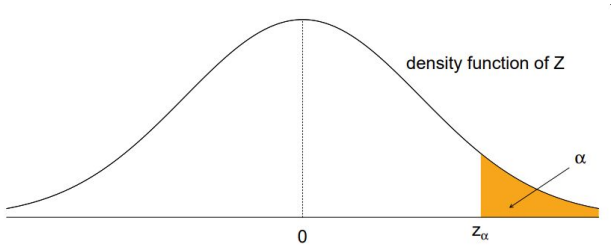
is an unbiased estimator of σ^2 since $E(S^2) = \sigma^2$.

Maximum Error of Estimate

Typically $\overline{X} \neq \mu$, so $\overline{X} - \mu$ measures the difference between the estimator and the true value of the parameter.

Recall that if the population is normal or if n is large, $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ follows a standard normal or an approximately standard normal distribution.

DEFINITION 3 (z_α)
Define z_α to be the number with an upper-tail probability of α for the standard normal distribution Z. That is, $P(Z > z_\alpha) = \alpha$.

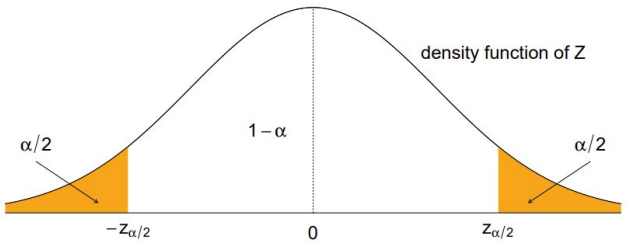


From the above definition, we then have

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

In other words,

$$P\left(\frac{|\overline{X} - \mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(|\overline{X} - \mu| \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$



This means that, with probability $1 - \alpha$, the error $|\overline{X} - \mu|$ is less than

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

DEFINITION 4 (MAXIMUM ERROR OF ESTIMATE)
The quantity

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

is called the ***maximum error of estimate****.*

EXAMPLE 6.4 (TV TIME FOR INTERNET USERS)
An investigator is interested in the amount of time internet users spend watching television per week.

Based on historical experience, he assumes that the standard deviation is $\sigma = 3.5$ hours.

He proposes to select a random sample of $n = 50$ internet users, poll them, and take the sample mean to estimate the population mean μ.

What can he assert with probability 0.99 about the maximum error of estimate?

Solution:
As $n = 50 \geq 30$ is large, $\dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is approximately normal.

So we can use the previous result, with $\sigma = 3.5$, $\alpha = 0.01$ and $z_{\alpha/2} = z_{0.005} = 2.576$.

With probability 0.99, the error is at most

$$E = 2.576 \times \frac{3.5}{\sqrt{50}} \approx 1.27.$$

REMARK
$z_{0.005}$ is the same as the 0.995 quantile of the standard normal. The value of 2.576 can be obtained from tables or software.

Use the command `qnorm(0.995)` or `qnorm(0.005, lower.tail=F)` to obtain the value via https://rdrr.io/snippets/.

Alternatively, you may use Radiant to get the same value as well.

Determination of Sample Size

We often want to know what the minimum sample size should be, so that with probability $1-\alpha$, the error is at most E_0.

To answer this, consider the fact that we want

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq E_0.$$

Solving for n, we have

$$n \geq \left(\frac{z_{\alpha/2}\cdot\sigma}{E_0}\right)^2.$$

Different Cases

We had previously understood the sampling distribution of $\overline{X}$ for a variety of cases. Repeating the same arguments above, we have the following table.

DIFFERENT CASES

	Population	σ	n	Statistic	E	n for desired E_0 and α
I	Normal	known	any	$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\cdot\sigma}{E_0}\right)^2$
II	any	known	large	$Z=\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	$z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\cdot\sigma}{E_0}\right)^2$
III	Normal	unknown	small	$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}$	$t_{n-1;\alpha/2}\cdot\frac{s}{\sqrt{n}}$	$\left(\frac{t_{n-1;\alpha/2}\cdot s}{E_0}\right)^2$
IV	any	unknown	large	$Z=\frac{\overline{X}-\mu}{S/\sqrt{n}}$	$z_{\alpha/2}\cdot\frac{s}{\sqrt{n}}$	$\left(\frac{z_{\alpha/2}\cdot s}{E_0}\right)^2$

2 Confidence Intervals for the Mean

Since a point estimate is almost never right, one might be interested in asking for an interval where the parameter lies in.

> **Definition 5 (Confidence Interval)**
> *An **interval estimator** is a rule for calculating, from the sample, an interval (a,b) in which you are fairly certain the parameter of interest lies in.*
>
> *This "fairly certain" can be quantified by the **degree of confidence** also known as **confidence level** $(1-\alpha)$, in the sense that*
>
> $$P(a < \mu < b) = 1 - \alpha.$$
>
> *(a,b) is called the $(1-\alpha)$ **confidence interval**.*

Case I: σ known, data normal

Consider the case where σ is known, and data comes from a normal population.

We learnt previously that

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

Rearranging, we have

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

So

$$\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

is a $(1-\alpha)$ confidence interval.

EXAMPLE 6.5
In order to set inventory levels, a computer company samples demand during lead time over 25 time periods:

```
235 374 309 499 253 421 361 514 462 369 394 439
348 344 330 261 374 302 466 535 386 316 296 332 334
```

It is known that the (population) standard deviation of demand over lead time is 75 computers. Given that $\bar{x} = 370.16$, estimate the mean demand over lead time with 95% confidence. Assume a normal distribution for the population.

Solution:
Note that $z_{\alpha/2} = z_{0.025} = 1.96$. The 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 370.16 \pm 1.96\frac{75}{\sqrt{25}} = 370.16 \pm 29.4$$

or $(340.76, 399.56)$.

REMARK
Notice that our $(1-\alpha)$ confidence interval can be written as $\overline{X} \pm E$.

This is not a coincidence: recall that there is $(1-\alpha)$ confidence that the error $|\overline{X} - \mu|$ is within E.

For the other cases, based on our understanding of the sampling distribution of $\overline{X}$, we can construct our confidence intervals for the different cases $\overline{X} \pm E$, based on the conditions given.

CONFIDENCE INTERVALS FOR THE MEAN
The table below gives the $(1-\alpha)$ confidence interval (formulas) for the population mean.

