Jian-Lun Xu

NOTES OF STAT420

Contents

1

Probability and Distributions

1.1 Sample space and operation of events

Definition 1.1.1. *The set of all possible outcomes of an experiment is called the sample space and is denoted by* Ω*. Any subset of* Ω *is called an event and usually denoted by* A, B, C*, etc.*

Definition 1.1.2. *The union of events* A *and* B*, denoted by* $A \cup B$*, is defined by*

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

In Figure 1.1.1 below, the union of events A and B is shown by the shaded region in the Venn diagram.

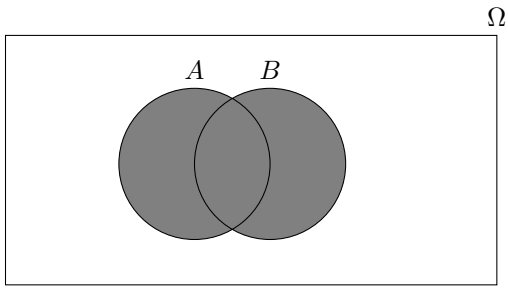


Figure 1.1.1 Union $A \cup B$ of A and B

Example 1.1.1. *Suppose that* $\Omega = \{1, 2, 3, 4, 5, 6\}$*. Let* $A = \{1, 2, 3\}$ *and* $B = \{2, 3, 4\}$*. Then* $A \cup B = \{1, 2, 3, 4\}$*.*

Example 1.1.2. *Suppose that* $\Omega = \mathbb{R}$*. Let* $A = (-\infty, 1]$ *and* $B = [0, 2)$*. Then* $A \cup B = (-\infty, 2)$*.*

Definition 1.1.3. *The intersection of events A and B, denoted by $A \cap B$ (or AB), is defined by*

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

In Figure 1.1.2 below, the intersection of events A and B is shown by the shaded region in the Venn diagram.

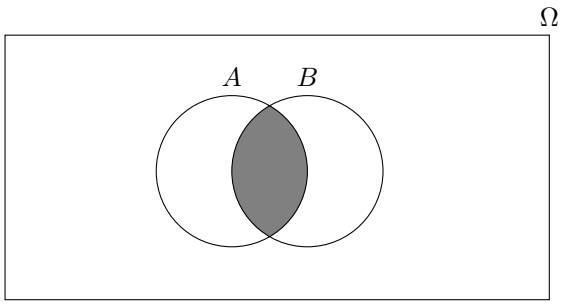


Figure 1.1.2 Intersection $A \cap B$ of A and B

Example 1.1.3. *Let A, B, and Ω be defined by Example 1.1.1. Then $A \cap B = \{2, 3\}$.*

Example 1.1.4. *Let A, B, and Ω be defined by Example 1.1.2. Then $A \cap B = [0, 1]$.*

Theorem 1.1.1. *Let A, B, and C be events. Then*

(a) $A \cup B = B \cup A$;	(a′) $A \cap B = B \cap A$;
(b) $A \cup A = A$;	(b′) $A \cap A = A$;
(c) $A \cup \emptyset = A$;	(c′) $A \cap \emptyset = \emptyset$;
(d) $A \cup (B \cup C) = (A \cup B) \cup C$;	(d′) $A \cap (B \cap C) = (A \cap B) \cap C$;
(e) $A \subseteq A \cup B$ *and* $B \subseteq A \cup B$;	(e′) $A \cap B \subseteq A$ *and* $A \cap B \subseteq B$;
(f) $A \subseteq B$ *iff* $A \cup B = B$;	(f′) $A \subseteq B$ *iff* $A \cap B = A$.

Theorem 1.1.2. *Let A, B, and C be events. Then*

(a) (*Distributive property of union over intersection*)

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

(b) (*Distributive property of intersection over union*)

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

Definition 1.1.4. *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events. We define*

$$\bigcup_{i=1}^{n} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for } \textbf{some} \text{ integer } i = 1, 2, \ldots, n\},$$
$$\bigcup_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for } \textbf{some} \text{ integer } i = 1, 2, 3, \ldots\},$$
$$\bigcap_{i=1}^{n} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for } \textbf{every} \text{ integer } i = 1, 2, \ldots, n\},$$
$$\bigcap_{i=1}^{\infty} A_i = \{\omega \in \Omega : \omega \in A_i \text{ for } \textbf{every} \text{ integer } i = 1, 2, 3, \ldots\}.$$

Example 1.1.5. *Suppose that $\Omega = \mathbb{R}$. For $n \in \mathbb{N}$, let $A_n = \big(0, n/(n+1)\big)$. Then*

$$\bigcup_{i=1}^{n} A_i = \big(0, n/(n+1)\big), \quad \bigcup_{i=1}^{\infty} A_i = \big(0, 1\big), \quad \bigcap_{i=1}^{n} A_i = \bigcap_{i=1}^{\infty} A_i = (0, 1/2)\,.$$

Example 1.1.6. *Suppose that $\Omega = \mathbb{R}$. For $n \in \mathbb{N}$, let $A_n = \big[-1/n, 1/n\big]$. Then*

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{\infty} A_i = [-1, 1], \quad \bigcap_{i=1}^{n} A_i = \big[-1/n, 1/n\big], \quad \bigcap_{i=1}^{\infty} A_i = \{0\}.$$

Definition 1.1.5. *Let A and B be events. The difference between A and B, denoted by $A\backslash B$, is defined as an event of all elements in A that are not in B, i.e.,*

$$A\backslash B = \{x \in \Omega : x \in A,\, x \notin B\}.$$

Clearly, $A\backslash B = \emptyset$ if $A \subseteq B$.

In Figure 1.1.3 below, the difference $A\backslash B$ between A and B is shown by the shaded region in the Venn diagram.

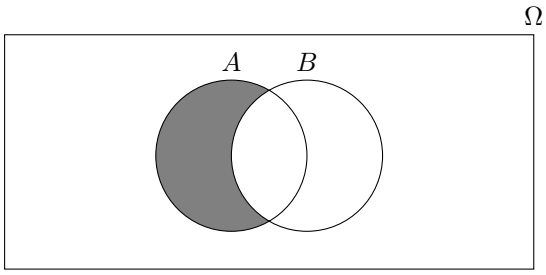


Figure 1.1.3 Difference $A\backslash B$ of A and B

Example 1.1.7. *Let A, B, and Ω be defined by Example 1.1.1. Then $A\backslash B = \{1\}$.*

Example 1.1.8. *Let A, B, and Ω be defined by Example 1.1.2. Then $A\backslash B = (-\infty, 0)$.*

Definition 1.1.6. *The complement of event A, denoted by A^c, is defined by*

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

That is, A^c is the event of all elements in Ω, but not in A.

In Figure 1.1.4 below, the complement A^c of A is shown by the shaded region in the Venn diagram.

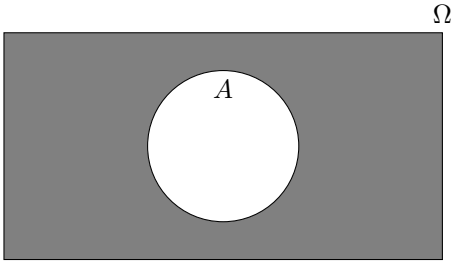


Figure 1.1.4 Complement A^c of A

Example 1.1.9. *Let A, B, and Ω be defined by Example 1.1.1. Then $A^c = \{4, 5, 6\}$ and $B^c = \{1, 5, 6\}$.*

Example 1.1.10. *Let A, B, and Ω be defined by Example 1.1.2. Then $A^c = (1, \infty)$ and $B^c = (-\infty, 0) \cup [2, \infty)$.*

Theorem 1.1.3. *(DeMorgan's laws) Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events. Then*

$$\begin{aligned}\left(\bigcup_{i=1}^{\infty} A_i\right)^c &= \bigcap_{i=1}^{\infty} A_i^c,\\ \left(\bigcap_{i=1}^{\infty} A_i\right)^c &= \bigcup_{i=1}^{\infty} A_i^c.\end{aligned} \tag{1.1}$$

Proof. We prove the first identity in (1.1) only because the proof of the second identity can be similarly carried out.

If $x \in \left(\bigcup_{i=1}^{\infty} A_i\right)^c$, then $x \notin \bigcup_{i=1}^{\infty} A_i$, which means that we cannot find an $n \in \mathbb{N}$ such that $x \in A_n$. That is, $x \in A_n^c$ for all $n \in \mathbb{N}$, which is equivalent to $x \in \bigcap_{i=1}^{\infty} A_i^c$. This proves that

$$\left(\bigcup_{i=1}^{\infty} A_i\right)^c \subseteq \bigcap_{i=1}^{\infty} A_i^c. \tag{1.2}$$

Conversely, if $y \in \bigcap_{i=1}^{\infty} A_i^c$, then $y \in A_i^c$ for all $i \in \mathbb{N}$. This means that $y \notin A_i$ for all $i \in \mathbb{N}$, which is equivalent to $y \in \left(\bigcup_{i=1}^{\infty} A_i\right)^c$. This proves that

$$\bigcap_{i=1}^{\infty} A_i^c \subseteq \left(\bigcup_{i=1}^{\infty} A_i\right)^c. \tag{1.3}$$

Combining (1.2) and (1.3) completes the proof. □

Definition 1.1.7. *Events A and B are said to be mutually exclusive (or disjoint) if $AB = \emptyset$ (impossible event).*

Example 1.1.11. *Let A, B, and Ω be defined by Example 1.1.1 and let $C = \{5, 6\}$. Then $AC = \emptyset$, That is, A and C are mutually exclusive. Similarly, B and C are also mutually exclusive.*

Example 1.1.12. *Let A, B, and Ω be defined by Example 1.1.2 and let $C = (-1, 0)$. Then $AC = (-1, 0) \neq \emptyset$. Thus, A and C are not mutually exclusive. However, B and C are mutually exclusive.*

Definition 1.1.8. *For a sequence $(A_n)_{n\in\mathbb{N}}$ of events, we define*

$$\inf_{i\geq n} A_i = \bigcap_{i=n}^{\infty} A_i,$$
$$\sup_{i\geq n} A_i = \bigcup_{i=n}^{\infty} A_i,$$
$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} \inf_{i\geq n} A_i = \bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} A_i,$$
$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \sup_{i\geq n} A_i = \bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} A_i.$$

Remark 1.1.1. *The event $\inf\limits_{i\geq 1} A_i = \bigcap\limits_{i=1}^{\infty} A_i$ is the largest event contained in each A_i and the event $\sup\limits_{i\geq 1} A_i = \bigcup\limits_{i=1}^{\infty} A_i$ is the smallest event containing all A_i. Both concepts are similar to those in calculus when we have a sequence of real numbers.*

Definition 1.1.9. *For a sequence $(A_n)_{n\in\mathbb{N}}$ of events, we call its limit, denoted by $\lim\limits_{n\to\infty} A_n$, exists if $\liminf\limits_{n\to\infty} A_n = \limsup\limits_{n\to\infty} A_n$.*

Example 1.1.13. *For $n \in \mathbb{N}$, let $A_n = [0, 1/n)$. Then*

$$\inf_{i\geq n} A_i = \bigcap_{i=n}^{\infty} A_i = \{0\},$$
$$\sup_{i\geq n} A_i = \bigcup_{i=n}^{\infty} A_i = [0, 1/n),$$
$$\liminf_{n\to\infty} A_n = \bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} A_i = \{0\},$$
$$\limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} A_i = \bigcap_{n=1}^{\infty} A_n = \{0\}.$$

Thus, $\lim\limits_{n\to\infty} A_n = \{0\}$.

Definition 1.1.10. *A sequence $(A_n)_{n\in\mathbb{N}}$ of events is nondecreasing if $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ and nonincreasing if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$. We denote nondecreasing sequence and nonincreasing sequence as $A_n \uparrow$ and $A_n \downarrow$, respectively. Furthermore, if $\lim\limits_{n\to\infty} A_n = A$, we denote this as $A_n \uparrow A$ and $A_n \downarrow A$, respectively.*

Example 1.1.14. *Let the sequence $(A_n)_{n\in\mathbb{N}}$ be defined by Example 1.1.13. Then it is nonincreasing.*

Definition 1.1.11. *A partition of event A is a collection of disjoint subsets $\{A_i\}$ of A such that $\bigcup_i A_i = A$.*

Example 1.1.15. *Let $A = [0, \infty)$ and let $A_i = [i-1, i)$ for $i \in \mathbb{N}$. Then $A_1, A_2, \ldots$ form a partition of A.*

Example 1.1.16. *Let Ω be defined by Example 1.1.1 and let $A_i = \{i\}$ for $i = 1, 2, 3, 4, 5, 6$. Then $A_1, A_2, \ldots, A_6$ form a partition of Ω.*

Example 1.1.17. *Let the sequence $(A_n)_{n\in\mathbb{N}}$ be defined by Example 1.1.13. Then $A_1, A_2, \ldots$ do not form a partition of $[0, 1)$ because the sets in the sequence $(A_n)_{n\in\mathbb{N}}$ are not disjoint.*

Example 1.1.18. *If we toss two identical coins once, then*

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

where the outcome (i, j) represents that the first coin comes up i and the second coin comes up j for $i, j = H, T$. Let

$$E = \{(H, T), (T, H)\} \quad \text{and} \quad F = \{(H, H), (T, T)\}.$$

Then $E \cup F = \Omega$ and $EF = \emptyset$. That means, E and F form a partition of Ω.

Example 1.1.19. *(Example 1.2.1 – Example 1.2.4)*

1.2 Probability

1.2.1 σ-field

Definition 1.2.1. *Let Ω be the sample space and let $\mathcal{F}$ be a nonempty collection of subsets of Ω. Then $\mathcal{F}$ is called a σ-field if it satisfies the following conditions:*

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

(b) *if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, and*

(c) *if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.*

Example 1.2.1. *Let A be a nonempty proper subset of Ω. Then $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$ is a σ-field. Indeed, $\mathcal{F}$ is the smallest σ-field containing A in the sense that $\mathcal{F}$ is a subcollection of $\mathcal{F}^*$ if $\mathcal{F}^*$ is any σ-field containing A. In this case, we usually say that $\mathcal{F}$ is generated by A.*

Remark 1.2.1. *For any Ω, there are two extreme examples of σ-field: the collection $\{\emptyset, \Omega\}$ and the power set $\mathcal{P}(\Omega)$ of Ω. Any other σ-field $\mathcal{F}$ based on Ω lies between these two extremes: $\{\emptyset, \Omega\} \subset \mathcal{F} \subset \mathcal{P}(\Omega)$.*

Remark 1.2.2. *The intersection of a collection of σ-fields is a σ-field, while the union of a collection of σ-fields may not be a σ-field.*

Example 1.2.2. *Suppose that $\Omega = \{1, 2, \ldots, 6\}$. Let $A = \{1, 2, 3\}$ and $B = \{1\}$. Then $\mathcal{F}_1 = \{\emptyset, A, A^c, \Omega\}$ and $\mathcal{F}_2 = \{\emptyset, B, B^c, \Omega\}$ are σ-fields. Meanwhile, $\mathcal{F}_1 \cap \mathcal{F}_2 = \{\emptyset, \Omega\}$ is a σ-field, while $\mathcal{F}_1 \cup \mathcal{F}_2 = \{\emptyset, A, A^c, B, B^c, \Omega\}$ is not a σ-field because $A^c \cup B = \{1, 4, 5, 6\} \notin \mathcal{F}_1 \cup \mathcal{F}_2$.*

Definition 1.2.2. *Let $\Omega = \mathbb{R}$. Then the σ-field $\mathcal{B}$ generated from all half-open intervals of the form $(a, b]$ is called the Borel σ-field. The elements of $\mathcal{B}$ are called Borel sets.*

Remark 1.2.3. *The Borel σ-field $\mathcal{B}$ contains the singleton $\{a\}$, semi-infinite intervals $(-\infty, b]$ and (b, ∞), open interval (a, b), closed interval $[a, b]$. This can be seen by observing the following facts:*

$$
\begin{aligned}
\{a\} &= \bigcap_{n=1}^{\infty} (a - 1/n, a] = \left(\bigcup_{n=1}^{\infty} (a - 1/n, a]^c \right)^c, \\
(-\infty, b] &= \bigcup_{n=1}^{\infty} (b - n, b], \\
(b, \infty) &= (-\infty, b]^c, \\
(a, b) &= \Big((-\infty, a] \cup (b, \infty) \cup \{b\} \Big)^c, \\
[a, b] &= (a, b] \cup \{a\}.
\end{aligned}
$$

Remark 1.2.4. *The Borel σ-field on $\mathbb{R}^n$ is the σ-field generated by n-dimensional rectangles of the form $\{(x_1, x_2, \ldots, x_n) : a_i < x_i \leq b_i, i = 1, 2, \ldots, n\}$.*

1.2.2 Probability

Definition 1.2.3. *Let Ω be the sample space of an experiment and let $\mathcal{F}$ be a σ-field of subsets of Ω. Then a real-valued set function P defined on $\mathcal{F}$ is called a probability if it satisfies the following conditions:*

(a) (*nonnegativity*) $P(E) \geq 0$ *for any* $E \in \mathcal{F}$,

(b) (*certainty*) $P(\Omega) = 1$, *and*

(c) (*countable additivity*) *If* $(E_n)_{n\in\mathbb{N}}$ *is an arbitrary sequence of mutually exclusive events, then*

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} P(E_n).$$

1.2.3 Property of probability

Property 1.2.1. $P(\emptyset) = 0$.

Property 1.2.2. $P(E^c) = 1 - P(E)$ *for any* $E \in \mathcal{F}$.

Property 1.2.3. (*Monotonicity and subtraction*) *If* $E_1 \subseteq E_2$, *then* $P(E_1) \leq P(E_2)$ *and* $P(E_2 \backslash E_1) = P(E_2) - P(E_1)$.

Property 1.2.4. *If* $E_i \in \mathcal{F}$ *for* $i = 1, 2$, *then*

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1E_2).$$

Property 1.2.5. (*Inclusion-Exclusion Principle*) *Let* $E_1, E_2, \ldots, E_n$ *be* n *events. Then*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum_{i<j}^{n} P(E_iE_j) + \sum_{i<j<k}^{n} P(E_iE_jE_k) + \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right).$$

Property 1.2.6. (*Boole's inequality*) *For any sequence of events* $(E_n)_{n\in\mathbb{N}}$,

$$P\left(\bigcup_{n=1}^{\infty} E_n\right) \leq \sum_{n=1}^{\infty} P(E_n).$$

Property 1.2.7. (*Bonferroni's inequality*) *For any events* $E_1, E_2, \ldots, E_n$,

$$P\left(\bigcap_{i=1}^{n} E_i\right) \geq \sum_{i=1}^{n} P(E_i) - (n-1).$$

Property 1.2.8. *Let* $(E_n)_{n\in\mathbb{N}}$ *be an either nondecreasing or nonincreasing sequence of events. Then* $\lim_{n\to\infty} P(E_n) = P(\lim_{n\to\infty} E_n)$.

Proof. If $(E_n)_{n\in\mathbb{N}}$ is a nondecreasing sequence of events, we define

$$\begin{aligned} F_1 &= E_1, \\ F_n &= E_n \backslash E_{n-1} = E_n E_{n-1}^c, \quad n \geq 2. \end{aligned}$$

Then $\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i = \lim_{n\to\infty} E_n$ and $\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i = E_n$ for $n \geq 1$. Furthermore, $(F_n)_{n\in\mathbb{N}}$ is a sequence of mutually exclusive events. By condition (c) in Definition 1.2.3, we have

$$\begin{aligned} P\left(\lim_{n\to\infty} E_n\right) &= P\left(\bigcup_{i=1}^{\infty} E_i\right) = P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i) \\ &= \lim_{n\to\infty} \sum_{i=1}^{n} P(F_i) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} F_i\right) = \lim_{n\to\infty} P(E_n). \end{aligned}$$

If $(E_n)_{n\in\mathbb{N}}$ is a nonincreasing sequence of events, we let $E_n^* = E_n^c$ for $n \in \mathbb{N}$. Then $(E_n^*)_{n\in\mathbb{N}}$ is a nondecreasing sequence of events. By proof above, we have

$$\lim_{n\to\infty} P\left(E_n^*\right) = P\left(\bigcup_{n=1}^{\infty} E_n^*\right),$$

which is equivalent to

$$1 - \lim_{n\to\infty} P(E_n) = 1 - P\left(\bigcap_{n=1}^{\infty} E_n\right),$$

i.e.,

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

because $\lim_{n\to\infty} E_n = \bigcap_{n=1}^{\infty} E_n$. □

Example 1.2.3. *Suppose that a fair coin is tossed twice. Let*

$$E = \{(H,T),(T,H)\} \quad \textit{and} \quad F = \{(H,H),(T,T)\}.$$

Then

$$\begin{aligned} P(E) &= P\big(\{(H,T)\}\big) + P\big(\{(T,H)\}\big) = 1/4 + 1/4 = 1/2, \\ P(F) &= P\big(\{(H,H)\}\big) + P\big(\{(T,T)\}\big) = 1/4 + 1/4 = 1/2, \\ P(\Omega) &= P(E) + P(F) - P(EF) = 1/2 + 1/2 - 0 = 1. \end{aligned}$$

Example 1.2.4. (*Example 1.3.1 – Example 1.3.2*)

1.3 Conditional probability and independence

Definition 1.3.1. *Let E and F be two events with $P(F) > 0$. Then the probability of the event E occurring given that the event F has occurred is called the conditional probability of E given F, which is calculated by*

$$P(E|F) = \frac{P(EF)}{P(F)}.$$

Example 1.3.1. *An urn contains 7 red balls and 5 white balls. We draw two balls from the urn without replacement. What is the probability that both balls are red?*

Solution. Let

$$\begin{aligned} E_1 &= \text{the event that the first ball is red,} \\ E_2 &= \text{the event that the second ball is red.} \end{aligned}$$

Then

$$P(E_1) = 7/12 \quad \text{and} \quad P(E_2|E_1) = 6/11.$$

Thus, the probability that both balls are red is equal to

$$P(E_1E_2) = P(E_2|E_1)P(E_1) = (6/11)(7/12) = 7/22.$$

Example 1.3.2. *(Examples 1.4.1, 1.4.3, 1.4.6)*

Definition 1.3.2. *Two events E and F are said to be independent if*

$$P(EF) = P(E)P(F).$$

Otherwise, E and F are called to be dependent.

Remark 1.3.1. *The concepts of independent events and mutually exclusive events are different. The independence of two events means that the occurrence of one event does not affect the occurrence of the other, while mutual exclusivity of two events means that if one event occurs then the other cannot occur.*

Remark 1.3.2. *Some textbooks use conditional probability to define the independence of two events. That is, E_1 and E_2 are said to be independent if $P(E_2|E_1) = P(E_2)$. This definition requires $P(E_1) > 0$, while Definition 1.3.2 does not have any restriction on $P(E_1)$ or $P(E_2)$.*

Two trivial cases for the independence of two events E_1 and E_2 are below

(i) E_1 and E_2 are independent if $P(E_1)$ or $P(E_2)$ is 1.
(ii) E_1 and E_2 are independent if $P(E_1)$ or $P(E_2)$ is 0.

Definition 1.3.3. *Events $E_1, E_2, \ldots, E_n$ are said to be mutually independent if for any subcollection $\{E_{i_1}, E_{i_2}, \ldots\} \subseteq \{E_1, E_2, \ldots, E_n\}$,*

$$P\left(\bigcap_{i_j} E_{i_j}\right) = \prod_{i_j} P(E_{i_j})$$

where $\{i_1, i_2, \ldots\} \subseteq \{1, 2, \ldots, n\}$.

Remark 1.3.3. *By Definition 1.3.3, mutually independent events are always pairwise independent. However, a set of events that is pairwise independent is unnecessary to be mutually independent.*

Example 1.3.3. *(Example 1.4.9)*

1.4 Random variable and its distribution

1.4.1 Basic concept

Definition 1.4.1. *Any real-valued function defined on the sample space Ω is*

called a random variable. It is usually denoted by the capital letter such as X, Y, Z, *etc.*

Definition 1.4.2. *Let* X *be a random variable. Then the cumulative distribution function (abbreviated by cdf)* F_X *of* X *is defined by* $F_X(x) = P(X \leq x)$ *for* $x \in \mathbb{R}$.

Property 1.4.1. *Any cumulative distribution function* $F(x)$ *has three properties below*

(a) $F(x)$ *is a nondecreasing function of* x,

(b) $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$,

(c) $F(x)$ *is right-continuous, i.e.,* $\lim_{x \to x_0+} F(x) = F(x_0)$ *for any* $x_0 \in \mathbb{R}$.

1.4.2 Discrete and continuous random variables

Definition 1.4.3. *If the random variable* X *takes on at most a countable number of possible values, then* X *is said to be discrete.*

When X is discrete, we often use the probability mass function (pmf) to describe X. The pmf $p(x)$ of X is defined by

$$p(x_i) = P(X = x_i), \quad i = 0, 1, 2, \ldots,$$

where $x_0, x_1, x_2, \ldots$ are possible values of X.

Property 1.4.2. *Any pmf* $p(x)$ *satisfies two conditions*

(i) $0 \leq p(x_i) \leq 1$, *i=0, 1, 2, ...,*

(ii) $\sum_{i=0}^{\infty} p(x_i) = 1$.

Definition 1.4.4. *The random variable* X *is continuous if there exists a nonnegative function* $f(x)$ *defined on* $\mathbb{R}$ *such that*

$$P(X \in D) = \int_D f(x)dx$$

for any Borel set D. *The function* $f(x)$ *is called the probability density function (pdf) of* X.

Property 1.4.3. *Any pdf $f(x)$ satisfies*

(i) $f(x) \geq 0$,

(ii) $\displaystyle\int_{-\infty}^{\infty} f(x)dx = P(X \in (-\infty, \infty)) = 1$,

(iii) $\displaystyle\int_{a}^{b} f(x)dx = P(a \leq X \leq b)$,

(iv) $P(X = a) = 0$ *for any* $a \in \mathbb{R}$,

(v) $\displaystyle F(x) = \int_{-\infty}^{x} f(t)dt$ *and* $\displaystyle\frac{dF(x)}{dx} = f(x)$.

1.4.3 Transformation $Y = g(X)$

(i) If $Y = g(X)$ is one-to-one, then

$$P(Y = y) = P\left(X = g^{-1}(y)\right) \qquad \text{if } X \text{ is discrete.}$$

$$f_Y(y) = f_X\big(g^{-1}(y)\big)\left|\frac{dx}{dy}\right| \qquad \text{if } X \text{ is continuous.}$$

(ii) If $Y = g(X)$ is not one-to-one, we start with the cdf of Y below

$$P(Y \leq y) = P\left(g(X) \leq y\right) \text{ for } y \in \mathbb{R}.$$

Example 1.4.1. *Assume that the cdf of X is given by*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ x^2 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

Let $Y = X^2$. Find the distribution of Y.

Solution. Note that the transformation $y = x^2$ is not one-to-one. For $y \in [0, 1]$, we have

$$\begin{aligned} P(Y \leq y) &= P(X^2 \leq y) \\ &= P(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\ &= F_X(\sqrt{y}) \\ &= y, \end{aligned}$$

which means that Y has a uniform distribution on the interval $[0, 1]$.

Example 1.4.2. *Assume that X has a uniform distribution on the interval $[0, 1]$. Let $Y = -2\ln(X)$. Find the distribution of Y.*

Solution. Note that the transformation $y = -2\ln(x)$ is one-to-one. Solving $y = -2\ln(x)$ yields $x = e^{-y/2}$, which implies that $dx/dy = (-1/2)e^{-y/2}$. Thus,

$$f_Y(y) = f_X\left(e^{-y/2}\right)\left|\frac{dx}{dy}\right| = 1 \cdot \left|\frac{-1}{2}e^{-y/2}\right| = \frac{1}{2}e^{-y/2}, \quad y \geq 0.$$

Note that the domain $y \geq 0$ is from $0 \leq e^{-y/2} \leq 1$.

1.5 Expectation and related quantities

1.5.1 Definition of the expectation

Definition 1.5.1. *If X is a random variable with a cdf $F(x)$, then the expected value (also called the expectation or mean) of X is defined by*

$$E(X) = \int_{-\infty}^{\infty} x dF(x)$$

provided $\int_{-\infty}^{\infty} |x| dF(x) < \infty$.

(i) *When X is discrete,* $E(X) = \sum_x xP(X = x)$.

(ii) *When X is continuous,* $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.

Property 1.5.1.

(i) *If $Y = g(X)$, then*

$$E(Y) = E[g(X)] = \begin{cases} \sum_x g(x)P(X = x) & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

provided $\sum_x |g(x)|P(X = x) < \infty$ *for the discrete case and* $\int_{-\infty}^{\infty} |g(x)|f(x)dx < \infty$ *for the continuous case.*

(ii) $E\Big(c_1g_1(X) + c_2g_2(X)\Big) = c_1E\Big(g_1(X)\Big) + c_2E\Big(g_2(X)\Big)$.

1.5.2 Variance

Definition 1.5.2. *The variance of random variable X is defined by*

$$\mathrm{Var}(X) = E\Big([X - E(X)]^2\Big)$$

provided $\int_{-\infty}^{\infty} x^2 dF_X(x) < \infty$. *We often use* σ^2 *to denote* $\mathrm{Var}(X)$, *where* $\sigma > 0$ *is called the standard deviation of* X.

Property 1.5.2.

(i) $\mathrm{Var}(X) = E(X^2) - [E(X)]^2$.
(ii) $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$, *where* a *and* b *are real numbers.*

Example 1.5.1. *Assume that the pdf of* X *is given by*

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

Find the mean and variance of X.

Solution. Direct calculations show that

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)dx = \frac{1}{2}\int_{-1}^{1} x(x+1)dx = \frac{1}{2}\left(\frac{x^3}{3} + \frac{x^2}{2}\right)\Big|_{-1}^{1} = \frac{1}{3},$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \frac{1}{2}\int_{-1}^{1} x^2(x+1)dx = \frac{1}{2}\left(\frac{x^4}{4} + \frac{x^3}{3}\right)\Big|_{-1}^{1} = \frac{1}{3},$$

which imply that

$$\sigma^2 = \mathrm{Var}(X) = 1/3 - (1/3)^2 = 2/9.$$

Example 1.5.2. *Assume that the pdf of* X *is given by*

$$f(x) = \frac{1}{x^2}, \quad x \geq 1.$$

Then $E(X)$ *does not exist.*

Proof. A direct calculation shows that

$$\int_{-\infty}^{\infty} |x| f(x)dx = \int_{1}^{\infty} x\frac{1}{x^2}dx = \int_{1}^{\infty} \frac{1}{x}dx = \lim_{b\to\infty}\int_{1}^{b} \frac{dx}{x}$$
$$= \lim_{b\to\infty} \ln(x)\Big|_{1}^{b} = \lim_{b\to\infty} \ln(b) = \infty. \qquad \square$$

1.5.3 Moment generating function

Definition 1.5.3. *Let X be a random variable such that $M_X(t) = E(e^{tX})$ exists for $|t| < h$ with some $h > 0$. Then $M_X(t)$, as a function of t, is called the moment generating function (abbreviated by mgf) of X.*

Property 1.5.3.

(i) $X \stackrel{d}{=} Y$ *iff* $M_X(t) = M_Y(t)$ *for all* $t \in (-h, h)$, *where* $h > 0$.

(ii) $M_X^{(m)}(0) = M_X^{(m)}(t)\Big|_{t=0} = E(X^m)$.

(iii) $E(X) = M_X^{(1)}(0)$ *and* $\mathrm{Var}(X) = M_X^{(2)}(0) - \left[M_X^{(1)}(0)\right]^2$.

(iv) *Let* $Y = aX + b$, *where* a *and* b *are real numbers. Then*

$$M_Y(t) = e^{bt} M_X(at).$$

2

Multivariate Distributions

2.1 Distribution of an n-dimensional random vector

2.1.1 Distribution of a random vector

Definition 2.1.1. *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ *be an* n*-dimensional random vector. Then the joint cdf of* $\mathbf{X}$ *is defined by*

$$F(\mathbf{x}) = F(x_1, x_2, \ldots, x_n) = P(X_i \leq x_i,\ i = 1, \ldots, n)$$

for $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$.

Definition 2.1.2. *Let* $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$, *where* $\mathbf{X}_1 \in \mathbb{R}^m$ *and* $\mathbf{X}_2 \in \mathbb{R}^{n-m}$. *Then the marginal cdfs of* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are defined by*

$$\begin{aligned} F_{\mathbf{X}_1}(\mathbf{x}_1) &= \lim_{\substack{x_j \to \infty \\ j=m+1,\ldots,n}} F(x_1, \ldots x_m, x_{m+1}, \ldots x_n) \\ &= P(X_i \leq x_i,\ i = 1, \ldots, m,\ X_j < \infty,\ j = m+1, \ldots, n), \\ F_{\mathbf{X}_2}(\mathbf{x}_2) &= \lim_{\substack{x_j \to \infty \\ j=1,\ldots,m}} F(x_1, \ldots x_m, x_{m+1}, \ldots x_n) \\ &= P(X_i < \infty,\ i = 1, \ldots, m,\ X_j \leq x_j,\ j = m+1, \ldots, n), \end{aligned}$$

for $\mathbf{x}_1 \in \mathbb{R}^m$ *and* $\mathbf{x}_2 \in \mathbb{R}^{n-m}$.

2.1.2 Distribution of a discrete random vector

Definition 2.1.3. *When* $\mathbf{X}$ *is discrete, the joint pmf of* $\mathbf{X}$ *is defined by*

$$p(\mathbf{x}) = P(X_i = x_i,\ i = 1, \ldots, n),$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T$ *is any possible value of* $\mathbf{X}$.

Remark 2.1.1. *Any joint pmf* $p(\mathbf{x})$ *satisfies two conditions below*

(i) $0 \le p(\mathbf{x}) \le 1, \quad \mathbf{x} \in \mathbb{R}^n$,

(ii) $\sum_{\mathbf{x}} p(\mathbf{x}) = 1$.

Definition 2.1.4. *Let* $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$. *Then the marginal pmf of* $\mathbf{X}_1$ *is defined by*

$$\begin{aligned} p_{\mathbf{X}_1}(\mathbf{x}_1) &= P(X_i = x_i,\ i = 1, \ldots, m) \\ &= \sum_{x_{m+1}, \ldots, x_n} p(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) \end{aligned}$$

for $\mathbf{x}_1 = (x_1, x_2, \ldots, x_m)^T \in \mathbb{R}^m$. *Similarly, the marginal pmf of* $\mathbf{X}_2$ *is defined by*

$$\begin{aligned} p_{\mathbf{X}_2}(\mathbf{x}_2) &= P(X_i = x_i,\ i = m+1, \ldots, n) \\ &= \sum_{x_1, \ldots, x_m} p(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) \end{aligned}$$

for $\mathbf{x}_2 = (x_{m+1}, \ldots, x_n)^T \in \mathbb{R}^{n-m}$.

Definition 2.1.5. *Let* $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$. *Then the conditional pmf of* $\mathbf{X}_1$ *given* $\mathbf{X}_2 = \mathbf{x}_2$ *is defined by*

$$p_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) = \frac{p(\mathbf{x})}{p_{\mathbf{X}_2}(\mathbf{x}_2)}$$

for $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. *Similarly, the conditional pmf of* $\mathbf{X}_2$ *given* $\mathbf{X}_1 = \mathbf{x}_1$ *is defined by*

$$p_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2|\mathbf{x}_1) = \frac{p(\mathbf{x})}{p_{\mathbf{X}_1}(\mathbf{x}_1)}$$

for $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

2.1.3 Distribution of a continuous random vector

Definition 2.1.6. *When* $\mathbf{X}$ *is continuous, the joint pdf of* $\mathbf{X}$ *is defined by*

$$f(\mathbf{x}) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n},$$

where $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

Remark 2.1.2. *Any joint pdf $f(\mathbf{x})$ satisfies two conditions below*

(i) $f(\mathbf{x}) \geq 0$ *for any* $\mathbf{x} \in \mathbb{R}^n$,

(ii) $\int\limits_{\mathbb{R}^n} f(\mathbf{x})d\mathbf{x} = 1.$

Definition 2.1.7. *Let* $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$. *Then the marginal pdf of* $\mathbf{X}_1$ *is defined by*

$$f_{\mathbf{X}_1}(\mathbf{x}_1) = \int\limits_{\mathbb{R}^{n-m}} f(\mathbf{x})d\mathbf{x}_2$$

for $\mathbf{x}_1 = (x_1, x_2, \ldots, x_m)^T \in \mathbb{R}^m$, *where* $\mathbf{x}_2 = (x_{m+1}, \ldots, x_n)^T \in \mathbb{R}^{n-m}$. *Similarly, the marginal pdf of* $\mathbf{X}_2$ *is defined by*

$$f_{\mathbf{X}_2}(\mathbf{x}_2) = \int\limits_{\mathbb{R}^m} f(\mathbf{x})d\mathbf{x}_1$$

for $\mathbf{x}_2 = (x_{m+1}, \ldots, x_n)^T \in \mathbb{R}^{n-m}$, *where* $\mathbf{x}_1 = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$.

Definition 2.1.8. *Let* $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$. *Then the conditional pdf of* $\mathbf{X}_1$ *given* $\mathbf{X}_2 = \mathbf{x}_2$ *is defined by*

$$f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) = \frac{f(\mathbf{x})}{f_{\mathbf{X}_2}(\mathbf{x}_2)}$$

for $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. *Similarly, the conditional pdf of* $\mathbf{X}_2$ *given* $\mathbf{X}_1 = \mathbf{x}_1$ *is defined by*

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2|\mathbf{x}_1) = \frac{f(\mathbf{x})}{f_{\mathbf{X}_1}(\mathbf{x}_1)}$$

for $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

2.2 Expectation and related quantities

Definition 2.2.1. *Let $\mathbf{X} \in \mathbb{R}^n$ be a random vector. Then the expected value and covariance matrix of $\mathbf{X}$ are defined by*

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix},$$

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right] = (\sigma_{ij})_{n \times n},$$

where

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = E\Big([X_i - E(X_i)][X_j - E(X_j)]\Big)$$

is the covariance between X_i and X_j, $i, j = 1, 2, \ldots, n$.

Remark 2.2.1. *The covariance matrix $\boldsymbol{\Sigma}$ is positive semi-definite, i.e., $\mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a} \geq 0$ for any $\mathbf{a} \in \mathbb{R}^n$.*

Definition 2.2.2. *The correlation coefficient ρ_{ij} between X_i and X_j is defined by*

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

for $i, j = 1, 2, \ldots, n$.

Remark 2.2.2. *$|\rho_{ij}| \leq 1$ for $i, j = 1, 2, \ldots, n$.*

Definition 2.2.3. *The joint moment generating function of $\mathbf{X}$ is defined by*

$$M_{\mathbf{X}}(\mathbf{t}) = E\big(e^{\mathbf{t}^T \mathbf{X}}\big) = E\left(e^{t_1 X_1 + \cdots + t_n X_n}\right)$$

provided $E\big(e^{t_1 X_1 + \cdots + t_n X_n}\big)$ exists for $|t_i| < h_i$, where $h_i > 0$ for $i = 1, 2, \ldots, n$ and $\mathbf{t} = (t_1, t_2, \ldots, t_n)^T$.

Remark 2.2.3. *For a random vector, there is a property that is similar to Property 1.5.3. For example, if $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, where $\mathbf{A}$ is a matrix and $\mathbf{b}$ is a vector, then we have*

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp\left(\mathbf{t}^T \mathbf{b}\right) M_{\mathbf{X}}(\mathbf{A}^T \mathbf{t}).$$

Theorem 2.2.1. *(Theorems 2.8.1 and 2.8.2) Let*

$$T = \sum_{i=1}^{n} a_i X_i \quad \textit{and} \quad W = \sum_{i=1}^{m} b_i Y_i.$$

Then

(i) $E(T) = \sum_{i=1}^{n} a_i E(X_i)$ *and* $E(W) = \sum_{i=1}^{m} b_i E(Y_i)$.

(ii) $\mathrm{Cov}(T, W) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j \mathrm{Cov}(X_i, Y_j)$.

(iii) $\mathrm{Var}(T) = \mathrm{Cov}(T, T) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2 \sum \sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j)$.

Proof. (see page 151) □

2.3 Independence

Definition 2.3.1. *The random variables $X_1, X_2, \ldots, X_n$ are said to be independent if and only if*

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i) \qquad (\textit{discrete case})$$

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \qquad (\textit{continuous case})$$

for $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

2.4 Special case – bivariate distribution

Definition 2.4.1. *Let* $\mathbf{X} = (X_1, X_2)^T$ *be a discrete random vector. Then*

$$\text{joint pmf of } \mathbf{X}: \ p(x_1, x_2) = P(X_1 = x_1, X_2 = x_2),$$
$$(i) \quad 0 \le p(x_1, x_2) \le 1,$$
$$(ii) \quad \sum_{x_1} \sum_{x_2} p(x_1, x_2) = 1,$$
$$\text{marginal pmf of } X_1: \ p_{X_1}(x_1) = \sum_{x_2} p(x_1, x_2),$$
$$\text{marginal pmf of } X_2: \ p_{X_2}(x_2) = \sum_{x_1} p(x_1, x_2),$$
$$\text{conditional pmf of } X_1 \text{ given } X_2 = x_2: p_{X_1|X_2}(x_1|x_2) = \frac{p(x_1, x_2)}{p_{X_2}(x_2)},$$
$$\text{conditional pmf of } X_2 \text{ given } X_1 = x_1: p_{X_2|X_1}(x_2|x_1) = \frac{p(x_1, x_2)}{p_{X_1}(x_1)}.$$

Definition 2.4.2. *Let* $\mathbf{X} = (X_1, X_2)^T$ *be a continuous random vector. Then*

$$\text{joint pdf of } \mathbf{X}: \ f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2},$$
$$(i) \quad f(x_1, x_2) \ge 0,$$
$$(ii) \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1,$$
$$\text{marginal pdf of } X_1: \ f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2,$$
$$\text{marginal pdf of } X_2: \ f_{X_2}(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1,$$
$$\text{conditional pdf of } X_1 \text{ given } X_2 = x_2: f_{X_1|X_2}(x_1|x_2) = \frac{f(x_1, x_2)}{f_{X_2}(x_2)},$$
$$\text{conditional pdf of } X_2 \text{ given } X_1 = x_1: f_{X_2|X_1}(x_2|x_1) = \frac{f(x_1, x_2)}{f_{X_1}(x_1)}.$$

Definition 2.4.3. *The joint mgf of* X_1 *and* X_2 *is defined by*

$$M(t_1, t_2) = E\left(e^{t_1 X_1 + t_2 X_2}\right)$$

provided $E\left(e^{t_1 X_1 + t_2 X_2}\right)$ *exists for* $|t_1| < h_1$, $|t_2| < h_2$, *where* $h_1 > 0, h_2 > 0$.

Remark 2.4.1. *$M(t_1, t_2)$ can be used to calculate $E(X_1), E(X_2), \mathrm{Var}(X_1), \mathrm{Var}(X_2)$, and $\mathrm{Cov}(X_1, X_2)$.*

Definition 2.4.4. *The conditional expected value of $u(X_1)$ given $X_2 = x_2$ is defined by*

$$E[u(X_1)|X_2 = x_2) = \begin{cases} \sum_{x_1} u(x_1) p_{X_1|X_2}(x_1|x_2) & \text{discrete case,} \\ \int_{-\infty}^{\infty} u(x_1) f_{X_1|X_2}(x_1|x_2) dx_1 & \text{continuous case.} \end{cases}$$

When $u(X_1) = X_1$, $E(X_1|X_2 = x_2)$ is called the conditional expected value of X_1 given $X_2 = x_2$.

When $u(X_1) = [X_1 - E(X_1|X_2 = x_2)]^2$, $E[u(X_1)|X_2 = x_2]$ is called the conditional variance of X_1 given $X_2 = x_2$.

Property 2.4.1.

(i) $E[E(X_2|X_1)] = E(X_2)$ *and* $E[E(X_1|X_2)] = E(X_1)$.

(ii) $\mathrm{Var}[E(X_2|X_1)] \leq \mathrm{Var}(X_2)$.

Proof. (see page 114) □

Example 2.4.1. *Let the joint pmf of (X_1, X_2) be defined by*

$$p(x_1, x_2) = (x_1 + x_2)/21, \quad x_1 = 1, 2, 3, \ x_2 = 1, 2.$$

Then

(i) $0 \leq p(x_1, x_2) \leq 1$.

(ii) $\sum_{x_1} \sum_{x_2} p(x_1, x_2) = 1$.

(iii) $p_{X_1}(x_1) = \sum_{x_2} p(x_1, x_2) = \begin{cases} 5/21 & \text{if} \quad x_1 = 1, \\ 7/21 & \text{if} \quad x_1 = 2, \\ 9/21 & \text{if} \quad x_1 = 3. \end{cases}$

(iv) $p_{X_2}(x_2) = \sum_{x_1} p(x_1, x_2) = \begin{cases} 9/21, & \textit{if} \quad x_2 = 1, \\ 12/21, & \textit{if} \quad x_2 = 2. \end{cases}$

(v) $p(x_1|x_2 = 1) = \begin{cases} 2/9, & \textit{if} \quad x_1 = 1, \\ 3/9, & \textit{if} \quad x_1 = 2, \\ 4/9, & \textit{if} \quad x_1 = 3. \end{cases}$

(vi) $p(x_1|x_2 = 2) = \begin{cases} 3/12 & \textit{if} \quad x_1 = 1, \\ 4/12 & \textit{if} \quad x_1 = 2, \\ 5/12 & \textit{if} \quad x_1 = 3. \end{cases}$

(vii) $p(x_2|x_1) = ...$

(viii) *Since* $p(1,1) = 2/21 \neq (5/21)(9/21) = p_{X_1}(1)p_{X_2}(1)$, X_1 *and* X_2 *are dependent.*

(ix)

$$\begin{aligned} \mu_1 = E(X_1) &= 1(5/21) + 2(7/21) + 3(9/21) = 46/21, \\ E(X_1^2) &= 1^2(5/21) + 2^2(7/21) + 3^2(9/21) = 114/21, \\ \sigma_1^2 = \sigma_{11} &= 114/21 - (46/21)^2 = 278/21^2, \end{aligned}$$

$$\begin{aligned} \mu_2 &= 1(9/21) + 2(12/21) = 33/21, \\ E(X_2^2) &= 1^2(9/21) + 2^2(12/21) = 57/21, \\ \sigma_2^2 = \sigma_{22} &= 57/21 - (33/21)^2 = 108/21^2, \end{aligned}$$

$$\begin{aligned} E(X_1X_2) &= 2/21 + 2(3/21) + 2(3/21) + 4(4/21) + 3(4/21) + 6(5/21) \\ &= 72/21, \\ \mathrm{Cov}(X_1, X_2) &= 72/21 - (46/21)(33/21) = -6/21^2, \\ \rho &= \frac{-6/21^2}{\sqrt{(278/21^2)(108/21^2)}} = \frac{-1}{\sqrt{3 \times 278}}. \end{aligned}$$

(x) $M(t_1, t_2) = E(e^{t_1X_1 + t_2X_2}) = \cdots\cdots$

(xi)

$$\begin{aligned}
E(X_1|X_2=1) &= 1(2/9)+2(3/9)+3(4/9) = 20/9,\\
E(X_1|X_2=2) &= 1(3/12)+2(4/12)+3(5/12) = 13/6,\\
E[E(X_1|X_2)] &= (20/9)P(X_2=1)+(26/12)P(X_2=2)\\
&= (20/9)(9/21)+(26/12)(12/21)\\
&= 46/21\\
&= E(X_1),\\
\mathrm{Var}[E(X_1|X_2)] &= (20/9)^2(9/21)+(26/12)^2(12/21)-(46/21)^2 = \frac{1}{3\times 21^2},\\
\mathrm{Var}(X_1) &\geq \mathrm{Var}\Big(E(X_1|X_2)\Big).
\end{aligned}$$

Example 2.4.2. *(Example 2.1.6 on page 91) Let the joint pdf of X_1 and X_2 be defined by*

$$f(x_1,x_2) = \begin{cases} x_1+x_2 & \text{if } 0<x_1<1,\ 0<x_2<1,\\ 0 & \text{otherwise.}\end{cases}$$

Then

(i) $f(x_1,x_2)\geq 0$.

(ii) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2)dx_1dx_2 = \int_0^1\int_0^1 (x_1+x_2)dx_1dx_2 = \int_0^1 (1/2+x_2)dx_2 = 1.$

(iii) $\displaystyle f_{X_1}(x_1) = \int_{-\infty}^{\infty} f(x_1,x_2)dx_2 = \int_0^1 (x_1+x_2)dx_2 = x_1+1/2, \quad 0<x_1<1.$

(iv) $f_{X_2}(x_2) = x_2+1/2, \quad 0<x_2<1.$

(v) $\displaystyle f(x_2|x_1) = \frac{f(x_1,x_2)}{f_{X_1}(x_1)} = \frac{x_1+x_2}{x_1+1/2}, \quad 0<x_1<1,\ 0<x_2<1.$

(vi) $\displaystyle f(x_1|x_2) = \frac{f(x_1,x_2)}{f_{X_2}(x_2)} = \frac{x_1+x_2}{x_2+1/2}, \quad 0<x_1<1,\ 0<x_2<1.$

(vii) *Since $f(x_1,x_2)\neq f_{X_1}(x_1)f_{X_2}(x_2)$, X_1 and X_2 are dependent.*

(viii)

$$\begin{aligned}
\mu_1 = E(X_1) &= \int_0^1 x_1(x_1 + 1/2)dx_1 = 7/12, \\
E(X_1^2) &= \int_0^1 x_1^2(x_1 + 1/2)dx_1 = 5/12, \\
\sigma_1^2 = \sigma_{11} &= 5/12 - (7/12)^2 = 11/12^2,
\end{aligned}$$

$$\begin{aligned}
\mu_2 = E(X_2) &= \int_0^1 x_2(x_2 + 1/2)dx_2 = 7/12, \\
E(X_2^2) &= \int_0^1 x_2^2(x_2 + 1/2)dx_2 = 5/12, \\
\sigma_2^2 = \sigma_{22} &= 5/12 - (7/12)^2 = 11/12^2,
\end{aligned}$$

$$\begin{aligned}
E(X_1X_2) &= \int_0^1 \int_0^1 x_1x_2(x_1 + x_2)dx_1dx_2 = ... = 1/3, \\
\mathrm{Cov}(X_1, X_2) &= 1/3 - (7/12)(7/12) = -1/12^2, \\
\rho &= \frac{-1/12^2}{\sqrt{(11/12^2)(11/12^2)}} = -\frac{1}{11}.
\end{aligned}$$

(ix) $M(t_1, t_2) = E(e^{t_1X_1+t_2X_2}) = \cdots\cdots$

(x)

$$\begin{aligned}
E(X_1|X_2 = x_2) &= \frac{1}{x_2 + 1/2}\int_0^1 x_1(x_1 + x_2)dx_1 = \frac{x_2/2 + 1/3}{x_2 + 1/2}, \\
E[E(X_1|X_2)] &= \int_0^1 \frac{x_2/2 + 1/3}{x_2 + 1/2}(x_2 + 1/2)dx_2 \\
&= \int_0^1 (x_2/2 + 1/3)dx_2 \\
&= \frac{1}{4}x_2^2 + \frac{1}{3}x_2\Big|_0^1 \\
&= 7/12 \\
&= E(X_1).
\end{aligned}$$

2.5 Transformation

Let $\mathbf{X} \in \mathbb{R}^n$ be a continuous random vector with a joint pdf $f_{\mathbf{X}}(\mathbf{x})$. For $i = 1, 2, \ldots, n$, let $Y_i = u_i(X_1, X_2, \ldots, X_n)$ be a one-to-one transformation. Then the joint pdf of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\Big(w_1(\mathbf{y}), w_2(\mathbf{y}), \ldots, w_n(\mathbf{y})\Big)|J|,$$

where $w_1(\mathbf{y}), w_2(\mathbf{y}), \ldots, w_n(\mathbf{y})$ are solutions of the system of equations

$$y_i = u_i(x_1, x_2, \ldots, x_n), \quad i = 1, 2, \ldots, n$$

for $x_1, x_2, \ldots, x_n$, i.e., $x_i = w_i(\mathbf{y})$ $(i = 1, 2 \ldots, n)$ and

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}.$$

Question: *How do you determine the domain of $f_{\mathbf{Y}}(\mathbf{y})$?*

Example 2.5.1. *(Example 2.2.1 – Example 2.2.4)*

Example 2.5.2. *(Example 2.2.5) Assume that the joint pdf of X_1 and X_2 is defined by*

$$f(x_1, x_2) = 10x_1x_2^2, \quad 0 < x_1 < x_2 < 1.$$

Let $Y_1 = X_1/X_2$ and $Y_2 = X_2$. Find the joint pdf of Y_1 and Y_2 and two marginal pdfs.

