STAT 712
MATHEMATICAL STATISTICS I

Fall 2017

Lecture Notes

Joshua M. Tebbs
Department of Statistics
University of South Carolina

Contents

1 Probability Theory

Complementary reading: Chapter 1 (CB). Sections 1.1-1.6.

1.1 Set Theory

Definitions: A **random experiment** is an experiment that produces outcomes which are not predictable with certainty in advance. The **sample space** S for a random experiment is the set of all possible outcomes.

Example 1.1: Consider the following random experiments and their associated sample spaces.

(a) Observe the high temperature for today:

$$S = \{\omega : -\infty < \omega < \infty\} = \mathbb{R}$$

(b) Record the number of planes landing at CAE:

$$S = \{\omega : \omega = 0, 1, 2, ..., \} = \mathbb{Z}^+$$

(c) Toss a coin three times:

$$S = \{\text{(HHH), (HHT), ..., (TTT)}\}$$

(d) Measure the length of a female subject's largest uterine fibroid:

$$S = \{\omega : \omega \geq 0\} = \mathbb{R}^+$$

Definitions: We say that a set (e.g., A, B, S, etc.) is **countable** if its elements can be put into a 1:1 correspondence with the set of natural numbers

$$\mathbb{N} = \{1, 2, 3, ..., \}.$$

If a set is not countable, we say it is **uncountable**. In Example 1.1,

(a) $S = \mathbb{R}$ is uncountable

(b) $S = \mathbb{Z}^+$ is countable (i.e., countably infinite); $|S| = +\infty$

(c) $S = \{\text{(HHH), (HHT), ..., (TTT)}\}$ is countable (i.e., countably finite); $|S| = 8$

(d) $S = \mathbb{R}^+$ is uncountable

Any **finite set** is countable. By "finite," we mean that $|A| < \infty$, that is, "the process of counting the elements in A comes to an end." An infinite set A can be countable or uncountable. By "infinite," we mean that $|A| = +\infty$. For example,

- $\mathbb{N} = \{1, 2, 3, ..., \}$ is countably infinite
- $A = \{\omega : 0 < \omega < 1\}$ is uncountable.

Definitions: Suppose that S is a sample space for a random experiment. An **event** A is a subset of S, that is, $A \subseteq S$.

- If $\omega \in A$, we say that "A occurs"
- If $\omega \notin A$, we say that "A does not occur."

The set A is a **subset** of B if

$$\omega \in A \Longrightarrow \omega \in B.$$

This is written $A \subset B$ or $A \subseteq B$. Two sets A and B are **equal** if each set is a subset of the other, that is,

$$A = B \Longleftrightarrow A \subseteq B \text{ and } B \subseteq A.$$

Set Operations: Suppose A and B are subsets of S.

- **Union:** $A \cup B = \{\omega \in S : \omega \in A \text{ or } \omega \in B\}$
- **Intersection:** $A \cap B = \{\omega \in S : \omega \in A \text{ and } \omega \in B\}$
- **Complementation:** $A^c = \{\omega \in S : \omega \notin A\}$

Theorem 1.1.4. Suppose A, B, and C are subsets of S.

(a) **Commutativity:**

$$\begin{aligned} A \cup B &= B \cup A \\ A \cap B &= B \cap A \end{aligned}$$

(b) **Associativity:**

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

(c) **Distributive Laws:**

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

(d) **DeMorgan's Laws:**

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

Extension: Suppose that $A_1, A_2, ..., A_n$ is a **finite sequence** of sets, where each $A_i \subseteq S$.

- **Union:**
$$\bigcup_{i=1}^{n} A_i = \{\omega \in S : \omega \in A_i \ \exists i\}$$
- **Intersection:**
$$\bigcap_{i=1}^{n} A_i = \{\omega \in S : \omega \in A_i \ \forall i\}$$

These operations are similarly defined for a **countable sequence** of sets $A_1, A_2, ...$, and also for an **uncountable** collection of sets; see pp 4 (CB).

Definitions: Suppose that A and B are subsets of S. We say that A and B are **disjoint** (or **mutually exclusive**) if

$$A \cap B = \emptyset.$$

We say that $A_1, A_2, ...$, are **pairwise disjoint** (or **pairwise mutually exclusive**) if

$$A_i \cap A_j = \emptyset \quad \forall i \neq j.$$

Definition: Suppose $A_1, A_2, ...$, are subsets of S. We say that $A_1, A_2, ...$, form a **partition** of S if

(a) $A_i \cap A_j = \emptyset \quad \forall i \neq j$ (i.e., the A_i's are pairwise disjoint)

(b) $\bigcup_{i=1}^{\infty} A_i = S$.

Example 1.2. Suppose $S = [0, \infty)$. Define $A_i = [i-1, i)$, for $i = 1, 2, ...,$. Clearly, the A_i's are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, so the sequence $A_1, A_2, ...$, partitions S.

Remark: The following topics are "more advanced" than those presented in CB's §1.1 but are needed to write proofs of future results.

Definitions: If $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$, we say that $\{A_n\}$ is an **increasing** sequence of sets. If $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, we say that $\{A_n\}$ is a **decreasing** sequence of sets.

1. If $\{A_n\}$ is increasing, then
$$\lim_{n \to \infty} A_n = \bigcup_{i=1}^{\infty} A_i.$$

2. If $\{A_n\}$ is decreasing, then
$$\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i.$$

3. If $\{A_n\}$ is neither increasing nor decreasing, $\lim_{n\to\infty} A_n$ may or may not exist.

Example 1.3. Suppose $S = \mathbb{R}^+ = [0, \infty)$. Define
$$A_n = [a - 1/n,\ b + 1/n],$$
where $1 < a < b < \infty$. For example, $A_1 = [a - 1,\ b + 1]$, $A_2 = [a - 1/2,\ b + 1/2]$, $A_3 = [a - 1/3,\ b + 1/3]$, etc. Clearly, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, that is, $\{A_n\}$ is monotone decreasing. Therefore,
$$\lim_{n\to\infty} A_n = \bigcap_{i=1}^{\infty} A_i = [a, b].$$

Example 1.4. Suppose $S = (-1, 1)$. Define
$$A_n = \begin{cases} (-1/n,\ 0], & n \text{ odd} \\ (0,\ 1/n), & n \text{ even.} \end{cases}$$
That is, $A_1 = (-1, 0]$, $A_2 = (0, 1/2)$, $A_3 = (-1/3, 0]$, etc. The sequence $\{A_n\}$ is neither increasing nor decreasing.

Question: In general, what does it mean for a sequence of sets $\{A_n\}$ to "converge?" Consider the following:
$$\begin{aligned} B_1 &= \bigcup_{k=1}^{\infty} A_k \\ B_2 &= \bigcup_{k=2}^{\infty} A_k \quad \text{Note: } B_2 \subseteq B_1 \\ B_3 &= \bigcup_{k=3}^{\infty} A_k \quad \text{Note: } B_3 \subseteq B_2 \\ &\vdots \end{aligned}$$
In general, define
$$B_n = \bigcup_{k=n}^{\infty} A_k.$$
Because $\{B_n\}$ is a decreasing sequence of sets, we know that $\lim_{n\to\infty} B_n$ exists. In particular,
$$\begin{aligned} \lim_{n\to\infty} B_n &= \bigcap_{n=1}^{\infty} B_n \\ &= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \equiv \limsup_n A_n = \overline{\lim}\ A_n. \end{aligned}$$

Interpretation: For a sequence of sets $\{A_n\}$,

$$\overline{\lim}\ A_n = \{\omega \in S : \forall n \geq 1\ \exists k \ni \omega \in A_k\}.$$

This is the set of all outcomes $\omega \in S$ that are in an infinite number of the A_n sets. We write

$$\omega \in \overline{\lim}\ A_n \Longleftrightarrow \omega \in A_n \text{ i.o. (infinitely often)}.$$

Now, let's return to our arbitrary sequence of sets $\{A_n\}$. Consider the following:

$$\begin{aligned} C_1 &= \bigcap_{k=1}^{\infty} A_k \\ C_2 &= \bigcap_{k=2}^{\infty} A_k \qquad \text{Note: } C_2 \supseteq C_1 \\ C_3 &= \bigcap_{k=3}^{\infty} A_k \qquad \text{Note: } C_3 \supseteq C_2 \\ &\vdots \end{aligned}$$

In general, define

$$C_n = \bigcap_{k=n}^{\infty} A_k.$$

Because $\{C_n\}$ is an increasing sequence of sets, we know that $\lim_{n\to\infty} C_n$ exists. In particular,

$$\begin{aligned} \lim_{n\to\infty} C_n &= \bigcup_{n=1}^{\infty} C_n \\ &= \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \equiv \liminf_n A_n = \underline{\lim}\ A_n. \end{aligned}$$

Interpretation: For a sequence of sets $\{A_n\}$,

$$\underline{\lim}\ A_n = \{\omega \in S : \exists n \geq 1 \ni \omega \in A_k\ \forall k \geq n\}.$$

This is the set of all outcomes $\omega \in S$ that are in A_n eventually. We write

$$\omega \in \underline{\lim}\ A_n \Longleftrightarrow \omega \in A_n \text{ e.v. (eventually)}.$$

Now, we return to our original question: In general, what does it mean for a sequence of sets $\{A_n\}$ to "converge?"

Answer: We say that $\lim_{n\to\infty} A_n$ exists if

$$\underline{\lim}\ A_n = \overline{\lim}\ A_n$$

and define

$$\lim_{n\to\infty} A_n = A$$

to be this common set. We also write $A_n \to A$, as $n \to \infty$. If

$$\underline{\lim}\ A_n \neq \overline{\lim}\ A_n,$$

we say that $\lim_{n\to\infty} A_n$ does not exist.

Example 1.5. Define $A_n = \{\omega : 1/n < \omega < 1 + 1/n\}$. Find $\underline{\lim}\ A_n$ and $\overline{\lim}\ A_n$. Does $\lim_{n\to\infty} A_n$ exist?

1.2 Basics of Probability Theory

Question: Given a random experiment and an associated sample space S, how do we assign a probability to $A \subseteq S$?

Question: How do we define probability?

1. **Relative Frequency:**
$$\text{pr}(A) = \lim_{n\to\infty} \frac{n(A)}{n}$$
2. **Subjective:** "confidence" or "measure of belief" that A will occur
3. **Modern Axiomatic:** $\text{pr}(A) = P(A)$, where P is a set function satisfying certain axioms.

Definition: Let $\mathcal{B}$ denote a collection of subsets of S. We say that $\mathcal{B}$ is a **σ-algebra** on S if

(i) $\emptyset \in \mathcal{B}$

(ii) $A \in \mathcal{B} \Longrightarrow A^c \in \mathcal{B}$; i.e., $\mathcal{B}$ is closed under complementation.

(iii) $A_1, A_2, ..., \in \mathcal{B} \Longrightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$; i.e., $\mathcal{B}$ is closed under countable unions.

Remark: At this point, we can think of $\mathcal{B}$ simply as a collection of events to which we can "assign" probability (so that mathematical contradictions are avoided). If $A \in \mathcal{B}$, we say that A is a $\mathcal{B}$-measurable set.

Results: Suppose that $\mathcal{B}$ is a σ-algebra on S.

- Because $S = \emptyset^c$, we see that $S \in \mathcal{B}$ (because $\mathcal{B}$ is closed under complementation).

- If $A_1, A_2, ..., \in \mathcal{B}$, then $A_1^c, A_2^c, ..., \in \mathcal{B}$; see Property (ii);

$$\implies \bigcup_{i=1}^{\infty} A_i^c \in \mathcal{B} \qquad \text{Property (iii)}$$

$$\implies \left(\bigcup_{i=1}^{\infty} A_i^c\right)^c = \bigcap_{i=1}^{\infty} A_i \in \mathcal{B} \qquad \text{DeMorgan's Law; Property (ii)}$$

  Therefore, a σ-algebra $\mathcal{B}$ is also closed under countable intersections.

Examples:

1. $\mathcal{B} = \{\emptyset, S\}$. This is called the "trivial σ-algebra."
2. $\mathcal{B} = \{\emptyset, A, A^c, S\}$. This is the smallest σ-algebra that contains A; denoted by $\sigma(A)$.
3. $|S| = n$; i.e., the sample space is finite and contains n outcomes. Define

$$\mathcal{B} = 2^S$$

   to be the set of all subsets of S. This is called the **power set** of S. If $|S| = n$, then $\mathcal{B} = 2^S$ contains 2^n sets (this can be proven using induction).

Example 1.6. Suppose $S = \{1, 2, 3\}$. The power set of S is

$$\mathcal{B} = 2^S = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, S, \emptyset\}$$

Note that $\mathcal{B} = 2^S$ contains $2^3 = 8$ sets.

Remark: For a given sample space S, there are potentially many σ-algebras. For example, in Example 1.6, both

$$\begin{aligned} \mathcal{B}_0 &= \{\emptyset, S\} \\ \mathcal{B}_1 &= \{\emptyset, \{1\}, \{2, 3\}, S\} \end{aligned}$$

are also σ-algebras on S. Note also that $\mathcal{B}_0 \subset \mathcal{B}_1 \subset \mathcal{B} = 2^S$. We call $\mathcal{B}_0$ and $\mathcal{B}_1$ **sub-σ-algebras** of S.

Remark: We will soon learn that a σ-algebra contains all sets (events) to which we can assign probability. To illustrate, with the (sub)-σ-algebra

$$\mathcal{B}_1 = \{\emptyset, \{1\}, \{2, 3\}, S\},$$

we could assign a probability to the set $\{2, 3\}$ because it is measurable with respect to this σ-algebra. However, we could not assign a probability to the event $\{2\}$ using $\mathcal{B}_1$; it is not $\mathcal{B}_1$-measurable. Of course, $\{2\}$ is $\mathcal{B}$-measurable.

Definition: Suppose S is a sample space. Let $\mathcal{A}$ be a collection of subsets of S. The σ-algebra **generated** by $\mathcal{A}$ is the smallest σ-algebra that contains $\mathcal{A}$. We denote this σ-algebra by $\sigma(\mathcal{A})$. In other words,

- $\mathcal{A} \subset \sigma(\mathcal{A})$, and
- if $\Sigma(\mathcal{A})$ is another σ-algebra on S that contains $\mathcal{A}$, then $\sigma(\mathcal{A}) \subseteq \Sigma(\mathcal{A})$.

Example 1.7. Suppose $S = \{1, 2, 3\}$. Suppose $\mathcal{A} = \{1\}$. Then

$$\sigma(\mathcal{A}) = \sigma(\{1\}) = \{\emptyset, \{1\}, \{2, 3\}, S\}.$$

Note that $\Sigma(\mathcal{A}) = 2^S$ (the power set) also contains $\mathcal{A}$. However, $\sigma(\mathcal{A}) \subseteq \Sigma(\mathcal{A})$.

Definition: Suppose $S = \mathbb{R} = (-\infty, \infty)$. Consider the collection of sets

$$\mathcal{A} = \{(a, b) : -\infty < a < b < \infty\},$$

that is, the collection of all open intervals on $\mathbb{R}$. An important σ-algebra on $\mathbb{R}$ is $\sigma(\mathcal{A})$. This σ-algebra is called the **Borel σ-algebra** on $\mathbb{R}$. Any set $B \in \sigma(\mathcal{A})$ is called a **Borel set**.

Remark: The Borel σ-algebra on $\mathbb{R}$ is commonly denoted by $\mathcal{B}(\mathbb{R})$. It contains virtually any subset of $\mathbb{R}$ that you could imagine; sets like $[a, b]$, $(a, b]$, $[a, b)$, $(-\infty, b]$, and $[a, \infty)$, and unions, intersections, and complements of these sets. It is possible to find subsets of $\mathbb{R}$ that are not Borel sets. However, these are pathological in nature, and we will ignore them.

Kolmogorov's Axioms: Suppose that S is a sample space and let $\mathcal{B}$ be a σ-algebra on S. Let P be a set function; i.e.,

$$P : \mathcal{B} \to [0, 1],$$

that satisfies the following axioms:

1. $P(A) \geq 0$, for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, ..., \in \mathcal{B}$ are pairwise disjoint, then

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

We call P a **probability set function** (or a **probability measure**). The third axiom is sometimes called the "Axiom of Countable Additivity." The domain of P must be restricted to a σ-algebra to avoid mathematical contradictions.

Definition: The triple $(S, \mathcal{B}, P)$ is called a **probability space**. It consists of

S = a sample space (collection of all outcomes)
$\mathcal{B}$ = a σ-algebra (collection of events to which you can assign probability)
P = a probability measure.

Example 1.8. Consider the probability space $(S, \mathcal{B}, P)$, where

- $S = \{\text{H}, \text{T}\}$
- $\mathcal{B} = \{\emptyset, \{\text{H}\}, \{\text{T}\}, S\}$
- $P : \mathcal{B} \to [0, 1]$, defined by $P(\{\text{H}\}) = 2/3$ and $P(\{\text{T}\}) = 1/3$.

This is a probability model for a random experiment where an unfair coin is flipped.

Example 1.9. Consider the probability space $(S, \mathcal{B}, P)$, where

- $S = (a, b)$, where $-\infty < a < b < \infty$
- $\mathcal{B} = \mathcal{B}(\mathbb{R}) \cap S = \{B \cap S : B \in \mathcal{B}(\mathbb{R})\}$; i.e., the Borel sets on S
- $P : \mathcal{B} \to [0, 1]$, defined for all $A \in \mathcal{B}$,

$$P(A) = \int_A \frac{1}{b-a}\, dx \quad (\text{"uniform" probability measure}).$$

For example, suppose $a = 0$, $b = 10$, and $A = \{\omega : 2 < \omega \leq 5\}$. Then $P(A) = 3/10$.

Theorem 1.2.6. Suppose that $S = \{\omega_1, \omega_2, ..., \omega_n\}$ is a finite sample space. Let $\mathcal{B}$ be a σ-algebra on S (e.g., $\mathcal{B} = 2^S$, etc.). We can construct a probability measure over $\mathcal{B}$ as follows:

1. Assign the "weight" or "mass" $p_i \geq 0$ to the outcome ω_i where $\sum_{i=1}^n p_i = 1$.
2. For any $A \in \mathcal{B}$, define

$$P(A) = \sum_{i=1}^n p_i I(\omega_i \in A),$$

where $I(\cdot)$ is the **indicator function**; i.e.,

$$I(\omega_i \in A) = \begin{cases} 1, & \omega_i \in A \\ 0, & \omega_i \notin A. \end{cases}$$

We now show that this "construction" of P satisfies the Kolmogorov Axioms on $(S, \mathcal{B})$.

Proof. Suppose $A \in \mathcal{B}$. By definition,

$$P(A) = \sum_{i=1}^n p_i I(\omega_i \in A) \geq 0$$

because both $p_i \geq 0$ and $I(\omega_i \in A) \geq 0$ $\forall i$. This establishes Axiom 1. To establish Axiom 2, simply note that

$$P(S) = \sum_{i=1}^{n} p_i I(\omega_i \in S) = \sum_{i=1}^{n} p_i = 1.$$

To establish Axiom 3, suppose that $A_1, A_2, ..., A_k \in \mathcal{B}$ are pairwise disjoint (a finite sequence suffices here because S is finite). Note that

$$\begin{aligned} P\left(\bigcup_{j=1}^{k} A_j\right) &= \sum_{i=1}^{n} p_i I\left(\omega_i \in \bigcup_{j=1}^{k} A_j\right) \\ &= \sum_{i=1}^{n} p_i \sum_{j=1}^{k} I(\omega_i \in A_j) \\ &= \sum_{j=1}^{k} \sum_{i=1}^{n} p_i I(\omega_i \in A_j) = \sum_{j=1}^{k} P(A_j). \end{aligned}$$

Therefore, the Kolmogorov Axioms are satisfied. □

Special case: Suppose $S = \{\omega_1, \omega_2, ..., \omega_n\}$, $\mathcal{B} = 2^S$, and $p_i = 1/n$ for each $i = 1, 2, ..., n$. That is, each outcome $\omega_i \in S$ receives the same probability weight. This is called an **equiprobability model**. Under an equiprobability model,

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

When outcomes (in a finite S) have the same probability, calculating $P(A)$ is "easy." We simply have two counting problems to solve; the first where we count the number of outcomes $\omega_i \in A$ and the second where we count the number of outcomes $\omega_i \in S$.

