# Chapter 1: Basic Concepts of Probability

## 1 Basic Probability Concepts and Definitions

- **Statistical Experiment**: Any procedure that obtains data (observations).
- **Sample Space** (denoted by $S$): The set of all possible outcomes of a statistical experiment.

  *It depends on the problem of interest*!
- **Sample Point**: Every outcome (element) in a sample space.
- **Events**: Subset of a sample space.

  **Example 1**
  Consider an experiment of **tossing a die**.

  - If the problem of interest is "the number 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.
    - Events: (1) An event that an odd number occurs = $\{1, 3, 5\}$; (2) An event that a number greater than 4 occurs = $\{5, 6\}$.
  - If the problem of interest is "whether the number is even or odd", then
    - Sample space: $S = \{\text{even}, \text{odd}\}$.
    - Sample point: "even" or "odd".
    - Events: An event that 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**. ■

**L–example 1.1**

Consider an experiment of **tossing two dice**. Suppose that the problem of interest is "the numbers that show on the top faces".

here is a tuple since the order matters

- If the dices are labelled, $S$ contains 36 elements:
  - Sample space: $S = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),$ $(2,1),(2,2),\ldots,(6,5),(6,6)\}$.
  - Sample point: $(1,1)$ or $(1,2)$ or ......
  - Events: event $A = \{\text{the sum of the dice equals } 7\}$,

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}.$$

- If the dices are NOT labelled, $S$ contains 21 elements:
  - Sample space:

$$S = \{\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\}\}$$
$$\{2,2\},\{2,3\},\ldots,\{5,6\},\{6,6\}$$

  - Sample point: $\{1,1\}$ or $\{1,2\}$ or ......
  - Events: event $A = \{\text{the sum of the dice equals } 7\}$,

$$A = \{\{1,6\},\{2,5\},\{3,4\}\}.$$

**L–example 1.2** Consider a two step experiment:

1. Flip a coin and observe whether the head $(H)$ or the tail $(T)$ is facing up.
2. If H is obtained in step 1, then flip it again; otherwise, roll a die once.

- Sample space: $S = \{(H,H),(H,T),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$.
- Sample point: $(H,H)$ or $(H,T)$ or ......
- Events: $A =$ no die is thrown,

$$A = \{(H,H),(H,T)\}.$$

**L–example 1.3** Consider an experiment of drawing **two** balls from a jar with a blue, a white, and a red ball.

If the problem of interest is the colours of the two drawn balls, then

- Sample space:

$$S = \{(B,W),(B,R),(W,B),(W,R),(R,B),(R,W)\}.$$

- Sample point: $(B,W)$ or $(B,R)$ or ......
- Events: $A = \{\text{a white ball is chosen}\}$,

$$A = \{(W,B),(W,R),(B,W),(R,W)\}$$

## 2 Event Operations

- Denote by $S$ the sample space; let $A$ and $B$ be two events. Event operations and possible relationships are summarized below.
- Event operations include:

  (1). Union: $A \cup B$; (2). Intersection: $A \cap B$; (3). Complement: $A'$.

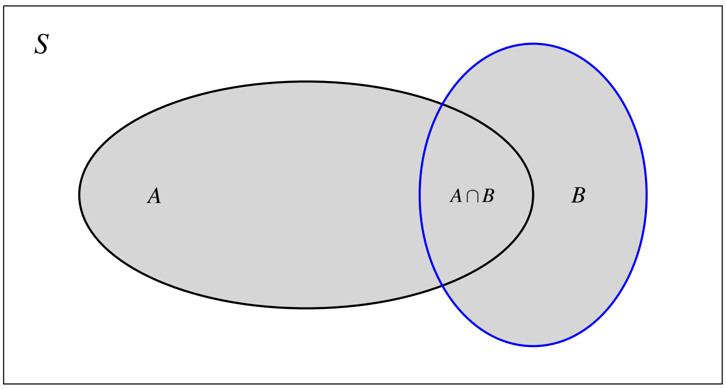
- Possible event relationships:

  (1). Contained: $A \subset B$; (2). Equivalent: $A = B$; (3) Mutually exclusive: $A \cap B = \emptyset$; (4) Independent: $A \perp B$ (postponed).

**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

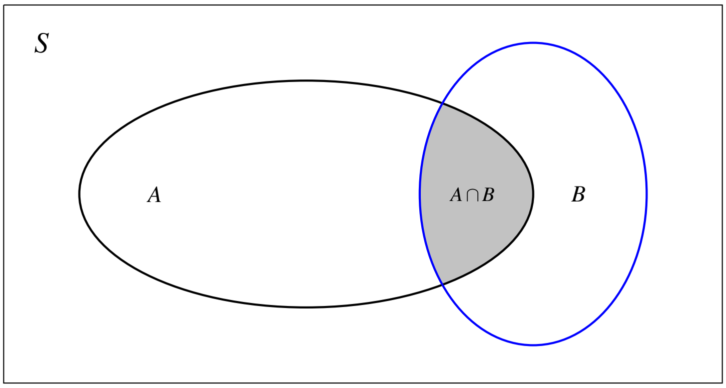
$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$



### 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

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




**Union** and **intersection** can be extended to $n$ events: $A_1, A_2, \ldots, A_n$.

- **Union**:

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

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

- **Intersection**:

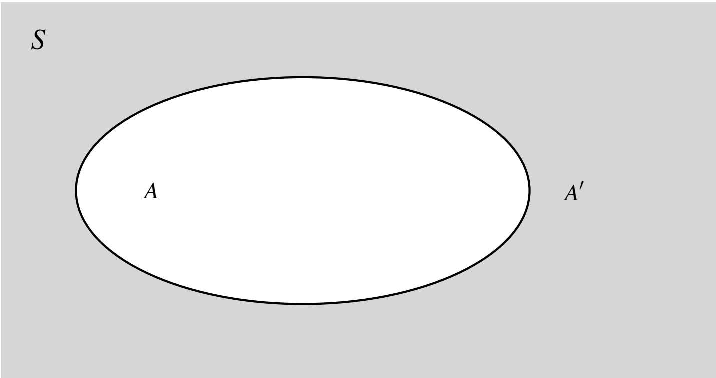
$$\bigcap_{i=1}^{n} A_i = A_1 \cap A_2 \ldots \cap A_n = \{x : x \in A_1 \text{ and } x \in A_2 \text{ and } \ldots \text{ and } x \in A_n\},$$

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

### Complement

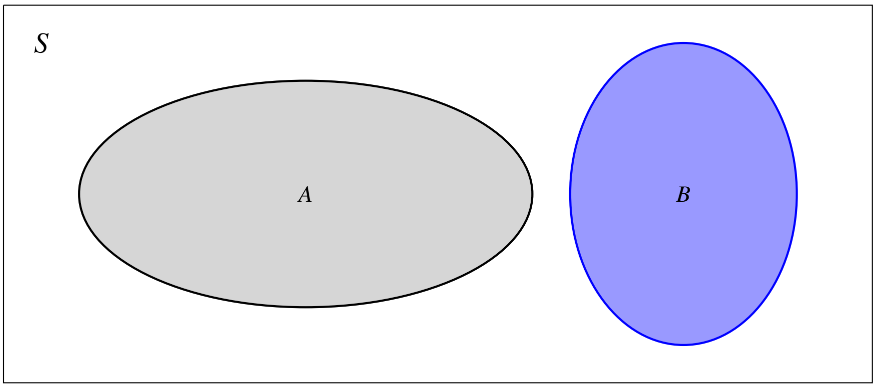
The **complement** of the event $A$ (with respect to $S$), denoted by $A'$, is the event with elements in $S$, which are not in $A$. That is

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



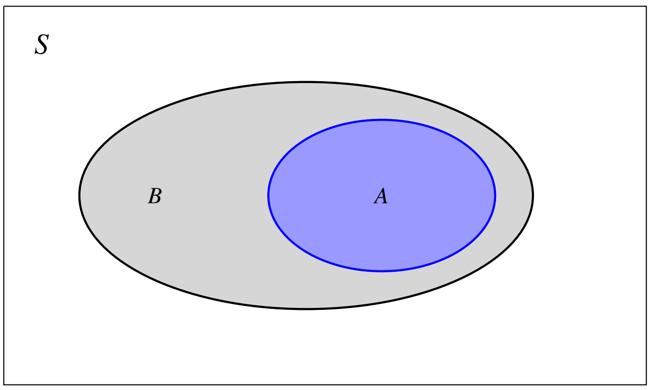

**Mutually Exclusive**

Events A and B are said to be mutually exclusive or disjoint, if $A \cap B = \emptyset$, i.e., $A$ 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 $B \supset A$).




$A \subset B$:
1) $A \subsetneq B$
2) $A = B$

- If $A \subset B$ and $B \subset A$, then $A = B$, i.e., set $A$ and $B$ are **equivalent**.

**EXAMPLE 1**
Consider the sample space and events: $S = \{1,2,3,4,5,6\}, A = \{1,2,3\}$, $B = \{1,3,5\}, C = \{2,4,6\}$. Then

- $A \cup B = \{1,2,3,5\}; \quad A \cup C = \{1,2,3,4,6\}; \quad B \cup C = S.$
- $A \cap B = \{1,3\}; \quad A \cap C = \{2\}; \quad B \cap C = \emptyset.$
- $A \cup B \cup C = S; \quad A \cap B \cap C = \emptyset.$
- $(A \cap B) \cup C = \{1,3\} \cup \{2,4,6\} = \{1,2,3,4,6\}.$
- $A' = \{4,5,6\}; \quad B' = \{2,4,6\} = C.$
- $B$ and $C$ are mutually exclusive, since $B \cap C = \emptyset$; $A$ and $B$ are not mutually exclusive since $A \cap B = \{1,3\} \neq \emptyset$.

## Some Basic Properties of Event Operations

(1). $A \cap A' = \emptyset$ (2). $A \cap \emptyset = \emptyset$

(3). $A \cup A' = S$ (4). $(A')' = A$

(5). $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(6). $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(7). $A \cup B = A \cup (B \cap A')$

(8). $A = (A \cap B) \cup (A \cap B')$

### De Morgan's Law

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

(9). $(A_1 \cup A_2 \cup \ldots \cup A_n)' = A_1' \cap A_2' \cap \ldots \cap A_n'$.
A special case: $(A \cup B)' = A' \cap B'$.

(10). $(A_1 \cap A_2 \cap \ldots \cap A_n)' = A_1' \cup A_2' \cup \ldots \cup A_n'$.
A special case: $(A \cap B)' = A' \cup B'$.

**EXAMPLE 2**
Adopt the setting of Example 1: $S = \{1,2,3,4,5,6\}$, $A = \{1,2,3\}$, $B = \{1,3,5\}$, $C = \{2,4,6\}$. We have

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

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

- Similarly, we can check $(A \cup B \cup C)' = \emptyset = A' \cap B' \cap C'$; and $(A \cap B \cap C)' = S = A' \cup B' \cup C'$.

**L–example 1.4** Revisit L–Example 1. Consider a two step experiment:

1. Flip a coin and observe whether the head $(H)$ or the tail $(T)$ is facing up.
2. If H is obtained in step 1, then flip it again; otherwise, roll a die once.

Then sample space:

$$S = \{(H,H),(H,T),(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\}$$

- Consider the events:

$$\begin{aligned} A &= \{\text{The die is rolled and the number is no more than 3}\} \\ &= \{(T,1),(T,2),(T,3)\} \\ B &= \{\text{The die is rolled and the number is even}\} \\ &= \{(T,2),(T,4),(T,6)\} \\ C &= \{\text{The die is not rolled}\} = \{(H,H),(H,T)\} \end{aligned}$$

- Then their complements are

$$\begin{aligned} A' &= \{(H,H),(H,T),(T,4),(T,5),(T,6)\} \\ B' &= \{(H,H),(H,T),(T,1),(T,3),(T,5)\} \\ C' &= \{(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\} \end{aligned}$$

- Some event operations:

$A \cup B = \{(T,1),(T,2),(T,3),(T,4),(T,6)\}$

$B \cup C = \{(T,2),(T,4),(T,6),(H,H),(H,T)\}$

$A \cap B = \{(T,2)\}$

$B \cap C = \emptyset$; so $B$ and $C$ are mutually exclusive

$A \cup B \cup C = \{(H,H),(H,T),(T,1),(T,2),(T,3),(T,4),(T,6)\}$

$A \cap B \cap C = \emptyset \qquad (A \cup B) \cap C = \emptyset$

$A' \cap B' = \{(H,H),(H,T),(T,5)\} = (A \cup B)'$

$A' \cup B' = \{(H,H),(H,T),(T,1),(T,3),(T,4),(T,5),(T,6)\} = (A \cap B)'$.

## 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 obtain some important counting methods: **permutation** and **combination**.

**Multiplication Principle**

Suppose that $r$ different experiments are to be performed sequentially.

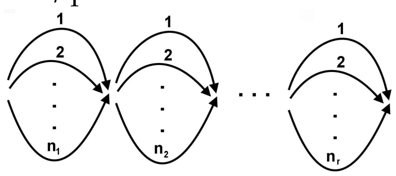
experiment 1 results in $n_1$ possible outcomes;

for each of the above result, experiment 2 results in $n_2$ possible outcomes;

$\cdots\ \cdots\ \cdots\ \cdots\ \cdots$

for each of the above result, experiment $r$ results in $n_r$ possible outcomes.

Together there are $n_1 n_2 \cdots n_r$ possible outcomes of the $r$ experiments.




**Example 1**

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

Solution:

- experiment 1: throwing a die; it has 6 possible outcomes: $\{1,2,3,4,5,6\}$.
- experiment 2: throwing a coin; for each outcome of experiment 1, it has 2 possible outcomes: $\{H, T\}$

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

In fact, the sample space is given by

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

**L–example 1.5** A small community consists of 10 men, each of whom has 3 sons. If one man and one of his sons are to be chosen as "father and son" of the year, how many different choices are possible?

Solution:

- experiment 1: choose the father; it has 10 possible choices.
- experiment 2: choose the son; for each of the father, there are 3 sons to choose from.

Together, there are $10 \times 3 = 30$ choices possible.

**L–example 1.6** How many **even three-digit numbers** can be formed from the digits 1,2,5,6, and 9 if each digit can be used **at most once**?

Solution: We can consider the whole task as 3 sequential experiments:

> experiment 1: choose the number for ones place; digits 2 and 6 can be used, so there are 2 possibilities.
>
> experiment 2: choose the number for tens place from digits left form experiment 1: 4 possibilities.
>
> experiment 3: choose the number for hundreds place from digits left from experiments 1 and 2: 3 possibilities.

Together, we have $2 \times 4 \times 3 = 24$ possibilities.

**L–example 1.7**

How many **even three-digit numbers** can be formed from the digits 1,2,5,6, and 9 if **no restriction** on how many times a digit is used?

Solution: Similar to the last L–Example:

> experiment 1: choose the number for ones place; digits 2 and 6 can be used, so there are 2 possibilities.
>
> experiment 2: choose the number for tens place from all digits provided: 5 possibilities.
>
> experiment 3: choose the number for hundreds place from all digits provided: 5 possibilities.

Together, we have $2 \times 5 \times 5 = 50$ possibilities.

**L–example 1.8**

In how many ways can 4 boys and 5 girls sit in a row if the boys and girls must alternate?

Solution: We must have the arrangement: G B G B G B G B G
The number of ways:

$$5(4)(4)(3)(3)(2)(2)(1)(1) = 5!4! = 2880,$$

where $n! = n(n-1)(n-2)\cdots(2)(1)$.

Question: What happens if there are 5 boys and 5 girls?

this will create 2 disjoint events

GBGBGBGBGB or BGBGBGBGBG

this is unlike the previous example where got more girls, so the girls must start the "queue" first

**Additional 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.

$\ldots \ldots \ldots \ldots \ldots$

Procedure $k$ can be carried out in $n_k$ ways.

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

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

**EXAMPLE 2**

We can take MRT or bus from home to Orchard road. If there are three bus routes and two MRT routes. How many ways we can go from home to Orchard road?

Solution: Consider that we go 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.

The ways are not overlapped. So the total number of ways that we can go from home to Orchard road is $2 + 3 = 5$.

**L–example 1.9** How many **even three-digit numbers** can be formed from the digits 0, 1, 2, 5, 6, and 9 if each digit can be used **at most once**?

Solution: We consider the whole task as two procedures based on ones place:

Procedure 1: 0 is used for the ones place. There are $5 \times 4 = 20$ ways to arrange the hundreds and tens place.

Procedure 2: 0 is not used for the ones place. (1) There are two ways (2 or 6) to fill in ones place; (2) as 0 can not be put at the hundreds place, we have 4 digits available for the hundreds place; (3) finally, we have 4 possible choices for the tens place. In summary, we have $2 \times 4 \times 4 = 32$ ways.

With the addition rule, we combine Procedures 1 and 2 to conclude that there are $20 + 32 = 52$ ways.

**L–example 1.10** Consider the digits 0,1,2,3,4,5, and 6. If each digit can be used **at most once**, how many 3-digit numbers, which are greater than 420, can be formed?

Solution: consider three procedures:

Procedure 1: the hundreds place is 4 and the tens place is 2: $(1)(1)(4) =$ 4 ways.

Procedure 2: the hundreds place is 4 and the tens place is 3,5,or 6: $(1)(3)(5) = 15$ ways.

Procedure 3: the hundreds place is 5 or 6: $(2)(6)(5) = 60$ ways

In total, $4+15+60 = 79$ ways.

**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 $n$ ($r \leq n$) is denoted by $P_r^n$, which has the value:

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

| ob1 | ob2 | ob3 | $\cdots$ | ob$r$ |
|---|---|---|---|---|
| $n$ ways | $(n-1)$ ways | $(n-2)$ ways | $\cdots$ | $(n-(r-1))$ ways |

- A special case: when $r = n$, $P_n^n = n!$. Practically, it is the number of ways to arrange $n$ objects in order.

  EXAMPLE 3
  Find the number of all possible four-letter code words in which all letters are different.

  Solution: $n = 26$, $r = 4$. So the number of all possible four-letter code words is

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

**L–example 1.11**

(a) How many ways can 6 persons line up to get on a bus?

(b) If certain 3 persons insisting on following each other, how many ways can these 6 persons line up?

(c) If 2 persons refuse to follow each other, how many ways of lining up are possible?

Solution:

(a) $n = r = 6$, so $P_6^6 = 6! = 720$ ways.

(b) Let $a, b, c, d, e, f$ be the names of 6 persons.

- Without loss of generality, assume that $a, b, c$ insist on following each other.
- Group them into one group, denoted by $A = \{a, b, c\}$, which can now be viewed as one single person.
- We need to line up four persons, i.e., $A, d, e, f$, in a row. So $P_4^4 = 4! = 24$ ways.

On the other hand, for each permutation above, such as $(A, d, f, e)$, $a, b, c$ within $A$ can be ordered differently. The number of ways of ordering them within $A$ is $P_3^3 = 3! = 6$.

Therefore, applying the multiplication rule, the number of ways to line them up is $24 \times 6 = 144$.

(c) With the same principle as Part (b), we can first count the number of ways of lining up if they two persons are following each other:

$$P_5^5 \times P_2^2 = 5! \times 2! = 240.$$

From Part (a), the total number of ways for lining up 6 persons is 720. Therefore, we have $720 - 240 = 480$ ways of lining up 6 persons such that two given persons are not following each other.

**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

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

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

- The derivation is as follows:

  - Based on permutation, **the number of ways to choose and arrange $r$ objects out** $n$ is $P_r^n$.
  - On the other hand, this permutation task can be achieved by conducting two experiments sequentially:

  Exp. 1 get a combination, i.e., select $r$ objects out $n$ without regard to the order; there are $\binom{n}{r}$ ways.

Exp. 2 for each combination, get a permutation on its $r$ objects; there are $P_r^r$ ways.

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

– As a consequent, $\binom{n}{r} \times P_r^r = P_r^n$, and we obtain

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

**EXAMPLE 4**
From 4 women and 3 men, find the number of committees of 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 $\binom{4}{2} \times \binom{3}{1} = 6 \times 3 = 18$.

**L–example 1.12** From a group of 4 men and 5 women, how many committees of size 3 are possible

(a) with no restriction?

(b) with 2 men and 1 woman if a certain man must be on the committee?

(c) with 2 men and 1 woman if 2 of the men are feuding and refuse to serve on the committee together?

Solution:

(a) The number of committee is $\binom{9}{3} = 9!/(3!6!) = 84$.

(b) Since a certain man must be on the committee, we only need to choose one man from the remaining 3 men and 1 woman from 5 woman: $\binom{1}{1}\binom{3}{1}\binom{5}{1} = 15$.

(c) The number of committees such that both "particular" men serve is $\binom{2}{2}\binom{5}{1} = 5$. As this includes all the "undesirable" cases, the number of "desirable" cases are $\binom{4}{2}\binom{5}{1} - 5 = 30 - 5 = 25$.

**L–example 1.13**

Shortly after being put into service, some buses manufactured by a certain company have developed cracks on the underside of the main frame.
Suppose a particular bus company has 20 of these buses, and the cracks have actually appeared in 8 of them.

(a) How many ways are there to select a sample of 5 buses from the 20 for a thorough inspection?

(b) In how many ways can a sample of 5 buses contain exactly 4 buses with visible cracks?

Solution:

(a) It is the number of ways for selecting 5 buses out of 20: $\binom{20}{5} = 20!/(5!15!) = 15504$.

(b) Now 4 buses need to be selected from the 8 buses with visible cracks: $\binom{8}{4} = 70$; then 1 bus need to be selected from the remaining 12 buses: $\binom{12}{1} = 12$. Applying the multiplication rule, the number of ways that satisfy the condition is $70 \times 12 = 840$.

## 4 Probability

- Intuitively, "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. We denote "$P(A)$".
- But..., how could we obtain such a number?
- Even more, as a fundamental concept, it has been extended from the intuition to more rigorous, abstract, and advanced mathematical theory and has wide applications in scientific disciplines.

### Interpretation of Probability by Relative Frequency

- Suppose that we repeat an experiment $E$ for $n$ 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 event $A$ in the $n$ repetition of $E$.
- Clearly, $f_A$ may not equal to $P(A)$ exactly. But when $n$ grows large, we may expect that $f_A$ may "approach" it; in a sense $f_A \approx P(A)$. Or more mathematically,
$$f_A \to P(A), \quad \text{as } n \to \infty.$$
- Therefore $f_A$ "mimics" $P(A)$, and it has the following properties:

  (1) $0 \le f_A \le 1$;

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

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

- Extending this, we 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:

(i) For any event $A$,
$$0 \le P(A) \le 1.$$

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

(iii) For any two mutually exclusive events $A$ and $B$, i.e., $A \cap B = \emptyset$,
$$P(A \cup B) = P(A) + P(B).$$

**EXAMPLE 1**

Let $H$ denote the event of getting head when tossing a coin. Find $P(H)$, if

(a) the coin is fair;

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

Solution:

(a) The sample space is $S = \{H, T\}$.

- "Fair" means $P(H) = P(T)$.
- The events $\{H\}$ and $\{T\}$ are mutually exclusive.
- Based on Axioms 2 and 3, we have

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

which implies $P(H) = 1/2$.

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

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

which leads to $P(T) = 1/3$ and $P(H) = 2/3$.

**L–example 1.14** A fair die is tossed. Let

$$A = \{\text{an even number turns up}\}$$

$$B = \{\text{1 or 3 turns up}\}$$

$$C = \{\text{a number divisible by 3 is obtained}\}$$

Find $P(A), P(B), P(C), P(A \cup B)$, and $P(A \cup C)$.
Solution: $A = \{2,4,6\}$, $B = \{1,3\}$, $C = \{3,6\}$. We have

- $P(A) = 3/6 = 1/2$; $P(B) = P(C) = 1/3$.
- Since $A \cap B = \emptyset$, based on Axiom 3, we have $P(A \cup B) = P(A) + P(B) = 5/6$.
- But $A \cap C = \{6\} \neq \emptyset$; Axiom 3 is not applicable. Instead, $A \cup C = \{2,3,4,6\}$ leads to $P(A \cup C) = 4/6 = 2/3$.

### Basic Properties of Probability

Using the axioms, we can derive the following propositions.

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

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

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

which implies $P(\emptyset) = 0$.

**Proposition 3**
*If $A_1, A_2, \ldots, A_n$ are mutually exclusive events ($A_i \cap A_j = \emptyset$ 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 is immediately observed based on the induction.

**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 property (8) of event operations, i.e., $A = (A \cap B) \cup (A \cap B')$, and $(A \cap B) \cap (A \cap B') = \emptyset$, 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 property (7) of event operations, i.e., $A \cup B = B \cup (A \cap B')$, and $B \cap (A \cap B') = \emptyset$; and proposition 5 that $P(A \cap B') = P(A) - P(A \cap B)$, 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 $A \cup B = B$; based on property (7) of the event operations, i.e, $A \cup B = A \cup (B \cap A')$; and based on $A \cap (B \cap A') = A \cap B \cap A' = \emptyset$, we have

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

**EXAMPLE 8**

- A retail establishment accepts either the American Express or the VISA credit card.
- A total of 24% of it customer 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 $A = \{$the customer carries an American Express Card$\}$; $V = \{$the customer carries an VISA Card$\}$.
- Then $P(A) = 0.24$; $P(V) = 0.61$; and $P(A \cap V) = 0.11$.
- The question is asking $P(A \cup V)$, and

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

**L–example 1.15** [Hall Pageant]

Audrey is taking part in her hall's pageant. The probability that she will **win the crown** is 0.14; the probability that she will **win Miss Photogenic** is 0.3; the probability that she will **win both** is 0.11.

(a) What is the probability that she wins at least one of two?

(b) What is the probability that she wins only one of two?

Solution: Let $A = \{$win the crown$\}$ and $B = \{$win Miss Photogenic$\}$.

(a) The probability that she wins at least one of the two titles

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.14 + 0.3 - 0.11 = 0.33.$$

(b) The event that she wins the crown but not the Miss Photogenic is $A \cap B'$; based on Proposition 5 $P(A \cap B') = P(A) - P(A \cap B) = 0.14 - 0.11 = 0.03$. Similarly the event that she wins the Miss Photogenic but not the crown is $B \cap A'$, and $P(B \cap A') = 0.3 - 0.11 = 0.19$. We have

$$P((A \cap B') \cup (B \cap A')) = 0.03 + 0.19 = 0.22,$$

since $(A \cap B') \cap (B \cap A') = \emptyset$.

**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.,
$$P(a_1) = P(a_2) = \ldots = P(a_k).$$
- 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 9**

- 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:

- Let $A = \{\text{the item is rusted}\}$, $B = \{\text{the item is a bolt}\}$, $S = \{\text{all the items}\}$.
- Since the item is selected at random, the elements in $S$ are equally likely. $S$ contains 200 elements. $A$ contain $25 + 75 = 100$ elements, $B$ contains 50, and $A \cap B$ contains 25.
- $P(A) = 100/200$, $P(B) = 50/200$, $P(A \cap B) = 25/200$;
$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 5/8.$$

**L–example 1.16** [Birthday Problem]

- Here's a useful party trick: walk into a room or bar with at least 50 people.
- Boldly claim that you sense two people sharing the same birthday. Act awesome afterwards.

- How often are you right?

Solution: We can cast this as a probability question:

> There are $n$ people in a room, what is the probability that there are at least two people with the same birthday?

Some assumptions:

- Each day is **equally likely** to be a birthday of everyone.
- there is no leap year.

We can then have the following:

- The sample space is

$$S = \{\text{all possible combinations of birthdays of } n \text{ people}\}.$$

  It is formed of equally likely sample points.

- Let

$$A = \{\text{at least two people share the same birthday}\},$$

  then

$$A' = \{\text{all people have different birthdays}\}.$$

- We count the number of sample points in $S$ and $A'$:
  - We call $n$ people as $p_1, p_2, \cdots, p_n$, and consider the number of choices of birthdays for them one by one.
  - If all people have different birthdays,
    * for $p_1$, it has 365 possibilities as his/her birthday;
    * for $p_2$, $365 - 1$;
    * $\cdots\cdots\cdots$;
    * for $p_n$, $365 - (n - 1)$.

    As a consequence, $\#(A') = 365(364)\cdots[365 - (n-1)]$.
  - Similarly $\#S = 365^n$.