Solution. *Solving $y_1 = x_1/x_2$ and $y_2 = x_2$ for x_1 and x_2 yields that*

$$x_1 = y_1y_2 \quad \textit{and} \quad x_2 = y_2,$$

which imply that

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2.$$

Thus, the joint pdf of Y_1 and Y_2 is equal to

$$g(y_1, y_2) = f(y_1y_2, y_2)|y_2| = 10(y_1y_2)y_2^2|y_2| = 10y_1y_2^4, \quad 0 < y_1 < 1,\ 0 < y_2 < 1.$$

Furthermore, two marginal pdfs are given by

$$g_{Y_1}(y_1) = \int_0^1 10y_1y_2^4 dy_2 = 2y_1, \quad 0 < y_1 < 1,$$
$$g_{Y_2}(y_2) = \int_0^1 10y_1y_2^4 dy_1 = 5y_2^4, \quad 0 < y_2 < 1.$$

Example 2.5.3. *Let X_1 and X_2 be independent and identically distributed random variables with a pdf $\lambda e^{-\lambda x}$, $x \geq 0$, where $\lambda > 0$ is a parameter. Find the pdf of $Y_1 = X_1 + X_2$.*

Solution. *We introduce a transformation by letting*

$$Y_1 = X_1 + X_2,$$
$$Y_2 = X_2.$$

Solving $y_1 = x_1 + x_2$ and $y_2 = x_2$ for x_1 and x_2 yields that

$$x_1 = y_1 - y_2,$$
$$x_2 = y_2,$$

which imply that

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 0 & 1 \end{vmatrix} = 1.$$

Since the joint pdf of X_1 and X_2 is

$$f(x_1, x_2) = \lambda e^{-\lambda x_1} \times \lambda e^{-\lambda x_2} = \lambda^2 e^{-\lambda(x_1+x_2)}, \quad x_1 \geq 0, x_2 \geq 0,$$

the joint pdf of Y_1 and Y_2 is equal to

$$g(y_1, y_2) = f(y_1 - y_2, y_2)|1| = \lambda^2 e^{-\lambda y_1}, \quad 0 \leq y_2 \leq y_1.$$

Furthermore, the marginal pdf of Y_1 is

$$g_{Y_1}(y_1) = \int_{-\infty}^{\infty} g(y_1, y_2) dy_2 = \int_0^{y_1} \lambda^2 e^{-\lambda y_1} dy_2 = \lambda^2 y_1 e^{-\lambda y_1}, \quad y_1 \geq 0.$$

Example 2.5.4. *(Example 2.7.2) Let X_1, X_2 and X_3 be independent and have the same pdf e^{-x}, $x > 0$. Let $Y_1 = X_1/(X_1+X_2+X_3)$, $Y_2 = X_2/(X_1+X_2+X_3)$ and $Y_3 = X_1 + X_2 + X_3$. Then Y_1, Y_2, and Y_3 are dependent.*

Proof. Note that the joint pdf of X_1, X_2 and X_3 is

$$f(x_1, x_2, x_3) = e^{-x_1}e^{-x_2}e^{-x_3} = e^{-(x_1+x_2+x_3)}, \quad x_1 > 0, x_2 > 0, x_3 > 0.$$

Solving $y_1 = x_1/(x_1+x_2+x_3)$, $y_2 = x_2/(x_1+x_2+x_3)$ and $y_3 = x_1+x_2+x_3$ for x_1, x_2 and x_3 yields that

$$\begin{aligned} x_1 &= y_1 y_3, \\ x_2 &= y_2 y_3, \\ x_3 &= (1-y_1-y_2)y_3, \end{aligned}$$

which imply that

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \frac{\partial x_1}{\partial y_3} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \frac{\partial x_2}{\partial y_3} \\ \frac{\partial x_3}{\partial y_1} & \frac{\partial x_3}{\partial y_2} & \frac{\partial x_3}{\partial y_3} \end{vmatrix} = \begin{vmatrix} y_3 & 0 & y_1 \\ 0 & y_3 & y_2 \\ -y_3 & -y_3 & 1-y_1-y_2 \end{vmatrix} = y_3^2.$$

Thus, the joint pdf of Y_1, Y_2 and Y_3 is equal to

$$g(y_1, y_2, y_3) = f\Big(y_1 y_3,\, y_2 y_3,\, (1-y_1-y_2)y_3\Big)|J| = y_3^2 e^{-y_3},$$

where $y_3 > 0$, $y_1 > 0$, $y_2 > 0$, $1-y_1-y_2 > 0$. Furthermore, three marginal pdfs are

$$\begin{aligned} g_{Y_1}(y_1) &= \int_0^{1-y_1}\int_0^{\infty} y_3^2 e^{-y_3} dy_3 dy_2 = \Gamma(3)\int_0^{1-y_1} dy_2 = 2(1-y_1),\ 0<y_1<1, \\ g_{Y_2}(y_2) &= 2(1-y_2), \quad 0<y_2<1, \\ g_{Y_3}(y_3) &= \int_0^1\int_0^{1-y_1} y_3^2 e^{-y_3} dy_2 dy_1 = y_3^2 e^{-y_3}\int_0^1\int_0^{1-y_1} dy_2 dy_1 \\ &= y_3^2 e^{-y_3}\int_0^1 (1-y_1)dy_1 = \frac{1}{2}y_3^2 e^{-y_3}, \quad y_3>0. \end{aligned}$$

Clearly, $g(y_1, y_2, y_3) \neq g_{Y_1}(y_1)g_{Y_2}(y_2)g_{Y_3}(y_3)$. Thus, Y_1, Y_2 and Y_3 are dependent.

It is worth mentioning that we can also use the joint conditional pdf of Y_2 and Y_3 given Y_1 to claim that Y_1, Y_2 and Y_3 are dependent. Indeed,

$$g(y_2, y_3|y_1) = \frac{g(y_1, y_2, y_3)}{g_{Y_1}(y_1)} = \frac{y_3^2 e^{-y_3}}{2(1-y_1)}, \quad y_3>0,\, y_1>0,\, 0<y_2<1-y_1,$$

which depends on y_1.

□

3

Some Special Distributions

3.1 Bernoulli and binomial distributions

Bernoulli trial is any statistical experiment or activity whose outcome can be classified as either a success or a failure. The experiment of tossing a coin is a typical example of the Bernoulli trial.

Definition 3.1.1. *Let $X = 1$ and $X = 0$ denote the outcomes of success and failure of the Bernoulli trial, respectively. Then the distribution of X is called a Bernoulli distribution with parameter $p \in [0, 1]$, where p represents the probability of success. The random variable X is called a Bernoulli random variable. We write $X \sim b(1, p)$.*

By Definition 3.1.1, the pmf and cdf of $X \sim b(1, p)$ are respectively given by

$$\begin{aligned} f_X(x) =& p^x(1-p)^{1-x}, \quad x = 0, 1, \\ F_X(x) =& \begin{cases} 0 & \text{if } x < 0, \\ 1-p & \text{if } 0 \le x < 1, \\ 1 & \text{if } x \ge 1. \end{cases} \end{aligned} \tag{3.1}$$

Property 3.1.1. *Let $X \sim b(1, p)$. Then the expectation, variance, and moment generating function of X are given by*

$$\begin{aligned} E(X) &= p, \\ \mathrm{Var}(X) &= p(1-p), \\ M_X(t) &= 1 - p + pe^t, \quad t \in (-\infty, \infty). \end{aligned} \tag{3.2}$$

We omit the proof here because it follows from the definitions of expectation, variance, and moment-generating function immediately. Furthermore, if $X \sim b(1, p)$, the raw and central moments of X are respectively given by

$$\begin{aligned} \mu'_k &= E(X^k) = p, \\ \mu_k &= E[(X-p)^k] = p(1-p)[(1-p)^{k-1} + (-1)^k p^{k-1}] \end{aligned}$$

for any $k \in \mathbb{N}$.

The Bernoulli distribution $b(1,p)$ *contains a parameter* p*, which is the probability of having the outcome success. In practice, we usually need to estimate this parameter. If a screening test for a specific cancer, for example, is defined as the Bernoulli trial and the result of the screening test is classified as false-positive (success) or non-false-positive (failure). There is no way to determine the probability of receiving a false-positive test if only one person takes the screening test. To estimate the probability of receiving a false-positive test, we need to recruit many people who are eligible to take the same screening test and use the sample proportion of a false-positive test to estimate the parameter* p*. Clearly, the numerator of this sample proportion is the number of successes (false-positive tests) after repeating the Bernoulli trial* n *times independently. What is its distribution? The answer to this question is related to the binomial distribution that is defined below.*

Definition 3.1.2. *Let* X *denote the number of successes after repeating the same Bernoulli trial* n *times independently. Then the distribution of* X *is called a binomial distribution with parameters* (n,p)*. The random variable* X *is called a binomial random variable. We write* $X \sim b(n,p)$*.*

By Definition 3.1.2, we see that the pmf and cdf of $X \sim b(n,p)$ *are given by*

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n,$$

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \displaystyle\sum_{i=0}^{[x]} \binom{n}{i} p^i (1-p)^{n-i} & \text{if } x \geq 0, \end{cases} \tag{3.3}$$

where $[x]$ *denotes the integer part of* x*.*

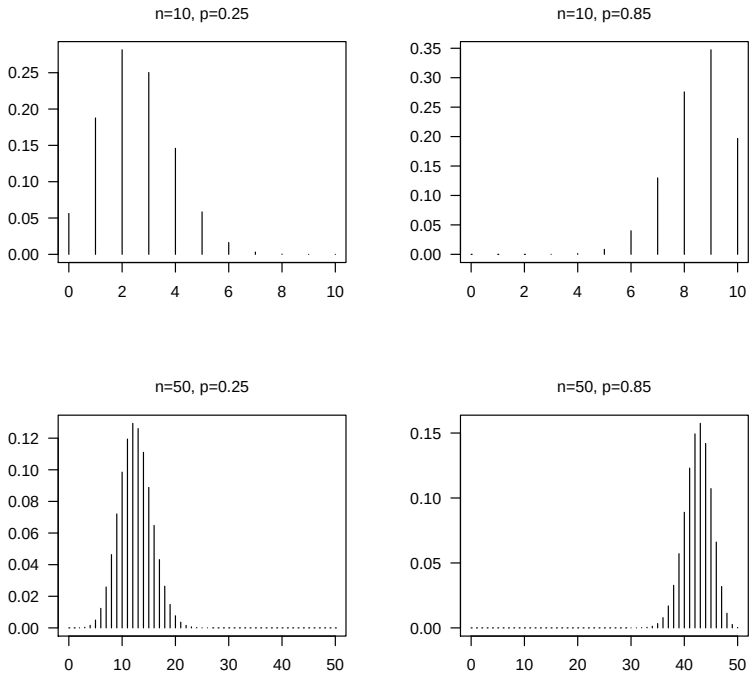


Figure 3.1.1 PMF of the binomial distribution

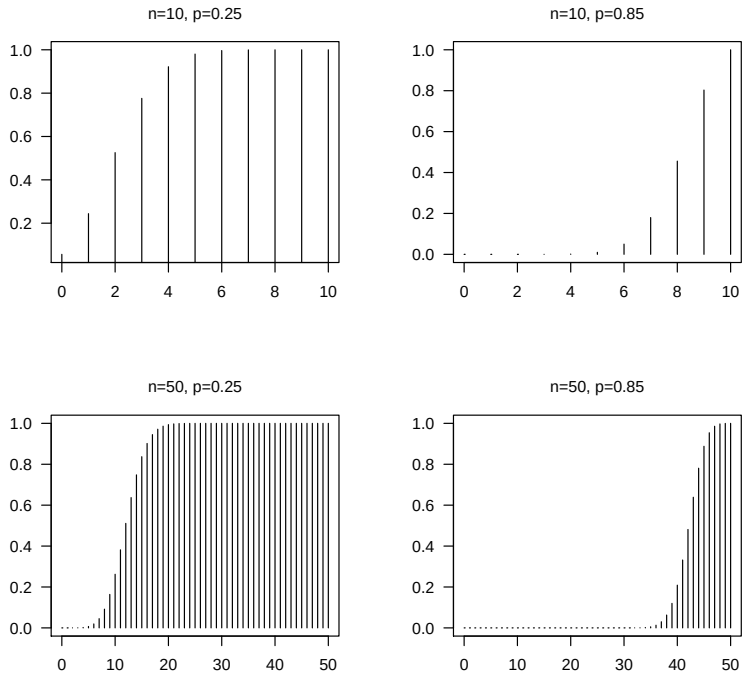


Figure 3.1.2 CDF of the binomial distribution

Property 3.1.2. *Let $X \sim b(n, p)$. Then the expectation and variance of X are given by*

$$E(X) = np \quad \textit{and} \quad \mathrm{Var}(X) = np(1-p), \tag{3.4}$$

respectively.

Proof. Using the pmf given by (3.3) and applying the definition of the expec-

tation, we have

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x\, f_X(x) \\
&= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \qquad (3.5) \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\
&= np \sum_{i=0}^{n-1} \frac{(n-1)!}{i!(n-1-i)!} p^i (1-p)^{n-1-i} \quad (\text{by } i = x-1) \\
&= np(p+1-p)^{n-1} \\
&= np.
\end{aligned}$$

Here the second-to-last equality in (3.5) follows from an application of the binomial theorem.

To show the variance of binomial distribution $b(n,p)$, we need $E(X^2)$, which can be similarly calculated below:

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{n} x^2\, f_X(x) \\
&= \sum_{x=0}^{n} [x(x-1)+x] \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} x(x-1) \binom{n}{x} p^x (1-p)^{n-x} + \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=2}^{n} x(x-1) \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} + np \qquad (3.6) \\
&= n(n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)!(n-x)!} p^{x-2} (1-p)^{n-x} + np \\
&= n(n-1)p^2 \sum_{i=0}^{n-2} \frac{(n-2)!}{i!(n-2-i)!} p^i (1-p)^{n-2-i} + np \\
&= n(n-1)p^2 (p+1-p)^{n-2} + np \\
&= n(n-1)p^2 + np.
\end{aligned}$$

Here the third-to-last equality of (3.6) follows from a transformation $i = x-2$. Combining (3.5) and (3.6) concludes that

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = n(n-1)p^2 + np - (np)^2 = np(1-p). \quad \square$$

Property 3.1.3. *Let $X \sim b(n,p)$. Then the moment generating function of X is given by*

$$M_X(t) = \left(1 - p + pe^t\right)^n, \quad t \in (-\infty, \infty). \tag{3.7}$$

Proof. Using the pmf in (3.3) and applying the definition of mgf will yield that

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= \left(1 - p + pe^t\right)^n \end{aligned} \tag{3.8}$$

for $t \in (-\infty, \infty)$. Here the last equality of (3.8) follows from an application of the binomial theorem. $\square$

Remark 3.1.1. *The proof of Property 3.1.2 for the expectation and variance of binomial distribution $b(n,p)$ is based on definitions of the expectation and variance. Both characteristics can also be obtained by using the mgf. Indeed, taking the first two derivatives of $M_X(t)$ with respect to t yields*

$$\begin{aligned} M_X'(t) &= np(1-p+pe^t)^{n-1}e^t, \\ M_X''(t) &= n(n-1)p^2(1-p+pe^t)^{n-2}e^{2t} + np(1-p+pe^t)^{n-1}e^t. \end{aligned} \tag{3.9}$$

Setting $t = 0$ in (3.9), we conclude that

$$\begin{aligned} E(X) &= M'(0) = np, \\ \mathrm{Var}(X) &= M''(0) - [M'(0)]^2 = n(n-1)p^2 + np - (np)^2 = np(1-p). \end{aligned}$$

Property 3.1.4. *(Additive property) Suppose that $X_1, X_2, \ldots, X_m$ are independent and $X_i \sim b(n_i, p)$ for $i = 1, 2, \ldots, m$. Then $X = \sum_{i=1}^m X_i \sim b(n, p)$, where $n = \sum_{i=1}^m n_i$.*

Proof. Let $n = \sum_{i=1}^{m} n_i$. Then we sequentially use the definition of the mgf, the property of the exponential function, the independence of $X_1, X_2, \ldots, X_n$, the mgf of the binomial distribution in Property 3.1.3 and the property of the exponential function to obtain the mgf of X below:

$$M_X(t) = E\left(e^{tX}\right) = E\left(\prod_{i=1}^{m} e^{tX_i}\right) = \prod_{i=1}^{m} E(e^{tX_i})$$
$$= \prod_{i=1}^{m} (1 - p + pe^t)^{n_i} = (1 - p + pe^t)^n,$$

which is the mgf of binomial distribution $b(n, p)$. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

Remark 3.1.2. *Since the binomial distribution describes the behavior of a count variable X from independent and repeated Bernoulli trials, it can be applied in many fields of science. When applying the binomial distribution to a specific problem in practice, however, we need to pay attention to four conditions involved in this distribution:*

(1) *each trial has two possible outcomes, success or failure;*

(2) *it consists of n identical trials;*

(3) *the probability p of success on any trial does not change from trial to trial, and*

(4) *all n trials are independent of each other.*

3.2 Poisson distribution

Definition 3.2.1. *The distribution of random variable X with the following pmf*

$$f_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, 2, \ldots \tag{3.10}$$

is called a Poisson distribution with parameter $\lambda > 0$. The random variable X is called a Poisson random variable. We write $X \sim \mathcal{P}(\lambda)$.

The cdf of $X \sim \mathcal{P}(\lambda)$ is given by

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \displaystyle\sum_{i=0}^{[x]} \frac{\lambda^i}{i!} e^{-\lambda} & \text{if } x \geq 0. \end{cases}$$

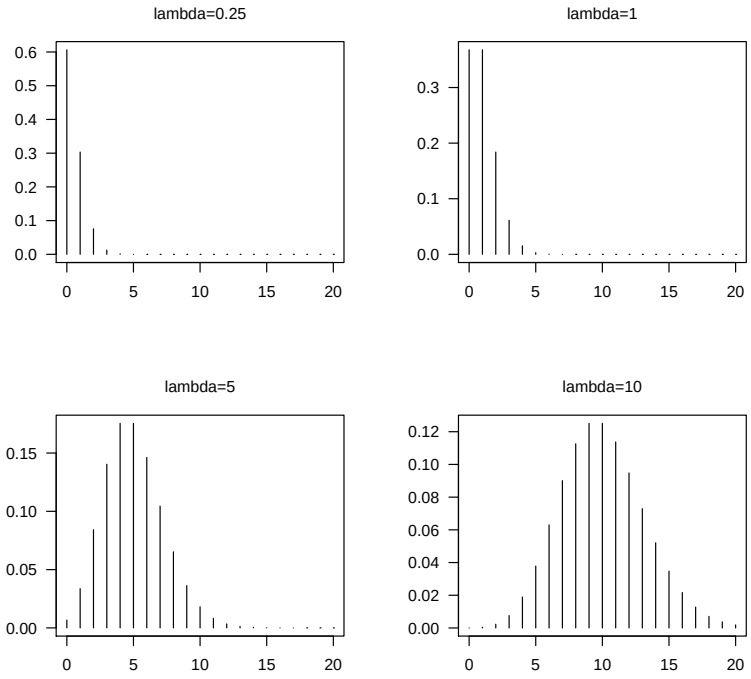


Figure 3.2.1 PMF of the Poisson distribution

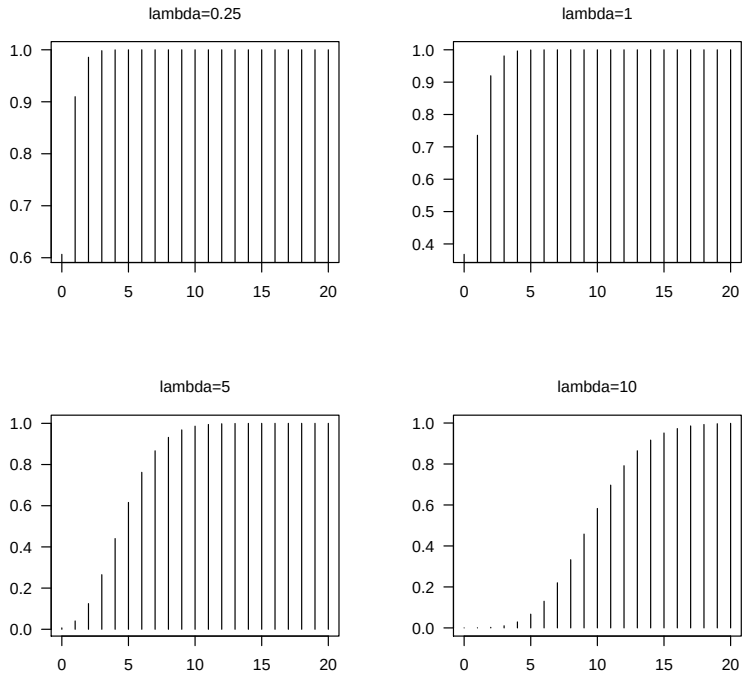


Figure 3.2.2 CDF of the Poisson distribution

Property 3.2.1. *Let $X \sim \mathcal{P}(\lambda)$. Then the expectation and variance of X are given by*

$$E(X) = \mathrm{Var}(X) = \lambda, \tag{3.11}$$

respectively.

Proof. Applying the definition of the expectation with the pmf (3.10), we have that

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\, f_X(x) = \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} \\ &= \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned} \tag{3.12}$$

Here the second-to-last equality of (3.12) follows from the Taylor's expansion of the exponential function $e^{\lambda} = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!}$.

To show the variance of Poisson distribution $\mathcal{P}(\lambda)$, we need $E(X^2)$, which can be similarly evaluated by

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{\infty} x^2 \, f_X(x) \\
&= \sum_{x=0}^{\infty} [x(x-1)+x] \frac{\lambda^x}{x!} e^{-\lambda} \\
&= \sum_{x=0}^{\infty} [x(x-1)] \frac{\lambda^x}{x!} e^{-\lambda} + \sum_{x=0}^{\infty} x \frac{\lambda^x}{x!} e^{-\lambda} \\
&= \lambda e^{-\lambda} \sum_{x=1}^{\infty} [x(x-1)] \frac{\lambda^{x-1}}{x!} + \lambda \qquad (3.13) \\
&= \sum_{x=2}^{\infty} [x(x-1)] \frac{\lambda^x}{x!} e^{-\lambda} + \lambda \\
&= \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} + \lambda \\
&= \lambda^2 e^{-\lambda} e^{\lambda} + \lambda \\
&= \lambda^2 + \lambda.
\end{aligned}$$

Combining (3.12) and (3.13) concludes that

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda. \qquad \square$$

Property 3.2.2. *Let $X \sim \mathcal{P}(\lambda)$. Then the moment generating function of X is given by*

$$M_X(t) = \exp\left(\lambda(e^t - 1)\right), \quad t \in (-\infty, \infty). \qquad (3.14)$$

Proof. We sequentially use the definition of the mgf, the pmf (3.10), and Taylor's series of the exponential function $e^z = \sum_{i=0}^{\infty} \frac{z^i}{i!}$ with $z = \lambda e^t$ to obtain that

$$\begin{aligned}
M_X(t) = E\left(e^{tX}\right) = \sum_x e^{tx} f_X(x) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\left(e^t \lambda\right)^x}{x!} \\
= e^{-\lambda} \exp\left(e^t \lambda\right) = \exp\left(\lambda(e^t - 1)\right)
\end{aligned}$$

for $t \in \mathbb{R}$. $\square$

Remark 3.2.1. *Taking the first two derivatives of $M_X(t)$ with respect to t, we have*

$$\begin{aligned} M'_X(t) &= \exp\left(\lambda(e^t-1)\right)\lambda e^t, \\ M''_X(t) &= \exp\left(\lambda(e^t-1)\right)\lambda e^t\left(1+\lambda e^t\right). \end{aligned} \tag{3.15}$$

Setting $t=0$ in (3.15), we conclude that

$$\begin{aligned} E(X) &= M'(0) = \lambda, \\ \mathrm{Var}(X) &= M''(0) - [M'(0)]^2 = \lambda(1+\lambda) - \lambda^2 = \lambda. \end{aligned}$$

Property 3.2.3. *(Additive property) Suppose that $X_1, X_2, \ldots, X_m$ are independent and $X_i \sim \mathcal{P}(\lambda_i)$ for $i = 1, 2, \ldots, m$. Then $X = \sum_{i=1}^m X_i \sim \mathcal{P}(\lambda)$, where $\lambda = \sum_{i=1}^m \lambda_i$.*

Proof. Let $\lambda = \sum_{i=1}^m \lambda_i$. Then we sequentially use the definition of the mgf, the property of the exponential function, the independence of $X_1, X_2, \ldots, X_n$, the mgf of the Poisson distribution in Property 3.2.2 and the property of the exponential function to obtain the mgf of X below:

$$\begin{aligned} M_X(t) = E\left(e^{tX}\right) &= E\left(\prod_{i=1}^m e^{tX_i}\right) = \prod_{i=1}^m E\left(e^{tX_i}\right) \\ &= \prod_{i=1}^m \exp\left(\lambda_i(e^t-1)\right) = \exp\left(\lambda(e^t-1)\right), \end{aligned}$$

which is the mgf of Poisson distribution $\mathcal{P}(\lambda)$. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

Remark 3.2.2. *The Poisson distribution has many applications in practice. It, for example, can be used to*

(1) *count the number of calls in a given period;*

(2) *count the number of bacteria in biology;*

(3) *determine the number of deaths by a rare disease in a district in a given period;*

(4) *count the number of car accidents on a highway in a given period.*

Example 3.2.1. *(Example 3.2.1 – Example 3.2.3)*

3.3 Geometric and negative binomial distributions

Definition 3.3.1. *Consider an experiment that consists of repeating Bernoulli trials independently until a success is obtained. Assume that the probability of success in each trial is p. Let X be the number of failures until the first success. Then the distribution of X is called a geometric distribution with parameter $p \in (0, 1]$. The random variable X is called a geometric random variable. We write $X \sim G(p)$.*

By Definition 3.3.1, we see that the pmf and cdf of $X \sim G(p)$ are given by

$$\begin{aligned} f_X(x) &= (1-p)^x p, \quad x = 0, 1, 2, \cdots, \\ F_X(x) &= \begin{cases} 0 & \text{if } x < 0, \\ 1 - (1-p)^{[x+1]} & \text{if } x \geq 0. \end{cases} \end{aligned} \tag{3.16}$$

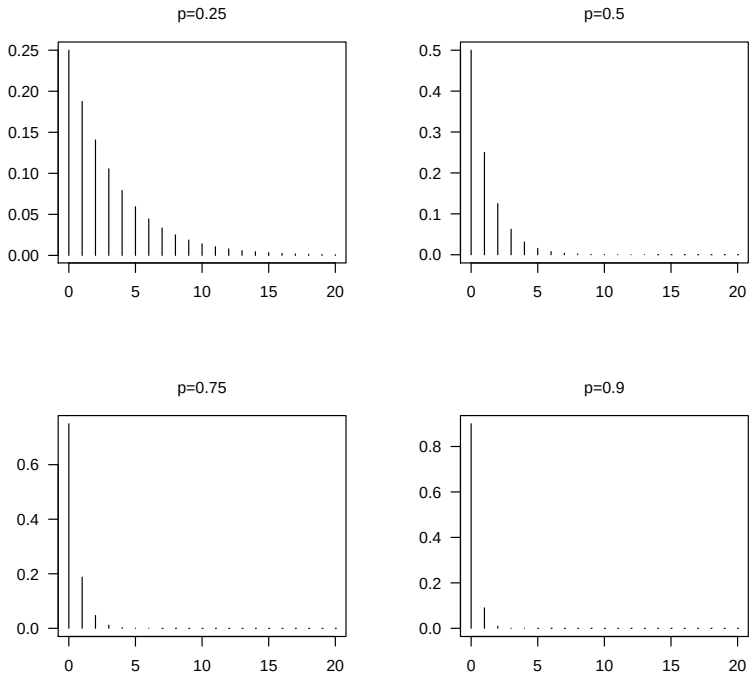


Figure 3.3.1 PMF of the geometric distribution

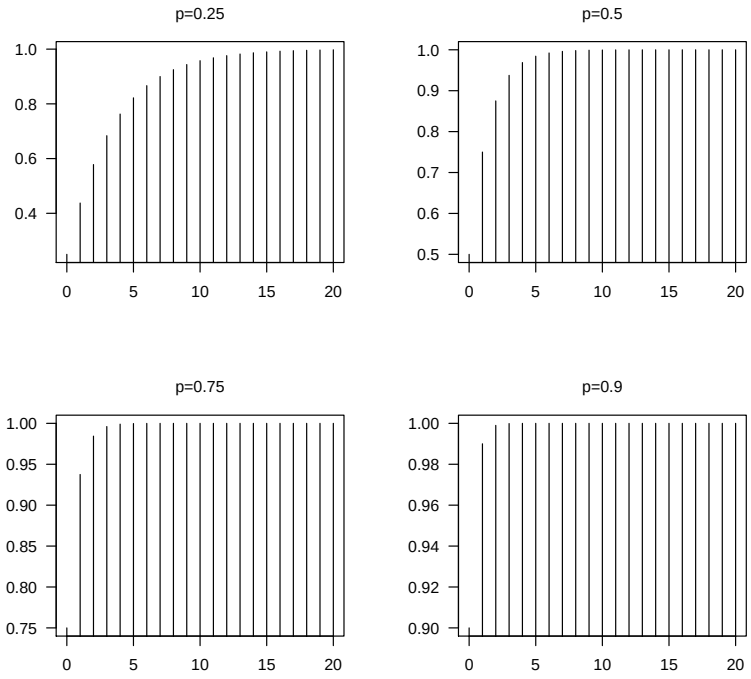


Figure 3.3.2 CDF of the geometric distribution

Remark 3.3.1. *Some textbooks define the distribution of $Y = X + 1$ as the geometric distribution. The random variable Y represents the number of trials required until the first success.*

Property 3.3.1. *Suppose that $X \sim G(p)$ with $p \in (0, 1)$. Then the pmf of X always decreases monotonically.*

Proof. For $x = 0, 1, \ldots$, we have from the pmf of X in (3.16) that

$$\frac{p_X(x+1)}{p_X(x)} = \frac{(1-p)^{x+1}p}{(1-p)^x p} = 1 - p < 1,$$

which means that $p_X(x+1) < p_X(x)$ for $x = 0, 1, 2, \ldots$. That is, $p_X(x)$ is a monotone decreasing function of x. □

Property 3.3.2. *Let $X \sim G(p)$. Then the expectation and variance of X are given by*

$$E(X) = \frac{1-p}{p} \quad \textit{and} \quad \mathrm{Var}(X) = \frac{1-p}{p^2}, \tag{3.17}$$

respectively.

Proof. Applying the definition of the expectation with the pmf given by (3.16), we have

$$\begin{aligned}
E(X) &= \sum_x x\, f_X(x) = \sum_{x=0}^{\infty} x(1-p)^x p = \sum_{x=1}^{\infty} x(1-p)^x p \\
&= \sum_{x=1}^{\infty} [(x-1)+1](1-p)^x p \\
&= \sum_{x=1}^{\infty} (x-1)(1-p)^x p + \sum_{x=1}^{\infty} (1-p)^x p \\
&= (1-p)\sum_{x=2}^{\infty} (x-1)(1-p)^{x-1} p + 1 - p \\
&= (1-p)\sum_{i=1}^{\infty} i(1-p)^i p + 1 - p \\
&= (1-p)\sum_{i=0}^{\infty} i(1-p)^i p + 1 - p \\
&= (1-p)E(X) + 1 - p,
\end{aligned} \tag{3.18}$$

which means that $\big(1-(1-p)\big)E(X) = 1-p$. This is equivalent to $E(X) = (1-p)/p$. Here the fourth-to-last equality in (3.18) follows from the geometric series.

To show the variance of geometric distribution $G(p)$, we need $E(X^2)$, which can be similarly evaluated by

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{\infty} x^2\, f_X(x) = \sum_{x=0}^{\infty} x^2(1-p)^x p = \sum_{x=1}^{\infty} [(x-1)+1]^2(1-p)^x p \\
&= \sum_{x=1}^{\infty} (x-1)^2(1-p)^x p + 2\sum_{x=1}^{\infty} (x-1)(1-p)^x p + \sum_{x=1}^{\infty} (1-p)^x p \\
&= (1-p)\sum_{i=0}^{\infty} i^2(1-p)^i p + 2(1-p)\sum_{i=0}^{\infty} i(1-p)^i p + 1 - p \\
&= (1-p)E(X^2) + 2(1-p)E(X) + 1 - p,
\end{aligned}$$

which means that $pE(X^2) = 2(1-p)E(X) + 1 - p$. Since $E(X) = (1-p)/p$, we have

$$E(X^2) = \frac{1}{p}\left(2(1-p)E(X) + 1 - p\right) = \frac{(2-p)(1-p)}{p^2}.$$

Thus,

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{(2-p)(1-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}. \quad \square$$

Property 3.3.3. *Let $X \sim G(p)$. Then the moment generating function of X is given by*

$$M_X(t) = \frac{p}{1-(1-p)e^t}, \quad t \in (-\infty, -\ln(1-p)). \tag{3.19}$$

Proof. Applying the definition of mgf with the pmf (3.16), we have

$$\begin{aligned} M_X(t) = E\left(e^{tX}\right) &= \sum_{x=0}^{\infty} e^{tx}(1-p)^x p = p\sum_{x=0}^{\infty}[(1-p)e^t]^x \\ &= \frac{p}{1-(1-p)e^t}, \end{aligned} \tag{3.20}$$

when $t \in (-\infty, -\log(1-p))$. Here the last equality of (3.20) is based on an application of the geometric series with a common ratio $(1-p)e^t$. $\square$

Remark 3.3.2. *The proof of Property 3.3.2 for the expectation and variance of geometric distribution $G(p)$ is based on definitions of the expectation and variance. Both characteristics can also be obtained by using the mgf. Indeed, taking the first two derivatives of $M_X(t)$ with respect to t yields*

$$\begin{aligned} M'_X(t) &= \frac{p(1-p)e^t}{\left(1-(1-p)e^t\right)^2}, \\ M''_X(t) &= \frac{p(1-p)e^t\left(1+(1-p)e^t\right)}{\left(1-(1-p)e^t\right)^3}. \end{aligned} \tag{3.21}$$

Setting $t = 0$ in (3.21) implies that

$$\begin{aligned} E(X) &= M'_X(0) = \frac{1-p}{p}, \\ \mathrm{Var}(X) &= M''_X(0) - [M'_X(0)]^2 = \frac{(2-p)(1-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}. \end{aligned}$$

Property 3.3.4. *(Memoryless property) Let $X \sim G(p)$. Then*

$$P(X \geq i+j | X \geq i) = P(X \geq j)$$

for any $i, j = 0, 1, 2, \ldots$.

Proof. We see from (3.16) that for any $x \geq 0$,

$$P(X > x) = (1-p)^{[x+1]}.$$

Thus, applying the definition of conditional probability yields that

$$\begin{aligned} P(X \geq i+j | X \geq i) &= \frac{P\big(X \geq i+j,\, X \geq i\big)}{P(X \geq i)} \\ &= \frac{P(X \geq i+j)}{P(X \geq i)} \\ &= \frac{(1-p)^{i+j}}{(1-p)^{i}} \\ &= (1-p)^{j} \\ &= P(X \geq j). \end{aligned}$$

□

Remark 3.3.3. *Memoryless property of the geometric distribution states that if there have been i failures initially, the probability of at least j more failures before the first success is the same as if we started the experiment for the first time and the information of initial i failures does not matter.*

Now we extend the geometric distribution to define a new distribution called the negative binomial distribution.

Definition 3.3.2. *Consider an experiment that consists of repeating independent Bernoulli trials until r successes are obtained. Assume that the probability of success in each trial is p. Let X be the number of failures until the rth success. Then the distribution of X is called a negative binomial distribution with a probability of success $p \in (0,1]$ and size r. The random variable X is called a negative binomial random variable. We write $X \sim nbinom(r,p)$.*

By Definition 3.3.2, the pmf and cdf of $X \sim nbinom(r,p)$ are given by

$$\begin{aligned} f_X(x) &= \binom{x+r-1}{r-1}(1-p)^x p^r, \quad x = 0,1,2,\cdots, \\ F_X(x) &= \begin{cases} 0, & \text{if } x < 0, \\ \displaystyle\sum_{i=0}^{[x]} \binom{i+r-1}{r-1}(1-p)^i p^r, & \text{if } x \geq 0. \end{cases} \end{aligned} \tag{3.22}$$

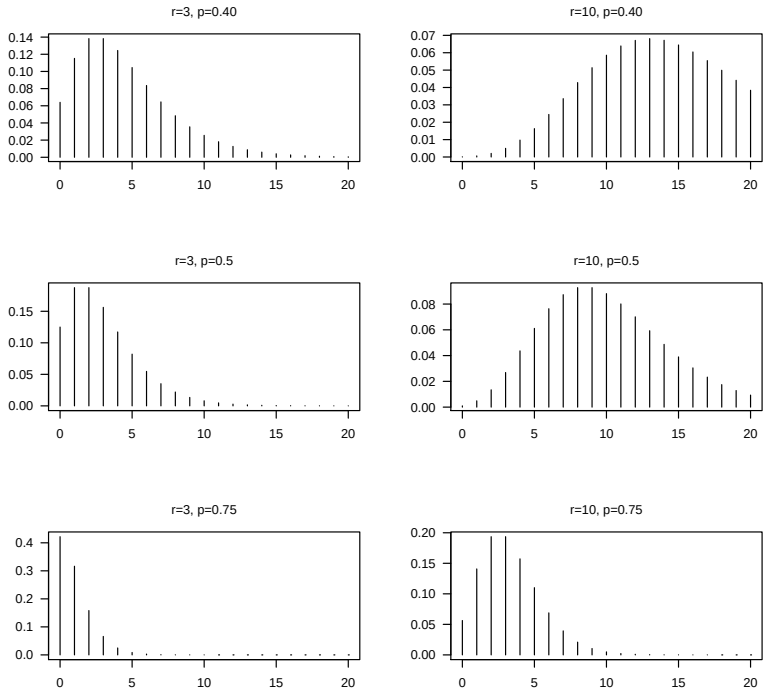


Figure 3.3.3 PMF of the negative binomial distribution

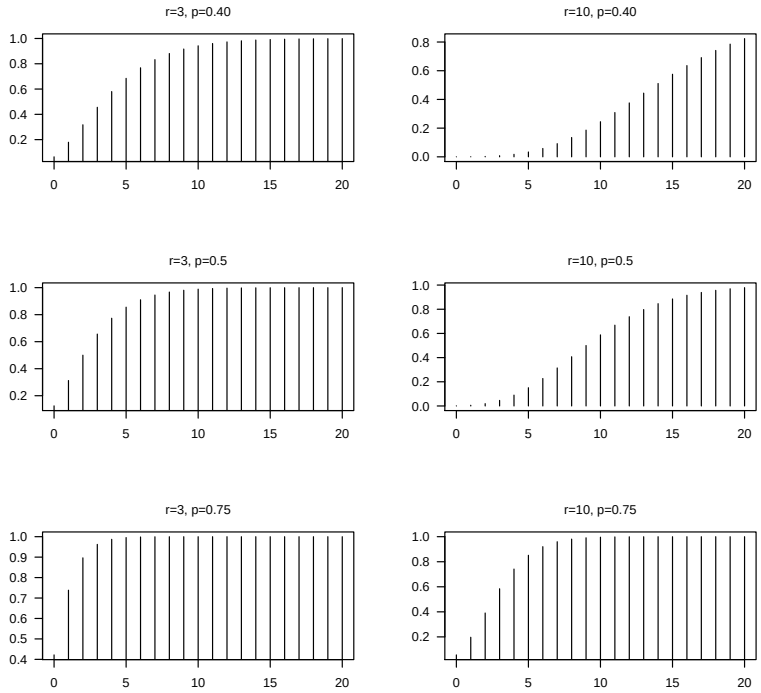


Figure 3.3.4 CDF of the negative binomial distribution

Property 3.3.5. *Let $X \sim nbinom(r, p)$. Then the expectation and variance of X are given by*

$$E(X) = r\frac{1-p}{p} \quad \textit{and} \quad \mathrm{Var}(X) = r\frac{1-p}{p^2}, \tag{3.23}$$

respectively.

Proof. Applying the definition of the expectation with the pmf given by (3.22),

we have

$$\begin{aligned}
E(X) &= \sum_x x\, f_X(x) \\
&= \sum_{x=0}^{\infty} x \binom{x+r-1}{r-1} (1-p)^x p^r \\
&= \sum_{x=1}^{\infty} x \binom{x+r-1}{r-1} (1-p)^x p^r \\
&= \sum_{x=1}^{\infty} \frac{(x+r-1)!}{(r-1)!(x-1)!} (1-p)^x p^r \\
&= \sum_{y=0}^{\infty} \frac{(y+r)!}{(r-1)!y!} (1-p)^{y+1} p^r \qquad (3.24) \\
&= \frac{r(1-p)}{p} \sum_{y=0}^{\infty} \frac{(y+r+1-1)!}{(r+1-1)!y!} (1-p)^y p^{r+1} \\
&= \frac{r(1-p)}{p} \sum_{y=0}^{\infty} \binom{y+r}{r} (1-p)^y p^{r+1} \\
&= \frac{r(1-p)}{p} \sum_{y=0}^{\infty} \binom{y+r}{r} (1-p)^y p^{r+1} \\
&= \frac{r(1-p)}{p}
\end{aligned}$$

because $\binom{y+r}{r}(1-p)^y p^{r+1}$ is the pmf of a negative binomial distribution $nbinom(r+1, p)$ and thus $\sum_{y=0}^{\infty} \binom{y+r}{r}(1-p)^y p^{r+1} = 1$.

To show the variance of negative binomial distribution $nbinom(r, p)$, we

need $E(X^2)$, which can be similarly evaluated by

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{\infty} x^2 \, f_X(x) \\
&= \sum_{x=0}^{\infty} x^2 \binom{x+r-1}{r-1} (1-p)^x p^r \\
&= \sum_{x=0}^{\infty} x^2 \frac{(x+r-1)!}{x!(r-1)!} (1-p)^x p^r \\
&= \sum_{x=1}^{\infty} x \frac{(x+r-1)!}{(x-1)!(r-1)!} (1-p)^x p^r \\
&= \sum_{x=1}^{\infty} [(x-1)+1] \frac{(x+r-1)!}{(x-1)!(r-1)!} (1-p)^x p^r \\
&= \sum_{x=1}^{\infty} (x-1) \frac{(x+r-1)!}{(x-1)!(r-1)!} (1-p)^x p^r \\
&\quad + \sum_{x=1}^{\infty} \frac{(x+r-1)!}{(x-1)!(r-1)!} (1-p)^x p^r \\
&= \sum_{x=2}^{\infty} \frac{(x+r-1)!}{(x-2)!(r-1)!} (1-p)^x p^r + E(X) \\
&= \frac{r(r+1)(1-p)^2}{p^2} \sum_{y=0}^{\infty} \frac{(y+r+1)!}{y!(r+1)!} (1-p)^y p^{r+2} + \frac{r(1-p)}{p} \\
&= \frac{r(r+1)(1-p)^2}{p^2} + \frac{r(1-p)}{p}.
\end{aligned} \tag{3.25}$$

Here we have used the fact that $\sum_{y=0}^{\infty} \frac{(y+r+1)!}{y!(r+1)!}(1-p)^y p^{r+2} = 1$ because $\frac{(y+r+1)!}{y!(r+1)!}(1-p)^y p^{r+2}$ for $y = 0, 1, 2, \ldots$ can be considered the pmf of negative binomial distribution $nbinom(r+2, p)$. Thus, combining (3.24) and (3.25) yields that

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - [E(X)]^2 \\
&= \frac{r(r+1)(1-p)^2}{p^2} + \frac{r(1-p)}{p} - \left(\frac{r(1-p)}{p}\right)^2 \\
&= \frac{r(1-p)}{p^2}.
\end{aligned}$$

□

Property 3.3.6. *Let $X \sim nbinom(r, p)$. Then the moment generating function of X is given by*

$$M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r, \quad t \in (-\infty, -\log(1-p)). \tag{3.26}$$

Proof. Applying the definition of mgf with the pmf (3.22), we have

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= \sum_{x=0}^{\infty} e^{tx}\binom{x+r-1}{r-1}(1-p)^x p^r \\ &= p^r \sum_{x=0}^{\infty}\binom{x+r-1}{r-1}[(1-p)e^t]^x \\ &= p^r\left(1-(1-p)e^t\right)^{-r} \\ &= \left(\frac{p}{1-(1-p)e^t}\right)^r, \end{aligned} \tag{3.27}$$

when $t \in (-\infty, -\log(1-p))$. Here the second-to-last equality of (3.27) is based on the fact that for $\theta \in (0, 1]$,

$$\sum_{x=0}^{\infty}\binom{x+r-1}{r-1}(1-\theta)^x = \theta^{-r},$$

which is from the pmf of the negative binomial distribution. □

Remark 3.3.4. *The proof of Property 3.3.5 for the expectation and variance of negative binomial distribution $nbinom(r, p)$ is based on definitions of the expectation and variance. Both characteristics can also be obtained by using the mgf. Indeed, taking the first two derivatives of $M_X(t)$ with respect to t yields*

$$\begin{aligned} M'_X(t) &= M_X(t)\frac{r(1-p)e^t}{1-(1-p)e^t}, \\ M''_X(t) &= M_X(t)\frac{r(1-p)e^t}{\left(1-(1-p)e^t\right)^2}\left(2-p+(r-1)(1-p)e^t\right). \end{aligned} \tag{3.28}$$

Setting $t = 0$ in (3.28) implies that

$$\begin{aligned} E(X) &= M'_X(0) = \frac{r(1-p)}{p}, \\ \mathrm{Var}(X) &= M''_X(0) - [M'_X(0)]^2 \\ &= \left(\frac{r(1-p)}{p}\right)^2 + \frac{r(1-p)}{p^2} - \left(\frac{r(1-p)}{p}\right)^2 = \frac{r(1-p)}{p^2}. \end{aligned}$$

Property 3.3.7. *Suppose that $X_1, X_2, \ldots, X_r$ are independent and $X_i \sim G(p)$ for $i = 1, 2, \ldots, r$. Then $X = \sum_{i=1}^{r} X_i \sim nbinom(r, p)$.*

Proof. We sequentially use the definition of the mgf, the property of the exponential function, the independence of $X_1, X_2, \ldots, X_r$, the mgf of the geometric distribution in Property 3.3.3 and the property of the exponential function to obtain the mgf of X below:

$$M_X(t) = E\left(e^{tX}\right) = E\left(\prod_{i=1}^{r} e^{tX_i}\right) = \prod_{i=1}^{r} E\left(e^{tX_i}\right)$$
$$= \prod_{i=1}^{r} \frac{p}{1-(1-p)e^t} = \left(\frac{p}{1-(1-p)e^t}\right)^r,$$

which is the mgf of negative binomial distribution $nbinom(r, p)$. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

Example 3.3.1. *(Example 3.1.6)*

3.4 Hypergeometric distribution

Definition 3.4.1. *Suppose that a population consists of $M+N$ objects, where M objects are characterized as successes and the other N objects are characterized as failures. A sample of k objects is randomly selected without replacement. Let X denote the number of successes in the sample. The distribution of X is called a hypergeometric distribution and X is called a hypergeometric random variable. We write $X \sim hyper(M, N, k)$, where M, N, and k are considered parameters.*

By Definition 3.4.1, the pmf and cdf of $X \sim hyper(M, N, k)$ are given by

$$f_X(x) = \frac{\binom{M}{x}\binom{N}{k-x}}{\binom{M+N}{k}}, \quad \max(0, k-N) \le x \le \min(k, M),$$
$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \sum_{i=0}^{[x]} \frac{\binom{M}{i}\binom{N}{k-i}}{\binom{M+N}{k}} & \text{if } x \ge 0. \end{cases} \tag{3.29}$$

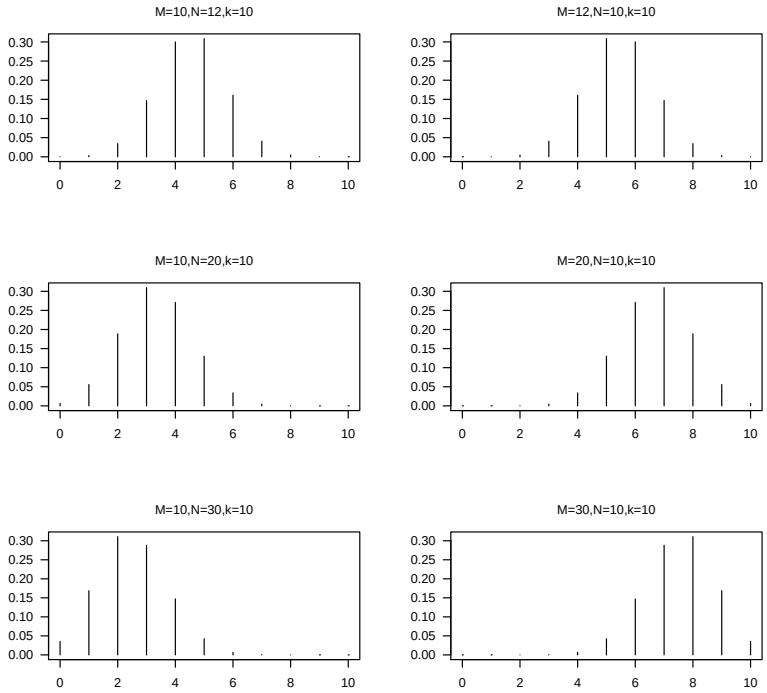


Figure 3.4.1 PMF of the hypergeometric distribution

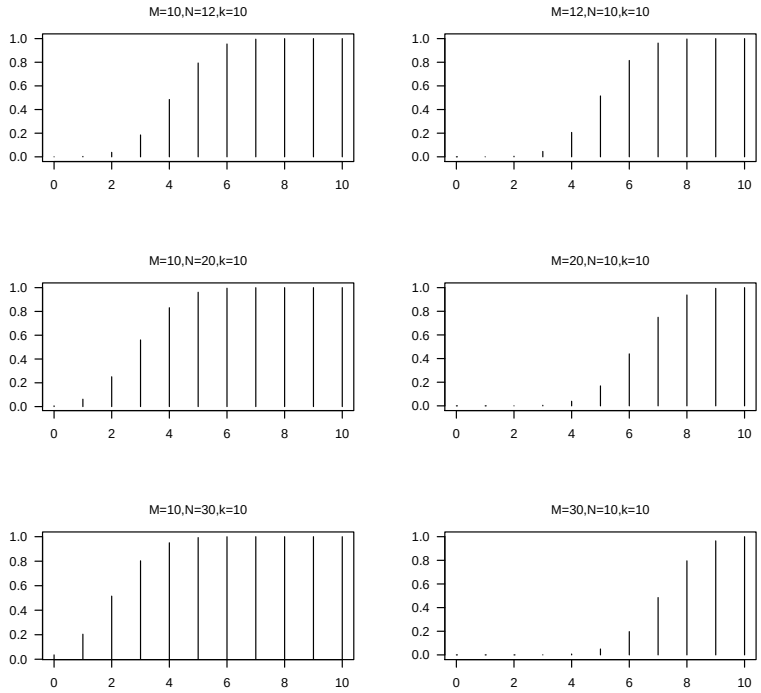


Figure 3.4.2 CDF of the hypergeometric distribution

Property 3.4.1. *Let $X \sim hyper(M, N, k)$. Then the expectation and variance of X are given by*

$$E(X) = k\frac{M}{M+N},$$
$$\mathrm{Var}(X) = k\frac{M+N-k}{M+N-1}\frac{M}{M+N}\frac{N}{M+N},$$

respectively.

3.5 Normal distribution

Definition 3.5.1. *If the pdf of random variable Z is given by*

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}, \tag{3.30}$$

then Z is called a standard normal random variable and its distribution is called a standard normal distribution. We write $Z \sim N(0,1)$.

Remark 3.5.1. *The cdf of $Z \sim N(0,1)$ is given by*

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)dt, \quad z \in \mathbb{R},$$

which does not have a closed form. To find probabilities from a standard normal distribution, one may either check its statistical table or use computer software such as R.