Case	Population	σ	n	Confidence Interval
I	Normal	known	any	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
II	any	known	large	$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$
III	Normal	unknown	small	$\bar{x} \pm t_{n-1;\alpha/2} \cdot s/\sqrt{n}$
IV	any	unknown	large	$\bar{x} \pm z_{\alpha/2} \cdot s/\sqrt{n}$

Note that n is considered large when $n \geq 30$.

EXAMPLE 6.6 (WHICH CASE?)
The following data set collects $n = 41$ randomly sampled waiting times of students from ST2334 to receive reply for their email from a survey in the day time.

```
 2.50  23.28  19.34   4.74   7.03  21.85   2.72
17.73  21.55   9.71  30.24   0.37  31.26  35.24
 7.81  16.69  66.54   1.88  14.14  46.59  28.17
 0.06   9.32   0.03  10.75   6.97  56.86   2.89
 7.67  30.16   0.33   0.44   3.77  25.07   7.05
 0.08  10.64  13.10   7.92 112.77  11.93
```

Given that $\bar{x} = 17.736$ and $s = 21.7$, construct a 98% confidence interval for the mean waiting time of *all ST2334 students*.

Solution:
Note that σ is unknown, and n is large. So we are in Case IV.

Note that $z_{\alpha/2} = z_{0.01} = 2.326$. So our 98% confidence interval is

$$\begin{aligned}\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} &= 17.736 \pm 2.326 \times \frac{21.7}{\sqrt{41}} \\ &= (9.85, 25.62).\end{aligned}$$

EXAMPLE 6.7 (WHICH CASE AGAIN?)
The contents of 7 similar containers of sulphuric acid (in litres) are

```
9.8  10.2  10.4  9.8  10.0  10.2  9.6
```

It can shown that $\bar{x} = 10$ and $s^2 = 0.08$. Find a 95% confidence interval for the mean content of all such containers, assuming an approximate normal distribution for container contents.

Solution:
We are in Case III.

Using software, we obtain $t_{6;0.025} = 2.447$.

Thus a 95% confidence interval for the mean content of all such containers is given as

$$\bar{x} \pm t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}} = 10 \pm 2.447 \cdot \frac{\sqrt{0.08}}{\sqrt{7}} = (9.738, 10.262).$$

INTERPRETING CONFIDENCE INTERVALS I

- We saw that $\overline{X} \pm E$ has probability $(1-\alpha)$ of containing μ.

  This is a probability statement about the procedure by which we compute the interval — the interval estimator.

- Each time we take a sample, and go through this construction, we get a different confidence interval.
- Sometimes we get a confidence interval that contains μ, and sometimes we get one that does not contain μ.
- Once an interval is computed, μ is either in it or not. There is no more randomness.

INTERPRETING CONFIDENCE INTERVALS II

- Since μ is typically not known, there is no way to determine whether a particular confidence interval succeeded in capturing the population mean.
- However, if we repeat this procedure of taking a sample and computing a confidence interval many times, about $(1-\alpha)$ of the many confidence intervals that we get will contain the true parameter.

  This is what "confidence" means — a confidence in the method used.

- The following R Shiny app allows us to explore this fact:
  `https://istats.shinyapps.io/ExploreCoverage/`

3 COMPARING TWO POPULATIONS

In real applications, it is quite common to compare the means of two populations.

Imagine that we have two populations

- Population 1 has mean μ_1, variance σ_1^2.
- Population 2 has mean μ_2, variance σ_2^2.

Experimental Design

In order to compare two populations, a number of observations from each population need to be collected. Experimental design refers to the manner in which samples from populations are collected.

TWO BASIC DESIGNS FOR COMPARING TWO TREATMENTS

- Independent samples — complete randomization.
- Matched pairs samples — randomization between matched pairs.

EXAMPLE 6.8 (INDEPENDENT SAMPLES)
In order to compare the examination scores of male and female students attending ST2334,

- 10 scores of female students are randomly sampled — Sample I,
- 8 scores of male students are randomly sampled — Sample II.

Note that all observations are independent —

- Sample I and Sample II are independent;
- Individuals within Sample I are independent;
- Individuals within Sample II are independent.

EXAMPLE 6.9 (MATCHED PAIRS SAMPLES)
In order to study whether there exists income difference between male and female, 100 married couples are sampled, and their monthly incomes are collected.

In this example, the treatment groups are the female group and male group.

Note that observations are dependent in a special way —

- Within the pair, the observations are dependent (since they are married to one another);
- Between pairs, observations are independent.

4 Independent Samples: Unequal variances

Our interest is to make statistical inference on $\mu_1 - \mu_2$. Consider the following assumptions:

Independent Samples (Known and unequal variances)

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are known and not the same: $\sigma_1^2 \neq \sigma_2^2$
5. Either one of the following conditions holds:
   - The two populations are normal; OR
   - Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

Consider $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, random samples from the two populations of interest. Let

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i, \text{ and } \overline{Y} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_i$$

be the means of random samples. Then,

$$E(\overline{X}) = \mu_1, \quad V(\overline{X}) = \frac{\sigma_1^2}{n_1}, \quad E(\overline{Y}) = \mu_2, \quad V(\overline{Y}) = \frac{\sigma_2^2}{n_2}.$$

Thus

$$E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 = \delta,$$

and, using the independence assumption,

$$V(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

When

- the two populations are normal, OR
- both samples are large,

we have

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1).$$

Confidence Intervals for $\mu_1 - \mu_2$

We are interested in the difference

$$\delta = \mu_1 - \mu_2,$$

with confidence $100(1-\alpha)\%$ for any $0 < \alpha < 1$.

If σ_1^2 and σ_2^2 are **known**, by the distributions above, we have

$$P\left(\left|\frac{(\overline{X}-\overline{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}\right| < z_{\alpha/2}\right) = 1-\alpha$$

or

$$P\left((\overline{X}-\overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1-\mu_2 < (\overline{X}-\overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) = 1-\alpha.$$

Thus the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\left((\overline{X}-\overline{Y}) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\ (\overline{X}-\overline{Y}) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

or

$$(\overline{X}-\overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}.$$

CONFIDENCE INTERVALS: KNOWN AND UNEQUAL VARIANCES

Suppose we have independent populations with known and unequal variances, and that either one of the following conditions holds:

- The two populations are normal; OR
- Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

The $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, is then given as

$$(\overline{x}-\overline{y}) \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

EXAMPLE 6.10
A study was conducted to compare two types of engines, A and B.