- Therefore

$$P(A') = \frac{\#(A')}{\#S} = \frac{365(364)\cdots[365-(n-1)]}{365^n},$$

  and hence

$$P(A) = 1 - P(A') = 1 - \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{n-1}{365}\right)$$

Let $q_n = P(A')$ when there are $n$ people, and

$$p_n = P(A) = 1 - q_n.$$

The values of $p_n$ and $q_n$ for selected values of $n$ are tabulated:

| n | $q_n$ | $p_n$ |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0.99726 | 0.00274 |
| 3 | 0.99180 | 0.00820 |
| 10 | 0.88305 | 0.11695 |
| 15 | 0.74710 | 0.25290 |
| 20 | 0.58856 | 0.41144 |
| 21 | 0.55631 | 0.44369 |
| 22 | 0.52430 | 0.47570 |
| 23 | 0.49270 | 0.50730 |
| 30 | 0.29368 | 0.70632 |
| 40 | 0.10877 | 0.89123 |
| 50 | 0.029626 | 0.979374 |
| 100 | $3.0725 \times 10^{-7}$ | $\approx 1$ |
| 253 | $6.9854 \times 10^{-53}$ | $\approx 1$ |

- For 50 people, 98% of the time you can find at least two people with the same birthday.
- The probability of having at least two people sharing the same birthday exceeds 1/2 once you have 23 people.
- When there are 100 people, almost for sure you can find two people sharing the same birthday!

**L–example 1.17** [Inverse Birthday Problem]

How large does a group of (randomly selected) people have to be such that the probability that someone shares his or her birthday **with you** is larger that 0.5?

Solution: The probability that $n$ persons all have different birthdays from you is $\left(\frac{364}{365}\right)^n$.

So we need $n$ such that $1 - (364/365)^n \geq 0.5$. Solving it, we obtain

$$n \geq \frac{\log(0.5)}{\log(364/365)} = 252.7.$$

We need at least 253 people (excluding youself).

**REMARK (BIRTHDAY PROBLEMS):**

Why there is a big difference in the answers between the two birthday problems?

- The inverse birthday problem requires the sharing of a **particular** day as the common birthday;
- The birthday problem allows that **any** day is the shared birthday. ■

## 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 1 (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 2**
A fair die is rolled twice.

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

(b) 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:

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

(a) The sample space is given by

| | | 2nd roll | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| 1st roll | 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) |
| | 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 $P(B) = 18/36$.

(b) 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 instances among 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):**

- $P(B|A)$ can also be 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 $A$.

**L–example 1.18**

- Suppose two fair dice are rolled.
- Given that the first die is less than 3, what is the probability that the sum of the 2 dice is more than 7?

Solution: Define the events:

$$\begin{aligned} B &= \{\text{the sum of the 2 dice is more than 7}\} \\ A &= \{\text{the first die is less than 3}\} \end{aligned}$$

Consider the reduced sample space, i.e., event $A$, with the following 12 equally likely sample points:

$$\{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6)\}$$

The required probability is $P(B|A) = 1/12$ since there is only one point $(2,6)$ in the reduced sample space that gives a sum more than 7.

**Multiplication Rule**

Rearranging the definition of the conditional probability, 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 together with the definition of the conditional probability, we have the inverse probability formula:

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

**EXAMPLE 3**

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

Solution:

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

**L–example 1.19**

Suppose that among 12 shirts, 3 are white. Two shirts are chosen randomly one by one **without replacement**

(a) What is the probability that **both shirts** that being picked are white?

(b) What is the probability that there is **only one** white shirt being picked?

Solution: Define the events:

$$\begin{aligned} A_1 &= \{\text{the first shirt is white}\} \\ A_2 &= \{\text{the second shirt is white}\}. \end{aligned}$$

(a) We have $P(A_1) = 3/12$; given that the first shirt is white, then there are 2 white shirts among the remaining 11 shirts, therefore

$$P(A_2|A_1) = 2/11.$$

$P(A_1 \cap A_2) = P(A_1)P(A_2|A_1) = (3/12)(2/11) = 1/22.$

(b) we need to compute the probability for $(A_1 \cap A_2') \cup (A_1' \cap A_2)$. We note that $(A_1 \cap A_2') \cap (A_1' \cap A_2) = \emptyset$.

$$P((A_1 \cap A_2') \cup (A_1' \cap A_2)) = P(A_1 \cap A_2') + P(A_1' \cap A_2).$$

On the other hand, with similar argument as Part (a), we have

$$\begin{aligned} P(A_1 \cap A_2') &= P(A_1)P(A_2'|A_1) = (3/12) \cdot (9/11) \\ P(A_1' \cap A_2) &= P(A_1')P(A_2|A_1') = (9/12) \cdot (3/11). \end{aligned}$$

As a consequence

$$P((A_1 \cap A_2') \cup (A_1' \cap A_2)) = (3/12) \cdot (9/11) + (9/12) \cdot (3/11) = 9/22.$$

## 6 Independence

Independence is one of the most important concepts in probability.

**Definition 1 (Independence)**
*Two events $A$ and $B$ are* ***independent*** *if and only if*

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

*We denote it as $A \perp B$.*

If $A$ and $B$ are not independent, they are **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 is observed from the definition of conditional probability:

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

- Intuitively, this is stating: $A$ and $B$ 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 2**
Suppose we roll 2 fair dice.

to check dependence must use the mathematical definition



(a) Let

$$\begin{aligned} A_6 &= \{\text{the sum of two dice is } 6\} \\ B &= \{\text{the first die equals } 4\}. \end{aligned}$$

Hence, $P(A_6) = 5/36$, $P(B) = 6/36 = 1/6$ and $P(A_6 \cap B) = 1/36$. Since

$$P(A_6 \cap B) \neq P(A_6)P(B),$$

we say that $A_6$ and $B$ are **dependent**.

(b) Let $A_7 = \{\text{the sum of two dice is } 7\}$. Then $P(A_7 \cap B) = 1/36$, $P(A_7) = 1/6$ and $P(B) = 1/6$. Hence

$$P(A_7 \cap B) = P(A_7)P(B),$$

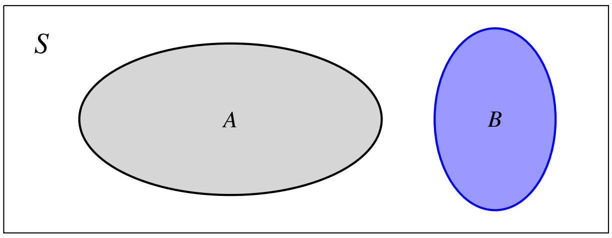
and we say that $A_7$ and $B$ are **independent**.

**REMARK (INDEPENDENT VS MUTUALLY EXCLUSIVE):**

**Independent** and **mutually exclusive** are totally different concepts:

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

"Mutually exclusive" can be illustrated by the Venn diagram, but "independent" can not.




■

**L–example 1.20** [Some properties of independence]

Decide whether the following statements are TRUE or FALSE.

(a) Suppose $P(A) > 0$, $P(B) > 0$, then if $A \perp B$, then $A$ and $B$ are not mutually exclusive.

(b) Suppose $P(A) > 0$, $P(B) > 0$, then if $A$ and $B$ are mutually exclusive, then $A \not\perp B$.

(c) $S$ and $\emptyset$ are independent of any event.

(d) If $A \perp B$, then $A \perp B'$, $A' \perp B$, and $A' \perp B'$.

Solution:

(a) TRUE; based on independence $P(A \cap B) = P(A)P(B) > 0$.

(b) TRUE; based on mutually exclusive $P(A \cap B) = 0 \neq P(A)P(B)$.

(c) TRUE; for any event $A$; $P(A \cap S) = P(A) = P(A)P(S)$; $P(A \cap \emptyset) = P(\emptyset) = 0 = P(A)P(\emptyset)$.

(d) TRUE; we derive one only. Note that $A = (A \cap B) \cup (A \cap B')$, we have

$$\begin{aligned} P(A \cap B') &= P(A) - P(A \cap B) = P(A) - P(A)P(B) \\ &= P(A)(1 - P(B)) = P(A)P(B'). \end{aligned}$$

**L–example 1.21**

- The probability that Tom will be alive in 20 years is 0.7.
- The probability that Jack will be alive in 20 years is 0.9.
- What is the probability that neither will be alive in 20 years?
- Define

$$\begin{aligned} A &= \{\text{Tom would be alive in 20 years.}\} \\ B &= \{\text{Jack would be alive in 20 years.}\} \end{aligned}$$

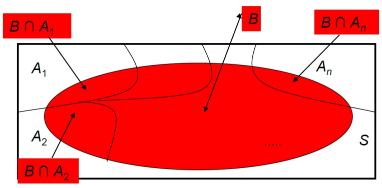
- $A$ and $B$ are independent, as whether one is alive would not affect the other.
- $A'$ and $B'$ are also independent.
- The desired probability is given by

$$P(A' \cap B') = P(A')P(B') = (1 - 0.7)(1 - 0.9) = 0.03.$$

## 7 The Law of Total Probability

**Definition 1 (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 $S$.*

in this case here it becomes a partition for B as well




$B = (A_1 \cap B) \cup (A_2 \cap B) \ldots$
$\ldots \cup (A_n \cap B)$

$\emptyset = (A_i \cap B) \cup (A_j \cap B)$
$i \neq j$

**Theorem 2 (The law of total probability)**

*If $A_1, A_2, \ldots, A_n$ is a partition of $S$, then for any event $B$, we have*

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

**A special case**: For any events $A$ and $B$, we have

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

**Example 3 (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 $P(B)$.
- Use the law of total probability

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

**L–example 1.22** [The Monty Hall Problem]

- Suppose you are on a game show, and you are given the choice of three doors: behind one door is a car; behind the others, goats.
- You pick a door, say No. 1, and the host Monty, who knows what is behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?"
- Is it to your advantage to switch your choice?

Solution: Let's formulate it as a probability question. Denote the events:

$$\begin{aligned} W &= \{\text{Win the car}\} \\ A &= \{\text{car is behind the door of the initial pick}\} \end{aligned}$$

Then, our interest is $P(W)$, i.e., the probability of wining the car.

Applying the law of total probability, we have

$$\begin{aligned} P(W) &= P(A)P(W|A) + P(A')P(W|A') \\ &= \frac{1}{3}P(W|A) + \frac{2}{3}P(W|A'). \end{aligned}$$

- If a "stick" strategy is used, i.e., not to switch your choice,
  - $P(W|A) = 1$; that is, you are sure to win the car if the initial choice is the car door;
  - $P(W|A') = 0$; that is, you are sure to lose the car if the initial choice is not the car door.

  As a consequence $P(W) = (1/3) \cdot 1 + (2/3) \cdot 0 = 1/3$.
- If a "switch" strategy is used, i.e., switch to another door when asked, then $P(W|A) = 0$ and $P(W|A') = 1$. We then have $P(W) = (1/3) \cdot 0 + (2/3) \cdot 1 = 2/3$.

Conclusion: "Switch" double the chance of wining the car!

**REMARK (MONTY HALL):**

Still confused? Watch the following videos:

http://www.youtube.com/watch?v=mhlc7peGlGg

http://www.youtube.com/watch?v=P9WFKmLK0dc ■

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

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

- A special case of Bayes' theorem: take $n = 2$. $A$ and $A'$ become a partition of $S$. We have

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

- This formula is practically meaningful. For example:
  - $A$ = disease status of a person;
  - $B$ = symptom observed;
  - $P(A)$: the probability of a disease in general;
  - $P(B|A)$: if diseased, probability of observing symptom;
  - $P(A|B)$: if symptom observed, probability of diseased.
- Bayes' theorem can be derived based on the conditional probability, multiplication rule, and the law of the total probability.
- In particular,

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

**EXAMPLE 2**
- Historically we observe some newly constructed house to collapse.
- The chance that the design is faulty is 1%.
- If the design is faulty, the chance that the house is to collapse is 75%, otherwise, the chance is 0.01%.
- If we observed that a newly constructed house collapsed, what is the probability that the design is faulty?

Solution:

- Let

$$\begin{aligned} B &= \{\text{The design is faulty}\}, \\ A &= \{\text{The house collapses}\}. \end{aligned}$$

- We have $P(B) = 0.01$, $P(A|B) = 0.75$, and $P(A|B') = 0.0001$.
- The question is asking to compute $P(B|A)$.
- We compute it based on Bayes' theorem. The denominator can be computed based on the law of total probability:

$$\begin{aligned} P(A) &= P(B)P(A|B) + P(B')P(A|B') \\ &= (0.01)(0.75) + (0.99)(0.00001) = 0.007599. \end{aligned}$$

- The numerator is

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

- As a consequence $P(B|A) = P(A|B)P(B)/P(A) = 0.9870$.

**L–example 1.23**

- An insurance company believes that people can be divided into two classes: **accident prone** and **not accident prone**.
- Historically, they observe that the probability that an accident-prone person will have an accident within a fixed 1-year period is 0.04; otherwise, the probability is 0.02.
- Assume that 30% of the population is accident prone.

(a) What is the probability that a new policyholder will have an accident within a year of purchasing a policy?

(b) If a new policyholder has an accident within a year of purchasing a policy, what is the probability that he or she is accident prone?

Solution:

- Define the events:

$$\begin{aligned} B &= \{\text{a new policy holder has an accident within a year}\} \\ A &= \{\text{a new policy holder is accident prone}\}. \end{aligned}$$

- Based on the question, $P(A) = 0.3$; $P(B|A) = 0.04$; $P(B|A') = 0.02$.

(a) $P(B) = P(A)P(B|A) + P(A')P(B|A') = 0.3(0.04) + 0.7(0.02) = 0.26$.

(b) $P(A|B) = P(A)P(B|A)/P(B) = 0.3(0.04)/0.26 = 6/13$.

# Chapter 2: Random Variables

## 1 Definition of Random Variable

- When an experiment is performed, we are often interested in some numerical characteristic of the result.
- Here are some practical examples.
  - An experiment is to examine 100 electronic components, our interest is "the number of defectives".
  - Flipping a coin 100 times, our interest could be the number of heads obtained, instead of the "H" and "T" sequence.
- This motivates us to assign a numerical value to a possible outcome of an experiment.

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

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

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

- For convenience, hereafter, we simplify "**random variable**" as "**RV**".

**Example 2**
- Let $S = \{HH, HT, TH, TT\}$ be a sample space associated with the experiment of flipping two coins.
- Define the RV:

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

- 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 $R_X = \{0, 1, 2\}$.

**L–example 2.1**

- A coin is thrown until a "head" occurs.

$$S = \{H, TH, TTH, TTTH, TTTTH, \cdots\}$$

- Let $X =$ the number of "trials" required. We then have

$$X(H) = 1,\ X(TH) = 2,\ X(TTH) = 3,\ \cdots,\ \text{and so on.}$$

- $R_X = \{1, 2, 3, \ldots, \}$

**REMARK:**

- We use upper case letters $X, Y, Z, X_1, X_2, \ldots$ to denote **random variables**.
- We use lower case letters $x, y, z, x_1, x_2$ to denote their **observed values** in the experiment.
- The set $\{X = x\}$ is a subset of $S$, in the sense:

$$\{X = x\} = \{s \in S : X(s) = x\}.$$

- Likewise, the set $\{X \in A\}$, for $A$ being a subset of $\mathbb{R}$, is also a subset of $S$:

$$\{s \in S : X(s) \in A\}.$$

- This gives $P(X = x)$ and $P(X \in A)$ based on probability defined on $S$:

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

**EXAMPLE 3**

- Revisit Example 2; $S = \{HH, HT, TH, TT\}$ is the sample space of flipping two coins. $X =$ number of heads obtained.
- Then $\{X = 0\} = \{TT\}$; $\{X = 1\} = \{HT, TH\}$; $\{X = 2\} = \{HH\}$; $\{X \geq 1\} = \{HT, TH, HH\}$.

- $P(X=0) = P(TT) = 1/4$; $P(X=1) = P(\{HT, TH\}) = 2/4$; $P(X = 2) = P(HH) = 1/4$; $P(X \geq 1) = P(\{HT, TH, HH\}) = 3/4$.
- We can summarize the probabilities of the RV $X$ as a table:

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

**L–example 2.2**

- When a pair of fair dice is rolled, what is the probability that a sum of 3 is obtained?

$$S = \Big\{(x_1, x_2)\Big| x_1 = 1,2,3,4,5,6;\ x_2 = 1,2,3,4,5,6\Big\}.$$

- $X =$ the sum of two dice. That is for any $(x_1, x_2) \in S$,

$$X((x_1, x_2)) = x_1 + x_2.$$

- The range of $X$ is
$$R_X = \{2, 3, 4, \ldots, 12\}.$$

- Since $\{X = 3\} = \{(1,2), (2,1)\}$, we have

$$P(X = 3) \quad = \quad P(\{(1,2), (2,1)\}) = 2/36.$$

- The probabilities of other possible values for $X$ can be found similarly, and are tabulated below:

| $x$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $P(X=x)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

## 2 Probability Distributions

- We introduce the distribution of two types of RVs: **discrete** and **continuous**.
- In particular, denote by $X$ the RV, and its range by $R_X$.
  - **Discrete**: 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**: $R_X$ is an **interval** or a **collection of intervals**.

**Discrete Probability Distributions**

- For a discrete RV $X$, we can always write $R_X = \{x_1, x_2, x_3, \ldots\}$.
- Each $x_i \in R_X$, there is a probability that $X$ takes this value, i.e., $P(X = x_i)$.
- We can define a function $f(x) = P(X = x)$.
  Note that $f(x_i) = P(X = x_i)$ for $x_i \in R_X$, and $f(x) = 0$ for $x \notin R_X$.
- $f(x)$ is called the **probability function**, **p.f.** (or **probability mass function, p.m.f.**) of $X$.
- The collection of pairs $(x_i, f(x_i)), i = 1, 2, 3, \ldots,$ is called the **probability distribution** of $X$.

The p.f. $f(x)$ of a discrete RV **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 1**

- Revisit Examples 2 and 3. RV $X$ is the number of heads when flipping two coins.
- The p.f. of $X$ is given below

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

- $f(x)$ satisfies (1) $f(x_i) \geq 0$ for $x_i = 0, 1,$ or 2; (2) $f(x) = 0$ for other $x$; (3) $f(0) + f(1) + f(2) = 1$.
- $B = [1, \infty)$; then $P(X \in B) = f(1) + f(2) = 3/4$.

**L–example 2.3**

Six lots of components are ready to be shipped by a certain supplier. The number of defective components in each lot is as follows:

| Lot | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| # of defectives | 0 | 2 | 0 | 1 | 2 | 0 |

- One of the lots is to be **randomly** selected and shipped to a customer.
- Let $X =$ # of defectives in the shipped lot.
- Then $R_X = \{0, 1, 2\}$.
- The lots are selected randomly, so each has the same probability to be chosen.
- Let $f(x)$ be the p.f. of $X$.
- We have
  - $f(0) = P(X = 0) = P(\text{lot 1 or 3 or 6 is selected}) = 3/6$.
  - $f(1) = P(X = 1) = P(\text{lot 4 is selected}) = 1/6$.
  - $f(2) = P(X = 2) = P(\text{lot 2 or 5 is selected}) = 2/6$.
- The probability function of $X$ can be summarized by

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $f(x)$ | $1/2$ | $1/6$ | $1/3$ |

- It satisfies all the properties of probability functions.
- If $B = \{0, 2\}$, $P(X \in B) = f(0) + f(2) = 1/2 + 1/3 = 5/6$.

**L–example 2.4**

(a) Find the constant $c$, such that

$$f(x) = cx, \quad \text{for } x = 1, 2, 3, 4,$$

and 0 otherwise, is a probability function of a random variable $X$.

(b) Compute $P(X \geq 3)$.

<u>Solution</u>:

(a) Based on the property $\sum_{i=1}^{\infty} f(x_i) = 1$, we have

$$f(x_1) + f(x_2) + f(x_3) + f(x_4) = 1,$$

which is

$$c + 2c + 3c + 4c = 1.$$

Therefore $c = 1/10$.

(b) $P(X \geq 3) = f(3) + f(4) = 3/10 + 4/10 = 7/10.$

**L–example 2.5**

- Consider a group of five potential blood donors: A, B, C, D and E, of whom only A and B have type O+ blood.
- Five blood samples, one from each individual , will be typed in random order until an O+ individual is identified.

Solution:

- Let $Y = \#$ of typing needed to identify an O+ individual.
- Let $O_i$ and $O_i'$ be the events that an O+ and a non-O+ individual is typed in the $i$th typing

$$\begin{aligned} f(1) &= P(Y=1) = P(O_1) = 2/5 = 0.4, \\ f(2) &= P(Y=2) = P(O_1' \cap O_2) = P(O_1')P(O_2|O_1') \\ &= \frac{3}{5} \cdot \frac{2}{4} = 0.3, \end{aligned}$$

$$\begin{aligned} f(3) &= P(O_1')P(O_2'|O_1')P(O_3|O_1' \cap O_2') \\ &= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = 0.2, \\ f(4) &= P(Y=4) \\ &= P(O_1')P(O_2'|O_1')P(O_3'|O_1' \cap O_2')P(O_4|O_1' \cap O_2' \cap O_3') \\ &= \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{2}{2} = 0.1, \end{aligned}$$

and $f(y) = 0$ if $y \neq 1,2,3,4$.

- Then the probability function of $Y$ is

| $y$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $f(y)$ | 0.4 | 0.3 | 0.2 | 0.1 |

**Continuous Probability Distributions**

- For a continuous RV $X$, $R_X$ is an interval or a collection of intervals.
- For any $x \in \mathbb{R}$, we must have $P(X = x) = 0$.
- The **probability function, p.f.**, (or **probability density function, p.d.f.**) is defined to quantify the probability that $X$ is in a certain range.

The **p.d.f.** of a continuous RV $X$, 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) $\displaystyle\int_{R_X} f(x)dx = 1.$

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

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

Note: (2) is equivalent to $\displaystyle\int_{-\infty}^{\infty} f(x)dx = 1$, since $f(x) = 0$ for $x \notin R_X$.

**Remark:**

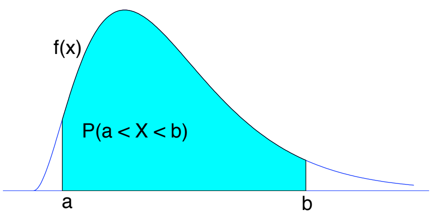
- For any arbitrary specific value $x_0$, we have

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

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

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

- 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 p.d.f., it suffices to check (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) $\int_{R_X} f(x)dx = 1$.

**EXAMPLE 2**

Let $X$ be a continuous RV with p.d.f. given by

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

(a) Find the value of $c$;

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

Solution:

(a) Since

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^1 cxdx = c \cdot \left. \frac{x^2}{2} \right|_0^1 = c/2,$$

we set $c/2 = 1$, and result in $c = 2$.

(b)

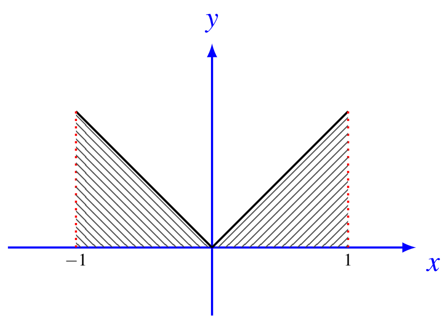
$$P(X \leq 1/2) = \int_{-\infty}^{1/2} f(x)dx = \int_0^{1/2} 2xdx = 1/4.$$

**L–example 2.6** Let $X$ be a random variable with probability function given by

$$f(x) = \begin{cases} c|x|, & |x| \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

Find $c$.

Solution: The area under the curve $|x|, |x| \leq 1$ is $2 \times (1 \times 1/2) = 1$.




Therefore $c \cdot 1 = 1$ results in $c = 1$.

**L–example 2.7**

- "Time headway" in traffic flow is the elapsed time between the time that one car finishes passing a fixed point and the instant that the next car begins to pass that point.
- Let $X =$ the time headway for two randomly chosen consecutive cars on a highway during a period of heavy flow.
- The following p.d.f. for $X$ was suggested:

$$f(x) = \begin{cases} 0.15e^{-0.15(x-0.5)}, & \text{for } x \geq 0.5; \\ 0, & \text{otherwise.} \end{cases}$$

(a) Verify that $f(x)$ is a legitimate p.d.f. for the RV $X$.

(b) Compute $P(X \leq 5)$.

<u>Solution</u>:

(a) To check that $f(x)$ is a p.d.f., we need only to verify (1) $f(x) \geq 0$ for any $x \in \mathbb{R}$; (2) $\int_{-\infty}^{\infty} f(x)dx = 1$. (1) is clearly satisfied, we prove (2):

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_{0.5}^{\infty} 0.15e^{-0.15(x-0.5)}dx \\ &= 0.15e^{0.075} \int_{0.5}^{\infty} e^{-0.15x}dx \\ &= 0.15e^{0.075} \left( -\frac{1}{0.15}e^{-0.15x} \right) \Big|_{0.5}^{\infty} = 1. \end{aligned}$$

(b)

$$\begin{aligned} P(X \leq 5) &= \int_{-\infty}^{5} f(x)dx = \int_{0.5}^{5} 0.15e^{-0.15(x-0.5)}dx \\ &= 0.15e^{0.075} \left( -\frac{1}{0.15}e^{-0.15x} \right) \Big|_{0.5}^{5} \\ &= e^{0.075} \left( -e^{-0.75} + e^{-0.075} \right) = 0.4908. \end{aligned}$$

## 3 Cumulative Distribution Function

**Definition 1**
*For any RV $X$, we define its cumulative distribution function (c.d.f.) by*

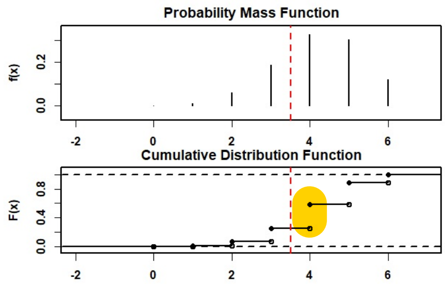
$$F(x) = P(X \leq x).$$

**Note**: This definition is applicable for $X$ to be either a discrete or a continuous RV.

**c.d.f. for Discrete RV**

- If $X$ is a **discrete RV**, 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 c.d.f. of a discrete RV 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 $< a$. More mathematically,

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

$\{X \le b\} = \{X < a\} \cup \{a \le X \le b\}$

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

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

**EXAMPLE 2**

- Revisit Examples 2 and 3. RV $X$ is the number of heads of flipping two fair coins, it has the p.f.:

| $x$ | 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 therefore obtain the c.d.f.:

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

**EXAMPLE 3**

Take the c.d.f. derived from Example 2:

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

Derive the corresponding p.f. (pretending that the c.d.f. is only information available for this distribution).

Solution:

- As $F(\cdot)$ only has four possible values, so 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 so that $f(x)$ is non-zero.
- We have

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

**L–example 2.8**

- Let $X$ = the number of days of sick leave taken by a randomly selected employee of a large company during a particular year.
- If the maximum number of allowable sick leave days per year is 14, possible values of $X$ are $0, 1, 2, \ldots, 14$.
- Suppose $F(0) = 0.58$, $F(1) = 0.72$, $F(2) = 0.76$, $F(3) = 0.81$, $F(4) = 0.88$, and $F(5) = 0.94$.
- We have

$$\begin{aligned} P(2 \le X \le 5) &= F(5) - F(2-) \\ &= F(5) - F(1) = 0.94 - 0.72 = 0.22. \end{aligned}$$

- and

$$\begin{aligned} P(X=3) &= F(3) - F(3-) = F(3) - F(2) \\ &= 0.81 - 0.76 = 0.05. \end{aligned}$$

**L–example 2.9** The p.f. for RV $X$ is given by

$$f(x) = \begin{cases} (1-p)^{x-1} p, & \text{for } x = 1, 2, 3, \ldots; \\ 0, & \text{otherwise,} \end{cases}$$

where $p \in (0, 1)$ is a fixed value. Find the c.d.f. for $X$.
<u>Solution</u>:

- For any $x = 1, 2, 3, \ldots$, set $q = 1 - p$

$$\begin{aligned} F(x) &= P(X \le x) = \sum_{t \le x} f(t) = \sum_{t=1}^{x} (1-p)^{t-1} p \\ &= p\left(1 + q + q^2 + \ldots + q^{x-1}\right) \\ &= p \cdot \frac{1-q^x}{1-q} = 1 - (1-p)^x. \end{aligned}$$

- Question: What is the value of $F(x)$, when $x$ is not a positive integer? For example, $x = 4.3$.