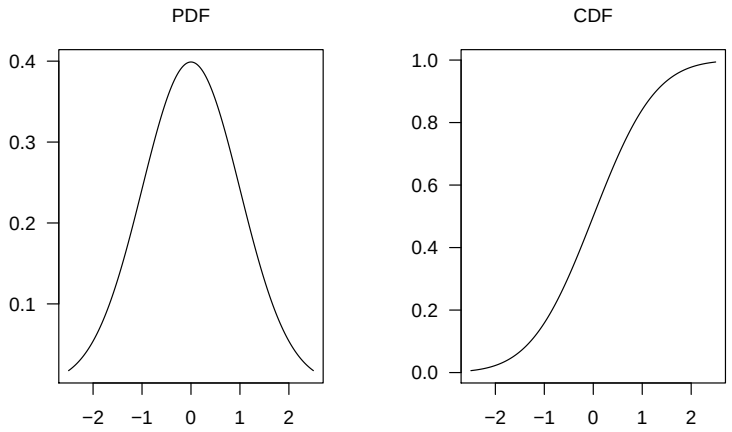


Figure 3.5.1 PDF and CDF of standard normal distribution

Property 3.5.1. *Suppose that $Z \sim N(0,1)$. Then the expectation, variance, and moment generating function of Z are given by*

$$\begin{aligned} E(Z) &= 0, \\ \text{Var}(Z) &= 1, \\ M_Z(t) &= e^{t^2/2}, \quad t \in (-\infty, \infty). \end{aligned} \tag{3.31}$$

Proof. By the pdf (3.30), applying the definition of the expectation will yield that

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z\phi(z)dz = \int_{-\infty}^{0} z\phi(z)dz + \int_{0}^{\infty} z\phi(z)dz \\ &= -\int_{0}^{\infty} z\phi(z)dz + \int_{0}^{\infty} z\phi(z)dz \\ &= 0 \end{aligned} \tag{3.32}$$

because

$$\begin{aligned} \int_{0}^{\infty} z\phi(z)dz &= \lim_{b\to\infty} \int_{0}^{b} z\phi(z)dz = \lim_{b\to\infty} \frac{-1}{\sqrt{2\pi}} e^{-z^2/2}\Big|_0^b = 1, \\ \int_{-\infty}^{\infty} |z|\phi(z)dz &= 2\int_{0}^{\infty} z\phi(z)dz = 2. \end{aligned}$$

To obtain the variance of normal distribution $N(0,1)$, we need $E(Z^2)$. Arguing along the same line as $E(Z)$ will have

$$\begin{aligned} E(Z^2) &= \int_{-\infty}^{\infty} z^2\phi(z)dz = 2\int_{0}^{\infty} z^2\phi(z)dz = \frac{2}{\sqrt{\pi}} \int_{0}^{\infty} t^{1/2} e^{-t} dt \\ &= \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{3}{2}\right) = \frac{2}{\sqrt{\pi}}\frac{1}{2}\Gamma\left(\frac{1}{2}\right) = 1 \end{aligned} \tag{3.33}$$

because $\Gamma(1/2) = \sqrt{\pi}$. Combining (3.32) and (3.33), we conclude that

$$\text{Var}(Z) = E(Z^2) - [E(Z)]^2 = 1.$$

To find the mgf of Z, we simplify $tz - \frac{z^2}{2}$ by completing the square

$$tz - \frac{z^2}{2} = \frac{t^2}{2} - \frac{(z-t)^2}{2}$$

and obtain

$$\begin{aligned} M_Z(t) = E\left(e^{tZ}\right) &= \int_{-\infty}^{\infty} e^{tz}\phi(z)dz = e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \phi(z-t)dz \\ &= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \phi(x)dx \quad (\text{by } x = z-t) \\ &= e^{\frac{1}{2}t^2} \end{aligned}$$

for $t \in \mathbb{R}$. □

Remark 3.5.2. *Taking the first two derivatives of $M_Z(t)$ with respect to t, we have*

$$M_Z'(t) = te^{\frac{1}{2}t^2} \quad \text{and} \quad M_Z''(t) = (1+t^2)e^{\frac{1}{2}t^2}. \tag{3.34}$$

Setting $t = 0$ in (3.34) yields that

$$\begin{aligned} E(Z) &= M'(0) = 0, \\ \mathrm{Var}(Z) &= M''(0) - [M'(0)]^2 = 1 - 0 = 1. \end{aligned}$$

Definition 3.5.2. *Let $X = \mu + \sigma Z$, where μ and $\sigma > 0$ are real numbers and $Z \sim N(0,1)$. Then the distribution of X is called a normal distribution with parameters μ and σ^2. We write $X \sim N(\mu, \sigma^2)$.*

According to Definition 3.5.2 and the pdf of Z, the pdf and cdf of X are given by

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}, \\ F_X(x) &= \Phi\left(\frac{x-\mu}{\sigma}\right), \quad x \in \mathbb{R}, \end{aligned} \tag{3.35}$$

where ϕ and Φ are pdf and cdf of $Z \sim N(0,1)$, respectively.

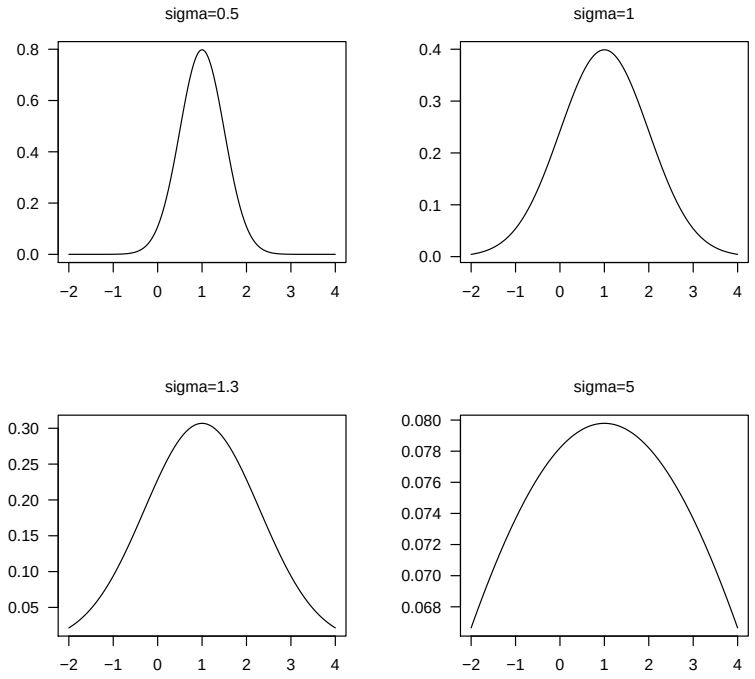


Figure 3.5.2 PDF of normal distribution (mu=1)

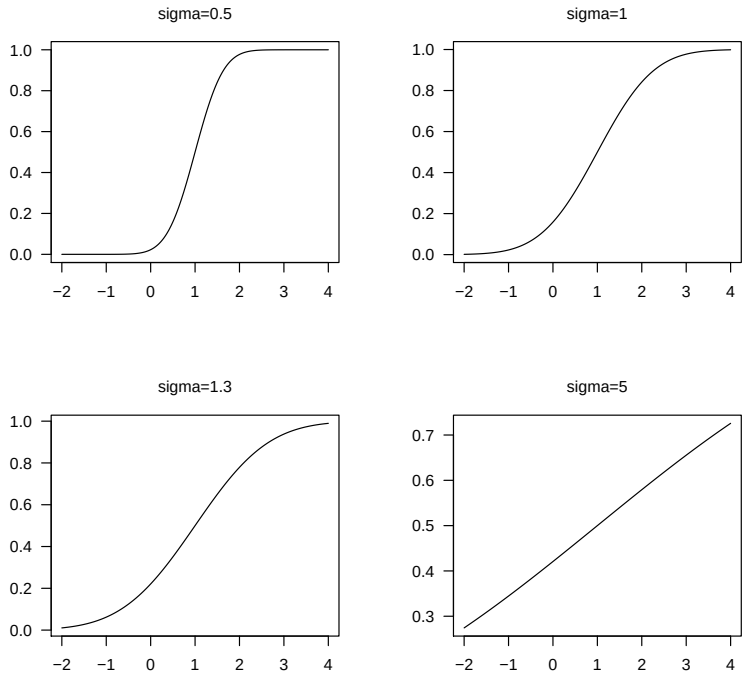


Figure 3.5.3 CDF of normal distribution (mu=1)

Property 3.5.2. *Let $X \sim N(\mu, \sigma^2)$. Then the expectation, variance, and moment generating function of X are given by*

$$\begin{aligned} E(X) &= \mu, \\ \mathrm{Var}(X) &= \sigma^2, \\ M_X(t) &= \exp\left(\mu t + \sigma^2 t^2/2\right), \quad t \in \mathbb{R}. \end{aligned}$$

Property 3.5.3. *Let $X \sim N(\mu, \sigma^2)$. Then*

$$\begin{aligned} P(-\sigma < X - \mu < \sigma) &\approx 0.6827, \\ P(-2\sigma < X - \mu < 2\sigma) &\approx 0.9545, \\ P(-3\sigma < X - \mu < 3\sigma) &\approx 0.9973. \end{aligned}$$

Property 3.5.4. *Suppose that $X_1, X_2, \ldots, X_n$ are independent and $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2, \ldots, n$. Let $X = \sum_{i=1}^n a_i X_i$, where $a_1, a_2, \ldots, a_n$ are real numbers. Then $X \sim N(\mu, \sigma^2)$, where $\mu = \sum_{i=1}^n a_i \mu_i$ and $\sigma^2 = \sum_{i=1}^n a_i^2 \sigma_i^2$.*

Proof. We sequentially use the definition of the mgf, the property of the exponential function, the independence of $X_1, X_2, \ldots, X_n$, the mgf of the normal distribution in Property 3.5.2 and the property of the exponential function to obtain the mgf of X below:

$$\begin{aligned} M_X(t) = E\left(e^{tX}\right) &= E\left(\prod_{i=1}^{n} e^{ta_iX_i}\right) = \prod_{i=1}^{n} E(e^{ta_iX_i}) \\ &= \prod_{i=1}^{n} \exp\left(a_i\mu_i t + \frac{a_i^2\sigma_i^2}{2}t^2\right) \\ &= \exp\left(t\sum_{i=1}^{n} a_i\mu_i + \frac{t^2}{2}\sum_{i=1}^{n} a_i^2\sigma_i^2\right) \\ &= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right), \end{aligned}$$

where $\mu = \sum_{i=1}^{n} a_i\mu_i$ and $\sigma^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2$. This is the mgf of normal distribution $N(\mu, \sigma^2)$. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

Corollary 3.5.1. *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with the normal distribution $N(\mu, \sigma^2)$. Then*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

3.6 χ^2-distribution

In statistics, we are often interested in knowing whether data follow a specific distribution. For example, how can we know whether the die is fair if it is rolled 100 times? The answer to this question is related to goodness-fit-test which uses an important distribution, called the χ^2-distribution.

Definition 3.6.1. *Suppose that $Z_1, Z_2, \ldots, Z_n$ are independent and $Z_i \sim N(0,1)$ for $i = 1, 2, \ldots, n$. Let $X = \sum_{i=1}^{n} Z_i^2$. Then the distribution of X is called a χ^2-distribution with n degrees of freedom and X is called a χ^2 random variable. We write $X \sim \chi_n^2$.*

Theorem 3.6.1. *The pdf of $X \sim \chi^2_n$ is given by*

$$f(x|n) = \begin{cases} \dfrac{x^{n/2-1}}{\Gamma(n/2)2^{n/2}} e^{-x/2}, & \text{if} \quad x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.36}$$

Proof. We prove the (3.36) by mathematical induction. Let $F(x|n)$ and $f(x|n)$ denote the cdf and pdf of $X \sim \chi^2_n$, respectively. Since $P(X > 0) = 1$, we only need to focus on $x > 0$.

(i) When $n = 1$, we have

$$F(x|1) = P(X \le x) = P(Z_1^2 \le x) = P\left(-\sqrt{x} \le Z_1 \le \sqrt{x}\right) = 2\Phi\left(\sqrt{x}\right) - 1,$$

which leads to

$$f(x|1) = \frac{dF(x|1)}{dx} = \frac{\phi(\sqrt{x})}{\sqrt{x}} = \frac{e^{-x/2}}{\sqrt{2\pi x}} = \frac{1}{\Gamma(1/2)2^{1/2}} x^{1/2-1} e^{-x/2}$$

because $\Gamma(1/2) = \sqrt{\pi}$. This shows that (3.36) is true when $n = 1$.

(ii) Assume that (3.36) is true when $n = m$. That is, we have

$$f(x|m) = \frac{1}{\Gamma(m/2)2^{m/2}} x^{m/2-1} e^{-x/2}, \quad x > 0.$$

(iii) When $n = m + 1$, using the convolution formula on page 108 and the

result in (i) will obtain that

$$\begin{aligned}
f(x|m+1) &= \int_0^x f(x-y|1)f(y|m)dy \\
&= \int_0^x \frac{e^{-(x-y)/2}}{\sqrt{2\pi(x-y)}} \frac{1}{\Gamma(m/2)2^{m/2}} y^{m/2-1}e^{-y/2}dy \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma(m/2)2^{m/2}} \int_0^x \frac{e^{-(x-y)/2}}{\sqrt{x-y}} y^{m/2-1}e^{-y/2}dy \\
&= \frac{e^{-x/2}}{\sqrt{2\pi}} \frac{1}{\Gamma(m/2)2^{m/2}} \int_0^x \frac{y^{m/2-1}}{\sqrt{x-y}}dy \\
&= \frac{e^{-x/2}}{\sqrt{2\pi}} \frac{1}{\Gamma(m/2)2^{m/2}} \int_0^x \frac{(x-z)^{m/2-1}}{\sqrt{z}}dy \\
&= \frac{e^{-x/2}}{\sqrt{2\pi}} \frac{x^{(m+1)/2-1}}{\Gamma(m/2)2^{m/2}} \int_0^x \left(\frac{z}{x}\right)^{1/2-1} \left(1-\frac{z}{x}\right)^{m/2-1} d\left(\frac{z}{x}\right) \\
&= \frac{e^{-x/2}}{\sqrt{2\pi}} \frac{x^{(m+1)/2-1}}{\Gamma(m/2)2^{m/2}} \int_0^1 u^{1/2-1}(1-u)^{m/2-1}du \\
&= \frac{e^{-x/2}}{\sqrt{2\pi}} \frac{x^{(m+1)/2-1}}{\Gamma(m/2)2^{m/2}} \frac{\Gamma(m/2)\Gamma(1/2)}{\Gamma((m+1)/2)} \\
&= \frac{1}{\Gamma((m+1)/2)2^{(m+1)/2}} x^{(m+1)/2-1}e^{-x/2}, \quad x>0.
\end{aligned}$$

Here the second-to-last equality follows from the Beta function in Calculus. This shows that (3.36) is true when $n = m+1$. Combining (i), (ii), and (iii) concludes that (3.36) is true for any $n \in \mathbb{N}$. □

Remark 3.6.1. *Theorem 3.6.1 can be proved by using the Gamma function.*

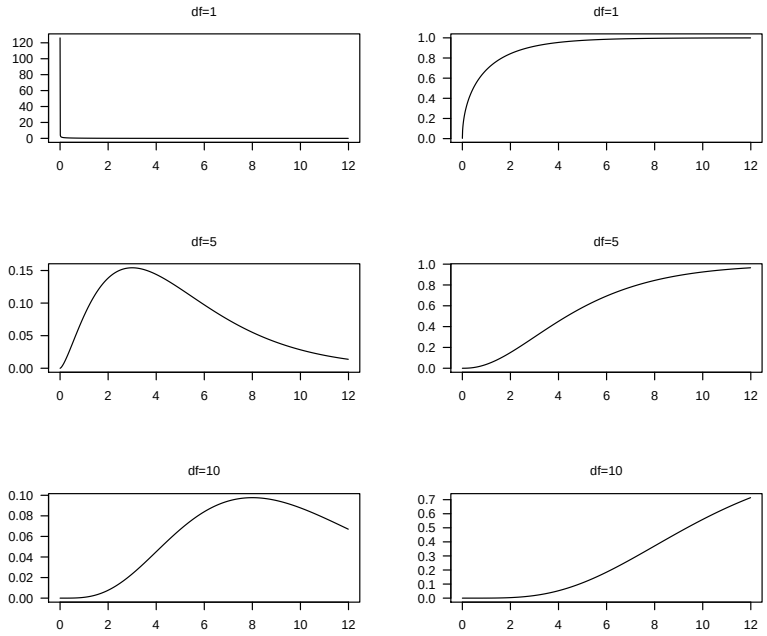


Figure 3.6.1 PDF and CDF of chi−square distribution

Property 3.6.1. *Suppose that* $X \sim \chi^2_n$. *Then*

$$E(X^r) = \frac{2^r \Gamma(n/2 + r)}{\Gamma(n/2)} \tag{3.37}$$

if $r > -n/2$.

Proof. By the pdf (3.36), applying the definition of the expectation will yield

that

$$\begin{aligned} E(X^r) &= \int_0^\infty x^r f(x|n)dx \\ &= \int_0^\infty x^r \frac{1}{\Gamma(n/2)2^{n/2}} x^{n/2-1}e^{-x/2}dx \\ &= \frac{1}{\Gamma(n/2)2^{n/2}} \int_0^\infty x^{(n+2r)/2-1}e^{-x/2}dx \qquad (3.38) \\ &= \frac{1}{\Gamma(n/2)2^{n/2}} \Gamma(n/2+r)2^{n/2+r} \\ &= \frac{2^r\Gamma(n/2+r)}{\Gamma(n/2)} \end{aligned}$$

provided that $r > -n/2$. □

Corollary 3.6.1. *Suppose that $X \sim \chi^2_n$. Then the expectation and variance of X are given by*

$$E(X) = n \quad \textit{and} \quad \mathrm{Var}(X) = 2n,$$

respectively.

Property 3.6.2. *Suppose that $X \sim \chi^2_n$. Then the moment generating function of X is given by*

$$M_X(t) = (1-2t)^{-n/2}, \quad t \in (-\infty, 1/2).$$

Proof. By the pdf (3.36), applying the definition of the mgf will yield that

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \int_0^\infty e^{tx} f_X(x)dx \\ &= \frac{(1-2t)^{-n/2}}{\Gamma(n/2)2^{n/2}} \int_0^\infty [(1-2t)x]^{n/2-1}e^{-x(1-2t)/2}d[(1-2t)x] \\ &= \frac{(1-2t)^{-n/2}}{\Gamma(n/2)2^{n/2}} \int_0^\infty y^{n/2-1}e^{-y/2}dy \qquad \big(\text{by } y = (1-2t)x\big) \\ &= (1-2t)^{-n/2} \end{aligned}$$

if $t < 1/2$. □

Remark 3.6.2. *Taking the first two derivatives of $M_X(t)$ with respect to t yields*

$$\begin{aligned} M'_X(t) &= n(1-2t)^{-n/2-1}, \\ M''_X(t) &= n(n+2)(1-2t)^{-n/2-2}. \end{aligned} \tag{3.39}$$

Setting $t = 0$ in (3.39) yields that

$$\begin{aligned} E(X) &= M'(0) = n, \\ \mathrm{Var}(X) &= M''(0) - [M'(0)]^2 = n(n+2) - n^2 = 2n. \end{aligned}$$

Property 3.6.3. *(Additive property) Suppose that $X_1, X_2, \ldots, X_m$ are independent and $X_i \sim \chi^2_{n_i}$ for $i = 1, \ldots, m$. Then $X = \sum_{i=1}^m X_i \sim \chi^2_n$, where $n = \sum_{i=1}^m n_i$.*

Proof. We sequentially use the definition of the mgf, the property of the exponential function, the independence of $X_1, X_2, \ldots, X_m$, the mgf of the χ^2-distribution in Property 3.6.2 and the property of the exponential function to obtain the mgf of X below:

$$\begin{aligned} M_X(t) = E\big(e^{tX}\big) &= E\left(\prod_{i=1}^m e^{tX_i}\right) = \prod_{i=1}^m E\big(e^{tX_i}\big) \\ &= \prod_{i=1}^m (1-2t)^{-n_i/2} = (1-2t)^{-n/2}, \end{aligned}$$

where $n = \sum_{i=1}^m n_i$. This is the mgf of χ^2_n-distribution. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

3.7 Student's t-distribution

Definition 3.7.1. *Suppose that the random variable $Z \sim N(0,1)$ and the random variable $V \sim \chi^2_n$ are independent. Let $T = Z/\sqrt{V/n}$. Then the distribution of T is called a student's t-distribution with n degrees of freedom and T is called a student's t random variable. We write $T \sim t_n$.*

Theorem 3.7.1. *The pdf of $T \sim t_n$ is given by*

$$f(t|n) = \frac{\Gamma\big((n+1)/2\big)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}, \quad t \in \mathbb{R}. \tag{3.40}$$

Proof. Let $F(t|n)$ and $f(t|n)$ denote the cdf and pdf of $T \sim t_n$, respectively. Then for $t \in \mathbb{R}$,

$$F(t|n) = P(T \le t) = P(Z \le t\sqrt{V/n}) = \int_0^\infty \Phi(t\sqrt{v/n}) \frac{v^{n/2-1}e^{-v/2}}{\Gamma(n/2)2^{n/2}} dv,$$

which leads to

$$\begin{aligned}
f(t|n) = \frac{dF(t|n)}{dt} &= \int_0^\infty \phi(t\sqrt{v/n})\sqrt{v/n}\, \frac{v^{n/2-1}e^{-v/2}}{\Gamma(n/2)2^{n/2}} dv \\
&= \frac{1}{\Gamma(n/2)2^{n/2}} \frac{1}{\sqrt{2\pi n}} \int_0^\infty v^{(n+1)/2-1} \exp\left\{-\frac{v}{2}\left(1+\frac{t^2}{n}\right)\right\} dv \\
&= \frac{1}{\Gamma(n/2)2^{n/2}} \frac{1}{\sqrt{2\pi n}} \left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \int_0^\infty u^{(n+1)/2-1} e^{-u/2} du \\
&= \frac{\Gamma((n+1)/2)2^{(n+1)/2}}{\Gamma(n/2)2^{n/2}} \frac{1}{\sqrt{2\pi n}} \left(1+\frac{t^2}{n}\right)^{-(n+1)/2} \\
&= \frac{\Gamma\big((n+1)/2\big)}{\sqrt{n\pi}\Gamma(n/2)} \left(1+\frac{t^2}{n}\right)^{-(n+1)/2}.
\end{aligned}$$

□

Remark 3.7.1. *The t-distribution is often called Student's t-distribution because it was discovered by William S. Gosset in 1908 when he worked as a statistician for the Guinness brewing company which had stipulated him not publish under his name. He, therefore, wrote under the pen name* **Student**.

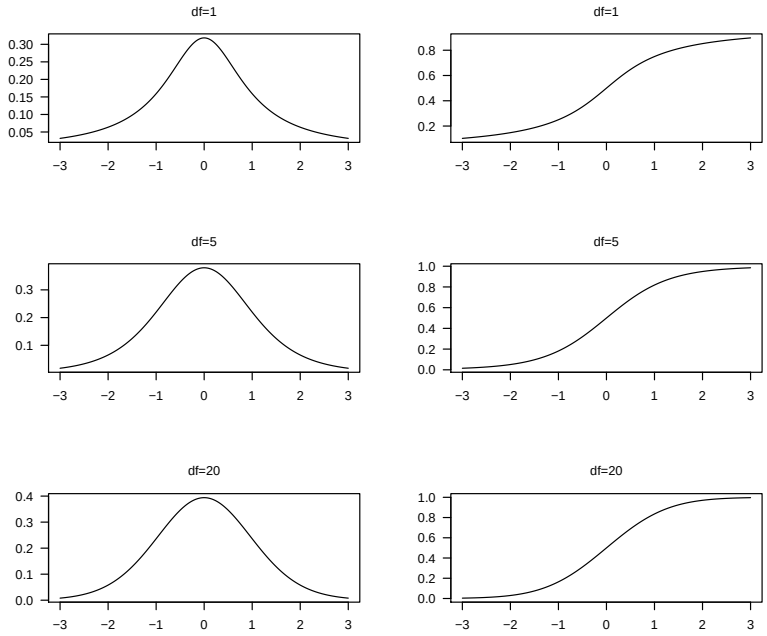


Figure 3.7.1 PDF and CDF of t−distribution

Property 3.7.1. *Let $T \sim t_n$. Then the expected value and variance of T are given by*

$$E(T) = 0 \quad (n > 1) \quad \textit{and} \quad \mathrm{Var}(T) = \frac{n}{n-2} \quad (n > 2),$$

respectively.

Remark 3.7.2. *As $n \to \infty$, t_n-distribution approaches to the standard normal distribution $N(0, 1)$.*

3.8 F-distribution

Definition 3.8.1. *Suppose that $X_1 \sim \chi^2_{n_1}$ and $X_2 \sim \chi^2_{n_2}$ are independent. Let $X = \dfrac{X_1/n_1}{X_2/n_2}$. Then the distribution of X is called an F-distribution with (n_1, n_2) degrees of freedom. We write $X \sim F_{n_1,n_2}$.*

Theorem 3.8.1. *The pdf of $X \sim F_{n_1,n_2}$ is given by*

$$f(x) = = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, \quad x>0.$$

Proof. Let $f(x|n)$ and $F(x|n)$ denote the pdf and cdf of χ^2_n-distribution respectively and let

$$C_n = \frac{1}{\Gamma(n/2)2^{n/2}}.$$

Then the cdf of $X \sim F_{n_1,n_2}$ can be expressed by

$$P(X \le x) = P\left(X_1 \le \frac{n_1}{n_2} x X_2\right).$$

Taking the derivative with respect to x yields that

$$\begin{aligned}
f(x) &= \int_0^\infty f\left(\frac{n_1}{n_2}xy\Big|n_1\right)\frac{n_1}{n_2}y f(y|n_2)dy \\
&= C_{n_1}C_{n_2}\int_0^\infty \left(\frac{n_1}{n_2}xy\right)^{n_1/2-1}\exp\left(-\frac{n_1}{2n_2}xy\right)\frac{n_1}{n_2}y\, y^{n_2/2-1}e^{-y/2}dy \\
&= C_{n_1}C_{n_2}\left(\frac{n_1}{n_2}x\right)^{\frac{n_1}{2}-1}\frac{n_1}{n_2}\int_0^\infty y^{\frac{n_1+n_2}{2}-1}\exp\left\{-\frac{y}{2}\left(\frac{n_1}{n_2}x+1\right)\right\}dy \\
&= C_{n_1}C_{n_2}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\Gamma\left(\frac{n_1+n_2}{2}\right)2^{\frac{n_1+n_2}{2}}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} \\
&= \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}
\end{aligned}$$

for $x > 0$. □

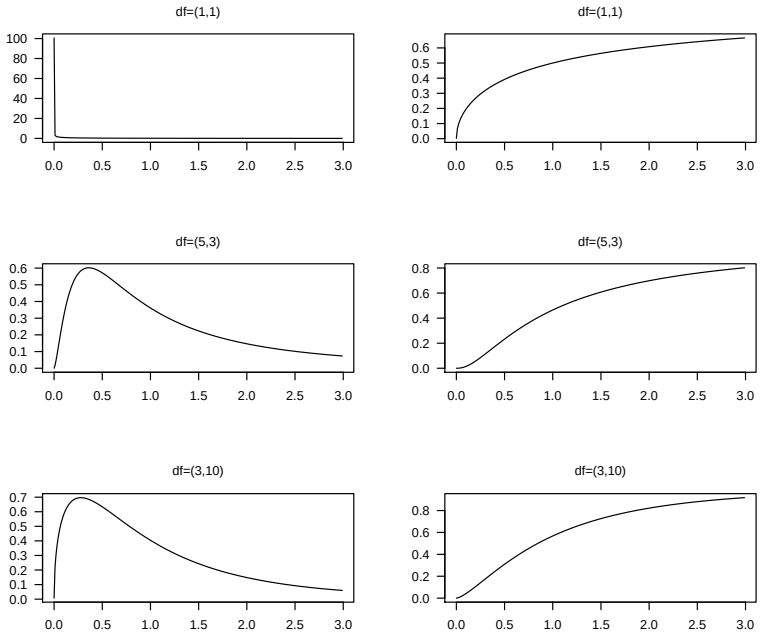


Figure 3.8.1 PDF and CDF of F−distribution

Property 3.8.1. *Suppose that* $X \sim F_{n_1,n_2}$. *Then*

$$E(X^k) = \left(\frac{n_2}{n_1}\right)^k \frac{\Gamma((2k+n_1)/2)\Gamma((n_2-2k)/2)}{\Gamma(n_1/2)\Gamma(n_2/2)}$$

provided $n_2 > 2k$. *In particular,*

$$E(X) = \frac{n_2}{n_2-2}, \quad (n_2 > 2),$$
$$\mathrm{Var}(X) = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}, \quad (n_2 > 4).$$

Remark 3.8.1. *If* $X \sim F_{n_1,n_2}$, *then* $1/X \sim F_{n_2,n_1}$.

3.9 Skewed normal distribution

Definition 3.9.1. *If the pdf of X has the form*

$$f_X(x) = 2\phi(x)\Phi(\lambda x), \quad x \in \mathbb{R}, \tag{3.41}$$

where $\lambda \in \mathbb{R}$ is a parameter, ϕ and Φ are the pdf and cdf of $N(0,1)$, respectively, the distribution of X is called a skew-normal distribution and X is called a skew-normal random variable. We write $X \sim SN(\lambda)$.

Remark 3.9.1. *We now show that $f_X(x)$ given by (3.41) is a pdf on $\mathbb{R}$. In fact, $f_X(x)$ is clearly nonnegative. Furthermore, let $I = \int_{-\infty}^{\infty} f_X(x)dx$. Then*

$$\begin{aligned}
I &= \int_{-\infty}^{\infty} f_X(x)dx \\
&= \int_{-\infty}^{\infty} 2\phi(x)\Phi(\lambda x)dx \\
&= 2\int_{-\infty}^{\infty} \phi(-y)\Phi(-\lambda y)dy \\
&= 2\int_{-\infty}^{\infty} \phi(y)[1-\Phi(\lambda y)]dy \\
&= 2\int_{-\infty}^{\infty} \phi(y)dy - 2\int_{-\infty}^{\infty} \phi(y)\Phi(\lambda y)dy \\
&= 2 - \int_{-\infty}^{\infty} 2\phi(y)\Phi(\lambda y)dy \\
&= 2 - \int_{-\infty}^{\infty} f_X(y)dy \\
&= 2 - I,
\end{aligned}$$

which is equivalent to $I = 1$.

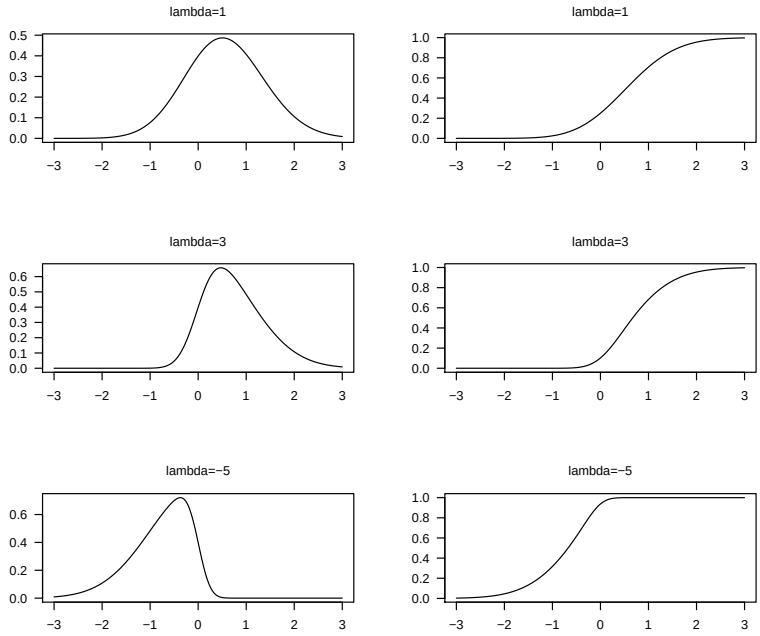


Figure 3.9.1 PDF and CDF of skew normal distribution

Property 3.9.1. *Let $X \sim SN(\lambda)$. Then $X^2 \sim \chi^2_1$.*

Property 3.9.2. *Let Y and Z be i.i.d. with $N(0,1)$ and let $X = [1 - I(Y \leq \lambda Z)]Z$. Then $X \sim SN(\lambda)$.*

Proof.

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P\Big([1 - 2I(Y > \lambda Z)]Z \leq x\Big) \\ &= P(Y > \lambda Z, -Z \leq x) + P(Y \leq \lambda Z, Z \leq x) \\ &= \int_{-x}^{\infty} [1 - \Phi(\lambda z)]\phi(z)dz + \int_{-\infty}^{x} \Phi(\lambda z)\phi(z)dz, \end{aligned}$$

which leads to

$$f(x) = \frac{dF(x)}{dx} = [1 - \Phi(-\lambda x)]\phi(-x) + \Phi(\lambda x)\phi(x) = 2\phi(x)\Phi(\lambda x)$$

for $x \in \mathbb{R}$. □

3.10 Exponential and gamma distributions

Definition 3.10.1. *Let X be a nonnegative random variable. If the pdf of X is given by*

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \tag{3.42}$$

the distribution of X is called an exponential distribution with rate parameter $\lambda > 0$. The random variable X is called an exponential random variable. We write $X \sim exp(\lambda)$.

By Definition 3.10.1, the cdf of $X \sim exp(\lambda)$ is given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \tag{3.43}$$

Remark 3.10.1. *When $\lambda = 1/2$, the exponential distribution $exp(1/2)$ is χ^2_2-distribution.*

Remark 3.10.2. *Some textbooks define the exponential distribution by using the scale parameter, which is equal to the reciprocal of the rate parameter.*

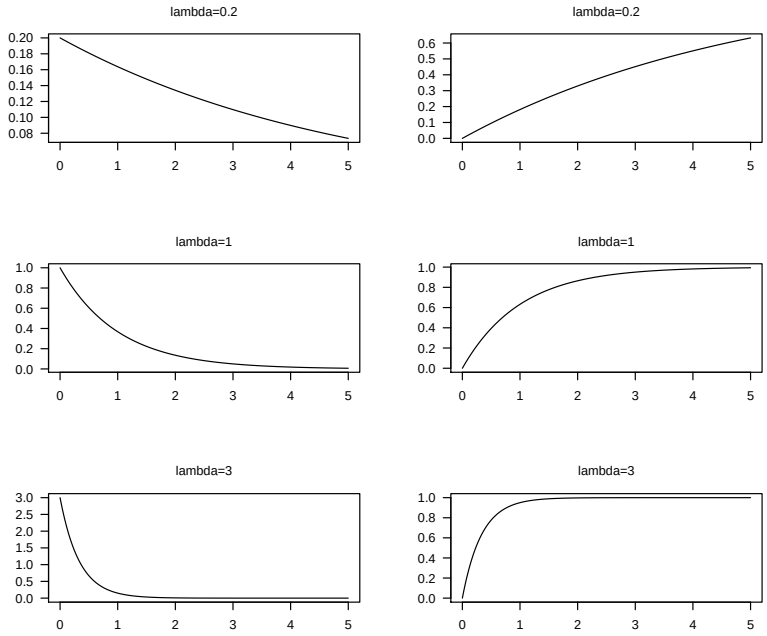


Figure 3.10.1 PDF and CDF of exponential distribution

Property 3.10.1. *Let $X \sim exp(\lambda)$. Then the expectation and the variance of X are respectively given by*

$$E(X) = 1/\lambda \quad \textit{and} \quad \mathrm{Var}(X) = 1/\lambda^2.$$

Proof. Applying the definition of the expectation with the pdf given by (3.42), we have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^{\infty} x\lambda e^{-\lambda x}dx = \frac{1}{\lambda}\int_0^{\infty} (\lambda x)e^{-\lambda x}d(\lambda x) \\ &= \frac{1}{\lambda}\int_0^{\infty} ye^{-y}dy = \frac{1}{\lambda}\Gamma(2) = \frac{1}{\lambda}. \end{aligned}$$

To show the variance of exponential distribution $exp(\lambda)$, we need $E(X^2)$. Arguing along the same line as $E(X)$, we have

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^{\infty} x^2\lambda e^{-\lambda x}dx = \frac{1}{\lambda^2}\int_0^{\infty} (\lambda x)^2 e^{-\lambda x}d(\lambda x) \\ &= \frac{1}{\lambda}\int_0^{\infty} y^2 e^{-y}dy = \frac{1}{\lambda^2}\Gamma(3) = \frac{2}{\lambda^2}. \end{aligned}$$

Thus,

$$\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \qquad \square$$

Property 3.10.2. *Let $X \sim exp(\lambda)$. Then the moment generating function of X is given by*

$$M_X(t) = \frac{\lambda}{\lambda - t}, \quad t \in (-\infty, \lambda). \tag{3.44}$$

Proof. By Definition 3.10.1, applying the definition of mgf will yield that

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\ &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx \\ &= \frac{\lambda}{\lambda - t} \end{aligned} \tag{3.45}$$

when $t \in (-\infty, \lambda)$. □

Remark 3.10.3. *The proof of Property 3.10.1 for the expectation and variance of exponential distribution $exp(\lambda)$ is based on definitions of the expectation and variance. Both characteristics can also be obtained by using the mgf. Indeed, taking the first two derivatives of $M_X(t)$ with respect to t yields*

$$\begin{aligned} M'_X(t) &= \frac{\lambda}{(\lambda - t)^2}, \\ M''_X(t) &= \frac{2\lambda}{(\lambda - t)^3}. \end{aligned} \tag{3.46}$$

Setting $t = 0$ in (3.46) implies that

$$\begin{aligned} E(X) &= M'_X(0) = 1/\lambda, \\ \mathrm{Var}(X) &= M''_X(0) - [M'_X(0)]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}. \end{aligned}$$

Theorem 3.10.1. *Suppose that $X_1, X_2, \ldots, X_m$ are independent and $X_i \sim exp(\lambda_i)$ for $i = 1, 2, \ldots, m$. Then $X = \min(X_1, X_2, \ldots, X_m) \sim exp(\lambda)$, where $\lambda = \sum_{i=1}^{m} \lambda_i$.*

Proof. Using the definition of cdf of the exponential distribution, we have

$$\begin{aligned} F_X(x) &= P(X \le x) \\ &= P\big(\min(X_1, X_2, \ldots, X_m) \le x\big) \\ &= 1 - P\big(\min(X_1, X_2, \ldots, X_m) > x\big) \\ &= 1 - \prod_{i=1}^{m} P(X_i > x) \\ &= 1 - \prod_{i=1}^{m} e^{-\lambda_i x} \\ &= 1 - e^{-\lambda x} \end{aligned}$$

for $x \ge 0$, where $\lambda = \sum_{i=1}^{m} \lambda_i$. Clearly, $F_X(x) = 0$ when $x < 0$. □

Definition 3.10.2. *If the pdf of random variable X is defined by*

$$f_X(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} & \text{if } x > 0, \\ 0 & \text{if } x \le 0, \end{cases} \tag{3.47}$$

the distribution of X is called a gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. The X is called a gamma random variable. We write $X \sim \Gamma(\alpha, \beta)$.

Remark 3.10.4. *When $\alpha = 1$, the gamma distribution becomes an exponential distribution, i.e., $\Gamma(1, \beta) = exp(1/\beta)$.*

Remark 3.10.5. *When $\alpha = n/2$ and $\beta = 2$, the gamma distribution reduces to χ^2_n-distribution, i.e., $\Gamma(n/2, 2) = \chi^2_n$.*

Remark 3.10.6. *Some textbooks define the gamma distribution by using the rate parameter, which is equal to the reciprocal of the scale parameter.*

Remark 3.10.7. *By Definition3.10.2, we see that the cdf of $X \sim \Gamma(\alpha, \beta)$ does not have a closed form in general.*

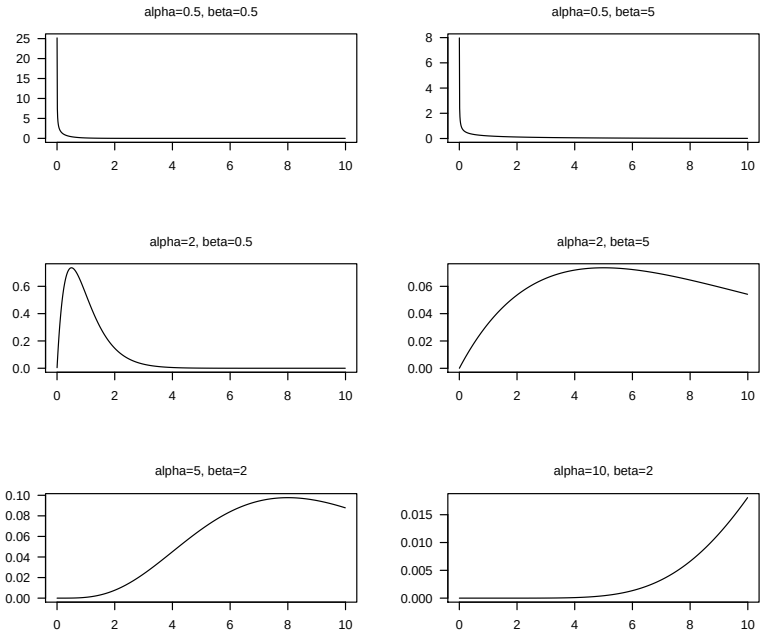


Figure 3.10.2 PDF of the gamma distribution

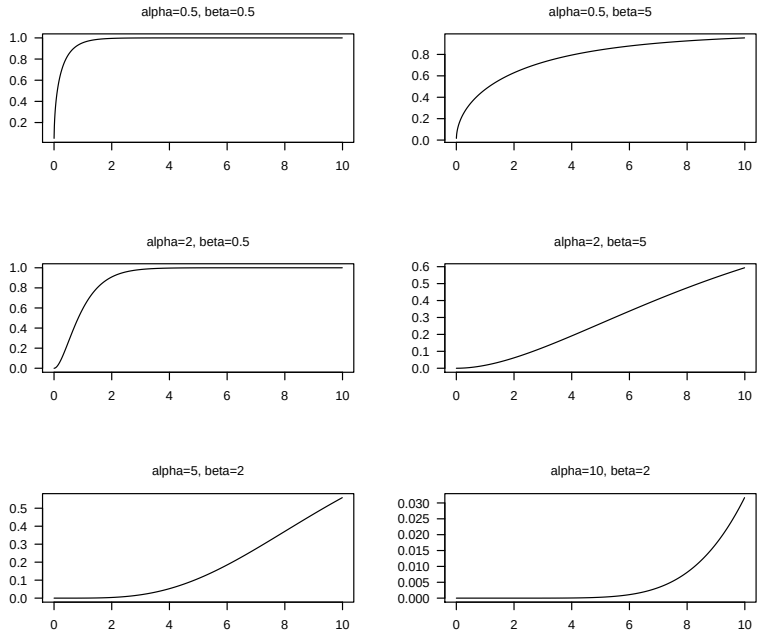


Figure 3.10.3 CDF of the gamma distribution

Property 3.10.3. *Let $X \sim \Gamma(\alpha, \beta)$. Then the expectation and variance of X are respectively given by*

$$E(X) = \alpha\beta \quad \textit{and} \quad \mathrm{Var}(X) = \alpha\beta^2. \tag{3.48}$$

Proof. Applying the definition of the expectation with the pdf given by (3.47),

we have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^{\infty} x \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x^{\alpha} e^{-x/\beta} dx \\ &= \frac{\beta^{\alpha+1}}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} y^{\alpha} e^{-y} dy \quad (\text{by} \quad y = x/\beta) \\ &= \beta \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\ &= \alpha\beta. \end{aligned} \tag{3.49}$$

To show the variance of gamma distribution $\Gamma(\alpha, \beta)$, we need to evaluate $E(X^2)$. Arguing along the same line as $E(X)$ will have

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &= \int_0^{\infty} x^2 \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x^{\alpha+1} e^{-x/\beta} dx \\ &= \frac{\beta^{\alpha+2}}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} y^{\alpha+1} e^{-y} dy \\ &= \beta^2 \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)} \\ &= \alpha(\alpha+1)\beta^2. \end{aligned} \tag{3.50}$$

Thus, combining (3.49) and (3.50) yields that

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2. \qquad \square$$

Property 3.10.4. *Let $X \sim \Gamma(\alpha, \beta)$. Then the moment generating function of X is given by*

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}, \quad t \in (-\infty, 1/\beta). \tag{3.51}$$

Proof. By Definition 3.10.2, applying the definition of mgf will yield that

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f_X(x)dx \\ &= \int_0^{\infty} e^{tx} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} x^{\alpha-1} e^{-(1/\beta - t)x} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \frac{\Gamma(\alpha)}{(1/\beta - t)^{\alpha}} \\ &= (1-\beta t)^{-\alpha} \end{aligned} \tag{3.52}$$

when $t \in (-\infty, 1/\beta)$. Here the second to last equality of (3.52) is based on an application of the gamma function. □

Remark 3.10.8. *The proof of Property 3.10.3 for the expectation and variance of gamma distribution $\Gamma(\alpha, \beta)$ is based on definitions of the expectation and variance. Both characteristics can also be obtained by using the mgf. Indeed, taking the first two derivatives of $M_X(t)$ with respect to t yields*

$$\begin{aligned} M'_X(t) &= \alpha\beta(1-\beta t)^{-\alpha-1}, \\ M''_X(t) &= \alpha(\alpha+1)\beta^2(1-\beta t)^{-\alpha-2}. \end{aligned} \tag{3.53}$$

Setting $t = 0$ in (3.53) implies that

$$\begin{aligned} E(X) &= M'_X(0) = \alpha\beta, \\ \mathrm{Var}(X) &= M''_X(0) - [M'_X(0)]^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2. \end{aligned}$$

Furthermore, we have

$$M_X^{(k)}(t) = \alpha(\alpha+1)\cdots(\alpha+k-1)\beta^k(1-\beta t)^{-\alpha-k},$$

which implies that

$$E(X^k) = \alpha(\alpha+1)\cdots(\alpha+k-1)\beta^k.$$

Property 3.10.5. *Suppose that $X_1, X_2, \ldots, X_m$ are independent and $X_i \sim \Gamma(\alpha_i, \beta)$ for $i = 1, \ldots, m$. Then $X = \sum_{i=1}^{m} X_i \sim \Gamma(\alpha, \beta)$, where $\alpha = \sum_{i=1}^{m} \alpha_i$.*

Proof. We sequentially use the definition of the mgf, the property of the exponential function, the independence of $X_1, X_2, \ldots, X_m$, the mgf of the gamma

distribution in Property 3.11.2 and the property of the exponential function to obtain the mgf of X below:

$$M_X(t) = E\left(e^{tX}\right) = E\left(\prod_{i=1}^{m} e^{tX_i}\right) = \prod_{i=1}^{m} E\left(e^{tX_i}\right) = \prod_{i=1}^{m}\left(\frac{1}{1-\beta t}\right)^{\alpha_i}$$
$$= \left(\frac{1}{1-\beta t}\right)^{\alpha},$$

where $\alpha = \sum_{i=1}^{m} \alpha_i$. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

3.11 Uniform distribution

Definition 3.11.1. *If the pdf of random variable X is given by*

$$f_X(x) = \frac{1}{b-a} I(a \le x \le b), \tag{3.54}$$

the distribution of X is called a uniform distribution on the interval $[a, b]$, where a and b with $b > a$ are parameters. The X is called a uniform random variable. We write $X \sim unif[a, b]$.

By Definition 3.11.1, we see that the cdf of $X \sim unif[a, b]$ has the form

$$F_X(x) = \frac{x-a}{b-a} I(a \le x < b) + I(x \ge b)$$

for any $x \in \mathbb{R}$.

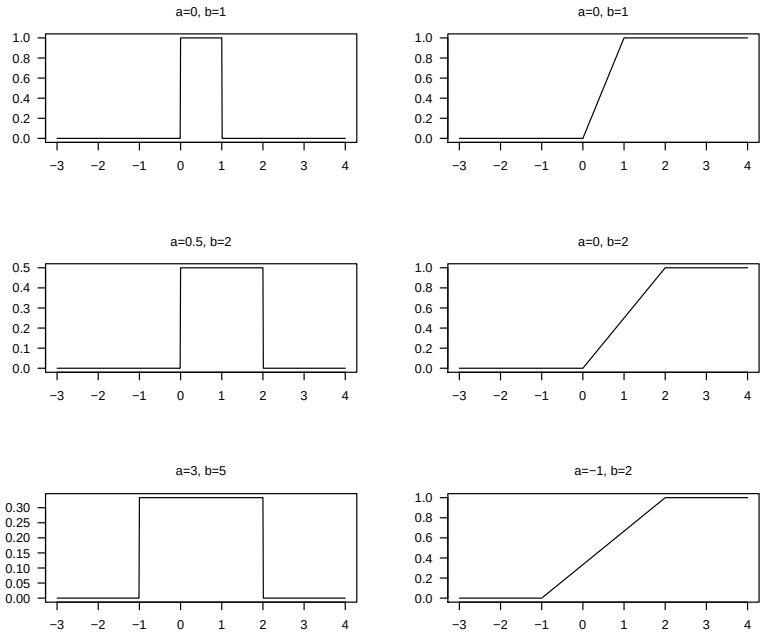


Figure 3.11.1 PDF and CDF of uniform distribution

Property 3.11.1. *Let $X \sim unif(a, b)$. Then the expectation and variance of X are given by*

$$E(X) = \frac{a+b}{2} \quad \textit{and} \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}. \tag{3.55}$$

Proof. Applying the definition of the expectation with the pdf given by (3.54), we have

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \frac{x^2}{2} \Big|_a^b = \frac{a+b}{2}. \tag{3.56}$$

To show the variance of uniform distribution $unif(a, b)$, we need $E(X^2)$.

Arguing along the same line as $E(X)$, we have

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_a^b x^2 \frac{1}{b-a}dx \\ &= \frac{1}{b-a}\frac{x^3}{3}\Big|_a^b = \frac{a^2+ab+b^2}{3}. \end{aligned} \tag{3.57}$$

Thus, combining (3.56) and (3.57) yields that

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}. \quad \square$$

Property 3.11.2. *Let $X \sim unif(a,b)$. Then the moment generating function of X is given by*

$$M_X(t) = \begin{cases} 1 & \text{if } t = 0, \\ \dfrac{e^{tb}-e^{ta}}{t(b-a)} & \text{if } t \neq 0, \end{cases}$$

Proof. By Definition 3.11.1, applying the definition of mgf will yield that

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \int_{-\infty}^{\infty} e^{tx} f_X(x)dx = \int_a^b e^{tx}\frac{1}{b-a}dx \\ &= \frac{1}{b-a}\frac{e^{tx}}{t}\Big|_a^b = \frac{e^{tb}-e^{ta}}{t(b-a)} \end{aligned} \tag{3.58}$$

if $t \neq 0$. Clearly, $M_X(0) = 1$ is true for any random variable X. $\square$

3.12 Beta distribution

Definition 3.12.1. *If the pdf of random variable X is given by*

$$f_X(x) = \begin{cases} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} & \text{if } x \in [0,1], \\ 0 & \text{otherwise}, \end{cases} \tag{3.59}$$

the distribution of X is called a beta distribution with shape parameters $a > 0$ and $b > 0$. We write $X \sim beta(a,b)$.

When $a = b = 1$, X has a uniform distribution on the interval $[0,1]$, i.e.,

$beta(1,1) = unif[0,1]$. Meanwhile, by Definition 3.12.1, we see that the cdf of $X \sim beta(a,b)$ does not have a closed form in general.

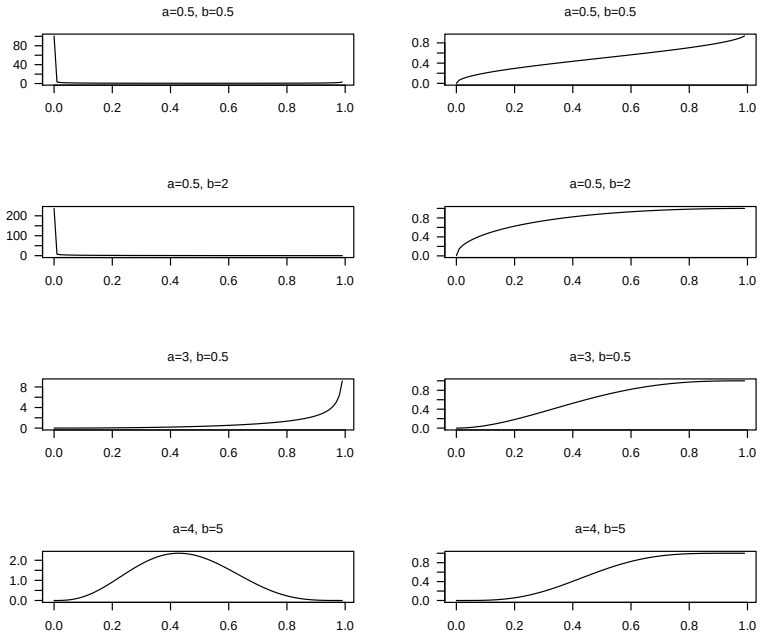


Figure 3.12.1 PDF and CDF of beta distribution

Property 3.12.1. *Let $X \sim beta(a,b)$. Then the expectation and variance of X are respectively given by*

$$E(X) = \frac{a}{a+b} \quad \textit{and} \quad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{3.60}$$

Proof. Applying the definition of the expectation with the pdf given by (3.59),

we have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^1 x \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^a (1-x)^{b-1} dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \\ &= \frac{a}{a+b}. \end{aligned} \tag{3.61}$$

To show the variance of beta distribution $beta(a, b)$, we need $E(X^2)$. Arguing along the same line as $E(X)$ will have

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &= \int_0^1 x^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 x^{a+1} (1-x)^{b-1} dx \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{a(a+1)\Gamma(a)\Gamma(b)}{(a+b)(a+b+1)\Gamma(a+b)} \\ &= \frac{a(a+1)}{(a+b)(a+b+1)}. \end{aligned} \tag{3.62}$$

Combining (3.61) and (3.62) concludes that

$$\begin{aligned} \text{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{a(a+1)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2 \\ &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

□

3.13 Weibull distribution

Definition 3.13.1. *If the pdf of random variable X is given by*

$$f_X(x) = \begin{cases} \dfrac{a}{b^a} x^{a-1} \exp\left(-\left(\dfrac{x}{b}\right)^a\right) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases} \tag{3.63}$$

the distribution of X is called a Weibull distribution with shape parameter $a > 0$ and scale parameter $b > 0$. The X is called a Weibull random variable. We write $X \sim weibull(a, b)$.