Example 1.10. Draw 5 cards from a standard deck of 52 cards (without replacement). What is the probability of getting "3 of a kind?" Here we can conceptualize the sample space as

$$S = \{[2_S, 2_D, 2_H, 2_C, 3_S], [2_S, 2_D, 2_H, 2_C, 3_D], ..., [\mathrm{A}_S, \mathrm{A}_D, \mathrm{A}_H, \mathrm{A}_C, \mathrm{K}_C]\}.$$

This is a finite sample space, and there are

$$|S| = \binom{52}{5} = 2598960$$

outcomes in S. Assume an equiprobability model with $\mathcal{B} = 2^S$. Define $A = \{\text{3 of a kind}\}$. How many ways can A occur?

$$|A| = \binom{13}{1}\binom{4}{3}\binom{12}{2}\binom{4}{1}^2 = 54912.$$

Therefore, assuming an equiprobability model,

$$P(A) = \frac{54912}{2598960} \approx 0.0211.$$

Remark: If $|S| = n < \infty$, we call S a **discrete sample space**. We also use this terminology if $|S| = +\infty$, but S is countable, e.g., $S = \{\omega_1, \omega_2, ..., \}$. For a σ-algebra on S (e.g., $\mathcal{B} = 2^S$, etc.), if we assign $P(\{\omega_i\}) = p_i \geq 0$, where $\sum_{i=1}^{\infty} p_i = 1$, the Kolmogorov Axioms are still satisfied for this construction.

Example 1.11. Suppose that $S = \{1, 2, 3, ..., \}$, $\mathcal{B} = 2^S$, and

$$P(\{i\}) = p_i = (1-p)^{i-1}p,$$

where $0 < p < 1$. This is called a **geometric probability measure**. Note that

$$P(S) = \sum_{i=1}^{\infty} p_i = \sum_{i=1}^{\infty} (1-p)^{i-1}p = p\sum_{j=0}^{\infty}(1-p)^j = p\left[\frac{1}{1-(1-p)}\right] = 1.$$

The Calculus of Probabilities: We now examine many results that follow from the Kolmogorov Axioms. In what follows, let S denote a sample space, $\mathcal{B}$ denote a σ-algebra on S, and P denote a probability measure. All events (e.g., A, B, C, etc.) are assumed to be measurable (i.e., $A \in \mathcal{B}$, etc).

Theorem 1.2.8.

(a) $P(\emptyset) = 0$

(b) $P(A) \leq 1$

(c) **Complement Rule:** $P(A^c) = 1 - P(A)$.

Proof. To prove part (c), write $S = A \cup A^c$ and apply Axioms 2 and 3. Part (b) then follows from Axiom 1. To prove part (a), note that $S^c = \emptyset$ and use part (c). □

Theorem 1.2.9.

(a) $P(A^c \cap B) = P(B) - P(A \cap B)$

(b) **Inclusion-Exclusion:** $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(c) **Monotonicity:** If $A \subseteq B$, then $P(A) \leq P(B)$.

Proof. To prove part (a), write $B = (A \cap B) \cup (A^c \cap B)$ and apply Axiom 3. To prove part (b), write $A \cup B = A \cup (A^c \cap B)$ and combine with part (a). To prove part (c), write $B = A \cup (A^c \cap B)$. This is true because $A \subseteq B$ by assumption. □

Remark: The identity

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

is called the **inclusion-exclusion** identity (for two events). Because $P(A \cup B) \leq 1$, it follows immediately that

$$P(A \cap B) \geq P(A) + P(B) - 1.$$

This is a special case of **Bonferroni's Inequality** (for two events).

Theorem 1.2.11.

(a) $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$, where $C_1, C_2, ...,$ is any partition of S.

(b) **Boole's Inequality:** For any sequence $A_1, A_2, ...,$

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

Proof. To prove part (a), write

$$A = A \cap S = A \cap \left(\bigcup_{i=1}^{\infty} C_i\right) = \bigcup_{i=1}^{\infty} (A \cap C_i),$$

and apply Axiom 3. We will prove Boole's Inequality later. □

Two additional results:

1. **Inclusion-Exclusion:** For any sequence $A_1, A_2, ..., A_n$,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum\sum_{i_1<i_2} P(A_{i_1} \cap A_{i_2}) + \sum\sum\sum_{i_1<i_2<i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} A_i\right).$$

2. **Continuity:** If $A_n \to A$, as $n \to \infty$, then $P(A_n) \to P(A)$; i.e.,

$$\lim_{n\to\infty} P(A_n) = P\left(\lim_{n\to\infty} A_n\right) = P(A).$$

Proof of Inclusion-Exclusion. If $\omega \notin A_i \ \forall i = 1, 2, ..., n$, then LHS = RHS = 0 and the result holds. Otherwise, suppose that ω is in exactly $m > 0$ of the events $A_1, A_2, ..., A_n$. Clearly, $\omega \in \cup_{i=1}^{n} A_i$ so the probability associated with ω is counted once on the LHS. For the RHS, consider the k-fold intersection $A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}$, where $i_1 < i_2 < \cdots < i_k$ and $k \leq m$. The

outcome ω is in exactly $\binom{m}{k}$ intersections of this type. Therefore, the probability associated with ω is counted

$$\binom{m}{1} - \binom{m}{2} + \binom{m}{3} - \cdots \pm \binom{m}{m}$$

times on the RHS. Therefore, it suffices to show that

$$1 = \binom{m}{1} - \binom{m}{2} + \binom{m}{3} - \cdots \pm \binom{m}{m}$$

or, equivalently, $\sum_{i=0}^{m} \binom{m}{i}(-1)^i = 0$. However, this is true from the **binomial theorem**; viz.,

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

by taking $a = -1$ and $b = 1$. Because ω was arbitrarily chosen, we are done. □

Proof of Continuity. Although this result does hold in general, we will assume that $\{A_n\}$ is monotone increasing (non-decreasing). Recall that if $\{A_n\}$ is increasing, then

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i.$$

Define the "ring-type" sets $R_1 = A_1$, $R_2 = A_2 \cap A_1^c$, $R_3 = A_3 \cap A_2^c$, ..., and so on. In general,

$$R_n = A_n \cap A_{n-1}^c,$$

for $n = 2, 3, ...,$. It is easy to see that $R_i \cap R_j = \emptyset \ \forall i \neq j$ (i.e., the ring sets are pairwise disjoint) and

$$\bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} R_n.$$

Therefore,

$$P\left(\lim_{n\to\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} R_n\right) = \sum_{n=1}^{\infty} P(R_n) = \lim_{n\to\infty} \sum_{j=1}^{n} P(R_j). \tag{1.1}$$

Now, recall that $P(R_1) = P(A_1)$ and that $P(R_j) = P(A_j \cap A_{j-1}^c) = P(A_j) - P(A_{j-1})$ by Theorem 1.2.9 (a). Therefore, the expression in Equation (1.1) equals

$$\lim_{n\to\infty} \left\{ P(A_1) + \sum_{j=2}^{n} [P(A_j) - P(A_{j-1})] \right\} = \lim_{n\to\infty} P(A_n).$$

We have shown that $P(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n)$. Thus, we are done. □

Remark: As an exercise, try to establish the continuity result when $\{A_n\}$ is a decreasing (non-increasing) sequence. The result does hold in general and would be proven in a more advanced course.

Proof of Boole's Inequality. Define $B_n = \cup_{i=1}^{n} A_i$. Clearly, $\{B_n\}$ is an increasing sequence of sets and $B_n \to \cup_{n=1}^{\infty} A_n$, as $n \to \infty$. Note that

$$\begin{aligned} B_j = B_{j-1} \cup A_j &\implies P(B_j) \leq P(B_{j-1}) + P(A_j) \\ &\implies P(A_j) \geq P(B_j) - P(B_{j-1}). \end{aligned}$$

We have

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = P\left(\bigcup_{n=1}^{\infty} B_n\right) = P\left(\lim_{n\to\infty} B_n\right) = \lim_{n\to\infty} P(B_n)$$

because $\{B_n\}$ is increasing. However, note that

$$P(B_n) = P(B_1) + \sum_{j=2}^{n}[P(B_j) - P(B_{j-1})] \leq P(A_1) + \sum_{j=2}^{n} P(A_j) = \sum_{j=1}^{n} P(A_j).$$

Taking limits, we have

$$\lim_{n\to\infty} P(B_n) \leq \lim_{n\to\infty} \sum_{j=1}^{n} P(A_j) = \sum_{n=1}^{\infty} P(A_n). \quad \square$$

Bonferroni's Inequality: Suppose that $A_1, A_2, ..., A_n$ is a sequence of events. Then

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

Proof. By Boole's Inequality (applied to the sequence $A_1^c, A_2^c, ..., A_n^c$),

$$P\left(\bigcup_{i=1}^{n} A_i^c\right) \leq \sum_{i=1}^{n} P(A_i^c).$$

Recalling that $\bigcup_{i=1}^{n} A_i^c = (\bigcap_{i=1}^{n} A_i)^c$ and $P(A_i^c) = 1 - P(A_i)$, the last inequality becomes

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

Rearranging terms gives the result. $\square$

Example 1.12. *The matching problem.* At a party, suppose that each of n men throws his hat into the center of a room. The hats are mixed up and then each man randomly selects a hat. What is the probability that at least one man selects his own hat? In other words, what is the probability that there is at least one "match?"

Solution: We can conceptualize the sample space as the set of all permutations of $\{1, 2, ..., n\}$. There are $n!$ such permutations. We assume that each of the $n!$ permutations is equally likely. In notation, $S = \{\omega_1, \omega_2, ..., \omega_N\}$, where ω_j is the jth permutation of $\{1, 2, ..., n\}$

and $N = n!$. Define the event $A_i = \{i\text{th man selects his own hat}\}$ and the event $A = \{\text{at least one match}\}$ so that

$$A = \bigcup_{i=1}^{n} A_i \implies P(A) = P\left(\bigcup_{i=1}^{n} A_i\right).$$

We now use inclusion-exclusion. Note the following:

$$\begin{aligned} P(A_i) &= \frac{(n-1)!}{n!} = \frac{1}{n} \qquad \forall i = 1, 2, ..., n \\ P(A_{i_1} \cap A_{i_2}) &= \frac{(n-2)!}{n!} \qquad 1 \leq i_1 < i_2 \leq n \\ P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) &= \frac{(n-3)!}{n!} \qquad 1 \leq i_1 < i_2 < i_3 \leq n \end{aligned}$$

This pattern continues; the probability of the n-fold intersection is

$$P\left(\bigcap_{i=1}^{n} A_i\right) = \frac{(n-n)!}{n!} = \frac{1}{n!}.$$

Therefore, by inclusion-exclusion, we have

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} P(A_i) - \sum\sum_{i_1<i_2} P(A_{i_1} \cap A_{i_2}) \\ &\quad + \sum\sum\sum_{i_1<i_2<i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} A_i\right) \\ &= n\left(\frac{1}{n}\right) - \binom{n}{2}\frac{(n-2)!}{n!} + \binom{n}{3}\frac{(n-3)!}{n!} - \cdots + (-1)^{n+1}\frac{1}{n!} \\ &= \sum_{k=1}^{n} (-1)^{k+1}\binom{n}{k}\frac{(n-k)!}{n!} \\ &= \sum_{k=1}^{n} (-1)^{k+1}\frac{n!}{k!(n-k)!}\frac{(n-k)!}{n!} = 1 - \sum_{k=0}^{n}\frac{(-1)^k}{k!}. \end{aligned}$$

Interestingly, note that as $n \to \infty$,

$$\lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} A_i\right) = \lim_{n\to\infty}\left[1 - \sum_{k=0}^{n}\frac{(-1)^k}{k!}\right] = 1 - \sum_{k=0}^{\infty}\frac{(-1)^k}{k!} = 1 - e^{-1} \approx 0.6321.$$

Some view this answer to be unexpected, believing that this probability should tend to zero (because the number of attendees becomes large) or that it should tend to one (because there are more opportunities for a match). The truth lies somewhere in the middle.

Recall: The McLaurin series expansion of $f(x) = e^x$ is

$$e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots + .$$

Exercise: Define $B = \{\text{exactly } k \text{ men select their own hats}\}$. Find $P(B)$ for n large (relative to k). *Answer:* $e^{-1}/k!$.

1.3 Conditional Probability and Independence

Definition: Consider a random experiment described by $(S, \mathcal{B}, P)$. The conditional probability of $A \in \mathcal{B}$, given that $B \in \mathcal{B}$ has occurred, can be computed

- over $(S, \mathcal{B}, P)$ using
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$
- over $(B, \mathcal{B}^*, P_B)$, where $\mathcal{B}^* = \{B \cap C : C \in \mathcal{B}\}$ and where P_B and $P(\cdot|B)$ are related by
$$P(A|B) = P_B(A \cap B) \quad \forall (A \cap B) \in \mathcal{B}^*.$$

Exercise: Show that $\mathcal{B}^*$ is a σ-algebra on B.

Example 1.13. Experiment: Toss two coins. Assume the model

$$\begin{aligned} S &= \{(\text{HH}), (\text{HT}), (\text{TH}), (\text{TT})\} \\ \mathcal{B} &= 2^S \\ P &= \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/4, \text{ for all } \omega \in S. \end{aligned}$$

Define

$$\begin{aligned} A &= \{(\text{HH}), (\text{HT})\} \\ B &= \{(\text{HH}), (\text{HT}), (\text{TH})\}. \end{aligned}$$

We can calculate $P(A|B)$ in our two ways:

- Over $(S, \mathcal{B}, P)$,
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{2/4}{3/4} = \frac{2}{3}.$$
- Over $(B, \mathcal{B}^*, P_B)$, where
$$\begin{aligned} \mathcal{B}^* &= \{B \cap C : C \in \mathcal{B}\} \\ &= \{\emptyset, \text{B}, \{(\text{HH})\}, \{(\text{HT})\}, \{(\text{TH})\}, \{(\text{HH}), (\text{HT})\}, \{(\text{HH}), (\text{TH})\}, \{(\text{HT}), (\text{TH})\}\} \end{aligned}$$
and P_B is an equiprobability measure; i.e., $P_B(\{\omega\}) = 1/3 \ \ \forall \omega \in B$. Note that $\mathcal{B}^*$ has $2^3 = 8$ sets and that $\mathcal{B}^* \subset \mathcal{B}$. We see that
$$P(A|B) = P_B(A \cap B) = P_B(\{(\text{HH}), (\text{HT})\}) = \frac{2}{3}.$$

Remark: In practice, it is often easier to work on $(S, \mathcal{B}, P)$, the original probability space, and compute conditional probabilities using

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

If $P(B) = 0$, then $P(A|B)$ is not defined. Provided that $P(B) > 0$, the probability measure $P(\cdot|B)$ satisfies the Kolmogorov Axioms; i.e.,

1. $P(A|B) \geq 0$, for all $A \in \mathcal{B}$
2. $P(B|B) = 1$
3. If $A_1, A_2, ..., \in \mathcal{B}$ are pairwise disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \sum_{i=1}^{\infty} P(A_i|B).$$

Proving this is left as an exercise.

Important: Because $P(\cdot|B)$ is a bona fide probability measure on $(S, \mathcal{B})$, it satisfies all of the probability rules that we stated and derived in §1.2. For example,

1. $P(A^c|B) = 1 - P(A|B)$
2. $P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B) - P(A_1 \cap A_2|B)$
3. For any sequence $A_1, A_2, ...,$

$$P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) \leq \sum_{i=1}^{\infty} P(A_i|B).$$

These are the "conditional versions" of the complement rule, inclusion-exclusion, and Boole's Inequality, respectively.

Law of Total Probability: Suppose $(S, \mathcal{B}, P)$ is a model for a random experiment. Let $B_1, B_2, ..., \in \mathcal{B}$ denote a partition of S; i.e., $B_i \cap B_j = \emptyset \; \forall i \neq j$ and $\cup_{i=1}^{\infty} B_i = S$. Then,

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i).$$

The first equality is simply Theorem 1.2.11(a). The second equality arises by noting that

$$P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)} \implies \underbrace{P(A \cap B_i) = P(A|B_i)P(B_i)}_{\text{"multiplication rule"}}.$$

Example 1.14. *Diagnostic testing.* A lab test is 95% effective at detecting a disease when it is present. It is 99% effective at declaring a subject negative when the subject is truly negative. If 8% of the population is truly positive, what is the probability a randomly selected subject will test positively?

Solution. Define the events

$$\begin{aligned} D &= \{\text{disease is present}\} \\ \maltese &= \{\text{test is positive}\}. \end{aligned}$$

We are given

$$\begin{aligned} P(\maltese|D) &= 0.95 \quad \text{(sensitivity)} \\ P(\maltese^c|D^c) &= 0.99 \quad \text{(specificity)} \\ P(D) &= 0.08 \quad \text{(prevalence)} \end{aligned}$$

The probability a randomly selected subject will test positively is

$$\begin{aligned} P(\maltese) &= P(\maltese|D)P(D) + P(\maltese|D^c)P(D^c) \\ &= (0.95)(0.08) + (0.01)(0.92) \approx 0.0852. \end{aligned}$$

Note that we have used LOTP with the partition $\{D, D^c\}$.

Question: What is the probability that a subject has the disease (D) if his test is positive?
Solution. We want to calculate

$$\begin{aligned} P(D|\maltese) &= \frac{P(D \cap \maltese)}{P(\maltese)} \\ &= \frac{P(\maltese|D)P(D)}{P(\maltese|D)P(D) + P(\maltese|D^c)P(D^c)} \\ &= \frac{(0.95)(0.08)}{(0.95)(0.08) + (0.01)(0.92)} \approx 0.892. \end{aligned}$$

Note: $P(D|\maltese)$ in this example is called the **positive predictive value** (PPV). As an exercise, calculate $P(D^c|\maltese^c)$, the **negative predictive value** (NPV).

Remark: We have just discovered a special case of **Bayes' Rule**, which allows us to "update" probabilities on the basis of observed information (here, the test result):

Prior probability		Test result		Posterior probability
$P(D) = 0.08$	$\longrightarrow$	$\maltese$	$\longrightarrow$	$P(D\|\maltese) \approx 0.892$
$P(D) = 0.08$	$\longrightarrow$	$\maltese^c$	$\longrightarrow$	$P(D\|\maltese^c) \approx 0.004$

Bayes' Rule: Suppose $(S, \mathcal{B}, P)$ is a model for a random experiment. Let $B_1, B_2, ..., \in \mathcal{B}$ denote a partition of S. Then,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{\infty} P(A|B_j)P(B_j)}.$$

This formula allows us to update our belief about the probability of B_i on the basis of observing A. In general,

$$P(B_i) \quad \longrightarrow \quad A \text{ occurs} \quad \longrightarrow \quad P(B_i|A).$$

Multiplication Rule: For two events A, $B \in \mathcal{B}$,

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

For n events $A_1, A_2, ..., A_n \in \mathcal{B}$,

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_i) \times P(A_2|A_1) \times P(A_3|A_1 \cap A_2) \times \cdots \times P\left(A_n \middle| \bigcap_{i=1}^{n-1} A_i\right).$$

Proving this is an easy induction argument.