**L–example 2.10** Suppose that the c.d.f. for RV $X$ is given by

$$F(x) = \begin{cases} 1-(1-p)^{\lfloor x \rfloor}, & \text{for } x \geq 1; \\ 0, & \text{for } x < 1, \end{cases}$$

where $\lfloor x \rfloor$ denotes the integer part of $x$. For example, $\lfloor 3.6 \rfloor = 3$, $\lfloor 4 \rfloor = 4$, $\lfloor 4.7 \rfloor = 4$. Find its p.f. $f(x)$.

Solution:

- $F(x)$ changes values only for $x = 1, 2, 3, \ldots$; therefore it is a discrete distribution.
- $R_X = \{1, 2, 3, \ldots, \}$, i.e., the set of positive integers.
- for any $x \in R_X$,

$$\begin{aligned} f(x) &= F(x) - F(x-) = (1-(1-p)^x) - \left(1-(1-p)^{x-1}\right) \\ &= (1-p)^{x-1}(1-(1-p)) = (1-p)^{x-1}p, \end{aligned}$$

  and $f(x) = 0$ otherwise.

**L–example 2.11**

- Many manufacturers have quality control programs that include inspection of incoming materials for defects.
- Suppose a computer manufacturer receives computer boards in lots of five.
- Two boards are selected from each lot for inspection.

(a) List all possible inspected boards for a lot.

(b) Suppose that boards 1 and 2 are the only defectives in a lot of five.

Define $X$ = # of defective boards observed among an inspection.

Find the probability distribution of $X$.

(c) Let $F(x)$ be the c.d.f. of $X$. Derive $F(x)$.

Solution:

(a) $\#(S) = \binom{5}{2} = 10$. The possible selections are

$$\Big\{\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{2,3\},\{2,4\},\{2,5\},\{3,4\},\{3,5\},\{4,5\}\Big\}.$$

(b) $X$ may take values of 0, 1, and 2.

$$\begin{aligned} f(0) &= P(X=0) = P(\{\{3,4\},\{3,5\},\{4,5\}\}) = 3/10, \\ f(2) &= P(X=2) = P(\{\{1,2\}\}) = 1/10, \\ f(1) &= P(X=1) = 1-[f(0)+f(2)] = 6/10, \end{aligned}$$

and $f(x) = 0$ elsewhere.

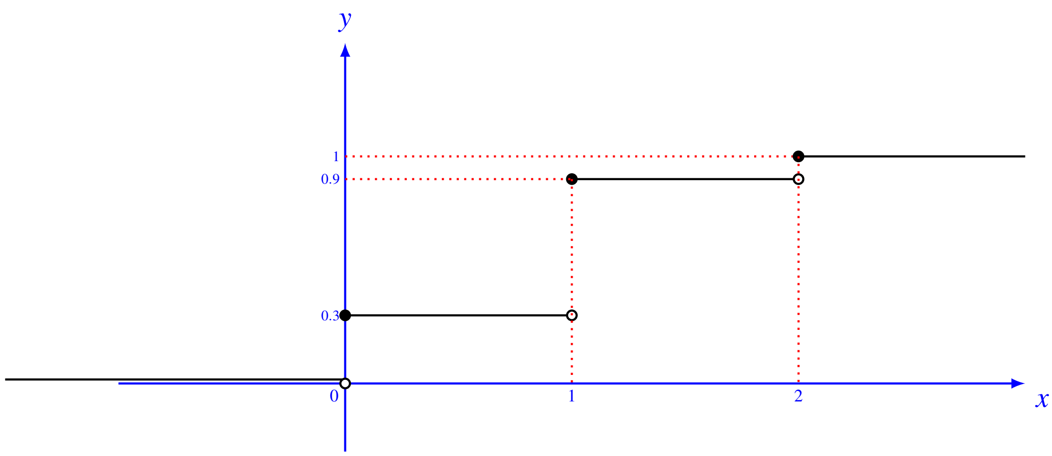
(c) It is sufficient to derive $F(0)$, $F(1)$, $F(2)$:

$$\begin{aligned} F(0) &= P(X \le 0) = f(0) = 0.3, \\ F(1) &= P(X \le 1) = f(0)+f(1) = 0.3+0.6 = 0.9 \\ F(2) &= P(X \le 2) = f(0)+f(1)+f(2) = 1. \end{aligned}$$

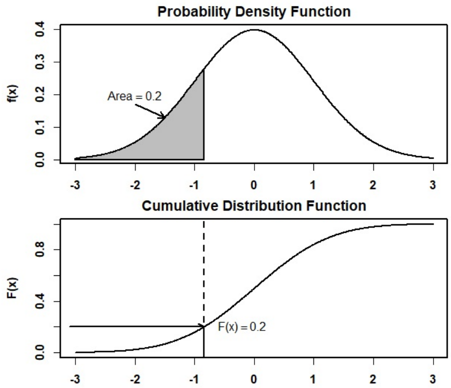
Therefore

$$F(x) = \begin{cases} 0, & x < 0, \\ 0.3, & 0 \le x < 1, \\ 0.9, & 1 \le x < 2, \\ 1, & 2 \le x. \end{cases}$$

This c.d.f. can be drawn as a figure below:



### c.d.f. for Continuous RV




- If $X$ is a continuous RV,

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

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

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

**Example 4**

- The p.d.f. of a RV $X$ is given by

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

- The c.d.f. of $X$ is

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

**Example 5**

Take the c.d.f. derived from Example 4:

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

Derive the corresponding p.f. (pretending that the c.d.f. is only information available for this distribution).

Solution:

- $F(x)$ is a c.d.f. 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 $\frac{d(0)}{dx} = \frac{d(1)}{dx} = 0$.

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

**L–example 2.12**

- Let $X$ be the vibratory stress (psi) on a wind turbine blade at a particular wind tunnel.
- The following p.d.f. for $X$ is proposed:
$$f(x) = \begin{cases} \dfrac{x}{\theta^2} e^{-x^2/(2\theta^2)}, & \text{for } x > 0, \\ 0, & \text{otherwise,} \end{cases}$$
where $\theta$ is a given constant.
- Verify that $f(x)$ is a legitimate p.d.f., and find its c.d.f. $F(x)$.

Solution:

- We first verify that $f(x)$ is a p.d.f.. It is obvious that $f(x) > 0$ for $x > 0$.
$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_0^{\infty} \frac{x}{\theta^2} e^{-x^2/(2\theta^2)} dx = -\int_0^{\infty} d\left(e^{-x^2/(2\theta^2)}\right) \\ &= -e^{-x^2/(2\theta^2)}\Big|_{x=0}^{\infty} \\ &= -0-(-1) = 1. \end{aligned}$$
This verifies that $f(x)$ is a valid p.d.f.
- For $x \le 0$, it is clearly $F(x) = 0$. For $x > 0$,
$$\begin{aligned} F(x) &= P(X \le x) = \int_{-\infty}^{x} f(t)dt = \int_0^x \frac{t}{\theta^2} e^{-t^2/(2\theta^2)} dt \\ &= -e^{-t^2/(2\theta^2)}\Big|_{t=0}^{x} \\ &= 1 - e^{-x^2/(2\theta^2)}. \end{aligned}$$

**L–example 2.13** With the c.d.f. given in the last example:
$$F(x) = 1 - e^{-x^2/(2\theta^2)},$$
for $x \ge 0$ and $F(x) = 0$ otherwise. Derive its p.f.

- As $F(x)$ assumes different values in the interval $x \ge 0$, therefore we have continuous distribution. For any $x \ge 0$, we have
$$\begin{aligned} f(x) &= \frac{dF(x)}{dx} = \frac{d\left[1 - e^{-x^2/(2\theta^2)}\right]}{dx} \\ &= \frac{-d\left[e^{-x^2/(2\theta^2)}\right]}{dx} = \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}, \end{aligned}$$
and $f(x) = 0$ for $x < 0$ since $d(F(x))/dx = d(0)/dx = 0$. This complies with the p.d.f. given in the last example.

**REMARK:**

- No matter whether $X$ is discrete or continuous, $F(x)$ is non-decreasing. In the sense that for any $x_1 < x_2$, $F(x_1) \leq F(x_2)$.
- p.f. and c.d.f. have a one-to-one correspondence. That is, with p.f. given, c.d.f. is uniquely determined; and vice versa.
- The ranges of $F(x)$ and $f(x)$ satisfy:
  - $0 \leq F(x) \leq 1$;
  - for discrete distribution, $0 \leq f(x) \leq 1$;
  - for continuous distribution, $f(x) \geq 0$, but **NO NEED** that $f(x) \leq 1$.

## 4 EXPECTATION AND VARIANCE OF A RV

- For a RV $X$, one natural practical question is: what is the **average value** of $X$, if the corresponding experiment is repeated many times.

  For example, $X$ is the number of heads for flipping a coin. We may want to know the average number of heads if we repeat to flip the coin "continuously".

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

**DEFINITION 1 (EXPECTATION OF DISCRETE RV)**
*Let $X$ be a discrete RV with $R_X = \{x_1, x_2, x_3, \dots\}$ and p.f. $f(x)$. 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 2 (EXPECTATION OF CONTINUOUS RV)**
*Let $X$ be a continuous RV with p.f. $f(x)$. The "**expectation**" or "**mean**" of $X$ is defined by*

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

**Note**: The expected value is not necessarily a possible value of the random variable $X$.

**EXAMPLE 3**
Suppose we toss a fair die and the upper face is recorded as $X$. We have $P(X=k)=1/6$ for $k=1,2,3,4,5,6$, and

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

**EXAMPLE 4**
The p.d.f. of weekly gravel sales $X$ is

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

Then we have

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

**L–example 2.14**

In a gambling game

- a man gains 5 if he gets all heads or all tails in tossing a fair coin 3 times;
- he pays 3 if either 1 or 2 heads show.

What is his expected gain?
Solution:

- Let $X$ be the amount he can gain in the game.
- Then $X=5$ or $-3$ with the following probabilities:

$$\begin{aligned}P(X=5) &= P(\{HHH,TTT\})=1/8+1/8=1/4;\\ P(X=-3) &= 1-P(X=5)=3/4.\end{aligned}$$

- $E(X)=5\left(\frac{1}{4}\right)+(-3)\left(\frac{3}{4}\right)=-1.$
- This means he will lose 1 per toss, if he **continuously play the game for a long run**.

**L–example 2.15**

- Suppose "$X =$ the total number of hours (in units of 100 hours) that a family runs a vacuum cleaner over a period of one year".
- The probability function of $X$ is given by

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

- Find the average number of hours per year that families run their vacuum cleaners.

Solution: The question is asking $100 \times E(X)$.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot x dx + \int_1^2 x(2-x) dx \\ &= \left(\frac{x^3}{3}\right)\Big|_0^1 + \left(x^2 - \frac{x^3}{3}\right)\Big|_1^2 \\ &= \left(\frac{1}{3} - 0\right) + \left[\left(4 - \frac{8}{3}\right) - \left(1 - \frac{1}{3}\right)\right] = 1. \end{aligned}$$

We conclude that on average, families run their vacuum cleaners 100 hours per year.

**Properties of Expectation**

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

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

(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** RV with p.m.f. $f(x)$ and range $R_X$,

$$E[g(X)] = \sum_{x \in R_X} g(x) f(x).$$

- If $X$ is a **continuous** RV with p.d.f. $f(x)$ and range $R_X$,

$$E[g(X)] = \int_{R_X} g(x) f(x) dx.$$

**L–example 2.16** Let $X$ be a random variable, and let $a$ and $b$ be any real numbers. Show that
$$E(aX+b) = aE(X)+b.$$

Solution:

- When $X$ is a discrete random variable with p.f. $f(x)$,

$$\begin{aligned} E(aX+b) &= \sum_{x\in R_X}(ax+b)f(x) \\ &= \sum_{x\in R_X} axf(x) + \sum_{x\in R_X} bf(x) \\ &= a\left(\sum_{x\in R_X} xf(x)\right) + b\left(\sum_{x\in R_X} f(x)\right) = aE(X)+b. \end{aligned}$$

- When $X$ is a continuous random variable with p.f. $f(x)$,

$$\begin{aligned} E(aX+b) &= \int_{-\infty}^{\infty}(ax+b)f(x)dx \\ &= \int_{-\infty}^{\infty}(ax)f(x)dx + \int_{-\infty}^{\infty} bf(x)dx \\ &= a\int_{-\infty}^{\infty} xf(x)dx + b\int_{-\infty}^{\infty} f(x)dx = aE(X)+b \end{aligned}$$

Note that based on properties (1) and (2), we have for constants $a_1, a_2, \ldots, a_k$ and RVs $X_1, X_2, \ldots, X_k$,

$$E(a_1X_1 + a_2X_2 + \ldots + a_kX_k) = a_1E(X_1) + a_2E(X_2) + \ldots + a_kE(X_k).$$

**Variance**

Let $g(x) = (x-\mu_X)^2$, this gives the definition of the **variance** for $X$.

> **DEFINITION 5 (VARIANCE)**
> *Let $X$ be a RV. The **variance** of $X$ is defined by*
> $$\sigma_X^2 = V(X) = E(X-\mu_X)^2.$$

**REMARK:**

- The definition is applicable no matter whether $X$ is discrete or continuous.
- If $X$ is a **discrete** RV with p.m.f. $f(x)$ and range $R_X$,

$$V(X) = \sum_{x\in R_X}(x-\mu_X)^2 f(x).$$

- If $X$ is a **continuous** RV with p.d.f. $f(x)$,

$$V(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)dx.$$

- For any $X$, $V(X) \geq 0$, and "=" holds if and only $P(X = E(X)) = 1$, or more intuitively, $X$ 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 6**
Let the p.f. of a RV $X$ 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:

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

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

**EXAMPLE 7**
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

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

Compute $E(X)$, $V(X)$, and $\sigma_X$.
Solution:

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

We use $V(X) = E(X^2) - [E(X)]^2$ to compute $V(X)$,

$$E(X^2) = \int_0^2 x^2 \cdot x/2 dx = \int_0^2 x^3/2 dx = \left.\frac{x^4}{8}\right|_0^2 = 2.$$

$$V(X) = E(X^2) - [E(X)]^2 = 2 - (4/3)^2 = 2/9.$$

$$\sigma_X = \sqrt{V(X)} = \sqrt{2}/3.$$

**L–example 2.17** Revisit Example 6. Let the p.f. of a RV $X$ be given by

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

(a) Compute $V(X)$ with the alternative formula.

(b) Define $Y = X^2 + 2$. Compute $E(Y)$ and $V(Y)$.

Solution:

(a) We shall use the formula $V(X) = E(X^2) - [E(X)]^2$ to compute the variance. We can use $E(X) = 1$.

$$\begin{aligned} E(X^2) &= \sum_{x \in R_X} x^2 f(x) \\ &= (-1)^2\left(\frac{1}{8}\right) + 0^2\left(\frac{2}{8}\right) + 1^2\left(\frac{1}{8}\right) + 2^2\left(\frac{4}{8}\right) = 9/4. \\ V(X) &= E(X^2) - [E(X)]^2 = 9/4 - 1^2 = 5/4. \end{aligned}$$

(b) $E(Y) = E(X^2) + 2 = 9/4 + 2 = 17/4$. We use $V(Y) = E(Y^2) - [E(Y)]^2$ to compute the variance.

$$\begin{aligned} E(Y^2) &= E[(X^2+2)^2] = E(X^4 + 4X^2 + 4) \\ &= E(X^4) + 4(9/4) + 4 = E(X^4) + 13 \\ &= (-1)^4\left(\frac{1}{8}\right) + 0^4\left(\frac{2}{8}\right) + 1^4\left(\frac{1}{8}\right) + 2^4\left(\frac{4}{8}\right) + 13 \\ &= 85/4; \end{aligned}$$

Therefore

$$V(Y) = E(Y^2) - [E(Y)]^2 = 85/4 - (17/4)^2 = 51/16.$$

**L–example 2.18** Show the property of variance:

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

Solution:

$$\begin{aligned} V(X) &= E[(X-\mu_X)^2] \\ &= E(X^2 - 2X\mu_X + \mu_X^2) \\ &= E(X^2) - E(2X\mu_X) + E(\mu_X^2) \\ &= E(X^2) - 2\mu_X E(X) + \mu_X^2 \\ &= E(X^2) - 2\mu_X^2 + \mu_X^2 = E(X^2) - \mu_X^2, \end{aligned}$$

since $\mu_X = E(X)$ is a constant.

**L–example 2.19** Show the property of the variance: $V(aX+b) = a^2V(X)$, where $a$ and $b$ are constants.

Solution: Note that this property is equivalent to the following two properties

(a) $V(aX) = a^2V(X)$, and

(b) $V(X+b) = V(X)$.

Therefore, we only need to show (a) and (b). For (a)

$$\begin{aligned} V(aX) &= E[(aX)^2] - [E(aX)]^2 = E(a^2X^2) - [aE(X)]^2 \\ &= a^2E(X^2) - a^2[E(X)]^2 = a^2V(X). \end{aligned}$$

For (b),

$$\begin{aligned} V(X+b) &= E[(X+b)^2] - [E(X+b)]^2 \\ &= E(X^2 + 2Xb + b^2) - [E(X)+b]^2 \\ &= E(X^2) + 2bE(X) + b^2 - \{[E(X)]^2 + 2bE(X) + b^2\} \\ &= E(X^2) - [E(X)]^2 = V(X). \end{aligned}$$

**L–example 2.20** Suppose that RV $X$ has p.f.

$$f(x) = \begin{cases} \dfrac{x}{225}, & 0 < x < 15 \\ \dfrac{30-x}{225}, & 15 \le x \le 30 \\ 0, & \text{otherwise} \end{cases}$$

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

<u>Solution</u>:

$$\begin{aligned}
E(X) &= \int_0^{15} x\left(\frac{x}{225}\right)dx + \int_{15}^{30} x\left(\frac{30-x}{225}\right)dx \\
&= \frac{1}{225}\left\{\left(\frac{x^3}{3}\right)\Big|_0^{15} + \left(15x^2 - \frac{x^3}{3}\right)\Big|_{15}^{30}\right\} \\
&= \frac{1}{225}\left\{\frac{15^3}{3} + \left(15(30)^2 - \frac{30^3}{3} - 15(15)^2 + \frac{15^3}{3}\right)\right\} = 15.
\end{aligned}$$

$$\begin{aligned}
E(X^2) &= \int_0^{15} x^2\left(\frac{x}{225}\right)dx + \int_{15}^{30} x^2\left(\frac{30-x}{225}\right)dx \\
&= \frac{1}{225}\left\{\left(\frac{x^4}{4}\right)\Big|_0^{15} + \left(10x^3 - \frac{x^4}{4}\right)\Big|_{15}^{30}\right\} = \frac{525}{2} = 262.5.
\end{aligned}$$

Therefore

$$V(X) = E(X^2) - [E(X)]^2 = 262.5 - 15^2 = 37.5.$$

# Chapter 3: 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 an individual 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**

- *Let $E$ be an experiment and $S$ be a corresponding sample space.*
- *Let $X$ and $Y$ be 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**.*

Similarly to one-dimensional situation, we can denote the **range space** of $(X,Y)$ by

$$R_{X,Y} = \left\{ (x,y) \middle| x = X(s), y = Y(s), s \in S \right\}.$$

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

**Definition 2**
*Let $X_1, X_2, \ldots, X_n$ be $n$ functions each assigning a real number to **every outcome** $s \in S$. We call $(X_1, X_2, \ldots, X_n)$ an **n-dimensional random variable** (or an **n-dimensional random vector**).*

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

**DEFINITION 3**

1 *$(X,Y)$ is a **discrete** two-dimensional RV 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*

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

2 *$(X,Y)$ is a **continuous** two-dimensional RV 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.

forward and backwards are true

- If both $X$ and $Y$ are discrete RVs, then $(X,Y)$ is a discrete RV.
- Likewise, if both $X$ and $Y$ are continuous random variables, then $(X,Y)$ is a continuous RV.
- Clearly, there are other cases. For example, $X$ is discrete, but $Y$ is continuous. These are not our focus in this module.

**EXAMPLE 4 ((DISCRETE RANDOM VECTOR))**

- Consider a TV set to be serviced.
- Let
$$X = \{\text{age to the nearest year of the set}\};$$
$$Y = \{\text{\# of defective components in the set}\}.$$
- $(X,Y)$ is a discrete 2-dimensional RV.
- $R_{X,Y} = \{(x,y)|x=0,1,2,\ldots;y=0,1,2,\ldots,n\}$, where $n$ is the total number of components in the TV.
- $(X,Y)=(5,3)$ means that the TV is 5 years old and has 3 defective components.

**L–example 3.1**

- A fast food restaurant operates a **drive-up facility** and a **walk-up window**.

- On a day, Let

$$X = \text{the proportion of time that the } \textbf{drive-up facility} \text{ is in use;}$$
$$Y = \text{the proportion of time that the } \textbf{walk-up window} \text{ is in use.}$$

- Then $R_{X,Y} = \{(x,y) | 0 \le x, 0 \le y \le 1\}$.
- $(X,Y)$ is a continuous 2-dimensional RV.

**Joint Probability Function**

- We introduce the probability function for the discrete and continuous RVs separately.
- For discrete random vector, similar to the one-dimensional case, we define its probability function by associate a number with each possible value of the RV.

**DEFINITION 5 (JOINT PROBABILITY FUNCTION FOR DISCRETE RV)**
*Let $(X,Y)$ be a 2-dimensional* ***discrete*** *RV, the* ***joint probability (mass) function*** *is defined by*

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

*for $x, y$ being possible values of $X$ and $Y$, or in the other words $(x,y) \in R_{X,Y}$.*

the set {X=x} is a subset of S, in the sense: {X=x} = {s E S : X(s) = x}

The joint probability mass 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) $\sum_{i=1}^{\infty}\sum_{j=1}^{\infty} f_{X,Y}(x_i,y_j) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = 1$;

or 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 6**
Find the value of $k$ such that $f(x,y) = kxy$ for $x = 1,2,3$ and $y = 1,2,3$ can serve as a joint probability function.

<u>Solution</u>: $R_{X,Y} = \{(x,y)|x = 1,2,3; y = 1,2,3\}$.

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

Based on property (3), we have

$$1 = \sum_{x=1}^{3}\sum_{y=1}^{3} kxy = k\sum_{x=1}^{3} x \sum_{y=1}^{3} y$$

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

which results in $k = 1/36$.

**L–example 3.2**

- A company has 2 production lines, $A$ and $B$, which produce at most 5 and 3 machines respectively.
- Let

$$\begin{aligned} X &= \text{number of machines produced by line } A \\ Y &= \text{number of machines produced by line } B. \end{aligned}$$

- The joint probability function $f(x,y)$ for $(X,Y)$ is given in the table, where each entry represents $f(x_i, y_j) = P(X = x_i, Y = y_j)$.
- What is the probability that in a day line $A$ produces more machines than line $B$?

**Table for the joint probability function** $f(x,y)$

| $y$ | $x$ | | | | | | Row |
|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | Total |
| 0 | 0 | 0.01 | 0.02 | 0.05 | 0.06 | 0.08 | 0.22 |
| 1 | 0.01 | 0.03 | 0.04 | 0.05 | 0.05 | 0.07 | 0.25 |
| 2 | 0.02 | 0.03 | 0.05 | 0.06 | 0.06 | 0.07 | 0.29 |
| 3 | 0.02 | 0.04 | 0.03 | 0.04 | 0.06 | 0.05 | 0.24 |
| Column Total | 0.05 | 0.11 | 0.14 | 0.20 | 0.23 | 0.27 | 1 |

Consider the event

$$A = \{\text{line } A \text{ produces more machines than line } B\} = \{X > Y\}.$$

Then we have

$$
\begin{aligned}
P(A) &= P(X>Y) \\
&= P\Big((X,Y)=(1,0) \text{ or } (X,Y)=(2,0) \text{ or } \\
&\quad (X,Y)=(2,1) \text{ or } \ldots \text{ or } (X,Y)=(5,3)\Big) \\
&= P\Big((X,Y)=(1,0)\Big)+\ldots+P\Big((X,Y)=(5,3)\Big) \\
&= f(1,0)+f(2,0)+\ldots+f(5,3)=0.73.
\end{aligned}
$$

**L–example 3.3**

- A company has 9 executives; 4 are married, 3 have never married, and 2 are divorced.
- Three executives are to be randomly selected for promotion.
- Among the selective executives, let

$$
\begin{aligned}
X &= \{\text{number of married executives}\} \\
Y &= \{\text{number of never married executives}\}.
\end{aligned}
$$

- Find the joint probability function of $X$ and $Y$.

Solution: Note that the executives are selected randomly; so every possible selection of the executives are equally likely.

- The total number of ways to select 3 executives out of 9 is $\binom{9}{3}$.
- The possible values of $x$ and $y$ are constrained by $x,y=0,1,2,3$ and $1 \leq x+y \leq 3$. The number of ways to select $x$ married and $y$ never married is given by $\binom{4}{x}\binom{3}{y}\binom{2}{3-x-y}$.
- Therefore, the joint probability function of $(X,Y)$ is given by

$$
\begin{aligned}
f_{X,Y}(x,y) &= P(X=x, Y=y) \\
&= \frac{\binom{4}{x}\binom{3}{y}\binom{2}{3-x-y}}{\binom{9}{3}},
\end{aligned}
$$

  for $x,y=0,1,2,3$ such that $1 \leq x+y \leq 3$ and $f_{X,Y}(x,y)=0$ otherwise.

- This joint p.f. can be summarized as a table.

| $x$ | $y$ = 0 | 1 | 2 | 3 | Row Total |
|---|---|---|---|---|---|
| 0 | 0 | $3/84$ | $6/84$ | $1/84$ | $10/84$ |
| 1 | $4/84$ | $24/84$ | $12/84$ | 0 | $40/84$ |
| 2 | $12/84$ | $18/84$ | 0 | 0 | $30/84$ |
| 3 | $4/84$ | 0 | 0 | 0 | $4/84$ |
| Column Total | $20/84$ | $45/84$ | $18/84$ | $1/84$ | 1 |

**DEFINITION 7 (JOINT PROBABILITY FUNCTION FOR CONTINUOUS RV)**
*Let $(X,Y)$ be a 2-dimensional **continuous** RV; its **joint probability (density) function** is a function $f_{X,Y}(x,y)$ such that*

$$P((X,Y) \in D) = \int\int_{(x,y)\in D} f_{X,Y}(x,y)dydx,$$

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

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

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) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$;

or equivalently $\int\int_{(x,y)\in R_{X,Y}} f_{X,Y}(x,y)dxdy = 1$.

**EXAMPLE 8**
Find the value $c$ such that $f(x,y)$ below can serve as a joint p.d.f. for a RV $(X,Y)$:

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

Solution: In order for $f(x,y)$ to be a p.d.f., we need

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

which implies $c = 12/13$.

**L–example 3.4**

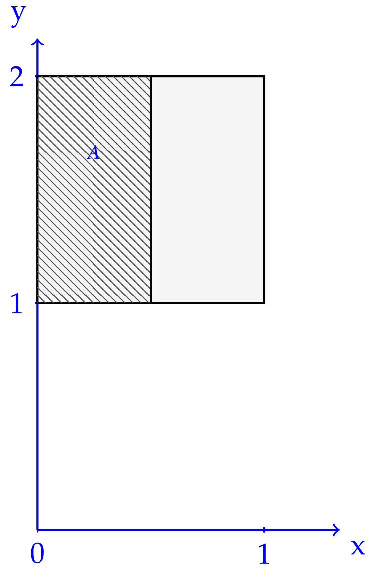
Reuse the p.d.f. of Example 8:

$$f(x,y) = \begin{cases} \frac{12}{13}x(x+y), & 0 \le x \le 1; 1 \le y \le 2 \\ 0, & \text{elsewhere} \end{cases}.$$

Assume that it is the joint p.d.f. of $(X,Y)$. Let $A = \{(x,y)|0 < x < 1/2; 1 < y < 2\}$. Compute $P((X,Y) \in A)$.