Remark 3.13.1. *When $a = 1$, the Weibull distribution reduces to an exponential distribution, i.e., $weibull(1, b) = exp(1/b)$.*

Remark 3.13.2. *Some textbooks define the Weibull distribution by using the rate parameter, which is equal to the reciprocal of the scale parameter.*

By Definition 3.13.1, we see that the cdf of $X \sim weibull(a, b)$ has the form

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - \exp\left(-\left(x/b\right)^a\right) & \text{if } x > 0. \end{cases}$$

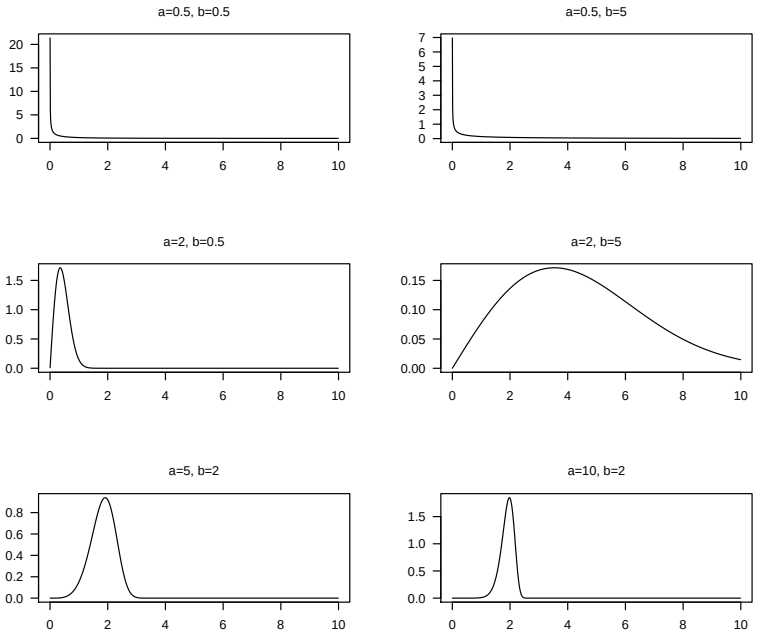


Figure 3.13.1 PDF of the Weibull distribution

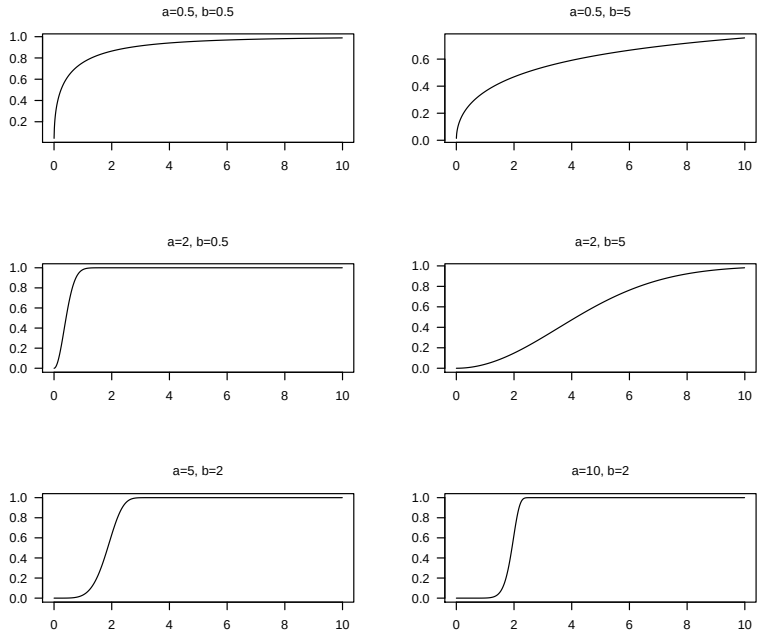


Figure 3.13.2 CDF of the Weibull distribution

Property 3.13.1. *Let $X \sim weibull(a, b)$. Then the expectation and variance of X are respectively given by*

$$\begin{aligned} E(X) &= b\Gamma(1 + 1/a), \\ \mathrm{Var}(X) &= b^2 \left(\Gamma(1 + 2/a) - [\Gamma(1 + 1/a)]^2\right). \end{aligned} \tag{3.64}$$

Proof. Applying the definition of the expectation with the pdf given by (3.63),

we have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^{\infty} x \frac{a}{b^a} x^{a-1} \exp\left(-(x/b)^a\right) dx \\ &= \int_0^{\infty} \frac{a}{b^a} x^a \exp\left(-(x/b)^a\right) dx \\ &= ab \int_0^{\infty} y^a \exp\left(-y^a\right) dy \\ &= b \int_0^{\infty} z^{1/a} e^{-z} dz \\ &= b\Gamma(1+1/a). \end{aligned} \tag{3.65}$$

To show the variance of gamma distribution $weibull(a, b)$, we need $E(X^2)$. Arguing along the same line as $E(X)$ will have

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &= \int_0^{\infty} x^2 \frac{a}{b^a} x^{a-1} \exp\left(-(x/b)^a\right) dx \\ &= \int_0^{\infty} \frac{a}{b^a} x^{a+1} \exp\left(-(x/b)^a\right) dx \\ &= ab^2 \int_0^{\infty} y^{a+1} \exp\left(-y^a\right) dy \\ &= b^2 \int_0^{\infty} z^{2/a} e^{-z} dz \\ &= b^2\Gamma(1+2/a). \end{aligned} \tag{3.66}$$

Thus, combining (3.65) and (3.66) yields that

$$\begin{aligned} \mathrm{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= b^2\Gamma(1+2/a) - \left(b\Gamma(1+1/a)\right)^2 \\ &= b^2\Big(\Gamma(1+2/a) - [\Gamma(1+1/a)]^2\Big). \end{aligned}$$

□

3.14 Cauchy distribution

Definition 3.14.1. *If the pdf of random variable X is given by*

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}, \quad x \in \mathbb{R}, \tag{3.67}$$

then X is called a standard Cauchy random variable and its distribution is called a standard Cauchy distribution. We write $X \sim C(0,1)$.

The cdf of $X \sim C(0,1)$ is given by

$$F_X(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}, \quad x \in \mathbb{R},$$

Remark 3.14.1. *Suppose that $X \sim C(0,1)$. Then the expectation is undefined.*

Proof. By the pdf (3.67), applying the definition of the expectation will yield that

$$\begin{aligned} E(|X|) &= \int_{-\infty}^{\infty} |x| f_X(x)dx \\ &= \lim_{a\to\infty}\lim_{b\to\infty}\int_{-a}^{b} |x| f_X(x)dx \\ &= \lim_{a\to\infty}\int_{-a}^{0} |x| f_X(x)dx + \lim_{b\to\infty}\int_{0}^{b} |x| f_X(x)dx \\ &= \lim_{a\to\infty}\int_{0}^{a} \frac{1}{\pi}\frac{y}{1+y^2}dy + \lim_{b\to\infty}\int_{0}^{b} \frac{1}{\pi}\frac{y}{1+y^2}dy \\ &= \frac{1}{2\pi}\left\{\lim_{a\to\infty}\ln(1+y^2)\Big|_0^a + \lim_{b\to\infty}\ln(1+y^2)\Big|_0^b\right\} \\ &= \infty. \end{aligned} \tag{3.68}$$

Thus, $E(X)$ is undefined. □

Definition 3.14.2. *Let $Y = \mu + \sigma X$, where μ and $\sigma > 0$ are real numbers and $X \sim C(0,1)$. Then the distribution of Y is called a Cauchy distribution with location parameter μ and scale parameter σ. We write $Y \sim C(\mu, \sigma)$.*

According to Definition 3.14.2, the pdf and cdf of Y are given by

$$\begin{aligned} f(y|\mu,\sigma^2) &= \frac{1}{\pi\sigma}\left\{1 + \left(\frac{y-\mu}{\sigma}\right)^2\right\}^{-1}, \quad y \in \mathbb{R}, \\ F(y|\mu,\sigma^2) &= \frac{1}{\pi}\arctan\left(\frac{y-\mu}{\sigma}\right) + \frac{1}{2}, \quad y \in \mathbb{R}, \end{aligned} \tag{3.69}$$

respectively.

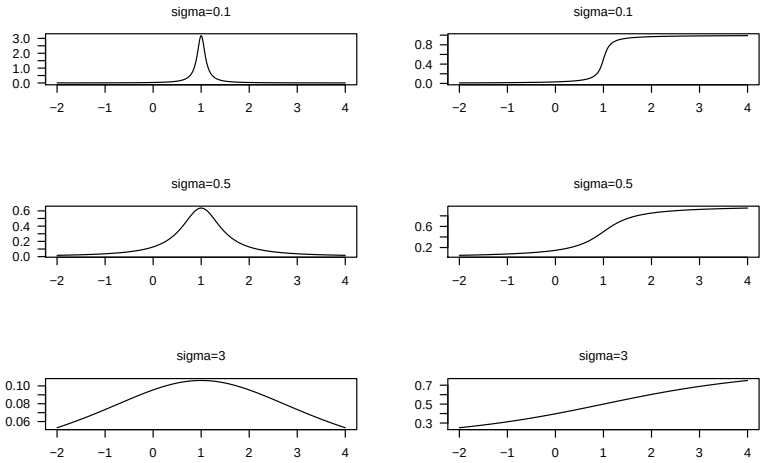


Figure 3.14.1 PDF and CDF of Cauchy distribution (location=1)

3.15 Multinomial distribution

Definition 3.15.1. *If the joint pmf of* $\mathbf{X} = (X_1, X_2, \ldots, X_k)^T$ *has the form*

$$P(\mathbf{X} = \mathbf{x}) = \frac{n!}{\prod_{j=1}^{k} x_j!} \prod_{j=1}^{k} p_j^{x_j},$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_k)^T$ *with* $\sum_{j=1}^{k} x_j = n$ *and* $\sum_{j=1}^{k} p_j = 1$, *then* $\mathbf{X}$ *is called a multinomial random vector and its distribution is called a multinomial distribution with parameters* n *and* $\mathbf{p} = (p_1, p_2, \ldots, p_k)^T$. *We write* $\mathbf{X} \sim multinom(n, \mathbf{p})$.

Property 3.15.1. *Let* $\mathbf{X} \sim multinom(n, \mathbf{p})$. *Then the expected value and*

covariance matrix are given by

$$E(\mathbf{X}) = n\mathbf{p} \quad \textit{and} \quad \mathrm{Cov}(\mathbf{X}) = n\big(\mathrm{diag}(p_1, p_2, \ldots, p_k) - \mathbf{p}\mathbf{p}^T\big).$$

respectively.

Property 3.15.2. *If* $\mathbf{X} \sim multinom(n, \mathbf{p})$*, then the joint mgf of* $\mathbf{X}$ *is given by*

$$M_{\mathbf{X}}(\mathbf{t}) = \left(\sum_{j=1}^{k} p_j e^{t_j}\right)^n.$$

Property 3.15.3. *If* $\mathbf{X} \sim multinom(n, \mathbf{p})$*, then* $X_i \sim b(n, p_i)$ *for* $i = 1, 2, \ldots, k$.

Property 3.15.4. *Suppose that* $\mathbf{X} \sim multinom(n, \mathbf{p})$*. If* $\{i_1, i_2, \ldots, i_m\}$ *and* $\{i_{m+1}, \ldots, i_k\}$ *form a partition of* $\{1, 2, \ldots, k\}$*, then the conditional distribution of* $X_{i_1}, \ldots, X_{i_m}$ *given* $X_{i_{m+1}}, \ldots, \ldots, X_{i_k}$ *is* $multinom(n - \sum_{j=m+1}^{k} X_{i_j}, \mathbf{q})$*, where* $\mathbf{q} = (q_1, q_2, \ldots, q_m)^T$ *with*

$$q_j = \frac{p_{i_j}}{\sum_{l=1}^{m} p_{i_l}}, \quad j = 1, 2, \ldots, m.$$

Property 3.15.5. *Suppose that* $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ *are independent and* $\mathbf{X}_i \sim multinom(n_i, \mathbf{p})$ *for* $i = 1, 2, \ldots, m$*. Then* $\mathbf{X} = \sum_{i=1}^{m} \mathbf{X}_i \sim multinom(n, \mathbf{p})$*, where* $n = \sum_{i=1}^{m} n_i$.

Proof. Let $n = \sum_{i=1}^{m} n_i$. Then we sequentially use the definition of the mgf, the property of the exponential function, the independence of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$, the mgf of the multinomial distribution in Theorem 3.1.3 and the property of the exponential function to obtain the mgf of X below:

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) = E\left(e^{\mathbf{t}^T\mathbf{X}}\right) = E\left(\prod_{i=1}^{m} e^{\mathbf{t}^T\mathbf{X}_i}\right) &= \prod_{i=1}^{m} E(e^{\mathbf{t}^T\mathbf{X}_i}) \\ = \prod_{i=1}^{m} \left(\sum_{j=1}^{k} p_j e^{t_j}\right)^{n_i} &= \left(\sum_{j=1}^{k} p_j e^{t_j}\right)^n. \end{aligned}$$

The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

3.16 Multivariate normal distribution

Definition 3.16.1. *Let* $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^T$, *where* $Z_1, Z_2, \ldots, Z_n$ *are i.i.d. with* $N(0,1)$. *Let* $\mathbf{X} = \boldsymbol{\mu} + A\mathbf{Z}$, *where* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T$ *is a real vector and* A *is an* $n \times n$ *matrix. Then the distribution of* $\mathbf{X}$ *is called a multivariate normal distribution. We write* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *where* $\boldsymbol{\Sigma} = AA^T$.

Property 3.16.1. *If* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then the joint pdf of* $\mathbf{X}$ *is equal to*

$$f(\mathbf{x}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad \mathbf{x} \in \mathbb{R}^n.$$

Property 3.16.2. *If* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then*

$$E(\mathbf{X}) = \boldsymbol{\mu} \quad \textit{and} \quad \mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}.$$

Property 3.16.3. *If* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *then the joint mgf of* $\mathbf{X}$ *is*

$$M_{\mathbf{X}}(\mathbf{t}) = \exp\left(\boldsymbol{\mu}^T\mathbf{t} + \frac{1}{2}\mathbf{t}^T\boldsymbol{\Sigma}\mathbf{t}\right), \quad \mathbf{t} \in \mathbb{R}^n.$$

Property 3.16.4. *If* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and* $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b} \in \mathbb{R}^q$, *then*

$$\mathbf{Y} \sim N_q(\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T).$$

Property 3.16.5. *Let* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *and let*

$$\mathbf{X} = \begin{pmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{pmatrix},$$

where $\mathbf{X}_1 \in \mathbb{R}^m$ *and* $\boldsymbol{\mu}_1 \in \mathbb{R}^m$, *and* $\boldsymbol{\Sigma}_{11}$ *is an* $m \times m$ *matrix. Then*

(i) $\mathbf{X}_1 \sim N_m(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ *and* $\mathbf{X}_2 \sim N_{n-m}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$,

(ii) $(\mathbf{X}_1|\mathbf{X}_2) \sim N_m\Big(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\Big)$,

(iii) $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent if and only if* $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Remark 3.16.1. $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is called a bivariate normal distribution, see page 198.*

4

Convergence of Random Variables

4.1 Some inequalities

4.1.1 Markov's inequality

Theorem 4.1.1. *(Markov's inequality) If X is a nonnegative random variable with $E(X) < \infty$, then for any $\epsilon > 0$,*

$$P(X \geq \epsilon) \leq \frac{1}{\epsilon} E(X). \tag{4.1}$$

Proof. For any $\epsilon > 0$, we define a random variable X_ϵ by

$$X_\epsilon = X - \epsilon I(X \geq \epsilon) = \begin{cases} X - \epsilon & \text{if } X \geq \epsilon, \\ X & \text{if } X < \epsilon, \end{cases} \tag{4.2}$$

where $I(X \geq \epsilon)$ is the indicator function of event $\{X \geq \epsilon\}$. Clearly, X_ϵ is nonnegative and $X_\epsilon = 0$ iff $X \in \{0, \epsilon\}$. Since $E(X) < \infty$, $E(X_\epsilon) < \infty$. Taking the expectation on both sides of (4.2) yields that

$$E(X_\epsilon) = E(X) - \epsilon P(X \geq \epsilon).$$

Since $P(X_\epsilon \geq 0) = 1$, $E(X_\epsilon) \geq 0$, which means that

$$E(X) - \epsilon P(X \geq \epsilon) \geq 0. \tag{4.3}$$

The inequality in (4.3) is clearly equivalent to

$$P(X \geq \epsilon) \leq \frac{1}{\epsilon} E(X). \qquad \square$$

Remark 4.1.1. *We need to mention that the positive real number ϵ in Markov's inequality (4.1) is usually chosen to be greater than $E(X)$ because if $\epsilon \leq E(X)$, the upper bound $E(X)/\epsilon$ will be greater than or equal to 1, which does not provide useful information for $P(X \geq \epsilon)$.*

Theorem 4.1.2. *Let X be a nonnegative random variable with $E(X) < \infty$. If there exists a positive real number ϵ such that Markov's inequality (4.1) becomes an equality, then $P(X \in \{0, \epsilon\}) = 1$.*

Proof. For $\epsilon > 0$ and nonnegative random variable X, we consider X_ϵ and its expectation $E(X_\epsilon)$ given by (4.2) and (4.3), respectively. We see from (4.3) that Markov's inequality (4.1) becomes an equality iff $E(X_\epsilon) = 0$. Since X_ϵ is nonnegative, $E(X_\epsilon) = 0$ is equivalent to $P(X_\epsilon = 0) = 1$, which means that $P(X \in \{0, \epsilon\}) = 1$. □

Remark 4.1.2. *Markov's inequality applies only to nonnegative random variables. Without this restriction, inequality can be false.*

Example 4.1.1. *Let the random variable X have a standard normal distribution $N(0, 1)$. Then $E(X) = 0$. For any $\epsilon > 0$, if we use Markov's inequality to bound the probability $P(X \geq \epsilon)$, we obtain that*

$$P(X \geq \epsilon) \leq \frac{1}{\epsilon} E(X) = 0,$$

which is clearly wrong because $P(X \geq \epsilon) > 0$.

4.1.2 Chebyshev's inequality

While Markov's inequality (4.1) takes $E(X)$ into account and provides an upper bound on $P(X \geq \epsilon)$ for the nonnegative random variable X, Chebyshev's inequality below uses $\mathrm{Var}(X)$ to present an upper bound on $P(|X - E(X)| \geq \epsilon)$ for any random variable X with finite variance.

Theorem 4.1.3. *(Chebyshev's inequality) Let X be a random variable with $\mathrm{Var}(X) < \infty$. Then for any $\epsilon > 0$,*

$$P(|X - E(X)| \geq \epsilon) \leq \frac{1}{\epsilon^2} \mathrm{Var}(X). \tag{4.4}$$

Proof. Let $\mu = E(X)$. Then for $\epsilon > 0$, we define a random variable X_ϵ below

$$\begin{aligned} X_\epsilon &= (X - \mu)^2 - \epsilon^2 I(|X - \mu| \geq \epsilon) \\ &= \begin{cases} (X - \mu)^2 - \epsilon^2 & \text{if } |X - \mu| \geq \epsilon, \\ (X - \mu)^2 & \text{if } |X - \mu| < \epsilon, \end{cases} \end{aligned} \tag{4.5}$$

where $I(|X - \mu| \geq \epsilon)$ is the indicator function of the event $\{|X - \mu| \geq \epsilon\}$.

Clearly, X_ϵ is nonnegative and $X_\epsilon = 0$ iff $X \in \{\mu-\epsilon, \mu, \mu+\epsilon\}$. Since $\mathrm{Var}(X) < \infty$, we take the expectation on both sides of (4.5) and obtain that

$$\begin{aligned} E(X_\epsilon) &= E[(X-\mu)^2] - \epsilon^2 P(|X-\mu| \geq \epsilon) \\ &= \mathrm{Var}(X) - \epsilon^2 P(|X-\mu| \geq \epsilon). \end{aligned} \tag{4.6}$$

Since $P(X_\epsilon \geq 0) = 1$, $E(X_\epsilon) \geq 0$. We conclude from (4.6) that

$$\mathrm{Var}(X) - \epsilon^2 P\big(|X - E(X)| \geq \epsilon\big) \geq 0,$$

which is equivalent to

$$P\big(|X - E(X)| \geq \epsilon\big) \leq \frac{1}{\epsilon^2}\mathrm{Var}(X). \qquad \square$$

Remark 4.1.3. *Chebyshev's inequality is often said to be a corollary of Markov's inequality because it is usually proved by applying Markov's inequality (4.1) to the random variable* $|X - E(X)|^2$, *i.e.,*

$$\begin{aligned} P(|X - E(X)| \geq \epsilon) &= P(|X - E(X)|^2 \geq \epsilon^2) \\ &\leq \frac{1}{\epsilon^2} E(|X - E(X)|^2) \\ &= \frac{1}{\epsilon^2}\mathrm{Var}(X). \end{aligned}$$

Remark 4.1.4. *Chebyshev's inequality is sometimes stated by an alternative form. Let* X *be a random variable with* $\sigma^2 = \mathrm{Var}(X) < \infty$. *Then for any* $\epsilon > 0$,

$$P(|X - E(X)| \geq \epsilon\sigma) \leq \frac{1}{\epsilon^2}.$$

Remark 4.1.5. *To make the Chebyshev's upper bound* $\mathrm{Var}(X)/\epsilon^2$ *on the probability* $P(|X - E(X)| \geq \epsilon)$ *useful, we should choose* ϵ *to be greater than* σ, *the standard deviation of* X.

Theorem 4.1.4. *Let* X *be a random variable with* $E(X) = \mu$ *and* $\mathrm{Var}(X) = \sigma^2 < \infty$. *If there exists a positive real number* ϵ *such that Chebyshev's inequality (4.4) becomes an equality, then* $P(X \in \{\mu - \epsilon, \mu, \mu + \epsilon\}) = 1$.

Proof. For $\epsilon > 0$ and random variable X, we consider X_ϵ and its expectation $E(X_\epsilon)$ given by (4.5) and (4.6), respectively. We see from (4.6) that Chebyshev's inequality (4.4) becomes an equality if and only if $E(X_\epsilon) = 0$. Since X_ϵ is nonnegative, $E(X_\epsilon) = 0$ is equivalent to $P(X_\epsilon = 0) = 1$, which means that

$$P(X \in \{\mu - \epsilon, \mu, \mu + \epsilon\}) = 1. \qquad \square$$

Theorem 4.1.5. *If* $\mathrm{Var}(X) = 0$*, then* $P\big(X = E(X)\big) = 1$.

Proof. For any $\epsilon > 0$, we obtain from the Chebyshev's inequality that

$$0 \le P\left(|X - E(X)| \ge \epsilon\right) \le \frac{1}{\epsilon^2}\mathrm{Var}(X) = 0,$$

i.e., $P\big(|X - E(X)| \ge \epsilon\big) = 0$ for any $\epsilon > 0$.

Let $\epsilon = 1/n$ and let $A_n = \{|X - E(X)| \ge 1/n\}$ for $n \in \mathbb{N}$. Then $(A_n)_{n\in\mathbb{N}}$ is a nondecreasing sequence of events with $P(A_n) = 0$ for any $n \in \mathbb{N}$ and

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} \{|X - E(X)| \ge 1/n\} = \{|X - E(X)| > 0\}.$$

Using the continuity of the probability, we obtain

$$\begin{aligned} 0 = P(A_n) &= \lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) \\ &= P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\big(|X - E(X)| > 0\big), \end{aligned}$$

which is equivalent to $P\big(X = E(X)\big) = 1$. □

4.1.3 Jensen's inequality

While inequalities in the previous two subsections provide bounds on the probabilities of certain events, we discuss a famous inequality in this subsection, which provides a bound on the expectation of a convex function of a random variable.

Theorem 4.1.6. *(Jensen's inequality) Let* X *be a random variable such that* $E(|X|)$ *and* $E[|\varphi(X)|]$ *exist, where* $\varphi(x)$ *is a convex function of* x*. Then*

$$E[\varphi(X)] \ge \varphi\big(E(X)\big) \tag{4.7}$$

with equality if and only if $P\big(\varphi(X) = aX + b\big) = 1$.

Proof. It can be shown that if φ is convex, then φ lies above any line that touches φ at some point.

Let $l(x) = \varphi(\mu) + s_\mu(x - \mu)$ be the equation of the tangent line of $\varphi(x)$ at μ, where $\mu = E(X)$. A line that touches a curve at a point without crossing over. Then

$$\varphi(x) \ge \varphi(\mu) + s_\mu(x - \mu)$$

for all x. Thus,

$$\varphi(X) \ge \varphi(\mu) + s_\mu(X - \mu). \tag{4.8}$$

Taking expectations on both sides of (4.8), we obtain

$$E[\varphi(X)] \geq \varphi(\mu) + s_\mu E(X - \mu) = \varphi(\mu)$$

as required.

Furthermore, if $\varphi(x)$ is linear, then equality follows from the linear property of expectation. Now we prove the converse by contradiction. Assume that $\varphi(x)$ is not linear. Since $\varphi(x)$ is convex, we have $\varphi(x) > l(x)$ for $x \neq \mu$, where $l(x)$ is the tangent line of $\varphi(x)$ at μ. Thus,

$$E[\varphi(X)] > E[l(X)] = \varphi(\mu) + s_\mu E(X - \mu) = \varphi(\mu),$$

which contradicts the condition $E[\varphi(X)] = \varphi(\mu)$. □

Example 4.1.2. *Let X be a random variable with $E(X^2) < \infty$. Then* $\mathrm{Var}(X) \geq 0$.

Proof. Since the function x^2 is convex, applying Jensen's inequality yields

$$E(X^2) \geq [E(X)]^2,$$

which means that $E(X^2) - [E(X)]^2 \geq 0$, i.e., $\mathrm{Var}(X) \geq 0$. □

Example 4.1.3. *For $i = 1, 2, \ldots, n$, let $a_i > 0$ and let*

$$AM = \frac{1}{n}\sum_{i=1}^{n} a_i, \quad GM = \left(\prod_{i=1}^{n} a_i\right)^{1/n}, \quad HM = \left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{a_i}\right)^{-1}.$$

represent the arithmetic mean, geometric mean, and harmonic mean, respectively. Then $HM \leq GM \leq AM$.

Proof. Let X be a random variable with a uniform distribution at points $a_1, a_2, \ldots, a_n$, i.e.,

$$P(X = a_i) = \frac{1}{n}, \quad i = 1, 2, \ldots, n.$$

(i) Let $\varphi(x) = -\ln(x)$ for $x > 0$. Then $\varphi(x)$ is convex because $\varphi''(x) > 0$. Applying Jensen's inequality, we obtain $E[\varphi(X)] \geq \varphi\big(E(X)\big)$. We now evaluate $E(X)$ and $E[\varphi(X)]$ below

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} a_i = AM,$$

$$E[\varphi(X)] = \frac{1}{n}\sum_{i=1}^{n} \varphi(a_i) = -\frac{1}{n}\sum_{i=1}^{n} \ln(a_i) = -\ln\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \varphi(GM).$$

Thus, $E[\varphi(X)] \geq \varphi\big(E(X)\big)$ is equivalent to $\varphi(GM) \geq \varphi(AM)$. Since $\varphi(x)$ is a monotonically decreasing function of x, $GM \leq AM$.

(ii) Let the random variable Y have a uniform distribution at points $1/a_i > 0, i = 1, 2, \ldots, n$. Applying Jensen's inequality to $\varphi(y) = -\ln(y)$ with random variable Y, we obtain

$$E[\varphi(Y)] \geq \varphi\big(E(Y)\big).$$

We now compute $E(Y)$ and $E[\varphi(Y)]$ below:

$$E(Y) = \frac{1}{n}\sum_{1}^{n}\frac{1}{a_i},$$

$$E[\varphi(Y)] = -\frac{1}{n}\sum_{1}^{n}\ln\left(\frac{1}{a_i}\right) = \ln\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \ln(GM).$$

Thus, $E[\varphi(Y)] \geq \varphi\big(E(Y)\big)$ means that

$$\ln(GM) \geq -\ln\left(\frac{1}{n}\sum_{1}^{n}\frac{1}{a_i}\right) = \ln(HM),$$

which is equivalent to $HM \leq GM$ because $\ln(t)$ is a monotonically increasing function of t.

Combining (i) and (ii) completes the proof. □

4.2 Convergence in distribution

Definition 4.2.1. *Consider a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables and a corresponding sequence $\big(F_{X_n}\big)_{n\in\mathbb{N}}$ of cdfs. If there exists a random variable X with cdf F_X such that*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for every continuity point x of F_X, we say that X_n converges in distribution (or in law, or weakly) to X and we usually write $X_n \xrightarrow{d} X$.

Remark 4.2.1. *Definition 4.2.1 does not require $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all $x \in \mathbb{R}$.*

Example 4.2.1. *Let $P(X = 0) = 1$ and let $P(X_n = 1/n) = 1$ for $n \in \mathbb{N}$. Then $X_n \xrightarrow{d} X$.*

Proof. Clearly, $x = 0$ is only discontinuity point of F_X and when $x = 0$, we have $\lim_{n\to\infty} F_{X_n}(0) = 0 \neq 1 = F_X(0)$. However, when $x \neq 0$,

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases} = F_X(x).$$

Because of the exception of discontinuity of F_X in Definition 4.2.1, we do have $X_n \xrightarrow{d} X$. □

Remark 4.2.2. *The limit of a sequence $(F_{X_n})_{n\in\mathbb{N}}$ of cdfs may not be a cdf.*

Example 4.2.2. *For $n \in \mathbb{N}$, let $X_n \sim N(0, 1/n^2)$. Since $nX_n \sim N(0, 1)$, the cdf F_{X_n} of X_n is given by*

$$F_{X_n}(x) = P(X_n \leq x) = P(nX_n \leq nx) = \Phi(nx),$$

which leads to

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} \Phi(nx) = \begin{cases} 0 & \text{if } x < 0, \\ 1/2 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

It is clear that this limit $\lim_{n\to\infty} F_{X_n}(x)$ is not a cdf because it is not right continuous at $x = 0$.

Remark 4.2.3. *Since the concept of convergence in distribution involves the distributions of random variables only, not the random variables themselves, it is possible that for all outcome $\omega \in \Omega$, the sequence $X_n(\omega)$ does not converge even though $X_n \xrightarrow{d} X$.*

Example 4.2.3. *For $n \in \mathbb{N}$, let $X_n = (-1)^n U$, where the random variable U represents the outcome of tossing a fair coin once and is defined by*

$$U = \begin{cases} 1 & \text{if outcome is a head,} \\ -1 & \text{if the outcome is a tail.} \end{cases}$$

Since $P(U = -1) = P(U = 1) = 1/2$, i.e., U is symmetric about zero, X_n has the same distribution as U for any $n \in \mathbb{N}$. Definitely, $X_n \xrightarrow{d} X$. However,

$$X_n(\{head\}) = (-1)^n U(\{head\}) = (-1)^n,$$
$$X_n(\{tail\}) = (-1)^n U(\{tail\}) = (-1)^{n+1}.$$

Clearly, both $X_n(\{head\})$ and $X_n(\{tail\})$ do not converge. This example is related to the relation between convergence in distribution and almost sure convergence.

Remark 4.2.4. *A sequence of continuous random variables may converge in distribution to a discrete random variable.*

Example 4.2.4. *Let $P(X = 1) = 1$ and let $X_n \sim beta(n, 1)$ for $n \in \mathbb{N}$. Clearly, X is a discrete random variable, while for each $n \in \mathbb{N}$, X_n is a continuous random variable. Furthermore, the cdf of X_n is given by*

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } x < 0, \\ x^n & \text{if } 0 \le x < 1, \\ 1 & \text{if } x \ge 1, \end{cases}$$

which leads to

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0 & \text{if } x < 1 \\ 1 & \text{if } x \ge 1 \end{cases} = F_X(x).$$

Remark 4.2.5. *It is possible that a sequence of discrete random variables converges in distribution to a continuous random variable, see examples after the central limit theorem.*

Remark 4.2.6. *Let X and X_n be continuous random variables for $n \in \mathbb{N}$. Then $X_n \xrightarrow{d} X$ does not imply convergence of the corresponding pdf.*

Example 4.2.5. *Let $X \sim unif[0, 1]$ and let X_n have the following cdf*

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } x < 0, \\ x - \sin(n\pi x)/(n\pi) & \text{if } 0 \le x < 1, \\ 1 & \text{if } x \ge 1 \end{cases}$$

for $n \in \mathbb{N}$. Clearly, X is a continuous random variable. Meanwhile, X_n is also a continuous random variable for each $n \in \mathbb{N}$. Furthermore,

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \le x < 1, \\ 1 & \text{if } x \ge 1, \end{cases}$$

which is the cdf of X. Thus, we have $X_n \xrightarrow{d} X$. The pdf of X_n, however, has the form

$$f_{X_n}(x) = 1 - \cos(n\pi x), \quad 0 \le x \le 1.$$

As $n \to \infty$, the limit of $f_{X_n}(x)$ doesn't exist.

Theorem 4.2.1. *Let all X_n and X be nonnegative integer-valued random variables. Then $X_n \xrightarrow{d} X$ if and only if $\lim_{n\to\infty} P(X_n = i) = P(X = i)$ for $i \in \mathbb{W}$ = set of nonnegative integers.*

Proof. Since all X_n and X are nonnegative integer-valued random variables, their cdfs are continuous for all $x \in \mathbb{R}\backslash\mathbb{W}$. Hence, $X_n \xrightarrow{d} X$ implies that

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad x \in \mathbb{R}\backslash\mathbb{W}.$$

Thus for $i \in \mathbb{W}$ and $\epsilon \in (0,1)$, we obtain

$$\begin{aligned}\lim_{n\to\infty} P(X_n = i) &= \lim_{n\to\infty}[F_{X_n}(i+\epsilon) - F_{X_n}(i-\epsilon)] \\ &= \lim_{n\to\infty} F_{X_n}(i+\epsilon) - \lim_{n\to\infty} F_{X_n}(i-\epsilon) \\ &= F_X(i+\epsilon) - F_X(i-\epsilon) \\ &= P(X = x).\end{aligned}$$

To prove the converse, let $x \in [0,\infty)\backslash\mathbb{W}$. Then

$$\begin{aligned}\lim_{n\to\infty} F_{X_n}(x) &= \lim_{n\to\infty} P(X_n \le x) = \lim_{n\to\infty} \sum_{i=0}^{[x]} P(X_n = i) \\ &= \sum_{i=0}^{[x]} \lim_{n\to\infty} P(X_n = i) = \sum_{i=0}^{[x]} P(X = i) \\ &= P(X \le x) = F_X(x).\end{aligned}$$

When $x < 0$, $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ is always true because $F_{X_n}(x) = F_X(x) = 0$ for all $n \in \mathbb{N}$. □

Example 4.2.6. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of binomial random variables with $X_n \sim b(n, \lambda/n)$ for $n \in \mathbb{N}$ and $n > \lambda$, where $\lambda > 0$ is a fixed constant. Then $X_n \xrightarrow{d} X$, where $X \sim \mathcal{P}(\lambda)$.*

Proof. Since $X_n \sim b(n, \lambda/n)$, we have

$$\begin{aligned}\lim_{n\to\infty} P(X_n = i) &= \lim_{n\to\infty} \binom{n}{i} \left(\frac{\lambda}{n}\right)^i \left(1 - \frac{\lambda}{n}\right)^{n-i} \\ &= \frac{\lambda^i}{i!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-i+1)}{n^i} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-i} \\ &= \frac{\lambda^i}{i!} e^{-\lambda}\end{aligned}$$

because for fixed $\lambda > 0$ and i,

$$\begin{aligned}\lim_{n\to\infty} \frac{n(n-1)\cdots(n-i+1)}{n^i} &= 1, \\ \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n &= e^{-\lambda}, \\ \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-i} &= 1.\end{aligned}$$

Since $(\lambda^i/i!)e^{-\lambda}$ for $i \in \mathbb{W}$ is the pmf of Poisson distribution $\mathcal{P}(\lambda)$, we can conclude from Theorem 4.2.1 that $X_n \xrightarrow{d} X$, where $X \sim \mathcal{P}(\lambda)$. □

We have used Definition 4.2.1 to show convergence in distribution in all the above examples. This method, however, may not be efficient for some problems. For example, suppose that $(U_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with uniform distribution on the interval $[0,1]$. For $n \in \mathbb{N}$, define $X_n = \sqrt{12n}\left(n^{-1}\sum_{i=1}^{n} U_i - 1/2\right)$. What is the limiting distribution of X_n as $n \to \infty$? Using the Definition 4.2.1 to answer this question appears to be a challenge. We now introduce a theorem that states the equivalence between convergence in distribution and convergence of their corresponding mgfs.

Theorem 4.2.2. *For $n \in \mathbb{N}$, let X_n have cdf F_{X_n} and mgf $M_{X_n}(t)$ for $|t| \le h$ with some $h > 0$. Let X have cdf F_X and mgf $M_X(t)$ for $|t| \le h^* \le h$. If $\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$ for all $|t| \le h^*$, then $X_n \xrightarrow{d} X$.*

Example 4.2.7. *(Use the mgf to do Example 5.2.6)*

Example 4.2.8. *For $n \in \mathbb{N}$, let $X_n = (Y_n - \lambda_n)/\sqrt{\lambda_n}$, where $Y_n \sim \mathcal{P}(\lambda_n)$ and $\lim_{n\to\infty} \lambda_n = \infty$. Show that $X_n \xrightarrow{d} X$, where $X \sim N(0,1)$.*

Proof. Since the mgf of $Y_n \sim \mathcal{P}(\lambda_n)$ is $M_{Y_n}(t) = \exp\left(\lambda_n(e^t - 1)\right)$ for $t \in \mathbb{R}$, the mgf of X_n is

$$\begin{aligned} M_{X_n}(t) &= e^{-t\sqrt{\lambda_n}} M_{Y_n}\left(\frac{t}{\sqrt{\lambda_n}}\right) \\ &= e^{-t\sqrt{\lambda_n}} \exp\left\{\lambda_n\left(e^{t/\sqrt{\lambda_n}} - 1\right)\right\} \\ &= e^{-t\sqrt{\lambda_n}} \exp\left\{\lambda_n\left(\frac{t}{\sqrt{\lambda_n}} + \frac{t^2}{2\lambda_n} + \frac{t^3}{6\lambda_n\sqrt{\lambda_n}} + \cdots\right)\right\} \\ &= \exp\left(\frac{t^2}{2} + \frac{t^3}{6\sqrt{\lambda_n}} + \cdots\right), \end{aligned}$$

which implies

$$\lim_{n\to\infty} M_{X_n}(t) = \exp\left(t^2/2\right), \quad t \in \mathbb{R}.$$

This is the mgf of standard normal distribution $N(0,1)$. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately. □

Example 4.2.9. *Suppose that $(U_n)_{n\in\mathbb{N}}$ is a sequence of independent and identically distributed (i.i.d.) random variables with a uniform distribution on the interval $[0,1]$. For $n \in \mathbb{N}$, define $X_n = \sqrt{12n}\left(n^{-1}\sum_{i=1}^{n} U_i - 1/2\right)$. Find the limiting distribution of X_n as $n \to \infty$.*

Solution. *Since the mgf of $U_1 \sim unif[0,1]$ is equal to*

$$M_{U_1}(t) = \begin{cases} 1 & \text{if } t = 0, \\ (e^t - 1)/t & \text{if } t \neq 0, \end{cases}$$

the mgf of X_n is

$$\begin{aligned} M_{X_n}(t) &= e^{-t\sqrt{12n}/2}\left\{M_{U_1}\left(\frac{\sqrt{12n}t}{n}\right)\right\}^n \\ &= \begin{cases} 1 & \text{if } t = 0, \\ e^{-t\sqrt{12n}/2}\left[(e^{\sqrt{12n}t/n} - 1)/(\sqrt{12n}t/n)\right]^n & \text{if } t \neq 0 \end{cases} \\ &= \begin{cases} 1 & \text{if } t = 0, \\ \left\{1 + \dfrac{t^2}{2n} + \dfrac{3t^4}{40n^2} + \cdots\right\}^n & \text{if } t \neq 0. \end{cases} \end{aligned}$$

Here we have used Taylor's expansion of the exponential function $e^z = \sum_{i=0}^{\infty} \frac{z^i}{i!}$. Thus, we obtain

$$\lim_{n\to\infty} M_{X_n}(t) = \exp\left(t^2/2\right), \quad t \in \mathbb{R},$$

which is the mgf of standard normal distribution $N(0,1)$. The desired result follows from the one-to-one correspondence between the distribution and the mgf immediately.

The most famous example of convergence in distribution is the central limit theorem (CLT). Example 4.2.9 is an application of the central limit theorem to a sequence of i.i.d. random variables from the uniform distribution $unif[0,1]$.

Theorem 4.2.3. *(Central limit theorem) Consider a sequence of i.i.d. random variables $X_1, X_2, \ldots$ from a population X with mgf M_X, $E(X) = \mu$, and $\mathrm{Var}(X) = \sigma^2$. For $n \in \mathbb{N}$, let*

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}}.$$

Then

$$\lim_{n\to\infty} P\left(Z_n \leq z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx = \Phi(z)$$

for $z \in \mathbb{R}$. That is, the limiting distribution of Z_n is a standard normal distribution $N(0,1)$.

Proof. We first assume that $\mu = 0$ and $\sigma^2 = 1$. Since the mgf of $X_i/\sqrt{n}$ is given by

$$E\left[\exp\left(tX_i/\sqrt{n}\right)\right] = M_X\left(t/\sqrt{n}\right),$$

the mgf of $Z_n = \sum_{i=1}^n X_i/\sqrt{n}$ is given by

$$M_{Z_n}(t) = \left[M_X\left(t/\sqrt{n}\right)\right]^n.$$

To prove the theorem, we must show that $[M_X(t/\sqrt{n})]^n \to \exp\left(t^2/2\right)$ as $n \to \infty$, or, equivalently, that $nL(t/\sqrt{n}) \to t^2/2$ as $n \to \infty$, where $L(t) = \ln[M_X(t)]$. In fact,

$$\begin{aligned}
L(0) &= 0, \\
L'(0) &= \frac{M_X'(0)}{M_X(0)} = \mu = 0, \\
L''(0) &= \frac{M_X(0)M_X''(0) - [M_X'(0)]^2}{[M_X(0)]^2} = \sigma^2 = 1,
\end{aligned}$$

which lead to

$$\begin{aligned}
\lim_{n\to\infty} \frac{L(t/\sqrt{n})}{n^{-1}} &= \lim_{n\to\infty} \frac{-L'(t/\sqrt{n})n^{-\frac{3}{2}}t}{-2n^{-2}} \\
&= \lim_{n\to\infty} \left[\frac{L'(t/\sqrt{n})t}{2n^{-\frac{1}{2}}}\right] \\
&= \lim_{n\to\infty} \left[L''\left(\frac{t}{\sqrt{n}}\right)\frac{t^2}{2}\right] \\
&= \frac{t^2}{2}.
\end{aligned}$$

Thus, the central limit theorem is proved when $\mu = 0$ and $\sigma^2 = 1$.

When $\mu \neq 0$ and $\sigma^2 \neq 1$, we consider the standardized random variables $X_i^* = (X_i - \mu)/\sigma$ for $i \in \mathbb{N}$. Then $E(X_i^*) = 0$ and $\mathrm{Var}(X_i^*) = 1$ for $i \in \mathbb{N}$ and the result follows by applying the preceding proof. □

Remark 4.2.7. *The CLT is sometimes stated by an alternative form. Consider a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with mgf M_{X_1}, $E(X_1) = \mu$, and $\mathrm{Var}(X_1) = \sigma^2 < \infty$. For $n \in \mathbb{N}$, let the random variable Z_n be defined by*

$$Z_n = \sqrt{n}\big(\bar{X}_n - \mu\big),$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then limiting distribution of Z_n is normal distribution $N(0,\sigma^2)$.

Remark 4.2.8. *The larger the sample size is, the better the approximation will be.*

Remark 4.2.9. *Continuity correction is often used when the sample is taken from a discrete distribution.*

Example 4.2.10. *Suppose that* 10 *fair dice are rolled. Find the approximate probability that the sum obtained is between* 30 *and* 40 *inclusively.*

Solution. *Let* X_i *denote the value of the* i*th die,* $i = 1, 2, \ldots, 10$*. Since*

$$E(X_i) = 7/2 \quad \textit{and} \quad \mathrm{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = 35/12,$$

the central limit theorem states that

$$\begin{aligned} P(30 \le X \le 40) &= P(29.5 \le X \le 40.5) \\ &= P\left(\frac{29.5 - 35}{\sqrt{\frac{350}{12}}} \le \frac{X - 35}{\sqrt{\frac{350}{12}}} \le \frac{40.5 - 35}{\sqrt{\frac{350}{12}}} \right) \\ &\approx \Phi(1.0184) - \Phi(-1.0184) \\ &\approx 0.6915. \end{aligned}$$

Hence around 69% *of the time* $\sum_{i=1}^{10} X_i$ *will be between* 30 *and* 40 *inclusively.*

Example 4.2.11. *Researchers over hundreds of years have consistently found that boys naturally outnumber girls at birth. The speculation is that this is nature's way of countering the relatively high mortality rates of males and creating more of a gender balance in the population. Historically, there have been about* 105 *boys born for every* 100 *girls worldwide. If we use this information as the probability of a newborn being a boy, What is the probability that at least half out of* 200 *newborns will be boys?*

Solution *For* $i = 1, 2, \ldots, 200$*, let* $X_i = 1$ *if the* i*th newborn is a boy and* $X_i = 0$ *otherwise. Then* $X_1, X_2, \ldots, X_{200}$ *are i.i.d. random variables with a Bernoulli distribution* $b(1, p)$*, where* $p = 105/205$*. Let* $X = \sum_{i=1}^{200} X_i$*. Then* $X \sim b(200, p)$*. If we use R to calculate* $P(X \ge 100)$*, we obtain*

$$P(X \ge 100) = 1 - P(X \le 99) = 1 - pbinom(99, 200, 105/205) \approx 0.6613.$$

Now we use the CLT with a continuity correction to approximate $P(X \ge 100)$

below

$$\begin{aligned} P(X \geq 100) &= P(X \geq 99.5) \\ &= P\left(\frac{X - 200(105/205)}{\sqrt{200(105/205)(100/205)}} \geq \frac{99.5 - 200(105/205)}{\sqrt{200(105/205)(100/205)}}\right) \\ &= P\left(\frac{X - 200(105/205)}{\sqrt{200(105/205)(100/205)}} \geq -0.4157645\right) \\ &\approx 1 - pnorm(-0.4157645, 0, 1) \\ &\approx 0.6612, \end{aligned}$$

which is very close to the result of using R directly.

Example 4.2.12. *Let $X \sim \mathcal{P}(100)$. Find $P(X \geq 120)$ by using R and the central limit theorem.*

Solution. *If we use R, we obtain*

$$P(X \geq 120) = 1 - P(X \leq 119) = 1 - ppois(119, 100) \approx 0.0282.$$

If we use the CLT with a continuity correction, we obtain

$$\begin{aligned} P(X \geq 120) = P(X \geq 119.5) &= P\left(\frac{X - 100}{\sqrt{100}} \geq \frac{119.5 - 100}{\sqrt{100}}\right) \\ &\approx 1 - \Phi(1.95) = 1 - pnorm(1.95, 0, 1) \approx 0.0256. \end{aligned}$$

Here we have used the fact that the variance of a Poisson random variable is equal to its mean and the additive property of Poisson distribution, i.e., X and $\sum_{i=1}^{100} X_i$ have the same distribution, where $X_1, X_2, \ldots, X_{100}$ form an i.i.d. sample from the Poisson distribution $\mathcal{P}(1)$.

Example 4.2.13. *Compute $\lim_{n\to\infty} \sum_{j=0}^{n} \frac{n^j}{j!} e^{-n}$.*

Solution. *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with a Poisson distribution $\mathcal{P}(1)$. By the additive property of Poisson distribution, $\sum_{i=1}^{n} X_i \sim \mathcal{P}(n)$. Applying the CLT, we obtain*

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n \cdot 1}{\sqrt{n \cdot 1}} \leq a\right) = \Phi(a)$$

for any $a \in \mathbb{R}$. Choosing $a = 0$ yields that

$$\lim_{n\to\infty} P\left(\sum_{i=1}^{n} X_i \leq n\right) = \frac{1}{2}.$$

Since $P\left(\sum_{i=1}^{n} X_i \leq n\right) = \sum_{j=0}^{n} \frac{n^j}{j!} e^{-n}$, *we can conclude that* $\lim_{n\to\infty} \sum_{j=0}^{n} \frac{n^j}{j!} e^{-n} = \frac{1}{2}$.

Remark 4.2.10. *The multivariate version of CLT can be stated as follows: Let* $\mathbf{X}_1, \mathbf{X}_2, \ldots$ *be a sequence of i.i.d. random vectors in* $\mathbb{R}^k$ *with mgf* $M_{\mathbf{X}}$, $E(\mathbf{X}_i) = \boldsymbol{\mu}$, *and* $\mathrm{Cov}(\mathbf{X}_i) = \boldsymbol{\Sigma}$. *Let the random vector*

$$\mathbf{Z}_n = \sqrt{n}\big(\bar{\mathbf{X}}_n - \boldsymbol{\mu}\big),$$

where $\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i$. *Then the limiting distribution of* $\mathbf{Z}_n$ *is a multivariate normal distribution* $N_k(\mathbf{0}, \boldsymbol{\Sigma})$.

4.3 Convergence in probability

Definition 4.3.1. *Consider a sequence* $(X_n)_{n\in\mathbb{N}}$ *of random variables and a random variable* X *defined on the same probability space. We say that* X_n *converges in probability to* X *and write* $X_n \xrightarrow{p} X$ *if for any* $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

Remark 4.3.1. *By Definition 4.3.1, we see that* $X_n \xrightarrow{p} X$ *is equivalent to* $X_n - X \xrightarrow{p} 0$, *which means that in the limit as* n *increases, almost all of the probability mass of* $X_n - X$ *will be concentrated in* $(-\epsilon, \epsilon)$, *no matter how small* ϵ *is. On the other hand, when* n *is fixed, it is possible to have a small probability mass outside* $(-\epsilon, \epsilon)$, *with a slowly decaying tail, which may have a strong impact on the expectation of* $X_n - X$. *Because of this, convergence in probability doesn't have any implication on the existence or convergence of* $E(X_n - X)$.

Example 4.3.1. *For* $n \in \mathbb{N}$, *let the joint pdf of* (X, X_n) *be defined by*

$$f(x,y) = \begin{cases} \dfrac{\lambda_n}{\pi} \dfrac{\exp\big(-\lambda_n(y-x)\big)}{1+x^2} & \text{if} \quad y \geq x, \\ \\ 0 & \text{if} \quad y < x, \end{cases}$$

where $\lim_{n\to\infty} \lambda_n = \infty$. *Then* $X_n \xrightarrow{p} X$ *and* $\lim_{n\to\infty} E(X_n - X) = 0$.

Proof. For any $\epsilon > 0$,

$$\begin{aligned} P(|X_n - X| \geq \epsilon) &= P(X_n \geq X + \epsilon) + P(X_n \leq X - \epsilon) \\ &= \int_{-\infty}^{\infty} \int_{x+\epsilon}^{\infty} f(x,y) dy dx + \int_{-\infty}^{\infty} \int_{-\infty}^{x-\epsilon} f(x,y) dy dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} \left(\int_{x+\epsilon}^{\infty} \lambda_n e^{-\lambda_n (y-x)} dy \right) dx + 0 \\ &= e^{-\lambda_n \epsilon} \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+x^2} dx \\ &= e^{-\lambda_n \epsilon} \\ &\to 0 \end{aligned}$$

as $n \to \infty$ because $\lim_{n\to\infty} \lambda_n = \infty$. This proves that $X_n \xrightarrow{p} X$. Furthermore, it can be shown that $X_n - X$ and X are independent, $X_n - X$ has an exponential distribution with rate parameter λ_n and X has a standard Cauchy distribution. By the property of exponential distribution, $E(X_n - X) = 1/\lambda_n$, which approaches to zero as $n \to \infty$.

It is worth mentioning that although $E(X_n - X)$ exists, $E(X_n)$ and $E(X)$ do not exist because X has a standard Cauchy distribution. □

The next example is similar to Example 4.3.1, but both $E(X_n)$ and $E(X)$ exist.

Example 4.3.2. *For $n \in \mathbb{N}$, let the joint pdf of (X, X_n) be defined by*

$$f(x,y) = \begin{cases} \lambda_n e^{-\lambda_n (y-x)} & \text{if} \quad y \geq x, \ 0 \leq x \leq 1, \\ 0 & \text{if} \quad y < x, \ 0 \leq x \leq 1, \end{cases}$$

where $\lim_{n\to\infty} \lambda_n = \infty$. That is, $(X_n | X = x)$ has a shifted-exponential distribution with rate parameter λ_n and $X \sim unif[0,1]$. Then $X_n \xrightarrow{p} X$. Furthermore, $\lim_{n\to\infty} E(X_n) = E(X) = 1/2$.

Proof. For any $\epsilon > 0$,

$$\begin{aligned} P(|X_n - X| \geq \epsilon) &= P(X_n \geq X + \epsilon) + P(X_n \leq X - \epsilon) \\ &= \int_{0}^{1} \int_{x+\epsilon}^{\infty} f(x,y) dy dx + \int_{0}^{1} \int_{-\infty}^{x-\epsilon} f(x,y) dy dx \\ &= \int_{0}^{1} \int_{x+\epsilon}^{\infty} \lambda_n e^{-\lambda_n (y-x)} dy dx + 0 \\ &= \int_{0}^{1} e^{-\lambda_n \epsilon} dx \\ &= e^{-\lambda_n \epsilon} \\ &\to 0 \end{aligned}$$

as $n \to \infty$ because $\lim_{n\to\infty} \lambda_n = \infty$. This proves that $X_n \xrightarrow{p} X$. Furthermore, it can be shown that $X_n - X$ and X are independent, $X_n - X$ has an exponential distribution with rate parameter λ_n and X has a uniform distribution $unif[0,1]$. By the property of exponential distribution, $E(X_n - X) = 1/\lambda_n$, which approaches to zero as $n \to \infty$ because $\lim_{n\to\infty} \lambda_n = \infty$. Here $E(X_n) = 1/\lambda_n + 1/2$ and $E(X) = 1/2$. □

Example 4.3.3. *For $n \in \mathbb{N}$, let the pmf of X_n be defined by*

$$P(X_n = 0) = 1 - n^{-b},$$
$$P(X_n = n^a) = n^{-b},$$

where real numbers a and b are known and satisfy $a > b > 0$. Then $X_n \xrightarrow{p} 0$ and $E(X_n)$ exists, but $\lim_{n\to\infty} E(X_n) = \infty$.