Definition: Two events A, $B \in \mathcal{B}$ are **independent** if

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

This definition implies (if conditioning events have strictly positive probability):

$$\begin{aligned} P(A|B) &= P(A) \\ P(B|A) &= P(B). \end{aligned}$$

Theorem 1.3.9. If A, $B \in \mathcal{B}$ are independent events, then so are

(a) A and B^c

(b) A^c and B

(c) A^c and B^c.

Generalization: A collection of events $A_1, A_2, ..., A_n \in \mathcal{B}$ are **mutually independent** if for any sub-collection $A_{i_1}, A_{i_2}, ..., A_{i_k}$, we have

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j}).$$

Special case: Three events A_1, A_2, and A_3. For these events to be mutually independent, we need them to be **pairwise independent**:

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1)P(A_2) \\ P(A_1 \cap A_3) &= P(A_1)P(A_3) \\ P(A_2 \cap A_3) &= P(A_2)P(A_3) \end{aligned}$$

and we also need $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$. Note that is possible for

- A_1, A_2, and A_3 to be pairwise independent but $P(A_1 \cap A_2 \cap A_3) \neq P(A_1)P(A_2)P(A_3)$. See Example 1.3.11 (pp 26 CB).

- $P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)$ but A_1, A_2, and A_3 are not pairwise independent. See Example 1.3.10 (pp 25-26 CB).

Example 1.15. Experiment: Observe the chlamydia status of $n = 30$ USC students. Here, we can conceptualize the sample space as

$$S = \{(0,0,0,...,0), (1,0,0,...,0), (0,1,0,...,0), ..., (1,1,1,...,1)\},$$

where "0" denotes a negative student and "1" denotes a positive student. Note that there are $|S| = 2^{30} = 1{,}073{,}741{,}824$ outcomes in S. Suppose that $\mathcal{B} = 2^S$ and that P is a probability measure that satisfies $P(A_i) = p$, where $A_i = \{i\text{th student is positive}\}$, for $i = 1, 2, ..., 30$, and $0 < p < 1$. Assume that $A_1, A_2, ..., A_{30}$ are mutually independent events.

Question: What is the probability that at least one student is positive?
Solution. Clearly, $P(A_i^c) = 1 - P(A_i) = 1 - p$, for $i = 1, 2, ..., 30$. Therefore,

$$\begin{aligned} P\left(\bigcup_{i=1}^{30} A_i\right) &= 1 - P\left(\bigcap_{i=1}^{30} A_i^c\right) \\ &= 1 - \prod_{i=1}^{30} P(A_i^c) \\ &= 1 - (1-p)^{30}. \end{aligned}$$

Question: What is the probability that exactly k students are positive?
Solution. Consider any outcome $\omega \in S$ containing exactly k 1's and $30 - k$ 0's. Any such outcome has probability $p^k(1-p)^{30-k}$ because individual statuses are mutually independent. Because there are $\binom{30}{k}$ such ω's in S that have exactly k 1's, the desired probability is

$$\binom{30}{k} p^k (1-p)^{30-k}.$$

This expression is valid for $k = 0, 1, 2, ..., 30$. For example, if $p = 0.10$ and $k = 3$, then $P(\text{exactly 3 positives}) \approx 0.236$.

1.4 Random Variables

Remark: In Example 1.15, the underlying sample space

$$S = \{(0,0,0,...,0), (1,0,0,...,0), (0,1,0,...,0), ..., (1,1,1,...,1)\}$$

contained $|S| = 2^{30} = 1{,}073{,}741{,}824$ outcomes. In most situations, it is easier to work with numerical valued functions of the outcomes, such as

$$X = \text{ number of positives (out of 30).}$$

We see that the "sample space" for X is

$$\mathcal{X} = \{x : x = 0, 1, 2, ..., 30\}.$$

Note that $\mathcal{X}$ is much easier to work with than S.

Definition: Let $(S, \mathcal{B}, P)$ be a probability space for a random experiment. The function

$$X : S \to \mathbb{R}$$

is called a **random variable** on $(S, \mathcal{B}, P)$ if

$$X^{-1}(B) \equiv \{\omega \in S : X(\omega) \in B\} \in \mathcal{B} \tag{1.2}$$

for all $B \in \mathcal{B}(\mathbb{R})$, where $\mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$. The set $X^{-1}(B)$ is called the **inverse image** of B (under the mapping X). The condition in (1.2) says that the inverse image of any Borel set B is measurable with respect to $\mathcal{B}$. Note that the notion of probability does not enter into the condition in (1.2).

Remark: The main point is that a random variable X, mathematically, is a function whose domain is S and whose range is $\mathbb{R}$. For example, in Example 1.15,

$$\begin{aligned} X((0,0,0,...,0)) &= 0 \\ X((1,0,0,...,0)) &= 1 \\ X((1,1,1,...,1)) &= 30. \end{aligned}$$

Notes: Suppose that X is a random variable on $(S, \mathcal{B}, P)$; i.e., $X : S \to \mathbb{R}$.

1. The collection of sets $\sigma(X) \equiv \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$ is a σ-algebra on S.

2. The measurability condition

   $$X^{-1}(B) \equiv \{\omega \in S : X(\omega) \in B\} \in \mathcal{B},$$

   for all $B \in \mathcal{B}(\mathbb{R})$, suggests that events of interest like $\{X \in B\}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ can be assigned probability in the same way that $\{\omega \in S : X(\omega) \in B\}$ can be assigned probability on $(S, \mathcal{B})$.

3. In a more advanced course, we might say that "X is a $\mathcal{B} - \mathcal{B}(\mathbb{R})$ measurable mapping from $S \to \mathbb{R}$."

Example 1.16. Consider a random experiment with

$$\begin{aligned} S &= \{1, 2, 3\} \\ \mathcal{B}_1 &= \{\emptyset, \{1\}, \{2, 3\}, S\} \\ P &= \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/3, \text{ for all } \omega \in S. \end{aligned}$$

Define the function X so that $X(1) = X(2) = 0$ and $X(3) = 1$. Consider the Borel set $B = \{0\}$. Note that

$$X^{-1}(B) = X^{-1}(\{0\}) = \{\omega \in S : X(\omega) = 0\} = \{1, 2\} \notin \mathcal{B}_1.$$

Therefore, the function X is not a random variable on $(S, \mathcal{B}_1)$. It does not satisfy the measurability condition. **Question:** Is X a random variable on $(S, \mathcal{B})$, where $\mathcal{B} = 2^S$?

Discrete sample spaces: Consider an experiment described by $(S, \mathcal{B}, P)$ where

$$\begin{aligned} S &= \{\omega_1, \omega_2, ..., \omega_n\} \\ \mathcal{B} &= 2^S \\ P &= \text{a valid probability measure.} \end{aligned}$$

Here, we allow for both cases:

- $n < \infty \implies S$ finite
- "$n = \infty$" $\implies S$ countable (i.e., countably infinite).

Suppose X is a random variable on $(S, \mathcal{B}, P)$ with range $\mathcal{X} = \{x_1, x_2, ..., x_m\}$. We call $\mathcal{X}$ the **support** of the random variable X; we allow for both cases:

- $m < \infty \implies$ "finite support"
- "$m = \infty$" $\implies$ "countably infinite support."

Define a new probability measure P_X according to

$$P_X(X = x_i) = \underbrace{P(\{\omega \in S : X(\omega) = x_i\})}_{\text{a probability on } (S, \mathcal{B})}.$$

We call P_X an **induced probability measure**, because it is a measure "induced" by the random variable X. It is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, and $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ is a probability space. We often use the terminology:

$$\begin{aligned} (S, \mathcal{B}, P) &\implies \text{domain space} \\ (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X) &\implies \text{range space.} \end{aligned}$$

Remark: The probability measure P_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfies the Kolmogorov Axioms (i.e., it is a valid probability measure).

Example 1.17. Experiment: Toss a fair coin twice. Consider the model described by

$$\begin{aligned} S &= \{(\text{HH}), (\text{HT}), (\text{TH}), (\text{TT})\} \\ \mathcal{B} &= 2^S \\ P &= \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/4, \text{ for all } \omega \in S. \end{aligned}$$

Define a random variable X on $(S, \mathcal{B}, P)$ by

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

The random variable X satisfies

ω	(HH)	(HT)	(TH)	(TT)
$X(\omega)$	2	1	1	0

Therefore, the support of X is $\mathcal{X} = \{x : x = 0, 1, 2\}$ and

$$\begin{aligned} P_X(X = 0) &= P(\{\omega \in S : X(\omega) = 0\}) = P(\{(\text{TT})\}) = \frac{1}{4} \\ P_X(X = 1) &= P(\{\omega \in S : X(\omega) = 1\}) = P(\{(\text{HT}), (\text{TH})\}) = \frac{2}{4} \\ P_X(X = 2) &= P(\{\omega \in S : X(\omega) = 2\}) = P(\{(\text{HH})\}) = \frac{1}{4}. \end{aligned}$$

We have the following **probability distribution** for the random variable X:

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

Important: We use upper case notation X to denote a random variable. A realization of X is denoted by $X(\omega) = x$ (i.e., lower case).

Remark: In practice, we are often given the probability measure P_X in the form of an assumed probability distribution for X (e.g., binomial, normal, etc.) and our "starting point" actually becomes $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$. For example, in Example 1.15, we calculated

$$P_X(X = x) = \binom{30}{x} p^x (1-p)^{30-x},$$

for $x \in \mathcal{X} = \{x : x = 0, 1, 2, ..., 30\}$. With this already available, there is little need to refer to the underlying experiment described by $(S, \mathcal{B}, P)$.

Example 1.18. Suppose that X denotes the systolic blood pressure for a randomly selected patient. Suppose it is assumed that for all $B \in \mathcal{B}(\mathbb{R})$,

$$P_X(B) \equiv P_X(X \in B) = \int_B \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}}_{= f_X(x), \text{ say.}} dx.$$

The function $f_X(x)$ is given without reference to the underlying experiment $(S, \mathcal{B}, P)$.

Remark: Suppose we have an experiment described by $(S, \mathcal{B}, P)$, and let $X : S \to \mathbb{R}$ be a random variable defined on this space. The induced probability measure P_X satisfies

$$\underbrace{P_X(X \in B)}_{\text{calculated on } (\mathbb{R}, \mathcal{B}(\mathbb{R}))} = \underbrace{P(\{\omega \in S : X(\omega) \in B\})}_{\text{calculated on } (S, \mathcal{B})}.$$

We should remember that although P and P_X are different probability measures, we will start to get "lazy" and write things like $P(X \in B)$, $P(0 < X \leq 4)$, $P(X = 3)$, etc. Although this is an abuse of notation, most textbook authors eventually succumb to this practice. In fact, the authors of CB stop writing P_X in favor of P after Chapter 1! In many ways, this is not surprising if we "start" by working on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to begin with. However, it is important to remember that P_X is a measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$; not P as we have defined it herein.

1.5 Distribution Functions

Definition: The **cumulative distribution function** (cdf) of a random variable X is

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x \in \mathbb{R}.$$

It is important to emphasize that $F_X(x)$ is defined for all $x \in \mathbb{R}$; not just for those values of $x \in \mathcal{X}$, the support of X.

Example 1.19. In Example 1.17, we worked with the random variable

$$X = \text{ number of heads observed (in two tosses)}$$

and calculated

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

The cdf of X is therefore

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

and is graphed in Figure 1.1 (next page).

Theorem 1.5.3. The function $F_X : \mathbb{R} \to [0, 1]$ is a cdf if and only if these conditions hold:

1. $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to \infty} F_X(x) = 1$
2. $F_X(x)$ is a non-decreasing function of x
3. $F_X(x)$ is right-continuous; i.e.,

$$\lim_{x \to x_0^+} F_X(x) = F_X(x_0) \quad \forall x_0 \in \mathbb{R}.$$

An alternate definition of right continuity is that $\lim_{n \to \infty} F_X(x_n) = F_X(x_0)$, for any real sequence $\{x_n\}$ such that $x_n \downarrow x_0$.

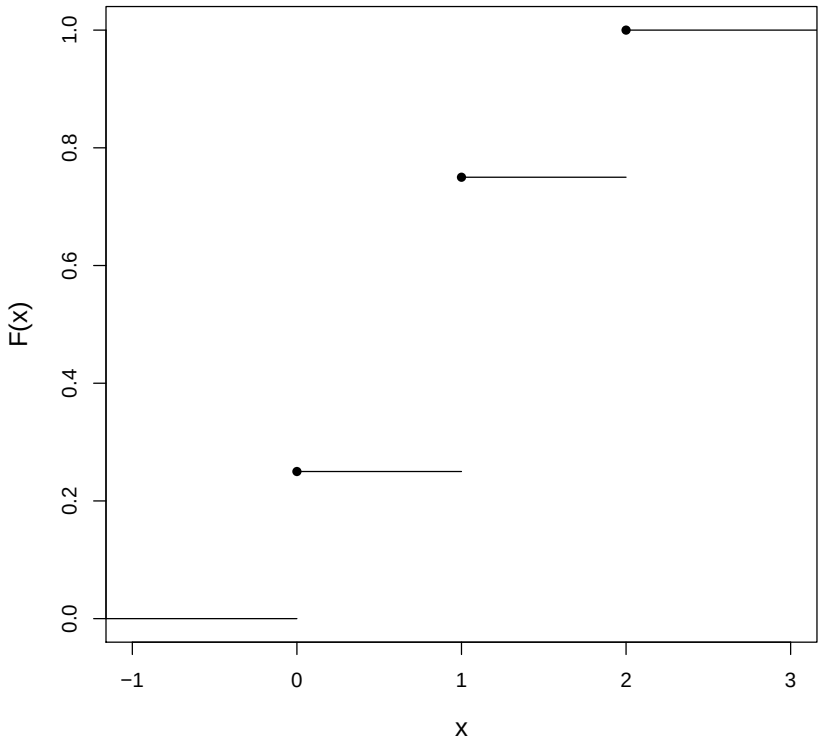


Figure 1.1: Cumulative distribution function $F_X(x)$ in Example 1.19.

Remark: Proving the necessity part ($\Longrightarrow$) part of Theorem 1.5.3 is not hard (we do this now). Establishing the sufficiency ($\Longleftarrow$) is harder. To do this, one would have to show there exists a sample space S, a probability measure P, and a random variable X on $(S, \mathcal{B}, P)$ such that $F_X(x)$ is the cdf of X. This argument involves measure theory, so we will avoid it.

Proof. ($\Longrightarrow$) Suppose that F_X is a cdf. To establish (1), suppose that $\{x_n\}$ is an increasing sequence of real numbers such that $x_n \to \infty$, as $n \to \infty$. Then $B_n = \{X \leq x_n\}$ is an increasing sequence of sets and

$$\lim_{n\to\infty} B_n = \lim_{n\to\infty} \{X \leq x_n\} = \bigcup_{n=1}^{\infty} \{X \leq x_n\} = \{X < \infty\}.$$

Therefore, using continuity of P_X, we have

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P_X(X \leq x_n) = P_X\left(\lim_{n\to\infty} \{X \leq x_n\}\right) = P_X(X < \infty) = P(S) = 1.$$

Because $\{x_n\}$ was arbitrary, we have established that $\lim_{x\to\infty} F_X(x) = 1$. Showing $\lim_{x\to-\infty} F_X(x) = 0$ is done similarly; just work with a decreasing sequence $\{x_n\}$. To establish (2), suppose that $x_1 \leq x_2$. Then $\{X \leq x_1\} \subseteq \{X \leq x_2\}$ and by monotonicity of P_X, we

have

$$F_X(x_1) = P_X(X \leq x_1) \leq P_X(X \leq x_2) = F_X(x_2).$$

Because x_1 and x_2 were arbitrary, this shows that $F_X(x)$ is a non-decreasing function of x. To establish (3), suppose that $\{x_n\}$ is a decreasing sequence of real numbers such that $x_n \to x_0$, as $n \to \infty$. Then $C_n = \{X \leq x_n\}$ is a decreasing sequence of sets and

$$\lim_{n\to\infty} C_n = \lim_{n\to\infty} \{X \leq x_n\} = \bigcap_{n=1}^{\infty} \{X \leq x_n\} = \{X \leq x_0\}.$$

Using continuity of P_X again, we have

$$\lim_{n\to\infty} F_X(x_n) = \lim_{n\to\infty} P_X(X \leq x_n) = P_X\left(\lim_{n\to\infty} \{X \leq x_n\}\right) = P_X(X \leq x_0) = F_X(x_0).$$

As x_0 was arbitrary, this establishes (3) and we are done. □

Example 1.20. Suppose that X is a random variable with cdf

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-x/\beta}, & x > 0, \end{cases}$$

where $\beta > 0$. This cdf corresponds to an **exponential distribution** and is graphed in Figure 1.2 (next page). It is easy to see that this function satisfies the three properties of a cdf stated in Theorem 1.5.3. First, we have $\lim_{x\to-\infty} F_X(x) = 0$, because $F_X(x) = 0 \; \forall x \leq 0$, and

$$\lim_{x\to\infty} F_X(x) = \lim_{x\to\infty} (1 - e^{-x/\beta}) = 1 - \lim_{x\to\infty} e^{-x/\beta} = 1,$$

because $e^{-x/\beta} \to 0$, as $x \to \infty$. Second, $F_X(x)$ is clearly non-decreasing when $x \leq 0$ (it is constant). When $x > 0$,

$$\frac{d}{dx} F_X(x) = \frac{d}{dx}(1 - e^{-x/\beta}) = \frac{1}{\beta} e^{-x/\beta} > 0 \;\; \forall x > 0.$$

Therefore, $F_X(x)$ is non-decreasing. Finally, $F_X(x)$ is a continuous function; therefore, it is clearly right-continuous.

Definition: A random variable is **discrete** if $F_X(x)$ is a step function of x (see Example 1.19). A random variable X is **continuous** if $F_X(x)$ is a continuous function of x (see Example 1.20). A random variable X whose cdf $F_X(x)$ contains both continuous and step function pieces can be categorized as a **mixture** random variable.

Definition: Suppose X and Y are random variables defined on the same probability space $(S, \mathcal{B}, P)$. We say that X and Y are identically distributed if

$$P_X(X \in B) = P_Y(Y \in B)$$

for all $B \in \mathcal{B}(\mathbb{R})$. We write $X \stackrel{d}{=} Y$.

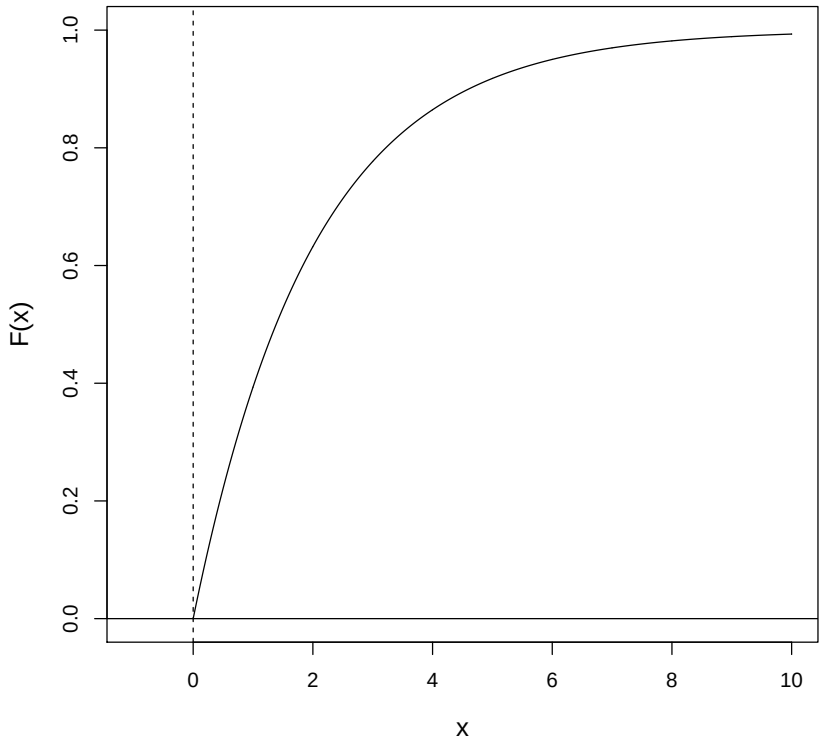


Figure 1.2: Cumulative distribution function $F_X(x)$ in Example 1.20 when $\beta = 2$.