- Set $A$ corresponds to the shaped area in the figure on the right.
- We have

$$\begin{aligned} P((X,Y) \in A) &= P(0 < X < 1/2; 1 < Y < 2) \\ &= \int_0^{1/2} \int_1^2 \frac{12}{13}x(x+y)dydx \\ &= \frac{12}{13}\int_0^{1/2} x(x+1.5)dx \\ &= \frac{12}{13}\left(\frac{1}{3}x^3 + 1.5 \cdot \frac{1}{2}x^2\right)\Big|_0^{1/2} \\ &= {}^{11}/_{52}. \end{aligned}$$




## 2 Marginal and Conditional Distributions

**Definition 1 (Marginal Probability Distribution)**
*Let $(X,Y)$ be a two-dimensional RV with joint p.f. $f_{X,Y}(x,y)$. We define the marginal distribution for $X$ as follows.*

- *If $Y$ is a discrete RV, then for any $x$,*

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

- *If $Y$ is a continuous RV, then for any $x$,*

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

**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.

- More intuitively, it is the distribution of $X$ by ignoring the presence of $Y$.

  For example, consider a person of a certain community,

  - suppose $X =$ body weight, $Y =$ height. $(X,Y)$ has a 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 2**

- Revisit Example 6. The joint p.f. is given by $f(x,y) = \frac{1}{36}xy$ for $x = 1,2,3$ 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 $x \in \{1,2,3\}$,

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

**L–example 3.5**

We reuse the joint p.f. of $(X,Y)$ derived in L–Example 1:

| $x$ | $y$ | | | | Row Total |
|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | |
| 0 | 0 | $3/84$ | $6/84$ | $1/84$ | $10/84$ |
| 1 | $4/84$ | $24/84$ | $12/84$ | 0 | $40/84$ |
| 2 | $12/84$ | $18/84$ | 0 | 0 | $30/84$ |
| 3 | $4/84$ | 0 | 0 | 0 | $4/84$ |
| Column Total | $20/84$ | $45/84$ | $18/84$ | $1/84$ | 1 |

Can we read out the marginal p.f. of $X$ and $Y$ from the table directly?

**L–example 3.6**

Reuse the p.d.f. of Example 8:

$$f(x,y) = \begin{cases} \frac{12}{13}x(x+y), & 0 \le x \le 1; 1 \le y \le 2 \\ 0, & \text{elsewhere} \end{cases}.$$

Assume that it is the joint p.d.f. of $(X,Y)$. Find the marginal distribution of $X$.

Solution: $(X,Y)$ is a continuous RV. For each $x \in [0,1]$, we have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy = \int_1^2 \frac{12}{13}x(x+y)dy \\ &= \frac{12}{13}x\left(x + \int_1^2 ydy\right) \\ &= \frac{12}{13}x(x+1.5); \end{aligned}$$

and for $x \notin [0,1]$, $f_X(x) = 0$.

**DEFINITION 3 (CONDITIONAL DISTRIBUTION)**
*Let $(X,Y)$ be a RV with joint p.f. $f_{X,Y}(x,y)$. Let $f_X(x)$ be the marginal p.f. 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 $Y = y$**:

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

- $f_{Y|X}(y|x)$ is defined only for $x$ such that $f_X(x) > 0$; likewise $f_{X|Y}(x|y)$ is defined only for $y$ such that $f_Y(y) > 0$.

$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \geq 0$

- The practical meaning of $f_{Y|X}(y|x)$: the distribution of $Y$ given that the random variable $X$ is observed to take the value $x$.

$\sum_y f_{Y|X}(y|x) = \frac{\sum_y f_{X,Y}(x,y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1$

- Considering $y$ as the variable ($x$ as a fixed value), $f_{Y|X}(y|x)$ is a p.f., so it must satisfy all the properties of p.f..

- But $f_{Y|X}(y|x)$ is not a p.f. for $x$; this means that there is **NO** requirement $\int_{-\infty}^{\infty} f_{Y|X}(y|x)dx = 1$ for $X$ continuous or $\sum_x f_{Y|X}(y|x) = 1$ for $X$ discrete.

- With the 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 RV,

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

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

For discrete case, the computation is similarly established based on $f_{Y|X}(y|x)$; please fill in the details on your own.

**EXAMPLE 4**
Revisit Examples 6 and 2.

- The joint p.f. for $(X, Y)$ is given by $f(x,y) = \frac{1}{36}xy$ for $x = 1,2,3$ and $y = 1,2,3$.
- The marginal p.f. for $X$ is $f_X(x) = \frac{1}{6}x$ for $x = 1,2,3$.
- Therefore, $f_{Y|X}(y|x)$ is defined for any $x = 1,2$, or 3:

$$
f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{(1/36)xy}{(1/6)x} = \frac{1}{6}y,
$$

for $y = 1,2,3$.

- We can compute

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

**L–example 3.7**

We reuse the joint p.f. of $(X,Y)$ derived in L–Example 1:

| $x$ | $y$: 0 | 1 | 2 | 3 | Row Total |
|---|---|---|---|---|---|
| 0 | 0 | $3/84$ | $6/84$ | $1/84$ | $10/84$ |
| 1 | $4/84$ | $24/84$ | $12/84$ | 0 | $40/84$ |
| 2 | $12/84$ | $18/84$ | 0 | 0 | $30/84$ |
| 3 | $4/84$ | 0 | 0 | 0 | $4/84$ |
| Column Total | $20/84$ | $45/84$ | $18/84$ | $1/84$ | 1 |

Can we read out the conditional p.f. $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$ from the table directly? How to compute $E(Y|X=x)$?

**L–example 3.8** Reuse Examples 8 and L–Example 2.

- The joint p.f. for $(X,Y)$ is given by

$$
f(x,y) = \begin{cases} \frac{12}{13}x(x+y), & 0\le x\le 1; 1\le y\le 2 \\ 0, & \text{elsewhere} \end{cases}.
$$

- The marginal p.f. for $X$ is given by

$$
f_X(x) = \frac{12}{13}x(x+1.5),
$$

  for $x\in[0,1]$.

- For each $x\in[0,1]$, the conditional p.f. $f_{Y|X}(y|x)$,

$$
\begin{aligned}
f_{Y|X}(y|x) &= \frac{f(x,y)}{f_X(x)} = \frac{(12/13)x(x+y)}{(12/13)x(x+1.5)} \\
&= \frac{x+y}{x+1.5},
\end{aligned}
$$

  for $y\in[1,2]$.

- We can compute

$$P(Y \le 1.5|X=0.5) = \int_1^{1.5} \frac{0.5+y}{0.5+1.5}dy = 0.5625.$$

- Furthermore

$$\begin{aligned} E(Y|X=0.5) &= \int_1^2 y\frac{0.5+y}{0.5+1.5}dy \\ &= \frac{1}{2}\int_1^2 (0.5y+y^2)dy \\ &= \frac{1}{2}\left(\frac{3}{4}+\frac{7}{3}\right) = 37/24. \end{aligned}$$

## 3 Independent Random Variables

**Definition 1 (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 no matter 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

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

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

  **<u>Conclusion</u>: 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 RVs.

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

**Example 2**

The joint p.f. 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.

- We conclude that $X$ and $Y$ are independent.

**L–example 3.9** Given that

$$f_{X,Y}(x,y) = \begin{cases} 2(x+y), & \text{for } 0 \le x \le 1, 0 < y < x \\ 0 & \text{elsewhere} \end{cases}$$

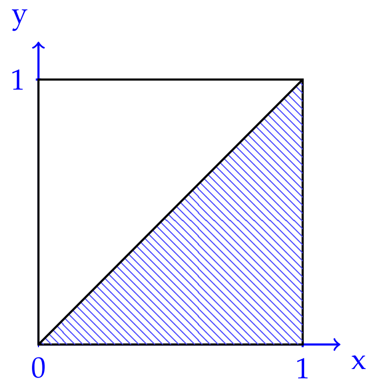
Are $X$ and $Y$ independent?
Solution:

- The direct way of checking the independence is to check whether

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

  holds for every $(x,y)$ combination. The detail of this method is left as an exercise.

- For this question, we can immediately conclude that $X$ and $Y$ are not independent by checking that $R_{X,Y}$ is not a product space.




**L–example 3.10** Suppose that $(X,Y)$ is a discrete RV. The joint p.f. is given by

| $x$ | $y$ | | | | $f_X(x)$ |
|---|---|---|---|---|---|
| | 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 |

Are $X$ and $Y$ independent?
Solution:
The zero entries in the table indicate that $R_{X,Y}$ is not a product space. Therefore, $X$ and $Y$ are not independent.

**L–example 3.11** We have a handy way to check independence when $f_{X,Y}(x,y)$ has an explicit formula in $R_{X,Y}$.

$X$ and $Y$ are independent if and only if both of the following hold:

- $R_{X,Y}$, the range that the p.f. is positive, is a product space.

- For any $(x,y) \in R_{X,Y}$, we have $f_{X,Y}(x,y) = C \cdot g_1(x) g_2(y)$; that is, it can be "factorized" as the product of two functions $g_1$ and $g_2$, where the former **depends on** $x$ **only**, the latter **depends on** $y$ **only**, and $C$ is a constant not depending on both $x$ and $y$.

**Note**: $g_1(x)$ and $g_2(y)$ on their own are NOT necessarily p.f.s.

- We use the joint p.d. in Example 6 to illustrate: $f(x,y) = \frac{1}{36}xy$ for $x = 1,2,3$ and $y = 1,2,3$.
- $A_1 = \{1,2,3\}$ and $A_2 = \{1,2,3\}$, so the $R_{X,Y}$ is a product space.
- $f_{X,Y}(x,y) = \frac{1}{36} \cdot (x) \cdot (y)$: $C = 1/36$, $g_1(x) = x$, $g_2(y) = y$.
- We conclude that $X$ and $Y$ are independent.
- The advantage of this method is that we don't need to find the marginal distributions $f_X(x)$ and $f_Y(y)$ and check $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

Following this strategy, we can get $f_X(x)$ and $f_Y(y)$ by standardizing $g_1(x)$ and $g_2(y)$. Consider $f_X(x)$ for illustration; $f_Y(y)$ is obtained similarly.

- If $X$ is a discrete RV, its p.m.f. is given by

$$f_X(x) = \frac{g_1(x)}{\sum_{t \in R_X} g_1(t)}.$$

- If $X$ is a continuous RV, its p.d.f. is given by

$$f_X(x) = \frac{g_1(x)}{\int_{t \in R_X} g_1(t)} dt.$$

- We continue to use the example above to illustrate. Here $X$ is a discrete RV, $R_X = A_1 = \{1,2,3\}$. We obtain its p.m.f.:

$$f_X(x) = \frac{g_1(x)}{\sum_{x \in R_X} g_1(x)} = \frac{x}{\sum_{x=1}^{3} x} = x/6.$$

- Similarly, we get $f_Y(y) = y/6$.

**L–example 3.12** Given that

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{3}x(1+y), & \text{for } 0 < x < 2, 0 < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

Are $X$ and $Y$ independent?
Solution:

- Set $A_1 = (0,2)$ and $A_2 = (0,1)$, then $R_{X,Y} = A_1 \times A_2$ is a product space.

- $f_{X,Y}(x,y)$ in $R_{X,Y}$ can be factorized by $C = 1/3$, $g_1(x) = x$, $g_2(y) = 1+y$. Therefore, we conclude that $X$ and $Y$ are independent.

- Furthermore,

$$f_X(x) = \frac{g_1(x)}{\int_{x \in A_1} g_1(x)dx} = \frac{x}{\int_0^2 xdx} = x/2;$$

$$f_Y(y) = \frac{g_2(y)}{\int_{y \in A_2} g_2(y)dy} = \frac{1+y}{\int_0^1 (1+y)dy} = \frac{2}{3}(1+y).$$

## 4 Expectation and Covariance

**Definition 1 (Expectation)**
*For any two variable function $g(x,y)$,*

- *if $(X,Y)$ is a discrete RV,*

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

- *if $(X,Y)$ is a continuous RV,*

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

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 2 (Covariance)**
*The **covariance** of $X$ and $Y$ is defined to be*

$$cov(X,Y) = E[(X - E(X))(Y - E(Y))]$$

$$= E[(X - \mu_x)(Y - \mu_y)]$$

- If $X$ and $Y$ are discrete RVs,

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

- If $X$ and $Y$ are continuous RVs,

$$\text{cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)(y-\mu_Y)f_{X,Y}(x,y)dxdy.$$

The covariance has the following properties.

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

$= E[(x-\mu_x)(y-\mu_y)]$

can take out miu outside of the expecation since its a constant

$= E[xy - \mu_y \cdot x - \mu_x \cdot y + \mu_x \mu_y]$

$= E(xy) - \mu_y E(x) - \mu_x E(y) + \mu_x \mu_y$

$= Exy - \mu_y \mu_x - \mu_x \mu_y + \mu_x \mu_y$

(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.

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

cov(X,Y) = cov(Y,X)

cov(X+b, Y) = cov(X,Y)

cov(aX,Y) = a cov(X,Y)

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

$V(ax) = a^2 V(X)$

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

**EXAMPLE 3**

Given the joint distribution for $(X,Y)$:

| $x$ | $y$ = 0 | 1 | 2 | 3 | $f_X(x)$ |
|---|---|---|---|---|---|
| 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 |

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

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

<u>Solution:</u>

(a) 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

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

(b) 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 (a).

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

Therefore

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

**L–example 3.13** Suppose that $(X,Y)$ has the p.f.

$$f_{X,Y}(x,y) = \begin{cases} x^2 + \dfrac{xy}{3}, & \text{for } 0 \le x \le 1, 0 \le y \le 2 \\ 0, & \text{otherwise} \end{cases}.$$

(a) Find $f_X(x)$, $f_Y(y)$ and $f_{Y|X}(y|x)$.

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

Solution:

(a) We first find the marginal density of $X$.

For $0 \le x \le 1$,

$$\begin{aligned} f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y &= \int_0^2 \left(x^2 + \frac{xy}{3}\right) \mathrm{d}y \\ &= \left(x^2 y + \frac{xy^2}{6}\right)\Bigg|_{y=0}^{2} = 2x^2 + \frac{2x}{3}. \end{aligned}$$

It is clear that $f_X(x) = 0$ for $x < 0$ or $x > 1$. Thus

$$f_X(x) = \begin{cases} 2x^2 + \dfrac{2x}{3}, & \text{for } 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

Similarly, the marginal density of $Y$ is given as

$$f_Y(y) = \begin{cases} \frac{1}{3} + \frac{y}{6}, & \text{for } 0 \le y \le 2 \\ 0, & \text{otherwise} \end{cases}.$$

The conditional probability density function of $Y$ given $X = x$ when $0 \le x \le 1$ is then given as

$$\begin{aligned} f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} &= \begin{cases} \frac{x^2 + xy/3}{2x^2 + 2x/3}, & \text{for } 0 \le y \le 2 \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} \frac{3x+y}{2(3x+1)}, & \text{for } 0 \le y \le 2 \\ 0, & \text{otherwise} \end{cases}. \end{aligned}$$

(b) We shall use the expression $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$.

Now

$$\begin{aligned} E(XY) &= \int_0^2 \int_0^1 xy\left(x^2 + \frac{xy}{3}\right) \text{dx}\,\text{dy} \\ &= \int_0^2 \int_0^1 \left(yx^3 + \frac{y^2x^2}{3}\right) \text{dx}\,\text{dy} \\ &= \int_0^2 \left(y\frac{x^4}{4} + \frac{y^2x^3}{9}\right)\bigg|_{x=0}^1 \text{dy} \\ &= \int_0^2 \left(\frac{y}{4} + \frac{y^2}{9}\right) \text{dy} \\ &= \frac{43}{54}. \end{aligned}$$

We have computed the marginal distributions for $X$ and $Y$ in Part (a). Thus

$$E(X) = \int_0^1 x\left(2x^2 + \frac{2x}{3}\right) \text{dx} = \left(\frac{2x^4}{4} + \frac{2x^3}{9}\right)\bigg|_{x=0}^1 = \frac{13}{18},$$

and

$$E(Y) = \int_0^2 y\left(\frac{1}{3} + \frac{y}{6}\right) \text{dy} = \left(\frac{y^2}{6} + \frac{y^3}{18}\right)\bigg|_{y=0}^2 = \frac{10}{9}.$$

This gives

$$\text{cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{43}{54} - \frac{13}{18} \times \frac{10}{9} = -\frac{1}{162}.$$

**L–example 3.14**

- Start from $V(X+Y) = V(X) + V(Y) + 2\,\mathrm{cov}(X,Y)$, we can have some interesting results.
- By induction, we have for any random variables $X_1, X_2, \ldots, X_n$,

$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \ldots + V(X_n) + 2\sum_{j>i} \mathrm{cov}(X_i, X_j).$$

- If $X$ and $Y$ are independent, we have

$$V(X \pm Y) = V(X) + V(Y).$$

- By induction, we have if $X_1, X_2, \ldots, X_n$ are independent,

$$V(X_1 \pm X_2 \pm \ldots \pm X_n) = V(X_1) + V(X_2) + \ldots + V(X_n).$$

# Chapter 4: Special Probability Distributions

## 1 Discrete Distributions

- Recall that for a discrete random variable $X$, the number of possible values (i.e., $R_X$) is **finite** or **countable**.
- The elements of $R_X$ can be listed as $x_1, x_2, x_3, \ldots$.
- In this section, we study some classes of discrete random variables.

### Discrete Uniform Distribution

**Definition 1**

- *If RV $X$ assumes the values $x_1, x_2, \ldots, x_k$ with equal probability, then $X$ follows a* ***discrete uniform distribution****.*
- *The p.f. for $X$ is given by*

$$f_X(x) = \frac{1}{k}, \qquad x = x_1, x_2, \ldots, x_k,$$

*and 0 otherwise.*

**Theorem 2**

*Suppose $X$ follows the discrete uniform distribution with $R_X = \{x_1, x_2, \ldots, x_k\}$, we have*

- *The expectation 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 is given by*

$\sum_{i=1}^{k} x_i^2 f_X(x)$

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

**EXAMPLE 3**

- A bulb is selected at random from a box that contains a 40-watt bulb, a 60-watt bulb, an 80-watt bulb, and a 100-watt bulb.
- Each bulb has $1/4$ probability of being selected.
- Let $X$ = the watts of the bulb being selected. Then $X$ follows a uniform distribution, and

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

$$f_X(x) = 1/4, \qquad \text{for } x = 40, 60, 80, 100,$$

  and 0 otherwise.
- We can compute the expectation:

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

- Variance can also be computed:

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

**L–example 4.1**

- Toss a fair die, $X$ = the number on the top face. Then $X$ follows a uniform distribution.
- $R_X = \{1, 2, 3, 4, 5, 6\}$, and

$$f_X(x) = 1/6, \qquad \text{for } x = 1, 2, 3, 4, 5, 6,$$

  and 0 otherwise.
- Expectation can be computed by

$$E(X) = \sum_i x_i f_X(x_i) = \sum_{i=1}^{6} i\left(\frac{1}{6}\right) = 3.5.$$

- Variance can be computed by

$$\begin{aligned} V(X) &= \sum_i x_i^2 f_X(x_i) - (E(X))^2 \\ &= \sum_{i=1}^{6} i^2 \left(\frac{1}{6}\right) - 3.5^2 = \frac{35}{12}. \end{aligned}$$

**Bernoulli Trial, Bernoulli Random Variable and Bernoulli Process**

**DEFINITION 4 (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 code the two outcomes as "1" (success) and "0" (failure).*

**DEFINITION 5 (BERNOULLI RANDOM VARIABLE)**

- *Let $X =$ 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 $p$ $(0 \le p \le 1)$ the probability of success of the Bernoulli trial. Then $X$ has the p.f.:*

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

*and $= 0$ for other values of $x$.*

- *This p.f. can also be written by*

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

- We often denote $X \sim \text{Bernoulli}(p)$, and denote $q = 1 - p$. Then the p.f. becomes $f_X(1) = p$ and $f_X(0) = q$.

**THEOREM 6**

*For a Bernoulli RV defined above, we have*

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

**REMARK (PARAMETERS):**

- In occasions, $f_X(x)$ may rely on one or more unknown quantities; different values of the quantities lead to different probability distributions.
- Such a quantity is called a **parameter** of the distribution.
- $p$ is the parameter in the Bernoulli distribution.
- The collection of the distributions that are determined by one or more unknown parameters is called a **family of probability distributions**.
- So the aforementioned Bernoulli distributions determined by the parameter $p$ is a family of probability distributions.

**EXAMPLE 7**
The following are all examples of Bernoulli trials:

- A coin toss
  Say we want heads, then H="heads" is success, and T="tails" is failure.
- Rolling a die
  Say we only care about rolling a 6. The outcome space is binarized to "success"$= \{6\}$ and "failure" $= \{1,2,3,4,5\}$.
- Polls
  Choosing a voter at random to ascertain whether that voter will vote "yes" in an upcoming referendum.

**EXAMPLE 8**

- 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.
- $P(X = 1) = 4/10 = 0.4$.
- Furthermore, the p.f. for $X$ is given by

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

**DEFINITION 9 (BERNOULLI PROCESS)**

- *A **Bernoulli process** consists of a sequence of repeatedly performed **independent and identical** Bernoulli trials.*

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

We are able to define several useful distributions based on Bernoulli trial and Bernoulli process. These distributions include:

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

### Binomial Distribution

If we have several (say $n$) i.i.d. Bernoulli trials, we can establish 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?
- Urn has 4 black balls and 3 white balls, draw 5 balls with replacement. How many black balls will there be?

**DEFINITION 10 (BINOMIAL RANDOM VARIABLE)**
*A* ***Binomial random variable*** *counts the number of successes in $n$ trials in 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 $X \sim B(n, p)$.*

*The probability of getting exactly $x$ successes is given as*

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

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

The theoretical development for Binomial distribution will be given in a lecture meeting.

**L–example 4.2 (Theory of the Binomial Distribution)**

- Based on the definition of binomial distribution: "$X$ is the number of successes in $n$ trials in a Bernoulli Process", so $X \sim B(n,p)$ **if and only if**
$$X = X_1 + X_2 + \ldots + X_n,$$
with $X_1, X_2, \ldots, X_n$ being i.i.d. Bernoulli($p$) RVs.
- We are able to derive the p.f. for $X$ as follows.
- Consider a specific realization of $X_1, X_2, \ldots, X_n$, namely $x_1, x_2, \ldots, x_n$ such that $\sum_{i=1}^{n} x_n = x$.
- Because $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli($p$) RVs, we have
$$\begin{aligned} & P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ = \ & P(X_1 = x_1)P(X_2 = x_2)\ldots P(X_n = x_n) \\ = \ & \prod_{i=1}^{n} p^{x_i} q^{1-x_i} = p^{\sum_{i=1}^{n} x_i} q^{n - \sum_{i=1}^{n} x_i} \\ = \ & p^x q^{n-x}. \end{aligned}$$
- Note that $\sum_{i=1}^{n} x_i = x$ means: out of $n$ trials, $x$ are observed as success and the rest as failure.
- For the collection of $n$ trials, how many such outcomes are possible? The answer is $\binom{n}{x}$, since it is equivalently to choosing $x$ trials out of $n$ to take success, and the rest take failure.
- Furthermore, for different choices of $x_1, x_2, \ldots, x_n$,
$$\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}$$
are mutually exclusive events.
- We have
$$\begin{aligned} P(X = x) &= P\left( \bigcup_{x_1,\ldots,x_n : \sum x_i = x} \{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \right) \\ &= \sum_{x_1,\ldots,x_n : \sum x_i = x} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= \sum_{x_1,\ldots,x_n : \sum x_i = x} p^x q^{n-x} = \binom{n}{x} p^x q^{n-x}. \end{aligned}$$

- We can also derive other characteristics of the binomial distribution based on the expression

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

- Expectation is given by

$$E(X) = E(X_1) + E(X_2) + \ldots + E(X_n) = p + p + \ldots + p = np.$$

- Because of the independence of $X_1, X_2, \ldots, X_n$, variance is

$$\begin{aligned} V(X) &= V(X_1 + X_2 + \ldots + X_n) = V(X_1) + V(X_2) + \ldots + V(X_n) \\ &= pq + pq + \ldots + pq = npq. \end{aligned}$$

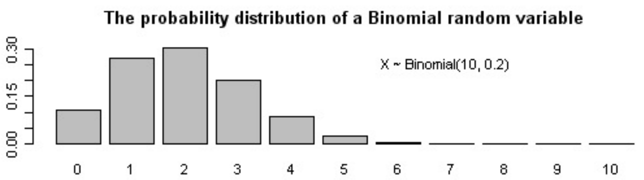
**REMARK:**

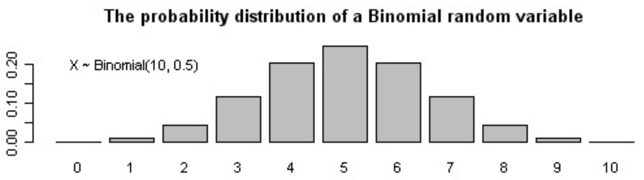
- When $n = 1$, the p.f. for the binomial RV $X$ is reduced to

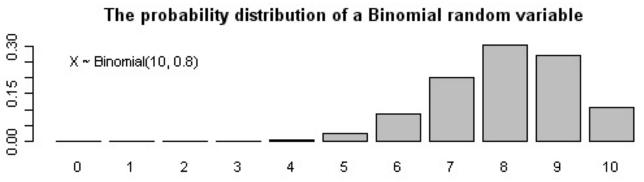
$$f_X(x) = p^x (1-p)^{1-x}, \quad \text{for } x = 0, 1.$$

- It is the p.f. for the Bernoulli distribution. Therefore Bernoulli distribution is a special case of the binomial distribution. ■

The p.f. for $B(10, 0.2)$, $B(10, 0.5)$, and $B(10, 0.8)$ are compared below.



**EXAMPLE 11**

- Flip a fair coin 10 independent times.
- What is the probability of observing exactly 6 heads?

Solution:

- Let $X =$ number of heads in 10 flips.
- Each flip of the coin can be observed as a Bernoulli trial, with probability of getting head (success) $p = 0.5$.
- Then $X$ is the number success out of 10 Bernoulli trials; so $X \sim B(10, 0.5)$.
- We can compute

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

**L–example 4.3** Pat Statsdud failed to study for the next statistics quiz. Pat's strategy is to rely on luck. The quiz consists of 10 multiple-choice questions. Each question has five possible answers, only one of which is correct. Pat plans to guess the answer to every question.

(a) What is the probability that Pat gets two answers correct?

(b) What is the probability that Pat fails the quiz? (suppose it is considered a failed quiz if a grade on the quiz is less than 50% , i.e. 5 questions out of 10).

Solution: Let $X$ denote the number of correct answers. Then $X \sim B(10, 0.2)$.

(a) The probability that he gets two correct answers is given by

$$P(X = 2) = \binom{10}{2}(0.2)^2(0.8)^8 \approx 0.302.$$

(b) The probability that he fails is given by

$$\begin{aligned} P(\text{fail quiz}) &= P(X \le 4) \\ &= P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \\ &\approx 0.967. \end{aligned}$$

To compute $P(X \le 4)$ for $X \sim B(10, 0.2)$:

(A) Method 1: use an online `R` compiler:

- Browse to `https://rdrr.io/snippets/`

- Enter the command

```
pbinom(4, 10, 0.2, lower.tail = TRUE)
```

  unto the compiler.
- `Ctrl-Enter` or `Run` to obtain the answer.
- For $X \sim B(n, p)$.
  - `pbinom(x, n, p)` gives $P(X \leq x)$.
  - `pbinom(x, n, p, lower.tail=FALSE)` gives $P(X > x)$.
  - `dbinom(x, n, p)` gives $P(X = x)$.

(B) Method 2: use `R` Shiny app `Radiant`:

- Browse to `https://vnijs.shinyapps.io/radiant`
- Select `Basics > Probability Calculator`.
- Select `Binomial` as the `Distribution`.
- Select $n$ as 10, $p$ as 0.2.
- Select `Values` as the `Input type`.
- Select 4 as the upper bound, $P(X = 4)$, $P(X \leq 4)$, $P(X > 4)$ are included.

**L–example 4.4**

- A man claims to have extrasensory perception (ESP).
- As a test, a fair coin is flipped 10 times, and he is asked to predict the outcome in advance.
- The man gets 7 out of 10 correct.
- What is the probability that he would have done at least this well if he had no ESP? That is, he gets 7 or more out of 10 correct.