Gas mileage, in miles per gallon, was measured. 50 experiments were conducted using engine A. 75 experiments were done for engine type B. The gasoline used and other conditions were held constant.

- The average gas mileage for 50 experiments using engine A was 36 miles per gallon and
- The average gas mileage for the 75 experiments using machine B was 42 miles per gallon.

Find a 96% confidence interval on $\mu_B - \mu_A$, where μ_A and μ_B are the population mean gas mileage for machine types A and B, respectively.

Assume that the population standard deviations are 6 and 8 for machine types A and B, respectively.

Solution:
For a 96% confidence interval, $\alpha = 0.04$ and $z_{0.02} = 2.05$. We are also given that

$$n_1 = 50, \bar{x}_A = 36, \ \sigma_1^2 = 6^2$$
$$n_2 = 75, \bar{x}_B = 42, \ \sigma_2^2 = 8^2$$

The sample sizes are large, so a 96% confidence interval for $\mu_B - \mu_A$ is

$$\begin{aligned}
&(\bar{x}_B - \bar{x}_A) \pm z_{\alpha/2}\sqrt{\sigma_2^2/n_2 + \sigma_1^2/n_1} \\
&= (42 - 36) \pm 2.05 \cdot \sqrt{8^2/75 + 6^2/50} \\
&= (3.428, 8.571).
\end{aligned}$$

We next consider the following assumptions/case:

INDEPENDENT SAMPLES (LARGE, WITH UNKNOWN VARIANCES)
1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are unknown and not the same: $\sigma_1^2 \neq \sigma_2^2$
5. Both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

Since σ_1 and σ_2 are unknown, let

$$S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_i - \overline{X})^2, \text{ and } S_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$$

and use

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1).$$

If σ_1^2 and σ_2^2 are **unknown**, the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$\left((\overline{X} - \overline{Y}) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},\ (\overline{X} - \overline{Y}) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$$

or

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

CONFIDENCE INTERVALS: LARGE, WITH UNKNOWN VARIANCES

Suppose we have independent populations with unknown and unequal variances, and that both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

The $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$, is then given as

$$(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2}.$$

5 INDEPENDENT SAMPLES: EQUAL VARIANCES

Consider the following assumptions:

INDEPENDENT SAMPLES: SMALL, WITH EQUAL VARIANCES

1. A random sample of size n_1 from population 1 with mean μ_1 and variance σ_1^2.
2. A random sample of size n_2 from population 2 with mean μ_2 and variance σ_2^2.
3. The two samples are independent.
4. The population variances are unknown and the same: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
5. Both samples are small: $n_1 < 30$, $n_2 < 30$
6. Both populations are normally distributed.

THE EQUAL VARIANCE ASSUMPTION
In real applications, the equal variance assumption is usually unknown and needs to be checked.

Based upon the normal distribution and equal variance assumptions

$$Z = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1).$$

Since σ is unknown, we shall estimate it.

Note that S_1^2 and S_2^2 are both unbiased estimators of σ^2 under the equal variance assumption.

We can use the pooled estimator to estimate σ^2 better.

DEFINITION 6 (THE POOLED ESTIMATOR: S_p^2)
*σ^2 can be estimated by the **pooled sample variance***

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

with S_1^2 and S_2^2 being the sample variances of the first and second samples respectively.

When we estimate σ^2 using S_p^2, the resulting statistic

$$T = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

follows a t-distribution with degrees of freedom $n_1 + n_2 - 2$.

We then have

$$P\left(-t_{n_1+n_2-2;\alpha/2} < \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{n_1+n_2-2;\alpha/2}\right) = 1 - \alpha.$$

CONFIDENCE INTERVALS: SMALL, WITH EQUAL VARIANCES
Suppose we have independent, normal populations with unknown and equal variances, and that both samples are small: $n_1 < 30$, $n_2 < 30$.

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given as

$$(\overline{X} - \overline{Y}) \pm t_{n_1+n_2-2;\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

EXAMPLE 6.11
A course in mathematics is taught to 12 students by the conventional classroom procedure. A second group of 10 students was given the same course by means of programmed materials.

At the end of the semester the same examination was given to each group.

- The 12 students meeting in the classroom made an average grade of 85 with standard deviation of 4.
- The 10 students using programmed materials made an average of 81 with a standard deviation of 5.

Find a 90% confidence interval for the difference between the population means, assuming the populations are approximately normally distributed with equal variances.

Solution:
Let μ_1 and μ_2 represent the average grades of all students who might take this course by the classroom and programmed presentations respectively.

So $\overline{x} - \overline{y} = 85 - 81 = 4$ is the point estimate for $\mu_1 - \mu_2$.

As we assume equal population variance, we estimate it by the pooled variance

$$s_p^2 = \frac{(12-1)\times 4^2 + (10-1)\times 5^2}{12+10-2} = 20.05.$$

In this case, $t_{n_1+n_2-2;\alpha/2} = t_{20;0.05} = 1.7247$. Thus a 90% confidence interval for $\mu_1 - \mu_2$ is given as

$$\begin{aligned}
&(\overline{x} - \overline{y}) \pm t_{n_1+n_2-2;\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= (85-81) \pm 1.7247 \times \sqrt{20.05} \times \sqrt{\frac{1}{12} + \frac{1}{10}} \\
&= (0.693, 7.307).
\end{aligned}$$

Independent Large Samples with Equal Variance

Note that for large samples such that $n_1 \geq 30$, $n_2 \geq 30$, we can replace $t_{n_1+n_2-2;\alpha/2}$ by $z_{\alpha/2}$ in the previous formula.

CONFIDENCE INTERVALS: LARGE, WITH EQUAL VARIANCES
Suppose we have independent populations with unknown and equal variances, and that both samples are large: $n_1 \geq 30$, $n_2 \geq 30$.

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given as

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

6 PAIRED DATA

Some times, like in the couple income example, it makes sense to take matched pairs instead of independent samples.
Because of dependence in the sample, the methods discussed previously are not applicable.