Note: Because $(-\infty, x]$ is a Borel set, we see that

$$X \stackrel{d}{=} Y \implies F_X(x) = F_Y(x) \; \forall x \in \mathbb{R}.$$

Does this relationship go the other way? The answer is "yes," but it is much harder to prove; see Theorem 1.5.10 (pp 34 CB). Because of this equivalence, a random variable's cdf $F_X(x)$ completely determines its distribution.

Remark: When two random variables have the same (identical) distribution, this does not mean that they are the same random variable! That is,

$$X \stackrel{d}{=} Y \not\Longrightarrow X = Y.$$

For example, suppose that

$$\begin{aligned} S &= \{(\text{HH}), (\text{HT}), (\text{TH}), (\text{TT})\} \\ \mathcal{B} &= 2^S \\ P &= \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/4, \text{ for all } \omega \in S. \end{aligned}$$

If X denotes the number of heads and Y denotes the number of tails, it is easy to see that X and Y have the same distribution, that is, $X \stackrel{d}{=} Y$. However, $2 = X((\text{HH})) \neq Y((\text{HH})) = 0$, for example, showing that X and Y are not everywhere equal.

1.6 Density and Mass Functions

Review: Suppose that X is a random variable with cdf $F_X(x)$. Recall that

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x \in \mathbb{R}$$

and that $F_X(x)$ completely determines the distribution of X. Recall that

$$\begin{aligned} X \text{ discrete} &\iff F_X(x) \text{ is a step function} \\ X \text{ continuous} &\iff F_X(x) \text{ is continuous.} \end{aligned}$$

Remark: Suppose X is a random variable with support $\mathcal{X}$. If $\mathcal{X}$ is a countable set, then X is discrete. This is an equivalent characterization to that given above. This implies that a cdf $F_X(x)$ can have at most a countable number of discontinuities.

Definition: The **probability mass function** (pmf) of a discrete random variable X is given by

$$f_X(x) = P_X(X = x), \quad \text{for all } x.$$

Example 1.21. Suppose that X is a random variable with pmf

$$f_X(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$. This is called a **Poisson distribution**. Note that $\mathcal{X} = \{x : x = 0, 1, 2, ..., \}$ is countable. The pmf and cdf of X is shown in Figure 1.3 when $\lambda = 5$ (next page). An expression for the cdf of X is

$$F_X(x) = \sum_{u:u \leq x} f_X(u) = \sum_{u:u \leq x} \frac{\lambda^u e^{-\lambda}}{u!}.$$

In other words, the cdf $F_X(x)$ "adds up" all probabilities less than or equal to x. Here are some calculations:

$$\begin{aligned} P_X(X = 1) &= f_X(1) = \frac{\lambda^1 e^{-\lambda}}{1!} = \lambda e^{-\lambda} \\ P_X(X \leq 1) &= F_X(1) = f_X(0) + f_X(1) = \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} = e^{-\lambda}(1 + \lambda). \end{aligned}$$

Note: In general, if X is a discrete random variable with pmf $f_X(x)$, then

$$P_X(X \in B) = \sum_{x \in B} f_X(x) = \sum_{x \in B} P_X(X = x).$$

That is, we "add up" all probabilities corresponding to the support points $x \in B$. Of course, if $x \notin \mathcal{X}$, then $f_X(x) = P_X(X = x) = 0$.

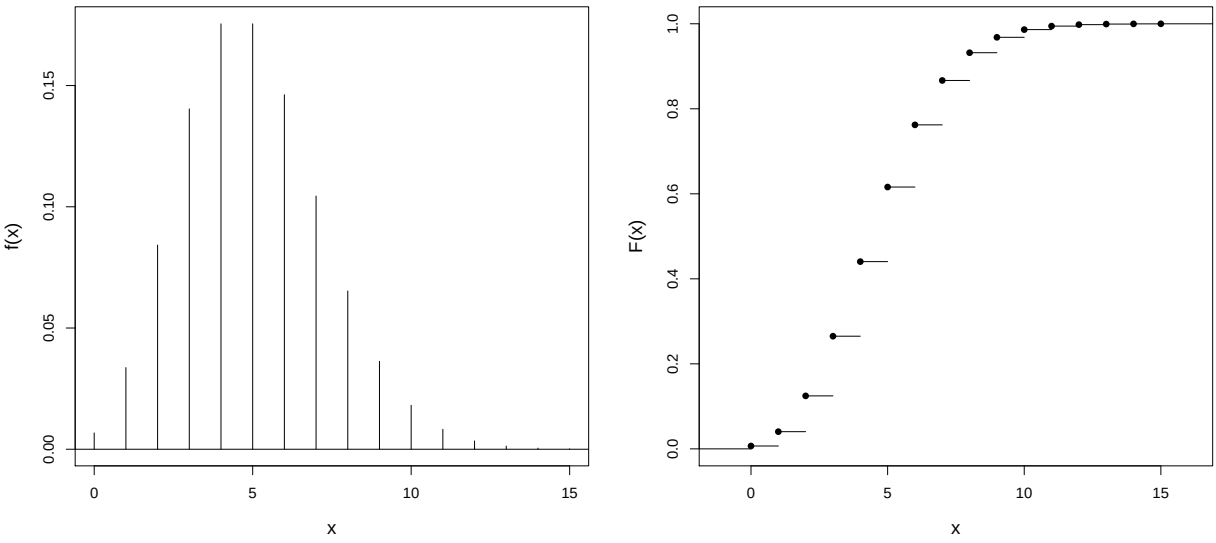


Figure 1.3: Pmf (left) and cdf (right) of $X \sim \text{Poisson}(\lambda = 5)$ in Example 1.21.

Remark: We now transition to continuous random variables and prove an interesting fact regarding them. Recall that the cdf of a continuous random variable is a continuous function.

Result: If X is a continuous random variable with cdf $F_X(x)$, then $P_X(X = x) = 0 \; \forall x \in \mathbb{R}$.

Proof. Suppose $\epsilon > 0$ and note that $\{X = x\} \subseteq \{x - \epsilon < X \leq x\}$. Therefore, by monotonicity,

$$P_X(X = x) \leq P_X(x - \epsilon < X \leq x) = P_X(X \leq x) - P_X(X \leq x - \epsilon) = F_X(x) - F_X(x - \epsilon).$$

Because P_X is a probability measure, we have

$$\begin{aligned} 0 \leq P_X(X = x) &\leq \lim_{\epsilon \to 0}[F_X(x) - F_X(x - \epsilon)] \\ &= F_X(x) - \lim_{\epsilon \to 0} F_X(x - \epsilon) \\ &= F_X(x) - F_X(x) = 0, \end{aligned}$$

because $F_X(x)$ is continuous. Therefore, we have shown that $0 \leq P_X(X = x) \leq 0$. Because ϵ was arbitrary, we are done. □

Remark: This result highlights the salient difference between discrete and continuous random variables. Discrete random variables have positive probability assigned to support points $x \in \mathcal{X}$. Continuous random variables do not.

Definition: The **probability density function** (pdf) of a continuous random variable X is function $f_X(x)$ that satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du,$$

for all $x \in \mathbb{R}$. If $f_X(x)$ is a continuous function, then

$$\frac{d}{dx}F_X(x) = \frac{d}{dx}\int_{-\infty}^{x} f_X(u)du = f_X(x).$$

This is a consequence of the Fundamental Theorem of Calculus. The **support** $\mathcal{X}$ of a continuous random variable X is the set of all $x \in \mathbb{R}$ such that $f_X(x) > 0$.

Example 1.22. The random variable X has probability density function (pdf)

$$f_X(x) = \frac{1}{2}e^{-|x|}, \text{ for } x \in \mathbb{R}.$$

(a) Find the cdf $F_X(x)$.
(b) Find $P_X(X > 5)$ and $P_X(-2 < X < 2)$.

Solution. (a) Recall that the cdf $F_X(x)$ is defined for all $x \in \mathbb{R}$. Also, recall the absolute value function

$$|x| = \begin{cases} -x, & x < 0 \\ x, & x \geq 0. \end{cases}$$

Case 1: For $x < 0$,

$$\begin{aligned} F_X(x) = \int_{-\infty}^{x} f_X(u)du &= \int_{-\infty}^{x} \frac{1}{2}e^u du \\ &= \frac{1}{2}e^u\Big|_{-\infty}^{x} = \frac{1}{2}(e^x - 0) = \frac{1}{2}e^x. \end{aligned}$$

Case 2: For $x \geq 0$,

$$\begin{aligned} F_X(x) = \int_{-\infty}^{x} f_X(u)du &= \int_{-\infty}^{0} f_X(u)du + \int_{0}^{x} f_X(u)du \\ &= \int_{-\infty}^{0} \frac{1}{2}e^u du + \int_{0}^{x} \frac{1}{2}e^{-u} du = \frac{1}{2} - \left(\frac{1}{2}e^{-u}\Big|_0^x\right) = 1 - \frac{1}{2}e^{-x}. \end{aligned}$$

Summarizing, the cdf of X is

$$F_X(x) = \begin{cases} \frac{1}{2}e^x, & x < 0 \\ 1 - \frac{1}{2}e^{-x}, & x \geq 0. \end{cases}$$

The pdf and cdf of X are shown in Figure 1.4 (next page). This is an example of a **LaPlace distribution**.
(b) The desired probabilities are

$$P_X(X > 5) = 1 - P_X(X \leq 5) = 1 - F_X(5) = 1 - \left(1 - \frac{1}{2}e^{-5}\right) \approx 0.0034$$

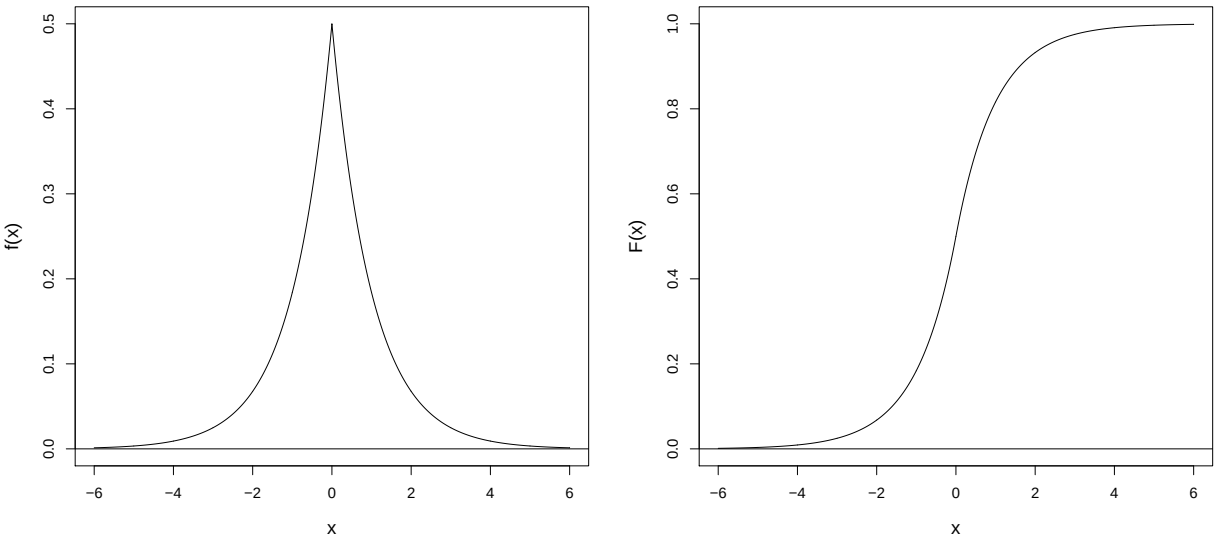


Figure 1.4: Pdf (left) and cdf (right) of X in Example 1.22.

and

$$\begin{aligned} P_X(-2 < X < 2) = P_X(-2 < X \leq 2) &= P_X(X \leq 2) - P_X(X \leq -2) \\ &= F_X(2) - F_X(-2) \\ &= \left(1 - \frac{1}{2}e^{-2}\right) - \frac{1}{2}e^{-2} = 1 - e^{-2} \approx 0.8647. \end{aligned}$$

Remark: In the last calculation, note that we wrote

$$\{X \leq 2\} = \{X \leq -2\} \cup \{-2 < X \leq 2\}$$

Therefore,

$$P_X(X \leq 2) = P_X(X \leq -2) + P_X(-2 < X \leq 2)$$

and, after rearranging,

$$\begin{aligned} P_X(-2 < X \leq 2) &= P_X(X \leq 2) - P_X(X \leq -2) \\ &= F_X(2) - F_X(-2). \end{aligned}$$

Result: If X is a **continuous** random variable with cdf $F_X(x)$ and pdf $f_X(x)$, then for any $a < b$,

$$P_X(a < X < b) = P_X(a \leq X < b) = P_X(a < X \leq b) = P_X(a \leq X \leq b)$$

and each one equals

$$F_X(b) - F_X(a) = \int_a^b f_X(x)dx.$$

Theorem 1.6.5. A function $f_X(x)$ is a pdf (pmf) of a random variable X if and only if

(a) $f_X(x) \geq 0$, for all $x \in \mathbb{R}$

(b) For a pmf or pdf, respectively,

$$\sum_{x \in \mathbb{R}} f_X(x) = 1 \quad \text{or} \quad \int_{\mathbb{R}} f_X(x)dx = 1.$$

Proof. We first prove the necessity ($\Longrightarrow$). Suppose $f_X(x)$ is a pdf (pmf). For the discrete case, $f_X(x) = P_X(X = x) \geq 0$ and

$$\sum_{x \in \mathbb{R}} f_X(x) = \sum_{x \in \mathcal{X}} P_X(X = x) = P(S) = 1.$$

For the continuous case, $f_X(x) = (d/dx)F_X(x) \geq 0$, because $F_X(x)$ is non-decreasing and

$$\int_{\mathbb{R}} f_X(x)dx = \lim_{x \to \infty} \int_{-\infty}^{x} f_X(u)du = \lim_{x \to \infty} F_X(x) = 1.$$

We have proven the necessity.

Remark: Proving the sufficiency ($\Longleftarrow$) is more difficult. For a function $f_X(x)$ satisfying (a) and (b), we recall that

$$F_X(x) = \sum_{u:u \leq x} f_X(u) \qquad \text{(discrete case)}$$

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \qquad \text{(continuous case)}.$$

In essence, we can write both of these expressions generally as the same expression

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

If X is discrete, then $F_X(x)$ is an integral with respect to a **counting measure**; that is, $F_X(x)$ is the sum over all u satisfying $u \leq x$. Thus, to establish the sufficiency part, it suffices to show that $F_X(x)$ defined above satisfies the three cdf properties in Theorem 1.5.3 (i.e., "end behavior" limits, non-decreasing, right-continuity). We do this now. First, note that

$$\begin{aligned} \lim_{x \to -\infty} F_X(x) &= \lim_{x \to -\infty} \int_{-\infty}^{x} f_X(u)du \\ &= \lim_{x \to -\infty} \int_{\mathbb{R}} f_X(u)I(u \leq x)du, \end{aligned}$$

where the **indicator function**

$$I(u \leq x) = \begin{cases} 1, & u \leq x \\ 0, & u > x \end{cases}$$

is regarded as a function of u. In the last integral, note that we can take the integrand and write

$$f_X(u)I(u \leq x) \leq f_X(u)$$

for all $u \in \mathbb{R}$ because $f_X(x) \geq 0$, by assumption, and also that $\int_{\mathbb{R}} f_X(x)dx = 1 < \infty$. Therefore, we have "dominated" the integrand in

$$\lim_{x \to -\infty} \int_{\mathbb{R}} f_X(u)I(u \leq x)du$$

above by a function that is integrable over $\mathbb{R}$. This, by means of the **Dominated Convergence Theorem**, allows us to interchange the limit and integral as follows:

$$\lim_{x \to -\infty} \int_{\mathbb{R}} f_X(u)I(u \leq x)du = \int_{\mathbb{R}} f_X(u) \underbrace{\lim_{x \to -\infty} I(u \leq x)}_{=\,0} du = 0.$$

We have shown that $\lim_{x \to -\infty} F_X(x) = 0$. Showing $\lim_{x \to \infty} F_X(x) = 1$ is done analogously and is therefore left as an exercise. To show that $F_X(x)$ is non-decreasing, suppose that $x_1 \leq x_2$. It suffices to show that $F_X(x_1) \leq F_X(x_2)$. Note that

$$F_X(x_1) = \int_{-\infty}^{x_1} f_X(u)du \leq \int_{-\infty}^{x_2} f_X(u)du = F_X(x_2),$$

because $f_X(u) \geq 0$, by assumption (i.e., if you integrate a non-negative function over a "larger" set, the integral cannot decrease). Finally, to prove that $F_X(x)$ is right-continuous, it suffices to show that

$$\lim_{x \to x_0^+} F_X(x) = F_X(x_0).$$

Note that

$$\begin{aligned} \lim_{x \to x_0^+} F_X(x) = \lim_{x \to x_0^+} \int_{-\infty}^{x} f_X(u)du &= \lim_{x \to x_0^+} \int_{\mathbb{R}} f_X(u)I(u \leq x)du \\ &\overset{\text{DCT}}{=} \int_{\mathbb{R}} f_X(u) \lim_{x \to x_0^+} I(u \leq x)du. \end{aligned}$$

It is easy to see that $I(u \leq x)$, now viewed as a function of x, is right-continuous. Therefore,

$$\begin{aligned} \int_{\mathbb{R}} f_X(u) \lim_{x \to x_0^+} I(u \leq x)du &= \int_{\mathbb{R}} f_X(u)I(u \leq x_0)du \\ &= \int_{-\infty}^{x_0} f_X(u)du = F_X(x_0). \end{aligned}$$

We have shown that $F_X(x)$ satisfies the three cdf properties in Theorem 1.5.3. Thus, we are done. $\square$

Remark: As noted on pp 37 (CB), there do exist random variables for which the relationship

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du$$

does not hold for any function $f_X(x)$. In a more advanced course, it is common to use the phrase "absolutely continuous" to refer to a random variable where this relationship holds; i.e., the random variable does, in fact, have a pdf $f_X(x)$.

2 Transformations and Expectations

Complementary reading: Chapter 2 (CB). Sections 2.1-2.3.

2.1 Distributions of Functions of a Random Variable

Remark: Suppose that X is a random variable defined over $(S, \mathcal{B}, P)$, that is, $X : S \to \mathbb{R}$ with the property that

$$X^{-1}(B) \equiv \{\omega \in S : X(\omega) \in B\} \in \mathcal{B}$$

for all $B \in \mathcal{B}(\mathbb{R})$. Going forward, we will rarely acknowledge explicitly the underlying probability space $(S, \mathcal{B}, P)$. In essence, we will consider the probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ to be the "starting point."

Question: If X is a random variable, then so is

$$Y = g(X),$$

where $g : \mathbb{R} \to \mathbb{R}$. A central question becomes this: "If I know the distribution of X, can I find the distribution of $Y = g(X)$?"

Note: For any $A \subseteq \mathbb{R}$ in the range space of g, note that

$$P_Y(Y \in A) = P_X(g(X) \in A) = P_X(X \in g^{-1}(A)),$$

where $g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}$, the inverse image of A under g. This shows that, in general, the distribution of Y depends on F_X (the distribution of X) and g.