Solution:

- Without ESP, the probability that he guesses correctly for each outcome is 0.5.
- Let $X =$ number of correct guesses out of 10 guesses. Then $X \sim B(10, 0.5)$.
- We have

$$\begin{aligned} & P(X \geq 7) = P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) \\ = & \binom{10}{7} 0.5^7 0.5^3 + \binom{10}{8} 0.5^8 0.5^2 + \binom{10}{9} 0.5^9 0.5^1 + \binom{10}{10} 0.5^{10} 0.5^0 \\ = & 0.1719. \end{aligned}$$

### Negative Binomial Distribution

- Consider a Bernoulli process, where the Bernoulli experiments can be repeated an arbitrary number of times.
- The interest could be how many trials are needed so that a certain number of successes occur.
- Set $X =$ number of trials until the $k$th success occurs. Then $X$ follows a **negative binomial distribution**; denoted by $X \sim NB(k, p)$, where $p$ is probability of success for each Bernoulli trial.
- In comparison with binomial distribution: the random variable "$X$" is the number of successes out of a fixed number $n$ of trials.

**Definition 12 (Negative Binomial Distribution)**

- *$X$ = number of i.i.d. Bernoulli(p) trials until the kth success occurs; then $X$ follows a **negative binomial distribution**, denoted by $X \sim NB(k, p)$.*
- *The p.f. of $X$ is given by*

$$f_X(x) = P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k},$$

*for $x = k, k+1, k+2, \ldots$*

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

### L–example 4.5

- We derive the probability function of the negative binomial distribution.
- We can interpret the event $X = x$ as follows,

$$\begin{aligned} \{X = x\} &= \{\text{used } x \text{ trials until the } k\text{th success occurs}\} \\ &= \{\text{observe } k-1 \text{ successes in the first } x-1 \text{ trials}\} \\ &\quad \cap \{x\text{th trial is a success}\} \\ &= A \cap B. \end{aligned}$$

- Based on binomial distribution,

$$\begin{aligned} P(A) &= P(\text{observe } k-1 \text{ successes in the first } x-1 \text{ trials}) \\ &= \binom{x-1}{k-1} p^{x-1} (1-p)^{x-k} \end{aligned}$$

- Since the last trial is the Bernoulli trial,

$$P(B) \quad = \quad P(x\text{th trial is a success}) = p$$

- $A$ and $B$ are independent; therefore, we have

$$P(X = x) = P(A \cap B) = P(A)P(B) = \binom{x-1}{k-1} p^{x-1}(1-p)^{x-k} \cdot p.$$

**EXAMPLE 13**

- 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 $X =$ number of rolls to get the 6th number 6. $X \sim NB(6, 1/6)$.
- Using the p.f. of negative binomial distribution:

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

**L–example 4.6** In an NBA championship series, the team that **wins four games out of seven is the winner**. Suppose that teams A and B face each other in the championship games and that **team A has probability 0.55 of winning a game over team B**.

(a) What is the probability that team A will win the series in 6 games?

(b) What is the probability that team A will win the series?

Solution: Suppose that Teams A and B can continuously play games. Let

$$X = \text{number of games that A needs to win 4 games}$$

and for each game, the chance that A will win is 0.55. Therefore $X \sim NB(4, 0.55)$.

(a) The question is asking

$$P(X = 6) = \binom{6-1}{4-1} 0.55^4 (1-0.55)^{6-4} = 0.1853.$$

(b) The probability that Team A will win is

$$\begin{aligned} & P(X=4) + P(X=5) + P(X=6) + P(X=7) \\ = & \binom{4-1}{4-1} 0.55^4 (1-0.55)^{4-4} + \binom{5-1}{4-1} 0.55^4 (1-0.55)^{5-4} \\ & + \binom{6-1}{4-1} 0.55^4 (1-0.55)^{6-4} + \binom{7-1}{4-1} 0.55^4 (1-0.55)^{7-4} \\ = & 0.6083. \end{aligned}$$

Question: Can Part (b) be solved using binomial distribution instead?

For $X \sim NB(k, p)$, we can use an online `R` compiler:

- Browse to `https://rdrr.io/snippets/`
- Command:
  - `dnbinom(x-k, k, p)` computes $P(X = x)$;
  - `pnbinom(x-k, k, p)` computes $P(X \leq x)$;
  - `pnbinom(x-k, k, p, lower.tail = F)` computes $P(X > x)$.

### Geometric Distribution

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

**DEFINITION 14 (GEOMETRIC DISTRIBUTION)**

- *$X$ = number of i.i.d. Bernoulli(p) trials until the first success occurs; then $X$ follows a* ***geometric distribution****, denote by $X \sim G(p)$.*
- *The p.f. of $X$ is given by*

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

- *We have $E(X) = 1/p$ and $V(X) = (1-p)/p^2$.*

### L–example 4.7

- At a "busy time", a telephone exchange is very near capacity, so callers have difficulty placing their calls.
- It may be of interest to know the number of attempts necessary in order to make a connection.
- Suppose that we let $p = 0.05$ be the probability of connection during a busy time.
- We are interested in knowing the probability that 5 attempts are necessary for a successful call.

Solution:

- Let $X$ = number of attempts needed for the first successful call.

- Then $X \sim G(p)$ or $X \sim NB(1,p)$, where $p = 0.05$.
- We have

$$P(X=5) = (1-p)^{5-1}p = 0.95^4(0.05) = 0.0407.$$

**Poisson Distribution**

**DEFINITION 15 (POISSON RANDOM VARIABLE)**
*The **Poisson random variable** $X$ denotes the number of events occurring in a **fixed period of time or fixed region**.*

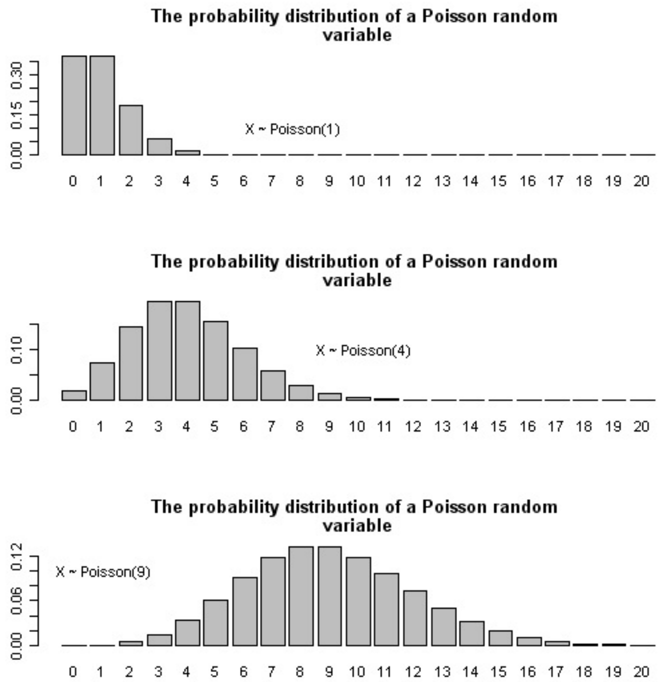
*We denote $X \sim \text{Poisson}(\lambda)$ where parameter $\lambda > 0$ is the expected number of occurrences during the given period/region; its p.m.f. is given by*

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

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

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

The p.f. for $\text{Poisson}(1)$, $\text{Poisson}(4)$, and $\text{Poisson}(9)$ are compared below.



**EXAMPLE 16**
The "fixed period of time or fixed region" given in the definition can be time period of any length, e.g., a minute, a day, a week, a month etc., and region of any size.

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.

**L–example 4.8** The number of infections $X$ in a hospital each week has been shown to follow a Poisson distribution with a mean of 3.0 infections per week. What is the probability that

(a) there is **no** infection for a week?

(b) there are **less than** 4 infections for a week?

Solution: It follows that

(a) $P(X=0)=e^{-3}$.

(b) $P(X<4)=e^{-3}\left(1+3+\frac{3^2}{2!}+\frac{3^3}{3!}\right)$.

Numerical computation for $X \sim \text{Poisson}(\lambda)$:

(A) Online `R` compiler: `https://rdrr.io/snippets/`

- `dpois(x, lambda)` computes $P(X=x)$;
- `ppois(x, lambda)` computes $P(X \le x)$;
- `ppois(x, lambda, lower.tail = F)` computes $P(X>x)$.

(B) use `R` Shiny app `Radiant`: `https://vnijs.shinyapps.io/radiant`; similar steps as Binomial distribution to do the computation.

**L–example 4.9**

- If the average number of oil tankers arriving each day at a port is known to be 10.
- The facilities at the port can handle at most 15 tankers per day.
- What is the probability that on a given day tankers will have to be sent away?

Solution:

- Let $X$ = number of tankers arriving each day.
- We have $X \sim \text{Poisson}(\lambda)$, where $\lambda = 10$.

$$\begin{aligned} P(X > 15) &= \sum_{x=16}^{\infty} \frac{e^{-10}10^x}{x!} = 1 - \sum_{x=0}^{15} \frac{e^{-10}10^x}{x!} \\ &= 1 - e^{-10}\left(1 + 10 + \frac{10^2}{2!} + \ldots + \frac{10^{15}}{15!}\right) \\ &= 0.0487. \end{aligned}$$

**L–example 4.10** We derive $E(X)$ and $V(X)$, for $X \sim \text{Poisson}(\lambda)$.

- For these derivation, the fundamental idea is to use the fact that for p.m.f. $f_X(x)$, we must have

$$\sum_{x \in R_X} f_X(x) = 1.$$

- We derive $E(X)$ first.

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x f_X(x) = \sum_{x=0}^{\infty} x \cdot \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-1)!} \\ &= \lambda \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \lambda, \quad \text{set } y = x-1. \end{aligned}$$

- We derive $V(X)$ next.

$$\begin{aligned} E(X(X-1)) &= \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=2}^{\infty} \frac{e^{-\lambda}\lambda^x}{(x-2)!} \\ &= \lambda^2 \sum_{y=0}^{\infty} \frac{e^{-\lambda}\lambda^y}{y!} = \lambda^2, \quad \text{set } y = x-2. \end{aligned}$$

We can compute $V(X)$ by

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 = E(X(X-1)) + E(X) - [E(X)]^2 \\ &= \lambda^2 + \lambda - \lambda^2 = \lambda. \end{aligned}$$

**DEFINITION 17 (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 $\alpha$ are*

- *the expected number of occurrences in an interval of length $T$ is $\alpha T$;*
- *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$(\alpha T)$ *distribution.*

**EXAMPLE 18**

- 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 $X_1$ = number of robberies in one day. Then $X_1 \sim \text{Poisson}(4)$ from the condition.
- Let $X$ = number of robberies in two days. Then $X \sim \text{Poisson}(2 \times 4) = \text{Poisson}(8)$.
- We have

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

**L–example 4.11**

A can company reports that the number of breakdowns per 8 hour shift on its machine-operated assembly line follows a Poisson distribution, with a mean of 1.5.

(a) What is the probability of exactly two breakdowns during the midnight shift?

(b) What is the probability of fewer than two breakdowns during the afternoon shift?

(c) What is the probability that no breakdown during three consecutive 8-hour shifts?

Solution: Let $X$ = number of breakdowns in an 8 hour shift. We have $X \sim$ Poisson($\lambda$) with $\lambda = 1.5$.

(a) The probability of exactly 2 breakdowns during the night shift is

$$P(X=2) = \frac{e^{-1.5}1.5^2}{2!} = 0.251.$$

(b) The probability of fewer than 2 breakdowns during the afternoon shift is

$$\begin{aligned} P(X<2) &= P(X=0)+P(X=1) \\ &= \frac{e^{-1.5}1.5^0}{0!} + \frac{e^{-1.5}1.5^1}{1!} = 0.5578. \end{aligned}$$

(c)
- Let $Y_1$ be a Bernoulli RV, where $Y_1 = 1$ if there is no breakdowns in the 1st 8 hour shift; and $Y_1 = 0$ otherwise. The probability of success is

$$p = P(Y_1 = 1) = P(X=0) = \frac{e^{-1.5}1.5^0}{0!} = 0.2231.$$

- Similarly define $Y_2$ and $Y_3$ as Bernoulli RVs, $Y_i = 1$ if no breakdown in the $i$th hour shift; and $Y_i = 0$ otherwise; for $i = 2,3$.
- Then $Y_1, Y_2, Y_3$ are i.i.d. Bernoulli($p$) RVs. Set $Y = Y_1 + Y_2 + Y_3$; then $Y \sim B(3, p)$. On the other hand $Y$ is counting the number of 8-hour shifts without breakdowns.
- "$Y = 3$" stands for the practical situation that no breakdown during three consecutive 8-hour shifts.

$$P(Y=3) = \binom{3}{3} p^3(1-p)^0 = 0.0111.$$

- An alternative method: using Poisson process, the number of breakdowns in $24 = 3 \times 8$ hours, denoted by RV $Z$, follows a Poisson($3 \times 1.5$) = Poisson($4.5$) distribution. The question is asking

$$P(Z=0) = \frac{e^{-4.5}4.5^0}{0!} = 0.0111.$$

**PROPOSITION 19 (POISSON APPROX. OF BINOMIAL DISTRIBUTION)**
*Let $X \sim B(n, p)$. 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*

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

The approximation is good when $n \geq 20$ and $p \leq 0.05$, or if $n \geq 100$ and $np \leq 10$.

**EXAMPLE 20**

- 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 $X =$ number of accidents among the 1000 cars.
- Then $X \sim B(1000, 0.0001)$. If we compute using binomial distribution,

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

- Computing these numbers is not easy.
- We solve the question using Poisson approximation.
- $n = 1000$ and $p = 0.0001$, hence, $np = \lambda = 0.1$.
- Thus

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

**L–example 4.12**

- In a manufacturing process in which glass items are being produced, defects or bubbles occur, occasionally rendering the piece undesirable for marketing.
- It is known that on the average 1 in every 1000 of these items produced has one or more bubbles.

- What is the probability that a random sample of 8000 will yield fewer than 7 items possessing bubbles?

Solution:

- Let $X =$ number of items processing bubbles.
- Then $X \sim B(8000, 0.001)$.
- Use the Poisson approximation, $\lambda = np = 8000 \times 0.001 = 8$, and hence $X \approx \text{Poisson}(\lambda)$.
- The (approximate) probability is

$$P(X < 7) = 1 - P(X \geq 7) \approx 1 - 0.6866 = 0.3134.$$

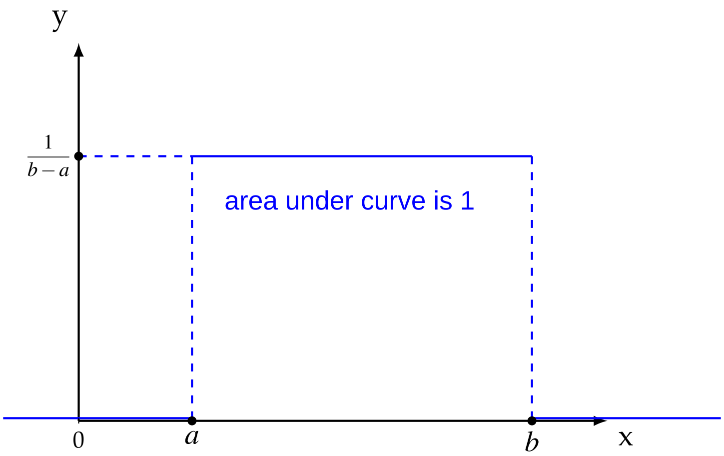
## 2 Continuous Distribution

- For a continuous random variable $X$, its range $R_X$ is an interval or a collection of multiple intervals.
- In this section, we study some classes of continuous random variables.

### Continuous Uniform Distribution

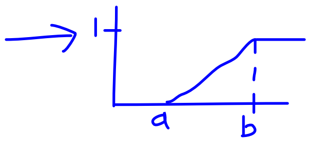
> **Definition 1 (Continuous Uniform Distribution)**
> *A random variable $X$ is said to follow a **uniform distribution** over the interval $(a, b)$ if its probability density function is given by*
>
> $$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & a \leq x \leq b \\ 0, & \textit{otherwise} \end{cases}.$$
>
> *We denote this by $X \sim U(a, b)$.*
>
> *It can be shown that $E(X) = \dfrac{a+b}{2}$ and $V(X) = \dfrac{(b-a)^2}{12}$.*

The p.d.f. for the continuous uniform distribution can be drawn as a figure below.




The c.d.f. for the continuous uniform distribution is given by

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




**EXAMPLE 2**

- A point is chosen at random on the line segment $[0,2]$.
- What is the probability that the chosen point lies between 1 and $3/2$?

Solution:

- Let $X =$ position of the point. $X \sim U(0,2)$.
- We have

$$f_X(x) = \frac{1}{2}, \quad \text{for } 0 \le x \le 2,$$

and 0 otherwise.

$$P\left(1 \le X \le \frac{3}{2}\right) = \int_1^{3/2} \frac{1}{2} dx = \frac{1}{2}\, x\Big|_1^{3/2} = 1/4.$$

**L–example 4.13**

- Buses arrive at a specified stop at 15-minute intervals starting at 7:00 am.
- That is, they arrive at 7:00, 7:15, 7:30, 7:45, and so on.

- If a passenger arrives at the stop at a time that is uniformly distributed between 7:00 and 7:30, find the probability that he waits less than 5 minutes for a bus.

<u>Solution</u>: Let $X$ denote the arrival time of the passenger (after 7:00am, in minutes). Then $X \sim U(0,30)$.

The passenger waits less than 5 minutes for a bus when and only when he arrives (a) between 7:10-7:15 or (b) 7:25-7:30. So the desired probability is

$$P(10 < X < 15) + P(25 < X < 30) = \frac{15-10}{30} + \frac{30-25}{30} = \frac{1}{3}.$$

**L–example 4.14** For the continuous uniform distribution, we derive

$$E(X) = \frac{a+b}{2}; \quad V(X) = \frac{1}{12}(b-a)^2.$$

- We derive $E(X)$ first

$$\begin{aligned} E(X) &= \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \left. \frac{x^2}{2} \right|_a^b \\ &= \frac{1}{b-a} \cdot \frac{b^2 - a^2}{2} = \frac{a+b}{2}. \end{aligned}$$

- We derive $V(X)$ next,

$$\begin{aligned} E(X^2) &= \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \left. \frac{x^3}{3} \right|_a^b \\ &= \frac{1}{b-a} \cdot \frac{b^3 - a^3}{3} = \frac{a^2 + ab + b^2}{3}. \end{aligned}$$

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{1}{12}(a^2 - 2ab + b^2) = \frac{(b-a)^2}{12}. \end{aligned}$$

**L–example 4.15** We derive the c.d.f. of a continuous uniform distribution.

- We take note that $F_X(x) = 0$ when $x < a$, and $F_X(x) = 1$ when $x > b$.
- When $a \le x \le b$,

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(t) dt = \int_{-\infty}^a 0\, dt + \int_a^x \frac{1}{b-a} dt \\ &= \frac{1}{b-a} \cdot t|_a^x = \frac{x-a}{b-a}. \end{aligned}$$

**Exponential Distribution**

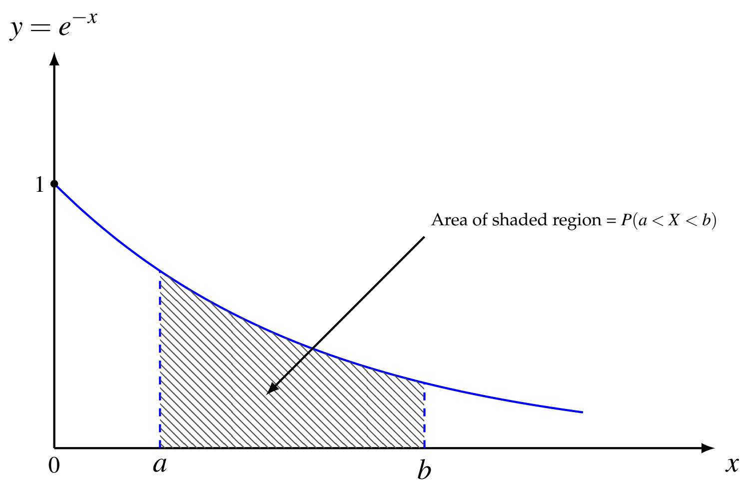
**Definition 3 (Exponential Distribution)**
*A continuous RV $X$ is said to follow an **exponential distribution** with parameter $\lambda > 0$ if its p.d.f. is given by*

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

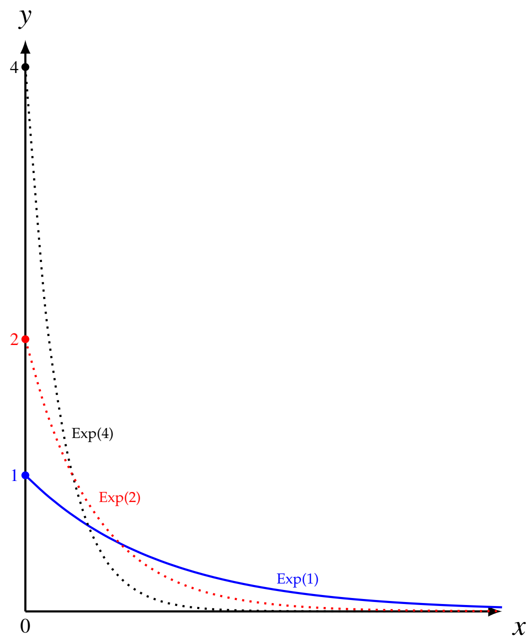
*We denote $X \sim \mathrm{Exp}(\lambda)$.*

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

The exponential p.d.f. with $\lambda = 1$ is shown below.




The shapes of the p.d.f.s of $\mathrm{Exp}(\lambda)$ for $\lambda = 1, 2, 4$.




The c.d.f. of $X \sim \text{Exp}(\lambda)$ is given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

**REMARK:**

- The p.d.f. can be written in the form

$$f_X(x) = \frac{1}{\mu} e^{-x/\mu}, \quad \text{for } x > 0,$$

  and 0 elsewhere.

- The parameters have the relationship $\mu = 1/\lambda$.

- We have

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

**EXAMPLE 4**

- 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 have $T \sim \text{Exp}(1/5)$,

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

- Let $X$ = # of systems out of 5 that are still functioning after 8 years.
- Then $X \sim B(5, 0.2)$. Hence,

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

**L–example 4.16** Let $X$ = response time at a certain on-line computer terminal (the elapsed time between the end of a user's inquiry and the beginning of the system's response to that inquiry). $X$ follows an exponential distribution with expected response time equal to 5 seconds.

(a) Find the probability that the response time is at most 10 seconds.

(b) Find the probability that the response time is between 5 and 10 seconds.

Solution: Since $E(X) = 5$, we have $X \sim \text{Exp}(1/5)$.

(a)

$$P(X \leq 10) = 1 - e^{-10/5} = 0.8647.$$

(b)

$$\begin{aligned} P(5 \leq X \leq 10) &= P(X \leq 10) - P(X < 5) \\ &= (1 - e^{-10/5}) - (1 - e^{-5/5}) = 0.2326. \end{aligned}$$

Numerical computation for $\text{Exp}(\lambda)$ distribution:

(A) Online `R` compiler: `https://rdrr.io/snippets/`

- `dexp(x, lambda)` computes $f_X(x)$;
- `pexp(x, lambda)` computes $P(X \leq x)$;
- `pexp(x, lambda, lower.tail = F)` computes $P(X > x)$.

(B) use `R` Shiny app `Radiant`: `https://vnijs.shinyapps.io/radiant`; similar steps as Binomial distribution to do the computation.

**L–example 4.17** We derive $E(X)$ and $V(X)$ for the exponential distribution.

$$\begin{aligned} E(X) &= \int_0^\infty x\lambda e^{-\lambda x}dx = \int_0^\infty xd\left(-e^{-\lambda x}\right) \\ &= -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty (-e^{-\lambda x})dx = \int_0^\infty \left(e^{-\lambda x}\right)dx \\ &= -\frac{1}{\lambda}e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}. \end{aligned}$$

$$\begin{aligned} E(X^2) &= \int_0^\infty x^2\lambda e^{-\lambda x}dx = \int_0^\infty x^2 d\left(-e^{-\lambda x}\right) \\ &= -x^2e^{-\lambda x}\Big|_0^\infty - \int_0^\infty \left(-e^{-\lambda x}\right)d(x^2) \\ &= \frac{2}{\lambda}\int_0^\infty x\lambda e^{-\lambda x}dx = \frac{2}{\lambda}\cdot\frac{1}{\lambda} = \frac{2}{\lambda^2}. \end{aligned}$$

Hence,

$$V(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

**L–example 4.18** Find the c.d.f. of the exponential distribution with parameter $\lambda$.
Solution:

- For $x \geq 0$,

  $$F_X(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda t}dt = -e^{-\lambda t}\Big|_0^x = 1 - e^{-\lambda x},$$

  and 0 otherwise.

- Also, we have

  $$P(X > x) = e^{-\lambda x}, \quad \text{for } x > 0.$$

**THEOREM 5**
*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**" in the sense:

- Let $X$ denote the life length of a bulb.
- Given that the bulb has lasted $s$ time units, i.e., $X > s$,
- the probability that it will last for the next $t$ units, i.e., $X > s+t$, is the same as the probability that it will last for the first $t$ units as brand new. ■

**L–example 4.19** We verify the no memory property of exponential distribution.

$$P(X > s+t | X > s) = \frac{P(\{X > s+t\} \cap \{X > s\})}{P(X > s)}$$
$$= \frac{P(X > s+t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t).$$

**Normal Distribution**

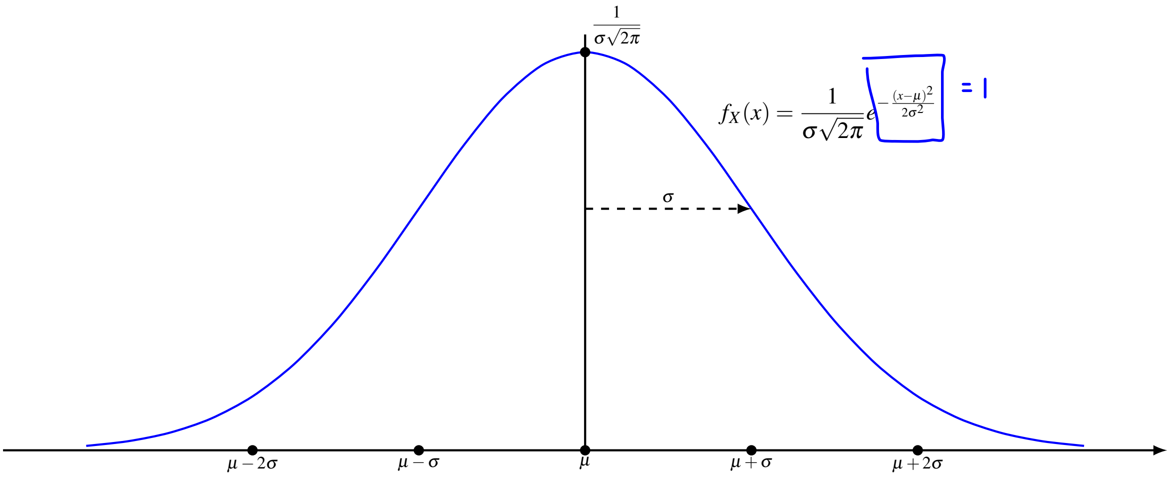
**DEFINITION 6 (NORMAL DISTRIBUTION)**
*A random variable $X$ is said to follow a* ***normal distribution*** *with parameters $\mu$ and $\sigma^2$ if its probability density function is given by*

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

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

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

The p.d.f. of normal distribution is positive over the whole real line, symmetric about $x = \mu$, and bell-shaped; see below.