Consider the assumptions that follows.

PAIRED DATA
1. $(X_1, Y_1), \ldots, (X_n, Y_n)$ are matched pairs, where $X_1, \ldots, X_n$ is a random sample from population 1, $Y_1, \ldots, Y_n$ is a random sample from population 2.
2. X_i and Y_i are dependent.
3. (X_i, Y_i) and (X_j, Y_j) are independent for any $i \neq j$.
4. For matched pairs, define $D_i = X_i - Y_i, \quad \mu_D = \mu_1 - \mu_2$.
5. Now we can treat $D_1, D_2, \ldots, D_n$ as a random sample from a single population with mean μ_D and variance σ_D^2.

All techniques derived for a single population can now be employed.

- We consider the statistic

$$T = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n}}, \quad \text{where} \quad \overline{D} = \frac{\sum_{i=1}^n D_i}{n}, \quad S_D^2 = \frac{\sum_{i=1}^n (D_i - \overline{D})^2}{n-1}.$$

- If $n < 30$ and the population is normally distributed then

$$T \sim t_{n-1}.$$

- If $n \geq 30$, then
$$T \sim N(0,1).$$

CONFIDENCE INTERVALS: PAIRED DATA
For paired data, if n is small ($n < 30$) and the population is normally distributed, a $(1-\alpha)100\%$ confidence interval for μ_D is
$$\overline{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

If n is large ($n \geq 30$), a $(1-\alpha)100\%$ confidence interval for μ_D is
$$\overline{d} \pm z_{\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

EXAMPLE 6.12
Twenty students were divided into 10 pairs, each member of the pair having approximately the same IQ.

One of each pair was selected at random and assigned to a mathematics section using programmed materials only. The other member of each pair was assigned to a section in which the professor lectured.

At the end of the semester each group was given the same examination and the following results were recorded.

Pair	1	2	3	4	5	6	7	8	9	10
P.M.	76	60	85	58	91	75	82	64	79	88
Lecture	81	52	87	70	86	77	90	63	85	83
d	–5	8	–2	–12	5	–2	–8	1	–6	5

Given that $\overline{d} = -1.6$ and $s_D^2 = 40.71$, compute a 98% confidence interval for the true difference in the two learning procedures.

Solution:
Since $\alpha = 0.02$, we have $t_{n-1;\alpha/2} = t_{9,0.01} = 2.821$. Thus a 98% confidence interval for the true difference μ_D is given as
$$\overline{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}} = -1.6 \pm 2.821 \times \sqrt{\frac{40.71}{10}} = (-7.292, 4.092).$$

Seven

Hypothesis Tests

1 Hypothesis Tests

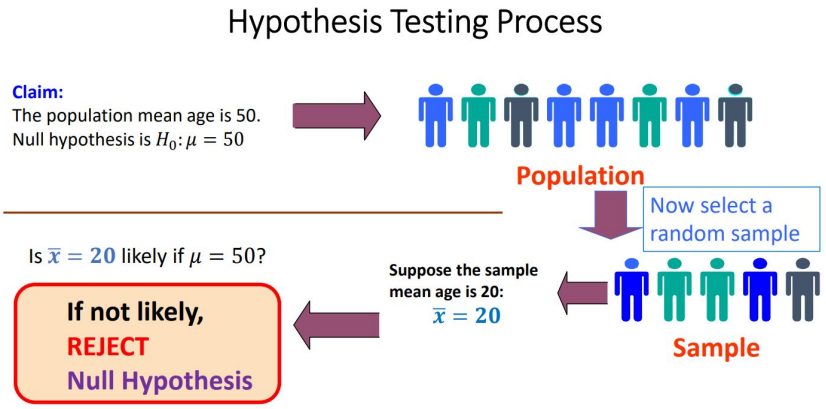
One of the most fundamental technique of statistical inference is the hypothesis test. There are many types of hypothesis tests but all follow the same logical structure, so we begin with hypothesis testing of a population mean.

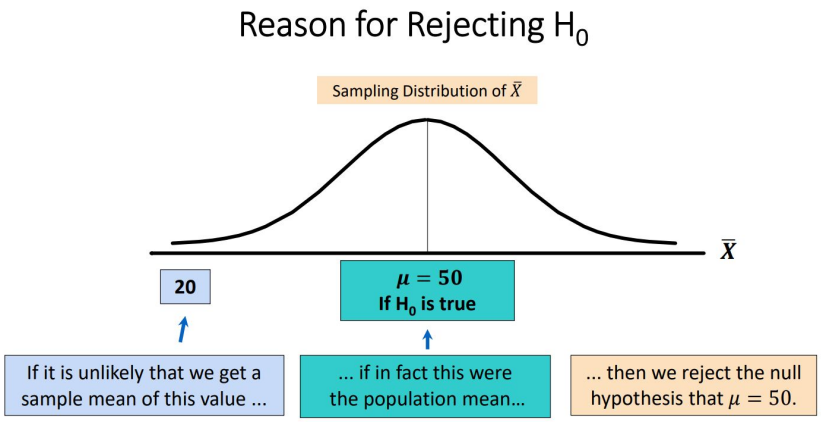
Hypothesis testing begins with a null hypothesis and an alternative hypothesis. Both the null and the alternative hypotheses are statements about a population. In this chapter, that statement will be a statement about the mean(s) of the population(s).

We will illustrate using an example.

Example 7.1 (Mean Age)
We are interested to check if the mean age of a population is $\mu = 50$.

Suppose we have no access to population data. So we take a sample from the population and obtained a sample mean age of $\bar{x} = 20$. Does this gives evidence for or against the hypothesis that $\mu = 50$?





EXAMPLE 7.2 (NUS STUDENTS' IQ)
Consider the statement

"NUS students have higher IQ than the general population (100)."

It is difficult/expensive to ask every NUS student to take an IQ test. So we take a sample.

Suppose the sample average is 110.

- Does that mean we're right?
- What if the sample average is 101? What about 100.1?
- Does the sample size matter?

HOW TO DO A HYPOTHESIS TEST
There are five main steps to hypothesis testing.

Step 1: Set your competing hypotheses: null and alternative.

Step 2: Set the level of significance.

Step 3: Identify the test statistic, its distribution and the rejection criteria.