Remark: In this course, we will consider g to be a real-valued function and will write $g : \mathbb{R} \to \mathbb{R}$ to emphasize this. However, the function g for our purposes is really a mapping from $\mathcal{X}$ (the support of X) to $\mathcal{Y}$, that is,

$$g : \mathcal{X} \to \mathcal{Y},$$

where $\mathcal{Y} = \{y : y = g(x),\ x \in \mathcal{X}\}$ is the support of Y.

Discrete case: Suppose that X is a discrete random variable (so that $\mathcal{X}$ is at most countable). Then Y is also a discrete random variable and the probability mass function (pmf) of Y is

$$f_Y(y) = P_Y(Y = y) = P_X(g(X) = y) = P_X(X = g^{-1}(y)) = \sum_{x \in \mathcal{X}: g(x) = y} f_X(x).$$

Above, the symbol $g^{-1}(y)$ is understood to mean

$$g^{-1}(y) = \{x \in \mathcal{X} : g(x) = y\},$$

the inverse image of the singleton $\{y\}$. In other words, $g^{-1}(y)$ is the set of all $x \in \mathcal{X}$ that get "mapped" into y under g. If there is always only one $x \in \mathcal{X}$ that satisfies $g(x) = y$, then $g^{-1}(y) = \{x\}$, also a singleton. This occurs when g is a one-to-one function on $\mathcal{X}$.

Example 2.1. Suppose that X is a discrete random variable with pmf

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, ..., n \\ 0, & \text{otherwise,} \end{cases}$$

where $0 < p < 1$. We say that X follows a **binomial distribution**.

Note: As a frame of reference (for where this distribution arises), envision a sequence of independent 0-1 "trials" (0 = failure; 1 = success), and let X denote the number of "successes" out of these n trials. We write $X \sim b(n, p)$. Note that the support of X is $\mathcal{X} = \{x : x = 0, 1, 2, ..., n\}$.

Question: What is the distribution of

$$Y = g(X) = n - X?$$

Note that the support of Y is given by

$$\mathcal{Y} = \{y : y = g(x),\ x \in \mathcal{X}\} = \{y : y = 0, 1, 2, ..., n\}.$$

Also, $g(x) = n - x$ is a one-to-one function over $\mathcal{X}$ (it is a linear function of x). Therefore,

$$y = g(x) = n - x \quad \Longleftrightarrow \quad x = g^{-1}(y) = n - y.$$

Therefore, the pmf of Y, for $y = 0, 1, 2, ..., n$, is given by

$$\begin{aligned} f_Y(y) = P_Y(Y = y) = P_X(n - X = y) &= P_X(X = n - y) \\ &= f_X(n - y) \\ &= \binom{n}{n-y} p^{n-y} (1-p)^{n-(n-y)} \\ &= \binom{n}{y} (1-p)^y p^{n-y}. \end{aligned}$$

That is,

$$f_Y(y) = \begin{cases} \binom{n}{y} (1-p)^y p^{n-y}, & y = 0, 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as a binomial pmf with "success probability" $1 - p$. We have therefore shown that

$$X \sim b(n, p) \quad \Longrightarrow \quad Y = g(X) = n - X \sim b(n, 1-p).$$

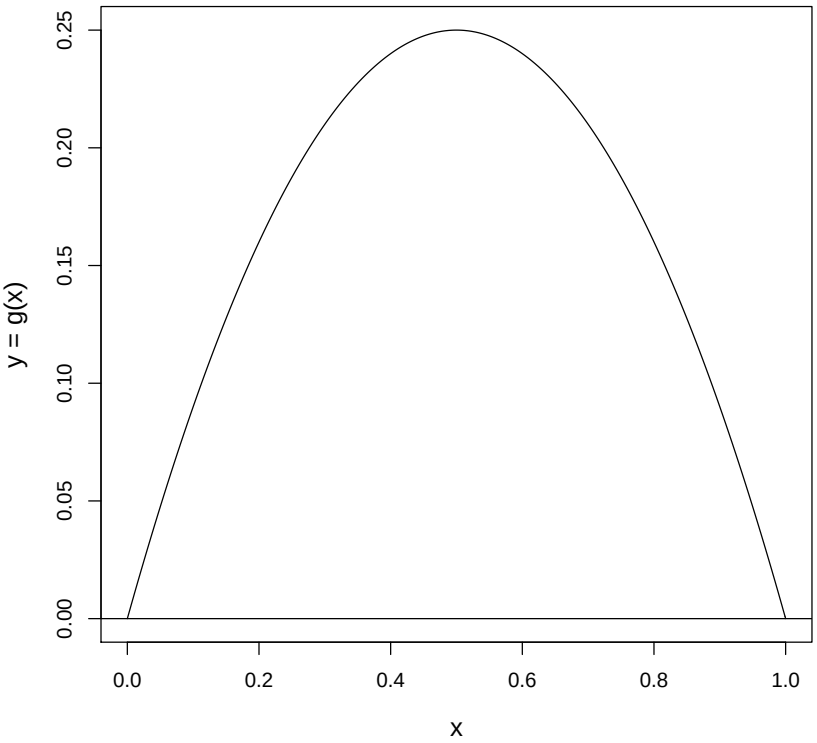


Figure 2.1: A graph of $g(x) = x(1-x)$ over $\mathcal{X} = \{x : 0 < x < 1\}$ in Example 2.2.

Continuous case: Suppose X and Y are continuous random variables, where $Y = g(X)$. The cumulative distribution function (cdf) of Y can be written as

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) &= P_X(g(X) \leq y) \\ &= \int_B f_X(x)dx, \end{aligned}$$

where the set $B = \{x \in \mathcal{X} : g(x) \leq y\}$. Therefore, finding the cdf of Y is straightforward conceptually. However, care must be taken in identifying the set B above.

Example 2.2. Suppose that X has pdf

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

This is a **uniform distribution** with support $\mathcal{X} = \{x : 0 < x < 1\}$. We now derive the distribution of

$$Y = g(X) = X(1-X).$$

Remark: Whenever you derive the distribution of a function of a random variable, it is helpful to first construct the graph of $g(x)$ over its domain; that is, over $\mathcal{X}$, the support of

X. See Figure 2.1. Doing so allows you to also determine the support of $Y = g(X)$. Note that $0 < x < 1 \Longrightarrow 0 < y < \frac{1}{4}$. Therefore, the support of Y is $\mathcal{Y} = \{y : 0 < y < \frac{1}{4}\}$.

Important: Note that, for $0 < y < \frac{1}{4}$,

$$\{Y \leq y\} = \{X \leq x_1\} \cup \{X \geq x_2\},$$

where $g^{-1}(\{y\}) = \{x_1, x_2\}$. Therefore, for $0 < y < \frac{1}{4}$, the cdf of Y is

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) &= P_X(X \leq x_1) + P_X(X \geq x_2) \\ &= \int_0^{x_1} 1dx + \int_{x_2}^1 1dx, \end{aligned}$$

where x_1 and x_2 both satisfy $y = g(x) = x(1-x)$. We can find x_1 and x_2 using the quadratic formula:

$$y = g(x) = x(1-x) \quad \Longrightarrow \quad -x^2 + x - y = 0.$$

The roots of this equation are

$$\begin{aligned} x &= \frac{-1 \pm \sqrt{(1)^2 - 4(-1)(-y)}}{2(-1)} \\ &= \frac{1}{2} \pm \frac{\sqrt{1-4y}}{2} \end{aligned}$$

(x_1 is the root with the negative sign; x_2 is the root with the positive sign). Therefore, for $0 < y < \frac{1}{4}$,

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) &= \int_0^{\frac{1}{2} - \frac{\sqrt{1-4y}}{2}} 1dx + \int_{\frac{1}{2} + \frac{\sqrt{1-4y}}{2}}^1 1dx \\ &= \frac{1}{2} - \frac{\sqrt{1-4y}}{2} + 1 - \frac{1}{2} - \frac{\sqrt{1-4y}}{2} \\ &= 1 - \sqrt{1-4y}. \end{aligned}$$

Summarizing,

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - \sqrt{1-4y}, & 0 < y < \frac{1}{4} \\ 1, & y \geq \frac{1}{4}. \end{cases}$$

It is easy to show that this cdf satisfies the three cdf properties in Theorem 1.5.3 (i.e., "end behavior" limits, non-decreasing, right-continuity). The probability density function (pdf) of Y is therefore

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \begin{cases} 2(1-4y)^{-1/2}, & 0 < y < \frac{1}{4} \\ 0, & \text{otherwise.} \end{cases} \end{aligned}$$

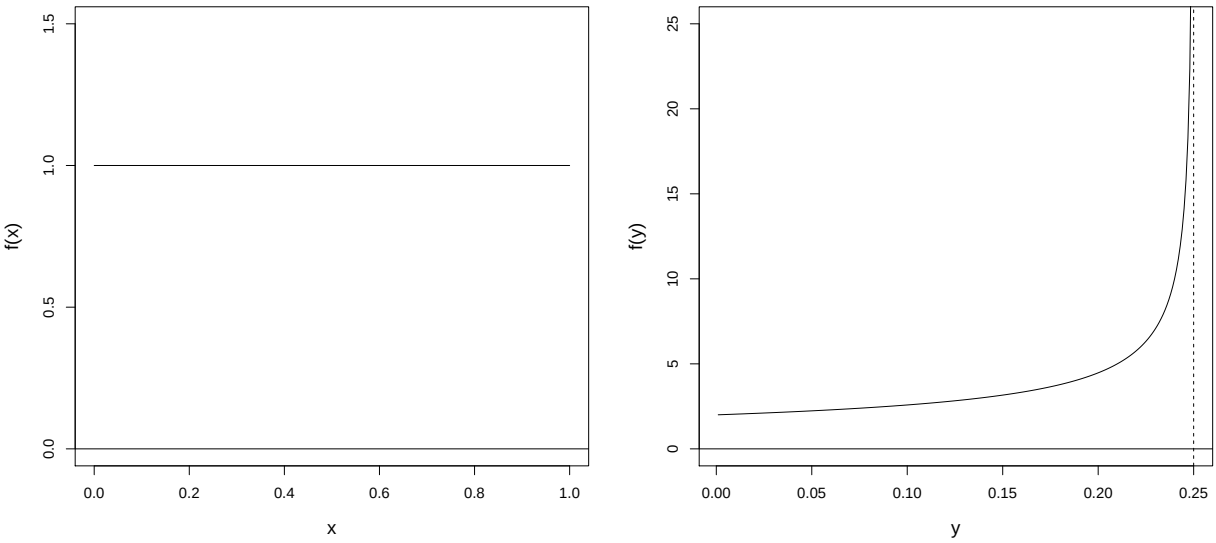


Figure 2.2: Pdf of X (left) and pdf of Y (right) in Example 2.2.

We can write this succinctly as

$$f_Y(y) = 2(1-4y)^{-1/2} I(0 < y < 1/4),$$

where $I(\cdot)$ is the indicator function. In Figure 2.2, we plot the pdf of X and the pdf of Y side by side. Doing so is instructive because it allows us to see what effect the transformation g has on the original distribution $f_X(x)$.

Monotone transformations: In Example 2.2, we see that $y = g(x) = x(1-x)$ is not a one-to-one function over $\mathcal{X} = \{x : 0 < x < 1\}$. In general, when $Y = g(X)$ and g is one-to-one over $\mathcal{X}$, we can get the pdf of Y easily (in terms of the pdf of X).

Recall: By "one-to-one," we mean that either (a) g is strictly increasing over $\mathcal{X}$ or (b) g is strictly decreasing over $\mathcal{X}$. Summarizing,

- Strictly increasing: $x_1 < x_2 \Rightarrow g(x_1) < g(x_2)$; if g is differentiable, $g'(x) > 0 \; \forall x \in \mathcal{X}$.
- Strictly decreasing: $x_1 < x_2 \Rightarrow g(x_1) > g(x_2)$; if g is differentiable, $g'(x) < 0 \; \forall x \in \mathcal{X}$.

Case 1: If g is strictly increasing, then

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) &= P_X(g(X) \leq y) \\ &= P_X(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)). \end{aligned}$$

The penultimate equality results from noting that $\{x : g(x) \leq y\} = \{x : x \leq g^{-1}(y)\}$. We have shown that

$$F_Y(y) = F_X(g^{-1}(y)), \text{ when } g \text{ is strictly increasing.}$$

Taking derivatives, the pdf of Y (where nonzero) is

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y(y) &= \frac{d}{dy}F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y)) \underbrace{\frac{d}{dy}g^{-1}(y)}_{>\,0}. \end{aligned}$$

Recall: From calculus, recall that if g is strictly increasing (decreasing), then g^{-1} is strictly increasing (decreasing).

Case 2: If g is strictly decreasing, then

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) &= P_X(g(X) \leq y) \\ &= P_X(X \geq g^{-1}(y)) \\ &= 1 - F_X(g^{-1}(y)). \end{aligned}$$

Again, the penultimate equality results from noting that $\{x : g(x) \leq y\} = \{x : x \geq g^{-1}(y)\}$. We have shown that

$$F_Y(y) = 1 - F_X(g^{-1}(y)), \text{ when } g \text{ is strictly decreasing.}$$

Taking derivatives, the pdf of Y (where nonzero) is

$$\begin{aligned} f_Y(y) = \frac{d}{dy}F_Y(y) &= \frac{d}{dy}\left[1 - F_X(g^{-1}(y))\right] \\ &= -f_X(g^{-1}(y)) \underbrace{\frac{d}{dy}g^{-1}(y)}_{<\,0}. \end{aligned}$$

Combining both cases, we arrive at the following result.

Theorem 2.1.5. Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is one-to-one over $\mathcal{X}$ (the support of X). If $f_X(x)$ is continuous on $\mathcal{X}$ and $g^{-1}(y)$ has a continuous derivative, then the pdf of Y is

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

for values of $y \in \mathcal{Y}$, the support of Y, $f_Y(y) = 0$, otherwise.

Remark: The quantity $(d/dy)g^{-1}(y)$ is sometimes called the **Jacobian** of the inverse transformation $x = g^{-1}(y)$.

Example 2.3. Suppose that $X \sim \mathcal{U}(0,1)$; i.e., X has pdf

$$f_X(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

For $\beta > 0$, find the pdf of

$$Y = g(X) = -\beta \ln X.$$

Solution. First, note that $g(x) = -\beta \ln x$ is strictly decreasing over $\mathcal{X} = \{x : 0 < x < 1\}$ because $g'(x) = -\beta/x < 0 \ \forall x \in \mathcal{X}$. The support of Y is $\mathcal{Y} = \{y : y > 0\}$. The inverse transformation $x = g^{-1}(y)$ is found as follows:

$$y = g(x) = -\beta \ln x \implies -\frac{y}{\beta} = \ln x \implies x = g^{-1}(y) = e^{-y/\beta}.$$

The Jacobian is

$$\frac{d}{dy} g^{-1}(y) = \frac{d}{dy}(e^{-y/\beta}) = -\frac{1}{\beta} e^{-y/\beta}.$$

Applying Theorem 2.1.5 directly, we have, for $y > 0$,

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= 1 \times \frac{1}{\beta} e^{-y/\beta} = \frac{1}{\beta} e^{-y/\beta}. \end{aligned}$$

Summarizing, the pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{\beta} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This is the pdf of an **exponential** random variable with parameter $\beta > 0$. We have therefore shown that

$$X \sim \mathcal{U}(0,1) \implies Y = g(X) = -\beta \ln X \sim \text{exponential}(\beta).$$

Example 2.4. Suppose that X is a continuous random variable with pdf

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, & x > 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ and $\beta > 0$. The function

$$\Gamma(\alpha) \stackrel{\alpha > 0}{=} \int_0^\infty u^{\alpha-1} e^{-u} du$$

is called the **gamma function** and will be discussed later. A random variable X with pdf $f_X(x)$ is said to follow a **gamma distribution** with

$$\begin{aligned} \alpha &\longrightarrow \text{“shape parameter”} \\ \beta &\longrightarrow \text{“scale parameter.”} \end{aligned}$$

We write $X \sim \text{gamma}(\alpha, \beta)$. We now find the distribution

$$Y = g(X) = \frac{1}{X}.$$

Solution. First, note that $g(x) = 1/x$ is strictly decreasing over $\mathcal{X} = \{x : x > 0\}$ because $g'(x) = -1/x^2 < 0 \, \forall x \in \mathcal{X}$. The support of Y is $\mathcal{Y} = \{y : y > 0\}$. The inverse transformation $x = g^{-1}(y)$ is found as follows:

$$y = g(x) = \frac{1}{x} \quad \Longrightarrow \quad x = g^{-1}(y) = \frac{1}{y}.$$

The Jacobian is

$$\frac{d}{dy} g^{-1}(y) = \frac{d}{dy}\left(\frac{1}{y}\right) = -\frac{1}{y^2}.$$

Applying Theorem 2.1.5 directly, we have, for $y > 0$,

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \left(\frac{1}{y}\right)^{\alpha-1} e^{-1/\beta y} \times \frac{1}{y^2} \\ &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \frac{1}{y^{\alpha+1}} e^{-1/\beta y}. \end{aligned}$$

Summarizing, the pdf of Y is

$$f_Y(y) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\beta^\alpha} \dfrac{1}{y^{\alpha+1}} e^{-1/\beta y}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This is called the **inverted gamma distribution** (not surprisingly). We have shown that

$$X \sim \text{gamma}(\alpha, \beta) \quad \Longrightarrow \quad Y = g(X) = \frac{1}{X} \sim \text{IG}(\alpha, \beta).$$

Q: What if g is not one-to-one over $\mathcal{X}$?
A: We can always use the general result that

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) &= P_X(g(X) \leq y) \\ &= \int_B f_X(x) dx, \end{aligned}$$

where $B = \{x \in \mathcal{X} : g(x) \leq y\}$. With an expression for the cdf $F_Y(y)$, we can then just differentiate it to find the pdf $f_Y(y)$. We already illustrated this "first principles" approach in Example 2.2.

Special case: Suppose that X is a continuous random variable with cdf $F_X(x)$ and pdf $f_X(x)$. Consider the transformation

$$Y = g(X) = X^2.$$

Note that $g(x) = x^2$ is not a one-to-one function over $\mathbb{R}$. However, it is one-to-one over $\mathcal{X} = \{x : 0 < x < 1\}$, for example. In general, the cdf of $Y = g(X) = X^2$ is, for $y > 0$,

$$\begin{aligned} F_Y(y) = P_Y(Y \leq y) &= P_X(X^2 \leq y) \\ &= P_X(-\sqrt{y} \leq X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}). \end{aligned}$$

Therefore, the pdf of $Y = X^2$ is, for $y > 0$,

$$\begin{aligned} f_Y(y) = \frac{d}{dy} F_Y(y) &= \frac{d}{dy}\left[F_X(\sqrt{y}) - F_X(-\sqrt{y})\right] \\ &= f_X(\sqrt{y})\frac{1}{2\sqrt{y}} - f_X(-\sqrt{y})\left(-\frac{1}{2\sqrt{y}}\right) \\ &= \frac{1}{2\sqrt{y}}\left[f_X(\sqrt{y}) + f_X(-\sqrt{y})\right]. \end{aligned}$$

Remark: This is a general formula for the pdf of $Y = g(X) = X^2$. Theorem 2.1.8 (pp 53 CB) generalizes this result.