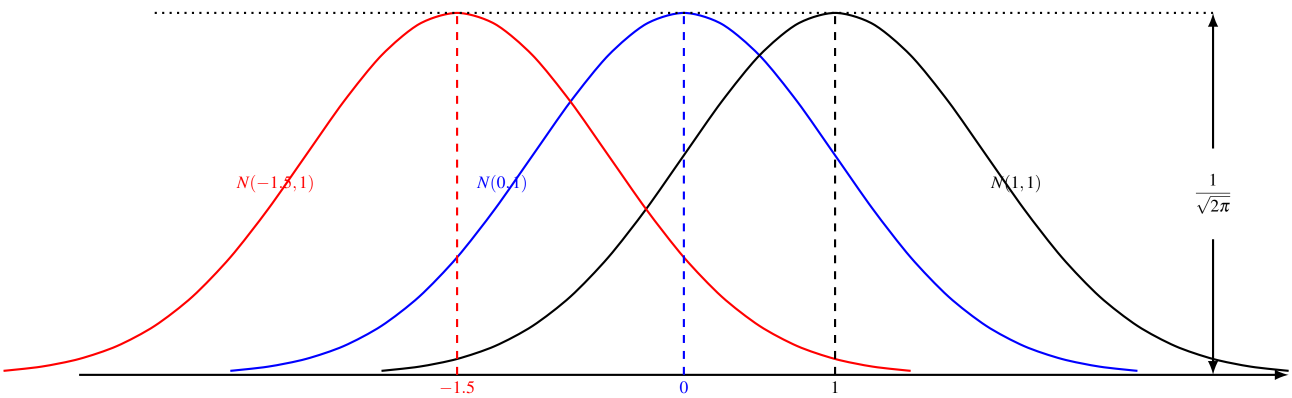

We give some properties of normal distribution.

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

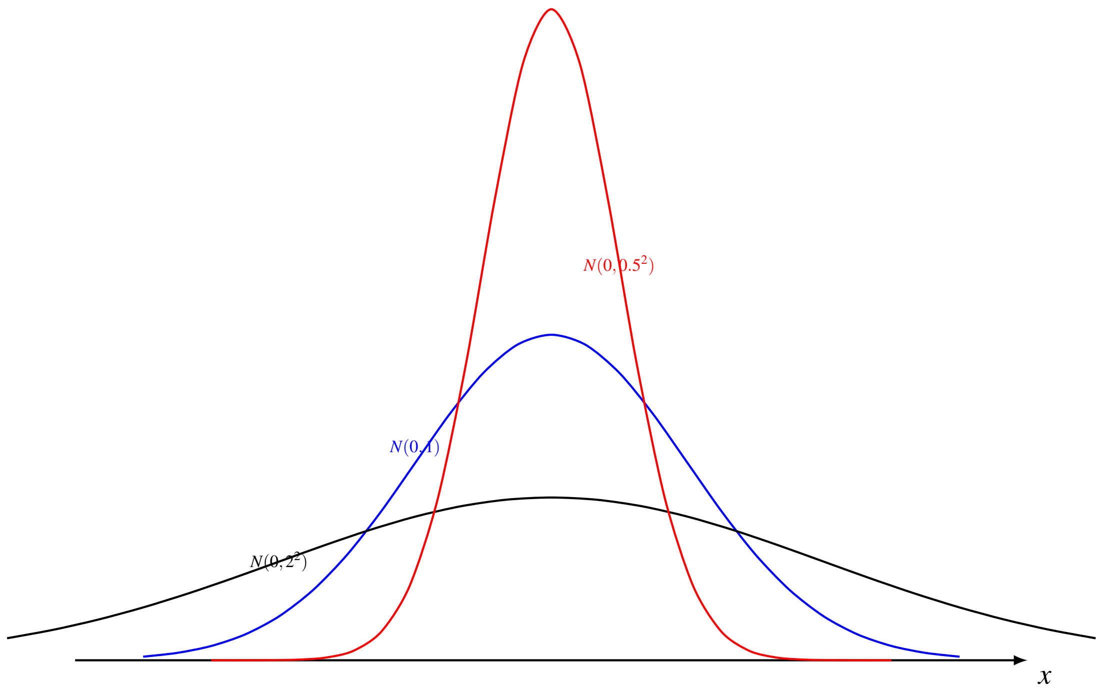
$$\int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = 1.$$

This validates that $f_X(\cdot)$ is a p.d.f.

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




(3) As $\sigma$ increases, the curve flattens; and vice versa.




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

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

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

We say that $Z$ has a standardized normal distribution; the p.d.f. of $Z$ is given by

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

**REMARK:**

- The importance of the standardized normal distribution is that it can be tabulated.
- Consider $X \sim N(\mu, \sigma^2)$; if we are to compute $P(x_1 < X < x_2)$ for any real values $x_1, x_2$, we can use the transformation $Z = (X - \mu)/\sigma$. In particular,

$$x_1 < X < x_2 \Longleftrightarrow \frac{x_1 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{x_2 - \mu}{\sigma}$$

- Let $z_1 = (x_1 - \mu)/\sigma$ and $z_2 = (x_2 - \mu)/\sigma$; 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 p.d.f. and c.d.f. of the standard normal distribution respectively. That is,

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

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

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

- However, calculating the probabilities for the normal probabilities is challenging because
  - there is no close formula for $\Phi(z)$;
  - so the computation relies on numerical integration.

- Instead, $\Phi(z)$ can be tabulated, or computed based on some statistical software.

- The standard normal distribution has the following properties:
  - ⋆ $P(Z \geq 0) = P(Z \leq 0) = \Phi(0) = 0.5$;
  - ⋆ For any $z$, $\Phi(z) = P(Z \leq z) = P(Z \geq -z) = 1 - \Phi(-z)$;
  - ⋆ $-Z \sim N(0, 1)$;
  - ⋆ If $Z \sim N(0, 1)$, then $\sigma Z + \mu \sim N(\mu, \sigma^2)$. ■

**Example 7**
Given $X \sim N(50, 100)$, find $P(45 < X < 62)$.

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

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

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

**L–example 4.20** When $X \sim N(65, 25)$, compute $P(47.5 < X \leq 80)$.

Solution: we have $\mu = 65$, $\sigma = 5$;

$$\begin{aligned} P(47.5 < X \le 80) &= P\left(\frac{47.5-65}{5} < \frac{X-65}{5} \le \frac{80-65}{5}\right) \\ &= P(-3.5 < Z \le 3) \\ &= P(Z \le 3) - P(Z \le -3.5) \\ &= P(Z \le 3) - P(Z \ge 3.5) \\ &= P(Z \le 3) - (1 - P(Z < 3.5)) \\ &= 0.99865 - 1 + 0.999767 = 0.998417. \end{aligned}$$

Numerical computation for $X \sim N(\mu, \sigma^2)$:

(A) Online `R` compiler: `https://rdrr.io/snippets/`

- `dnorm(x, mu, sigma)` computes $f_X(x)$;
- `pnorm(x, mu, sigma)` computes $P(X \le x)$;
- `pnorm(x, mu, sigma, lower.tail = F)` computes $P(X > x)$.

(B) use `R` Shiny app `Radiant`: `https://vnijs.shinyapps.io/radiant`; similar steps as Binomial distribution to do the computation.

**L–example 4.21**

- An expert witnesses in a paternity suit testifies that the length (in days) of pregnancy is approximately normally distributed with parameters $\mu = 270$ and $\sigma = 10$.
- The defendant in the suit is able to prove that he was out of the country during a period that began 290 days before the birth of the child and ended 240 days before the birth.
- If the defendant was, in fact, the father of the child, what is the probability that the mother could have had a very long or a very short pregnancy indicated by the testimony?

Solution: Let $X$ denote the length of the pregnancy and assume that the defendant is the father; then $X \sim N(270, 10^2)$. The probability of the birth could occur within the indicated duration is

$$\begin{aligned} & P(X > 290 \text{ or } X < 240) \\ = \; & P(X > 290) + P(X < 240) \\ = \; & P\left(\frac{X-270}{10} > \frac{290-270}{10}\right) + P\left(\frac{X-270}{10} < \frac{240-270}{10}\right) \\ = \; & 1 - \Phi(2) + \Phi(-3) \\ = \; & 1 - \Phi(2) + [1 - \Phi(3)] = 0.0241. \end{aligned}$$

**DEFINITION 8 (QUANTILE)**
*The $\alpha$th (upper) quantile ($0 \leq \alpha \leq 1$) of the RV $X$ is the number $x_\alpha$ that satisfies*

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

- Specifically, we denote by $z_\alpha$ the $\alpha$th (upper) quantile (or $100\alpha$ percentage point) of $Z \sim N(0,1)$. That is

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

- For example, $z_{0.05} = 1.645$, $z_{0.01} = 2.326$.
- Since the p.d.f. of $Z$, i.e., $\phi(z)$, is symmetrical about 0, therefore

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

**EXAMPLE 9**
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

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

which implies $P(Z > z) = 0.01$; therefore $z = z_{0.01} = 2.326$.

**L–example 4.22**

- On a common test, the average grade was 74 and the standard deviation was 7. Suppose that the grades are given as integers.
- If 12% of the class are given A's, and the grades are assumed to follow a normal distribution,

- what is the lowest possible A and the highest possible B?

Solution:

- We want to find $x$ such that $P(X > x) = 0.12$.

$$P(X > x) = P\left(Z > \frac{x - 74}{7}\right) = 0.12,$$

  where $Z = (X - 74)/7$.

- On the other hand, using a statistical software, $P(Z > z) = 0.12$ implies $z = 1.175$.
- By setting $(x - 74)/7 = 1.175$, we obtain

$$x = 74 + (1.175)7 = 82.225.$$

- Hence, the lowest possible A is 83 and the highest possible B is 82.

Compute the $\alpha$th (upper) quantile of $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$:

(A) Online `R` compiler: `https://rdrr.io/snippets/`

- `qnorm(alpha, mu, sigma, lower.tail = False)` computes $x_\alpha$;
- `qnorm(alpha, mu, sigma)` computes $x_{1-\alpha}$;
- `qnorm(alpha, lower.tail = False)` computes $z_\alpha$;
- `qnorm(alpha)` computes $z_{1-\alpha}$.

(B) use `R` Shiny app `Radiant`: `https://vnijs.shinyapps.io/radiant`.

**L–example 4.23**

- Let $X$ = the amount of sugar which a filling machine puts into "500g" packets.
- The actual amount of sugar filled varies from packets to packets.
- Suppose $X \sim N(\mu, 4^2)$.
- If only 2% of the packets contain less than 500g of sugar.
- What is the actual mean fill of these packets?

Solution: We need

$$0.02 = P(X < 500) = P\left(Z < \frac{500-\mu}{4}\right) = PP\left(Z > -\frac{500-\mu}{4}\right),$$

where $Z = (X-\mu)/4$.
On the other hand, from a statistical software, we have $P(Z > 2.0537) = 0.02$. Therefore

$$-\frac{500-\mu}{4} = 2.0537,$$

which leads to $\mu = 508.2$. That is, the mean fill should be 508.2g.

**L–example 4.24** The width of a slot of a duralumin in forging is (in inches) normally distributed with $\mu = 0.9000$ and $\sigma = 0.0030$. The specification limits were given as $0.9000 \pm 0.0050$.

(a) What percentage of forgings will be defective?

(b) What is the maximum allowable value of $\sigma$ that will permit no more than 1 in 100 defectives when the widths are normally distributed with $\mu = 0.9000$ and $\sigma$?

Solution:

(a) Let $X$ be the width of our normally distributed slot. The probability that a forging is acceptable is given by

$$\begin{aligned} P(0.895 < X < 0.905) &= P\left(\frac{0.895-0.9}{0.003} < Z < \frac{0.905-0.9}{0.003}\right) \\ &= P(-1.67 < Z < 1.67) \\ &= 2\Phi(1.67) - 1 = 0.905. \end{aligned}$$

So that the probability that a forging is defective is $1 - 0.905 = 0.095$. Thus 9.5 percent of forgings are defective.

(b) We need to find the value of $\sigma$ such that

$$P(0.895 < X < 0.905) \geq \frac{99}{100}.$$

Now

$$P(0.895 < X < 0.905) = \ldots = 2P\left(Z < \frac{0.005}{\sigma}\right) - 1.$$

We thus have to solve for $\sigma$ so that

$$2P\left(Z < \frac{0.005}{\sigma}\right) - 1 \geq 0.99.$$

or

$$P\left(Z < \frac{0.005}{\sigma}\right) \geq (1+0.99)/2 = 0.995.$$

From a statistical software, we have $P(Z \geq 2.576) = 0.005$ so we can use $\frac{0.005}{\sigma} \geq 2.576$ which gives $\sigma \leq 0.0019$.

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

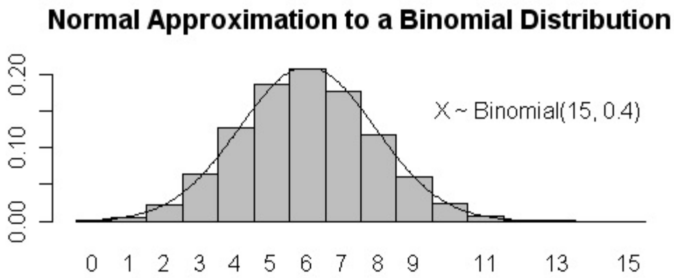
- A good rule of thumb is to use the normal approximation only when

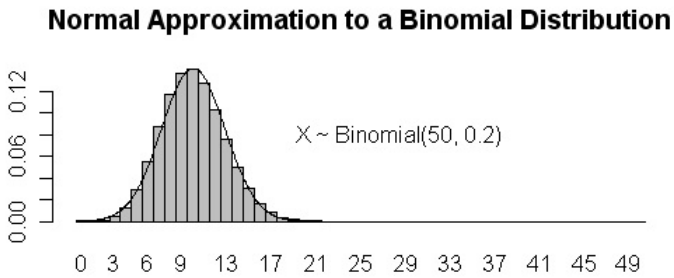
$$np > 5 \quad \text{and} \quad n(1-p) > 5.$$

**PROPOSITION 10 (NORMAL APPROX. TO BINOMIAL DISTRIBUTION)**
*Let $X \sim B(n, p)$; so that $E(X) = np$ and $V(X) = np(1-p)$. Then as $n \to \infty$,*

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

Normal Approximation to the Binomial Distribution




**L–example 4.25**

- If $X \sim B(15, 0.4)$, then

$$P(X=4) = \binom{15}{4} 0.4^4 (0.6)^{11} = 0.1268.$$

- By normal approximation, we may consider

$$Y \sim N(\mu, \sigma^2),$$

  with $\mu = np = 6$ and $\sigma^2 = npq = 3.6$.

Hence,

$$\begin{aligned} P(X=4) &= P(3.5 < X < 4.5) \approx P(3.5 < Y < 4.5) \\ &= P\left(\frac{3.5-6}{\sqrt{3.6}} < Z < \frac{4.5-6}{\sqrt{3.6}}\right) \\ &\approx P(-1.32 < Z < -0.79) \\ &= \Phi(-0.79) - \Phi(-1.32) \\ &= 0.1214. \end{aligned}$$

In this example, we have made the **continuity correction** to improve the approximation. In general, we have

(a) $P(X = k) \approx P\left(k - 1/2 < X < k + 1/2\right)$;

(b) $P(a \le X \le b) \approx P\left(a - 1/2 < X < b + 1/2\right)$;
$P(a < X \le b) \approx P\left(a + 1/2 < X < b + 1/2\right)$;
$P(a \le X < b) \approx P\left(a - 1/2 < X < b - 1/2\right)$;
$P(a < X < b) \approx P\left(a + 1/2 < X < b - 1/2\right)$.

(c) $P(X \le c) = P(0 \le X \le c) \approx P\left(-1/2 < X < c + 1/2\right)$.

(d) $P(X > c) = P(c < X \le n) \approx P\left(c + 1/2 < X < n + 1/2\right)$.

**L–example 4.26**

- A system is made up of 100 components, and each of which has a reliability equal to 0.90.
- These components function independently of one another, and the entire system functions only when at least 80 components function.
- What is the probability that the system functioning?

Solution:

- Let $X$ = number of components functioning.
- Then $X \sim B(100, 0.9)$.
- Thus $E(X) = (100)(0.9) = 90$ and $V(X) = (100)(0.9)(0.1) = 9$.
- The system is functioning if $80 \le X \le 100$,

$$\begin{aligned} P(80 \le X \le 100) &\approx P\left(\frac{79.5 - 90}{3} < \frac{X - 90}{3} < \frac{100.5 - 90}{3}\right) \\ &= P(-3.5 < Z < 3.5) \\ &= \Phi(3.5) - \Phi(-3.5) = 0.9995. \end{aligned}$$

# *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, $p$, $\mu$ or $\sigma$) 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 $\mu$ and the standard deviation $\sigma$ 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 1 (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 2 (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. ■

**L-EXAMPLE 5.1 (SRS USING R)**
We want to choose a simple random sample of size 5 from a group of 20 mice, to be used in studying the growth rate of tumors in a cancer research experiment.

We can tag the mice with numbers 1 to 20 and enter the `R` command
```
sample(1:20, 5)
```
unto the online `R` compiler `https://rdrr.io/snippets/` to select a SRS of size 5 from the population $\{1, 2, \ldots, 20\}$.

**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.*

**L-EXAMPLE 5.2**

- Take note that we are sampling from an infinite population if we sample with replacement from a finite population, and the sample is random if
  (1) in each draw, every element in the population has the **same probability of being selected**, and
  (2) successive draws are **independent**.
- Specifically, let's consider the population of the sums of all possible rolls of a pair of dice.
- For each roll, the outcome is finite. Let's denote the random $X$ as the sum of the two dice in a single roll. The p.m.f. of $X$ is given by

| $x$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $f_X(x)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

- Conceptually, we can keep rolling the die for an arbitrary number of times. Let's denote the rolling outcome as $X_1, X_2, X_2, \ldots$
- Therefore, the population is considered to be infinite.

## 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 1 (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 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 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 4 (Sampling Distribution)**
*The probability distribution of a statistic is called a* ***sampling distribution****.*

**L-Example 5.3**
We look at an example of the sampling distribution of the sample mean. Consider a discrete uniform population consisting of the values

$$\{3, 5, 7, 9, 11\}.$$

The population size is $N = 5$.

Note that

$$f_X(x) = 1/5, \quad \text{for } x = 3, 5, 7, 9, 11.$$

The population mean and variance can be computed to be

$$\mu_X = E(X) = 7 \text{ and } \sigma_X^2 = \text{var}(X) = 8.$$

Suppose we list all possible samples of size 2 with replacement, and then for each sample we compute $\overline{X}$. There are $5^2 = 25$ possible distinct samples and their means are as follows:

| Sample | $\overline{X}$ | Sample | $\overline{X}$ | Sample | $\overline{X}$ | Sample | $\overline{X}$ | Sample | $\overline{X}$ |
|---|---|---|---|---|---|---|---|---|---|
| (3, 3) | 3 | (5, 3) | 4 | (7, 3) | 5 | (9, 3) | 6 | (11, 3) | 7 |
| (3, 5) | 4 | (5, 5) | 5 | (7, 5) | 6 | (9, 5) | 7 | (11, 5) | 8 |
| (3, 7) | 5 | (5, 7) | 6 | (7, 7) | 7 | (9, 7) | 8 | (11, 7) | 9 |
| (3, 9) | 6 | (5, 9) | 7 | (7, 9) | 8 | (9, 9) | 9 | (11, 9) | 10 |
| (3,11) | 7 | (5,11) | 8 | (7,11) | 9 | (9,11) | 10 | (11,11) | 11 |

The sampling distribution of $\overline{X}$ is now found to be:

| $\bar{x}$ | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|
| $P(\overline{X} = \bar{x})$ | 1/25 | 2/25 | 3/25 | 4/25 | 5/25 | 4/25 | 3/25 | 2/25 | 1/25 |

Similarly, we can compute the mean and variance of $\overline{X}$ to be

$$\mu_{\overline{X}} = E(\overline{X}) = 7 \text{ and } \sigma^2_{\overline{X}} = Var(\overline{X}) = 4.$$

Thus we have

$$\mu_{\overline{X}} = \mu_X \text{ and } \sigma^2_{\overline{X}} = \frac{\sigma^2_X}{2},$$

where 2 is the sample size.

**Two Results**

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

- Theorem 5 provides formulas for the center and the spread of the sampling distribution.
- Theorem 1 describes the shape of the sampling distribution, showing that it is often approximately normal.

**THEOREM 5 (MEAN AND VARIANCE OF $\overline{X}$)**
*For random samples of size n taken from an infinite population with mean $\mu_X$ and variance $\sigma^2_X$, the sampling distribution of the sample mean $\overline{X}$ has mean $\mu_X$ and variance $\frac{\sigma^2_X}{n}$. That is,*

$$\mu_{\overline{X}} = E\left(\overline{X}\right) = \mu_X \quad \textit{and} \quad \sigma^2_{\overline{X}} = \mathrm{var}\left(\overline{X}\right) = \frac{\sigma^2_X}{n}.$$

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

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

**DEFINITION 6 (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 $\sigma$) and the terminology standard error (instead of standard deviation) distinguishes this measure from the standard deviation $\sigma$ of the population. ■

**L-EXAMPLE 5.4**
Let's derive the results:

$$E(\overline{X}) = \mu_X; \qquad \mathrm{var}(\overline{X}) = \frac{\sigma_X^2}{n}.$$

Based on the definition of $\overline{X}$,

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_X = \mu_X;$$

using the independence

$$\mathrm{var}(\overline{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma_X^2 = \frac{\sigma_X^2}{n}.$$

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

**THEOREM 7 (LAW OF LARGE NUMBERS (LLN))**
*If $X_1, \ldots, X_n$ are independent random variables with the same mean $\mu$ and variance $\sigma^2$, 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 $\mu_X$, 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 1 (CENTRAL LIMIT THEOREM (CLT))**
*If $\overline{X}$ is the mean of a random sample of size $n$ taken from a population having mean $\mu$ and finite variance $\sigma^2$, then, as $n \to \infty$,*

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

*or equivalently*

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

**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 $\chi^2$, 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 2 (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

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

**L-EXAMPLE 5.5**
For CLT, we make the following remarks.

- The convergence in CLT is "**convergence in distribution**", or more rigorously, for any $x$,

$$\lim_{n\to\infty} P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x),$$

  where we recall that $\Phi(x)$ denotes the c.d.f. of $N(0,1)$.

- So, for finite but large sample size $n$, we say $\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ **approximately** follows a standard normal distribution.

- However, if $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$, then

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{or} \quad \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

  **exactly**, regardless of the sample size $n$.

**L-EXAMPLE 5.6 (CALCULATING NORMAL PROBABILITIES)**
In an earlier example, we calculated the probability

$$P(Z > 1.18) = 0.119.$$

We demonstrate two ways how this can be obtained using software.

(A) Using an online `R` compiler.

- Browse to `https://rdrr.io/snippets/`
- Enter the command

  `pnorm(1.18, lower.tail=FALSE)`

  unto the compiler.
- `Ctrl-Enter` or `Run` to obtain the answer.
- By default, `pnorm(y)` gives the probability $P(Z < y)$. The argument `lower.tail=FALSE` then gives us $P(Z > y)$.

(B) Using the `R` Shiny app `Radiant`.

- Browse to `https://vnijs.shinyapps.io/radiant`
- Select `Basics > Probability Calculator`.
- Select `Normal` as the `Distribution`.
- Select `Values` as the `Input type`.
- Enter `1.18` as `Lower bound`.
- The probability $P(X > 1.18)$ appears as one of the answers.


**L-EXAMPLE 5.7 (A CLT APP)**
The following application, written using `R` Shiny, illustrates how the sampling distribution of the sample mean is approximately normal, as the sample size gets larger.

https://david-chew.shinyapps.io/CLT4means/

**L-EXAMPLE 5.8 (NICOTINE CONTENT)**
The nicotine content in a single cigarette of a particular brand is a random variable with mean $\mu = 0.8$ mg and standard deviation $\sigma = 0.1$ mg.

If an individual smokes five packs (20 cigarettes per pack) of these cigarettes per week, what is the probability that the total amount of nicotine consumed in a week is at least 82 mg?

***Solution:***
Note that 5 packs consist of 100 cigarettes. Let $X_i$, $i = 1, \ldots, 100$ denote the nicotine contents of the 100 cigarettes.

Then the $X_i$'s form a random sample from a distribution with mean $\mu = 0.8$ and standard deviation $\sigma = 0.1$.

We apply the CLT to get, approximately

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

or equivalently

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(0.8, \frac{0.1^2}{100}\right).$$

The required probability is then given as

$$\begin{aligned} P\left(\sum_{i=1}^{100} X_i \geq 82\right) &= P\left(\overline{X} \geq 0.82\right) \\ &\approx P\left(Z \geq \frac{0.82 - 0.8}{0.01}\right) = P(Z \geq 2) = 0.0228. \end{aligned}$$

**L-EXAMPLE 5.9 (CHEMICAL IMPURITY)**

- When a batch of a certain chemical product is prepared, the amount of a particular impurity in the batch is a random variable with mean value 4.0 g and standard deviation 1.5 g.
- If 50 batches are independently prepared, what is the approximate probability that the sample average amount of impurity is between 3.5 g and 3.8 g?

***Solution:***
Since $n$ is large, we apply the Central Limit Theorem and we have

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(4, \frac{1.5^2}{50}\right).$$

Hence

$$
\begin{aligned}
P\left(3.5 \leq \overline{X} \leq 3.8\right) &\approx P\left(\frac{3.5-4.0}{1.5/\sqrt{50}} \leq Z \leq \frac{3.8-4.0}{1.5/\sqrt{50}}\right) \\
&= P(-2.357 \leq Z \leq -0.943) = 0.1636.
\end{aligned}
$$

## 5 Other Sampling Distributions

We next describe the $\chi^2$, $t$, 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 1 (The $\chi^2$ Distribution)**
*Let $Z$ be a standard normal random variable. A random variable with the same distribution as $Z^2$ is called a* ***$\chi^2$ 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* ***$\chi^2$ random variable with $n$ degrees of freedom****.*

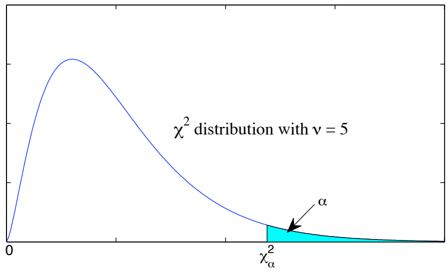
**Remark:**
We denote a $\chi^2$ random variable with $n$ degrees of freedom as $\chi^2(n)$. ■

**Properties of $\chi^2$ Distributions**

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

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

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




**L-EXAMPLE 5.10**

- Based on the definition, $Y \sim \chi^2(n)$ **if and only if** we have i.i.d. $Z_1, \ldots, Z_n$ standard normal random variable, such that

$$Y = Z_1^2 + Z_2^2 + \ldots + Z_n^2.$$

  This is useful when we derive some properties of the $\chi^2$ distribution. Properties 1 and 3 are resulted from this definition.

- For i.i.d. $N(\mu, \sigma^2)$ RVs $X_1, X_2, \ldots, X_n$, if we define

$$Y = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2},$$

  then $Y \sim \chi^2(n)$.

**L-EXAMPLE 5.11 (COMPUTATIONS WITH THE $\chi^2$-DISTRIBUTION)**
We show how you can use software to compute $\chi^2$ probabilities and obtain $\chi^2(n;\alpha)$ values.

Suppose $Y \sim \chi^2(5)$, and we are interested to compute/obtain

1. $P(Y > 2)$
2. $P(1 < Y < 2)$
3. $\chi^2(5; 0.05)$

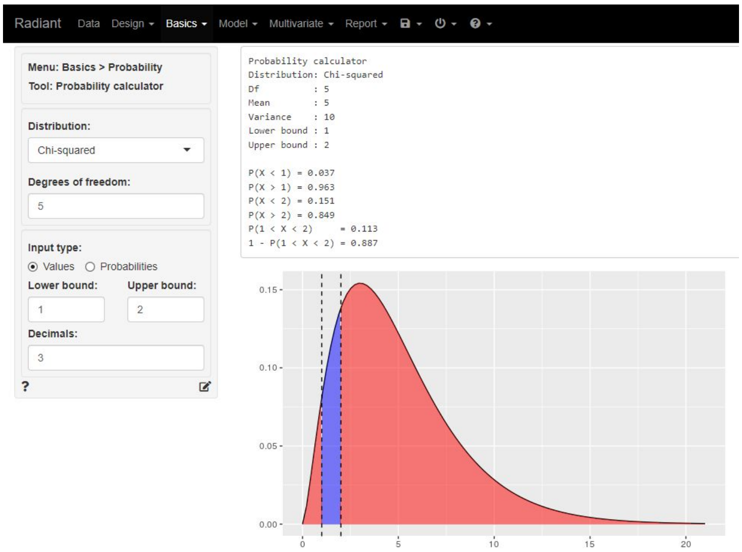
Here's how you can do that using the `R` Shiny app `Radiant`.