Proof. For any $\epsilon > 0$,

$$P(|X_n| \geq \epsilon) = P(X_n = n^a) = n^{-b},$$

which approaches to 0 as $n \to \infty$. This proves that $X_n \xrightarrow{p} 0$. Furthermore,

$$E(X_n) = 0\left(1 - n^{-b}\right) + n^a\left(n^{-b}\right) = n^{a-b},$$

which approaches to ∞ as $n \to \infty$ because $a > b$. □

Remark 4.3.2. *The random variables $X_1, X_2, \ldots$ in Definition 4.3.1 do not have to be independent. Furthermore, the distribution of X_n can vary as n changes. If the distribution of $X_n - X$ is known, we may use it to determine whether convergence in probability holds as we did in the previous three examples. If only available information of $X_n - X$ is its moments, we may use either Markov's or Chebyshev's inequality to show convergence in probability.*

Example 4.3.4. *For $n \in \mathbb{N}$, let X_n be a positive random variable with $E(X_n) = 1/n^a$, where real number a is positive and known. Then $X_n \xrightarrow{p} 0$.*

Proof. By Markov's inequality,

$$P(|X_n - 0| \geq \epsilon) = P(X_n \geq \epsilon) + P(X_n \leq -\epsilon) = P(X_n \geq \epsilon) \leq \frac{E(X_n)}{\epsilon} = \frac{1}{n^a \epsilon},$$

which approaches to 0 as $n \to \infty$ because $a > 0$. By the squeeze/sandwich theorem, we obtain

$$\lim_{n\to\infty} P(|X_n - 0| \geq \epsilon) = 0.$$

This proves that $X_n \xrightarrow{p} 0$. □

Example 4.3.5. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables from a population X with $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. For $n \in \mathbb{N}$, let*

$$Y_n = A_n^{-1} \sum_{i=1}^{n} a_i X_i,$$

where $A_n = \sum_{i=1}^{n} a_i$ and $(a_n)_{n\in\mathbb{N}}$ is a sequence of positive real numbers such that

$$\lim_{n\to\infty} A_n^{-2} \sum_{i=1}^{n} a_i^2 = 0.$$

Then $Y_n \xrightarrow{p} \mu$.

Proof. Since $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables from a population X with $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, we obtain

$$E(Y_n) = E\left(A_n^{-1} \sum_{i=1}^{n} a_i X_i\right) = A_n^{-1} \sum_{i=1}^{n} a_i E(X_i) = \mu A_n^{-1} \sum_{i=1}^{n} a_i = \mu,$$
$$\mathrm{Var}(Y_n) = \mathrm{Var}\left(\frac{1}{A_n} \sum_{i=1}^{n} a_i X_i\right) = \frac{1}{A_n^2} \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) = \frac{\sigma^2}{A_n^2} \sum_{i=1}^{n} a_i^2.$$

By Chebyshev's inequality,

$$P(|Y_n - \mu| \geq \epsilon) \leq \frac{1}{\epsilon^2} \mathrm{Var}(Y_n) = \frac{\sigma^2}{A_n^2 \epsilon^2} \sum_{i=1}^{n} a_i^2,$$

which approaches to 0 as $n \to \infty$ because $\lim_{n\to\infty} A_n^{-2} \sum_{i=1}^{n} a_i^2 = 0$. Thus, we have

$$\lim_{n\to\infty} P(|Y_n - \mu| \geq \epsilon) = 0,$$

which means that $Y_n \xrightarrow{p} \mu$. □

Example 4.3.6. *Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with $unif[0, 1]$ distribution. For $n \in \mathbb{N}$, define $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then $\bar{X}_n \xrightarrow{p} 1/2$.*

Proof. Although we know that $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with $unif[0, 1]$ distribution, it appears to be a challenge to find the distribution of $\bar{X}_n$. We now use Chebyshev's inequality to prove the desired result. Indeed, since $E(X_1) = 1/2$ and $\mathrm{Var}(X_1) = 1/12$, we have

$$E(\bar{X}_n) = \frac{1}{2} \quad \text{and} \quad \mathrm{Var}(\bar{X}_n) = \frac{1}{12n}.$$

By Chebyshev's inequality,

$$P(|\bar{X}_n - 1/2| \geq \epsilon) \leq \frac{1}{\epsilon^2}\text{Var}(\bar{X}_n) = \frac{1}{12n\epsilon^2},$$

which approaches to 0 as $n \to \infty$. Thus, we have

$$\lim_{n\to\infty} P(|\bar{X}_n - 1/2| \geq \epsilon) = 0.$$

which means that $\bar{X}_n \xrightarrow{p} 1/2$. □

The most famous example of convergence in probability is the weak law of large numbers (WLLN), which states that the sample mean converges in probability to the population mean. Example 4.3.6 is an application of the weak law of large numbers to a random sample from the uniform distribution $unif[0, 1]$.

Theorem 4.3.1. *(Weak law of large numbers) Let* $(X_n)_{n\in\mathbb{N}}$ *be a sequence of i.i.d. random variables with* $E(X_i) = \mu$ *and* $\text{Var}(X_i) = \sigma^2 < \infty$. *Define* $S_n = \sum_{i=1}^n X_i$. *Then* $S_n/n \xrightarrow{p} \mu$.

Proof. To prove that $S_n/n \xrightarrow{p} \mu$, we need to show that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0.$$

Indeed, a direct calculation shows that

$$E\left(\frac{S_n}{n}\right) = \mu \quad \text{and} \quad \text{Var}\left(\frac{S_n}{n}\right) = \frac{\sigma^2}{n}.$$

Applying Chebyshev's inequality, we obtain that for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}. \tag{4.9}$$

Taking the limit in (4.9) as $n \to \infty$ and using the fact that $P(|S_n/n - \mu| \geq \epsilon) \geq 0$ conclude that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0,$$

which means that $S_n/n \xrightarrow{p} \mu$. □

Remark 4.3.3. *When sample size n gets larger, the WLLN states that the distribution of the sample mean S_n/n becomes more concentrated around the population mean μ. In other words, for any $\delta > 0$ and $\epsilon > 0$, there exists an N such that for all $n \geq N$,*

$$P\left(\Big|\frac{S_n}{n} - \mu\Big| \geq \epsilon\right) < \delta,$$

which means that for any given level of accuracy ϵ and confidence δ, S_n/n is likely to be approximately equal to μ, within these levels of accuracy and confidence, provided that the sample size n is sufficiently large.

Remark 4.3.4. *The WLLN stated in Theorem 4.3.1 requires a finite variance, i.e., $\mathrm{Var}(X_i) = \sigma^2 < \infty$. In fact, this condition can be dropped. We present proof without this condition below for interested readers.*

Theorem 4.3.2. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $E(|X_1|) < \infty$. Define $S_n = \sum_{i=1}^n X_i$. Then $S_n/n \xrightarrow{p} \mu$.*

Proof. For $n \in \mathbb{N}$, let $Y_n = X_n - \mu$. Then $(Y_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with $E(Y_1) = 0$ and

$$\frac{S_n}{n} - \mu = \frac{1}{n}\sum_{i=1}^n Y_i.$$

We write $\bar{Y}_n = \frac{1}{n}\sum\limits_{i=1}^n Y_i$. Then $S_n/n \xrightarrow{p} \mu$ is equivalent to $\bar{Y}_n \xrightarrow{p} 0$.

To prove that $\bar{Y}_n \xrightarrow{p} 0$, we need to show that for any $\epsilon > 0$ and $\delta > 0$, there exists an $N \in \mathbb{N}$ such that for every $n \geq N$, we have

$$P(|\bar{Y}_n| \geq \epsilon) < \delta. \tag{4.10}$$

Since $E(|X_1|) < \infty$, $E(|Y_1|) = E(|X_1 - \mu|) \leq E(|X_1|) + |\mu| < \infty$, which allows us to choose a constant L large enough such that

$$E[|Y_1|I(|Y_1| \geq L)] < \frac{\epsilon\delta}{3}. \tag{4.11}$$

Meanwhile, we choose N large enough such that $N \geq \frac{36L^2}{\epsilon^2\delta^2}$. For $n \in \mathbb{N}$, define

$$\begin{aligned}
U_n &= Y_n I(|Y_n| < L) - E\big(Y_n I(|Y_n| < L)\big),\\
V_n &= Y_n I(|Y_n| \geq L) - E\big(Y_n I(|Y_n| \geq L)\big),\\
\bar{U}_n &= \frac{1}{n}\sum_{i=1}^n U_i,\\
\bar{V}_n &= \frac{1}{n}\sum_{i=1}^n V_i.
\end{aligned}$$

Since $(Y_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with mean zero, $(U_n)_{n\in\mathbb{N}}$ is also a sequence of i.i.d. random variables with mean zero. In addition, we have $|U_n| \leq 2L$ for any $n \in \mathbb{N}$. Thus, when $n \geq N$,

$$\begin{aligned} E(\bar{U}_n^2) &= E\left(\frac{1}{n^2}\sum_{i=1}^{n} U_i^2 + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} U_i U_j\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} E(U_i^2) + \frac{2}{n^2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E(U_i)E(U_j) \\ &= \frac{1}{n}E(U_i^2) \leq \frac{4L^2}{n} \leq \frac{4L^2}{N} \leq \frac{\epsilon^2\delta^2}{9}, \end{aligned}$$

which leads to

$$\left(E(|\bar{U}_n|)\right)^2 \leq \frac{\epsilon^2\delta^2}{9}.$$

because $E\left(\bar{U}_n^2\right) \geq \left(E(|\overline{U}_n|)\right)^2$.

Similarly, $(V_n)_{n\in\mathbb{N}}$ is a also sequence of i.i.d. random variables with mean zero because $(Y_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with mean zero. Applying Jensen's inequality to the function $f(t) = |t|$, we have

$$|E[Y_n I(|Y_n| \geq L)]| \leq E[|Y_n| I(|Y_n| \geq L)] = E[|Y_1| I(|Y_1| \geq L)] < \frac{\epsilon\delta}{3}, \quad (4.12)$$

where the last inequality of (4.12) follows from (4.11).

$$\begin{aligned} E\left(|\bar{V}_n|\right) = E\left(\left|\frac{1}{n}\sum_{i=1}^{n} V_i\right|\right) &\leq E\left(\frac{1}{n}\sum_{i=1}^{n}|V_i|\right) = E(|V_1|) \\ &\leq E|Y_1 I(|Y_1| \geq L)| + |E[Y_1 I(|Y_1| \geq L)]| \\ &\leq 2E[|Y_1| I(|Y_1| \geq L)] \\ &< \frac{2\epsilon\delta}{3}. \end{aligned} \quad (4.13)$$

Finally, applying Markov's inequality, using the fact that $\bar{Y}_n = \bar{U}_n + \bar{V}_n$ with the triangle inequality, and combining (4.12) and (4.13) yield that

$$P\left(|\bar{Y}_n| \geq \epsilon\right) \leq \frac{E(|\bar{Y}_n|)}{\epsilon} \leq \frac{E(|\bar{U}_n|) + E(|\bar{V}_n|)}{\epsilon} \leq \frac{1}{\epsilon}\left(\frac{\epsilon\delta}{3} + \frac{2\epsilon\delta}{3}\right) = \delta.$$

This proves that $\bar{Y}_n \xrightarrow{p} 0$. □

Example 4.3.7. *Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with a pmf $f_X(x) = c_p x^{-p}$ for $x \in \mathbb{N}$, where $p \in (2, 3]$ and $c_p = \left(\sum_{x=1}^{\infty} x^{-p}\right)^{-1}$ is a normalized constant. For $n \in \mathbb{N}$, let $Y_n = n^{-1}\sum_{i=1}^{n} X_i$. Then $Y_n \xrightarrow{p} c_p/c_{p-1}$.*

Proof. By the p-series, we obtain

$$E(X_1) = \sum_{x=1}^{\infty} x f_X(x) = c_p \sum_{x=1}^{\infty} x^{-(p-1)} = \frac{c_p}{c_{p-1}}$$

because $p \in (2, 3]$. By Theorem 4.3.2, we conclude that

$$Y_n \xrightarrow{p} \frac{c_p}{c_{p-1}}.$$

However,

$$E(X_1^2) = \sum_{x=1}^{\infty} x^2 f_X(x) = c_p \sum_{x=1}^{\infty} x^{-(p-2)}$$

diverges because $p - 2 \in (0, 1]$. Thus, $\mathrm{Var}(X_1)$ does not exist. □

Theorem 4.3.3. *Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of random variables and c is a constant such that*

$$\lim_{n\to\infty} E(X_n) = c \quad \textit{and} \quad \lim_{n\to\infty} \mathrm{Var}(X_n) = 0.$$

Then $X_n \xrightarrow{p} c$.

Proof. For $t \geq 0$, let $f(t) = \sqrt{t}$. Since $f(t)$ is concave, applying Jensen's inequality to f yields

$$\begin{aligned} E|X_n - E(X_n)| &= E\left(\sqrt{\left[X_n - E(X_n)\right]^2}\right) \\ &\leq \sqrt{E\left(\left[X_n - E(X_n)\right]^2\right)} \\ &= \sqrt{\mathrm{Var}(X_n)}. \end{aligned} \tag{4.14}$$

For any $\epsilon > 0$, using Markov's inequality, the triangle inequality, and (4.14) sequentially, we obtain

$$\begin{aligned} P\left(|X_n - c\,| \geq \epsilon\right) &\leq \frac{1}{\epsilon} E(|X_n - c\,|) \\ &= \frac{1}{\epsilon} E\Big(\Big|X_n - E(X_n) + E(X_n) - c\Big|\Big) \\ &\leq \frac{1}{\epsilon}\Big(E\big(|X_n - E(X_n)|\big) + |E(X_n) - c\,|\Big) \\ &\leq \frac{1}{\epsilon}\Big(\sqrt{\mathrm{Var}(X_n)} + |E(X_n) - c\,|\Big). \end{aligned} \tag{4.15}$$

Applying (4.14) with the fact that $P\left(|X_n - c| \geq \epsilon\right) \geq 0$ to (4.15) concludes that $\lim_{n\to\infty} P\left(|X_n - c| \geq \epsilon\right) = 0$ because $\lim_{n\to\infty} \mathrm{Var}(X_n) = 0$ and

$$\lim_{n\to\infty} |E(X_n) - c\,| = |\lim_{n\to\infty} E(X_n) - c\,| = 0.$$

This proves that $X_n \xrightarrow{p} c$. □

Example 4.3.8. *For $n \in \mathbb{N}$, let the random variable X_n have a beta$(n, 10)$ distribution. Then $X_n \xrightarrow{p} 1$.*

Proof. If $X \sim beta(a, b)$, then

$$E(X) = \frac{a}{a+b} \quad \text{and} \quad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Applying this fact with $a = n$ and $b = 10$, we obtain

$$\lim_{n\to\infty} E(X_n) = \lim_{n\to\infty} \frac{n}{n+10} = 1,$$
$$\lim_{n\to\infty} \text{Var}(X_n) = \lim_{n\to\infty} \frac{10n}{(10+n)^2(n+11)} = 0.$$

By Theorem 4.3.3, we can conclude that $X_n \xrightarrow{p} 1$. □

Theorem 4.3.4. *If $X_n \xrightarrow{p} X$ and h is a continuous function defined on $\mathbb{R}$, then $h(X_n) \xrightarrow{p} h(X)$.*

Proof. Since $h(x)$ is continuous on $\mathbb{R}$, we have that for any $x_0 \in \mathbb{R}$ and $\epsilon > 0$, there exists a $\delta > 0$ such that

$$|h(x) - h(x_0)| < \epsilon$$

whenever $|x - x_0| < \delta$. Furthermore, when n is large enough, we have

$$P(|X_n - X| < \delta) > 1 - \epsilon$$

because $X_n \xrightarrow{p} X$. Thus, when n is large enough, we have

$$\begin{aligned} P(|h(X_n) - h(X)| < \epsilon) &= P(|h(X_n) - h(X)| < \epsilon, |X_n - X| < \delta) \\ &\quad + P(|h(X_n) - h(X)| < \epsilon, |X_n - X| \geq \delta) \\ &\geq P(|h(X_n) - h(X)| < \epsilon, |X_n - X| < \delta) \\ &= P(|X_n - X| < \delta) \\ &> 1 - \epsilon, \end{aligned}$$

which leads to

$$\lim_{n\to\infty} P(|h(X_n) - h(X)| < \epsilon) = 1,$$

or equivalently,

$$\lim_{n\to\infty} P(|h(X_n) - h(X)| \geq \epsilon) = 0.$$ □

Theorem 4.3.5. *If* $X_n \xrightarrow{p} X$, *then* $X_n \xrightarrow{d} X$.

Proof. Let F_{X_n} and F_X denote the cdfs of X_n and X, respectively. Then for any $\epsilon > 0$ and $x \in \mathbb{R}$, we have

$$\begin{aligned} F_{X_n}(x) &= P(X_n \leq x) \\ &= P(X_n \leq x, X < x + \epsilon) + P(X_n \leq x, X \geq x + \epsilon) \\ &\leq P(X < x + \epsilon) + P(X_n - X \leq x - X, x - X \leq -\epsilon) \\ &\leq P(X < x + \epsilon) + P(X_n - X \leq -\epsilon) \\ &\leq P(X < x + \epsilon) + P(X_n - X \leq -\epsilon) + P(X_n - X \geq \epsilon) \\ &= P(X < x + \epsilon) + P(|X_n - X| \geq \epsilon). \end{aligned} \tag{4.16}$$

Similarly, we have

$$F_X(x - \epsilon) \leq P(X_n < x) + P(|X_n - X| \geq \epsilon). \tag{4.17}$$

Combining (4.16) and (4.17), we have

$$\begin{aligned} F_X(x - \epsilon) - P(|X_n - X| \geq \epsilon) &\leq P(X_n < \epsilon) \\ &\leq F_{X_n}(x) \\ &\leq P(X < x + \epsilon) + P(|X_n - X| \geq \epsilon). \end{aligned} \tag{4.18}$$

Taking the limit as $n \to \infty$ in (4.18) will yield

$$F_X(x - \epsilon) \leq \lim_{n \to \infty} F_{X_n}(x) \leq P(X < x + \epsilon), \tag{4.19}$$

which is true for any $\epsilon > 0$ and $x \in \mathbb{R}$. Thus, when x is a continuity point of F_X, letting $\epsilon \to 0+$ in (4.19) and applying the squeeze theorem will obtain

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x).$$

That means, $X_n \xrightarrow{d} X$. □

Remark 4.3.5. *The converse of Theorem 4.3.5 is not true. That is,* $X_n \xrightarrow{d} X$ *does not imply that* $X_n \xrightarrow{p} X$ *in general.*

Example 4.3.9. *For* $n \in \mathbb{N}$, *let* $X_n = (-1)^n X$, *where the random variable* X *is symmetric about zero with* $P(|X| \geq 1) = 1$. *Then* $X_n \xrightarrow{d} X$, *but* X_n *does not converge in probability to* X.

Proof. The result that $X_n \xrightarrow{d} X$ is true because X is symmetric about zero

and $X_n \stackrel{d}{=} X$. For any $\epsilon \in (0,1)$, however,

$$\begin{aligned} P(|X_n - X| \geq \epsilon) &= P(|(-1)^n X - X| \geq \epsilon) \\ &= P(|(-1)^n - 1||X| \geq \epsilon) \\ &= \begin{cases} P(|X| \geq \epsilon/2), & \text{if } n \text{ is odd,} \\ 0, & \text{if } n \text{ is even,} \end{cases} \\ &= \begin{cases} 1, & \text{if } n \text{ is odd,} \\ 0, & \text{if } n \text{ is even,} \end{cases} \end{aligned}$$

which does not converge. □

Theorem 4.3.6. *Let c be a constant. Then $X_n \xrightarrow{p} c$ if and only if $X_n \xrightarrow{d} c$.*

Proof. Since the constant c can be considered a special random variable whose cdf is given by

$$F_c(x) = I(x \geq c), \quad x \in \mathbb{R},$$

applying Theorem 4.3.5 concludes that $X_n \xrightarrow{p} c$ implies $X_n \xrightarrow{d} c$. Now we prove the converse, i.e., $X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{p} c$. Indeed, for any $\epsilon > 0$,

$$\begin{aligned} P(|X_n - c| < \epsilon) &= P(c - \epsilon < X_n < c + \epsilon) \\ &= P(X_n < c + \epsilon) - P(X_n \leq c - \epsilon) \\ &= P(X_n < c + \epsilon) - F_{X_n}(c - \epsilon) \\ &= P\left(\bigcup_{m=1}^{\infty} \{X_n \leq c + \epsilon - 1/m\} \right) - F_{X_n}(c - \epsilon) \\ &= \lim_{m \to \infty} P(X_n \leq c + \epsilon - 1/m) - F_{X_n}(c - \epsilon) \\ &= \lim_{m \to \infty} F_{X_n}(c + \epsilon - 1/m) - F_{X_n}(c - \epsilon). \end{aligned} \tag{4.20}$$

Since c is the only discontinuity point of F_c, using the condition that $X_n \xrightarrow{d} c$ will obtain

$$\begin{aligned} \lim_{n \to \infty} F_{X_n}(c + \epsilon - 1/m) &= F_c(c + \epsilon - 1/m), \\ \lim_{n \to \infty} F_{X_n}(c - \epsilon) &= F_c(c - \epsilon), \end{aligned} \tag{4.21}$$

provided that $m > 1/\epsilon$. Taking the limit in (4.20) as $n \to \infty$ and using (4.21), we have

$$\lim_{n \to \infty} P(|X_n - c| < \epsilon) = \lim_{m \to \infty} F_c(c + \epsilon - 1/m) - F_c(c - \epsilon) = 1 - 0 = 1,$$

which is equivalent to $\lim_{n \to \infty} P(|X_n - c| \geq \epsilon) = 0$. □

4.4 Convergence in rth mean

Definition 4.4.1. *Consider a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ with finite rth moment and a random variable X with finite rth moment defined on the same probability space. We say that X_n converges in rth mean to X if*

$$\lim_{n\to\infty} E(|X_n - X|^r) = 0.$$

We write $X_n \xrightarrow{rm} X$. In particular, when $r = 2$, the convergence, which is a widely used one, goes by the special name of convergence in quadratic mean and is usually written $X_n \xrightarrow{qm} X$.

Theorem 4.4.1. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with finite variance and let $\mu = E(X_1)$. Then $\bar{X}_n \xrightarrow{qm} \mu$.*

Proof. Since $E(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_1)/n$,

$$\lim_{n\to\infty} E(|\bar{X}_n - \mu|^2) = \lim_{n\to\infty} \mathrm{Var}(\bar{X}_n) = \lim_{n\to\infty} \frac{\mathrm{Var}(X_1)}{n} = 0.$$

This proves that $\bar{X}_n \xrightarrow{qm} \mu$. □

Theorem 4.4.2. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables with finite rth moment and let X be a random variable X with finite rth moment defined on the same probability space. Then $X_n \xrightarrow{p} X$ if $X_n \xrightarrow{rm} X$.*

Proof. For any $\epsilon > 0$, applying Markov's inequality to the random variable $|X_n - X|^r$ yields that

$$P(|X_n - X| \geq \epsilon) = P(|X_n - X|^r \geq \epsilon^r) \leq \frac{1}{\epsilon^r} E(|X_n - X|^r),$$

which approaches to zero because $\lim_{n\to\infty} E(|X_n - X|^r) = 0$. Applying the squeeze/sandwich theorem and using the fact that $P(|X_n - X| \geq \epsilon) \geq 0$ yield that $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$. This proves that $X_n \xrightarrow{p} X$. □

Remark 4.4.1. *The converse of Theorem 4.4.2 is not true, i.e., $X_n \xrightarrow{p} X$ does not imply $X_n \xrightarrow{rm} X$ in general.*

Example 4.4.1. *Let $P(X_n = n) = n^{-a}$ and $P(X_n = 0) = 1 - n^{-a}$, where $0 < a < r$. Then $X_n \xrightarrow{p} 0$, but $\lim_{n\to\infty} E(X_n) = \infty$.*

Proof. For any $\epsilon > 0$,

$$P(|X_n| \geq \epsilon) = P(X_n = n) = n^{-a},$$

which approaches to 0 as $n \to \infty$ because $a > 0$. This proves that $X_n \xrightarrow{p} 0$. On the other hand,

$$E(X_n^r) = 0^r(1 - n^{-a}) + n^r(n^{-a}) = n^{r-a},$$

which approaches to ∞ as $n \to \infty$ because $a < r$. Furthermore, $E(X^r) = 0$ because $P(X = 0) = 1$. Thus, $E(X_n^r) \nrightarrow E(X^r)$ in general. □

4.5 Almost sure convergence

Definition 4.5.1. *Consider a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables and a random variable X defined on the same probability space. We say that X_n converges almost surely (a.s.) to X and write $X_n \xrightarrow{a.s.} X$ if*

$$P\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1.$$

Example 4.5.1. *For $n \in \mathbb{N}$, let $X_n = \frac{1}{n}X$, where X is a random variable. Then X_n converges almost surely to zero.*

Remark 4.5.1. *It is generally difficult to verify almost sure convergence. Some theorems in the next subsection will be helpful to verify almost sure convergence.*

Lemma 4.5.1. *(Borel-Cantelli Lemma)*

(a) *If $(E_n)_{n\in\mathbb{N}}$ be an infinite sequence of events such that $\sum_{n=1}^{\infty} P(E_n) < \infty$, then $P\big(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\big) = 0$.*

(b) *If $(E_n)_{n\in\mathbb{N}}$ is a sequence of independent events such that $\sum_{n=1}^{\infty} P(E_n) = \infty$, then $P\big(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\big) = 1$.*

Proof. (a) For $n \in \mathbb{N}$, let $F_n = \bigcup_{i=n}^{\infty} E_i$. Then $(F_n)_{n\in\mathbb{N}}$ is a nonincreasing sequence of events and thus

$$\lim_{n\to\infty} F_n = \bigcap_{n=1}^{\infty} F_n = \bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i.$$

By the continuity of probability, we have

$$\begin{aligned} P\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\right) &= P(\lim_{n\to\infty} F_n) = \lim_{n\to\infty} P(F_n) \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=n}^{\infty} E_i\right) \le \lim_{n\to\infty}\sum_{i=n}^{\infty} P(E_i) \\ &= 0. \end{aligned} \tag{4.22}$$

Here the inequality in (4.22) follows from Boole's inequality, while the last equality in (4.22) is based on the condition $\sum_{n=1}^{\infty} P(E_n) < \infty$.

(b) By DeMorgan's rule, $\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\right)^c = \bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} E_i^c$. Using the continuity of probability, we obtain

$$P\left(\bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} E_i^c\right) = P\left(\lim_{n\to\infty}\bigcap_{i=n}^{\infty} E_i^c\right) = \lim_{n\to\infty} P\left(\bigcap_{i=n}^{\infty} E_i^c\right). \tag{4.23}$$

When $m \ge n$, $\bigcap_{i=n}^{\infty} E_i^c \subseteq \bigcap_{i=n}^{m} E_i^c$. By the monotonicity of probability, Property 1.2.3, we have

$$P\left(\bigcap_{k=n}^{\infty} E_i^c\right) \le P\left(\bigcap_{i=n}^{m} E_i^c\right)$$

for $m \ge n$, which implies that

$$P\left(\bigcap_{k=n}^{\infty} E_i^c\right) \le \lim_{m\to\infty} P\left(\bigcap_{i=n}^{m} E_i^c\right) = \lim_{m\to\infty}\prod_{i=n}^{m}\big(1 - P(E_i)\big),$$

because $(E_n)_{n\in\mathbb{N}}$ is a sequence of independent events. Applying well-known inequality $1 - t \le e^{-t}$ for $t \ge 0$, we conclude that

$$P\left(\bigcap_{i=n}^{\infty} E_i^c\right) \le \lim_{m\to\infty}\exp\left(-\sum_{i=n}^{m} P(E_i)\right) = \exp\left(-\sum_{i=n}^{\infty} P(E_i)\right) = 0 \tag{4.24}$$

because $\sum_{n=1}^{\infty} P(E_n)$ diverges. Since the probability is always nonnegative, we obtain from (4.24) that $P\left(\bigcap_{i=n}^{\infty} E_i^c\right) = 0$ for any $n \in \mathbb{N}$. Based on (4.23), we can claim that $P\left(\bigcup_{n=1}^{\infty}\bigcap_{i=n}^{\infty} E_i^c\right) = 0$, which is equivalent to $P\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\right) = 1$. □

Corollary 4.5.1. *Let $(E_n)_{n\in\mathbb{N}}$ be a sequence of independent events. Then $P\left(\bigcap_{n=1}^{\infty}\bigcup_{i=n}^{\infty} E_i\right)$ is either 0 or 1.*

Theorem 4.5.1. *(Strong law of large numbers) Suppose that $(X_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. random variables with a finite $E(X_i) = \mu$ and finite fourth moment. For $n \in \mathbb{N}$, let $S_n = \sum_{i=1}^n X_i$. Then*

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1.$$

Proof. For $n \in \mathbb{N}$, let $X_n^+ = \max(0, X_n)$ and $X_n^- = \max(0, -X_n)$. Then $X_n = X_n^+ - X_n^-$. We, without loss of generality, assume that $X_n \geq 0$ for $n \in \mathbb{N}$. For $n \in \mathbb{N}$, let $Y_n = X_n I(X_n \leq n)$ and $S_n^* = \sum_{i=1}^n Y_i$. For any $\epsilon > 0$, let $k_n = [a^n]$, where $a > 1$. By Chebyshev's inequality, we have

$$\begin{aligned}
\sum_{n=1}^{\infty} P\left(\left|\frac{S_{k_n}^* - E(S_{k_n}^*)}{k_n}\right| \geq \epsilon\right) &\leq \epsilon^{-2} \sum_{n=1}^{\infty} \frac{\mathrm{Var}(S_{k_n}^*)}{k_n^2} \\
&= \epsilon^{-2} \sum_{n=1}^{\infty} \frac{1}{k_n^2} \sum_{i=1}^{k_n} \mathrm{Var}(Y_i) \\
&\leq c_1 \sum_{n=1}^{\infty} \frac{E(Y_n^2)}{n^2} \\
&\leq c_1 \sum_{n=1}^{\infty} \frac{1}{n^2} \int_0^n x^2 dF_X(x) \\
&\leq c_1 \sum_{n=1}^{\infty} \frac{1}{n^2} \sum_{k=0}^{n-1} \int_k^{k+1} x^2 dF_X(x) \\
&\leq c_2 \sum_{k=0}^{\infty} \frac{1}{k=1} \int_k^{k+1} x^2 dF_X(x) \\
&\leq c_3 \sum_{k=0}^{\infty} \int_k^{k=1} x dF_X(x) \\
&= c_3 E(X_1) \\
&< \infty,
\end{aligned}$$

where F_X is the cdf of X_1 and c_1, c_2,and c_3 are constants. Furthermore,

$$E(X_1) = \lim_{n\to\infty} \int_0^n x dF_X(x) = \lim_{n\to\infty} E(Y_n) = \lim_{n\to\infty} \frac{E(S_{k_n}^*)}{k_n}.$$

Thus by the Borel-Cantelli Lemma,

$$\lim_{n\to\infty} \frac{S_{k_n}^*}{k_n} = E(X_1)$$

almost surely. In addition,

$$\begin{aligned}\sum_{n=1}^{\infty} P(Y_n \neq X_n) &= \sum_{n=1}^{\infty} P(X_n > n) = \sum_{n=1}^{\infty}\sum_{i=n}^{\infty}\int_i^{i+1} dF_X(x) \\ &= \sum_{i=1}^{\infty} i\int_i^{i+1} dF_X(x) \leq \sum_{i=1}^{\infty}\int_i^{i+1} x dF_X(x) \\ &\leq E(X_1) \\ &< \infty.\end{aligned}$$

Hence by Borel-Cantelli Lemma, $X_n \neq Y_n$ only finitely many times. Consequently,

$$\lim_{n\to\infty} \frac{S_{k_n}}{k_n} = E(X_1)$$

almost surely. Now we can conclude from the monotonicity of S_n that

$$\frac{1}{a}E(X_1) \leq \liminf_{n\to\infty} \frac{S_n}{n} \leq \limsup_{n\to\infty} \frac{S_n}{n} \leq aE(X_1)$$

for all $a > 1$, which gives us the desired result. □

Example 4.5.2. *Suppose that a fair coin is tossed n times independently. Let S_n represent the number of heads. Then SLLN states that*

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = \frac{1}{2}\right) = 1,$$

while WWLN says that

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \frac{1}{2}\right| \geq \epsilon\right) = 0.$$

Example 4.5.3. *For $n \in \mathbb{N}$, let X_n have a normal distribution $N(0, 1/n)$. Then X_n converges to zero almost surely.*

Solution. *Note that X_n can be considered $\sum_{i=1}^{n} Z_i/n$, where $(Z_i)_{i\in\mathbb{N}}$ is a sequence of i.i.d. $N(0,1)$ random variables. By SLLN, X_n converges to $E(Z_1) = 0$ almost surely.*

Theorem 4.5.2. *Consider a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables and a random variable X defined on the same probability space. If for every $\epsilon > 0$,*

$$\sum_{n=1}^{\infty} P(|X_n - X| \geq \epsilon) < \infty, \tag{4.25}$$

then X_n converges to X almost surely.

Theorem 4.5.3. *If* $X_n \xrightarrow{a.s.} X$, *then* $X_n \xrightarrow{p} X$.

Proof. For any $\epsilon > 0$, we define

$$E_n = \bigcap_{i=n}^{\infty} \{\omega \in \Omega : |X_i(\omega) - X(\omega)| < \epsilon\}, \quad n = 1, 2, \cdots,$$
$$E = \left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}.$$

Clearly, $(E_n)_{n\in\mathbb{N}}$ is a sequence of nondecreasing events, which implies

$$\lim_{n\to\infty} E_n = \bigcup_{n=1}^{\infty} E_n.$$

If $\omega \in E$, then there exists an N such that whenever $n \geq N$,

$$|X_n(\omega) - X(\omega)| < \epsilon.$$

Thus, $\omega \in E_N$, which implies that $E \subseteq \bigcup_{n=1}^{\infty} E_n$. Since $X_n \xrightarrow{a.s.} X$, $P(E) = 1$. Using the monotonicity and continuity of probability, we obtain

$$1 = P(E) \leq P\left(\bigcup_{n=1}^{\infty} E_n\right) = \lim_{n\to\infty} P(E_n) \leq 1,$$

which means that $\lim_{n\to\infty} P(E_n) = 1$. Since $E_n \subseteq \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \epsilon\}$, applying monotonicity of the probability yields that

$$P(E_n) \leq P(|X_n - X| < \epsilon) \leq 1. \tag{4.26}$$

Taking the limit in (4.26) as $n \to \infty$ and applying the squeeze theorem in calculus conclude that

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

which is equivalent to

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0.$$

This proves that $X_n \xrightarrow{p} X$. □

Remark 4.5.2. *The converse of Theorem 4.5.3 is not true in general.*

Remark 4.5.3. *Convergence in rth mean does not imply convergence almost surely.*

Example 4.5.4. *For $n \in \mathbb{N}$, consider the random variable X_n with pmf $P(X_n = 0) = 1 - n^{-p}$ and $P(X_n = 1) = n^{-p}$, where $p \in (0, 1)$. In addition, let $P(X = 0) = 1$. Then*

$$E(|X_n - X|^r) = E(X_n^r) = n^{-p} \longrightarrow 0$$

as $n \to \infty$. This shows that $X_n \xrightarrow{rm} X$. However, it does not converge almost surely to X because for any $\epsilon \in (0, 1)$,

$$\sum_{n=1}^{\infty} P(|X_n - X| \geq \epsilon) = \sum_{n=1}^{\infty} \frac{1}{n^p} = \infty$$

and by Borel-Cantelli Lemma,

$$P(|X_n - X| \geq \epsilon \text{ infinitely often}) = 1.$$

Remark 4.5.4. *Almost sure convergence does not imply convergence in rth mean.*

Example 4.5.5. *For $n \in \mathbb{N}$, consider the random variable X_n with pmf $P(X_n = 0) = 1 - 1/n^p$ and $P(X_n = n^{p+1}) = 1/n^p$, where $p \in [2, \infty)$. Let $P(X = 0) = 1$. Then*

$$E(|X_n - X|^r) = E(X_n^r) = n^{r(p+1)-p} \longrightarrow \infty$$

as $n \to \infty$ when $r > p/(p + 1)$. On the other hand, for any $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P(|X_n - X| \geq \epsilon) = \sum_{n=1}^{\infty} \frac{1}{n^p} < \infty$$

and by Borel-Cantelli Lemma,

$$P(|X_n - X| \geq \epsilon \text{ infinitely often}) = 0.$$

This shows that $X_n \xrightarrow{a.s.} X$.

5

Estimation

5.1 Introduction

- **Distributions:** *binomial distribution* $b(n,p)$, *geometric distribution* $G(p)$, *Poisson distribution* $\mathcal{P}(\lambda)$, *Gamma distribution* $\Gamma(\alpha,\beta)$, *Normal distribution* $N(\mu,\sigma^2)$, *etc.*
- **One sample data:** $X_1, X_2, \ldots, X_n$ *(i.i.d.)*
- **Questions related to one sample data:**
  (1) *You may know from the history that* $X_1, X_2, \ldots, X_n$ *follow a normal distribution* $N(\mu,\sigma^2)$. *What is the mean* μ? *What is the variance* σ^2? *(estimation problem)*
  (2) *If you claim that data* $X_1, X_2, \ldots, X_n$ *follow the normal distribution* $N(0,1)$, *is your claim true? (testing hypothesis problem)*
- **Two sample data:** $\{X_1, X_2, \ldots, X_n\}$ *are i.i.d. from* $N(\mu_1,\sigma^2)$ *and* $\{Y_1, Y_2, \ldots, Y_n\}$ *are i.i.d. from* $N(\mu_2,\sigma^2)$
- **Questions related to two sample data:**
  (1) *How can we estimate* μ_1, μ_2 *and* σ^2? *(estimation problem)*
  (2) *Can we claim that* $\mu_1 = \mu_2$? *(testing hypothesis problem)*

5.2 Random sample and its order statistics

Definition 5.2.1. *If the random variables* $X_1, X_2, \ldots, X_n$ *are independent and identically distributed, i.e., each* X_i *has the same distribution, then we say that these random variables constitute a random (or an i.i.d.) sample of size n from that common distribution.*

Definition 5.2.2. *Given a random sample $X_1, X_2, \ldots, X_n$, any function $T = T(X_1, X_2, \ldots, X_n)$ of the sample is called a statistic.*

Theorem 5.2.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with $E(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Define*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{(sample mean)},$$
$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \textit{(sample variance)}.$$

Then

(i) $E(\bar{X}) = \mu$ *and* $\mathrm{Var}(\bar{X}) = \sigma^2/n$, *i.e.,* $\bar{X}$ *is an unbiased estimator of* μ,

(ii) $E(S^2) = \sigma^2$, *i.e.,* S^2 *is an unbiased estimator of* σ^2,

(iii) $\mathrm{Cov}(X_i - \bar{X}, \bar{X}) = 0, \quad i = 1, 2, \cdots, n.$

Proof. (i) Direct calculations show that

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right) = \frac{1}{n}E(X_1) + \frac{1}{n}E(X_2) + \cdots + \frac{1}{n}E(X_n) \\ &= \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu, \\ \mathrm{Var}(\bar{X}) &= \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

(ii) Note that

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Thus,

$$\begin{aligned} E[(n-1)S^2] &= E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = E\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - nE[(\bar{X} - \mu)^2] \\ &= \sum_{i=1}^{n}\mathrm{Var}(X_i) - n\mathrm{Var}(\bar{X}) = n\sigma^2 - n(\sigma^2/n) = (n-1)\sigma^2. \end{aligned}$$

Dividing by $n-1$ on both sides concludes that $E(S^2) = \sigma^2$.

(iii) For $i = 1, 2 \ldots, n$,

$$\begin{aligned}
\mathrm{Cov}(X_i - \bar{X}, \bar{X}) &= \mathrm{Cov}(X_i, \bar{X}) - \mathrm{Cov}(\bar{X}, \bar{X}) \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) - \frac{\sigma^2}{n} \\
&= \frac{1}{n}\mathrm{Cov}(X_i, X_i) - \frac{\sigma^2}{n} \\
&= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} \\
&= 0.
\end{aligned}$$

□

Theorem 5.2.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$. Then $\bar{X}$ and $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ are independent.*

Proof. Write $\bar{t} = \sum_{i=1}^{n} t_i/n$. We compute the joint mgf of $\bar{X}$ and $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ below:

$$\begin{aligned}
& M_{\bar{X}, X_1 - \bar{X}, \ldots, X_n - \bar{X}}(t, t_1, \ldots, t_n) \\
&= E \exp\{t\bar{X} + \sum_{i=1}^{n} t_i (X_i - \bar{X})\} \\
&= E \exp\left\{\sum_{i=1}^{n} t_i X_i - \left(\sum_{i=1}^{n} t_i - t\right)\bar{X}\right\} \\
&= E \exp\left\{\sum_{i=1}^{n} X_i \left(t_i - \frac{n\bar{t} - t}{n}\right)\right\} \\
&= \prod_{i=1}^{n} E \exp\left\{\frac{X_i[t + n(t_i - \bar{t})]}{n}\right\} \\
&= \prod_{1}^{n} \exp\left\{\frac{\mu[t + n(t_i - \bar{t})]}{n} + \frac{\sigma^2}{2}\frac{1}{n^2}[t + n(t_i - \bar{t})]^2\right\} \\
&= \exp(\mu t) \exp\left\{\frac{\sigma^2}{2n^2}\left(nt^2 + n^2 \sum_{i=1}^{n} (t_i - \bar{t})^2\right)\right\} \\
&= \exp\left(\mu t + \frac{\sigma^2/n}{2} t^2\right) \exp\left(\frac{\sigma^2}{2} \sum_{i=1}^{n} (t_i - \bar{t})^2\right) \\
&= M_{(\bar{X}, X_1 - \bar{X}, \ldots, X_n - \bar{X})}(t, 0, \ldots, 0) M_{(\bar{X}, X_1 - \bar{X}, \ldots, X_n - \bar{X})}(0, t_1, \ldots, t_n) \\
&= M_{\bar{X}}(t) M_{X_1 - \bar{x}, \cdots, X_n - \bar{X}}(t_1, t_2, \cdots, t_n),
\end{aligned}$$

which means that $\bar{X}$ and $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ are independent. □

Corollary 5.2.1. $\bar{X} \sim N(\mu, \sigma^2/n)$.

Corollary 5.2.2. *$\bar{X}$ and S^2 are independent.*

Corollary 5.2.3. $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Proof. Let

$$\begin{aligned} Z_i &= (X_i - \mu)/\sigma, \quad i = 1, 2, \ldots, n, \\ W_1 &= \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2, \\ W_2 &= \frac{(n-1)S^2}{\sigma^2}. \end{aligned}$$

Then $Z_1, Z_2, \ldots, Z_n$ are i.i.d. from $N(0,1)$, which implies that $Z_1^2, Z_2^2, \ldots, Z_n^2$ are i.i.d. from χ^2_1. We also see from Corollary 5.2.2 that W_1 and W_2 are independent. Furthermore, we have

$$\sum_{i=1}^n Z_i^2 = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2 + \frac{(n-1)S^2}{\sigma^2} = W_1 + W_2. \qquad (5.1)$$

Since the mgf of χ^2_k is $(1-2t)^{-k/2}$, calculating the mgf for both sides of (5.1) and using the fact that $W_1 \sim \chi^2_1$ yield that

$$(1-2t)^{-n/2} = (1-2t)^{-1/2} M_{W_2}(t).$$

Thus,

$$M_{W_2}(t) = (1-2t)^{-(n-1)/2}$$

which is the mgf of χ^2_{n-1}-distribution. Thus, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. □

Definition 5.2.3. *Given a random sample $X_1, X_2, \ldots, X_n$, we order $X_1, X_2, \ldots, X_n$ from the smallest to the largest below*

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}.$$

Then $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are called order statistics of $X_1, X_2, \ldots, X_n$ and $X_{(i)}$ is called the ith order statistic of the random sample $X_1, X_2, \ldots, X_n$.

Theorem 5.2.3. *For $i = 1, 2, \ldots, n$, let $Y_i = X_{(i)}$ denote the order statistics of the random sample $X_1, X_2, \ldots, X_n$ from a distribution with the pdf $f(x)$.*

(a) *The joint pdf of $(Y_1, Y_2, \ldots, Y_n)$ is given by*

$$g(y_1, \cdots, y_n) = \begin{cases} n! \prod_{i=1}^n f(y_i), & \text{if } y_1 < y_2 < \cdots < y_n, \\ 0, & \text{otherwise.} \end{cases}$$

(b) *The joint pdf of* (Y_i, Y_j) $(i < j)$ *is given by*

$$g(y_i, y_j) = \frac{n! f(y_i) f(y_j) [F(y_i)]^{i-1} [F(y_j) - F(y_i)]^{j-i-1} [1 - F(y_j)]^{n-j}}{(i-1)!(j-i-1)!(n-j)!}$$

if $y_i < y_j$, *zero otherwise.*

(c) *The pdf of* Y_i *for* $i = 1, 2, \ldots, n$ *is given by*

$$g(y_i) = \frac{n!}{(i-1)!(n-i)!} [F(y_i)]^{i-1} [1 - F(y_i)]^{n-i} f(y_i), \quad y_i \in \mathbb{R}.$$

Proof. (a) The joint pdf of $(X_1, X_2, \ldots, X_n)$ is given by

$$h(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

For $i = 1, 2, \ldots, n$, let $Y_i = X_{(i)}$. Then there are $n!$ possible transformations and each of them has a Jacobian of 1. Thus,

$$g(y_1, y_2, \ldots, y_n) = n! \prod_{i=1}^{n} f(y_i), \quad y_1 < y_2 < \cdots < y_n.$$

(b) For $y_i < y_j$ with $i < j$, we have

$$\begin{aligned}
g(y_i, y_j) &= \int \cdots \int g(y_1, y_2, \ldots, y_n) \prod_{k \neq i,j}^{i-1} dy_k \\
&= \int_{-\infty}^{y_i} \cdots \int_{-\infty}^{y_2} \int_{y_i}^{y_j} \cdots \int_{y_{j-2}}^{y_j} \int_{y_j}^{\infty} \cdots \int_{y_{n-1}}^{\infty} n! \prod_{l=1}^{n} f(y_l) \prod_{k \neq i,j}^{i-1} dy_k \\
&= \frac{n! f(y_i) f(y_j)}{(i-1)!} [F(y_i)]^{i-1} \int_{y_i}^{y_j} \cdots \int_{y_{j-2}}^{y_j} \int_{y_j}^{\infty} \cdots \int_{y_{n-1}}^{\infty} \prod_{k=i+1, k \neq j}^{n} f(y_k) dy_k \\
&= \frac{n! f(y_i) f(y_j) [F(y_i)]^{i-1} [F(y_j) - F(y_i)]^{j-i-1}}{(i-1)!(j-i-1)!} \int_{y_j}^{\infty} \cdots \int_{y_{n-1}}^{\infty} \prod_{k=j+1}^{n} f(y_k) dy_k \\
&= \frac{n! f(y_i) f(y_j)}{(i-1)!(j-i-1)!(n-j)!} [F(y_i)]^{i-1} [F(y_j) - F(y_i)]^{j-i-1} [1 - F(y_j)]^{n-j}.
\end{aligned}$$

(c) For $i = 1, 2, \ldots, n$, we have from (ii) above that

$$\begin{aligned}
g(y_i) &= \int_{y_i}^{\infty} g(y_i, y_j) dy_j \\
&= \frac{n![F(y_i)]^{i-1} f(y_i)}{(i-1)!(j-i-1)!(n-j)!} \\
&\quad \times \int_{y_i}^{\infty} [F(y_j) - F(y_i)]^{j-i-1} [1 - F(y_j)]^{n-j} dF(y_j) \\
&= \frac{n![F(y_i)]^{i-1} f(y_i)}{(i-1)!(j-i-1)!(n-j)!} \int_{z_i}^{1} (x - z_i)^{j-i-1} (1-x)^{n-j} dx \\
&\qquad \Big(x = F(y_j), \;\; z_i = F(y_i) \Big) \\
&\overset{y=\frac{x-z_i}{1-z_i}}{=} \frac{n![F(y_i)]^{i-1} f(y_i)(1-z_i)^{n-i}}{(i-1)!(j-i-1)!(n-j)!} \int_0^1 y^{j-i-1} (1-y)^{n-j} dy \\
&= \frac{n![F(y_i)]^{i-1} f(y_i)(1-z_i)^{n-i}}{(i-1)!(j-i-1)!(n-j)!} \frac{\Gamma(j-i)\Gamma(n-j+1)}{\Gamma(j-1+n-j+1)} \\
&= \frac{n!}{(i-1)!(n-i)!} [F(y_i)]^{i-1} [1 - F(y_i)]^{n-i} f(y_i)
\end{aligned}$$

for $y_i \in \mathbb{R}$. □

Example 5.2.1. *(Example 4.4.1 – special case of Theorem with $n = 3$)*

Example 5.2.2. *(Example 4.4.2) Suppose $Y_1 < Y_2 < Y_3 < Y_4$ are order statistics of the random sample from a distribution having pdf*

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

Find $P(Y_3 > \frac{1}{2})$.

Solution. *Direction calculations show that*

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

Thus, the pdf of Y_3 is given by

$$\begin{aligned}
g(y_3) &= \frac{4!}{(3-1)!(4-3)!} [F(y_3)]^{3-1} [1 - F(y_i)]^{4-3} f(y_i) \\
&= 12(y_3^2)^2 (1 - y_3^2)^1 2y_3 \\
&= 24 y_3^5 (1 - y_3^2), \quad 0 < y_3 < 1,
\end{aligned}$$

which leads to

$$\begin{aligned} P\left(Y_3 > \frac{1}{2}\right) &= \int_{\frac{1}{2}}^{1} g(y_3)dy_3 = 24\int_{\frac{1}{2}}^{1} y_3^5(1-y_3^2)dy_3 \\ &= 24\left(\frac{1}{6}y_3^6 - \frac{1}{8}y_3^8\right)\Big|_{\frac{1}{2}}^{1} \\ &= \frac{243}{256}. \end{aligned}$$

Example 5.2.3. *(Example 4.4.3) Find the pdf of $Z = Y_3 - Y_1$ for a random sample of size 3 from the uniform distribution $unif[0,1]$.*

Example 5.2.4. *(Quantiles) Let the cdf $F(x)$ of X be continuous. For any $0 < p < 1$, the p-th quantile of X is defined by $\xi_p = F^{-1}(p)$. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample and let $Y_1 < Y_2 < \ldots < Y_n$ be its order statistics. Let $k = [p(n+1)]$ be integer part of $p(n+1)$. Then*

$$\begin{aligned} E[F(Y_k)] &= \int_{-\infty}^{\infty} F(y_k)g(y_k)dy_k \\ &= \int_{-\infty}^{\infty} F(y_k)\frac{n!}{(k-1)!(n-k)!}[F(y_k)]^{k-1}(1-F(y_k))^{n-k}f(y_k)dy_k \\ &= \frac{n!}{(k-1)!(n-k)!}\int_{-\infty}^{\infty}[F(y_k)]^k[1-F(y_k)]^{n-k}dF(y_k) \\ &= \frac{n!}{(k-1)!(n-k)!}\frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \\ &= \frac{k}{n+1}. \end{aligned}$$

Since $p \approx \frac{k}{n+1}$, it is reasonable to take Y_k as an estimator of ξ_p.

5.3 Method of moments estimation

The method of moments is probably the oldest method of deriving a point estimator, dating back to Karl Pearson, an English mathematical statistician, in the late 1800s. The method of moments estimation is quite simple, but it may not be the best estimator.

Definition 5.3.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample taken from a population having the pdf or pmf $f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. Then the method of moments estimator (MME) of $\boldsymbol{\theta}$ is given by the statistic*

$$\mathbf{T}(X_1, X_2, \ldots, X_n) = \mathbf{h}\left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2, \ldots, \frac{1}{n}\sum_{i=1}^{n} X_i^k\right)$$

where $\boldsymbol{\theta} = \mathbf{h}(\mu_1', \mu_2', \ldots, \mu_k')$, $\mathbf{h}$ *is a known function and* $\mu_j' = E(X^j)$, $j = 1, 2, \ldots, k$.

Remark 5.3.1. *The procedure of finding the MME consists of three steps below:*

(1) Calculate lower-order population moments and find expressions in terms of parameters

$$\begin{aligned} \mu_1' &= h_1(\theta_1, \theta_2, \ldots, \theta_k), \\ \mu_2' &= h_2(\theta_1, \theta_2, \ldots, \theta_k), \\ &\vdots \\ \mu_k' &= h_k(\theta_1, \theta_2, \ldots, \theta_k). \end{aligned}$$

(2) Invert the expressions found in Step 1 and find solutions for the parameters in terms of the population moments $\mu_1', \mu_2', \ldots, \mu_k'$.

(3) Replace the population moments μ_i' *by the sample moments* $\frac{1}{n}\sum_{j=1}^n X_j^i$ *in Step 2 for* $i = 1, 2, \ldots, k$.

Example 5.3.1. *Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from the binomial distribution* $b(m, p)$, *where* m *is known and* p *is unknown. Find the MME of* p.