Example 2.5. *Standard normal-χ^2 relationship.* Suppose the random variable X has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} I(x \in \mathbb{R}).$$

This is the **standard normal distribution**; we write $X \sim \mathcal{N}(0, 1)$. The support of X is $\mathcal{X} = \{x : -\infty < x < \infty\}$. Consider the transformation

$$Y = g(X) = X^2.$$

The support of Y is $\mathcal{Y} = \{y : y \geq 0\}$. However, because Y is continuous, $P_Y(Y = 0) = 0$. We can therefore proceed assuming that $y > 0$. By the last result, we have, for $y > 0$,

$$\begin{aligned} f_Y(y) = \frac{1}{2\sqrt{y}}\left[\frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} + \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2}\right] &= \frac{1}{2\sqrt{y}} \frac{2e^{-y/2}}{\sqrt{2\pi}} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}. \end{aligned}$$

Summarizing, the pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This is the pdf of a χ^2 **distribution** with $\nu = 1$ degree of freedom. We have shown that

$$X \sim \mathcal{N}(0, 1) \quad \Longrightarrow \quad Y = g(X) = X^2 \sim \chi^2_1.$$

We will use this fact repeatedly in this course.

Interesting: Recall that we defined the gamma function

$$\Gamma(\alpha) \stackrel{\alpha>0}{=} \int_0^\infty u^{\alpha-1}e^{-u}du.$$

The gamma function satisfies certain properties (see Chapter 3). We will later show that $\Gamma(1/2) = \sqrt{\pi}$. Rewriting the χ^2_1 pdf, we see that, for $y > 0$,

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}}e^{-y/2} \\ &= \frac{1}{\Gamma(\frac{1}{2})2^{1/2}}y^{\frac{1}{2}-1}e^{-y/2}, \end{aligned}$$

which we recognize as a gamma(α, β) pdf with parameters $\alpha = 1/2$ and $\beta = 2$. Therefore, the χ^2_1 distribution is a special member of the gamma(α, β) family; it is the gamma distribution arising when $\alpha = 1/2$ and $\beta = 2$.

Probability Integral Transformation: Suppose that X is a continuous random variable with cdf $F_X(x)$. Define the random variable

$$Y = F_X(X).$$

The random variable $Y \sim \mathcal{U}(0, 1)$, that is, the pdf and cdf of Y, respectively, are given by

$$f_Y(y) = I(0 < y < 1) \qquad \text{and} \qquad F_Y(y) = \begin{cases} 0, & y < 0 \\ y, & 0 \le y \le 1 \\ 1, & y > 1. \end{cases}$$

Proof. Suppose that $F_X(x)$ is strictly increasing. Regardless of the support of X, the random variable $Y = F_X(X)$ has support $\mathcal{Y} = \{y : 0 \le y \le 1\}$. The cdf of Y, for $0 \le y \le 1$, is given by

$$F_Y(y) = P_Y(Y \le y) = P_X(F_X(X) \le y) = P_X(X \le F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

In the third equality above, we used the fact that $\{x : F_X(x) \le y\} = \{x : x \le F_X^{-1}(y)\}$. This is true because $F_X(x)$ is strictly increasing (i.e., a unique inverse exists). Therefore,

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ y, & 0 \le y \le 1 \\ 1, & y > 1, \end{cases}$$

proving the result. □

Remark: The Probability Integral Transformation remains true when X is continuous but has a cdf $F_X(x)$ that is not strictly increasing (i.e., it could have flat regions over $\mathcal{X}$). In this situation, we just have to redefine what we mean by "inverse" over these flat regions; see pp 54-55 (CB).

Remark: The novelty of this result is that it holds for any continuous distribution, that is, a continuous random variable's cdf, when viewed as random itself, follows a $\mathcal{U}(0, 1)$ distribution, regardless of what the random variable's distribution actually is. This result is useful in numerous instances, for example, in the theoretical development of probability values used in hypothesis testing.

2.2 Expected Values

Definition: Suppose that X is a random variable. The **expected value** (or **mean**) of X is defined as

$$\begin{aligned} E(X) &= \sum_{x \in \mathcal{X}} x f_X(x) \qquad \text{(discrete case)} \\ E(X) &= \int_{\mathbb{R}} x f_X(x) dx \qquad \text{(continuous case)} \end{aligned}$$

Note: If $E(|X|) = +\infty$, then we say that "$E(X)$ does not exist." This occurs when the sum (integral) above does not converge absolutely. In other words, for $E(X)$ to exist in the discrete case, we need $\sum_{x \in \mathcal{X}} |x| f_X(x)$ to converge. In the continuous case, we need $\int_{\mathbb{R}} |x| f_X(x) dx < \infty$.

Example 2.6. A discrete random variable is said to have a **Poisson distribution** if its probability mass function (pmf) is given by

$$f_X(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$. The expected value of X is

$$E(X) = \sum_{x=0}^{\infty} x \, \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=1}^{\infty} x \, \frac{\lambda^x e^{-\lambda}}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} \underbrace{\sum_{y=0}^{\infty} \frac{\lambda^y}{y!}}_{= e^{\lambda}} = \lambda,$$

because $\sum_{y=0}^{\infty} \lambda^y / y!$ is the McLaurin series expansion of e^{λ}. Therefore, if $X \sim \text{Poisson}(\lambda)$, then $E(X) = \lambda$.

Example 2.7. A continuous random variable is said to have a **Pareto distribution** if its probability density function (pdf) is given by

$$f_X(x) = \frac{\beta \alpha^{\beta}}{x^{\beta+1}} I(x > \alpha),$$

where $\alpha > 0$ and $\beta > 0$. The expected value of X is

$$\begin{aligned} E(X) = \int_{\mathbb{R}} x f_X(x) dx &= \int_{\alpha}^{\infty} x \, \frac{\beta \alpha^{\beta}}{x^{\beta+1}} dx \\ &= \beta \alpha^{\beta} \int_{\alpha}^{\infty} \frac{1}{x^{\beta}} dx \\ &= \beta \alpha^{\beta} \left(-\frac{1}{\beta - 1} \, \frac{1}{x^{\beta-1}} \bigg|_{x=\alpha}^{\infty} \right) \\ &= \frac{\beta \alpha^{\beta}}{\beta - 1} \left(\frac{1}{\alpha^{\beta-1}} - \lim_{x \to \infty} \frac{1}{x^{\beta-1}} \right) = \frac{\beta \alpha}{\beta - 1}, \end{aligned}$$

provided that $\beta > 1$. Note that if $\beta = 1$, then $E(X) = \alpha \int_\alpha^\infty (1/x)dx$, which is not finite. Also, if $0 < \beta < 1$, then the limit above becomes

$$\lim_{x\to\infty} \frac{1}{x^{\beta-1}} = \lim_{x\to\infty} x^{1-\beta} = +\infty$$

showing that $E(X)$ does not exist either. Therefore, if $X \sim \text{Pareto}(\alpha, \beta)$, then

$$E(X) = \frac{\beta\alpha}{\beta - 1}, \quad \text{provided that } \beta > 1.$$

If $0 < \beta \leq 1$, then $E(X)$ does not exist.

Functions of Random Variables: Suppose X is a random variable (discrete or continuous). The expected value of $g(X)$ is

$$\begin{aligned} E[g(X)] &= \sum_{x\in\mathcal{X}} g(x) f_X(x) \qquad \text{(discrete case)} \\ E[g(X)] &= \int_{\mathbb{R}} g(x) f_X(x) dx \qquad \text{(continuous case)} \end{aligned}$$

Note: If $E[|g(X)|] = +\infty$, then we say that "$E[g(X)]$ does not exist." This occurs when the sum (integral) above does not converge absolutely. In other words, for $E[g(X)]$ to exist in the discrete case, we need $\sum_{x\in\mathcal{X}} |g(x)| f_X(x)$ to converge. In the continuous case, we need $\int_{\mathbb{R}} |g(x)| f_X(x) dx < \infty$.

Law of the Unconscious Statistician: Suppose X is a random variable and let $Y = g(X)$, $g : \mathbb{R} \to \mathbb{R}$. In the continuous case, we can calculate $E(Y) = E[g(X)]$ in two ways:

$$\begin{aligned} E[g(X)] &= \int_{\mathbb{R}} g(x) f_X(x) dx \\ E(Y) &= \int_{\mathbb{R}} y f_Y(y) dy, \end{aligned}$$

where $f_Y(y)$ is the pdf (pmf) of Y. If X and Y are discrete random variables, the integrals above are simply sums. The Law of the Unconscious Statistician says that $E(Y) = E[g(X)]$ in the sense that if one expectation exists, so does the other and they are equal.

Example 2.8. Suppose that $X \sim \mathcal{U}(0, 1)$, and let $Y = g(X) = -\ln X$. We will show that $E(Y) = E[g(X)]$. With respect to the distribution of X,

$$E[g(X)] = E(-\ln X) = \int_0^1 -\ln x \; dx.$$

Let

$$\begin{aligned} u &= -\ln x \qquad & du &= -\tfrac{1}{x} dx \\ dv &= dx \qquad & v &= x. \end{aligned}$$

Integration by parts shows that the last integral

$$\begin{aligned}\int_0^1 -\ln x \; dx &= -x\ln x\Big|_0^1 - \int_0^1 (-1)dx \\ &= (0-0)+1 = 1.\end{aligned}$$

To calculate $E(Y)$, recall from Example 2.3 that $Y \sim$ exponential(1). Therefore, $f_Y(y) = e^{-y}I(y>0)$ and

$$E(Y) = \int_0^\infty ye^{-y}dy.$$

Let

$$\begin{array}{ll} u = y & du = dy \\ dv = e^{-y} & v = -e^{-y} \end{array}$$

Integration by parts shows that the last integral

$$\begin{aligned}\int_0^\infty ye^{-y}dy &= -ye^{-y}\Big|_0^\infty - \int_0^\infty -e^{-y}dy \\ &= (0-0)+1 = 1.\end{aligned}$$

Therefore, $E(Y) = E[g(X)]$, as claimed.

Note: The process of taking expectations is a **linear operation**. For constants a and b,

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

For example, in Example 2.8,

$$E(2Y-3) = 2E(Y)-3 = 2(1)-3 = -1.$$

Theorem 2.2.5. Let X be a random variable and let a, b, and c be constants. For any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

(a) $E[ag_1(X)+bg_2(X)+c] = aE[g_1(X)]+bE[g_2(X)]+c$.

(b) if $g_1(x) \geq 0$ for all x, then $E[g_1(X)] \geq 0$.

(c) if $g_1(x) \geq g_2(x)$ for all x, then $E[g_1(X)] \geq E[g_2(X)]$.

(d) if $a \leq g_1(x) \leq b$ for all x, then $a \leq E[g_1(X)] \leq b$.

Proof. Assume X is continuous with pdf $f_X(x)$ and support $\mathcal{X}$. To prove (a), note that

$$\begin{aligned}E[ag_1(X)+bg_2(X)+c] &= \int_{\mathbb{R}} [ag_1(x)+bg_2(x)+c]f_X(x)dx \\ &= a\int_{\mathbb{R}} g_1(x)f_X(x)dx + b\int_{\mathbb{R}} g_2(x)f_X(x)dx + c\int_{\mathbb{R}} f_X(x)dx \\ &= aE[g_1(X)]+bE[g_2(X)]+c.\end{aligned}$$

To prove (b), note that $g_1(x)f_X(x) \geq 0$ for all $x \in \mathcal{X}$. Therefore,

$$E[g_1(X)] = \int_{\mathbb{R}} g_1(x)f_X(x)dx \geq 0.$$

To prove (c), note that $g_1(x) \geq g_2(x) \implies g_1(x) - g_2(x) \geq 0$. From part (b), we know $E[g_1(X) - g_2(X)] = E[g_1(X)] - E[g_2(X)] \geq 0$. To prove part (d), note that

$$E[g_1(X)] = \int_{\mathbb{R}} g_1(x)f_X(x)dx \geq \int_{\mathbb{R}} af_X(x)dx = a\int_{\mathbb{R}} f_X(x)dx = a.$$

An analogous argument shows that $E[g_1(X)] \leq b$. □

Interesting characterization: Suppose that X is a random variable and suppose $E(X)$ exists. Then

$$E(X) = \arg\min_{b\in\mathbb{R}} E[(X-b)^2].$$

Proof. Let

$$h(b) = E[(X-b)^2] = E(X^2 - 2bX + b^2) = E(X^2) - 2bE(X) + b^2.$$

Note that

$$\frac{d}{db}\, h(b) = -2E(X) + 2b \stackrel{\text{set}}{=} 0 \implies b = E(X).$$

Because $(d^2/db^2)h(b) = 2 > 0$, the solution $b = E(X)$ is a minimizer. □

Interpretation: Suppose that you would like to predict the value of X and will use the value b as this prediction. Therefore, the quantity $X - b$ can be thought of as the "error" in your prediction. Prediction errors can be positive or negative, so we could consider the quantity $(X-b)^2$ instead because it is always non-negative. Choosing $b = E(X)$ minimizes the expected squared error of prediction.

Special Expectations: We list below special expectations of the form $E[g(X)]$.

1. $g(X) = X^k$. The expectation

   $$E[g(X)] = E(X^k) \equiv \mu_k'$$

   is called the k**th moment** of X.

2. $g(X) = (X-\mu)^k$, where $\mu = E(X)$. The expectation

   $$E[g(X)] = E[(X-\mu)^k] \equiv \mu_k$$

   is called the k**th central moment** of X.

3. $g(X) = e^{tX}$, where t is a constant. The expectation

   $$E[g(X)] = E(e^{tX}) \equiv M_X(t)$$

   is called the **moment generating function** of X. Note: The function $\kappa_X(t) = \ln M_X(t)$ is called the **cumulant generating function** of X.

4. $g(X) = t^X$, where t is a constant. The expectation

$$E[g(X)] = E(t^X)$$

is called the **factorial moment generating function** of X; see pp 83 (CB).

5. $g(X) = e^{itX}$, where t is a constant and $i = \sqrt{-1}$. The expectation

$$E[g(X)] = E(e^{itX}) \equiv \psi_X(t)$$

is called the **characteristic function** of X. In this case, the function $g : \mathbb{R} \to \mathbb{C}$.

2.3 Moments and Moment Generating Functions

Definition: Suppose that X is a random variable. The k**th** (uncentered) **moment** of X is

$$\mu_k' = E(X^k).$$

The k**th central moment** of X is

$$\mu_k = E[(X - \mu)^k],$$

where $\mu = E(X)$. Usually when talking about moments, k is a positive integer.

- The 1st moment of X is $\mu_1' = E(X)$, which is the **mean** of X.
- The 2nd central moment of X is $\mu_2 = E[(X - \mu)^2]$. We call this the **variance** of X and usually denote this by σ^2 or $\text{var}(X)$. That is,

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2].$$

Remark: Note that the variance of X can be computed as

$$\begin{aligned} \text{var}(X) = E[(X - \mu)^2] &= E(X^2 - 2\mu X + \mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 \\ &= E(X^2) - \mu^2 \\ &= E(X^2) - [E(X)]^2. \end{aligned}$$

This is called the **variance computing formula.**

Remark: Because $g(x) = (x - \mu)^2 \geq 0$ for all $x \in \mathbb{R}$, we know that

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] \geq 0$$

by Theorem 2.2.5(b). The only time $\text{var}(X) = 0$ is when $X = \mu$ with probability one; i.e., $P_X(X = \mu) = 1$. In this case, the distribution of X is **degenerate** at μ; in other words, all of the probability associated with X is located at the single value $x = \mu$.

Definition: The positive square root of the variance of X is the **standard deviation** of X, that is,

$$\sigma = \sqrt{\sigma^2} = \sqrt{\text{var}(X)}.$$

In practice, the standard deviation σ is easier to interpret because its units are the same as those for X. The variance of X is measured in (units)^2.

Example 2.9. Suppose that $X \sim \text{Poisson}(\lambda)$; i.e., the pmf of X is

$$f_X(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$. In Example 2.6, we showed $E(X) = \lambda$. We now calculate $\text{var}(X)$. The 2nd (uncentered) moment of X is

$$\begin{aligned} E(X^2) = \sum_{x=0}^{\infty} x^2 \, \frac{\lambda^x e^{-\lambda}}{x!} &= \lambda \sum_{x=1}^{\infty} x \, \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} \\ &= \lambda \sum_{y=0}^{\infty} (y+1) \, \frac{\lambda^y e^{-\lambda}}{y!} \\ &= \lambda E(Y+1), \end{aligned}$$

where the random variable $Y \sim \text{Poisson}(\lambda)$. Therefore,

$$E(X^2) = \lambda E(Y+1) = \lambda[E(Y)+1] = \lambda(\lambda+1)$$

and the variance of X is

$$\text{var}(X) = E(X^2) - [E(X)]^2 = \lambda(\lambda+1) - \lambda^2 = \lambda.$$

Summary: If $X \sim \text{Poisson}(\lambda)$, then $E(X) = \text{var}(X) = \lambda$.

Example 2.10. Suppose that $X \sim \text{Pareto}(\alpha, \beta)$; i.e., the pdf of X is

$$f_X(x) = \frac{\beta \alpha^\beta}{x^{\beta+1}} I(x > \alpha),$$

where $\alpha > 0$ and $\beta > 0$. In Example 2.7, we showed $E(X) = \beta\alpha/(\beta-1)$, provided that $\beta > 1$. We now calculate $\text{var}(X)$. The 2nd (uncentered) moment of X is

$$\begin{aligned} E(X^2) = \int_{\mathbb{R}} x^2 f_X(x) dx &= \int_{\alpha}^{\infty} x^2 \, \frac{\beta \alpha^\beta}{x^{\beta+1}} dx \\ &= \beta \alpha^\beta \int_{\alpha}^{\infty} \frac{1}{x^{\beta-1}} dx \\ &= \beta \alpha^\beta \left(-\frac{1}{\beta-2} \, \frac{1}{x^{\beta-2}} \Big|_{x=\alpha}^{\infty} \right) \\ &= \frac{\beta \alpha^\beta}{\beta-2} \left(\frac{1}{\alpha^{\beta-2}} - \lim_{x \to \infty} \frac{1}{x^{\beta-2}} \right) = \frac{\beta \alpha^2}{\beta-2}, \end{aligned}$$

provided that $\beta > 2$. If $0 < \beta \leq 2$, then $E(X^2)$ does not exist. Therefore, the variance of X is

$$\begin{aligned} \text{var}(X) = E(X^2) - [E(X)]^2 &= \frac{\beta\alpha^2}{\beta - 2} - \left(\frac{\beta\alpha}{\beta - 1}\right)^2 \\ &= \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}. \end{aligned}$$

Note that this formula only applies if $\beta > 2$. If $0 < \beta \leq 2$, then $\text{var}(X)$ does not exist.

Theorem 2.3.4. If X is a random variable with finite variance, i.e., $\text{var}(X) < \infty$, then for constants a and b,

$$\text{var}(aX + b) = a^2\text{var}(X).$$

Proving this is easy; apply the variance computing formula to $\text{var}(Y)$, where $Y = aX + b$.

Remarks: Note that this formula is different than the analogous result for expected values; i.e.,

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

The result for variances says that additive (location) shifts through b do not affect the variance. Also, if $a = 0$, then $\text{var}(b) = 0$. In other words, the variance of a constant is zero.

Definition: Suppose that X is a random variable with cdf $F_X(x)$. The **moment generating function** (mgf) of X is

$$M_X(t) = E(e^{tX}),$$

provided this expectation is finite for all t in an open neighborhood about $t = 0$; i.e., $\exists h > 0 \ni E(e^{tX}) < \infty \ \forall t \in (-h, h)$. If no such $h > 0$ exists, then the moment generating function of X does not exist. A general expression for $M_X(t)$ is

$$M_X(t) = E(e^{tX}) = \int_{\mathbb{R}} e^{tx} dF_X(x),$$

written as a Riemann-Stiljes integral, which is understood to mean

$$\begin{aligned} M_X(t) &= \sum_{x \in \mathcal{X}} e^{tx} f_X(x) \qquad \text{(discrete case)} \\ M_X(t) &= \int_{\mathbb{R}} e^{tx} f_X(x) dx \qquad \text{(continuous case)} \end{aligned}$$

Example 2.11. Suppose that $X \sim \text{Poisson}(\lambda)$; i.e., the pmf of X is

$$f_X(x) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$.