- Browse to `https://vnijs.shinyapps.io/radiant`
- Select `Basics > Probability Calculator`.
- Select `Chi-squared` as the `Distribution`.

- Enter `5` as the `Degrees of freedom`.

To obtain the probabilities,

- Select `Values` as the `Input type`.
- Enter `1` as `Lower bound`, `2` as `Upper bound`.
- The output shows that $P(X > 2) = 0.849$, $P(1 < X < 2) = 0.113$ amongst others.




To obtain $\chi^2(5;0.05)$,

- Select `Probabilities` as the `Input type`.
- Enter `0.95` as `Upper bound`.
  Note here that $0.95 = 1 - 0.05$.
- The output shows that $P(X > 11.07) = 0.05$. Thus we conclude that $\chi^2(5;0.05) = 11.07$.


### 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 $\text{var}(X) = \sigma^2$, 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 $i$.

**L-EXAMPLE 5.12**

- The sample variance has an alternative formula based on the fact:

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

  for $X_1, X_2, \ldots, X_n$ being arbitrary real numbers.

- Based on this, we can quickly derive $E(S^2) = \sigma^2$.
- We can make the problem a little more general. Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. RVs with mean $\mu$ and variance $\sigma^2$.
- Denote $Y_i = X_i - \mu$, then $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E(Y_i) = 0$, $\text{var}(Y_i) = \sigma^2$; and since $\overline{Y} = \overline{X} - \mu$, we have $E(\overline{Y}) = 0$, $\text{var}(\overline{Y}) = \sigma^2/n$.

- Now that $X_i - \overline{X} = X_i - \mu - (\overline{X} - \mu) = Y_i - \overline{Y}$, we have

$$\begin{aligned} E\left\{\sum_{i=1}^{n}(X_i - \overline{X})^2\right\} &= E\left\{\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right\} = E\left\{\sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2\right\} \\ &= \sum_{i=1}^{n} E(Y_i^2) - nE(\overline{Y}^2) = \sum_{i=1}^{n} \mathrm{var}(Y_i) - n\,\mathrm{var}(\overline{Y}) \\ &= n\sigma^2 - n\sigma^2/n = (n-1)\sigma^2, \end{aligned}$$

which immediately implies $E(S^2) = \sigma^2$.

**THEOREM 3**
*If $S^2$ is the variance of a random sample of size $n$ taken from a normal population having the variance $\sigma^2$, 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 $\chi^2$ distribution with $n-1$ degrees of freedom.*

**L-EXAMPLE 5.13**
Suppose 6 random samples are drawn from a normal population $N(\mu, 4)$. Define the sample variance

$$S^2 = \frac{1}{5}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Find $c$ such that $P(S^2 > c) = 0.05$.

***Solution:***
We know that $\frac{5S^2}{4} \sim \chi^2(5)$. Hence,

$$\begin{aligned} & P(S^2 > c) = 0.05 \\ \Leftrightarrow\ & P(5S^2/4 > 5c/4) = 0.05 \\ \Leftrightarrow\ & 5c/4 = \chi^2(5; 0.05) = 11.07 \\ \Leftrightarrow\ & c = 8.86. \end{aligned}$$

**Definition 4 (The $t$-Distribution)**
*Suppose $Z \sim N(0,1)$ and $U \sim \chi^2(n)$. If $Z$ and $U$ are independent, then*

$$T = \frac{Z}{\sqrt{U/n}}$$

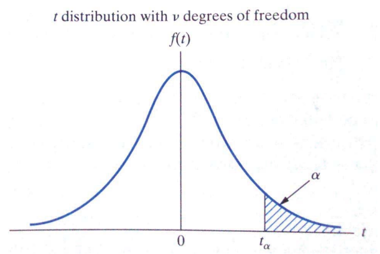
*follows the **$t$-distribution with $n$ degrees of freedom**.*

**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 $\text{var}(T) = n/(n-2)$ 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 5**
*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 6**
*If $X_1, \ldots, X_n$ are independent and identically distributed normal random variables with mean $\mu$ and variance $\sigma^2$, then*

$$\frac{X - \mu}{S/\sqrt{n}}$$

*follows a t-distribution with $n - 1$ degrees of freedom.*

**EXAMPLE 7 (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 $\overline{x} = 22$ 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?

**L-EXAMPLE 5.14 (COMPUTATIONS WITH THE $t$-DISTRIBUTION)**
We show how you can use software to compute $t$ probabilities and obtain $t(n;\alpha)$ values.

Suppose $Y \sim t(4)$, and we are interested to compute/obtain

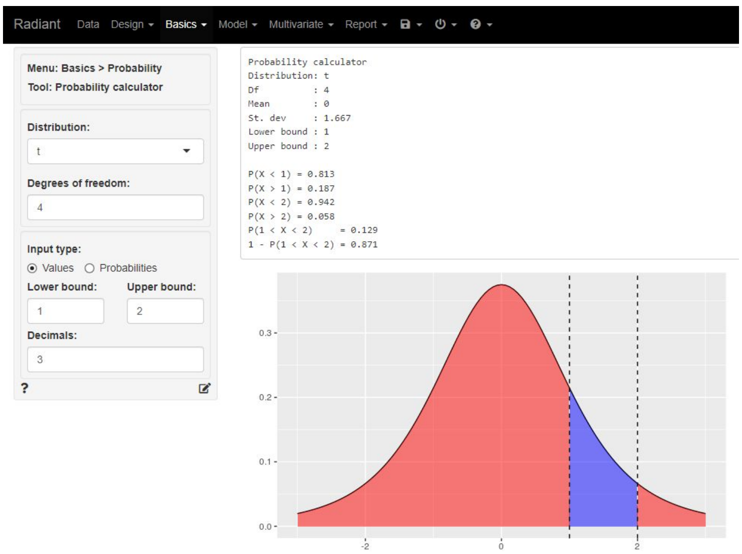
1. $P(Y > 2)$
2. $P(1 < Y < 2)$
3. $t(4; 0.005)$

Here's how you can do that using the `R` Shiny app `Radiant`.

- Browse to `https://vnijs.shinyapps.io/radiant`
- Select `Basics > Probability Calculator`.
- Select `t` as the `Distribution`.
- Enter `4` as the `Degrees of freedom`.

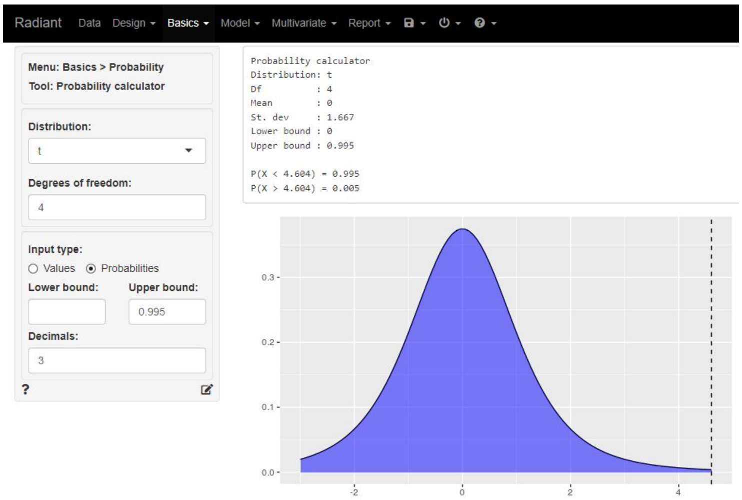
To obtain the probabilities,

- Select `Values` as the `Input type`.
- Enter `1` as `Lower bound`, `2` as `Upper bound`.
- The output shows that $P(X > 2) = 0.058$, $P(1 < X < 2) = 0.129$ amongst others.




To obtain $t(4; 0.005)$,

- Select `Probabilities` as the `Input type`.
- Enter `0.995` as `Upper bound`.
  Note here that $0.995 = 1 - 0.005$.
- The output shows that $P(X > 4.604) = 0.005$. Thus we conclude that $t(4; 0.005) = 4.604$.




**L-EXAMPLE 5.15**

A manufacturer of light bulbs claims that his light bulbs will burn on the average $\mu = 500$ hours. To maintain this average, he tests 25 bulbs each month.

If the computed $t$ value, $\frac{\bar{x}-\mu}{s/\sqrt{n}}$, falls between $-t_{24;0.05}$ and $t_{24;0.05}$, he is satisfied with his claim.

What conclusion should be drawn from a sample that has a mean ≡518 hours and a standard deviation $s = 40$ hours? Assume that the distribution of burning times in hours is approximately normal.

***Solution:***

From the $t$-table or software, $t_{24;0.05} = 1.711$.

Therefore, the manufacturer is satisfied with his claim if a sample of 25 bulbs yields a $t$-value between $-1.711$ and $1.711$.

If $\mu = 500$, then

$$t = \frac{518 - 500}{40/5} = 2.25 > 1.711.$$

Note that if $\mu > 500$, then the value of $t$ computed from the sample would be more reasonable. Hence the manufacturer is likely to conclude that his bulbs are a better product than he thought.

**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 8 (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*

$$\mathrm{var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}, \quad \text{for } n > 4.$$

- *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*

$$P(F > F(n,m;\alpha)) = \alpha,$$

*where $F \sim F(m,n)$.*

- *It can be shown that*

$$F(m,n;1-\alpha) = 1/F(n,m;\alpha).$$

**EXAMPLE 9**
For example,

$$F(4,5;0.05) = 5.19$$

means that $P(F > 5.19) = 0.05$, where $F \sim F(4,5)$.

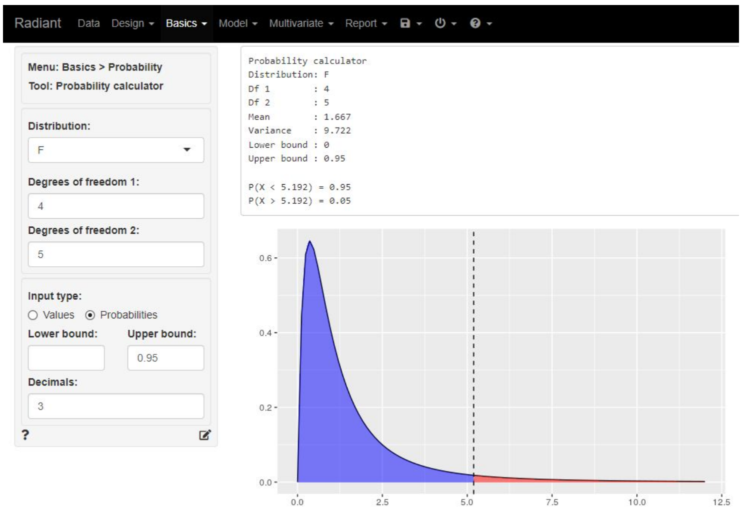
**L-EXAMPLE 5.16 (COMPUTATIONS WITH THE $F$-DISTRIBUTION)**
We show how you can use software to obtain $F(n,m;\alpha)$ values.

Suppose $Y \sim F(5,4)$, and we are interested to obtain

1. $F(4,5;0.05)$

Here's how you can do that using the `R` Shiny app `Radiant`.

- Browse to `https://vnijs.shinyapps.io/radiant`
- Select `Basics > Probability Calculator`.
- Select `F` as the `Distribution`.
- Enter `4,5` as the `Degrees of freedom 1/2` respectively.
- Select `Probabilities` as the `Input type`.
- Enter `0.95` as `Upper bound`.
  Note here that $0.95 = 1 - 0.05$.
- The output shows that $P(X > 5.192) = 0.05$. Thus we conclude that $F(4,5;0.05) = 5.192$.




**L-EXAMPLE 5.17**
Suppose that random samples of sizes $n_1$ and $n_2$ are selected from two normal populations with variances $\sigma_1^2$ and $\sigma_2^2$ respectively.

From an earlier section, we know that

$$U = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1)$$

and

$$V = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

are independent random variables.

Therefore we have

$$F = \frac{U/(n_1 - 1)}{V/(n_2 - 1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

**L-EXAMPLE 5.18**

Let $S_1^2$ and $S_2^2$ be the sample variances of independent random samples of sizes $n_1 = 25$ and $n_2 = 31$, taken from normal populations with variances $\sigma_1^2 = 10$ and $\sigma_2^2 = 15$ respectively. Find $P(S_1^2/S_2^2 > 1.26)$.

***Solution:***

Note that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1),$$

which gives

$$\frac{S_1^2/10}{S_2^2/15} \sim F(24, 30).$$

Thus

$$\begin{aligned} P\left(\frac{S_1^2}{S_2^2} > 1.26\right) &= P\left(\frac{S_1^2/10}{S_2^2/15} > 1.26 \times \frac{15}{10}\right) \\ &= P(F > 1.89) = 0.05. \end{aligned}$$

Note that here $F \sim F(24, 30)$.

# 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.

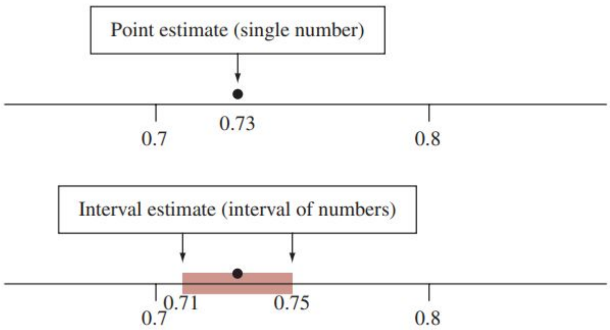
**Interval estimation**
: Based on sample data, two numbers are calculated to form an interval within which the parameter is expected to lie. ■

**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 $\mu$, 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 ($\mu$) 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 $\mu$. In this case, $\overline{X}$ is the estimator (for $\mu$), 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?*

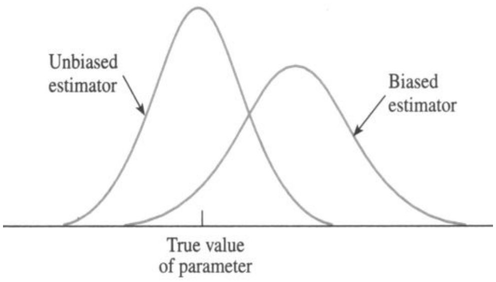
### Unbiased Estimator

One of the reasons we think $\overline{X}$ is a good estimator of $\mu$ is because $E(\overline{X}) = \mu$. That is, "on average", the estimator is right.

In general, we represent the parameter of interest by $\theta$. For example, $\theta$ can be $p, \mu$, or $\sigma$.

**DEFINITION 2 (UNBIASED ESTIMATOR)**
*Let $\hat{\Theta}$ be an estimator of $\theta$. Then $\hat{\Theta}$ is a random variable based on the sample. If $E(\hat{\Theta}) = \theta$, we call $\hat{\Theta}$ an* ***unbiased estimator*** *of $\theta$.*

**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 $\mu$ and variance $\sigma^2$. Then

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

is an unbiased estimator of $\sigma^2$ since $E(S^2) = \sigma^2$.

**L-EXAMPLE 6.1 (UNBIASED ESTIMATOR)**
A bus arrives at the bus stop according to a $U(0,\theta)$ distribution. The lecturer wants to estimate $\theta$. So this morning he randomly selected 4 students and asked their waiting time for a bus. The following values are obtained:

$$X_1, \ldots, X_4.$$

(a) Is $\overline{X}$ an unbiased estimator of $\theta$?

(b) Using (a), construct an unbiased estimator of $\theta$.

(c) Is there another unbiased estimator of $\theta$?

***Solution:***
We know that $X_1, \ldots, X_4$ have the same distribution as $X \sim U(0,\theta)$.

(a) No. This is because

$$E(\overline{X}) = E(X) = \int_0^\theta \frac{x}{\theta}\,\mathrm{dx} = \frac{\theta}{2} \neq \theta.$$

(b) $2\overline{X}$ is one. This is because $E(2\overline{X}) = 2E(\overline{X}) = \theta$.

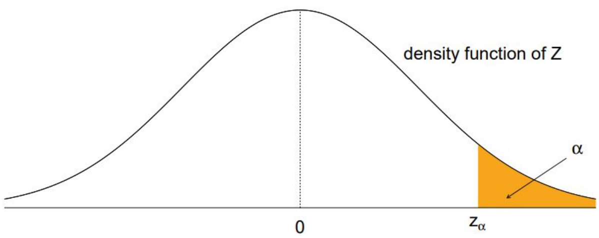
(c) Yes. $2X_1$ is another since $E(2X_1) = E(2X) = \theta$.

### 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 ($z_\alpha$)**
*Define $z_\alpha$ to be the number with an upper-tail probability of $\alpha$ for the standard normal distribution $Z$. That is, $P(Z > z_\alpha) = \alpha$.*

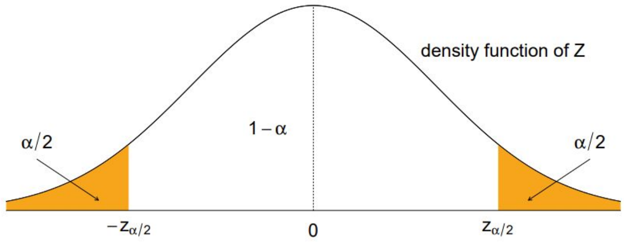


From the above definition, we then have

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

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 $\mu$.

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

$$n \geq \left( \frac{z_{\alpha/2} \cdot \sigma}{E_0} \right)^2.$$

**L-EXAMPLE 6.2 (TV TIME FOR INTERNET USERS II)**
What is the sample size $n$ required such that the television investigator can assert with 99% probability that his estimation error is at most 0.5 hour?

***Solution:***
The sample size required is

$$n = \left( \frac{2.576 \times 3.5}{0.5} \right)^2 \approx 325.$$

### 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 | $\sigma$ | $n$ | Statistic | $E$ | $n$ for desired $E_0$ and $\alpha$ |
|---|---|---|---|---|---|---|
| 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 | $Z = \frac{\overline{X}-\mu}{S/\sqrt{n}}$ | $z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$ | $\left(\frac{z_{\alpha/2} \cdot s}{E_0}\right)^2$ |

■

**L-EXAMPLE 6.3 (TV TIME FOR INTERNET USERS III)**
Back to the case where the investigator polls $n = 50$ internet users.

Suppose we do not trust the historical assumption that $\sigma = 3.5$. Instead, we estimate $\sigma$ using the sample standard deviation $s = 2.6$.

With 99% confidence, what is $E$, the maximum error of our estimate?

***Solution:***
We are in Case IV. So $E$ is given as

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 2.576 \cdot \frac{2.6}{\sqrt{50}} \approx 0.947.$$

## 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: $\sigma$ known, data normal**

Consider the case where $\sigma$ 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 | $\sigma$ | $n$ | 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 $\sigma$ 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

$$\bar{x} \pm t_{n-1;\alpha/2} \cdot \frac{s}{\sqrt{n}} = 10 \pm 2.447 \cdot \frac{\sqrt{0.08}}{\sqrt{7}} = (9.738, 10.262).$$

**L-EXAMPLE 6.4 (AGAIN, WHICH CASE?)**
A major department store chain is interested in estimating the average amount its credit card customers spent on their first visit to the chain's new store in the mall.

Fifty credit card accounts were randomly sampled and analyzed with the following results:

$$\bar{x} = \$62.56 \quad \text{and} \quad s = \$20.$$

(a) Identify the population the department store chain is interested in learning about.

(b) Which population parameter does the chain wish to estimate?

(c) Construct a 90% confidence interval for the parameter identified in the previous part.

***Solution:***

(a) The population is all its credit card customers.

(b) We are interested in $\mu$, the average amount credit card customers spent on their first visit to the chain's new store in the mall.

(c) As $n$ is large and $\sigma$ is unknown, we are in Case IV.

Note that $z_{0.05} = 1.645$. Thus the 90% confidence interval for $\mu$ is

$$x \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} = 62.56 \pm 1.645 \cdot \frac{20}{\sqrt{50}} = (57.907, 67.213).$$

**L-EXAMPLE 6.5 (BUYING PHONES ON EBAY)**
eBay is a popular Internet company for auctioning just about anything.

The following is a random sample of 11 completed auctions for an unlocked Apple iPhone 5s with 16GB storage in new condition (item not used, but original packaging might be missing), obtained from eBay in July 2014.

Closing Price (in $):
`570 620 610 590 540 590 565 590 580 570 595`

We are given that $\bar{x} = 583.64$, $s = 22.15$.

The retail price of of this phone is $649.

By constructing a 95% confidence interval for the mean closing price (of this phone) on eBay, find out how much can you save by buying items on eBay compared to their actual retail price. Assume an approximate normal distribution for the closing price.

***Solution:***
Let $\mu$ denote the population mean for the closing price of the auction.

The point estimate of $\mu$ is the sample mean $\bar{x} = 583.64$.

We are in Case III, and a 95% confidence interval for $\mu$ is given as

$$\bar{x} \pm t_{10;0.025} \cdot \frac{s}{\sqrt{n}} = 583.64 \pm 2.228 \cdot \frac{22.15}{\sqrt{11}} = (568.8, 598.5).$$

The upper bound of this confidence interval is \$50 below the retail price of \$649, so some potential savings can be made when buying this phone from eBay.

**INTERPRETING CONFIDENCE INTERVALS I:**

- We saw that $\overline{X} \pm E$ has probability $(1-\alpha)$ of containing $\mu$.

  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 $\mu$, and sometimes we get one that does not contain $\mu$.
- Once an interval is computed, $\mu$ is either in it or not. There is no more randomness. ■

**INTERPRETING CONFIDENCE INTERVALS II:**

- Since $\mu$ 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/` ■

**L-EXAMPLE 6.6**
We give two correct ways to interpret the meaning of the level $1-\alpha$ confidence interval.

- The first is based on how confidence interval is defined:

$$P(\overline{X} - E < \mu < \overline{X} + E) = 1 - \alpha.$$

  Here, $\overline{X}$ is the random variable that supplies the randomness such that we can talk about "probability". However,

$$P(\bar{x} - z_{\alpha/2}\sigma/\sqrt{n} < \mu < \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha$$

  makes nonsense, as there is no random variable in the $P(\cdot)$ statement. More specifically, if your computed confidence interval is $(3,6)$, it is **INVALID** to report "the probability that $\mu$ is contained in $(3,6)$ is 95%."

- The second is the empirical explanation. Suppose that we can sample the data infinitely many times from the population.
  - ⋆ Get the first sample $\left(X_1^{(1)}, X_2^{(1)}, \ldots, X_n^{(1)}\right)$ from the population, and compute the confidence interval $(a_1, b_1)$.
  - ⋆ Get the second sample $\left(X_1^{(2)}, X_2^{(2)}, \ldots, X_n^{(2)}\right)$ from the population, and compute the confidence interval $(a_2, b_2)$.
  - ⋆ Continue with this procedure $\ldots, \ldots, \ldots$
  - ⋆ Get the $K$th sample $\left(X_1^{(K)}, X_2^{(K)}, \ldots, X_n^{(K)}\right)$ from the population, and compute the confidence interval $(a_K, b_K)$.
  - ⋆ For a sufficiently large $K$, the proportion of these intervals that contain the true value of the parameter, $\mu$ say, is $1 - \alpha$.

- In practice, for a computed confidence interval, we can only claim that with a certain "confidence" the interval will cover the true value. For example, if the computed 95% CI for $\mu$ is $(3,6)$, we can only claim that we have 95% "confidence" that the true value of $\mu$ will be contained in $(3,6)$.

## 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 $\mu_1$, variance $\sigma_1^2$.
- Population 2 has mean $\mu_2$, variance $\sigma_2^2$.

**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 $\mu_1$ and variance $\sigma_1^2$.
2. A random sample of size $n_2$ from population 2 with mean $\mu_2$ and variance $\sigma_2^2$.
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,

$$E(\overline{X}) = \mu_1, \quad \text{var}(\overline{X}) = \frac{\sigma_1^2}{n_1}, \quad E(\overline{Y}) = \mu_2, \quad \text{var}(\overline{Y}) = \frac{\sigma_2^2}{n_2}.$$

Thus

$$E(\overline{X} - \overline{Y}) = \mu_1 - \mu_2 = \delta,$$

and, using the independence assumption,

$$\text{var}(\overline{X} - \overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

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 $\sigma_1^2$ and $\sigma_2^2$ 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

$$(\bar{x}-\bar{y})\pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}.$$

**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 $\mu_A$ and $\mu_B$ 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 $\mu_1$ and variance $\sigma_1^2$.
2. A random sample of size $n_2$ from population 2 with mean $\mu_2$ and variance $\sigma_2^2$.
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 $\sigma_1$ and $\sigma_2$ 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 $\sigma_1^2$ and $\sigma_2^2$ 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

$$(\overline{x}-\overline{y}) \pm z_{\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2}.$$

## 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 $\mu_1$ and variance $\sigma_1^2$.
2. A random sample of size $n_2$ from population 2 with mean $\mu_2$ and variance $\sigma_2^2$.
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 $\sigma$ is unknown, we shall estimate it.

Note that $S_1^2$ and $S_2^2$ are both unbiased estimators of $\sigma^2$ under the equal variance assumption.

We can use the pooled estimator to estimate $\sigma^2$ better.

**DEFINITION 6 (THE POOLED ESTIMATOR: $S_p^2$)**
*$\sigma^2$ 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 $\sigma^2$ 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 $\mu_1$ and $\mu_2$ represent the average grades of all students who might take this course by the classroom and programmed presentations respectively.

So $\bar{x} - \bar{y} = 85 - 81 = 4$ 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

$$\begin{aligned}
&(\bar{x} - \bar{y}) \pm t_{n_1+n_2-2;\alpha/2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
&= (85-81) \pm 1.7247 \times \sqrt{20.05} \times \sqrt{\frac{1}{12} + \frac{1}{10}} \\
&= (0.693, 7.307).
\end{aligned}$$

**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

$$(\overline{X} - \overline{Y}) \pm z_{\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

**L-EXAMPLE 6.7 (ELECTRICAL USAGE)**
As a baseline for a study on the effects of changing electrical pricing for electricity during peak hours, July usage during peak hours was obtained for $n_1 = 45$ homes with air-conditioning and $n_2 = 55$ homes without. The summarized results are provided below

| | Samples | | |
|---|---|---|---|
| population | Size | Mean | Variance |
| With | 45 | 204.4 | 13,825.3 |
| Without | 55 | 130.0 | 8,632.0 |

Construct a 95% confidence interval for $\delta = \mu_1 - \mu_2$.

***Solution:***
For a 95% confidence interval, $\alpha = 0.05$ and $z_{0.025} = 1.96$. The information provided by the question includes:

$$n_1 = 45, \bar{x} = 204.4,\ s_1^2 = 13825.3$$
$$n_2 = 55, \bar{y} = 130.0,\ s_2^2 = 8632.0$$

If we adopt the assumption that $\sigma_1^2 \neq \sigma_2^2$, the 95% confidence interval is then constructed via the formula:

$$\begin{aligned}
&(\bar{x} - \bar{y}) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \\
= \quad & (204.4 - 130.0) \pm 1.96\sqrt{\frac{13825.3}{45} + \frac{8632.0}{55}} \\
= \quad & (32.1724, 116.6276).
\end{aligned}$$

If we assume that $\sigma_1^2 = \sigma_2^2$, the 95% confidence interval is then constructed via the formula:

$$\begin{aligned}
&(\bar{x} - \bar{y}) \pm z_{\alpha/2}\, s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\
= \quad & (204.4 - 130) \pm 1.96 \cdot \sqrt{10963.69}\sqrt{\frac{1}{45} + \frac{1}{55}} \\
= \quad & (33.1478,\ 115.6522).
\end{aligned}$$

Here

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 10963.69.$$

**REMARK:**
We can roughly assume the equal variance if

$$1/2 \le S_1/S_2 \le 2.$$

This is because the statistic is not overly sensitive to small difference between the population variances.

## 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 $\mu_D$ and variance $\sigma_D^2$.

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 $\mu_D$ is

$$\bar{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

If $n$ is large ($n \geq 30$), a $(1-\alpha)100\%$ confidence interval for $\mu_D$ is

$$\bar{d} \pm z_{\alpha/2} \cdot \frac{s_D}{\sqrt{n}}.$$

■

**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 $\bar{d} = -1.6$ 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 $\mu_D$ is given as

$$\bar{d} \pm t_{n-1;\alpha/2} \cdot \frac{s_D}{\sqrt{n}} = -1.6 \pm 2.821 \times \sqrt{\frac{40.71}{10}} = (-7.292, 4.092).$$

**L-EXAMPLE 6.8 (WATER TREATMENT)**
A state law requires municipal waste water treatment plants to monitor their discharges into rivers and streams. A treatment plant could choose to send its samples to a commercial laboratory of its choosing.