Step 4: Compute the observed test statistic value, based on your data.

Step 5: Conclusion.

Let us have a closer look at each step.

Step 1: Null Hypothesis vs Alternative Hypothesis

Our goal is to decide between two competing hypotheses.

NULL VS ALTERNATIVE

In general, we adopt the position of the null hypothesis unless there is overwhelming evidence against it.

The null hypothesis is typically the default assumption, or the conventional wisdom about a population. Often it is exactly the thing that a researcher is trying to show is false.

We usually let the hypothesis that we want to prove be the alternative hypothesis. The alternative hypothesis states that the null hypothesis is false, often in a particular way.

The outcome of hypothesis testing is to either reject or fail to reject the null hypothesis.

A researcher would collect data relating to the population being studied and use a hypothesis test to determine whether the evidence against the null hypothesis (if any) is strong enough to reject the null hypothesis in favor of the alternative hypothesis.

We usually phrase the hypotheses in terms of population parameters.

EXAMPLE 7.3 (ONE-SIDED TEST)

Let μ be the average IQ of NUS students. Consider

$$H_0 : \mu = 100 \quad \text{vs} \quad H_1 : \mu > 100.$$

This is an example of a one-sided hypothesis test.

For this alternative hypothesis, we do not care if $\mu < 100$: the goal here is just to show NUS students have IQ higher than 100.

EXAMPLE 7.4 (TWO-SIDED TEST)

Sometimes it is more natural to do a two-sided hypothesis test.

For example, let p be the probability of heads for a particular coin. You want to test if the coin is fair (that is, $p = 0.5$), as it is equally problematic if p was larger or smaller.

Hence you set your hypotheses to be

$$H_0 : p = 0.5 \quad \text{vs} \quad H_1 : p \neq 0.5.$$

Step 2: Level of Significance

For any test of hypothesis, there are two possible conclusions:

- Reject H_0 and therefore conclude H_1;
- Do not reject H_0 and therefore conclude H_0.

Whatever decision is made, there is a possibility of making an error.

	Do not reject H_0	Reject H_0
H_0 is true	Correct Decision	**Type I error**
H_0 is false	**Type II error**	Correct Decision

DEFINITION 1 (TYPE I VS TYPE II ERROR)
*The rejection of H_0 when H_0 is true is called a **Type I error**.*

*Not rejecting H_0 when H_0 is false is called a **Type II error**.*

DEFINITION 2 (SIGNIFICANCE LEVEL VS POWER)
*The probability of making a Type I error is called the **level of significance**, denoted by α. That is,*

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 \mid H_0 \text{ is true}).$$

Let

$$\beta = P(\text{Type II error}) = P(\text{Do not reject } H_0 \mid H_0 \text{ is false}).$$

*We define $1 - \beta = P(\text{Reject } H_0 \mid H_0 \text{ is false})$ to be the **power of the test**.*

REMARK
The Type I error is considered a serious error, so we want to control the probability of making such an error.

Thus prior to conducting a hypothesis test, we set the significance level α to be small, typically at $\alpha = 0.05$ or 0.01.

Step 3: Test Statistic, Distribution and Rejection Region

To test the hypothesis, we first select a suitable test statistic for the parameter under the hypothesis.

The test statistic serves to quantify just how unlikely it is to observe the sample, assuming the null hypothesis is true.

As the significance level α is given, a decision rule can be found such that it divides the set of all possible values of the test statistic into two regions, one being the rejection region (or critical region) and the other, the acceptance region.

Step 4 & 5: Calculation and Conclusion

Once a sample is taken, the value of the test statistic is obtained.

We check if it is within our rejection region.

- If it is, our sample was too improbable assuming H_0 is true, hence we reject H_0.
- If it is not, we did not accomplish anything. We failed to reject H_0 and hence fall back to our original assumption of H_0.

Note that in the latter case, we did not "prove" that H_0 is true. Hence, it is prudent to use the term "fail to reject H_0" instead of "accept H_0."

2 Hypotheses concerning the Mean

Let's apply our hypothesis steps to testing a population mean.

Case: Known variance

Let us consider the case where

- the population variance σ^2 is known; AND
- where
  - the underlying distribution is normal; OR
  - n is sufficiently large (say, $n \geq 30$).

Step 1: We set the null and alternatives hypotheses as

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

Note that in this case we are considering a two-sided alternative hypothesis.

Step 2: Set level of significance: α is typically set to be 0.05.

Step 3: Statistic & its distribution:
With σ^2 known and population normal (or $n \geq 30$),

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

When H_0 is true, $\mu = \mu_0$, the above becomes

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1),$$

and will serve as our test statistic.

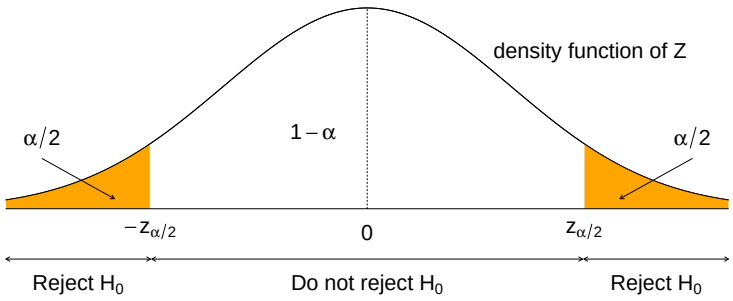
Rejection region:
Intuitively, we should reject H_0 when $\overline{X}$ is too large or too small compared with μ_0.

This is the same as when Z is too large or too small. In theory,

$$P(|Z| > z_{\alpha/2}) = \alpha.$$

Let the observed value of Z be z. Then the rejection region is defined by $|z| > z_{\alpha/2}$, which is

$$z < -z_{\alpha/2} \quad \text{or} \quad z > z_{\alpha/2}.$$



Step 4: Computations: z should be computed from the statistic above based upon the observed sample.

Step 5: Conclusion: check whether z is located within rejection region. If so, reject H_0, otherwise do not reject H_0.

WHERE DID THE VALUE 0.05 COME FROM?

In 1931, in a famous book called The Design of Experiments, *Sir Ronald Fisher discussed the amount of evidence needed to reject a null hypothesis.*

He said that it was situation dependent, but remarked, somewhat casually, that for many scientific applications, 1 out of 20 might be a reasonable value.