The mgf of X is

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \sum_{x=0}^{\infty} e^{tx} \, \frac{\lambda^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}. \end{aligned}$$

Note we have used the fact that

$$\sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{\lambda e^t};$$

i.e., the LHS is the McLaurin series expansion of $h(t) = e^{\lambda e^t}$. This expansion is valid for all $t \in \mathbb{R}$. Hence, the mgf of X exists.

Example 2.12. Suppose that $X \sim b(n, p)$; i.e., the pmf of X is

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, ..., n \\ 0, & \text{otherwise,} \end{cases}$$

where $0 < p < 1$. The mgf of X is

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

where $q = 1 - p$. Note that we have used the binomial expansion formula above; i.e.,

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

with $a = pe^t$ and $b = q = 1 - p$. This expansion holds for any a and b. Therefore, the expansion is valid for all $t \in \mathbb{R}$. Hence, the mgf of X exists.

Example 2.13. Suppose that $X \sim$ exponential(β); i.e., the pdf of X is

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta} I(x > 0),$$

where $\beta > 0$. The mgf of X is

$$\begin{aligned} M_X(t) = E(e^{tX}) = \int_0^{\infty} e^{tx} \, \frac{1}{\beta} e^{-x/\beta} dx &= \frac{1}{\beta} \int_0^{\infty} e^{-x\left(\frac{1}{\beta} - t\right)} dx \\ &= \frac{1}{\beta} \left[-\frac{1}{\frac{1}{\beta} - t} \, e^{-x\left(\frac{1}{\beta} - t\right)} \Big|_{x=0}^{\infty} \right] \\ &= \frac{1}{1 - \beta t} \left[e^{-x\left(\frac{1}{\beta} - t\right)} \Big|_{\infty}^{0} \right] \\ &= \frac{1}{1 - \beta t} \left[1 - \lim_{x \to \infty} e^{-x\left(\frac{1}{\beta} - t\right)} \right]. \end{aligned}$$

Note that

$$\begin{aligned}\lim_{x\to\infty} e^{-x\left(\frac{1}{\beta}-t\right)} &= 0 \qquad \text{if } \frac{1}{\beta}-t>0\\ \lim_{x\to\infty} e^{-x\left(\frac{1}{\beta}-t\right)} &= +\infty \qquad \text{if } \frac{1}{\beta}-t<0.\end{aligned}$$

Therefore, provided that

$$\frac{1}{\beta}-t>0 \iff t<\frac{1}{\beta},$$

the mgf of X exists and is given by

$$M_X(t)=\frac{1}{1-\beta t}.$$

Note that $\exists h>0$ (e.g., $h=1/\beta$) such that $M_X(t)=E(e^{tX})<\infty\ \forall t\in(-h,h)$.

Generalization: If $X\sim$ gamma(α,β), then

$$M_X(t)=\left(\frac{1}{1-\beta t}\right)^{\alpha}, \qquad t<\frac{1}{\beta}.$$

When $\alpha=1$, the gamma(α,β) distribution reduces to the exponential(β) distribution. The gamma mgf is derived on pp 63-64 (CB).

Why are mgfs useful?

Reason 1: Moment generating functions are functions that generate moments.

Theorem 2.3.7. If X is a random variable with mgf $M_X(t)$, then

$$E(X^k)=M_X^{(k)}(0),$$

where

$$M_X^{(k)}(0)=\frac{d^k}{dt^k}M_X(t)\bigg|_{t=0}.$$

This result shows that moments of X can be found by differentiation.

Proof. Set $k=1$. The mgf of X can be written generally as

$$M_X(t)=E(e^{tX})=\int_{\mathbb{R}} e^{tx}dF_X(x).$$

Taking the first derivative, we have

$$\begin{aligned}\frac{d}{dt}M_X(t) &= \frac{d}{dt}\int_{\mathbb{R}} e^{tx}dF_X(x)\\ &\stackrel{?}{=} \int_{\mathbb{R}}\frac{d}{dt}e^{tx}dF_X(x)\\ &= \int_{\mathbb{R}} xe^{tx}dF_X(x)=E(Xe^{tX}).\end{aligned}$$

Therefore,

$$\frac{d}{dt}M_X(t)\bigg|_{t=0} = E(Xe^{0X}) = E(X).$$

Showing this for $k = 2, 3, ...$, is done similarly. □

Remark: In the argument above, we needed to assume that the interchange of the derivative and integral (sum) is justified. When the mgf exists, this interchange is justified. See also §2.4 (CB) for a more general discussion on this topic.

Interesting: Writing $M_X(t)$ in its McLaurin series expansion (i.e., a Taylor series expansion about $t = 0$), we see that

$$\begin{aligned} M_X(t) &= M_X(0) + \frac{M_X^{(1)}(0)}{1!}(t-0) + \frac{M_X^{(2)}(0)}{2!}(t-0)^2 + \frac{M_X^{(3)}(0)}{3!}(t-0)^3 + \cdots \\ &= 1 + E(X)t + \frac{E(X^2)}{2}t^2 + \frac{E(X^3)}{6}t^3 + \frac{E(X^4)}{24}t^4 + \cdots \\ &= \sum_{k=0}^{\infty} \frac{E(X^k)}{k!}t^k. \end{aligned}$$

You can also convince yourself that Theorem 2.3.7 is true, that is,

$$E(X^k) = \frac{d^k}{dt^k}M_X(t)\bigg|_{t=0},$$

by differentiating the RHS of $M_X(t)$ written in its expansion (and evaluating derivatives at $t = 0$). This argument would not relieve you from having to justify an interchange of the derivative; the interchange now would involve an infinite sum. As in our proof of Theorem 2.3.7, this interchange is justified provided that the mgf exists.

Example 2.14. Suppose that $X \sim b(n, p)$; i.e., the pmf of X is

$$f_X(x) = \begin{cases} \binom{n}{x}p^x(1-p)^{n-x}, & x = 0, 1, 2, ..., n \\ 0, & \text{otherwise,} \end{cases}$$

where $0 < p < 1$. In Example 2.12, we derived the mgf of X to be

$$M_X(t) = (q + pe^t)^n,$$

where $q = 1 - p$. Differentiating $M_X(t)$, we have

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}(q + pe^t)^n = n(q + pe^t)^{n-1}pe^t.$$

Therefore,

$$E(X) = \frac{d}{dt}M_X(t)\bigg|_{t=0} = n(q + pe^0)^{n-1}pe^0 = np.$$

Exercise: Show $\text{var}(X) = np(1 - p)$.

Discussion: In general, a random variable's first four moments describe important physical characteristics of its distribution.

1. $E(X) = \mu$ describes the "center" of the distribution of X.
2. $\sigma^2 = \text{var}(X) = E(X^2) - [E(X)]^2$ describes the "spread" or "variability" in the distribution of X.
3. The **skewness** of X is defined as
$$\xi = \frac{E[(X-\mu)^3]}{(\sigma^2)^{3/2}}$$
and describes the "skewness" in the distribution of X (i.e., the departure from symmetry).
   - $\xi = 0 \Longrightarrow f_X(x)$ is symmetric about μ
   - $\xi > 0 \Longrightarrow f_X(x)$ is skewed right
   - $\xi < 0 \Longrightarrow f_X(x)$ is skewed left.
4. The **kurtosis** of X is defined as
$$\kappa = \frac{E[(X-\mu)^4]}{(\sigma^2)^2}$$
and describes the "peakedness" of a distribution relative to the amount of variability in the tails of the distribution of X.
   - $\kappa = 3 \Longrightarrow$ mesokurtic; normal distribution (as a reference)
   - $\kappa > 3 \Longrightarrow$ leptokurtic; $f_X(x)$ has a more acute peak around μ and fatter tails
   - $\kappa < 3 \Longrightarrow$ platykurtic; $f_X(x)$ has a broader peak around μ and thinner tails.

Remarks:

- Obviously, we need the appropriate moments to exist for these quantities to be relevant; for example, we need $E(X^3)$ to exist to talk about a random variable's skewness.
- If a random variable's mgf exists, then it characterizes an infinite set of moments. However, not all random variables have mgfs.
- Higher order moments existing implies the existence of lower order moments, as the follow result shows.

Result: Suppose X is a random variable. If $E(X^m)$ exists, so does $E(X^k)$ for all $k \leq m$.

Proof. The kth moment of X is
$$E(X^k) = \int_{\mathbb{R}} x^k dF_X(x).$$

To prove that $E(X^k)$ exists, it suffices to show that

$$E(|X|^k) = \int_{\mathbb{R}} |x|^k dF_X(x) < \infty.$$

Toward this end, note that we can write

$$\begin{aligned}
\int_{\mathbb{R}} |x|^k dF_X(x) &= \int_{|x|\leq 1} |x|^k dF_X(x) + \int_{|x|>1} |x|^k dF_X(x) \\
&\leq \int_{|x|\leq 1} dF_X(x) + \int_{|x|>1} |x|^m dF_X(x) \\
&\leq \int_{\mathbb{R}} dF_X(x) + \int_{\mathbb{R}} |x|^m dF_X(x) \\
&= 1 + E(|X|^m).
\end{aligned}$$

The first inequality results because $|x|^k \leq 1$ whenever $|x| \leq 1$ and $|x|^k \leq |x|^m$ whenever $|x| > 1$. The second inequality results because, in both integrals, we are integrating positive functions over a larger region. We have shown that $E(|X|^k) \leq 1 + E(|X|^m)$. However, $E(X^m)$ exists by assumption so $E(|X|^m) < \infty$. Thus, we are done. □

Why are mgfs useful?

Reason 2: Moment generating functions uniquely determine a random variable's distribution.

Theorem 2.3.11. Suppose X and Y are random variables, defined on the same probability space $(S, \mathcal{B}, P)$, with moment generating functions $M_X(t)$ and $M_Y(t)$, respectively, which exist. Then

$$F_X(x) = F_Y(x)\ \forall x \in \mathbb{R} \iff M_X(t) = M_Y(t)\ \ \forall t \in (-h, h),\ \exists h > 0.$$

Remarks: The practical implication of Theorem 2.3.11 is that the mgf of a random variable completely determines its distribution. Proving the necessity ($\Longrightarrow$) of Theorem 2.3.11 is easy. Proving the sufficiency ($\Longleftarrow$) is not. Note that if X is continuous,

$$\begin{aligned}
M_X(t) &= \int_{\mathbb{R}} e^{tx} dF_X(x) \\
&= \int_{\mathbb{R}} e^{tx} f_X(x) dx,
\end{aligned}$$

is a LaPlace transform of $f_X(x)$. The sufficiency part stems from the uniqueness of LaPlace transforms.

Recall: In Section 2.1, recall that we posed the general question:

> "If I know the distribution of X, can I find the distribution of $Y = g(X)$?"

In the light of Theorem 2.3.11, we now have another approach on how to answer this question. Specifically, we can derive the mgf of $Y = g(X)$. Because mgfs are unique, the distribution identified by this mgf must be the answer.

Example 2.15. Suppose that $X \sim \text{gamma}(\alpha, \beta)$. Find the distribution of

$$Y = g(X) = cX,$$

where $c > 0$. Recall that the mgf of X is

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}, \qquad t < \frac{1}{\beta}.$$

The mgf of Y is

$$\begin{aligned} M_Y(t) = E(e^{tY}) &= E(e^{tcX}) \\ &= M_X(ct) \\ &= \left(\frac{1}{1-\beta ct}\right)^{\alpha}, \end{aligned}$$

which exists for $ct < 1/\beta \Longleftrightarrow t < 1/\beta c$. We recognize $M_Y(t)$ as the mgf of a gamma distribution with shape parameter α and scale parameter βc. Because mgfs are unique (i.e., they uniquely identify a distribution), it must be true that $Y = cX \sim \text{gamma}(\alpha, \beta c)$.

Remark: When finding the distribution of a function of a random variable $Y = g(X)$, we have three ways to approach this problem:

1. CDF technique: derive $F_Y(y)$ directly; I call this the "first principles" approach
2. Transformation: requires g to be one-to-one (Theorem 2.1.5)
3. MGF technique: derive $M_Y(t)$ and identify the corresponding distribution.

Theorem 2.3.15. Suppose X is a random variable with mgf $M_X(t)$. For any constants a and b, the mgf of $Y = g(X) = aX + b$ is given by

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

Proof. The mgf of Y is

$$\begin{aligned} M_Y(t) = E(e^{tY}) &= E\left[e^{t(aX+b)}\right] \\ &= E(e^{bt}e^{atX}) \\ &= e^{bt}E(e^{atX}) = e^{bt}M_X(at). \ \square \end{aligned}$$

Example 2.16. Suppose that X has pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} I(x \in \mathbb{R}),$$

where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. A random variable with this pdf is said to have a **normal distribution** with mean μ and variance σ^2, written $X \sim \mathcal{N}(\mu, \sigma^2)$. In Chapter 3, we will show that the mgf of X is

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}.$$

Here, we derive the distribution of $Y = g(X) = aX + b$, where a and b are constants. To do this, simply note that

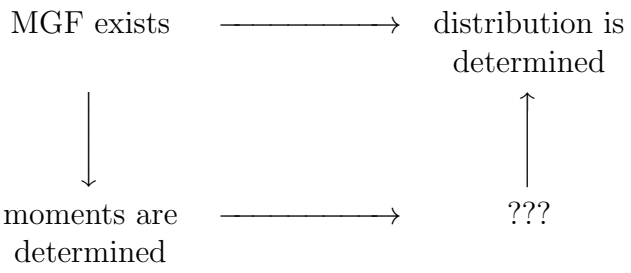
$$\begin{aligned} M_Y(t) = e^{bt} M_X(at) &= e^{bt} e^{\mu(at) + \sigma^2 (at)^2/2} \\ &= e^{(a\mu + b)t + a^2\sigma^2 t^2/2}, \end{aligned}$$

which we recognize as the mgf of a normal distribution with mean $a\mu + b$ and variance $a^2\sigma^2$. We have therefore shown that

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Longrightarrow \quad Y = g(X) = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2).$$

This result, which is important in its own right, is actually just a special case of a more general result stating that **linear combinations** of normal random variables are also normally distributed, a fact that we will prove more generally in Chapter 4.

Interesting relationships: The following diagram describes the relevant relationships between mgfs and their associated distributions and moments:



Q: Does an infinite set of moments uniquely determine a distribution?
A: Yes, if $\mathcal{X}$ is bounded. No, otherwise. That is, it is possible for two different distributions to have the same (infinite) set of moments, as the following example shows.

Example 2.17. Suppose that X and Y have pdfs

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}x} e^{-(\ln x)^2/2} I(x > 0) \\ f_Y(y) &= f_X(y)[1 + \sin(2\pi \ln y)]. \end{aligned}$$

A random variable $X \sim f_X(x)$ is said to have a **lognormal distribution** (actually this is just one member of the lognormal family of distributions). For these two distributions, it is possible to show that

$$E(X^r) = E(Y^r) = e^{r^2/2}, \quad \text{for } r = 1, 2, 3, ..., .$$

However, these two distributions are very different distributions; see Figure 2.3.2 (pp 65 CB). This example illustrates that even if two random variables have the same (infinite) set of moments, they do not necessarily have the same distribution.

Interesting: Another interesting fact about the lognormal distribution in Example 2.17 is that X has all of its moments, given by $E(X^r) = e^{r^2/2}$, for $r = 1, 2, 3, ...,$. However, the mgf of X does not exist, because

$$E(e^{tX}) = \int_0^\infty e^{tx} \frac{1}{\sqrt{2\pi}x} e^{-(\ln x)^2/2} dx$$

is not finite. See Exercise 2.36 (pp 81 CB).

Why are mgfs useful?

Reason 3: Moment generating functions can help to establish convergence results.

Theorem 2.3.12. Suppose $\{X_n\}$ is a sequence of random variables, where X_n has mgf $M_{X_n}(t)$. Suppose that

$$M_{X_n}(t) \to M_X(t),$$

as $n \to \infty$ for all $t \in (-h, h)$ $\exists h > 0$; i.e., the sequence of functions $M_{X_n}(t)$ converges pointwise for all t in an open neighborhood about $t = 0$. Then

1. There exists a unique cdf $F_X(x)$ whose moments are determined by $M_X(t)$.
2. The sequence of cdfs
$$F_{X_n}(x) \to F_X(x),$$
as $n \to \infty$, for all $x \in C_{F_X}$, the set of points $x \in \mathbb{R}$ where $F_X(\cdot)$ is continuous.

In other words, convergence of mgfs implies convergence of cdfs. We write $X_n \xrightarrow{d} X$, as $n \to \infty$, and say that "X_n **converges in distribution** to X."

Aside: When discussing convergence results in mathematical statistics, we will often be asked to evaluate a limit of the form

$$\lim_{n\to\infty} \left[1 + \frac{b}{n} + \frac{g(n)}{n}\right]^{cn},$$

where b and c are constants (free of n) and $\lim_{n\to\infty} g(n) = 0$. A L'Hôpital's rule argument shows that

$$\lim_{n\to\infty} \left[1 + \frac{b}{n} + \frac{g(n)}{n}\right]^{cn} = e^{bc}.$$

A special case of this result arises when $g(n) = 0$ and $c = 1$; i.e.,

$$\lim_{n\to\infty} \left(1 + \frac{b}{n}\right)^n = e^b.$$

Example 2.18. Suppose that $X_n \sim b(n, p_n)$, where $np_n = \lambda$ for all n. For this sequence of random variables, we have

$$\begin{aligned} M_{X_n}(t) = E(e^{tX_n}) &= (q_n + p_n e^t)^n \\ &= \left[1 + \frac{\lambda(e^t - 1)}{n}\right]^n, \end{aligned}$$

where $q_n = 1 - p_n$. Therefore, with $b = \lambda(e^t - 1)$, $c = 1$, and $g(n) = 0$, we have

$$\lim_{n\to\infty} M_{X_n}(t) = e^{\lambda(e^t-1)},$$

which we recognize as the mgf of a Poisson distribution with mean λ. Therefore, the **limiting distribution** of $X_n \sim b(n, p_n)$, where $np_n = \lambda$ for all $n \in \mathbb{N}$, is Poisson(λ). We write $X_n \xrightarrow{d} X$, as $n \to \infty$, where $X \sim$ Poisson(λ).

Example 2.19. Suppose that $Y_n \sim$ gamma(n, β), where β is free of n, so that

$$M_{Y_n}(t) = \left(\frac{1}{1-\beta t}\right)^n, \qquad t < \frac{1}{\beta}.$$

Find the limiting distribution of

$$X_n = \frac{Y_n}{n}.$$

Solution. The mgf of X_n is

$$\begin{aligned} M_{X_n}(t) = E(e^{tX_n}) &= E\left[e^{t\left(\frac{Y_n}{n}\right)}\right] \\ &= M_{Y_n}(t/n) \\ &= \left[\frac{1}{1-\beta(t/n)}\right]^n \\ &= \left(1 - \frac{\beta t}{n}\right)^{-n}, \end{aligned}$$

provided that $t/n < 1/\beta \Longleftrightarrow t < n/\beta$. Taking $b = -\beta t$, $c = -1$, and $g(n) = 0$ in the general limit result stated earlier, we see that

$$\lim_{n\to\infty} M_{X_n}(t) = e^{\beta t}.$$

The limiting mgf $M_X(t) = e^{\beta t}$ is the mgf of a **degenerate** random variable X with all of its probability mass at a single point, namely, $x = \beta$. That is, the cdf of X is

$$F_X(x) = \begin{cases} 0, & x < \beta \\ 1, & x \geq \beta. \end{cases}$$

We have therefore shown that $X_n \xrightarrow{d} X$, as $n \to \infty$, where X has a degenerate distribution at β. In Chapter 5, we will refer to this type of convergence as "convergence in probability" and will write $X_n \xrightarrow{p} \beta$, as $n \to \infty$.