Concern over this self-monitoring led a civil engineer to design a matched pairs experiment. Exactly the same bottle of effluent cannot be sent to two different laboratories. To match "identical" as closely as possible, she would take a sample of effluent in a large sample bottle and pour it back and forth over two open specimen bottles.

When they were filled and capped, a coin was flipped to see if the one on the right was sent to commercial laboratory or the state laboratory.

This process was repeated 11 times. The results, for the response suspended solids (SS) are

| Sample | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Commercial lab | 27 | 23 | 64 | 44 | 30 | 75 | 26 | 124 | 54 | 30 | 14 |
| State lab | 15 | 13 | 22 | 29 | 31 | 64 | 30 | 64 | 56 | 20 | 21 |
| Difference $X_i - Y_i$ | 12 | 10 | 42 | 15 | $-1$ | 11 | $-4$ | 60 | -2 | 10 | $-7$ |

Obtain a 95% confidence interval for the difference in SS from the two labs.

***Solution:***
We assume a normal distribution for the population. From the data, we compute the following

$$n = 11, \overline{d} = 13.27, s_D^2 = 418.61.$$

Further, since $\alpha = 0.05$, we have $t_{n-1;\alpha/2} = t_{10,0.025} = 2.228$.

The 95% confidence interval is given as

$$13.27 \pm 2.228\sqrt{\frac{418.61}{11}} = 13.27 \pm 2.228 \cdot 6.16891 = (-0.47, 27.01).$$

If we **WRONGLY** assume the independence between populations, and construct the confidence interval using the formula

$$(\overline{X} - \overline{Y}) \pm t_{n_1+n_2-2;\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

We can compute $s_p^2 = 688.9182$, and the 95% confidence interval is given by

$$\begin{aligned} & 13.27 \pm 2.086\sqrt{688.9182}\sqrt{\frac{1}{11} + \frac{1}{11}} = 13.27 \pm 2.086 \cdot 11.19187 \\ = \;& (-10.07582, 36.61582). \end{aligned}$$

# *Seven*

# Hypothesis Tests

## 1 Hypothesis Tests

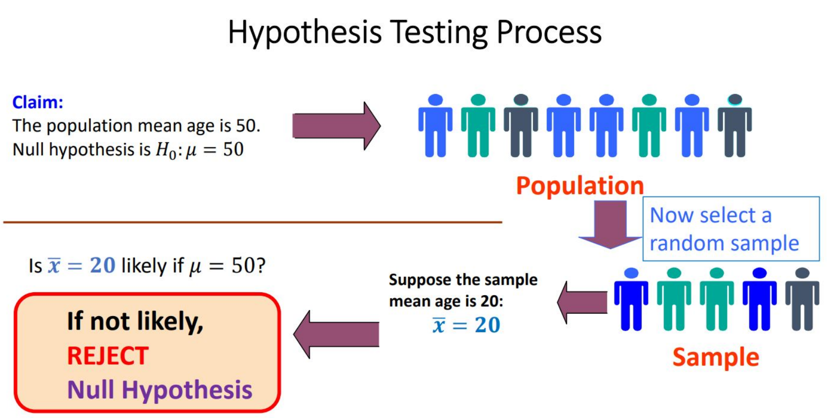
One of the most fundamental technique of statistical inference is the hypothesis test. There are many types of hypothesis tests but all follow the same logical structure, so we begin with hypothesis testing of a population mean.

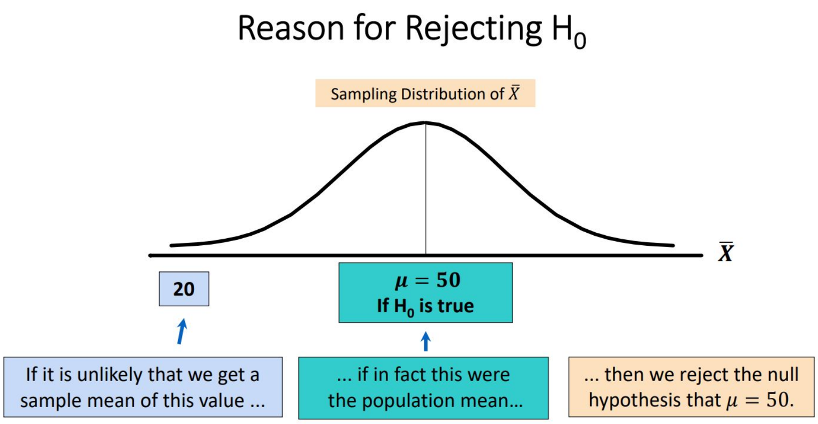
Hypothesis testing begins with a null hypothesis and an alternative hypothesis. Both the null and the alternative hypotheses are statements about a population. In this chapter, that statement will be a statement about the mean(s) of the population(s).

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 $\mu$ 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 $\mathrm{H}_0$ |
|---|---|---|
| $H_0$ 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 $\alpha$. That is,*

$$\alpha = P(\textit{Type I error}) = P(\textit{Reject } H_0 \mid H_0 \textit{ is true}).$$

*Let*

$$\beta = P(\textit{Type II error}) = P(\textit{Do not reject } H_0 \mid H_0 \textit{ is false}).$$

*We define $1-\beta = P(\textit{Reject } H_0 \mid H_0 \textit{ is false})$ 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 $\alpha$ 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 $\alpha$ 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 $H_0$."

**L-EXAMPLE 7.1 (POSSIBLE STRUCTURES OF $H_0$ AND $H_1$)**

For a valid set of hypotheses (i.e., $H_0$ versus $H_1$), they need to be disjoint. For example "$H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$" is one valid set; "$H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$" is another valid set.

By convention (and theoretically supported), we usually write the null hypothesis in the "equal" form, i.e., $H_0 : \theta = \theta_0$, with $\theta_0$ being a given value. The hypotheses have three possible forms:

⋆ $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$; in this case $H_0 : \theta_0 = \theta_0$ in fact means $\theta \leq \theta_0$;

⋆ $H_0 : \theta = \theta_0$ versus $H_1 : \theta < \theta_0$; in this case $H_0 : \theta_0 = \theta_0$ in fact means $\theta \geq \theta_0$;

⋆ $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

So we need to check the form of $H_1$ to ensure the real meaning of $H_0 : \theta = \theta_0$.

**L-EXAMPLE 7.2**

A certain type of cold vaccine is known to be only 25% effective after a period of 2 years.

In order to determine if a new and somewhat more expensive vaccine is superior in providing protection against the same virus for a longer period of time, 20 people are chosen at random and inoculated with the new vaccine.

If more than 8 of those receiving the new vaccine surpass the 2-year period without contracting the virus, the new vaccine will be considered superior to the one presently in use.

This is equivalent to testing the hypothesis that the binomial parameter for the probability of a success on a given trial is $p = \frac{1}{4}$ against the alternative that $p > \frac{1}{4}$.

In other words, we are testing

$$H_0 : p = \frac{1}{4} \quad \text{vs} \quad H_1 : p > \frac{1}{4}.$$

Let $X$ be the number of individuals who remain free of the virus for at least 2 years.

For the conditions given, we think of the "acceptance region" and "rejection region" as

$$\overbrace{0, 1, 2, \ldots, 7, 8,}^{\text{acceptance region}} \; \overbrace{9, 10, 11, \ldots, 19, 20}^{\text{rejection region}}$$

The above decision rule has the level of significance given by

$$\begin{aligned}
\alpha &= P(\text{Type I error}) \\
&= P(\text{Reject } H_0 \mid H_0 \text{ is true}) \\
&= P(X > 8 \mid p = \frac{1}{4}) \\
&= \sum_{i=9}^{20} \binom{20}{i} \left(\frac{1}{4}\right)^i \left(\frac{3}{4}\right)^{20-i} = 0.0409.
\end{aligned}$$

The probability of committing a Type II error is impossible to compute unless we have a specific alternative hypothesis. So let's consider

$$H_0 : p = \frac{1}{4} \quad \text{vs} \quad H_1 : p = \frac{1}{2}.$$

Note that $p = \frac{1}{2}$ satisfies $p > \frac{1}{4}$.

With this,

$$\begin{aligned}
\beta &= P(\text{Type II error}) \\
&= P(\text{Do not reject } H_0 \mid H_1 \text{ is true}) \\
&= P(X \le 8 \mid p = \frac{1}{2}) \\
&= \sum_{i=0}^{8} \binom{20}{i} \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{20-i} = 0.2517.
\end{aligned}$$

## 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 $\sigma^2$ 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: $\alpha$ is typically set to be 0.05.

Step 3: Statistic & its distribution:
With $\sigma^2$ 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.

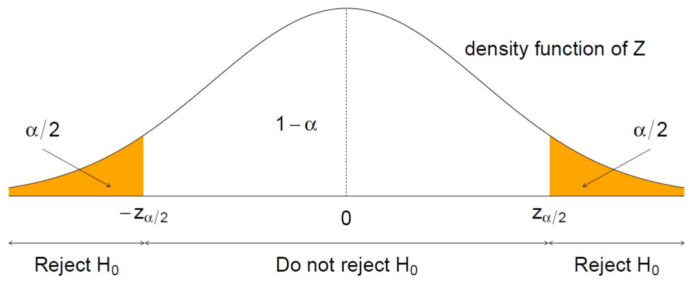
Rejection region:
Intuitively, we should reject $H_0$ when $\overline{X}$ is too large or too small compared with $\mu_0$.

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 $\mu$ 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 $\sigma^2$ 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

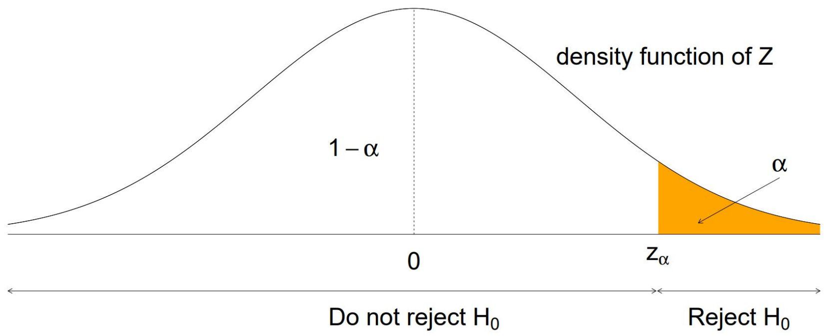
$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

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 $\sigma^2$ 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 $\alpha$.

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.

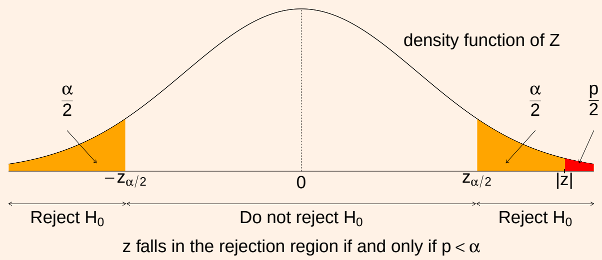
**DEFINITION 3 ($p$-VALUE)**
*The* ***p-value*** *is the probability of obtaining a test statistic at least as extreme ($\leq$ or $\geq$) than the observed sample value, given $H_0$ is true.*

*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, $|Z| > |z|$.

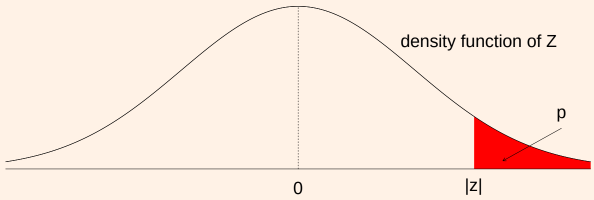
So the $p$-value is given by

$$p\text{-value} = P(|Z| > |z|) = 2P(Z > |z|) = 2P(Z < -|z|)$$




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.




One−sided test: p−value is represented by the shaded area

■

**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 $\mu$ be the average midterm score for the whole class.

Step 1: $H_0 : \mu = 16 \quad \text{vs} \quad H_1 : \mu \neq 16.$

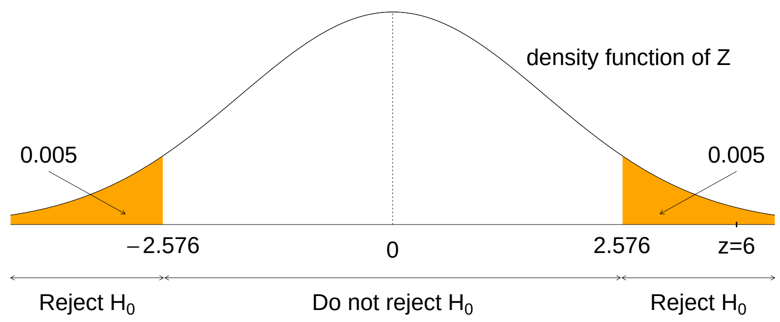
Step 2: Choose $\alpha = 0.01$.

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: $z = (22 - 16)/(\sqrt{5}/\sqrt{5}) = 6 > 2.576$.

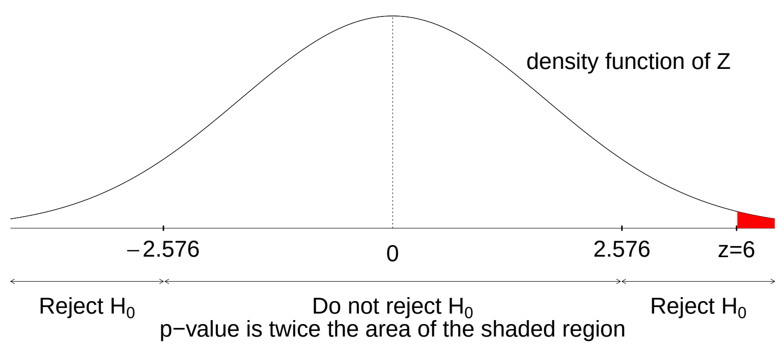
Step 5: As $z = 6$ falls in rejection region, $H_0$ is rejected.

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 $\sigma^2$ 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. ■

**L-EXAMPLE 7.3 (MIDTERM EXAM SCORE II)**
Continuing from the previous example. Let's say the lecturer didn't announce the variance, that is, $\sigma$ is unknown.

The data given has $\bar{x} = 22$ and $s = 2.55$.

Test again at $\alpha = 0.01$ significance level whether the average midterm score is different from 16.

***Solution:***
Since $\sigma$ is unknown, we need to perform a t-test.

Again, we let $\mu$ be the average midterm score for the whole class.

Step 1: $H_0 : \mu = 16$ vs $H_1 : \mu \neq 16$.

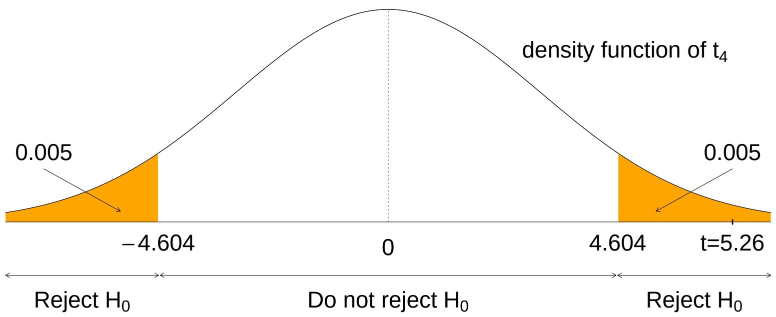
Step 2: Choose $\alpha = 0.01$.

Step 3: Since $\sigma$ is unknown, data are normal, and $n = 5$, the test statistics is

$$T = \frac{\overline{X} - 16}{S/\sqrt{n}} \sim t_{n-1} = t_4.$$

Now $t_{n-1,\alpha/2} = t_{4,0.005} = 4.604$. So the rejection region is

$$t < -4.604 \quad \text{or} \quad t > 4.604.$$



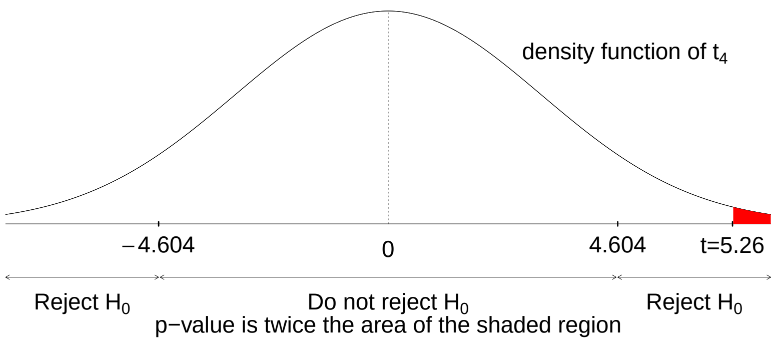

Step 4: $t = (22 - 16)/(2.55/\sqrt{5}) = 5.26$.

Step 5: Since $t = 5.26$ falls within the rejection region, we reject $H_0$.

Alternatively, the $p$-value can be found to be

$$2P(t_4 > 5.26) = 0.0063,$$

which is smaller than $\alpha = 0.01$, so we reject $\mathrm{H}_0$.



**L-EXAMPLE 7.4 (DEPARTMENT STORE)**
A department store manager determines that a new billing system will be cost-effective only if the mean monthly account is more than \$170. It is known that the accounts has standard deviation \$65.

A random sample of 400 monthly accounts is drawn, for which the sample mean is \$178. Can we conclude that the new system will be cost-effective at 5% level of significance?

***Solution:***
Let $\mu$ be the mean monthly account.

Step 1: We test

$$H_0 : \mu = 170 \quad \text{vs} \quad H_1 : \mu > 170.$$

Step 2: Choose $\alpha = 0.05$.

Step 3: Since $n$ is large and $\sigma$ is known, we use the test statistic:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Under $H_0$, by the Central Limit Theorem, we have, $Z \sim N(0, 1)$.

At a 5% significance level ($\alpha = 0.05$), we get

$$z_\alpha = z_{0.05} = 1.645.$$

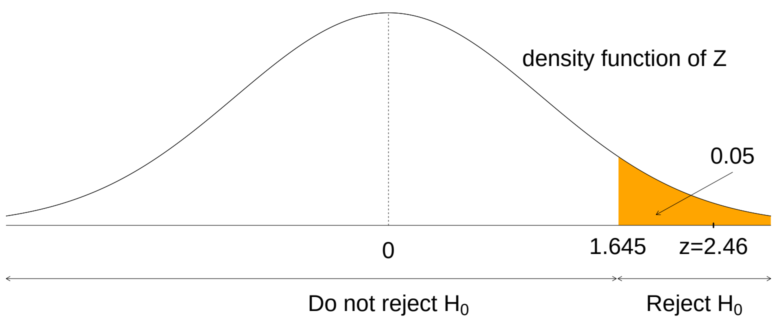
Step 4: We are given that

$$n = 400, \quad \bar{x} = 178, \quad \sigma = 65, \quad \alpha = 0.05$$

and so

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{178 - 170}{65/\sqrt{400}} = 2.46 > z_\alpha = 1.645.$$

Step 5: Therefore, we reject the null hypothesis and conclude that the mean monthly account is more than \$170.



## 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 $\mu$.

We illustrate using Case III: normal population, small $n$, unknown $\sigma$.

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 $\mu$ in this case is given by

$$\left(\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \, \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right).$$

If the $100(1-\alpha)\%$ confidence interval contains $\mu_0$, we will have

$$\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

Rearranging the above inequality, we obtain

$$-t_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \leq t_{\alpha/2}.$$

This means that the computed test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ satisfies

$$-t_{\alpha/2} \leq t \leq t_{\alpha/2}.$$

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 $\mu_0$, $H_0$ will not be rejected at level $\alpha$.

Similarly, when the confidence interval does not contain $\mu_0$, 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., $\sigma$ 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

$$H_0 : \mu = 17 \quad \text{vs} \quad H_1 : \mu \neq 17?$$

**L-Example 7.5**
A study based on a sample size of 36 reported a mean of 87 with a margin of error of 10 for 95% confidence.

Give the 95% confidence interval for the population mean $\mu$.

You are then asked to test the hypothesis that $\mu = 80$ against a two sided alternative at $\alpha = 0.05$. What is your conclusion?

***Solution:***
The 95% confidence interval for $\mu$ is given as

$$\bar{x} \pm E = 87 \pm 10 = (77, 97).$$

The 95% confidence interval contains the the value 80 so there is no evidence to reject the null at $\alpha = 0.05$.

## 4 Tests Comparing Means: Independent Samples

Suppose two independent samples are drawn from two populations with means $\mu_1$ and $\mu_2$. 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 $\sigma_1^2$ and $\sigma_2^2$ 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 $\sigma_1^2$ and $\sigma_2^2$ 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 $\mu_1$ and $\mu_2$ are different?
Use $\alpha = 0.01$.

***Solution:***
Let $\mu_1$ and $\mu_2$ 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.

**L-EXAMPLE 7.6 (ELECTRICAL USAGE II)**

As a baseline for a study on the effects of changing electrical pricing for electricity during peak hours, July usage during peak hours was obtained for $n_1 = 45$ homes with air-conditioning and $n_2 = 55$ homes without. The summarized results are provided below

| | Samples | | |
|---|---|---|---|
| population | Size | Mean | Variance |
| With | 45 | 204.4 | 13,825.3 |
| Without | 55 | 130.0 | 8,632.0 |

Perform a hypothesis test at $\alpha = 0.05$ that the mean on-peak usage for homes with air-conditioning is higher than that for homes without.

***Solution:***

Let $\mu_1$ and $\mu_2$ be the mean on-peak usage for homes with and without air-conditioning respectively.

Step 1: Again we have $\delta_0 = 0$. So we test

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 > 0.$$

Step 2: Set $\alpha = 0.05$.

Step 3: Test statistic and its distribution is given below:

$$Z = \frac{(\overline{X} - \overline{Y}) - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1).$$

The rejection region is $z > z_{0.05} = 1.645$.

Step 4: Plug in the data,

$$z = \frac{204.4 - 130.0 - 0}{\sqrt{\frac{13{,}825.3}{45} + \frac{8632.0}{55}}} = 3.45.$$

Step 5: We reject $H_0$ since $z = 3.45 > z_{0.05} = 1.645$.

**COMPARING MEANS: INDEPENDENT SAMPLES II:**

Consider the case where

- the population variances $\sigma_1^2$ and $\sigma_2^2$ 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

$$Z = \frac{(\overline{X} - \overline{Y}) - \delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

**L-EXAMPLE 7.7 (COAL SPECIMENS)**

The following are measurements of heat-producing capacity (millions of calories per ton) of sample specimens of coal from two mines:

```
Mine 1: 8260 8130 8350 8070 8340
Mine 2: 7950 7890 7900 8140 7920 7840
```

The sample summary statistics are

$$\bar{x} = 8230, \quad s_1 = 125.5, \quad \bar{y} = 7940, \quad s_2 = 104.5.$$

Assume that both populations are normal with equal variance. Test at $\alpha = 0.01$ level if the means between these two mines are different.

***Solution:***

Let $\mu_1$ and $\mu_2$ be the mean heat-producing capacity for the two mines.

Step 1: $H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \mu_1 - \mu_2 \neq 0.$

Step 2: $\alpha = 0.01$.

Step 3: We are given that $s_1 = 125.5, s_2 = 104.5$ and that the equal variance assumption holds. The test statistics is

$$T = \frac{(\overline{X} - \overline{Y}) - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Since $t_{n_1+n_2-2,\alpha/2} = t_{9,0.005} = 3.250$, the rejection region is

$$t < -3.250 \quad \text{or} \quad t > 3.250.$$

Step 4: Plug in everything, we get

$$\begin{aligned} s_p^2 &= \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \\ &= \frac{(5-1)\times 125.5^2 + (6-1)\times 104.5^2}{5+7-2} = 13066.92, \end{aligned}$$

and

$$t = \frac{(8230-7940)-0}{\sqrt{13066.92}\times\sqrt{\frac{1}{5}+\frac{1}{6}}} = 4.18963.$$

Step 5: Since $t = 4.18963$ is in the rejection region, we reject $H_0$.

## 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.

**L-EXAMPLE 7.8 (WATER TREATMENT)**
A state law requires municipal waste water treatment plants to monitor their discharges into rivers and streams. A treatment plant could choose to send its samples to a commercial laboratory of its choosing.

Concern over this self-monitoring led a civil engineer to design a matched pairs experiment. Exactly the same bottle of effluent cannot be sent to two different laboratories. To match "identical" as closely as possible, she would take a sample of effluent in a large sample bottle and pour it back and forth over two open specimen bottles.

When they were filled and capped, a coin was flipped to see if the one on the right was sent to commercial laboratory or the state laboratory.

This process was repeated 11 times. The results, for the response suspended solids (SS) are

| Sample | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Commercial lab | 27 | 23 | 64 | 44 | 30 | 75 | 26 | 124 | 54 | 30 | 14 |
| State lab | 15 | 13 | 22 | 29 | 31 | 64 | 30 | 64 | 56 | 20 | 21 |
| Difference $X_i - Y_i$ | 12 | 10 | 42 | 15 | $-1$ | 11 | $-4$ | 60 | -2 | 10 | $-7$ |

The summary statistics for $D = X_i - Y_i$ are

$$\overline{d} = 13.27, s_D^2 = 418.61.$$

Conduct a hypothesis test to check if the SS from the commercial lab is higher than those from state lab at significance level 0.05. Assume a normal distribution for the population.

***Solution:***
We shall test

$$H_0 : \mu_D = 0 \quad \text{vs} \quad H_1 : \mu_D > 0.$$

The test statistics is

$$T = \frac{\overline{D} - 0}{S_D/\sqrt{n}},$$

and the rejection region is $t > t_{10,0.05} = 1.812$.

Computations gives the observed test statistics as

$$t = \frac{\overline{d} - 0}{\sqrt{418.61/11}} = 2.15 > 1.812.$$

Since $t = 2.15 > t_{10,0.05} = 1.812$, we reject $H_0$ and conclude that the response from commercial lab is higher than those from the state lab.

**L-EXAMPLE 7.9 (MEAN RELATED INFERENCE)**
Assume that we are to make inference, including constructing the confidence interval or perform a hypothesis test, concerning a mean-related parameters $\theta$. A general procedure is as follows:

- ✓ Step 1: Look for an estimator $\widehat{\theta}$ for $\theta$, e.g., $\overline{X}$ for $\mu$.
- ✓ Step 2: Derive the formula for $\text{var}(\widehat{\theta})$.
- ✓ Step 3: We then construct

$$T = \frac{\widehat{\theta} - \theta}{\sqrt{V}}. \tag{7.1}$$

For $V$, we consider the following possibilities.

- ⋆ If $\text{var}(\widehat{\theta})$ does not depend on any unknown parameter, e.g., when $\sigma^2$ is known, $\text{var}(\overline{X}) = \sigma^2/n$, we set $V = \text{var}(\widehat{\theta})$. The statistic $T$ given in (7.1) (approximately) follows the $N(0,1)$ distribution, when the data are normal or the sample size is sufficiently large.

⋆ If $\text{var}(\widehat{\theta})$ contains some other unknown parameters, e.g., $\sigma^2$, we replace the parameter with its estimator, e.g., $S^2$ can be used to replace $\sigma^2$, and result in $\widehat{\text{var}}(\widehat{\theta})$. We set $V = \widehat{\text{var}}(\widehat{\theta})$. The distribution of $T$ given in (7.1) has two possibilities:

(1) The sample size is sufficiently large; then $T \sim N(0,1)$ approximately.

(2) If the sample size is small, but the observations are normally distributed, then $T \sim t_{df}$, where $df$ is the degrees of the freedom of the parameter estimated in $\text{var}(\widehat{\theta})$.

Now, statistical inference for $\theta$ can be done based on $T$ and its associated distribution discussed above.

- Construct $(1-\alpha)$ confidence interval:

$$\widehat{\theta} \pm M\sqrt{V},$$

where $M$, based on the distribution of $T$, is either $z_{\alpha/2}$ or $t_{df,\alpha/2}$.

- Test statistic for hypothesis test with $H_0 : \theta = \theta_0$ is given by

$$T = \frac{\widehat{\theta} - \theta_0}{\sqrt{V}},$$

with its distribution resulted from above.