Since then, some people — indeed some entire disciplines — have treated the number 0.05 as sacrosanct.

Sir Ronald Fisher (1890 – 1962) was one of the founders of modern Statistics. For a biography of Fisher, browse to
http://www-history.mcs.st-andrews.ac.uk/Biographies/Fisher.html

EXAMPLE 7.5

The director of a factory wants to determine if a new machine A is producing cloths with a breaking strength of 35 kg with a standard deviation of 1.5 kg.

A random sample of 49 pieces of cloths is tested and found to have a mean breaking strength of 34.5 kg. Is there evidence that the machine is not meeting the specifications for mean breaking strength?
Use $\alpha = 0.05$.

Solution:

Note that $n > 30$ and $\sigma = 1.5$.

Let μ be the mean breaking strength of cloths manufactured by the new machine.

Step 1: We test

$$H_0 : \mu = 35 \quad \text{vs} \quad H_1 : \mu \neq 35.$$

Step 2: Set $\alpha = 0.05$.

Step 3: As σ^2 is known and $n \geq 30$,

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

will serve as our test statistic.

Since $z_{\alpha/2} = z_{0.025} = 1.96$, the critical/rejection region is

$$z < -1.96 \quad \text{or} \quad z > 1.96.$$

Step 4: z is computed to be

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{34.5 - 35}{1.5/\sqrt{49}} = -2.3333 < -1.96.$$

Step 5: The observed z value, $z = -2.3333$, falls inside the critical region. Hence the null hypothesis $H_0 : \mu = 35$ is rejected at the 5% level of significance.

One-sided alternatives

Now the above procedures are establish under

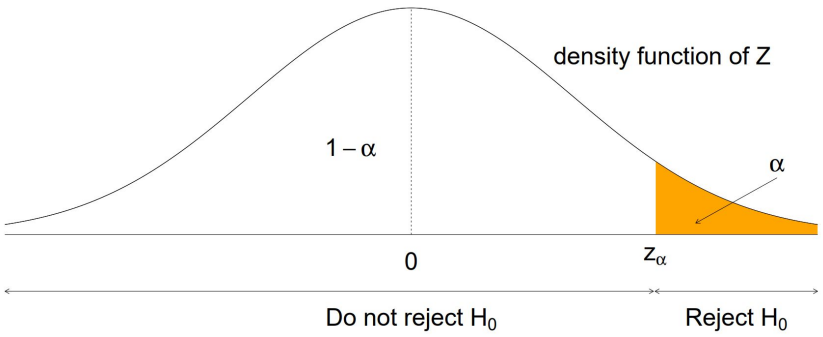
$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

Suppose instead we are considering

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu > \mu_0.$$

Similar steps can be used to address this problem, we only need to do the following changes:

- Step 1: H_1 is replaced with $H_1 : \mu > \mu_0$.
- Step 3: The test statistic and its distribution are kept the same. The rejection region should be replaced with $z > z_\alpha$, since now, we should reject only when $\bar{x}$ (and therefore z) is large.



The case for

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu < \mu_0$$

should be self-evident.

HYPOTHESIS TEST FOR THE MEAN: KNOWN VARIANCE
Consider the case where

- the population variance σ^2 is known; AND
- where
  - the underlying distribution is normal; OR
  - n is sufficiently large (say, $n \geq 30$).

For the null hypothesis $H_0 : \mu = \mu_0$, the test statistics is given by

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

Let z be the observed Z value. For the alternative hypothesis

- $H_1 : \mu \neq \mu_0$, the rejection region is

$$z < -z_{\alpha/2} \quad \text{or} \quad z > z_{\alpha/2}.$$

- $H_1 : \mu < \mu_0$, the rejection region is

$$z < -z_{\alpha}.$$

- $H_1 : \mu > \mu_0$, the rejection region is

$$z > z_{\alpha}.$$

p-value approach to testing

The above technique introduced by Fisher is based on a pre-declared significance level α.

Today, there is little reason to stick to the arbitrary 1% or 5% levels that Fisher suggested. We can instead use the idea of the p-value.

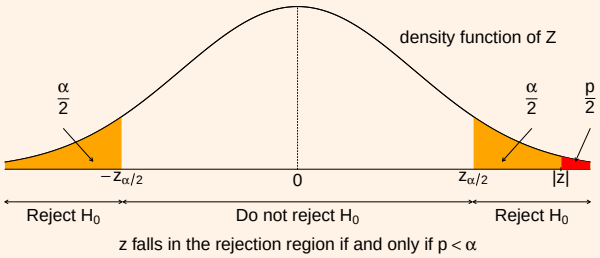
> **DEFINITION 3 (p-VALUE)**
> *The **p-value** is the probability of obtaining a test statistic at least as extreme ($\leq$ or $\geq$) than the observed sample value, given H_0 is true.*
>
> *It is also called the observed level of significance.*

p-VALUE FOR HYPOTHESIS TESTS

Suppose our computed test statistic was z. For a two sided test, a "worse" result would be if $Z > |z|$ or $Z < -|z|$, in other words, $|Z| > |z|$.

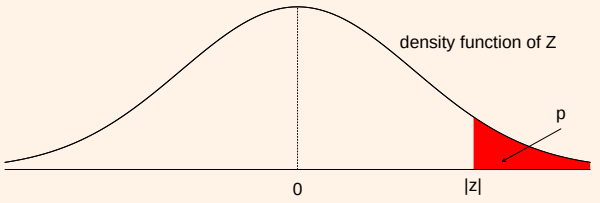
So the p-value is given by

$$p\text{-value} = P(|Z| > |z|) = 2P(Z > |z|) = 2P(Z < -|z|)$$



For the alternative hypothesis $H_1 : \mu < \mu_0$, the p-value is $P(Z < -|z|)$.
That is, only the area in the left tail is used.

For the alternative hypothesis $H_1 : \mu > \mu_0$, the p-value is $P(Z > |z|)$.
That is, only the area in the right tail is used.



REJECTION CRITERIA USING *p*-VALUE

We see that the p-value is smaller than the significance level *if and only if* our test statistic is in the rejection region.

Thus our rejection criteria would be

- If p-value $< \alpha$, reject H_0; else
- If p-value $\geq \alpha$, do not reject H_0.