Solution. *Since* $\mu_1' = E(X) = mp$, $p = \mu_1'/m$. *Replacing* μ_1' *by the sample mean obtains the MME of* p *below*

$$\hat{p} = \frac{1}{m}\hat{\mu}_1' = \frac{1}{m}\bar{X}_n = \frac{1}{mn}\sum_{i=1}^n X_i.$$

Example 5.3.2. *Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from the normal distribution* $N(\mu, \sigma^2)$, *where* μ *and* σ^2 *are unknown. Find the MMEs of* μ *and* σ^2.

Solution. *Since*

$$\begin{aligned} \mu_1' &= E(X) = \mu, \\ \mu_2' &= E(X^2) = \mathrm{Var}(X) + [E(X)]^2 = \sigma^2 + \mu^2, \end{aligned} \tag{5.2}$$

solving (5.2) for μ *and* σ^2 *in terms of* μ_1' *and* μ_2' *will obtain that*

$$\begin{aligned} \mu &= \mu_1', \\ \sigma^2 &= \mu_2' - (\mu_1')^2. \end{aligned} \tag{5.3}$$

Replacing μ_1' *and* μ_2' *by their corresponding sample parts in (5.3) will yield the MMEs of* μ *and* σ^2 *below*

$$\begin{aligned} \hat{\mu} &= \bar{X}_n, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2. \end{aligned}$$

Example 5.3.3. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Gamma distribution $\Gamma(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$ are unknown. Find the MMEs of α and β.*

Solution. *Since*

$$\begin{aligned} \mu_1' &= E(X) = \alpha\beta, \\ \mu_2' &= E(X^2) = \alpha(\alpha+1)\beta^2. \end{aligned} \tag{5.4}$$

Solving the system of equations (5.4) for α and β, we obtain that

$$\begin{aligned} \alpha &= \frac{(\mu_1')^2}{\mu_2' - (\mu_1')^2}, \\ \beta &= \frac{\mu_2' - (\mu_1')^2}{\mu_1'}. \end{aligned} \tag{5.5}$$

Replacing μ_1' and μ_2' by their corresponding sample parts in (5.5) will yield the MMEs of α and β below

$$\begin{aligned} \hat{\alpha} &= \frac{(\hat{\mu}_1')^2}{\hat{\mu}_2' - (\hat{\mu}_1)^2} = \frac{n\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}, \\ \hat{\beta} &= \frac{1}{n\bar{X}_n} \sum_{i=1}^n (X_i - \bar{X}_n)^2. \end{aligned}$$

Example 5.3.4. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution $\mathcal{P}(\lambda)$, where $\lambda > 0$ is unknown. Find the MME of λ.*

Solution. *Since $\mu_1' = E(X) = \lambda$, the MME of λ is given by $\hat{\lambda} = \bar{X}_n$. The Poisson distribution has a nice property that $\mathrm{Var}(X) = E(X) = \lambda$, which leads to another estimator*

$$\hat{\hat{\lambda}} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$

This estimator is based on $\mathrm{Var}(X) = \lambda = \mu_2' - (\mu_1')^2$. Can we consider $\hat{\hat{\lambda}}$ an MME of λ? The answer is negative because the MME is an estimator involving the lowest-order sample moments.

Example 5.3.5. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the uniform distribution $unif[-\theta, \theta]$, where $\theta > 0$ is unknown. Find the MME of θ.*

Solution. *Since $\mu_1' = E(X) = 0$, which does not depend on the parameter θ, we evaluate $\mu_2' = E(X^2)$, which is equal to $\theta^2/3$. Thus, $\theta = \sqrt{3\mu_2'}$. Replacing μ_2' by the second sample moment obtains the MME of θ below*

$$\hat{\theta} = \left(\frac{3}{n} \sum_{i=1}^n X_i^2 \right)^{1/2}.$$

Remark 5.3.2. *This example shows that although the population distribution* $unif[-\theta, \theta]$ *contains only one parameter* θ*, we need to evaluate the second population moment* μ_2' *in order to find the MME of* θ*.*

Example 5.3.6. *Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from* X *with a uniform distribution* $unif[0, \theta]$*. How can we estimate* θ*?*

Solution. *We consider two methods to estimate the parameter* θ*.*

(i) *By method of moments estimation, we calculate*

$$\mu_1' = E(X) = \int_0^\theta \frac{xdx}{\theta} = \frac{\theta}{2},$$

which is equivalent to $\theta = 2\mu_1'$*. Replacing* μ_1' *by* $\bar{X}$*, we obtain that the MME of* θ *is* $\hat{\theta} = 2\bar{X}$*. Now we study the property of* $\hat{\theta}$ *below.*

Since $E(\hat{\theta}) = 2E(\bar{X}) = 2E(X_1) = 2(\theta/2) = \theta$*,* $\hat{\theta}$ *is unbiased. Furthermore,*

$$\mathrm{Var}(\hat{\theta}) = 4\mathrm{Var}(\bar{X}) = \frac{4}{n}\mathrm{Var}(X_1) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Through Chebyshev's inequality, we have

$$P(|\hat{\theta} - \theta| \geq \epsilon) \leq \frac{\mathrm{Var}(\hat{\theta})}{\epsilon^2} = \frac{\theta^2}{3n\epsilon^2} \to 0$$

as $n \to \infty$*. Thus,* $\hat{\theta} \xrightarrow{p} \theta$*.*

(ii) *We now consider another estimator of* θ *defined by* $X_{(n)} = \max(X_1, X_2, \ldots, X_n)$ *because it is the closest to* θ*. To study properties of* $X_{(n)}$*, we need to find its distribution. A direct calculation shows that the cdf of* $X_{(n)}$ *has the form*

$$G(t) = \begin{cases} 0, & \text{if } t < 0, \\ (t/\theta)^n, & \text{if } 0 \leq t < \theta, \\ 1, & \text{if } t \geq \theta. \end{cases}$$

Thus, the pdf of $X_{(n)}$ *is given by*

$$g(t) = \begin{cases} nt^{n-1}/\theta^n, & \text{if } 0 \leq t < \theta, \\ 0, & \text{otherwise,} \end{cases}$$

which leads to

$$E(X_{(n)}) = \int_0^\theta t\frac{nt^{n-1}}{\theta^n}dt = \frac{n}{n+1}\theta \neq \theta.$$

That means, $X_{(n)}$ is biased. Furthermore, we have

$$\begin{aligned} E(X_{(n)}^2) &= \int_0^\theta t^2 \frac{nt^{n-1}}{\theta^n} dt = \frac{n}{n+2}\theta^2, \\ \mathrm{Var}(X_{(n)}) &= \frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2 \\ &= \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right]\theta^2 \\ &= \frac{n}{(n+1)^2(n+2)}\theta^2. \end{aligned}$$

Applying Markov's inequality to the random variable $(X_{(n)} - \theta)^2$, we have

$$\begin{aligned} P(|X_{(n)} - \theta| \geq \epsilon) &\leq \frac{E(X_{(n)} - \theta)^2}{\epsilon^2} \\ &= \frac{1}{\epsilon^2}\left(\mathrm{Var}(X_{(n)}) + [E(X_{(n)}) - \theta]^2\right) \\ &= \left(\frac{n}{(n+1)^2(n+2)} + \frac{1}{(n+1)^2}\right)\frac{\theta^2}{\epsilon^2} \\ &= \frac{2}{(n+1)(n+2)}\frac{\theta^2}{\epsilon^2} \\ &\to 0 \end{aligned}$$

as $n \to \infty$. This means that $X_{(n)} \xrightarrow{p} \theta$.

Now we compare variance of $\hat{\theta}$ with variance of $X_{(n)}$ and see that

$$\mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{3n} > \frac{n\theta^2}{(n+1)^2(n+2)} = \mathrm{Var}(X_{(n)})$$

for any $\theta > 0$ and $n \geq 1$, i.e., the variance of $\hat{\theta}$ is larger than the variance of $X_{(n)}$ even though $\hat{\theta}$ is unbiased. The estimate $X_{(n)}$ is based on another method of estimation, called maximum likelihood estimation. Figure 5.1.1 below presents two scatter plots of two estimators based on 2000 simulations in which the sample is generated from $unif[0,1]$ with size 10. We can see that MME has a larger variety of around 1 than MLE.

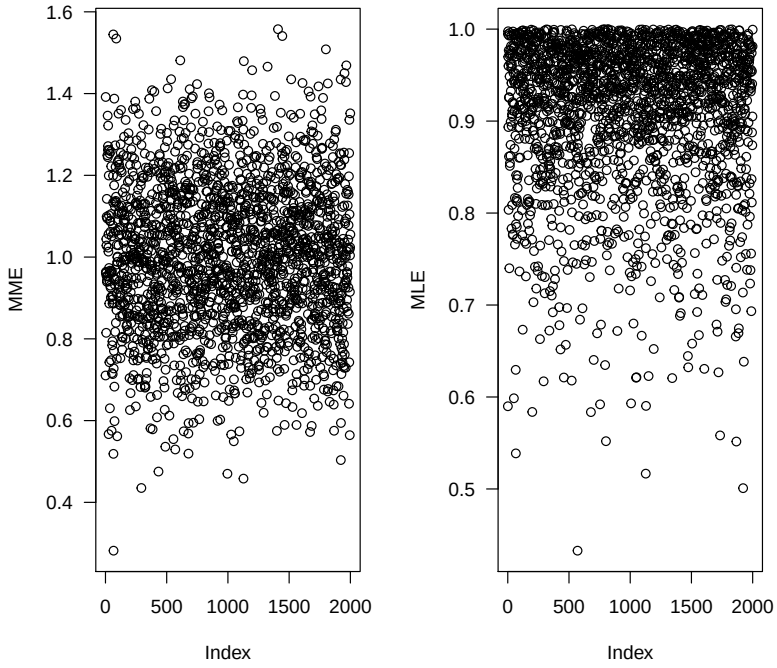


Figure 5.1.1 Comparison of MME and MLE

5.4 Maximum likelihood estimation

Suppose that an i.i.d. sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ *of a random variable* X *is chosen according to one of a family of probability distributions* $\{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$*, where* Θ *is the parameter space, which consists of all possible values of* $\boldsymbol{\theta}$*. In addition, we use* $f(\mathbf{x}; \boldsymbol{\theta})$ *to denote the pdf or pmf for the data when* $\boldsymbol{\theta}$ *is the true value. Then the principle of maximum likelihood proposed by geneticist/statistician Sir Ronald A. Fisher in 1922 is to choose the estimator* $\hat{\boldsymbol{\theta}}$ *as the value for the parameter that makes the observed data most probable.*

Definition 5.4.1. *Given* $\mathbf{X} = \mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ *is observed, the function*

of $\boldsymbol{\theta}$ defined by

$$L(\boldsymbol{\theta};\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Theta \tag{5.6}$$

is called the likelihood function of the observed data $\mathbf{x}$.

Remark 5.4.1. *Although the likelihood function has the same form as the joint pdf or pmf of the observed data, it is not a pdf or pmf because it is a function of* $\boldsymbol{\theta}$. *The likelihood function* $L(\boldsymbol{\theta};\mathbf{x})$ *is a measure of how likely is to have produced* $\mathbf{x}$, *i.e., a certain value of* $\boldsymbol{\theta}$ *appears to be more likely than others. If* $L(\boldsymbol{\theta}_1;\mathbf{x}) > L(\boldsymbol{\theta}_2;\mathbf{x})$, *it means that the sample* $\mathbf{X} = \mathbf{x}$ *we observed is more likely to have occurred if* $\boldsymbol{\theta} = \boldsymbol{\theta}_1$ *than* $\boldsymbol{\theta} = \boldsymbol{\theta}_2$.

Definition 5.4.2. *The maximum likelihood estimator (MLE)* $\hat{\boldsymbol{\theta}}$ *of* $\boldsymbol{\theta}$ *is a value of* $\boldsymbol{\theta}$ *that maximizes the likelihood function* $L(\boldsymbol{\theta};\mathbf{x})$, *i.e.,*

$$L(\hat{\boldsymbol{\theta}};\mathbf{x}) = \sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta};\mathbf{x}). \tag{5.7}$$

Remark 5.4.2. *The method of maximum likelihood generally results in the problem of maximizing* $L(\boldsymbol{\theta};\mathbf{x})$ *with respect to* $\boldsymbol{\theta}$. *A way from calculus to do the maximization is to take a derivative or partial derivative with respect to* $\boldsymbol{\theta}$ *and then equate the derivative or partial derivatives to zero. The solution* $\hat{\boldsymbol{\theta}}$ *will be the MLE of* $\boldsymbol{\theta}$ *if it is a global maximizer. It should be mentioned that rather than maximizing* $L(\boldsymbol{\theta};\mathbf{x})$ *that can be quite tedious, we often maximize the log-likelihood function* $l(\boldsymbol{\theta};\mathbf{x}) = \ln[L(\boldsymbol{\theta};\mathbf{x})]$ *because the natural logarithm is an increasing function.*

Example 5.4.1. *Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from* X *with a Bernoulli distribution* $b(1,\theta)$, *where* $\theta \in [0,1]$ *is unknown. Find the MLE of* θ.

Solution. *Since the pmf of Bernoulli distribution with parameter* θ *has the form*

$$f(x;\theta) = \theta^x(1-\theta)^{1-x}, \quad x = 0, 1,$$

the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ *is given by*

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i},$$

and thus the log-likelihood function is equal to

$$l(\theta;\mathbf{x}) = \left(\sum_{i=1}^{n} x_i\right)\ln(\theta) + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-\theta).$$

Taking the derivative of $l(\theta;\mathbf{x})$ with respect to θ and setting it to be zero, we obtain that

$$l'(\theta;\mathbf{x}) = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} x_i\right) = 0,$$

whose solution is $\frac{1}{n}\sum_{i=1}^{n} x_i$. Since $l(\theta;\mathbf{x})$ is concave, we can conclude that the MLE of θ is $\hat{\theta} = \bar{X}$.

Example 5.4.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from X with a Poisson distribution $\mathcal{P}(\theta)$, where $\theta > 0$ is unknown. Find the MLE of θ.*

Solution. *Since the pmf of $\mathcal{P}(\theta)$ has the form*

$$f(x;\theta) = \frac{\theta^x}{x!}e^{-\theta}, \quad x = 0, 1, 2, \ldots,$$

the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ is given by

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}}{x_i!}e^{-\theta} = \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}e^{-n\theta},$$

and thus the log-likelihood function is equal to

$$l(\theta;\mathbf{x}) = \sum_{i=1}^{n} x_i \ln(\theta) - n\theta - \sum_{i=1}^{n} \ln(x_i!).$$

Taking the derivative of $l(\theta;\mathbf{x})$ with respect to θ and setting it to be zero, we obtain that

$$l'(\theta;\mathbf{x}) = \frac{1}{\theta}\sum_{i=1}^{n} x_i - n = 0,$$

whose solution is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Since $l(\theta;\mathbf{x})$ is concave, we can conclude that the MLE of θ is $\hat{\theta} = \bar{X}$.

Remark 5.4.3. *For Example 5.4.2, what is the mle of θ if all observed values $x_1, x_2, \ldots, x_n$ are 0? In this case, the likelihood function is $e^{-n\theta}$, which is monotonically decreasing in θ. Since $\Theta = (0, \infty)$, the mle of θ in this case does not exist.*

Example 5.4.3. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from X with a geometric distribution $G(\theta)$, where $\theta \in (0, 1]$ is unknown. Find the MLE of θ.*

Solution. *Since the pmf of geometric distribution* $\mathrm{geom}(\theta)$ *has the form*

$$f(x;\theta) = \theta(1-\theta)^x, \quad x = 0, 1, \ldots,$$

the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ *is given by*

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \theta) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{\sum_{i=1}^{n} x_i},$$

and thus the log-likelihood function is equal to

$$l(\theta; \mathbf{x}) = n \ln(\theta) + \left(\sum_{i=1}^{n} x_i \right) \ln(1-\theta).$$

Taking the derivative of $l(\theta; \mathbf{x})$ *with respect to* θ *and setting it to be zero, we obtain that*

$$l'(\theta; \mathbf{x}) = \frac{n}{\theta} - \frac{1}{1-\theta} \sum_{i=1}^{n} x_i = 0,$$

whose solution is $(1+\bar{x})^{-1}$. *Since* $l(\theta; \mathbf{x})$ *is concave, we can conclude that the MLE of* θ *is* $\hat{\theta} = (1+\bar{X})^{-1}$.

Example 5.4.4. *Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from* X *with a uniform distribution* $unif[0, \theta]$, *where* $\theta \in [0, \infty)$ *is unknown. Find the MLE of* θ.

Solution. *Since the pdf of uniform distribution* $unif[0, \theta]$ *has the form*

$$f(x; \theta) = \frac{1}{\theta} I(0 \leq x \leq \theta).$$

the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ *is given by*

$$\begin{aligned} L(\theta; \mathbf{x}) &= \prod_{i=1}^{n} f(x_i; \theta) \\ &= \left(\frac{1}{\theta} \right)^n \prod_{i=1}^{n} I(0 \leq x_i \leq \theta) \\ &= \left(\frac{1}{\theta} \right)^n I(0 \leq x_{(1)} \leq x_{(n)} \leq \theta). \end{aligned}$$

The graph of $L(\theta; \mathbf{x})$ *shows that when* $\theta < x_{(n)}$, $L(\theta|\mathbf{x}) = 0$ *and* $\theta \geq x_{(n)}$, $L(\boldsymbol{\theta}; \mathbf{x}) = (1/\theta)^n$ *that is monotonically decreasing. That means,* $L(\theta; \mathbf{x})$ *is maximized at* $x_{(n)}$. *Thus, the MLE of* θ *is* $\hat{\theta} = X_{(n)}$. *This estimator is actually used in Example 5.3.6.*

Remark 5.4.4. *The maximum likelihood estimator may not be unique. The likelihood function* $L(\boldsymbol{\theta}; \mathbf{x})$ *can have many local maxima when it is defined in high dimensional space. Thus, finding the global maximum can be a major computational challenge.*

Example 5.4.5. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from X with a uniform distribution $unif[\theta-1/2, \theta+1/2]$, where $\theta \in (-\infty, \infty)$ is unknown. Find the MLE of θ.*

Solution. *Since the pdf of uniform distribution $unif[\theta-1/2, \theta+1/2]$ has the form*

$$f(x;\theta) = I(\theta - 1/2 \le x \le \theta + 1/2),$$

the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ is given by

$$\begin{aligned} L(\theta;\mathbf{x}) &= \prod_{i=1}^{n} f(x_i;\theta) \\ &= \prod_{i=1}^{n} I(\theta - 1/2 \le x_i \le \theta + 1/2) \\ &= I(\theta - 1/2 \le x_{(1)} \le x_{(n)} \le \theta + 1/2) \\ &= I(x_{(n)} - 1/2 \le \theta \le x_{(1)} + 1/2). \end{aligned}$$

Thus, any value in the closed interval $[X_{(n)} - 1/2, X_{(1)} + 1/2]$ is an MLE of θ.

Remark 5.4.5. *When the parameter $\boldsymbol{\theta}$ is a vector, we often use the following methods to obtain the MLE of $\boldsymbol{\theta}$:*

(1) If $\boldsymbol{\theta}$ is a continuous variable and $\dfrac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ exists, solve $\dfrac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$ and check $\dfrac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \le \mathbf{0}$, where

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1} \quad \cdots \quad \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_k} \right)^T,$$
$$\left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right) = \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right)_{k \times k}$$

if $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^T$.

(2) Use the definition of the MLE for other cases.

Example 5.4.6. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, where parameters $\mu \in (-\infty, \infty)$ and $\sigma^2 > 0$ are unknown. Find the MLEs of μ and σ^2.*

Solution. *Since the pdf of $N(\mu, \sigma^2)$ has the form*

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), \quad -\infty < x < \infty,$$

where $\boldsymbol{\theta} = (\mu, \sigma^2)^T$, *the likelihood function of observed data* $\mathbf{X} = \mathbf{x}$ *is given by*

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right),$$

and thus the log-likelihood function is equal to

$$l(\boldsymbol{\theta}; \mathbf{x}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Taking the partial derivatives of $l(\boldsymbol{\theta}; \mathbf{x})$ *with respect to* μ *and* σ^2 *and setting them to be zero, we obtain that*

$$\frac{\partial}{\partial \mu} l(\theta; \mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0,$$
$$\frac{\partial}{\partial \sigma^2} l(\theta; \mathbf{x}) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0.$$

Solving this system of equations for (μ, σ^2) *obtains that*

$$(\hat{\mu}, \hat{\sigma}^2) = \left(\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right).$$

Note that

$$\begin{aligned}
\mathbf{A} &= \left(\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right) \\
&= \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mu^2} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \sigma^2 \partial \sigma^2} \end{pmatrix} \\
&= \begin{pmatrix} -n/\sigma^2 & -\sum_{i=1}^{n}(x_i - \mu)/\sigma^4 \\ -\sum_{i=1}^{n}(x_i - \mu)/\sigma^4 & n/(2\sigma^4) - \sum_{i=1}^{n}(x_i - \mu)^2/\sigma^6 \end{pmatrix}.
\end{aligned}$$

When $\mu = \hat{\mu} = \bar{x}$ *and* $\sigma^2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, *the matrix* $\mathbf{A}$ *becomes*

$$\begin{aligned}
\mathbf{A}\big|_{\mu=\hat{\mu}, \sigma^2=\hat{\sigma}^2} &= \begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & n/(2\hat{\sigma}^4 - \sum_{i=1}^{n}(x_i - \bar{x})^2/\hat{\sigma}^6 \end{pmatrix} \\
&= \begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -n/(2\hat{\sigma}^4) \end{pmatrix},
\end{aligned}$$

which is negative definite. Thus, we can conclude that the MLEs of μ *and* σ^2 *are* $\bar{X}$ *and* $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, *respectively.*

Example 5.4.7. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from X with a gamma distribution $\Gamma(\alpha, \beta)$, where $\boldsymbol{\theta}^T = (\alpha, \beta) \in \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$ is unknown. Find the MLE of $\boldsymbol{\theta}$.*

Solution. *Since the pdf of gamma distribution $\Gamma(\alpha, \beta)$ is*

$$f(x; \boldsymbol{\theta}) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-x/\beta), \quad x > 0,$$

the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ is given by

$$L(\boldsymbol{\theta}; \mathbf{x}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} \exp(-x_i/\beta),$$

and thus the log-likelihood function is equal to

$$l(\boldsymbol{\theta}; \mathbf{x}) = -n \ln\big(\Gamma(\alpha)\big) - n\alpha \ln(\beta) + (\alpha - 1) \sum_{i=1}^{n} \ln(x_i) - \frac{1}{\beta} \sum_{i=1}^{n} x_i.$$

Taking the partial derivatives of $l(\theta; \mathbf{x})$ with respect to α and β and setting them to be zero, we obtain that

$$\frac{\partial}{\partial \alpha} l(\boldsymbol{\theta}; \mathbf{x}) = -n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} - n \ln(\beta) + \sum_{i=1}^{n} \ln(x_i) = 0,$$

$$\frac{\partial}{\partial \beta} l(\boldsymbol{\theta}; \mathbf{x}) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x_i = 0.$$

Although we can solve for β in terms of α explicitly, we have a nonlinear equation in α, which cannot be solved in a closed form. There are two methods and they are called root-finding methods, which are based on Bolzano's theorem in calculus that states that if a continuous function defined on an interval changes its signs on an interval, it must be zero at some point on that interval.

Example 5.4.8. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution $b(1, \theta)$. The $\hat{\theta} = \bar{X}$ is the MLE of θ provided $\Theta = [0, 1]$. If $0 \leq \theta \leq \frac{1}{3}$, $\bar{X}$ might not be the MLE of θ because of possibility that $\bar{X} > \frac{1}{3}$.*

Theorem 5.4.1. *(Invariance property) MLE is preserved by parameterization $\boldsymbol{\eta} = g(\boldsymbol{\theta})$ with a known function $g : \Theta \to g(\Theta)$. That is, the MLE of $\boldsymbol{\eta}$ is $g(\hat{\boldsymbol{\theta}})$ if the MLE of $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}$.*

Proof. See page 359. □

Theorem 5.4.2. *Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. and satisfy $(R0) - (R2)$. Let θ_0 be the true parameter. Then the likelihood function*

$$\frac{\partial}{\partial\theta}L(\theta) = 0 \text{ or } \frac{\partial}{\partial\theta}l(\theta) = 0$$

has a solution $\hat{\theta}_n$ such that $\hat{\theta}_n \xrightarrow{p} \theta_0$.

Proof. See page 359. □

Definition 5.4.3. *Fisher information is defined by*

$$I(\theta) = -E\left(\frac{\partial^2 \ln[f(X;\theta)]}{\partial\theta^2}\right) = \text{Var}\left(\frac{\partial \ln[f(X;\theta)]}{\partial\theta}\right).$$

Since $1 = \int f(x;\theta)dx$, taking the derivative with respect to θ on the both sides will yield that

$$\begin{aligned}
0 &= \int \frac{\partial f(x;\theta)}{\partial\theta}dx \\
&= \int \frac{\partial f(x;\theta)/\partial\theta}{f(x;\theta)} f(x;\theta)dx \\
&= E\left(\frac{\partial \ln[f(X;\theta)]}{\partial\theta}\right) \\
&= \int \frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta} f(x;\theta)dx \\
&+ \int \frac{\partial \ln[f(x;\theta)]}{\partial\theta}\frac{\partial \ln[f(x;\theta)]}{\partial\theta} f(x;\theta)dx \\
&= \int \frac{\partial^2 \ln[f(x;\theta)]}{\partial\theta} f(x;\theta)dx \\
&+ \int \left(\frac{\partial \ln[f(x;\theta)]}{\partial\theta^2}\right)^2 f(x;\theta)dx \\
&= E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right) + E\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2,
\end{aligned}$$

which is equivalent to

$$-E\left(\frac{\partial^2 \ln f(X;\theta)}{\partial\theta^2}\right) = E\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)^2 = \text{Var}\left(\frac{\partial \ln f(X;\theta)}{\partial\theta}\right)$$

Theorem 5.4.3. *(Asymptotic normality of MLE) Assume that $X_1, X_2, \ldots, X_n$ are i.i.d. with pdf $f(x;\theta_0)$ for $\theta_0 \in \Theta$ such that the regularity conditions (R0)–(R5) are satisfied. Furthermore, suppose that the Fisher information $I(\theta_0)$ exists. Then any consistent sequence of solutions of the MLE equations satisfies*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N\big(0, 1/I(\theta_0)\big).$$

Proof. See page 369. □

Corollary 5.4.1. *If $g'(\theta_0) \neq 0$, then*

$$\sqrt{n}[g(\hat{\theta}_n) - g(\theta_0)] \xrightarrow{d} N\left(0, \frac{[g'(\theta_0)]^2}{I(\theta_0)}\right).$$

5.5 Efficient estimator

Definition 5.5.1. *Let $T(X_1, X_2, \ldots, X_n)$ be a statistic. We say that T is an unbiased estimator of θ if $E(T) = \theta$.*

Theorem 5.5.1. *(Rao-Cramér Lower Bound) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a common pdf $f(x;\theta)$, where $\theta \in \Theta$. Assume that $(R0) - (R4)$ holds. Let $Y = u(X_1, X_2, \ldots, X_n)$ be a statistic with mean $E(Y) = E[u(X_1, X_2, \ldots, X_n)] = k(\theta)$. Then*

$$\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)},$$

where $I(\theta)$ is the Fisher information.

Proof. (continuous case) Since $E(Y) = k(\theta)$,

$$k(\theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2 \ldots, x_n) \prod_{i=1}^{n} f(x_i;\theta) dx_1 dx_2 \cdots dx_n,$$

which implies that

$$\begin{aligned} k'(\theta) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \ldots, x_n) \left(\sum_{i=1}^{n} \frac{1}{f(x_i;\theta)} \frac{\partial f(x_i;\theta)}{\partial \theta} \right) \prod_{i=1}^{n} f(x_i;\theta) dx_i \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} u(x_1, x_2, \ldots, x_n) \left(\sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta} \right) \prod_{i=1}^{n} f(x_i;\theta) dx_i. \end{aligned}$$

Let $Z = \sum_{i=1}^{n} \frac{\partial \ln[f(X_i;\theta)]}{\partial \theta}$. Then Z is a random variable and

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{i=1}^{n} \frac{\partial \ln f(x_i;\theta)}{\partial \theta} \prod_{j=1}^{n} f(x_j;\theta) dx_j \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{\partial f(x_i;\theta)}{\partial \theta} \prod_{j \neq i} f(x_j;\theta) dx_j \\ &= \sum_{i=1}^{n} \int_{-\infty}^{\infty} \frac{\partial f(x_i;\theta)}{\partial \theta} dx_i \\ &= \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x_i;\theta) dx_i \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \mathrm{Var}(Z) &= \mathrm{Var}\left(\sum_{i=1}^{n} \frac{\partial \ln[f(X_i;\theta)]}{\partial \theta} \right) \\ &= \sum_{i=1}^{n} \mathrm{Var}\left(\frac{\partial \ln[f(X_i;\theta)]}{\partial \theta} \right) \\ &= n \cdot \mathrm{Var}\left(\frac{\partial \ln[f(X;\theta)]}{\partial \theta} \right) \\ &= n \cdot E\left(\frac{\partial \ln[f(X;\theta)]}{\partial \theta} \right)^2 \\ &= n \cdot I(\theta). \end{aligned}$$

Moreover,

$$k'(\theta) = E(YZ) = E(Y)E(Z) + \rho \sigma_Y \sqrt{nI(\theta)} = \rho \sigma_Y \sqrt{nI(\theta)},$$

where ρ is the correlation coefficient between Y and Z. Thus,

$$[k'(\theta)]^2 = \rho^2 \sigma_Y^2 nI(\theta) = \rho^2 \mathrm{Var}(Y) nI(\theta),$$

which is equivalent to

$$\rho^2 = \frac{[k'(\theta)]^2}{\mathrm{Var}(Y) nI(\theta)}.$$

Since $\rho^2 \leq 1$, then $\mathrm{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)}$. □

Corollary 5.5.1. *If Y is an unbiased estimator of θ, then* $\mathrm{Var}(Y) \geq \frac{1}{nI(\theta)}$.

Definition 5.5.2. *An unbiased estimator Y of the parameter θ is called an efficient estimator if and only if the variance of Y is equal to* $1/[nI(\theta)]$.

Definition 5.5.3. *Efficiency of an unbiased estimator Y of θ is defined by*

$$\frac{1/[nI(\theta)]}{\text{Var}(Y)} = \frac{1}{nI(\theta)\text{Var}(Y)}.$$

Example 5.5.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution $\mathcal{P}(\lambda)$. Then MLE $\hat{\lambda}$ of λ is efficient.*

Proof. We know from Example 5.4.2 that the MLE $\hat{\lambda}$ of λ is $\bar{X}$, i.e., $\hat{\lambda} = \bar{X}$. Thus,

$$\text{Var}(\hat{\lambda}) = \text{Var}(\bar{X}) = \frac{\text{Var}(X)}{n} = \frac{\lambda}{n}.$$

Meanwhile, the pmf of X has the form

$$f(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}, \quad x = 0, 1, \ldots,$$

which leads to

$$\ln[f(x;\lambda)] = x\,\ln(\lambda) - \lambda - \ln(x!)$$

and

$$\frac{d\ln[f(x;\lambda)]}{d\lambda} = \frac{x}{\lambda} - 1.$$

Thus,

$$\begin{aligned} I(\lambda) &= E\left(\frac{d\ln[f(X;\lambda)]}{d\lambda}\right)^2 = E\left(\frac{X}{\lambda} - 1\right)^2 \\ &= \frac{1}{\lambda^2}E(X-\lambda)^2 = \frac{1}{\lambda^2}\text{Var}(X) = \frac{1}{\lambda}. \end{aligned}$$

Since $\text{Var}(\hat{\lambda}) = 1/[nI(\lambda)] = \lambda/n$, $\hat{\lambda} = \bar{X}$ is an efficient estimator of λ. □

Example 5.5.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the beta distribution beta$(\theta, 1)$. Then the MLE of θ is not efficient.*

Proof. We first find the MLE of θ. Since the pdf of beta distribution *beta*$(\theta, 1)$ has the form

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 < x < 1,$$

where $\theta \in (0, \infty)$, the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ is given by

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1}$$

and thus the log-likelihood function is equal to

$$l(\theta;\mathbf{x}) = n\ln(\theta) + (\theta-1)\ln\left(\prod_{i=1}^{n} x_i\right).$$

Taking the derivative of $l(\theta;\mathbf{x})$ with respect to θ and setting it to be zero, we obtain that

$$\frac{dl}{d\theta} = \frac{n}{\theta} + \ln\left(\prod_{i=1}^{n} x_i\right) = 0,$$

whose solution is $-n/\sum_{i=1}^{n}\ln(x_i)$. Since $l(\theta|\mathbf{x})$ is concave, the MLE of θ is $\hat{\theta} = -n/\sum_{i=1}^{n}\ln(X_i) = 1/\bar{Y}$, where $\bar{Y} = \sum_{i=1}^{n} Y_i/n$ and $Y_i = -\ln(X_i)$ for $i = 1, 2, \ldots, n$.

We now show that $\hat{\theta}$ is not efficient. To do so, we need to find the distribution of $Y = -\ln(X)$. Indeed, the cdf of Y is given by

$$G(y) = P(Y \le y) = P(-\ln(X) \le y) = P(X \ge e^{-y}).$$

Taking the derivative of $G(y)$ yields the pdf of Y below

$$g(y) = f(e^{-y};\theta)e^{-y} = \theta e^{-\theta y}, \quad y > 0,$$

i.e., $Y \sim \Gamma(1, 1/\theta)$. Using this fact and the additivity of the gamma distribution, we can conclude that $W = \sum_{i=1}^{n} Y_i \sim \Gamma(n, 1/\theta)$. Thus,

$$\begin{aligned} E(W^k) &= \int_0^\infty w^k \frac{\theta^n}{\Gamma(n)} w^{n-1} e^{-\theta w} dw \\ &= \frac{\theta^n}{\Gamma(n)} \int_0^\infty w^{k+n-1} e^{-\theta w} dw \\ &= \frac{\theta^n}{\Gamma(n)} \frac{\Gamma(n+k)}{\theta^{n+k}} \int_0^\infty \frac{\theta^{n+k}}{\Gamma(n+k)} w^{n+k-1} e^{-\theta w} dw \\ &= \frac{\Gamma(n+k)}{\Gamma(n)} \frac{1}{\theta^k}. \end{aligned} \tag{5.8}$$

When $k = -1$, $E(W^{-1}) = \frac{\theta}{n-1}$, which implies that $E(\hat{\theta}) = E(nW^{-1}) = \frac{n}{n-1}\theta \neq \theta$, $\hat{\theta}$ is biased and thus not efficient.

Although $\hat{\theta}$ is biased, it can be modified to form an unbiased estimator of θ. Let $\theta^* = \frac{n-1}{n}\hat{\theta}$. Then $E(\theta^*) = \frac{n-1}{n}E(\hat{\theta}) = \frac{n-1}{n}\frac{n}{n-1}\theta = \theta$, i.e., θ^* is unbiased. Now we will determine if θ^* is efficient. To do so, we evaluate the variance of θ^*. Taking $k = -2$ in (5.8) yields that

$$E(W^{-2}) = \frac{\theta^2}{(n-1)(n-2)}.$$

Thus,

$$\mathrm{Var}(W^{-1}) = \frac{\theta^2}{(n-1)(n-2)} - \left(\frac{\theta}{n-1}\right)^2 = \frac{\theta^2}{(n-1)^2(n-2)}$$

and

$$\begin{aligned}\mathrm{Var}(\theta^*) &= \left(\frac{n-1}{n}\right)^2 \mathrm{Var}(\hat{\theta}) \\ &= \left(\frac{n-1}{n}\right)^2 n^2 \mathrm{Var}(W^{-1}) \\ &= \left(\frac{n-1}{n}\right)^2 \frac{n^2}{(n-1)^2(n-2)}\theta^2 \\ &= \frac{\theta^2}{n-2}.\end{aligned}$$

On the other hand, a direct calculation shows that

$$\begin{aligned}\ln[f(x;\theta)] &= \ln(\theta) + (\theta-1)\ln(x), \\ \frac{d\ln[f(x;\theta)]}{d\theta} &= \frac{1}{\theta} + \ln(x), \\ \frac{d^2\ln[f(x;\theta)]}{d\theta^2} &= -\frac{1}{\theta^2},\end{aligned}$$

which implies that

$$I(\theta) = -E\left(\frac{d^2\ln[f(X;\theta)]}{d\theta^2}\right) = \frac{1}{\theta^2}.$$

Clearly,

$$\mathrm{Var}(\theta^*) = \frac{\theta^2}{n-2} > \frac{\theta^2}{n} = \frac{1}{nI(\theta)}.$$

Thus, θ^* is also not efficient even though it is unbiased. □

5.6 Interval estimator

Definition 5.6.1. *Let $X_1, X_2, \ldots, X_n$ be i.i.d. from X with a pdf or pmf $f(x;\theta)$, $\theta \in \Theta = \mathbb{R}$. Then the interval*

$$[\hat{\theta}_L(X_1, X_2, \ldots, X_n), \hat{\theta}_U(X_1, X_2, \ldots, X_n)]$$

is called a $(1-\alpha)100\%$ confidence interval if

$$P\left[\hat{\theta}_L(X_1, X_2, \ldots, X_n) \le \theta \le \hat{\theta}_U(X_1, X_2, \ldots, X_n)\right] = 1-\alpha,$$

where $1-\alpha$ is called the confidence level and α, which is usually given such as 1%, 5%, or 10%, is called significance level.

Method: *Suppose that $f(x;\theta)$ is known except the parameter θ. Let $T(X_1, X_2, \ldots, X_n, \theta)$ be a real-valued function such that*

(i) *$T(X_1, X_2, \ldots, X_n, \theta)$ is a statistic for every θ,*

(ii) *as a function of θ, $T(X_1, X_2, \ldots, X_n, \theta)$ is monotone increasing or decreasing for every $(X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ and thus $\lambda = T(X_1, X_2, \ldots, X_n, \theta)$ is solvable for θ.*

If the distribution of T is independent of θ, one can use T, called a pivot or pivotal quantity, to construct a confidence interval for θ.

5.6.1 One sample case

Example 5.6.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$. Find the $(1-\alpha)100\%$ confidence interval for μ.*

Solution. *By Corollary 5.2.1, $\bar{X} \sim N(\mu, \sigma^2/n)$, which is equivalent to*

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

Meanwhile, we know from Corollary 5.2.3 that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

By Corollary 5.2.2, Z and S^2 are independent. Thus,

$$T = \frac{Z}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1},$$

which will be the pivot, i.e.,

$$P(|T| \le t_{\alpha/2,n-1}) = 1-\alpha.$$

Solving $|T| \le t_{\alpha/2,n-1}$ for μ, we obtain that

$$\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2,n-1} \le \mu \le \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2,n-1}.$$

Therefore, the $(1-\alpha)100\%$-confidence interval for μ is

$$\left[\bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2,n-1}, \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2,n-1}\right].$$

Example 5.6.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$. Find the $(1-\alpha)100\%$-confidence interval for σ^2.*

Solution. *By Corollary 5.2.3,*

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

which leads to

$$P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha.$$

This is equivalent to

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha.$$

Therefore, the $(1-\alpha)100\%$-confidence interval for σ^2 is

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right].$$

If the population is unknown, how can we construct the confidence interval for the population mean? The answer to this question is to use the central limit theorem, which states that if $X_1, X_2, \ldots, X_n$ form an i.i.d. from X with $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$, then

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \longrightarrow N(0,1)$$

as $n \to \infty$. We divide into two cases below:

(1) When σ^2 is known,

$$P\left(\left|\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\right| \leq z_{\alpha/2}\right) = 1-\alpha,$$

which is equivalent to

$$P\left(\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1-\alpha.$$

(2) When σ^2 is unknown, we replace σ by $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}$

Example 5.6.3. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution $b(1,\theta)$. It is known that MLE of θ is $\hat{\theta} = \bar{X}$.*

$$\frac{\sqrt{n}(\bar{X}-\theta)}{\sqrt{\theta(1-\theta)}} \longrightarrow N(0,1)$$

as $n \to \infty$. The $100(1-\alpha)\%$ confidence for θ is given by

$$\left[\bar{X} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \bar{X} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right]$$

5.6.2 Two samples

(1) Normal distributions with equal variance

Suppose that $X_1, X_2, \ldots, X_{n_1}$ is an i.i.d. sample from the normal distribution $N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_{n_2}$ is also an i.i.d. sample from the normal distribution $N(\mu_2, \sigma^2)$. Assume that two samples are independent. We want to derive the MLEs of μ_1, μ_2, and σ^2 and the confidence interval for $\mu_1 - \mu_2$.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_{n_1})^T$ and $\mathbf{y} = (y_1, y_2, \ldots, y_{n_2})^T$. Then the likelihood function of observed values $\mathbf{x}$ and $\mathbf{y}$ is given by

$$\begin{aligned} L(\mu_1, \mu_2, \sigma^2 | \mathbf{x}, \mathbf{y}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_1} \exp\left(-\frac{\sum_{i=1}^{n_1}(x_i - \mu_1)^2}{2\sigma^2}\right) \\ &\quad \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_2} \exp\left(-\frac{\sum_{i=1}^{n_2}(y_i - \mu_2)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_1+n_2} \exp\left(-\frac{\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{i=1}^{n_2}(y_i - \mu_2)^2}{2\sigma^2}\right) \end{aligned}$$

Thus, the log-likelihood function has the form

$$\begin{aligned} l(\mu_1, \mu_2, \sigma^2) &= \ln L \\ &= -\frac{n_1 + n_2}{2}\ln(2\pi) - \frac{n_1 + n_2}{2}\ln(\sigma^2) \\ &\quad - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{i=1}^{n_2}(y_i - \mu_2)^2\right). \end{aligned}$$

Taking the partial derivatives with respect to μ_1, μ_2, and σ^2 first and then setting them to be zero, we obtain the MLEs of three parameters below:

$$\begin{aligned} \hat{\mu}_1 &= \bar{X}, \\ \hat{\mu}_2 &= \bar{Y}, \\ \hat{\sigma}^2 &= \frac{1}{n_1 + n_2}\left(\sum_{i=1}^{n_1}(X_i - \bar{X})^2 + \sum_{i=1}^{n_2}(Y_i - \bar{Y})^2\right) \\ &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2}, \end{aligned}$$

where S_1^2 and S_2^2 are two sample variances.

(i) Confidence interval for $\mu_1 - \mu_2$

It is known from the normal sampling theory that

$$\bar{X} \sim N(\mu_1, \sigma^2/n_1),$$
$$\bar{Y} \sim N(\mu_2, \sigma^2/n_2),$$

which lead to

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$$

or equivalently

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0, 1).$$

Meanwhile, we also know from the normal sampling theory that

$$(n_1 + n_2)\frac{\hat{\sigma}^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

because of independence between $(n_1 - 1)S_1^2/\sigma^2 \sim \chi^2_{n_1-1}$ and $(n_2 - 1)S_2^2/\sigma^2 \sim \chi^2_{n_2-1}$ and additive property of χ^2 distribution.

Let

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

denote the pooled sample variance. Then $(n_1 + n_2 - 2)(S_p^2/\sigma^2) \sim \chi^2_{n_1+n_2-2}$. By the definition of t-distribution, we have

$$\frac{[(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)]/\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}}{\sqrt{(n_1 + n_2 - 2)S_p^2/[\sigma^2(n_1 + n_2 - 2)]}} \sim t_{n_1+n_2-2},$$

i.e.,

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

Thus, the $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\left[\bar{X} - \bar{Y} - S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{\alpha/2,n_1+n_2-2}, \bar{X} - \bar{Y} + S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{\alpha/2,n_1+n_2-2}\right]$$

(ii) Confidence for the ratio of two variances

Let $X_1, X_2, \ldots, X_{n_1}$ be an i.i.d. sample from the normal distribution $N(\mu_1, \sigma_1^2)$, and let $Y_1, Y_2, \ldots, Y_{n_2}$ be an i.i.d. sample from the normal distribution $N(\mu_2, \sigma_2^2)$. Assume that two samples are independent. We want to find a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2, i.e., we need to find a and b such that

$$P\left(a < \frac{\sigma_1^2}{\sigma_2^2} < b\right) = 1 - \alpha.$$

By the normal sampling theory, we know that $(n_1-1)S_1^2/\sigma_1^2 \sim \chi^2_{n_1-1}$ and $(n_2-1)S_2^2/\sigma_2^2 \sim \chi^2_{n_2-1}$ are independent. Thus, by the definition of F-distribution, we have

$$\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} \sim F_{n_1-1,n_2-1},$$

which is equivalent to

$$\frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \sim F_{n_1-1,n_2-1}.$$

Thus,

$$P\left(F_{1-\alpha/2,n_1-1,n_2-1} \leq \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \leq F_{\alpha/2,n_1-1,n_2-1}\right) = 1 - \alpha.$$

Thus, a $100(1-\alpha)\%$ confidence interval for σ_1^2/σ_2^2 is

$$\left[\frac{S_1^2/S_2^2}{F_{\alpha/2,n_1-1,n_2-1}}, \frac{S_1^2/S_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}\right].$$

(2) Use CLT for non-normal distributions

Example 5.6.4. *Let $X_1, X_2, \ldots, X_{n_1}$ be an i.i.d. sample from the Bernoulli distribution $b(1, p_1)$, and let $Y_1, Y_2, \ldots, Y_{n_2}$ be an i.i.d. sample from the Bernoulli distribution $b(1, p_2)$. Assume that two samples are independent.*

$$\bar{X} \longrightarrow N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$$

$$\bar{Y} \longrightarrow N\left(p_2, \frac{p_2(1-p_2)}{n_1}\right)$$

$$\frac{\sqrt{n_1}(\bar{X} - p_1)}{\sqrt{p_1(1-p_1)}} \longrightarrow N(0,1)$$

$$\frac{\sqrt{n_2}(\bar{Y} - p_2)}{\sqrt{p_2(1-p_2)}} \longrightarrow N(0,1)$$

$$\frac{\bar{X}-\bar{Y}-(p_1-p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \longrightarrow N(0,1)$$

$$\frac{\bar{X}-\bar{Y}-(p_1-p_2)}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n_1}+\frac{\bar{Y}(1-\bar{Y})}{n_2}}} \longrightarrow N(0,1)$$

$$P\left(\left|\frac{\bar{X}-\bar{Y}-(p_1-p_2)}{\sqrt{\frac{\bar{X}(1-\bar{X})}{n_1}+\frac{\bar{Y}(1-\bar{Y})}{n_2}}}\right| \leq z_{\alpha/2}\right) = 1-\alpha.$$

Thus, $100(1-\alpha)\%$ *confidence interval for* $p_1 - p_2$ *is*

$$\left(\bar{X}-\bar{Y}-z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n_1}+\frac{\bar{Y}(1-\bar{Y})}{n_2}}, \bar{X}-\bar{Y}+z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n_1}+\frac{\bar{Y}(1-\bar{Y})}{n_2}}\right).$$

5.6.3 Paired sample

Let (X_i, Y_i), $i = 1, 2, \ldots, n$ *be an i.i.d. sample from* (X, Y) *with* $E(X_1) = \mu_1, E(Y_1) = \mu_2$.

Assume that $D_i = X_i - Y_i$ *i=1,2,. . . ,n form an i.i.d. sample from* D *with a normal distribution* $N(\mu_1 - \mu_2, \sigma^2)$. *Then* $100(1-\alpha)\%$ *confidence interval for* $\mu_1 - \mu_2$ *is*

$$\left[\bar{X}-\bar{Y}-t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}, \bar{X}-\bar{Y}+t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}\right],$$

where S_D^2 *is the sample variance based on* $D_1, D_2, \ldots, D_n$.

6

Testing Statistical Hypotheses

6.1 Some basic concepts

- **Sample:** $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$
- **Joint distribution of the sample:**

$$\mathbf{X} \sim f(\mathbf{x}; \boldsymbol{\theta}), \quad \textit{where} \quad \mathbf{x} = (x_1, \ldots, x_n)^T, \quad \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$$

- **Assumptions:**

  (i) $f(\mathbf{x}; \boldsymbol{\theta})$ *is known except for the parameter* $\boldsymbol{\theta}$.

  (ii) Θ *contains at least two points.*

Definition 6.1.1. *A parametric hypothesis is an assertion about the unknown parameter* $\boldsymbol{\theta}$. *The null hypothesis is given by* $H_0 : \boldsymbol{\theta} \in \Theta_0 \subset \Theta$ *and the alternative hypothesis is given by* $H_1 : \boldsymbol{\theta} \in \Theta_1 = \Theta \backslash \Theta_0$.

Definition 6.1.2. *If* Θ_0 (Θ_1) *contains only one point, we say that* Θ_0 (Θ_1) *is simple; otherwise, composite. Thus, if a hypothesis is simple, the probability distribution of* $\mathbf{X}$ *is completely specified under the hypothesis.*

Example 6.1.1. *Let* $X \sim N(\mu, \sigma^2)$, *where* μ *and* σ^2 *are unknown.*

$$\Theta = \left\{ \boldsymbol{\theta} = (\mu, \sigma^2) : \mu \in \mathbb{R},\, \sigma^2 > 0 \right\}.$$

(i) *If* $H_0 : \mu \leq \mu_0,\, \sigma^2 > 0$ *versus* $H_1 : \mu > \mu_0, \sigma^2 > 0$, *where* μ_0 *is a known constant, then* Θ_0 *and* Θ_1 *are composite.*

(ii) *If* $H_0 : \mu = \mu_0,\, \sigma^2 > 0$ *versus* $H_1 : \mu \neq \mu_0,\, \sigma^2 > 0$ *where* μ_0 *is a known constant, then* Θ_0 *and* Θ_1 *are composite.*

(iii) *Let* $\sigma^2 = \sigma_0^2$ *be known. Then* $\Theta = \{\mu : \mu \in \mathbb{R}\}$. *If* $H_0 : \mu = \mu_0$ *versus* $H_1 : \mu \neq \mu_0$, *where* μ_0 *is a known constant, then* Θ_0 *is simple and* Θ_1 *is composite.*

Example 6.1.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from Bernoulli distribution $b(1,p)$ and let $p_0 \in (0,1)$ be known. Clearly, $\Theta = \{p : p \in (0,1)\}$.*

(i) *If $H_0 : p = p_0$ versus $H_1 : p \neq p_0$, then Θ_0 is simple and Θ_1 is composite.*

(ii) *If $H_0 : p \leq p_0$ versus $H_1 : p > p_0$, then Θ_0 and Θ_1 are composite.*

(iii) *If $H_0 : p \geq p_0$ versus $H_1 : p < p_0$, then Θ_0 and Θ_1 are composite.*

Definition 6.1.3. *Let $\mathbf{X} \sim f(\mathbf{x};\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. A subset C of $\mathbb{R}^n$ such that if $\mathbf{x} \in C$, then H_0 is rejected is called the critical (or rejection) region. That is,*

$$C = \{\mathbf{x} \in \mathbb{R}^n : H_0 \text{ is rejected if } \mathbf{x} \in C\}.$$

Definition 6.1.4. *Type I error, also known as a false positive, is the error of rejecting a null hypothesis when it is true, while Type II error, also known as a false negative is the error of not rejecting a null hypothesis when the alternative hypothesis is true.*

	State of Nature	
Hypothesis Testing Decision	H_0 is true	H_1 is true
Fail to reject H_0	Correct	Type II error
Reject H_0	Type I error	Correct

Remark 6.1.1. *Probabilities of making Type I error and Type II error are calculated by*

$$P(\textit{Type I error}) = P(C \,|\, H_0 \textit{ is true}) = P_{H_0}(C),$$
$$P(\textit{Type II error}) = P(C^c \,|\, H_1 \textit{ is true}) = P_{H_1}(C^c).$$

Remark 6.1.2. *To determine the rejection region C, we usually limit $P_{H_0}(C) \leq \alpha$ and minimize $P_{H_1}(C^c)$, where α is called the significance level of the test.*

Remark 6.1.3. *When we use computer software, the probability of Type I error is generally reported as the p-value.*

Remark 6.1.4. *Rejection region has the same structure as the H_1.*

Definition 6.1.5. *The power of a test is defined as*

$$P_{H_1}(C) = 1 - P(\textit{Type II error}).$$

6.2 Relationship between confidence intervals and two-sided hypothesis tests

The confidence interval for μ_0 consists of all those values μ for which $H_0 : \mu = \mu_0$ is accepted.

6.3 Sample from $N(\mu, \sigma^2)$

Example 6.3.1. *(Example 4.5.4) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$. We want to test*

$$H_0 : \mu = \mu_0 \ \textit{versus} \ H_1 : \mu > \mu_0.$$

Determine the rejection region with the significance level α.

Solution. *When the sample is taken from the normal distribution $N(\mu, \sigma^2)$, we know that*

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \textit{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

We also know that $\bar{X}$ and S^2 are independent. By the definition of t-distribution, we have

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

for any $\mu \in \mathbb{R}$. Thus, under H_0,

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}.$$

If the value of T is too large, we have to favor H_1 and reject H_0. Thus, the rejection region has the form $C = \{(X_1, \cdots, X_n) : T > c\}$. Since $P_{H_0}(C) = \alpha$, $c = t_{\alpha,n-1}$, which leads to

$$C = \{(X_1, \cdots, X_n) : T > t_{\alpha,n-1}\}.$$

Remark 6.3.1. *If we want to perform a two-sided test below*

$$H_0 : \mu = \mu_0 \quad versus \quad H_1 : \mu \neq \mu_0,$$

the rejection region is

$$C = \{(X_1, \cdots, X_n) : |T| > t_{\alpha/2, n-1}\}.$$

Remark 6.3.2. *If we want to perform a one-sided test below*

$$H_0 : \mu = \mu_0 \quad versus \quad H_1 : \mu < \mu_0,$$

the rejection region is

$$C = \{(X_1, \cdots, X_n) : T < t_{1-\alpha, n-1}\}.$$

Example 6.3.2. *(Example 4.6.2) Let $X_1, X_2, \ldots, X_{n_1}$ be an i.i.d. sample from the normal distribution $N(\mu_1, \sigma^2)$ and let $Y_1, Y_2, \ldots, Y_{n_2}$ be an i.i.d. sample from the normal distribution $N(\mu_2, \sigma^2)$. Assume that two samples are independent. We want to test*

$$H_0 : \mu_1 = \mu_2 \quad versus \quad H_1 : \mu_1 > \mu_2.$$

Determine the rejection region with the significance level α.