REMARK
In practice, it is better to report the p-value than to indicate whether H_0 is rejected.

- The p-values of 0.049 and 0.001 both result in rejecting H_0 when $\alpha = 0.05$, but the second case provides much stronger evidence.
- p-values of 0.049 and 0.051 provide, in practical terms, the same amount of evidence about H_0.

Most research articles report the p-value rather than a decision about H_0. From the p-value, readers can view the strength of evidence against H_0 and make their own decision, if they want to.

EXAMPLE 7.6 (MIDTERM EXAM SCORE)
Recall the midterm exam scores example in an earlier chapter. The data obtained are

$$20, 19, 24, 22, 25.$$

We were told that the exam scores are approximately normal.

The lecturer announced that the variance of the exam score over the class is 5 (just believe that this is the truth). Test at $\alpha = 0.01$ significance level whether the average midterm score is different from 16.

Solution:
Let μ be the average midterm score for the whole class.

Step 1: $H_0 : \mu = 16 \quad \text{vs} \quad H_1 : \mu \neq 16.$

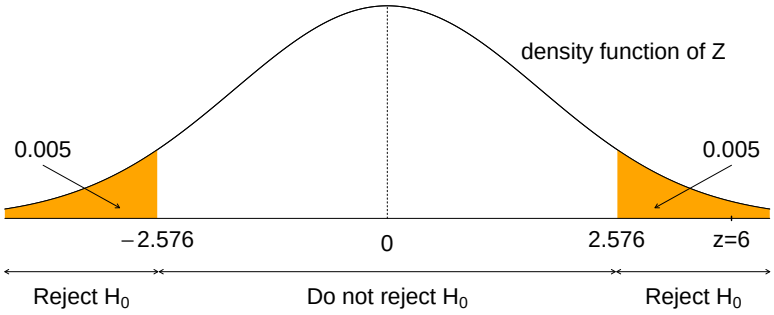
Step 2: Choose $\alpha = 0.01$.

Step 3: In this example $\sigma = \sqrt{5}$ is known, data are normal, and $n = 5$. Therefore the test statistic and its distribution is

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

Now $z_{\alpha/2} = z_{0.005} = 2.576$. Thus the rejection region is

$$z < -2.576 \quad \text{or} \quad z > 2.576.$$



Step 4: $z = (22-16)/(\sqrt{5}/\sqrt{5}) = 6 > 2.576$.

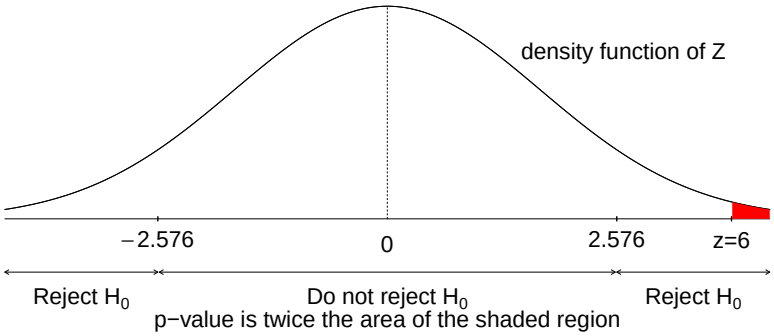
Step 5: As $z = 6$ falls in rejection region, H_0 is rejected.

Alternatively, we can use the p-value approach.

Note that the p-value is given, using a computer, as

$$2P(Z > 6) = 1.973175 \times 10^{-9},$$

which is smaller than $\alpha = 0.01$. So we reject H_0.



We can use our knowledge of the sampling distribution to determine the test statistic for other situations.

HYPOTHESIS TEST FOR THE MEAN: UNKNOWN VARIANCE
Consider the case where

- the population variance σ^2 is unknown; AND
- the underlying distribution is normal.

For the null hypothesis $H_0 : \mu = \mu_0$, the test statistics is given by

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

Let t be the observed T value. For the alternative hypothesis

- $H_1 : \mu \neq \mu_0$, the rejection region is

$$t < -t_{n-1,\alpha/2} \quad \text{or} \quad t > t_{n-1,\alpha/2}.$$

- $H_1 : \mu < \mu_0$, the rejection region is

$$t < -t_{n-1,\alpha}.$$

- $H_1 : \mu > \mu_0$, the rejection region is

$$t > t_{n-1,\alpha}.$$

REMARK
When $n \geq 30$, we can replace t_{n-1} by Z, the standard normal distribution.

3 TWO-SIDED TESTS AND CONFIDENCE INTERVALS

In this section, we establish that the two-sided hypothesis test procedure is equivalent to finding a $100(1-\alpha)\%$ confidence interval for μ.

We illustrate using Case III: normal population, small n, unknown σ.

Once again, consider

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

The $100(1-\alpha)\%$ confidence interval for μ in this case is given by

$$\left(\overline{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right).$$

If the $100(1-\alpha)\%$ confidence interval contains μ_0, we will have

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu_0 \le \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Rearranging the above inequality, we obtain

$$-t_{\alpha/2} \le \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \le t_{\alpha/2}.$$

This means that the computed test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ satisfies

$$-t_{\alpha/2} \le t \le t_{\alpha/2}.$$

Note that the rejection region for this case is

$$t < -t_{\alpha/2} \quad \text{or} \quad t > t_{\alpha/2}.$$

This means that when the confidence interval contains μ_0, H_0 will not be rejected at level α.

Similarly, when the confidence interval does not contain μ_0, then

$$t > t_{\alpha/2} \quad \text{or} \quad t < -t_{\alpha/2}.$$

Thus t falls within the rejection region and so H_0 will be rejected.

Therefore confidence intervals can be used to perform two-sided tests.

EXAMPLE 7.7 (MIDTERM EXAM SCORE III)
Back to Example 7.6, regarding midterm exam scores. Assume that the lecturer did not announce the variance, i.e., σ is unknown.

The student constructed a 99% ($\alpha = 0.01$) confidence interval for the average score of students for the midterm:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 22 \pm 4.604 \times \frac{2.55}{\sqrt{5}} = (16.75, 27.25).$$

The interval does not contain 16, so the following test of hypothesis should be rejected at $\alpha = 0.01$:

$$H_0 : \mu = 16 \quad \text{vs} \quad H_1 : \mu \neq 16.$$

What about

$$H_0 : \mu = 17 \quad \text{vs} \quad H_1 : \mu \neq 17?$$

4 Tests Comparing Means: Independent Samples

Suppose two independent samples are drawn from two populations with means μ_1 and μ_2. We are interested in testing

$$H_0 : \mu_1 - \mu_2 = \delta_0$$

against a suitable alternative hypothesis.

Comparing Means: Independent Samples I

(A) Consider the case where

- the population variances σ_1^2 and σ_2^2 are known; AND
- where
  - the underlying distributions are normal; OR
  - n_1, n_2 are sufficiently large (say, $n_1 \geq 30, n_2 \geq 30$).

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, the test statistics is given by

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).$$

(B) Consider the case where

- the population variances σ_1^2 and σ_2^2 are unknown; AND
- n_1, n_2 are sufficiently large (say, $n_1 \geq 30, n_2 \geq 30$).

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, the test statistics is given by

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1).$$

The rejection regions or p-values can be established similarly as before.

Rejection regions and p-values

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, and specified alternative H_1, the rejection regions and p-values are given below.

H_1	Rejection Region	p-value
$\mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$P(Z > \lvert z \rvert)$
$\mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$P(Z < -\lvert z \rvert)$
$\mu_1 - \mu_2 \neq \delta_0$	$z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$	$2P(Z > \lvert z \rvert)$

EXAMPLE 7.8
Analysis of a random sample consisting of $n_1 = 20$ specimens of cold-rolled steel to determine yield strengths resulted in a sample average strength of $\bar{x} = 29.8$ ksi.

A second random sample of $n_2 = 25$ two-side galvanized steel specimens gave a sample average strength of $\bar{y} = 34.7$ ksi.

Assuming that the two yield strength distributions are normal with $\sigma_1 = 4.0$ and $\sigma_2 = 5.0$, does the data indicate that the corresponding true average yield strengths μ_1 and μ_2 are different?
Use $\alpha = 0.01$.

Solution:
Let μ_1 and μ_2 be the mean strength of cold-rolled steel and two-side galvanized steel respectively.

Step 1: Note that $\delta_0 = 0$ in this example. So the hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0.$$

Step 2: Set $\alpha = 0.01$.

Step 3: Test statistic and its distribution is given below:

$$Z = \frac{(\overline{X} - \overline{Y}) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0,1).$$

Note that $z_{\alpha/2} = z_{0.005} = 2.5782$. Thus the rejection region is

$$z > 2.5782 \quad \text{or} \quad z < -2.5782.$$

Step 4: Plug in the data,

$$z = \frac{(29.8 - 34.7) - 0}{\sqrt{\frac{16}{20} + \frac{25}{25}}} = -3.652 < -2.5782 = -z_{\alpha/2}.$$

Step 5: Since $z = -3.652$ falls inside the critical region, hence $H_0 : \mu_1 = \mu_2$ is rejected at the 1% level of significance. We conclude that the sample data strongly suggest that the true average yield strength for cold-rolled steel differs from that for galvanized steel.

Alternatively, we can compute the p-value to be

$$2 \times P(Z < -3.652) = 0.00026 < 0.01 = \alpha.$$

Thus we reject the null hypothesis at $\alpha = 0.01$ level.

COMPARING MEANS: INDEPENDENT SAMPLES II
Consider the case where

- the population variances σ_1^2 and σ_2^2 are unknown but equal;
- the underlying distributions are normal;
- n_1, n_2 are small (say, $n_1 < 30, n_2 < 30$).

For the null hypothesis $H_0 : \mu_1 - \mu_2 = \delta_0$, the test statistics is given by

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

5 TESTS COMPARING MEANS: PAIRED DATA

Comparing means with matched-pairs data is easy. We merely use methods we have already learned for single samples.

COMPARING MEANS: PAIRED DATA
For paired data, define $D_i = X_i - Y_i$.

For the null hypothesis $H_0 : \mu_D = \mu_{D_0}$, the test statistics is given by

$$T = \frac{\overline{D} - \mu_{D_0}}{S_D/\sqrt{n}}.$$

- If $n < 30$ and the population is normally distributed then

$$T \sim t_{n-1}.$$

- If $n \geq 30$, then

$$T \sim N(0,1).$$

EXAMPLE 7.9 (TREATING CATALYST SURFACES)
Prof X developed a new procedure for treating catalyst surfaces which he claims will result in a significant enhancement in the number of active sites.

The number of active sites can be determined by absorption of H_2 gas.

Prof X tested each sample before and after the treatment and obtained the following H_2 uptake in terms of mmol/g.

Sample No.	Before treatment (X)	After treatment (Y)	Difference (D)
1	165	172	7
2	146	189	43
3	174	168	−6
4	186	176	−10
5	147	198	51
6	153	184	31
7	132	188	56
8	175	197	22

The summary statistics for the variable D are $\overline{d} = 24.25$ and $s_D = 25.34$.
Has the treatment resulted in an increase in the number of active sites on the catalyst surfaces? Assume normality, and test at $\alpha = 0.05$ level.

Solution:
Note that in such a setup the two samples are not independent, and so the two sample t-test does not apply.

Define $D_i = Y_i - X_i$, where X_i and Y_i are the "before treatment" and "after treatment" readings.

The question is now reduced to:

Do the data give any evidence that $\mu_D > 0$?

Step 1: We set the null and alternative to be

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D > 0.$$

Step 2: Set $\alpha = 0.05$.

Step 3: We use the paired t-test with the test statistics

$$T = \frac{\overline{D} - 0}{S_D/\sqrt{n}}.$$

The rejection region is $t > t_{7,0.05} = 1.895$.

Step 4: The observed t value is

$$t = \frac{\overline{d} - 0}{s_D/\sqrt{n}} = \frac{24.25 - 0}{25.34/\sqrt{8}} = 2.70 > 1.895.$$

Step 5: Since $t = 2.70 > t_{7,0.05} = 1.895$, we reject H_0 and conclude that there is evidence that treatment of catalysts increases the number of active sites.

As an aside, the p-value is

$$P(t_7 > t) = P(t_7 > 2.70) = 0.0153,$$

which is smaller than 0.05.