Solution. *We know from the normal sampling theory that*

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

for any $\mu_1, \mu_2 \in \mathbb{R}$, where S_p^2 is the pooled sample variance. Thus, under H_0,

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}.$$

If the value of T is too large, we have to favor H_1 and reject H_0. Thus, the rejection region has the form $C = \{(\mathbf{X}, \mathbf{Y}) : T > c\}$. Since $P_{H_0}(C) = \alpha$, $c = t_{\alpha, n_1+n_2-2}$, which leads to

$$C = \{(X_1, \cdots, X_n) : T > t_{\alpha, n_1+n_2-2}\}.$$

Remark 6.3.3. *If we want to perform a two-sided test below*

$$H_0 : \mu_1 = \mu_2 \quad versus \quad H_1 : \mu_1 \neq \mu_2,$$

the rejection region is

$$C = \{(X_1, \cdots, X_n) : |T| > t_{\alpha/2, n_1+n_2-2}\}.$$

Remark 6.3.4. *If we want to perform a one-sided test below*

$$H_0 : \mu_1 = \mu_2 \quad \textit{versus} \quad H_1 : \mu_1 < \mu_2,$$

the rejection region is

$$C = \{(X_1, \cdots, X_n) : T < t_{1-\alpha, n_1+n_2-2}\}.$$

6.4 Sample from a population whose distribution is unknown

Example 6.4.1. *(Example 4.5.3) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a population X with $E(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Assume that n is large. We want to test*

$$H_0 : \mu = \mu_0 \ \textit{versus} \ H_1 : \mu > \mu_0,$$

where μ_0 is known. Determine the rejection region with the significance level α.

Solution. *By the CLT, we know that*

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \longrightarrow N(0, 1)$$

as $n \to \infty$. We replace σ by its consistent estimator S and still have

$$\frac{\sqrt{n}(\bar{X} - \mu)}{S} \longrightarrow N(0, 1)$$

as $n \to \infty$. Thus, under H_0,

$$Z = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \longrightarrow N(0, 1)$$

as $n \to \infty$. If the value of Z is too large, we have to favor H_1 and reject H_0. Thus, the rejection region has the form $C = \{(X_1, \cdots, X_n) : Z > c\}$. Since $P_{H_0}(C) = \alpha$, $c = z_\alpha$, which leads to

$$C = \{(X_1, \cdots, X_n) : Z > z_\alpha\}.$$

Remark 6.4.1. *If we want to perform a two-sided test below*

$$H_0 : \mu = \mu_0 \quad \textit{versus} \quad H_1 : \mu \neq \mu_0,$$

the rejection region is

$$C = \{(X_1, \cdots, X_n) : |Z| > z_{\alpha/2}\}.$$

Remark 6.4.2. *If we want to perform a one-sided test below*

$$H_0 : \mu = \mu_0 \quad versus \quad H_1 : \mu < \mu_0,$$

the rejection region is

$$C = \{(X_1, \cdots, X_n) : Z < z_{1-\alpha}\}.$$

Remark 6.4.3. *The confidence interval for μ_0 will be*

$$\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right),$$

which is obtained by solving the inequality $|Z| < z_{\alpha/2}$.

Example 6.4.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution $b(1, p)$ and let $\hat{p}$ denote the sample mean. Assume that n is large.*

(1) *Suppose that we want to test $H_0 : p = p_0$ versus $H_1 : p > p_0$. We reject H_0 at α level if*

$$\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} > z_\alpha.$$

(2) *Suppose that we want to test $H_0 : p = p_0$ versus $H_1 : p < p_0$. We reject H_0 at α level if*

$$\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}} < -z_\alpha.$$

(3) *Suppose that we want to test $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. We reject H_0 at level α if*

$$\left|\frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}\right| > z_{\alpha/2}.$$

Example 6.4.3. *(Paired t-test) Let (X_i, Y_i), $i = 1, 2, \ldots, n$ be an i.i.d. sample from (X, Y) with $E(X_1) = \mu_1$ and $E(Y_1) = \mu_2$. We want to test $H_0 : \mu_1 = \mu_2$.*

Assume that $D_i = X_i - Y_i$ i=1,2,...,n form an i.i.d. sample from the normal distribution $N(\mu_1 - \mu_2, \sigma^2)$.

6.5 χ^2 test

6.5.1 Test variance of $N(\mu, \sigma^2)$

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. We want to test

$$H_0 : \sigma^2 = \sigma_0^2 \quad vs. \quad H_1 : \sigma^2 \neq \sigma_0^2,$$

where σ_0^2 is known.

It is known from the normal sampling theory that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Thus, $(n-1)S^2/\sigma_0^2 \sim \chi^2_{n-1}$ under H_0. Since

$$P\left(\frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{1-\alpha/2,n-1}\right) + P\left(\frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{\alpha/2,n-1}\right) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,$$

the rejection region C is

$$C = \left\{\mathbf{X} : S^2 < \frac{\sigma_0^2}{n-1}\chi^2_{1-\alpha/2,n-1} \text{ or } S^2 > \frac{\sigma_0^2}{n-1}\chi^2_{\alpha/2,n-1}\right\}.$$

Remark 6.5.1. *When $H_0 : \sigma^2 \leq \sigma_0^2$ and $H_0 : \sigma^2 \geq \sigma_0^2$, rejection regions are given by*

$$C = \left\{\mathbf{X} : S^2 > \frac{\sigma_0^2}{n-1}\chi^2_{\alpha,n-1}\right\}$$

and

$$C = \left\{\mathbf{X} : S^2 < \frac{\sigma_0^2}{n-1}\chi^2_{1-\alpha,n-1}\right\},$$

respectively.

Remark 6.5.2. *It is worth pointing out that the above method assumes that μ is unknown. If μ is known, the method should be modified. To be specific, when μ is known, we have that $\frac{X_i-\mu}{\sigma} \sim N(0,1)$ for $i = 1, 2 \ldots, n$, which leads to*

$$\sigma^{-2}\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n.$$

To test

$$H_0 : \sigma^2 = \sigma_0^2 \quad \textit{versus} \quad H_1 : \sigma^2 \neq \sigma_0^2,$$

where σ_0^2 is known, the rejection region C should be

$$C = \left\{\mathbf{X} : W < \sigma_0^2\chi^2_{1-\alpha/2,n} \text{ or } W > \sigma_0^2\chi^2_{\alpha/2,n}\right\},$$

where $W = \sum_{i=1}^{n}(X_i-\mu)^2$. Similarly, when we perform one-sided test $H_0 : \sigma^2 \leq \sigma_0^2$ or $H_0 : \sigma^2 \geq \sigma_0^2$,the corresponding rejection regions should also be modified.

6.5.2 Goodness of fit test (one-sample problem)

If $T_1, T_2, \ldots, T_n$ form an i.i.d. sample from the Bernoulli distribution $b(1, p_1)$, then CLT states that

$$\frac{\sqrt{n}(\bar{T}-p_1)}{\sqrt{p_1(1-p_1)}} \longrightarrow N(0,1)$$

as $n \to \infty$, *i.e.,*

$$Y = \frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \longrightarrow N(0,1)$$

where $X_1 = \sum_{i=1}^{n} T_i \sim b(n, p_1)$. *Thus,*

$$Q_1 = Y^2 \longrightarrow \chi_1^2$$

as $n \to \infty$. *Note that*

$$\begin{aligned} Q_1 &= Y^2 \\ &= \frac{(X_1 - np_1)^2}{np_1(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_1 - np_1)^2}{n(1-p_1)} \\ &= \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2}, \end{aligned}$$

where $p_2 = 1 - p_1$ *and* $X_2 = n - X_1$.

Suppose that we have a multinomial distribution. To be specific, let $(X_1, X_2, \ldots, X_k)$ *follows a multinomial distribution with parameters* $p_1, \ldots, p_k$. *Then*

$$Q_{k-1} = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \longrightarrow \chi_{k-1}^2 \quad as \ n \to \infty.$$

If we want to test

$$H_0 : p_i = p_{i0} \, (i = 1, 2, \ldots, k) \quad versus \quad H_1 : H_0 \, is \, not \, true,$$

where $p_{i0} \geq 0 \, (i = 1, 2, \ldots, k)$ *are known and satisfy* $\sum_{i=1}^{k} p_{i0} = 1$. *Under* H_0, $Q_{k-1} \to \chi_{k-1}^2$ *as* $n \to \infty$. *Thus, the rejection region has the form*

$$C = \left\{ \mathbf{X} : Q_{k-1} > \chi_{\alpha, k-1}^2 \right\}.$$

Example 6.5.1. *A die is rolled 60 times with the following outcomes*

Outcome	1	2	3	4	5	6
Frequency	13	19	11	8	5	4

Is the die fair at $\alpha = 5\%$?

Solution. *We use a multinomial distribution with $p_{10} = \cdots = p_{60} = 1/6$ to fit the data, i.e., we want to test*

$$H_0 : p_i = 1/6\,(i = 1, 2, \ldots, 6) \quad \textit{versus} \quad H_1 : H_0 \textit{ is not true},$$

Now we calculate Q_5 below

$$\begin{aligned} Q_5 &= \sum_{i=1}^{6} \frac{(x_i - np_{i0})^2}{np_{i0}} \\ &= \frac{(13-10)^2}{10} + \frac{(19-10)^2}{10} + \frac{(11-10)^2}{10} \\ &\quad + \frac{(8-10)^2}{10} + \frac{(5-10)^2}{10} + \frac{(4-10)^2}{10} \\ &= 15.6. \end{aligned}$$

When $\alpha = 5\%$, $\chi^2_{0.05,5} = 11.07$. Since $15.6 > 11.07$, H_0 is rejected. Thus, the die is biased at $\alpha = 5\%$.

Example 6.5.2. *(Example 4.7.2) Let $X_1, X_2, \ldots, X_{80}$ be an 80 observations on $[0, 1]$. We partition $(0, 1)$ by A_1, A_2, A_3, A_4, which are defined by*

$$\begin{aligned} A_1 &= \left\{ x : 0 < x \leq \frac{1}{4} \right\}, \\ A_2 &= \left\{ x : \frac{1}{4} < x \leq \frac{1}{2} \right\}, \\ A_3 &= \left\{ x : \frac{1}{2} < x \leq \frac{3}{4} \right\}, \\ A_4 &= \left\{ x : \frac{3}{4} < x < 1 \right\}. \end{aligned}$$

The observed frequencies falling into A_1, A_2, A_3, and A_4 are $6, 18, 20$, and 36, respectively. Let $p_i = P(A_i)$ for $i = 1, 2, 3, 4$. We want to test

$$H_0 : \textit{data follow Beta(2,1) distribution} \quad \textit{versus} \quad H_1 : H_0 \textit{ is false.}$$

Solution. *When H_0 is true, the pdf of X_1 is $2x_1$, $0 < x_1 < 1$, which leads to*

$$p_{10} = \int_0^{1/4} 2x dx = \frac{1}{16},$$
$$p_{20} = \int_{1/4}^{1/2} 2x dx = \frac{3}{16},$$
$$p_{30} = \int_{1/2}^{3/4} 2x dx = \frac{5}{16},$$
$$p_{40} = \int_{3/4}^{1} 2x dx = \frac{7}{16}.$$

Thus,

$$\begin{aligned} Q_3 &= \sum_{i=1}^{4} \frac{(X_i - np_{i0})^2}{np_{i0}} \\ &= \frac{(6-5)^2}{5} + \frac{(18-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(36-35)^2}{35} \\ &= \frac{64}{35} \\ &= 1.83. \end{aligned}$$

When $\alpha = 2.5\%$, $\chi^2_{0.025,3} = 9.348$. Since $1.83 < 9.348$, we cannot reject H_0 at $\alpha = 2.5\%$.

6.5.3 Test for homogeneity (two-sample problem)

We have two multinomial distributions below:

Category	1	2	$\cdots$	k
Observed frequency	X_{11}	X_{21}	$\cdots$	X_{k1}
Pr.	p_{11}	p_{21}	$\cdots$	p_{k1}

Category	1	2	$\cdots$	k
Observed frequency	X_{12}	X_{22}	$\cdots$	X_{k2}
Pr.	p_{12}	p_{22}	$\cdots$	p_{k2}

We want to see if there is any difference between the two distributions:

$$H_0 : p_{i1} = p_{i2}, \quad i = 1, 2, \ldots, k \quad \textit{versus} \quad H_1 : H_0 \textit{ is false.}$$

Under H_0, $p_{i1} = p_{i2}$ *is estimated by* $(X_{i1} + X_{i2})/(n_1 + n_2)$. *We reject* H_0 *at level* α *if*

$$\sum_{j=1}^{2} \sum_{i=1}^{k} \frac{\left(X_{ij} - n_j \frac{X_{i1}+X_{i2}}{n_1+n_2}\right)^2}{n_j \frac{X_{i1}+X_{i2}}{n_1+n_2}} > \chi^2_{\alpha,k-1},$$

where $\sum_{i=1}^{k} X_{ij} = n_j$ *for* $j = 1, 2$.

Example 6.5.3. *To compare the effectiveness of two diets A and B, 150 infants were included in a study. Diet A was given to 80 randomly selected infants and diet B was given to the other 70 infants. At a later time, the health of each infant was observed and classified into one of three categories "excellent", "average" and "poor".*

Health under Two Different Diets

	Excellent	Average	Poor	Sample size
Diet A	37	24	19	80
Diet B	17	33	20	70
Total	54	57	39	150

We wish to test the null hypothesis that there is no difference between the quality of the two diets

Solution. *The expected frequencies are given by*

	Excellent	Average	Poor
Diet A	$\frac{80\times54}{150}$	$\frac{80\times57}{150}$	$\frac{80\times39}{150}$
Diet B	$\frac{70\times54}{150}$	$\frac{70\times57}{150}$	$\frac{70\times39}{150}$

Thus,

$$\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = 2.335+1.347+0.156+2.668+1.54+0.178 \approx 8.224.$$

Since $\chi^2_{0.05,k-1} = \chi^2_{0.05,2} \approx 5.9915 < 8.224$, *we reject* H_0 *at* $\alpha = 5\%$. *That is, there is a significant difference between the quality of the two diets.*

6.5.4 Test for independence

Let $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ *be an i.i.d. sample from the random vector* (X,Y). *We want to test whether* X *and* Y *are independent.*

We partition the space of X*'s into* $A_1,A_2,\ldots,A_a$ *and partition the space of* Y*'s into* $B_1,B_2,\ldots,B_b$. *Let*

$$p_{ij} = P(A_iB_j) = P(X\in A_i, Y\in B_j)$$

for $i=1,2,\ldots,a, j=1,2,\ldots,b$. *We want to test*

$$H_0: p_{ij} = p_{i\cdot}p_{\cdot j} \ \text{ versus } \ H_1: H_0 \text{ is not true,}$$

where $p_{i\cdot} = \sum_{j=1}^b p_{ij}$ *and* $p_{\cdot j} = \sum_{i=1}^a p_{ij}$ *for* $i=1,2,\ldots,a$, $j=1,2,\ldots,b$. *We reject* H_0 *at* α *if*

$$\sum_{i=1}^a\sum_{j=1}^b \frac{[X_{ij}-n(X_{i\cdot}/n)(X_{\cdot j}/n)]^2}{n(X_{i\cdot}/n)(X_{\cdot j}/n)} > \chi^2_{\alpha,(a-1)(b-1)},$$

where X_{ij} *denotes the observed frequency in* $A_i\cap B_j$, $X_{i\cdot} = \sum_{j=1}^b X_{ij}$, *and* $X_{\cdot j} = \sum_{i=1}^a X_{ij}$ *for* $i=1,2,\ldots,a$ *and* $j=1,2,\ldots,b$.

Remark 6.5.3. *Under* H_0, *the number of parameters is* $a-1+b-1 = a+b-2$. *Thus, the degrees of freedom is equal to* $df = ab-1-(a+b-2) = (a-1)(b-1)$.

Remark 6.5.4. *The* a *and* b *should be chosen appropriately such that* $\min(X_{ij}) \geq 5$.

6.6 Likelihood ratio test

- *We want to test*

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \textit{versus} \quad H_1 : \boldsymbol{\theta} \in \Theta_1 = \Theta \backslash \Theta_0.$$

- *We reject H_0 if and only if $\Lambda(\mathbf{x}) < c$, where c is a constant and*

$$\Lambda(\mathbf{x}) = \frac{\max\limits_{\boldsymbol{\theta}\in\Theta_0} \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}\in\Theta} \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})}$$

Example 6.6.1. *(Example 6.3.1) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the exponential distribution with a pdf $f(x;\theta) = \theta^{-1}e^{-x/\theta}$, $x > 0$. We want to test*

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta \neq \theta_0.$$

where θ_0 is known. Determine the rejection region at level α.

Solution. *Since the likelihood function of observed data $\mathbf{X} = \mathbf{x}$ is*

$$L(\theta) = \theta^{-n} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{n} x_i\right),$$

thus the log-likelihood function has the form

$$l(\theta) = -n\ln(\theta) - \frac{1}{\theta}\sum_{i=1}^{n} x_i.$$

Taking the derivative of $l(\theta)$ with respect to θ, we have

$$l'(\theta) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i.$$

Solving $l'(\theta) = 0$ yields that $\hat{\theta} = \bar{x}$. That is, $\bar{X}$ is the MLE of θ because $l''(\theta) < 0$ when $\theta = \bar{X}$.

On the other hand, when H_0 is true, we have

$$L(\theta_0) = \theta_0^{-n} \exp\left(-\frac{1}{\theta_0}\sum_{i=1}^{n} x_i\right).$$

Thus,

$$\Lambda(\mathbf{x}) = \frac{\theta_0^{-n} \exp\left(-\frac{1}{\theta_0}\sum_{i=1}^{n} x_i\right)}{\bar{x}^{-n}\exp(-n)} < c$$

if and only if

$$\left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left(-\frac{n\bar{x}}{\theta_0}\right) < c^*,$$

where $c^* = ce^{-n}$. *To solve the above inequality, we consider a function defined by* $g(t) = t^n e^{-nt}$ *for* $t > 0$. *Then* $g(t)$ *attains its maximum at* $t = 1$ *because*

$$g'(t) = nt^{n-1}e^{-nt} - nt^n e^{-nt} = nt^{n-1}(1-t)e^{-nt}$$

Thus, $\Lambda(\mathbf{x}) < c$ *is equivalent to* $\bar{x}/\theta_0 \le c_1$ *or* $\bar{x}/\theta_0 \ge c_2$. *Since*

$$\frac{2}{\theta_0}\sum_{i=1}^n X_i \sim \chi^2_{2n},$$

we reject H_0 *if* $\frac{2}{\theta_0}\sum_{i=1}^n X_i \le \chi^2_{1-\alpha/2,2n}$ *or* $\frac{2}{\theta_0}\sum_{i=1}^n X_i \ge \chi^2_{\alpha/2,2n}$.

Example 6.6.2. *(Example 6.3.2) Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from the normal distribution* $N(\mu, \sigma^2)$, *where* $\sigma^2 > 0$ *is known. We want to test*

$$H_0 : \mu = \mu_0 \ \textit{versus}\ H_1 : \mu \neq \mu_0.$$

Determine the rejection region at level α.

Solution. *For the observed data* $\mathbf{X} = \mathbf{x}$, *the likelihood function under* H_0 *is*

$$L(\mu_0) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2\right).$$

Under Θ,

$$L(\mu) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right),$$

the MLE of $\mu = \bar{x}$. *Since*

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i - \mu_0 + \mu_0 - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - \mu_0)^2 + 2(\mu_0 - \bar{x})(n\bar{x} - n\mu_0) + n(\mu_0 - \bar{x})^2 \\ &= \sum_{i=1}^n (x_i - \mu_0)^2 - n(\bar{x} - \mu_0)^2,\end{aligned}$$

we obtain

$$\begin{aligned}L(\bar{x}) &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2\right) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu_0)^2\right) \exp\left(\frac{n(\bar{x} - \mu_0)^2}{2\sigma^2}\right),\end{aligned}$$

which leads to

$$\Lambda(\mathbf{x}) = \frac{L(\mu_0)}{L(\bar{x})} = \exp\left(-\frac{1}{2\sigma^2}n(\bar{x}-\mu_0)^2\right).$$

Thus, $\Lambda(\mathbf{x}) < c$ *if and only if*

$$\frac{n}{2\sigma^2}(\bar{x}-\mu_0)^2 > c_1$$

or equivalently

$$\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > c_2,$$

where $c_2 = 2c_1$. *Under* H_0, *we know*

$$\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1.$$

Thus, we reject H_0 *if*

$$\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)^2 > \chi^2_{\alpha,1}.$$

Example 6.6.3. *Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from the normal distribution* $N(\mu, \sigma^2)$, *where* $\sigma^2 > 0$ *is unknown. We want to test*

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0.$$

Determine the rejection region at level α.

Solution. *For the observed data* $\mathbf{X} = \mathbf{x}$, *the likelihood function under* H_0 *is*

$$L(\mu_0, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_0)^2\right).$$

The MLE of σ^2 *is* $\hat{\sigma}^2 = (1/n)\sum_{i=1}^{n}(x_i-\mu_0)^2$. *Under* $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$, *the MLEs of* μ *and* σ^2 *are*

$$\hat{\hat{\mu}} = \bar{x},$$
$$\hat{\hat{\sigma}}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

respectively, which implies that

$$L(\hat{\hat{\mu}}, \hat{\hat{\sigma}}^2) = \left(\frac{1}{2\pi\hat{\hat{\sigma}}^2}\right)^{n/2} e^{-n/2}.$$

Thus,

$$\Lambda(\mathbf{x}) = \frac{L(\mu_0, \hat{\sigma}^2)}{L(\hat{\hat{\mu}}, \hat{\hat{\sigma}}^2)} = \left(\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \mu_0)^2} \right)^{n/2}.$$

Applying the fact that

$$\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2 = \sum_{i=1}^n (x_i - \mu_0)^2,$$

we obtain

$$\begin{aligned} \Lambda(\mathbf{x}) < c &\iff \frac{1}{\Lambda(\mathbf{x})} > \frac{1}{c} \\ &\iff \left(1 + \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)^{n/2} > \frac{1}{c} \\ &\iff \frac{n(\bar{x} - \mu_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} > \left(\frac{1}{c} \right)^{2/n} - 1 \end{aligned}$$

Under $H_0 : \sqrt{n}(\bar{X} - \mu_0)/\sigma \sim N(0,1)$. *Furthermore, it is known from the normal sampling theory that*

$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi^2_{n-1}$$

is independent of $\bar{X}$. *By the definition of t-distribution, we obtain*

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/(n-1)}} \sim t_{n-1}.$$

Thus, we reject H_0 *if* $|T| > t_{\alpha/2, n-1}$.

Theorem 6.6.1. *Let* $X_1, X_2, \ldots, X_n$ *be i.i.d. with pdf* $f(\mathbf{x}; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$. *Under some regularity conditions of* $f(\mathbf{x}; \boldsymbol{\theta})$, *we have*

$$-2 \ln[\Lambda(\mathbf{x})] \xrightarrow{d} \chi^2_q$$

where q *is the difference of the number of independent parameters in* Θ *and the number of independent parameters in* Θ_0

Example 6.6.4. *Let* $(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})$ *be an i.i.d. sample from* (X_1, X_2) *with a multinomial distribution* $(1, p_1, p_2, 1 - p_1 - p_2)$, *i.e., the joint pmf of* (X_1, X_2) *is given by*

$$f(\mathbf{x}; p_1, p_2) = p_1^{x_1} p_2^{x_2} (1 - p_1 - p_2)^{1 - x_1 - x_2},$$

where $x_1, x_2 = 0, 1$ *and*

$$\Theta = \{(p_1, p_2) : 0 < p_1 < 1,\ 0 < p_2 < 1,\ p_1 + p_2 < 1\}.$$

We want to test

$$H_0 : p_1 = p_2 \quad \textit{versus} \quad H_1 : p_1 \neq p_2.$$

Under Θ, *the MLEs of* p_1 *and* p_2 *are given by*

$$\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \quad \textit{and} \quad \hat{p}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i},$$

which leads to

$$L(\hat{p}_1, \hat{p}_2) = \hat{p}_1^{n\hat{p}_1}\hat{p}_2{}^{n\hat{p}_2}(1 - \hat{p}_1 - \hat{p}_2)^{n - n\hat{p}_1 - n\hat{p}_2}.$$

On the other hand, the likelihood function under H_0 *becomes*

$$L(p_1, p_1) = p_1^{\sum_{i=1}^{n}(x_{1i}+x_{2i})}(1 - 2p_1)^{n - \sum_{i=1}^{n}(x_{1i}+x_{2i})},$$

which is maximized at

$$\hat{\hat{p}}_1 = \frac{\hat{p}_1 + \hat{p}_2}{2}.$$

Thus,

$$L(\hat{\hat{p}}_1, \hat{\hat{p}}_2) = \left(\frac{\hat{p}_1 + \hat{p}_2}{2}\right)^{n(\hat{p}_1+\hat{p}_2)} (1 - \hat{p}_1 - \hat{p}_2)^{n(1-\hat{p}_1-\hat{p}_2)},$$

$$\Lambda^{-1} = \frac{L(\hat{p}_1, \hat{p}_2)}{L(\hat{\hat{p}}_1, \hat{\hat{p}}_2)} = \left(\frac{2\hat{p}_1}{\hat{p}_1 + \hat{p}_2}\right)^{n\hat{p}_1} \left(\frac{2\hat{p}_2}{\hat{p}_1 + \hat{p}_2}\right)^{n\hat{p}_2}.$$

We reject H_0 *if*

$$2\ln[\Lambda^{-1}] > \chi^2_{\alpha,1}.$$

7

Sufficiency

7.1 Introduction

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a pmf or pdf $f(x;\theta)$ and let T be a statistic. Then

(i) *T is an unbiased estimator of θ if $E(T) = \theta$, and*

(ii) *if T_1 and T_2 are unbiased estimators of θ with $\mathrm{Var}(T_1) < \mathrm{Var}(T_2)$, then we would choose T_1 over T_2.*

For example, let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$. Then

(i) the sample mean $\bar{X}$ and X_1 are unbiased estimators of μ. We choose $\bar{X}$ because

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} < \mathrm{Var}(X_1) = \sigma^2 \quad \textit{if} \quad n > 1.$$

(ii) the sample variance $\hat{\sigma}_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of σ^2 with

$$\begin{aligned}
\mathrm{Var}(\hat{\sigma}_n^2) &= \mathrm{Var}\left(\frac{\sigma^2}{n-1}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}\right) \\
&= \mathrm{Var}\left(\frac{\sigma^2}{n-1}\frac{(n-1)S^2}{\sigma^2}\right) \\
&= \frac{\sigma^4}{(n-1)^2}\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) \\
&= \frac{\sigma^4}{(n-1)^2}2(n-1) \quad \left(b/c \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}\right) \\
&= \frac{2\sigma^4}{n-1}.
\end{aligned}$$

We would choose $\hat{\sigma}_n^2$ *over* $\hat{\sigma}_{n-1}^2$, *where*

$$\hat{\sigma}_{n-1}^2 = \frac{1}{n-2}\sum_{i=1}^{n-1}(X_i - \bar{X}_*)^2,$$
$$\bar{X}_* = \frac{1}{n-1}\sum_{i=1}^{n-1} X_i.$$

Question 1: *Among the class of all unbiased estimators of* θ, *how can we find an estimator such that its variance is the smallest?*

Question 2: *Except for the variance, are there other criteria to judge estimators?*

7.2 Uniformly minimum variance unbiased estimator (UMVUE) and minimax estimator

7.2.1 UMVUE

Definition 7.2.1. *Let* $\mathcal{U}$ *be the class of all unbiased estimators* T *of* $\theta \in \Theta$ *such that* $E_\theta(T^2) < \infty$ *for all* $\theta \in \Theta$. *An estimator* $T_0 \in U$ *is called a uniformly minimum variance unbiased estimator (UMVUE) of* θ *if*

$$E_\theta\left[(T_0 - \theta)^2\right] \leq E_\theta\left[(T - \theta)^2\right]$$

for all $\theta \in \Theta$ *and every* $T \in U$.

Remark 7.2.1. *Let* $X_1, X_2, \ldots, X_n$ *be an i.i.d. sample from the pdf/pmf* $f(x;\theta)$. *If the UMVUE of* θ *exists, then it is a symmetric function of* $X_1, X_2, \ldots, X_n$.

Example 7.2.1. *Let* $X \sim \mathcal{P}(\theta)$. *Then* X *is the UMVUE of* θ.

Proof. This example can be considered an i.i.d. sample from the Poisson distribution $\mathcal{P}(\theta)$ with a sample size of 1. We prove the desired result by showing two parts below.

(i) Since $X \sim \mathcal{P}(\theta)$, we have that $E(X) = \theta$, which means that X is an unbiased estimator of θ.

(ii) Let $g(X)$ be an unbiased estimator of 0. Then $T = X + g(X)$ is an unbiased estimator of θ because

$$E(T) = E[X + g(X)] = E(X) + E[g(X)] = \theta + 0 = \theta.$$

Note that

$$E_\theta[g(X)] = 0 \iff \sum_{i=0}^{\infty} g(i)\frac{\theta^i}{i!}e^{-\theta} = 0 \quad \text{for all} \quad \theta$$
$$\iff g(i) = 0 \quad \text{for} \quad i = 0, 1, 2, \ldots \quad \text{(Why?)}$$

Hence X is the UMVUE of θ. □

7.2.2 Minimax estimator

Suppose that $X_1, X_2, \ldots, X_n$ form an i.i.d. from a pdf/pmf $f(x;\theta)$, where $\theta \in \Theta$. Let $Y = U(X_1, X_2, \ldots, X_n)$ be a statistic and $\delta(y)$ be a point estimate of θ, where $y = U(x_1, x_2, \ldots, x_n)$ be the observed value of Y. Furthermore, let

$$\mathcal{A} = \textit{the set of all actions or decisions open to a statistician.}$$

Definition 7.2.2. *A decision function/a decision rule δ is a statistic that takes values in $\mathcal{A}$, i.e., δ maps $\mathbb{R}^n$ into $\mathcal{A}$: $\mathbb{R}^n \xrightarrow{\delta} \mathcal{A}$.*

Definition 7.2.3. *A nonnegative function L that maps $\Theta \times \mathcal{A}$ into $\mathbb{R}$ is called a loss function, i.e., $L(\theta, \delta(y))$ represents the loss to the statistician if he takes action $\delta(y)$ when θ is the true parameter.*

Example 7.2.2. *Loss functions:* $L(\theta, \delta) = |\theta - \delta|; \quad L(\theta, \delta) = |\theta - \delta|^2$.

Definition 7.2.4. *$R(\theta, \delta) = E_\theta[L(\theta, \delta(Y))]$ is called the risk function associated with δ.*

Definition 7.2.5. *The minimax principle is to choose δ^* so that*

$$\max_\theta R(\theta, \delta^*) \leq \max_\theta R(\theta, \delta)$$

for all $\delta \in \mathcal{A}$.

Example 7.2.3. *Let X have a Bernoulli distribution $b(1, \theta)$, where $\theta \in \Theta = \{\frac{1}{4}, \frac{1}{2}\}$ and $\mathcal{A} = \{a_1, a_2\}$. Let the loss function L be defined by*

	a_1	a_2
$\theta = 1/4$	1	4
$\theta = 1/2$	3	2

That is,

$$L(1/4, a_1) = 1, \quad L(1/4, a_2) = 4,$$
$$L(1/2, a_1) = 3, \quad L(1/2, a_2) = 2.$$

Decision rules are given by

$$\delta_1(0) = a_1,\ \delta_2(0) = a_1,\ \delta_3(0) = a_2,\ \delta_4(0) = a_2,$$
$$\delta_1(1) = a_1\ , \delta_2(1) = a_2,\ \delta_3(1) = a_1,\ \delta_4(1) = a_2.$$

The risk function is presented in the table below:

i	$R(1/4, \delta_i)$	$R(1/2, \delta_i)$	$\max_\theta R(\theta, \delta_i)$	$\min_i \max_\theta R(\theta, \delta_i)$
1	*1*	*3*	*3*	
2	*7/4*	*5/2*	*5/2*	**5/2**
3	*13/4*	*5/2*	*13/4*	
4	*4*	*2*	*4*	

The values of the risk function are calculated by using its definition. For example,

$$\begin{aligned}
R(1/4, \delta_1) &= E[L(1/4, \delta_1(X))] \\
&= L(1/4, \delta_1(0)) \left(1 - \frac{1}{4}\right) + L(1/4, \delta_1(1)) \frac{1}{4} \\
&= L(1/4, a_1) \frac{3}{4} + L(1/4, a_1) \frac{1}{4} \\
&= 1 * \frac{3}{4} + 1 * \frac{1}{4} \\
&= 1, \\
R(1/4, \delta_2) &= E[L(1/4, \delta_2(X))] \\
&= L(1/4, \delta_2(0)) \frac{3}{4} + L(1/4, \delta_2(1)) \frac{1}{4} \\
&= L(1/4, a_1) \frac{3}{4} + L(1/4, a_2) \frac{1}{4} \\
&= \frac{3}{4} + 4 * \frac{1}{4} \\
&= \frac{7}{4}.
\end{aligned}$$

Thus, the minimax solution is

$$\delta^*(x) = \delta_2(x) = \begin{cases} a_1, & \text{if } x = 0, \\ a_2, & \text{if } x = 1. \end{cases}$$

7.3 Sufficient Statistics

7.3.1 Definition and examples

Definition 7.3.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a pdf $f(x;\theta)$ or pmf $p(x;\theta)$, where $\theta \in \Theta$. Let $T = T(X_1, X_2, \ldots, X_n)$ be a statistic. Then T is a sufficient statistic for θ if and only if*

$$P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n \,|\, T = t) \qquad (*)$$

does not depend on θ for all $\theta \in \Theta$

Remark 7.3.1. *The $(*)$ is equivalent to the joint conditional pdf/pmf.*

Remark 7.3.2. *Θ does not have to be in 1-dimensional space.*

Remark 7.3.3. *If T is sufficient for θ, any one-to-one function of T is also sufficient for θ.*

Example 7.3.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution $b(1,\theta)$, where $0 < \theta < 1$. Then $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.*

Proof. Since the pmf of the Bernoulli distribution $b(1,\theta)$ is given by

$$p(x;\theta) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1,$$

the joint pmf of the $(X_1, X_2, \ldots, X_n)$ is equal to

$$p(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^n p(x_i;\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i}.$$

To evaluate the following conditional probability

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|\, T = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^n X_i = t)}{P(T = t)},$$

we use the fact that $T \sim b(n,\theta)$ and discuss two cases below:

(i) When $\sum_{i=1}^n x_i \neq t$, then $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|\, T = t) = 0$.

(ii) When $\sum_{i=1}^n x_i = t$, then

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|\, T = t) &= \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)}{P(T = t)} \\ &= \frac{\prod_{i=1}^n P(X_i = x_i)}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{1}{\binom{n}{t}}. \end{aligned}$$

Combining (i) and (ii) concludes that T is a sufficient statistic for θ because $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \,|\, T = t)$ does not depend on θ for all $\theta \in (0, 1)$. This proves that $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ. □

Example 7.3.2. *(Example 7.2.2)*

Example 7.3.3. *(Example 7.2.3) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the shifted exponential distribution whose pdf is given by*

$$f(x;\theta) = e^{-(x-\theta)} I(x > \theta) = \begin{cases} 0, & x \leq \theta, \\ e^{-(x-\theta)}, & x > \theta. \end{cases}$$

Let $Y_1 = \min(X_1, X_2, \ldots, X_n)$. Then Y_1 is a sufficient statistic for θ.

Proof. The joint pdf of $(X_1, X_2, \ldots, X_n)$ is equal to

$$\begin{aligned} g(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^n f(x_i;\theta) \\ &= \prod_{i=1}^n e^{-(x_i-\theta)} I(x_i > \theta) \\ &= \exp\left(n\theta - \sum_{i=1}^n x_i\right) I(\min(x_i) > \theta). \end{aligned}$$

Furthermore, using (4.4.2) on page 255 will obtain the pdf of Y_1 given by

$$g_{Y_1}(y_1) = ne^{-n(y_1-\theta)} I(y_1 > \theta) = ne^{n\theta - n\min(x_i)} I(\min(x_i) > \theta).$$

Thus, the conditional pdf of the sample given Y_1 has the form

$$g(x_1, x_2, \ldots, x_n | Y_1 = y_1) = \frac{e^{-\sum_{i=1}^n x_i}}{ne^{-n\min(x_i)}},$$

which does not depend on θ. This proves that $Y_1 = \min(X_i)$ is a sufficient statistic for θ. □

Example 7.3.4. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution $\mathcal{P}(\theta)$. Then $T = \sum_{i=1}^n X_i$ is a sufficient statistic for θ.*

Proof. Given $T = t$, the conditional pmf of $(X_1, \ldots, X_n)$ is given by

$$P(X_1 = x_1, \ldots, X_n = x_n \,|\, T = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)}.$$

To evaluate this conditional probability, we use the fact that $T \sim \mathcal{P}(n\theta)$ and discuss two cases below:

(i) When $\sum_{i=1}^n x_i \neq t$, then $P(X_1 = x_1, \ldots, X_n = x_n \,|\, T = t) = 0$.

(ii) When $\sum_{i=1}^n x_i = t$, then

$$\begin{aligned} P(X_1 = x_1, \ldots, X_n = x_n \,|\, T = t) &= \frac{\prod_{i=1}^n \frac{\theta^{x_i}}{x_i!} e^{-\theta}}{\frac{(n\theta)^t}{t!} e^{-n\theta}} \\ &= \frac{t!}{\prod_{i=1}^n x_i!} \frac{\theta^{\sum_{i=1}^n x_i}}{n^t \theta^t} \\ &= \frac{t!}{\prod_{i=1}^n x_i!} \left(\frac{1}{n}\right)^t \\ &= \frac{t!}{\prod_{i=1}^n x_i!} \prod_{i=1}^n \left(\frac{1}{n}\right)^{x_i}, \end{aligned}$$

which is the pmf of multinomial distribution $(t, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$.

Combining (i) and (ii) concludes that T is a sufficient statistic for θ because $P(X_1 = x_1, \ldots, X_n = x_n \,|\, T = t)$ does not depend on θ for all $\theta > 0$. □

Example 7.3.5. *Let X_1 and X_2 be i.i.d. from the Poisson distribution $\mathcal{P}(\theta)$. Then $X_1 + 2X_2$ is NOT a sufficient statistic for θ.*

Proof. We show that $P(X_1 = x_1, X_2 = x_2 \,|\, X_1 + 2X_2 = t)$ depends on θ in

general. Indeed,

$$\begin{aligned}
&P(X_1 = 0, X_2 = 1 \,|\, X_1 + 2X_2 = 2) \\
&= \frac{P(X_1 = 0, X_2 = 1, X_1 + 2X_2 = 2)}{P(X_1 + 2X_2 = 2)} \\
&= \frac{P(X_1 = 0, X_2 = 1)}{P(X_1 = 0, X_2 = 1) + P(X_1 = 2, X_2 = 0)} \\
&= \frac{P(X_1 = 0)P(X_2 = 1)}{P(X_1 = 0)P(X_2 = 1) + P(X_1 = 2)P(X_2 = 0)} \\
&= \frac{P(X_1 = 0)P(X_1 = 1)}{P(X_1 = 0)P(X_1 = 1) + P(X_1 = 2)P(X_1 = 0)} \\
&= \frac{P(X_1 = 1)}{P(X_1 = 1) + P(X_1 = 2)} \\
&= \frac{\theta e^{-\theta}}{\theta e^{-\theta} + \theta^2 e^{-\theta}/2!} \\
&= \frac{2}{2 + \theta},
\end{aligned}$$

which depends on θ. Thus, $X_1 + 2X_2$ is not a sufficient statistic for θ. □

7.3.2 The Factorization Criterion (Neyman)

Theorem 7.3.1. *(Theorem 7.2.1.) Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a distribution with a pmf or pdf $f(x;\theta)$, where $\theta \in \Theta$. Then $T = T(X_1, X_2, \ldots, X_n)$ is sufficient for θ* **iff**

$$\prod_{i=1}^{n} f(x_i;\theta) = k_1\big(t(x_1, x_2, \ldots, x_n);\theta\big)\, k_2(x_1, x_2, \ldots, x_n),$$

where $k_1 \geq 0$ depends on t and θ and $k_2 \geq 0$ depends on the sample only.
The proof can be seen on page 422.

Remark 7.3.4. *Theorem also holds when θ is a vector.*

Remark 7.3.5. *Sufficient statistic for θ may be a vector while $\theta \in \mathbb{R}$.*

Example 7.3.6. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution $b(1,\theta)$. Then $T = \sum_{i=1}^{n} X_i$ is sufficient for θ.*

Proof. Since the pmf of the Bernoulli distribution $b(1,\theta)$ is given by

$$p(x;\theta) = \theta^x (1-\theta)^{1-x}, \quad x = 0, 1,$$

the joint pmf of $(X_1, X_2, \ldots, X_n)$ is equal to

$$\begin{aligned}\prod_{i=1}^{n} p(x_i;\theta) &= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\ &= (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i} \\ &= k_1\big(t,\theta\big)k_2(x_1,\ldots,x_n)\end{aligned}$$

where

$$\begin{aligned}&k_1(t,\theta) = (1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^t, \\ &t = \sum_{i=1}^{n} x_i, \\ &k_2(x_1,\ldots,x_n) = 1.\end{aligned}$$

This proves that $T = \sum_{i=1}^n X_i$ is sufficient for θ. □

Example 7.3.7. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, where σ^2 is known. Then $\bar{X}$ is sufficient for μ.*

Proof. Since the pdf of the normal distribution $N(\mu, \sigma^2)$ is given by

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R},$$

using the fact that

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2$$

can write the joint pdf of $(X_1, X_2, \ldots, X_n)$ below:

$$\begin{aligned}\prod_{i=1}^{n} f(x_i;\mu) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}\right)\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\right) \\ &= k_1\big(t,\mu\big)k_2(x_1,\ldots,x_n),\end{aligned}$$

where

$$\begin{aligned}&k_1(t,\mu) = \exp\left(-\frac{n(t-\mu)^2}{2\sigma^2}\right), \\ &t = \bar{x}, \\ &k_2(x_1,\ldots,x_n) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\right).\end{aligned}$$

This proves that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sufficient for μ. □

Example 7.3.8. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Beta distribution* Beta$(\theta, 1)$, *where $\theta > 0$ is a parameter. Then $T = \prod_{i=1}^{n} X_i$ is sufficient for θ.*

Proof. Since the pdf of the Beta distribution Beta$(\theta, 1)$ is given by

$$f(x;\theta) = \theta x^{\theta-1}, \quad x \in (0,1),$$

the joint pdf of $(X_1, X_2, \ldots, X_n)$ is equal to

$$\prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1} = k_1\big(t,\theta\big)k_2(x_1,\ldots,x_n),$$

where

$$\begin{aligned} k_1(t,\theta) &= \theta^n t^{\theta-1}, \\ t &= \prod_{i=1}^{n} x_i, \\ k_2(x_1,\ldots,x_n) &= 1. \end{aligned}$$

This proves that $T = \prod_{i=1}^{n} X_i$ is sufficient for θ. □

Example 7.3.9. *(Example 7.2.6 on page 425–the shifted exponential distribution)*

Example 7.3.10. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the discrete uniform distribution at points $\{1, 2, \ldots, N\}$, where N is a positive and unknown integer. Then $T = \max(X_i)$ is sufficient for N.*

Proof. Since the pmf of the uniform distribution at points $\{1, 2, \cdots, N\}$ is given by

$$p(x;N) = \frac{1}{N}, \quad x \in \{1, 2, \cdots, N\},$$

the joint pmf of $(X_1, X_2, \ldots, X_n)$ is equal to

$$\begin{aligned} \prod_{i=1}^{n} p(x_i;N) &= \prod_{i=1}^{n} \frac{1}{N} I(x_i \in \{1, 2, \cdots, N\}) \\ &= \frac{1}{N^n}\phi(\max(x_i), N)\,\phi(1, \min(x_i)) \\ &= k_1\big(t, N\big)k_2(x_1,\ldots,x_n), \end{aligned}$$

where

$$
\begin{aligned}
&k_1(t,N) = \frac{1}{N^n}\phi(t,N),\\
&t = \max(x_i),\\
&k_2(x_1,\ldots,x_n) = \phi(1,\min(x_i)),\\
&\phi(a,b) = \begin{cases} 1, & \text{if } b \geq a,\\ 0, & \text{if } b < a. \end{cases}
\end{aligned}
$$

This proves that $T = \max(X_i)$ is sufficient for N. □

Example 7.3.11. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$, where μ and σ^2 are unknown. Then $\mathbf{T} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is jointly sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$.*

Proof. Since the pdf of the normal distribution $N(\mu, \sigma^2)$ is given by

$$
f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R},
$$

the joint pdf of $(X_1, X_2, \ldots, X_n)$ is equal to

$$
\begin{aligned}
\prod_{i=1}^n f(x_i;\boldsymbol{\theta}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}\right)\\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{\mu\sum_{i=1}^n x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right)\\
&= k_1\big(\mathbf{t},\boldsymbol{\theta}\big)k_2(x_1,\ldots,x_n),
\end{aligned}
$$

where

$$
\begin{aligned}
&k_1(\mathbf{t},\boldsymbol{\theta}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2} + \frac{\mu\sum_{i=1}^n x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right),\\
&\mathbf{t} = \left(\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right),\\
&k_2(x_1,\ldots,x_n) = 1.
\end{aligned}
$$

This proves that $\mathbf{T} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is jointly sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$. □

Example 7.3.12. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the uniform distribution $U(-\theta/2, \theta/2)$, where $\theta > 0$. Then $\big(\min(X_i), \max(X_i)\big)$ is sufficient for θ.*

Proof. Since the pdf of the uniform distribution $U(-\theta/2, \theta/2)$ is given by

$$f(x;\theta) = \frac{1}{\theta}\, I\left(-\frac{\theta}{2} \le x \le \frac{\theta}{2}\right),$$

the joint pdf of $(X_1, X_2, \ldots, X_n)$ is equal to

$$\begin{aligned}\prod_{i=1}^{n} f(x_i;\theta) &= \prod_{i=1}^{n} \frac{1}{\theta}\, I\left(-\frac{\theta}{2} \le x_i \le \frac{\theta}{2}\right) \\ &= \left(\frac{1}{\theta}\right)^n I\left(-\frac{\theta}{2} \le \min(x_i) \le \max(x_i) \le \frac{\theta}{2}\right) \\ &= k_1\big(\mathbf{t}, \theta\big) k_2(x_1, \ldots, x_n),\end{aligned}$$

where

$$\begin{aligned}&k_1(\mathbf{t}, \theta) = \left(\frac{1}{\theta}\right)^n I\left(-\frac{\theta}{2} \le \min(x_i) \le \max(x_i) \le \frac{\theta}{2}\right), \\ &\mathbf{t} = \big(\min(x_i), \max(x_i)\big), \\ &k_2(x_1, \cdots, x_n) = 1.\end{aligned}$$

This proves that $\mathbf{T} = (\min(X_i), \max(X_i))$ is sufficient for θ. □

7.3.3 Property of a sufficient statistic

Theorem 7.3.2. *(Theorem 7.3.1.– Rao-Blackwell) Let $X_1, \ldots, X_n$ be an i.i.d. sample from a pdf/pmf $f(x;\theta)$, where $\theta \in \Theta$. Let $h(\mathbf{X}) \in \mathcal{U}$ denote the class of all unbiased estimators of θ. Let T be a sufficient statistic for θ. Then $E(h(\mathbf{X})|T)$ does not depend on θ, is an unbiased estimator of θ, and $\mathrm{Var}\big(E(h|T)\big) \le \mathrm{Var}(h)$ for all $\theta \in \Theta$.*

The proof is omitted.

Theorem 7.3.3. *(Theorem 7.3.2.) Let $X_1, \ldots, X_n$ be an i.i.d. sample from a pdf/pmf $f(x;\theta)$, $\theta \in \Theta$. If T is a sufficient statistic for θ and the mle of θ exists, then $\hat{\theta}$ is a function of T.*

The proof can be seen on page 427.

7.4 Complete family

7.4.1 Definitions

Definition 7.4.1. *Let $\{f(x;\theta) : \theta \in \Theta\}$ be a family of pdfs or pmfs. We say that this family is complete if*

$$E_\theta[g(X)] = 0 \quad \textit{for all} \quad \theta \in \Theta$$

implies that

$$P_\theta[g(X) = 0] = 1 \quad \textit{for all} \quad \theta \in \Theta.$$

Definition 7.4.2. *A statistic $T(\mathbf{X})$ is said to be complete if the family of distributions of $T(\mathbf{X})$ is complete.*

Example 7.4.1. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution $b(1,\theta)$, where $\theta \in (0,1)$. Then $T = \sum_{i=1}^n X_i$ is complete.*

Proof. Since $T \sim b(n,\theta)$, we have

$$\begin{aligned} E[g(T)] = \sum_{i=0}^{n} g(i)\binom{n}{i}\theta^i(1-\theta)^{n-i} = 0 \quad \text{for} \quad 0 < \theta < 1 \\ \iff (1-\theta)^n \sum_{i=0}^{n} g(i)\binom{n}{i}\left(\frac{\theta}{1-\theta}\right)^i = 0 \\ \iff \sum_{i=0}^{n} g(i)\binom{n}{i}\eta^i = 0, \quad \text{where } \eta = \frac{\theta}{1-\theta} \in (0,\infty). \end{aligned}$$

This is a polynomial in $\eta \in (0,\infty)$. Hence $g(i) = 0$ for $i = 0, 1, \ldots, n$, which means that T is complete. □

Example 7.4.2. *Let X have the uniform distribution $U[0,\theta]$, where $\theta \in (0,\infty)$. Then the family $\{U[0,\theta] : \theta > 0\}$ is complete.*

Proof. Direct calculation shows that

$$\begin{aligned} E_\theta[g(X)] = \int_0^\theta \frac{1}{\theta} g(x)dx = 0 \quad \text{for} \quad \theta > 0 \\ \iff \int_0^\theta g(x)dx = 0 \quad \text{for} \quad \theta > 0 \\ \iff g(\theta) = 0 \quad \text{for} \quad \theta > 0. \quad \text{(Why?)} \end{aligned}$$

This proves that the family $\{U(0,\theta) : \theta > 0\}$ is complete. □

Example 7.4.3. *Example 7.4.1 on page 432.*

Example 7.4.4. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\theta, \theta^2)$. Then $\mathbf{T} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2 \right)$ is sufficient for θ but not complete.*

Proof. Similar to Example 14, the sufficiency of $\mathbf{T}$ can be proved by using the Factorization criterion.

Now we show that $\mathbf{T}$ is not complete for θ. Let

$$g(\mathbf{T}) = 2 \left(\sum_{i=1}^n X_i \right)^2 - (n+1) \sum_{i=1}^n X_i^2.$$

Then $E[g(\mathbf{T})] = 0$. Indeed, since $\sum_{i=1}^n X_i \sim N(n\theta, n\theta^2)$, we have

$$\begin{aligned} E_\theta[g(\mathbf{T})] &= E \left[2 \left(\sum_{i=1}^n X_i \right)^2 - (n+1) \sum_{i=1}^n X_i^2 \right] \\ &= 2[n\theta^2 + (n\theta)^2] - (n+1)n(\theta^2 + \theta^2) \\ &= 0. \end{aligned}$$

However, $g(\mathbf{t}) \neq 0$. □

7.4.2 Exponential family

Consider a family $\{f(x;\theta) : \theta \in \Theta\}$ of pdfs/pmfs, where Θ is an interval on the real line $\mathbb{R}$.

Definition 7.4.3. *If there exist real-valued functions $p(\theta)$ and $q(\theta)$ on Θ and functions $K(x)$ and $S(x)$ such that*

$$f(x;\theta) = \exp\{p(\theta)K(x) + S(x) + q(\theta)\},$$

we say that the family $\{f(x;\theta) : \theta \in \Theta\}$ is a one-parameter exponential family.

Remark 7.4.1. *$K(X)$ is a sufficient statistic for θ because*

$$f(x;\theta) = \exp\{p(\theta)K(x) + S(x) + q(\theta)\} = \exp\{p(\theta)K(x) + q(\theta)\} \exp\{S(x)\}.$$

Taking $k_1(t,\theta) = \exp\{p(\theta)t + q(\theta)\}$ and $k_2(x) = \exp\{S(x)\}$ will see the desired result.

Remark 7.4.2. *$K(X)$ is also complete for θ.*

Remark 7.4.3. *If $X_1, \ldots, X_n$ form an i.i.d. sample from a one-parameter exponential family, then the joint pdf/pmf is also a one-parameter exponential family. Meanwhile $\sum_{i=1}^n K(X_i)$ is complete and sufficient for θ.*

Example 7.4.5. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution $\mathcal{P}(\theta)$, where $\theta > 0$. Then $T = \sum_{i=1}^n X_i$ is complete and sufficient for θ.*

Proof. Since the pmf of $X \sim \mathcal{P}(\theta)$ can be expressed by

$$f(x;\theta) = \frac{\theta^x}{x!}e^{-\theta} = \exp\{x\ln(\theta) - \ln(x!) - \theta\},$$

which is a special case of the exponential family with

$$\begin{aligned} p(\theta) &= \ln(\theta), \quad & q(\theta) &= -\theta, \\ K(x) &= x, & S(x) &= -\ln(x!). \end{aligned}$$

By Remark 7.4.3, we can conclude that $Y = \sum_{i=1}^n K(X_i) = \sum_{i=1}^n X_i$ is complete and sufficient for θ. □

Now we let $\Theta \subset \mathbb{R}^m$ be an m-dimensional interval.

Definition 7.4.4. *If there exist real-valued functions $p_1(\boldsymbol{\theta}), \ldots, p_m(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta})$ on Θ and functions $K_1(x), \ldots, K_m(x)$ and $S(x)$ such that*

$$f(x;\boldsymbol{\theta}) = \exp\left\{\sum_{j=1}^m p_j(\boldsymbol{\theta})K_j(x) + S(x) + q(\boldsymbol{\theta})\right\},$$

we say that the family $\{f(x;\boldsymbol{\theta} : \boldsymbol{\theta} \in \Theta)\}$ is an m-parameter exponential family.

Remark 7.4.4. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from an m-parameter exponential family. Then the joint pdf/pmf is also an m-parameter exponential family. Meanwhile, $(Y_1, \ldots Y_m)$ is jointly complete and sufficient for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T$, where $Y_j = \sum_{i=1}^n K_j(X_i)$, $j = 1, \ldots, m$.*

Example 7.4.6. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\theta_1, \theta_2)$ and let $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Then*

$$\begin{aligned} f(x;\boldsymbol{\theta}) &= \frac{1}{\sqrt{2\pi\theta_2}}\exp\left(-\frac{(x-\theta_1)^2}{2\theta_2}\right) \\ &= \exp\left(-\frac{1}{2\theta_2}x^2 + \frac{\theta_1}{\theta_2}x - \frac{\theta_1^2}{2\theta_2} - \ln\sqrt{2\pi\theta_2}\right) \\ &= \exp\left\{\sum_{j=1}^2 p_j(\boldsymbol{\theta})K_j(x) + S(x) + q(\boldsymbol{\theta})\right\}, \end{aligned}$$

where

$$p_1(\boldsymbol{\theta}) = -\frac{1}{2\theta_2}, \quad p_2(\boldsymbol{\theta}) = \frac{\theta_1}{\theta_2}, \quad q(\boldsymbol{\theta}) = -\frac{\theta_1^2}{2\theta_2} - \ln\sqrt{2\pi\theta_2},$$
$$K_1(x) = x^2, \qquad K_2(x) = x, \qquad S(x) = 0.$$

According Remark 4, we conclude that Y_1 and Y_2 are jointly complete and sufficient for $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where

$$Y_1 = \sum_{i=1}^{n} K_1(X_i) = \sum_{i=1}^{n} X_i^2,$$
$$Y_2 = \sum_{i=1}^{n} K_2(X_i) = \sum_{i=1}^{n} X_i.$$

Remark 7.4.5. *Let $Z_1 = Y_2/n = \bar{X}$ and $Z_2 = \dfrac{Y_1 - Y_2^2/n}{n-1} = S^2$. Then*

$$(Y_1, Y_2) \longleftrightarrow (Z_1, Z_2).$$

7.4.3 Finding the UMVUE

Theorem 7.4.1. *(Theorem 7.4.1. – Lehmann-Scheffe) If T is a complete and sufficient statistic and there exists an unbiased estimator h of θ, then $E(h|T)$ is the unique UMVUE of θ.*

Example 7.4.7. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution $\mathcal{P}(\theta)$, where $\theta > 0$. Then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the UMVUE of θ.*

Proof. By Example 7.4.5, we know that $Y = \sum_{i=1}^{n} X_i$ is a complete and sufficient statistic for θ. Furthermore, $h = Y/n$ is an unbiased estimator of θ. Thus,

$$E(h \,|\, Y) = E\left(\frac{Y}{n}\,\middle|\, Y\right) = \frac{Y}{n} = \bar{X}$$

is the UMVUE of θ. □

Example 7.4.8. (*Example 7.4.6 - continued*) *It is known that $\bar{X}$ and S^2 are jointly complete and sufficient for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and $\mathbf{h} = (\bar{X}, S^2)$ is an unbiased estimator of $\boldsymbol{\theta}$. Thus, $\bar{X}$ is the UMVUE of θ_1 and S^2 is the UMVUE of θ_2.*

Example 7.4.9. *Example 7.4.2 on page 433.*

Example 7.4.10. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\theta, 1)$, where $\theta \in \mathbb{R}$. Find the UMVUE of $\eta = P(X_1 \leq c)$, where c is a fixed constant.*

Solution. *Note that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is complete and sufficient for θ and $\bar{X}$ is the UMVUE of θ.*

Let $h(\mathbf{X}) = I(X_1 \leq c)$, where

$$\begin{aligned} \mathbf{X} &= (X_1, \ldots, X_n)^T \sim N_n(\theta \mathbf{\Pi}_n, I_n), \\ \mathbf{\Pi}_n &= (1, \ldots, 1)^T, \\ I_n &= \operatorname{diag}(1, \cdots, 1) \quad (n \times n \ \textit{identity matrix}). \end{aligned}$$

Then

$$E[h(\mathbf{X})] = E[I(X_1 \leq c)] = P(X_1 \leq c) = \Phi(c - \theta),$$

i.e., $h(\mathbf{X})$ is an unbiased estimator of η. Now we evaluate $E[h(\mathbf{X}) \,|\, \bar{X}]$. To do so, we let $Y_1 = X_1, Y_2 = \bar{X}$. Then

$$\begin{aligned} \mathbf{Y} &= \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \frac{1}{n} & \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} \\ &= A\mathbf{X} \\ &\sim N_2(\theta A \Pi_n, AA^T). \end{aligned}$$

Since

$$A\Pi_n = \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \Pi_2 \quad \textit{and} \quad AA^T = \begin{pmatrix} 1 & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} \end{pmatrix},$$

the conditional distribution of Y_1 given $Y_2 = y_2$ is

$$N\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - \mu_2), \sigma_1^2(1 - \rho^2)\right) = N\left(y_2, \frac{n-1}{n}\right),$$

i.e., $(Y_1 | Y_2 = y_2) \sim N(y_2, \frac{n-1}{n})$. Here we have used the fact that

$$\begin{aligned} &\mu_1 = \mu_2 = \theta, \\ &\sigma_1 = 1, \\ &\sigma_2 = \rho = 1/\sqrt{n}. \end{aligned}$$

Therefore,

$$\begin{aligned} E[I(Y_1 \le c) \,|\, Y_2 = y_2] &= P(Y_1 \le c \,|\, Y_2 = y_2) \\ &= P\left(\frac{Y_1 - y_2}{\sqrt{\frac{n-1}{n}}} \le \frac{c - y_2}{\sqrt{\frac{n-1}{n}}} \,\middle|\, Y_2 = y_2 \right) \\ &= P\left(Z \le \frac{c - y_2}{\sqrt{\frac{n-1}{n}}} \,\middle|\, Y_2 = y_2 \right) \qquad \big(Z \sim N(0,1)\big) \\ &= \Phi\left(\sqrt{\frac{n}{n-1}}(c - y_2) \right). \end{aligned}$$

Thus, $\Phi\left(\sqrt{\frac{n}{n-1}}(c - \bar{X}) \right)$ *is the UMVUE of* $\Phi(c - \theta)$

Remark 7.4.6. *If* $\hat{\theta}$ *is the UMVUE of* θ, $g(\hat{\theta})$ *is not the UMVUE of* $g(\theta)$ *in general, where the function* g *is known.*

Example 7.4.11. *Examples 7.6.1 an 7.6.2 on pages 440 and 441.*

Summary. *Three steps to find the UMVUE of the parameter* θ *(or a function* $\eta = g(\theta)$*):*

Step 1. *Find the complete and sufficient statistic* T *for* θ.

Step 2. *Find an unbiased estimator* h *of* θ *(or* η*).*

Step 3. *Evaluate* $E(h|T)$.

8

Optimal Tests of Hypotheses

8.1 Basic Concepts

In chapter 6, we studied the following concepts:

- *Sample: Let $X_1, \ldots, X_n$ be an i.i.d. sample from a pdf/pmf $f(x;\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k$. Write $\mathbf{X} = (X_1, \ldots, X_n)$.*
- *Assumption:*
  1. *$f(x;\boldsymbol{\theta})$ is known except $\boldsymbol{\theta}$.*
  2. *Θ contains at least two points.*
- *Hypotheses*
  1. *Null hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0 \subset \Theta$*
  2. *Alternative hypothesis $H_1 : \boldsymbol{\theta} \in \Theta_1 = \Theta - \Theta_0$*
  3. *Simple/composite hypothesis*
- *Rejection region: A subset C of $\mathbb{R}^n$ is called a rejection region if $\mathbf{X} \in C$.*
- *Two types of errors:*
  1. *Type I error: Reject H_0 when in fact it is true.*
  2. *Type II error: Not reject H_0 when in fact it is false.*
  3. *Pr. of Type I error $= P_{\boldsymbol{\theta}}(C), \quad \boldsymbol{\theta} \in \Theta_0$.*
  4. *Pr. of Type II error $= P_{\boldsymbol{\theta}}(C^c), \quad \boldsymbol{\theta} \in \Theta_1$.*
  5. *We usually limit $P_{H_0}(C) \leq \alpha$ and minimize $P_{H_1}(C^c) = P_{\boldsymbol{\theta}}(C^c), \ \boldsymbol{\theta} \in \Theta_1$.*
- *Power: $P_{\boldsymbol{\theta}}(C) = 1 - P_{\boldsymbol{\theta}}(\text{Type II Error}), \quad \boldsymbol{\theta} \in \Theta_1$.*
- *Likelihood ratio test*

8.2 Best critical region

Definition 8.2.1. *Let C be a subset of the sample space $\mathbb{R}^n$. Then we say C is a best critical region of size α for testing*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta = \theta_1$$

if

1. $P_{\theta_0}(C) = \alpha \quad (\textit{or } P_{H_0}(C) = \alpha)$, *and*
2. *for every subset A of $\mathbb{R}^n$, $P_{\theta_0}(A) = \alpha$ implies that $P_{\theta_1}(C) \geq P_{\theta_1}(A)$.*

8.3 Neyman-Pearson Theorem

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a pdf/pmf $f(x;\theta)$, where $\theta \in \Theta = \{\theta_0, \theta_1\}$. Let C be a subset of the sample space $\mathbb{R}^n$ such that

1. $\dfrac{L(\theta_0;\mathbf{X})}{L(\theta_1;\mathbf{X})} \leq k \quad \textit{for} \quad \mathbf{X} \in C,$
2. $\dfrac{L(\theta_0;\mathbf{X})}{L(\theta_1;\mathbf{X})} > k \quad \textit{for} \quad \mathbf{X} \in C^c,$
3. $\alpha = P_{H_0}(\mathbf{X} \in C).$

Then C is the best critical region of size α for testing

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta = \theta_1.$$

Proof. We prove the desired result by showing the following two cases separately.

(i) If C is the only critical region of size α, the theorem is proved by checking the definition of the best critical region.

(ii) If there exists another critical region A of size α, we denote

$$\int \cdots \int_{\mathbb{R}^n} L(\theta; x_1, \cdots, x_n) dx_1 \cdots dx_n = \int_{\mathbb{R}^n} L(\theta;\mathbf{x}) d\mathbf{x}$$

and need to show that

$$J(\theta_1) = \int_C L(\theta_1;\mathbf{x}) d\mathbf{x} - \int_A L(\theta_1;\mathbf{x}) d\mathbf{x} \geq 0.$$

Since $C = (C \cap A) \cup (C \cap A^c)$ and $A = (A \cap C) \cup (A \cap C^c)$, we have

$$\begin{aligned} J(\theta_1) &= \int_C L(\theta_1; \mathbf{x}) d\mathbf{x} - \int_A L(\theta_1; \mathbf{x}) d\mathbf{x} \\ &= \int_{C \cap A} L(\theta_1; \mathbf{x}) d\mathbf{x} + \int_{C \cap A^c} L(\theta_1; \mathbf{x}) d\mathbf{x} \\ &\quad - \int_{A \cap C} L(\theta_1; \mathbf{x}) d\mathbf{x} - \int_{A \cap C^c} L(\theta_1; \mathbf{x}) d\mathbf{x} \\ &= \int_{C \cap A^c} L(\theta_1; \mathbf{x}) d\mathbf{x} - \int_{A \cap C^c} L(\theta_1; \mathbf{x}) d\mathbf{x}. \end{aligned}$$

When $\mathbf{x} \in C \cap A^c \subseteq C$, we have

$$\begin{aligned} L(\theta_1; \mathbf{x}) &\geq \frac{1}{k} L(\theta_0; \mathbf{x}), \\ \int_{C \cap A^c} L(\theta_1; \mathbf{x}) d\mathbf{x} &\geq \frac{1}{k} \int_{C \cap A^c} L(\theta_0; \mathbf{x}) d\mathbf{x}. \end{aligned}$$

When $\mathbf{x} \in A \cap C^c \subseteq C^c$, we have

$$\begin{aligned} L(\theta_1; \mathbf{x}) &< \frac{1}{k} L(\theta_0; \mathbf{x}), \\ \int_{A \cap C^c} L(\theta_1; \mathbf{x}) d\mathbf{x} &< \frac{1}{k} \int_{A \cap C^c} L(\theta_0; \mathbf{x}) d\mathbf{x}. \end{aligned}$$

Thus,

$$\begin{aligned} J(\theta_1) &> \frac{1}{k} \left\{ \int_{C \cap A^c} L(\theta_0; \mathbf{x}) d\mathbf{x} - \int_{A \cap C^c} L(\theta_0; \mathbf{x}) d\mathbf{x} \right\} \\ &= \frac{1}{k} \left\{ \int_{C \cap A} L(\theta_0; \mathbf{x}) d\mathbf{x} + \int_{C \cap A^c} L(\theta_0; \mathbf{x}) d\mathbf{x} \right. \\ &\quad \left. - \int_{A \cap C} L(\theta_0; \mathbf{x}) d\mathbf{x} - \int_{A \cap C^c} L(\theta_1; \mathbf{x}) d\mathbf{x} \right\} \\ &= \frac{1}{k} \left\{ \int_C L(\theta_0; \mathbf{x}) d\mathbf{x} - \int_A L(\theta_0; \mathbf{x}) d\mathbf{x} \right\} \\ &= \frac{1}{k} (\alpha - \alpha) \\ &= 0. \end{aligned}$$

□

Example 8.3.1. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\theta, 1)$, where $\theta \in \Theta = \{\theta_0, \theta_1\}$ and $\theta_0 < \theta_1$ are known. We want to perform the following test*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta = \theta_1.$$

Determine the best critical region and the power of the test.

Solution. *The likelihood functions of the sample* $(X_1, \ldots, X_n)$ *under* H_0 *and* H_1 *are*

$$L(\theta_0; \mathbf{X}) = \prod_{i=1}^{n} f(X_i; \theta_0) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \theta_0)^2}{2}\right),$$
$$L(\theta_1; \mathbf{X}) = \prod_{i=1}^{n} f(X_i; \theta_1) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(X_i - \theta_1)^2}{2}\right),$$

respectively. Thus,

$$\begin{aligned}
\frac{L(\theta_0; \mathbf{X})}{L(\theta_1; \mathbf{X})} &= \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left[(X_i - \theta_0)^2 - (X_i - \theta_1)^2\right]\right) \\
&= \exp\left(-\frac{1}{2}\sum_{i=1}^{n}\left(-2\theta_0 X_i + 2\theta_1 X_i + \theta_0^2 - \theta_1^2\right)\right) \\
&= \exp\left(-\frac{n(\theta_0^2 - \theta_1^2)}{2}\right)\exp[n\bar{X}(\theta_0 - \theta_1)] \\
&\leq k \\
&\iff \bar{X} \geq k^*,
\end{aligned}$$

where

$$k^* = \frac{\ln(k) + n(\theta_0^2 - \theta_1^2)/2}{n(\theta_1 - \theta_0)}.$$

Thus,

$$\begin{aligned}
P_{H_0}\left(\frac{L(\theta_0; \mathbf{X})}{L(\theta_1; \mathbf{X})} \leq k\right) = \alpha &\iff P_{H_0}(\bar{X} \geq k^*) = \alpha \\
&\iff P_{H_0}\left(\frac{\bar{X} - \theta_0}{1/\sqrt{n}} \geq \frac{k^* - \theta_0}{1/\sqrt{n}}\right) = \alpha \\
&\iff \frac{k^* - \theta_0}{1/\sqrt{n}} = z_\alpha \\
&\iff k^* = \theta_0 + z_\alpha/\sqrt{n}.
\end{aligned}$$

Thus, we reject H_0 *if* $\bar{X} \geq \theta_0 + \frac{z_\alpha}{\sqrt{n}}$. *That is, the best critical region* C *is given by*

$$C = \left\{\mathbf{X} : \bar{X} \geq \theta_0 + \frac{z_\alpha}{\sqrt{n}}\right\}.$$

To compute the power of the test, we calculate the probability of Type II

error first. Indeed,

$$\begin{aligned} P(\textit{Type II error}) &= P_{H_1}(C^c) \\ &= 1 - P_{H_1}(C) \\ &= 1 - P_{H_1}(\bar{X} \geq \theta_0 + z_\alpha/\sqrt{n}) \\ &= 1 - P_{H_1}\left(\frac{\bar{X} - \theta_1}{1/\sqrt{n}} \geq \frac{\theta_0 - \theta_1 + \frac{1}{\sqrt{n}} z_\alpha}{1/\sqrt{n}}\right) \\ &= \Phi\big(\sqrt{n}(\theta_0 - \theta_1) + z_\alpha\big), \end{aligned}$$

which leads to

$$\textit{Power} = 1 - P(\textit{Type II error}) = 1 - \Phi(\sqrt{n}(\theta_0 - \theta_1) + z_\alpha).$$

Remark 8.3.1. *Example 8.1.2 is a special case of above example with* $\theta_0 = 0$, $\theta_1 = 1$, $\alpha = 5\%$ *and* $n = 25$. *We reject* H_0 *if*

$$\bar{X} \geq 0 + 1.645/\sqrt{25} \approx 0.0329$$

and have a

$$\textit{Power} = 1 - \Phi\big(\sqrt{25}(0 - 1) + 1.645\big) \approx 0.9996.$$

Example 8.3.2. *Let* $X_1, \ldots, X_n$ *be an i.i.d. sample from the exponential distribution with a pdf*

$$f(x; \theta) = \theta e^{-\theta x}, \quad x \geq 0,$$

where $\theta \in \Theta = \{1, 2\}$. *We want to perform the following test*

$$H_0 : \theta = 1 \quad \textit{versus} \quad H_1 : \theta = 2.$$

Determine the best critical region and find the power of the test.

Solution. *The likelihood functions of the sample* $X_1, \ldots, X_n$ *under* H_0 *and* H_1 *are*

$$\begin{aligned} L(1; \mathbf{X}) &= \prod_{i=1}^{n} f(X_i; 1) = \prod_{i=1}^{n} e^{-X_i} = \exp\left(-\sum_{i=1}^{n} X_i\right), \\ L(2; \mathbf{X}) &= \prod_{i=1}^{n} f(X_i; 2) = \prod_{i=1}^{n} 2e^{-2X_i} = 2^n \exp\left(-2\sum_{i=1}^{n} X_i\right), \end{aligned}$$

respectively. Thus,

$$\frac{L(1; \mathbf{X})}{L(2; \mathbf{X})} = 2^{-n} \exp\left(\sum_{i=1}^{n} X_i\right) \leq k \iff \sum_{i=1}^{n} X_i \leq k^*,$$

where

$$k^* = \ln(k) + n\ln(2).$$

Thus,

$$P_{H_0}\left(\frac{L(1;\mathbf{X})}{L(2;\mathbf{X})} \le k\right) = \alpha \iff P_{H_0}\left(\sum_{i=1}^{n} X_i \le k^*\right) = \alpha.$$

Since $\sum_{i=1}^{n} X_i \sim \Gamma(n, 1/\theta)$ *for any* $\theta > 0$, *we have that* $Y = \sum_{i=1}^{n} X_i \sim \Gamma(n,1)$ *when* H_0 *is true. That is, the pdf of* Y *is equal to*

$$\frac{y^{n-1}e^{-y}}{\Gamma(n)}, \quad y \ge 0.$$

Now we are going to make a connection between the distribution of Y *and* χ^2*-distribution. To do so, we let* $Z = 2Y$. *Then the pdf of* Z *is given by*

$$\frac{1}{2^n\Gamma(n)} z^{n-1}e^{-z/2} = \frac{1}{2^n\Gamma((2n)/2)} z^{(2n)/2-1}e^{-z/2}, \quad z \ge 0,$$

i.e., $Z \sim \chi^2_{2n}$. *Thus,*

$$\begin{aligned}
\alpha = P_{H_0}\left(\sum_{i=1}^{n} X_i \le k^*\right) &= P_{H_0}(2Y \le 2k^*) = P_{H_0}(Z \le 2k^*) \\
&\iff 2k^* = \chi^2_{1-\alpha,\,2n} \\
&\iff k^* = \frac{1}{2}\chi^2_{1-\alpha,\,2n}
\end{aligned}$$

Thus, we reject H_0 *if* $\sum_{i=1}^{n} X_i \le \frac{1}{2}\chi^2_{1-\alpha,\,2n}$, *which means that the best critical region* C *is*

$$C = \left\{\mathbf{X} : \sum_{i=1}^{n} X_i \le \frac{1}{2}\chi^2_{1-\alpha,\,2n}\right\}.$$

To compute the power of the test, we calculate the probability of Type II error first. Indeed,

$$P(\textit{Type II error}) = P_{H_1}(C^c) = 1 - P_{H_1}(C) = 1 - P_{H_1}\left(\sum_{i=1}^{n} X_i \le \frac{1}{2}\chi^2_{1-\alpha,\,2n}\right).$$

Under H_1, *we know that* $T = \sum_{i=1}^{n} X_i \sim \Gamma(n, \frac{1}{2})$ *having a pdf*

$$\frac{2^n}{\Gamma(n)} t^{n-1}\, e^{-2t}, \quad t \ge 0.$$

Let $W = 4T$. *Then* $W \sim \chi^2_{2n}$. *(Why?) Thus,*

$$P(\textit{Type II error}) = 1 - P(W \le 2\chi^2_{1-\alpha,\,2n}),$$

which leads to

$$Power = P(W \leq 2\chi^2_{1-\alpha,\,2n}).$$

For example, when $n = 10$ *and* $\alpha = 5\%$, *we have that* $\chi^2_{0.95,20} = 10.851$. *Thus,*

$$Power = P(W \leq 2 * 10.851) = P(W \leq 21.702) \approx 0.643 = \text{pchisq}(21.702, 20).$$

Question: *Do the rejection region* C *and* H_1 *have the same structure?*

Example 8.3.3. *Let* $X_1, \ldots, X_n$ *be an i.i.d. sample from a pmf* $p(x;\theta)$. *We want to perform the following test*

$$\begin{aligned} H_0: &\quad p(x;\theta_0) = \frac{e^{-1}}{x!}, && x = 0, 1, 2, \ldots, \\ H_1: &\quad p(x;\theta_1) = \left(\frac{1}{2}\right)^{x+1}, && x = 0, 1, 2, \ldots. \end{aligned}$$

Determine the best critical region C.

Solution. *The likelihood functions of the sample* $X_1, \ldots, X_n$ *under* H_0 *and* H_1 *are*

$$\begin{aligned} L(\theta_0; \mathbf{X}) &= \prod_{i=1}^{n} f(X_i;\theta_0) = \prod_{i=1}^{n} \frac{e^{-1}}{X_i!} = \frac{e^{-n}}{\prod_{i=1}^{n} X_i!}, \\ L(\theta_1; \mathbf{X}) &= \prod_{i=1}^{n} f(X_i;\theta_1) = \prod_{i=1}^{n} \left(\frac{1}{2}\right)^{X_i+1} = 2^{-n-\sum_{i=1}^{n} X_i}, \end{aligned}$$

respectively. Thus,

$$\frac{L(\theta_0; \mathbf{X})}{L(\theta_1; \mathbf{X})} = (2e^{-1})^n \frac{2^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!} \leq k.$$

Can we find the best critical region C *explicitly?*

Remark 8.3.2. *Sometimes it is hard to determine the best critical region* C *when* α *is given.*

Remark 8.3.3. *Neyman-Pearson Lemma tests two specific distributions, one is for* H_0 *and the other is for* H_1. *They do not have to come from the same family.*

8.4 Uniformly most powerful (UMP) test

8.4.1 Concept

Definition 8.4.1. *The critical region C is a uniformly most powerful (UMP) critical region of size α for testing H_0 (simple) against H_1(composite) if the set C is the best critical region of size α for testing H_0 (simple) against each simple hypothesis in H_1. Furthermore, a test defined by this critical region C is called a uniformly most powerful (UMP) test with the significant level α for testing H_0 (simple) against H_1 (composite), i.e.,*

$$\begin{aligned}&\Theta_0 = \{\theta_0\},\\ &\Theta_1 \quad \textit{contains more than one point},\\ &H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta \in \Theta_1.\end{aligned}$$

Remark 8.4.1. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from a pdf/pmf $f(x;\theta), \theta \in \Theta$. Suppose that $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ. If the UMP test exists, then the test statistic depends on $T(X_1, \ldots, X_n)$.*

By the Factorization theorem, one has

$$L(\theta; \mathbf{X}) = \prod_{i=1}^{n} f(X_i;\theta) = k_1(t(X_1, \ldots, X_n), \theta)\, k_2(X_1, \ldots, X_n).$$

Thus,

$$\frac{L(\theta_0; \mathbf{X})}{L(\theta_1; \mathbf{X})} = \frac{k_1(t(X_1, \ldots, X_n), \theta_0)}{k_1(t(X_1, \ldots, X_n), \theta_1)},$$

which depends on $t(X_1, \ldots, X_n)$.

8.4.2 Method

For <u>each</u> $\theta_1 \in \Theta_1$, we use the Neyman-Pearson Theorem to test

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta = \theta_1.$$

If the best critical region of the above test does not depend on the choice of θ_1, we can claim that the above test is a UMP test.

Example 8.4.1. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(0, \theta)$. Show that there exists a UMP test with significance level α*

for testing

$$H_0 : \theta = \theta_0 \quad vs \quad H_1 : \theta > \theta_0,$$

where θ_0 is known.

Proof. Let $\theta_1 \in \Theta_1$. Then $\theta_1 > \theta_0$. Since the pdf of $X \sim N(0, \theta)$ is given by

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right), \quad x \in \mathbb{R},$$

the likelihood function of $(X_1, \cdots, X_n)$ is equal to

$$L(\theta; \mathbf{X}) = \prod_{i=1}^{n} f(X_i;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{X_i^2}{2\theta}\right) = \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left(-\frac{\sum_{i=1}^n X_i^2}{2\theta}\right),$$

which leads to

$$\begin{aligned} \frac{L(\theta_0;\mathbf{X})}{L(\theta_1;\mathbf{X})} &= \left(\frac{\theta_1}{\theta_0}\right)^{n/2} \exp\left(-\frac{\theta_1 - \theta_0}{2\theta_0\theta_1} \sum_{i=1}^{n} X_i^2\right) \le k \\ &\iff \sum_{i=1}^{n} X_i^2 \ge c = \frac{2\theta_0\theta_1}{\theta_1 - \theta_0}\left(\frac{n}{2}\ln(\theta_1/\theta_0) - \ln(k)\right). \end{aligned}$$

Thus, the best critical region C will be

$$C = \left\{\mathbf{X} : \sum_{i=1}^{n} X_i^2 \ge c\right\},$$

where c needs to be determined. Under H_0,

$$\frac{1}{\theta_0}\sum_{i=1}^{n} X_i^2 \sim \chi_n^2.$$

Thus,

$$\begin{aligned} P_{H_0}(C) = \alpha &\iff P_{H_0}\left(\sum_{i=1}^{n} X_i^2 \ge c\right) = \alpha \\ &\iff P_{H_0}\left(\frac{1}{\theta_0}\sum_{i=1}^{n} X_i^2 \ge \frac{c}{\theta_0}\right) = \alpha \\ &\iff \frac{c}{\theta_0} = \chi^2_{\alpha,n} \\ &\iff c = \theta_0\chi^2_{\alpha,n} \end{aligned}$$

Thus, we reject H_0 if $\sum_{i=1}^n X_i^2 \ge \theta_0\chi^2_{\alpha,n}$, i.e., the best critical region C is given by

$$C = \left\{\mathbf{X} : \sum_{i=1}^{n} X_i^2 \ge \theta_0\chi^2_{\alpha,\,n}\right\}.$$

Since C does not depend on the choice of θ_1, C is the rejection region of a UMP test. □

Example 8.4.2. *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the normal distribution $N(\theta, 1)$. Show that the UMP test does not exist for testing*

$$H_0 : \theta = \theta_0 \quad \textit{versus} \quad H_1 : \theta \neq \theta_0,$$

where θ_0 is known.

Proof. Note that $\Theta = \mathbb{R}$, $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \mathbb{R} - \Theta_0$. The likelihood function of $(X_1, X_2, \ldots, X_n)$ is equal to

$$L(\theta; \mathbf{X}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \theta)^2\right).$$

Let $\theta_1 \in \Theta_1$. Then

$$\begin{aligned}
\frac{L(\theta_0; \mathbf{X})}{L(\theta_1; \mathbf{X})} = \exp\left(-(\theta_1 - \theta_0)\sum_{i=1}^{n} X_i + \frac{n}{2}(\theta_1^2 - \theta_0^2)\right) &\leq k \\
\iff -(\theta_1 - \theta_0)\sum_{i=1}^{n} X_i + \frac{n}{2}(\theta_1^2 - \theta_0^2) &\leq \ln(k) \qquad (8.1) \\
\iff -(\theta_1 - \theta_0)\left(\sum_{i=1}^{n} X_i - \frac{n}{2}(\theta_1 + \theta_0)\right) &\leq \ln(k)
\end{aligned}$$

If $\theta_1 > \theta_0$, then (8.1) is equivalent to

$$\sum_{i=1}^{n} X_i \geq \frac{\ln(k)}{\theta_0 - \theta_1} + \frac{n}{2}(\theta_1 + \theta_0). \qquad (8.2)$$

If $\theta_1 < \theta_0$, then (8.1) is equivalent to

$$\sum_{i=1}^{n} X_i \leq \frac{\ln(k)}{\theta_0 - \theta_1} + \frac{n}{2}(\theta_1 + \theta_0). \qquad (8.3)$$

We see from (8.2) and (8.3) that the critical region C depends on the choice of $\theta_1 \in \Theta_1$. Thus, the UMP test for H_0 versus H_1 does not exist.

□

Definition 8.4.2. *We say that the likelihood $L(\theta; \mathbf{X})$ has monotone likelihood ratio (mlr) in statistic $y = u(\mathbf{X})$, if for $\theta_1 < \theta_2$, then the ratio*

$$\frac{L(\theta_1; \mathbf{X})}{L(\theta_2; \mathbf{X})}$$

is a monotone function of $y = u(\mathbf{X})$.

Example 8.4.3. *Let $X_1, \ldots, X_n$ be an i.i.d. sample from the uniform distribution $U(0, \theta)$, where $\theta > 0$. Then $L(\theta; \mathbf{X})$ has mlr in* $\max(X_i)$.

Proof. Since the pdf of the uniform distribution $U(0, \theta)$ has the form

$$f(x; \theta) = \frac{1}{\theta} I(0 < x < \theta),$$

the likelihood function of the sample $(X_1, \ldots, X_n)$ is equal to

$$L(\theta; \mathbf{X}) = \left(\frac{1}{\theta}\right)^n I\left(0 < \min(X_i) \leq \max(X_i) < \theta\right).$$

When $\theta_1 < \theta_2$, we have

$$\begin{aligned} \frac{L(\theta_1; \mathbf{X})}{L(\theta_2; \mathbf{X})} &= \left(\frac{\theta_2}{\theta_1}\right)^n \frac{I(0 < \min(X_i) \leq \max(X_i) < \theta_1)}{I(0 < \min(X_i) \leq \max(X_i) < \theta_2)} \\ &= \left(\frac{\theta_2}{\theta_1}\right)^n \begin{cases} 1, & \text{if} \quad \max(X_i) \leq \theta_1, \\ 0, & \text{if} \quad \theta_1 < \max(X_i) \leq \theta_2, \\ 0, & \text{if} \quad \max(X_i) > \theta_2. \end{cases} \end{aligned}$$

Thus, $\frac{L(\theta_1; \mathbf{X})}{L(\theta_2; \mathbf{X})}$ is a monotone function of $\max(X_i)$. □

Example 8.4.4. *Let $X \sim$ Cauchy distribution having a pdf*

$$f(x; \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}, \quad x \in \mathbb{R}.$$

Then when $\theta_1 < \theta_2$, we have

$$\frac{L(\theta_1; X)}{L(\theta_2; X)} = \frac{f(X; \theta_1)}{f(X; \theta_2)} = \frac{1 + (X - \theta_2)^2}{1 + (X - \theta_1)^2} \longrightarrow 1$$

as $X \to \infty$ or $X \to -\infty$. Thus, the Cauchy distribution does not have a monotone likelihood ratio.

Remark 8.4.2. *If $f(x; \theta)$ has a monotone likelihood ratio $U(\mathbf{X})$, then UMP test depends on $U(\mathbf{X})$.*

9

Selected Topics – Linear Regression Model and Quadratic Forms

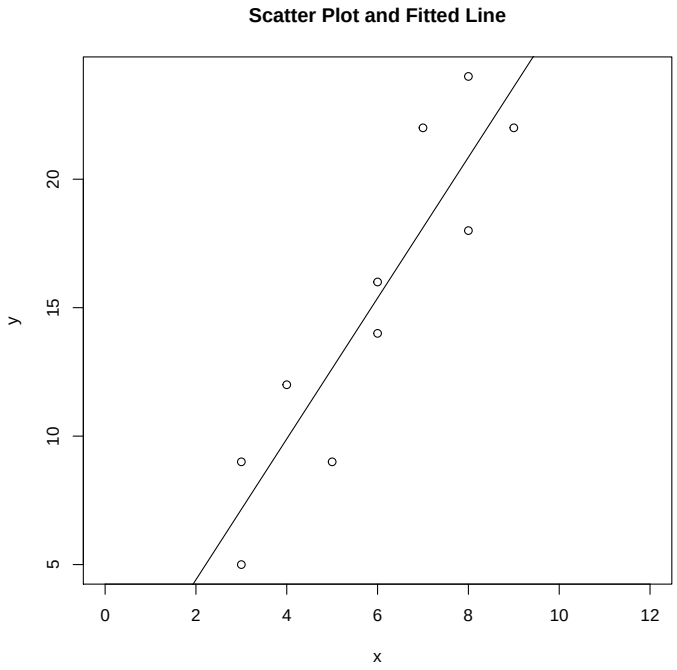
9.1 Linear regression

Example 9.1.1. *In one stage of the development of a new drug for an allergy, an experiment is conducted to study how different dosages of the drug affect the duration of relief from the allergic symptoms. 10 patients are included in the experiment. Each patient receives a specified dosage of the drug and is asked to report back as soon as the protection of the drug seems to wear off. Is there any linear relationship between x and y?*

Dosage (x) (in milligrams)	Duration of Relief (y) (in days)
3	9
3	5
4	12
5	9
6	14
6	16
7	22
8	18
8	24
9	22

It appears that y is increasing in x. To find a possible linear relationship between x and y, we first plot points (x_i, y_i), $i = 1, 2, \ldots, n$ in the plane and then try to find a line to fit all points according to certain criteria.

Scatter Plot: *Plot* (x_i, y_i), $i = 1, 2, \ldots, n$ *on the plane below:*



If there is a linear relationship between x *and* y*, we may write*

$$y = a + bx = \alpha + \beta(x - \bar{x}),$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$*. To determine this line, we need to know the* α *and* β*. The idea is to minimize*

$$D = \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \bar{x})]^2.$$

9.1.1 Statistical model and assumptions

(i) Notation

$x =$ *predictor variable (or independent variable),*
$y =$ *response variable (or dependent variable).*

(ii) Model and assumption

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ *are i.i.d. with* $E(\epsilon_1) = 0$ *and* $\mathrm{Var}(\epsilon_1) = \sigma^2$, α, β *and* σ^2 *are unknown. The observed data are* (x_i, y_i), $i = 1, 2, \ldots, n$.

(iii) Least squares estimators of α and β

The principle of least squares*: Determine the values for the parameters* α *and* β *so that the overall discrepancy* D *defined by*

$$\begin{aligned} D &= \sum (\textit{observed-predicted response})^2 \\ &= \sum_{i=1}^{n} d_i^2 \\ &= \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \bar{x})]^2 \end{aligned}$$

is minimized. The parameter values thus determined are called the least squares estimators. In fact

$$\begin{aligned} \frac{\partial D}{\partial \alpha} &= -2 \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \bar{x})] = 0, \\ \frac{\partial D}{\partial \beta} &= -2 \sum_{i=1}^{n} [y_i - \alpha - \beta(x_i - \bar{x})](x_i - \bar{x}) = 0, \end{aligned}$$

which are equivalent to

$$\begin{aligned} \sum_{i=1}^{n} y_i - n\alpha - \beta \sum_{i=1}^{n} (x_i - \bar{x}) &= 0, \\ \sum_{i=1}^{n} y_i (x_i - \bar{x}) - \alpha \sum_{i=1}^{n} (x_i - \bar{x}) - \beta \sum_{i=1}^{n} (x_i - \bar{x})^2 &= 0. \end{aligned}$$

The solutions to α *and* β *are called the least squares estimators of* α *and* β, *which are given by*

$$\begin{aligned} \hat{\alpha} &= \bar{y}, \\ \hat{\beta} &= \frac{S_{xy}}{S_{xx}}, \end{aligned}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i(y_i - \bar{y}).$$

(iv) Fitted (or estimated) regression line

$$\hat{y} = \hat{\alpha} + \hat{\beta}(x - \bar{x})$$

(v) Residuals

$$\hat{d}_i = y_i - [\hat{\alpha} + \hat{\beta}(x_i - \bar{x})], \quad i = 1, 2, \ldots, n,$$

where

$$\sum_{i=1}^{n} \hat{d}_i = n\bar{y} - n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

(vi) Sum of squares due to error

$$SSE = \sum_{i=1}^{n} \hat{d}_i{}^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

$$\begin{aligned}
SSE &= \sum_{i=1}^{n} \hat{d}_i^{\,2} \\
&= \sum_{i=1}^{n} [y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2 \\
&= \sum_{i=1}^{n} [y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})]^2 \\
&= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{\beta} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= S_{yy} - 2\frac{S_{xy}}{S_{xx}} S_{xy} + \frac{S_{xy}^2}{S_{xx}^2} S_{xx} \\
&= S_{yy} - \frac{S_{xy}^2}{S_{xx}}
\end{aligned}$$

Example 9.1.2. *(Example 9.1.1 – continued)*

- $\bar{x} = 5.9, \quad \bar{y} = 15.1, \quad S_{xx} = 40.9, \quad S_{yy} = 370.9, \quad S_{xy} = 112.1$
- *Fitted regression line:* $y = -1.07 + 2.74x$.
- $SSE \approx 63.6528$.

9.1.2 Properties of $\hat{\alpha}$ and $\hat{\beta}$

Theorem 9.1.1. *$E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$.*

Proof. Direct calculations show that

$$\begin{aligned}
E(\hat{\alpha}) = E(\bar{y}) &= \frac{1}{n} \sum_{i=1}^{n} E(y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} E[\alpha + \beta(x_i - \bar{x}) + \epsilon_i] \\
&= \frac{1}{n} \sum_{i=1}^{n} [\alpha + \beta(x_i - \bar{x}) + E(\epsilon_i)] \\
&= \frac{1}{n} \sum_{i=1}^{n} [\alpha + \beta(x_i - \bar{x})] \\
&= \alpha,
\end{aligned}$$

$$
\begin{aligned}
E(\hat{\beta}) = E(S_{xy}/S_{xx}) &= \frac{1}{S_{xx}} E(S_{xy}) \\
&= \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})\right) \\
&= \frac{1}{S_{xx}} E\left(\sum_{i=1}^{n} y_i(x_i - \bar{x})\right) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})E(y_i) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})[\alpha + \beta(x_i - \bar{x})] \\
&= \frac{\alpha}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x}) + \frac{\beta}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \beta.
\end{aligned}
$$

□

Theorem 9.1.2. $\mathrm{Var}(\hat{\alpha}) = \sigma^2/n$, $\mathrm{Var}(\hat{\beta}) = \sigma^2/S_{xx}$, *and* $\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = 0$.

Proof. Direct calculations show that

(i) $\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$

(ii)

$$
\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\left(\frac{S_{xy}}{S_{xx}}\right) \\
&= \frac{1}{S_{xx}^2}\mathrm{Var}\left(\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]\right) \\
&= \frac{1}{S_{xx}^2}\mathrm{Var}\left(\sum_{i=1}^{n}(x_i - \bar{x})y_i\right) \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\mathrm{Var}(y_i) \\
&= \frac{1}{S_{xx}^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma^2 \\
&= \frac{\sigma^2}{S_{xx}}.
\end{aligned}
$$

(iii)

$$\begin{aligned}
\mathrm{Cov}(\hat{\alpha}, \hat{\beta}) &= \mathrm{Cov}\left(\sum_{i=1}^{n} \frac{y_i}{n}, \sum_{j=1}^{n} \frac{x_j - \bar{x}}{S_{xx}} y_j \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{n} \frac{x_j - \bar{x}}{S_{xx}} \mathrm{Cov}(y_i, y_j) \\
&= \sum_{i=1}^{n} \frac{1}{n} \frac{x_i - \bar{x}}{S_{xx}} \mathrm{Cov}(y_i, y_i) \\
&= \sum_{i=1}^{n} \frac{1}{n} \frac{x_i - \bar{x}}{S_{xx}} \sigma^2 \\
&= \frac{\sigma^2}{n} \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \bar{x}) \\
&= 0.
\end{aligned}$$

□

9.1.3 Estimation of σ^2

Definition 9.1.1. *If* $\mathbf{U}$ *and* $\mathbf{V}$ *are two (column) random vectors, then the covariance matrix between* $\mathbf{U}$ *and* $\mathbf{V}$ *is defined by*

$$\mathrm{Cov}(\mathbf{U}, \mathbf{V}) = E\left[(\mathbf{U} - E\mathbf{U})(\mathbf{V} - E\mathbf{V})^T \right].$$

(i) Idempotent matrix P

Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, *where*

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}.$$

Then $\mathbf{P}^T = \mathbf{P}$ *and* $\mathbf{P}^2 = \mathbf{P}$. *In fact,*

$$\begin{aligned}
\mathbf{P}^T &= [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\
&= [\mathbf{X}^T]^T[(\mathbf{X}^T\mathbf{X})^{-1}]^T\mathbf{X}^T \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \mathbf{P}, \\
\mathbf{P}^2 &= [\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T][\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T] \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\
&= \mathbf{P}.
\end{aligned}$$

(ii) Unbiased estimator of σ^2

Let

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \hat{\mathbf{d}} = \begin{pmatrix} \hat{d}_1 \\ \hat{d}_2 \\ \vdots \\ \hat{d}_n \end{pmatrix}, \quad \boldsymbol{\nu} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \quad \hat{\boldsymbol{\nu}} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}.$$

Then

$$\begin{aligned}
\hat{\boldsymbol{\nu}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \\
\hat{\mathbf{d}} &= \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\nu}} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}, \\
SSE &= ||\hat{\mathbf{d}}||^2 \\
&= ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\nu}}||^2 \\
&= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\nu}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\nu}}) \\
&= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y} \\
&= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})^2\mathbf{Y} \\
&= \mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y},
\end{aligned}$$

which leads to

$$\begin{aligned}
E(SSE) &= E[\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y}] \\
&= E\{\mathrm{tr}[\mathbf{Y}^T(\mathbf{I}_n - \mathbf{P})\mathbf{Y}]\} \\
&= E\{\mathrm{tr}[(\mathbf{I}_n - \mathbf{P})\mathbf{Y}\mathbf{Y}^T]\} \\
&= \mathrm{tr}\{(\mathbf{I}_n - \mathbf{P})[E(\mathbf{Y} - \mathbf{X}\boldsymbol{\nu} + \mathbf{X}\boldsymbol{\nu})(\mathbf{Y} - \mathbf{X}\boldsymbol{\nu} + \mathbf{X}\boldsymbol{\nu})^T]\} \\
&= \mathrm{tr}\{(\mathbf{I}_n - \mathbf{P})[E(\mathbf{Y} - \mathbf{X}\boldsymbol{\nu})(\mathbf{Y} - \mathbf{X}\boldsymbol{\nu})^T + \mathbf{X}\boldsymbol{\nu}\boldsymbol{\nu}^T\mathbf{X}^T]\} \\
&= \mathrm{tr}\{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\Sigma} + (\mathbf{I}_n - \mathbf{P})\mathbf{X}\boldsymbol{\nu}\boldsymbol{\nu}^T\mathbf{X}^T]\} \\
&= \mathrm{tr}\{(\mathbf{I}_n - \mathbf{P})\boldsymbol{\Sigma}\} + \boldsymbol{\nu}^T\mathbf{X}^T(\mathbf{I}_n - \mathbf{P})\mathbf{X}\boldsymbol{\nu} \\
&= \sigma^2\mathrm{tr}(\mathbf{I}_n - \mathbf{P}) \\
&= (n-2)\sigma^2.
\end{aligned}$$

Thus, one unbiased estimator of σ^2 *is*

$$\hat{\sigma^2} = \frac{SSE}{n-2}.$$

9.1.4 Evaluation of fit

- *Model:* $Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad i = 1, 2, \ldots, n.$
- *Data:* $(x_i, y_i), \quad i = 1, 2, \ldots, n.$
- *Assumption:* $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ *are i.i.d. with* $E(\epsilon_1) = 0$ *and* $\mathrm{Var}(\epsilon_1) = \sigma^2$.

Two questions should be investigated:

Question 1. *Why do we assume that* $\mathrm{Var}(\epsilon_i) = \sigma^2$ *for* $i = 1, 2, \ldots, n$?

Question 2. *Why do we use a linear model* $\alpha + \beta(x - \bar{x})$ *to fit the data? If we use a linear model to fit the data, will all the information in the data set be explained?*

(i) Evaluation of $\mathrm{Var}(\epsilon_i) = \sigma^2, \ i = 1, 2, \ldots, n.$

Plot the residuals versus the values of x_i *at* $i = 1, 2, \ldots, n$. *The assumption* $\mathrm{Var}(\epsilon_i) = \sigma^2$ *that is independent of* x *values is reasonable if the plot looks like a horizontal band.*

(ii) The strength of a linear relation

The observed value y_i can always be expressed by

$$y_i = \left[\hat{\alpha} + \hat{\beta}(x_i - \bar{x})\right] + y_i - \left[\hat{\alpha} + \hat{\beta}(x_i - \bar{x})\right],$$

where

$$\hat{\alpha} + \hat{\beta}(x_i - \bar{x}) \longleftrightarrow \textit{explained or predicted linear relation},$$
$$y_i - [\hat{\alpha} + \hat{\beta}(x_i - \bar{x})] \longleftrightarrow \textit{residual from a linear relation}.$$

Ideally, all points lie exactly on the line. However, this probably will not happen in practice. Thus, we hope that the observed SSE is small because it is an overall measure of departure from linearity. Note that

$$SSE = \sum_{i=1}^{n} \hat{d}_i^2 = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

which is equivalent to

$$S_{yy} = \frac{S_{xy}^2}{S_{xx}} + SSE,$$

$$S_{yy} \longleftrightarrow \textit{total variability of } y,$$
$$\frac{S_{xy}}{S_{xx}} \longleftrightarrow \textit{variability explained by a linear relation},$$
$$SSE \longleftrightarrow \textit{unexplained variability}.$$

$$SSE \approx 0 \iff S_{yy} \approx \frac{S_{xy}^2}{S_{xx}} \iff \frac{S_{xy}^2}{S_{xx}S_{yy}} \approx 1.$$

Definition 9.1.2. *The quantity*

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

is called the sample correlation coefficient.

Remark 9.1.1. *The strength of a linear relation is measured by r^2.*

Example 9.1.3. *(Example 9.1.1 –continued) Based on data in Example 9.1.1, we have $S_{xx} = 40.9$, $S_{yy} = 370.9$, $S_{xy} = 112.1$, which leads to $r^2 \approx 0.83$. This means that around 83% of variability in y is explained by the linear regression $y = -1.07 + 2.74x$.*

Remark 9.1.2. *When r^2 is small, we can only conclude that a line does not give a good fit to the data.*

(i) *There is little relation between the variables in the sense that the scatter diagram fails to exhibit any pattern.*

(ii) *There is a prominent relation but it is nonlinear; that is, the scatter is banded around a curve rather than a line.*

9.2 Quadratic forms and their independence

Definition 9.2.1. *Let* $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)^T$, *where* $Z_1, Z_2, \ldots, Z_n$ *are i.i.d. with a standard normal distribution* $N(0,1)$. *Let*

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{A}\mathbf{Z}.$$

Then the distribution of $\mathbf{X}$ *is called a multivariate normal distribution with mean* $\boldsymbol{\mu}$ *and variance-covariance matrix* $\boldsymbol{\Sigma}$, *where* $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}^T$. *We write* $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- *pdf:* $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),\ \mathbf{x} \in \mathbb{R}^n.$
- $E(\mathbf{X}) = \boldsymbol{\mu},\ \mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}.$
- *When* $\boldsymbol{\mu} = \mathbf{0}$ *and* $\boldsymbol{\Sigma} = \mathbf{I}_n$, *it is called a standard normal distribution*
- $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2) \iff X_1, X_2, \ldots, X_n$ *are independent.*
- *Let* $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T)$. *The conditional distribution of* $\mathbf{X}_1$ *given* $\mathbf{X}_2$ *is also a normal distribution.*

  Assume that $\boldsymbol{\mu} = \mathbf{0}$ *and* $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$. *Let*

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} = \frac{X_i - 0}{\sigma} = \frac{X_i}{\sigma} \sim N(0,1), \quad i = 1, 2, \ldots, n.$$

Then $Z_1, Z_2, \ldots, Z_n$ *are i.i.d.,* $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$, *and*

$$\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^n Z_i^2 \sim \chi_n^2,$$

where $\mathbf{Z}^T = (Z_1, Z_2, \ldots, Z_n)$. *What is the distribution of* $\mathbf{Z}^T\mathbf{A}\mathbf{Z}$ *when* $\mathbf{A}$ *is a symmetric matrix?*

Theorem 9.2.1. *Let* $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$ *and* $\mathbf{A}$ *is a symmetric matrix with rank* r*. Then*

$$\mathbf{Z}^T\mathbf{A}\mathbf{Z} \sim \chi^2_r \iff \mathbf{A} \text{ is idempotent.}$$

Proof. Let $\lambda_1, \ldots, \lambda_r$ be the nonzero eigenvalues of $\mathbf{A}$. Then

$$\mathbf{\Gamma}\mathbf{A}\mathbf{\Gamma}^T = \text{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0) = \mathbf{\Lambda},$$

where $\mathbf{\Gamma}$ is an orthogonal matrix. Let $\mathbf{Y} = \mathbf{\Gamma}\mathbf{Z}$. Then

$$\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{\Gamma}\mathbf{I}_n\mathbf{\Gamma}^T) = N_n(\mathbf{0}, \mathbf{I}_n)$$

and

$$\mathbf{Z}^T\mathbf{A}\mathbf{Z} = \mathbf{Z}^T\mathbf{\Gamma}^T\mathbf{\Lambda}\mathbf{\Gamma}\mathbf{Z} = (\mathbf{\Gamma}\mathbf{Z})^T\mathbf{\Lambda}(\mathbf{\Gamma}\mathbf{Z}) = \mathbf{Y}^T\mathbf{\Lambda}\mathbf{Y} = \sum_{i=1}^{r} \lambda_i Y_i^2.$$

Since $Y_1, Y_2, \ldots, Y_r$ are i.i.d. from $N(0,1)$, $Y_i^2 \sim \chi^2_1$ for $i = 1, 2, \ldots, r$. Thus, the mgf of $\mathbf{Z}^T\mathbf{A}\mathbf{Z}$ is given by

$$M(t) = E\left(e^{t\mathbf{Z}^T\mathbf{A}\mathbf{Z}}\right) = \prod_{i=1}^{r} E\left(e^{t\lambda_i Y_i^2}\right) = \prod_{i=1}^{r} (1-2t\lambda_i)^{-1/2} = (1-2t)^{-r/2}$$

if and only if $\lambda_1 = \lambda_2 = \cdots = \lambda_r = 1$. □

Theorem 9.2.2. *Let* $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_n)$*. Assume that* $\mathbf{A}$ *and* $\mathbf{B}$ *are two real symmetric matrices. Then* $\mathbf{Z}^T\mathbf{A}\mathbf{Z}$ *and* $\mathbf{Z}^T\mathbf{B}\mathbf{Z}$ *are independent if and only if* $\mathbf{A}\mathbf{B} = \mathbf{0}$.

The proof can be seen in the textbook (Theorem 9.9.1).