3 Common Families of Distributions

Complementary reading: Chapter 3 (CB). Sections 3.1-3.6.

3.1 Introduction

Definition: A **parametric model** (or **parametric family**) is a set of distributions indexed by a finite-dimensional parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)'$, where $d \geq 1$. Unless otherwise stated, the parameter $\boldsymbol{\theta}$ is regarded as fixed (i.e., it is not random).

Example 3.1. Suppose $X \sim$ exponential(β). Because $\beta > 0$, we see that a collection of distributions emerges; i.e.,

$$\left\{ f_X(x|\beta) = \frac{1}{\beta} e^{-x/\beta} I(x > 0);\ \beta > 0 \right\}.$$

Here, the parameter $\boldsymbol{\theta} = \beta$, a scalar ($d = 1$). One member of this collection (i.e., family) arises when $\beta = 2$, for example,

$$f_X(x|2) = \frac{1}{2} e^{-x/2} I(x > 0).$$

The pdf

$$f_X(x|3) = \frac{1}{3} e^{-x/3} I(x > 0)$$

corresponds to another member of this family.

Example 3.2. Suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$; i.e., the pdf of X is

$$f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} I(x \in \mathbb{R}),$$

where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Here, the parameter $\boldsymbol{\theta} = (\mu, \sigma^2)'$ is two-dimensional ($d = 2$). One very important member of the $\mathcal{N}(\mu, \sigma^2)$ family arises when $\mu = 0$ and $\sigma^2 = 1$. This member is called the **standard normal distribution** and is denoted by $\mathcal{N}(0, 1)$.

Remark: A common format for a first-year sequence in probability and mathematical statistics (like STAT 712-713) is to accept a given parametric family of distributions as being appropriate and then proceed to develop what is exclusively **model-dependent, parametric statistical inference** (in contrast to nonparametric statistical inference). We therefore endeavor to investigate various "named" families of distributions that will be relevant for future use (we have seen many already). We will examine families of distributions that correspond to both discrete and continuous random variables.

3.2 Discrete Distributions

Recall: A random variable X is **discrete** if its cdf $F_X(x)$ is a step function. An equivalent characterization is that the support of X, denoted by $\mathcal{X}$, is at most countable.

1. Discrete Uniform. A random variable is said to have a discrete uniform distribution if its pmf is given by

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

where $N \in \mathbb{N}$. Note that this distribution puts the same weight $1/N$ on each outcome $x \in \mathcal{X} = \{x : x = 1, 2, ..., N\}$. **Notation:** $X \sim \text{DU}(1, N)$.

Mean/Variance: The relevant moments of $X \sim \text{DU}(1, N)$ are

$$\begin{aligned} E(X) &= \frac{N+1}{2} \\ \text{var}(X) &= \frac{(N+1)(N-1)}{12}. \end{aligned}$$

MGF: The mgf of $X \sim \text{DU}(1, N)$ is

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \sum_{x=1}^{N} e^{tx} \ \frac{1}{N} \\ &= \frac{1}{N}e^t + \frac{1}{N}e^{2t} + \cdots + \frac{1}{N}e^{Nt}. \end{aligned}$$

Generalization: The discrete uniform distribution can be generalized easily. The pmf of $X \sim \text{DU}(N_0, N_1)$ is

$$f_X(x|N_0, N_1) = \begin{cases} \frac{1}{N_1 - N_0 + 1}, & x = N_0, N_0 + 1, ..., N_1 \\ 0, & \text{otherwise,} \end{cases}$$

where N_0 and N_1 are integers satisfying $N_0 < N_1$.

2. Hypergeometric. A random variable X is said to have a hypergeometric distribution if its pmf is given by

$$f_X(x|N, M, K) = \begin{cases} \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, & x = 0, 1, 2, ..., K \\ 0, & \text{otherwise,} \end{cases}$$

where $N, M, K \in \mathbb{N}$, $M < N$, $K < N$. The support $\mathcal{X} = \{x : x = 0, 1, 2, ..., K\}$ is appropriate when K is "small" when compared to both N and M. **Notation:** $X \sim \text{hyper}(N, K, M)$.

Remark: To understand this distribution, it is easiest to conceptualize a finite population of N objects, where the objects are classified as either of "Type I" or "Type II."

$$\begin{aligned} N &= \text{population size} \\ K &= \text{sample size} \\ M &= \text{number of Type I objects in the population.} \end{aligned}$$

We sample K objects from the population at random and without replacement (SRSWOR). The random variable X records

$$X = \text{ number of Type I objects in the sample (i.e., out of } K).$$

Mean/Variance: The relevant moments of $X \sim \text{hyper}(N, K, M)$ are

$$\begin{aligned} E(X) &= \frac{KM}{N} \\ \text{var}(X) &= \frac{KM}{N}\left(1 - \frac{M}{N}\right)\left(\frac{N-K}{N-1}\right). \end{aligned}$$

The term $(\frac{N-K}{N-1})$ is called the "finite population correction factor" and arises in sampling contexts. The mgf of $X \sim \text{hyper}(N, K, M)$ exists but not in a convenient form.

Curiosity: If the population size $N \to \infty$ so that $\frac{M}{N} \to p \in (0, 1)$, note that for fixed K,

$$E(X) \to Kp \quad \text{and} \quad \text{var}(X) \to Kp(1-p),$$

which are the corresponding moments of the $b(n, p)$ distribution. Not only do the moments converge, but the hyper(N, K, M) pmf also converges to the $b(n, p)$ pmf under the same conditions; see Exercise 3.11 (pp 129 CB). When N is "large" (i.e., large relative to K), probability calculations in a finite population (or when sampling without replacement) should be "close" to those in a population viewed as infinite in size (or when sampling is done with replacement).

Terminology: A **Bernoulli trial** is an experiment with two possible outcomes, where

- the outcomes can be thought of as "success" or "failure"
- $p = \text{pr}(\text{"success"})$ is the same for each trial.

3. Binomial. A random variable X is said to have a binomial distribution if its pmf is given by

$$f_X(x|n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, 2, ..., n \\ 0, & \text{otherwise,} \end{cases}$$

where $n \in \mathbb{N}$ and $0 < p < 1$. The random variable X counts the number of "successes" in n independent Bernoulli trials. **Notation:** $X \sim b(n, p)$.

Mean/Variance: The relevant moments of $X \sim b(n,p)$ are

$$\begin{aligned} E(X) &= np \\ \text{var}(X) &= np(1-p). \end{aligned}$$

MGF: The mgf of $X \sim b(n,p)$ is

$$M_X(t) = (q + pe^t)^n, \quad \text{where } q = 1-p.$$

Remark: When $n = 1$, the $b(n,p)$ distribution reduces to the **Bernoulli distribution** with pmf

$$f_X(x|p) = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

We write $X \sim \text{Bernoulli}(p)$.

4. Geometric. A random variable X is said to have a geometric distribution if its pmf is given by

$$f_X(x|p) = \begin{cases} (1-p)^{x-1}p, & x = 1, 2, 3, ..., \\ 0, & \text{otherwise,} \end{cases}$$

where $0 < p < 1$. **Conceptualization:** Suppose independent Bernoulli trials are performed. The random variable X counts the number of trials needed to observe the 1st success. The support of X is $\mathcal{X} = \{x : x = 1, 2, 3, ..., \} = \mathbb{N}$. **Notation:** $X \sim \text{geom}(p)$.

Mean/Variance: The relevant moments of $X \sim \text{geom}(p)$ are

$$\begin{aligned} E(X) &= \frac{1}{p} \\ \text{var}(X) &= \frac{q}{p^2}, \end{aligned}$$

where $q = 1 - p$.

MGF: The mgf of $X \sim \text{geom}(p)$ is

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

for $qe^t < 1 \Longleftrightarrow t < -\ln q$.

Memoryless Property: For integers $s > t$,

$$P_X(X > s|X > t) = P_X(X > s - t).$$

The geometric distribution is the only discrete distribution that has this property.

5. Negative Binomial. A random variable X is said to have a negative binomial distribution if its pmf is given by

$$f_X(x|r,p) = \begin{cases} \binom{x-1}{r-1}p^r(1-p)^{x-r}, & x = r, r+1, r+2, ..., \\ 0, & \text{otherwise}, \end{cases}$$

where $0 < p < 1$. **Conceptualization:** Suppose independent Bernoulli trials are performed. The random variable X counts the number of trials needed to observe the rth success, where $r \geq 1$. The support of X is $\mathcal{X} = \{x : x = r, r+1, r+2, ..., \}$. **Notation:** $X \sim \text{nib}(r,p)$. When $r = 1$, the nib(r,p) distribution reduces to the geom(p) distribution. The value r is called the **waiting parameter**.

Mean/Variance: The relevant moments of $X \sim \text{nib}(r,p)$ are

$$\begin{aligned} E(X) &= \frac{r}{p} \\ \text{var}(X) &= \frac{rq}{p^2}, \quad \text{where } q = 1-p. \end{aligned}$$

MGF: The mgf of $X \sim \text{nib}(r,p)$ is

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \sum_{x=r}^{\infty} e^{tx}\binom{x-1}{r-1}p^r(1-p)^{x-r} \\ &= (pe^t)^r \underbrace{\sum_{x=r}^{\infty}\binom{x-1}{r-1}(qe^t)^{x-r}}_{=\,(1-qe^t)^{-r}} \\ &= \left(\frac{pe^t}{1-qe^t}\right)^r, \end{aligned}$$

for $qe^t < 1 \iff t < -\ln q$. That $\sum_{x=r}^{\infty}\binom{x-1}{r-1}(qe^t)^{x-r} = (1-qe^t)^{-r}$ follows from the lemma below.

LEMMA. Suppose that r is a nonnegative integer. Then

$$\sum_{x=r}^{\infty}\binom{x-1}{r-1}(qe^t)^{x-r} = (1-qe^t)^{-r}.$$

Proof. Consider the function $f(w) = (1-w)^{-r}$, where r is a nonnegative integer. It is easy to show that

$$\begin{aligned} f'(w) &= r(1-w)^{-(r+1)} \\ f''(w) &= r(r+1)(1-w)^{-(r+2)} \\ &\vdots \end{aligned}$$

In general, $f^{(z)}(w) = r(r+1)\cdots(r+z-1)(1-w)^{-(r+z)}$, where $f^{(z)}(w)$ denotes the zth derivative of f with respect to w. Note that

$$f^{(z)}(w)\Big|_{w=0} = r(r+1)\cdots(r+z-1).$$

Now, consider writing the McLaurin Series expansion of $f(w)$; i.e., a Taylor Series expansion of $f(w)$ about $w=0$; this expansion is given by

$$f(w) = \sum_{z=0}^{\infty} \frac{f^{(z)}(0)}{z!} w^z = \sum_{z=0}^{\infty} \frac{r(r+1)\cdots(r+z-1)}{z!} w^z = \sum_{z=0}^{\infty} \binom{r+z-1}{r-1} w^z.$$

Letting $w = qe^t$ and $z = x - r$ proves the lemma. □

Alternative definition: Suppose independent Bernoulli trials are performed. We have defined $X \sim \text{nib}(r,p)$ to record

$$X = \text{ number of trials needed to observe the } r\text{th success.}$$

Define the random variable $Y = X - r$. Note that

$$Y = \text{ number of failures observed before the } r\text{th success.}$$

We can derive $f_Y(y) = f_Y(y|r,p)$ by performing a transformation for discrete random variables. First note that $\mathcal{Y} = \{y : y = 0, 1, 2, ...,\}$. Therefore, the pmf of $Y = g(X) = X - r$, for $y = 0, 1, 2, ...$, is given by

$$\begin{aligned} f_Y(y) = P_Y(Y=y) = P_X(X-r=y) &= P_X(X=y+r) \\ &= \binom{y+r-1}{y} p^r (1-p)^y. \end{aligned}$$

We can get the mean and variance of $Y = X - r$ easily:

$$E(Y) = E(X-r) = E(X) - r = \frac{r}{p} - r = \frac{rq}{p}$$

and

$$\text{var}(Y) = \text{var}(X-r) = \text{var}(X) = \frac{rq}{p^2}.$$

6. Poisson. A random variable X is said to have a Poisson distribution if its pmf is given by

$$f_X(x|\lambda) = \begin{cases} \dfrac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise,} \end{cases}$$

where $\lambda > 0$. The support of X is $\mathcal{X} = \{x : x = 0, 1, 2, ...,\}$. **Notation:** $X \sim \text{Poisson}(\lambda)$.

Mean/Variance: The relevant moments of $X \sim \text{Poisson}(\lambda)$ are

$$\begin{aligned} E(X) &= \lambda \\ \text{var}(X) &= \lambda. \end{aligned}$$

MGF: The mgf of $X \sim \text{Poisson}(\lambda)$ is

$$M_X(t) = e^{\lambda(e^t-1)}.$$

Conceptualization: A Poisson random variable X can be interpreted as counting the number of "occurrences" in a unit interval of time (or space), where the occurrences arise according to a **Poisson process**; see pp 135-136 (CB).

Recall: In Chapter 2, we showed that if $X_n \sim b(n, p_n)$, where $np_n = \lambda$ for all n, then $X_n \xrightarrow{d} X$, as $n \to \infty$, where $X \sim \text{Poisson}(\lambda)$. Therefore, if $X_n \sim b(n, p)$ and n is large,

$$P_{X_n}(X_n = x) = \binom{n}{x} p^x (1-p)^{n-x} \approx \frac{\lambda^x e^{-\lambda}}{x!},$$

where $\lambda = np$. The approximation is best when n is large (not surprising) and p is small. See pp 94 (CB) for a numerical example.

New Result: Suppose $\{Y_r\}$ is a sequence of random variables, where

$$f_{Y_r}(y) = \binom{y+r-1}{y} p^r (1-p)^y,$$

for $y = 0, 1, 2, ...,$. That is, Y_r follows a negative binomial distribution, but where Y_r records the number of failures before the rth success (i.e., under our alternative definition). This negative binomial distribution is linked to the Poisson distribution in the following way: If $r \to \infty$ and $p \to 1$ such that $r(1-p) \to \lambda > 0$, then $Y_r \xrightarrow{d} Y$, where $Y \sim \text{Poisson}(\lambda)$. This result can be established by first deriving $M_{Y_r}(t)$, the mgf of Y_r, and then showing

$$M_{Y_r}(t) \to e^{\lambda(e^t-1)},$$

the mgf of Y. See Exercise 3.15 (pp 130 CB).

3.3 Continuous Distributions

Recall: A random variable X is **continuous** if its cdf $F_X(x)$ is a continuous function.

1. Uniform. A random variable X is said to have a uniform distribution if its pdf is given by

$$f_X(x|a,b) = \begin{cases} \frac{1}{b-a}, & a < x < b \\ 0, & \text{otherwise,} \end{cases}$$

where $-\infty < a < b < \infty$. **Notation:** $X \sim \mathcal{U}(a, b)$.

Mean/Variance: The relevant moments of $X \sim \mathcal{U}(a, b)$ are

$$\begin{aligned} E(X) &= \frac{a+b}{2} \\ \text{var}(X) &= \frac{(b-a)^2}{12}. \end{aligned}$$

MGF: The mgf of $X \sim \mathcal{U}(a, b)$ is

$$M_X(t) = \begin{cases} \dfrac{e^{bt} - e^{at}}{(b-a)t}, & t \neq 0 \\ 1, & t = 0. \end{cases}$$

CDF: The cdf of $X \sim \mathcal{U}(a, b)$ is

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

Remark: A special member of the $\mathcal{U}(a, b)$ family arises when $a = 0$ and $b = 1$. It is called the "standard" uniform distribution; $X \sim \mathcal{U}(0, 1)$.

2. Gamma. A random variable X is said to have a gamma distribution if its pdf is given by

$$f_X(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} I(x > 0),$$

where $\alpha > 0$ and $\beta > 0$. **Notation:** $X \sim \text{gamma}(\alpha, \beta)$. Recall that

$$\begin{aligned} \alpha &\longrightarrow \text{"shape parameter"} \\ \beta &\longrightarrow \text{"scale parameter."} \end{aligned}$$

The cdf of X can not be written in closed form; i.e., it can be expressed as an integral of $f_X(x|\alpha, \beta)$, but it can not be simplified.

Gamma function: For $\alpha > 0$, define the function

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du.$$

The gamma function satisfies certain properties:

1. $\Gamma(1) = 1$
2. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
3. $\Gamma(1/2) = \sqrt{\pi}$.

Note that if $\alpha \in \mathbb{N} = \{1, 2, 3, ..., \}$, then second (recursive) property implies

$$\Gamma(\alpha) = (\alpha - 1)!$$

Mean/Variance: The relevant moments of $X \sim \text{gamma}(\alpha, \beta)$ are

$$\begin{aligned} E(X) &= \alpha\beta \\ \text{var}(X) &= \alpha\beta^2. \end{aligned}$$

MGF: The mgf of $X \sim \text{gamma}(\alpha, \beta)$ is

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}, \qquad t < \frac{1}{\beta}.$$

Connection with Poisson distribution: Suppose that we observe events according to a Poisson process (with intensity parameter $\lambda > 0$). Define

$$W = \text{ time until the } \alpha\text{th event.}$$

Then $W \sim \text{gamma}(\alpha, \beta)$, where $\beta = 1/\lambda$.
Proof. Clearly, W is a non-negative random variable that is continuous. The cdf of W, for $w > 0$, is given by

$$\begin{aligned} F_W(w) = P_W(W \leq w) &= 1 - P_W(W > w) \\ &= 1 - \text{pr}(\{\text{fewer than } \alpha \text{ events in } [0, w]\}) \\ &= 1 - \sum_{j=0}^{\alpha-1} \frac{(\lambda w)^j e^{-\lambda w}}{j!}. \end{aligned}$$

Result: If $X \sim \text{Poisson}(\lambda)$, then X counts the number of events over a unit interval of time. Over an interval of length $w > 0$, the number of events is Poisson with mean λw.

The pdf of W, for $w > 0$, is given by

$$\begin{aligned} f_W(w) = \frac{d}{dw} F_W(w) &= \lambda e^{-\lambda w} - e^{-\lambda w} \underbrace{\sum_{j=1}^{\alpha-1} \left[\frac{j(\lambda w)^{j-1}\lambda}{j!} - \frac{(\lambda w)^j \lambda}{j!} \right]}_{\text{telescoping sum}} \\ &= \lambda e^{-\lambda w} - e^{-\lambda w}\left[\lambda - \frac{\lambda(\lambda w)^{\alpha-1}}{(\alpha-1)!}\right] \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\lambda w}, \end{aligned}$$

which is the pdf of $W \sim \text{gamma}(\alpha, \beta)$, where $\beta = 1/\lambda$. □

Integration Trick: Because the gamma(α, β) pdf integrates to one, we have

$$\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx = 1 \quad \Longrightarrow \quad \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \Gamma(\alpha)\beta^\alpha.$$

This result is extremely useful and will be used repeatedly.

3. Exponential. A random variable X is said to have an exponential distribution if its pdf is given by

$$f_X(x|\beta) = \frac{1}{\beta}e^{-x/\beta}I(x > 0),$$

where $\beta > 0$. **Notation:** $X \sim$ exponential(β). The exponential(β) distribution is a special case of the gamma(α, β) distribution when $\alpha = 1$.

Mean/Variance: The relevant moments of $X \sim$ exponential(β) are

$$\begin{aligned} E(X) &= \beta \\ \text{var}(X) &= \beta^2. \end{aligned}$$

MGF: The mgf of $X \sim$ exponential(β) is

$$M_X(t) = \frac{1}{1-\beta t}, \qquad t < \frac{1}{\beta}.$$

CDF: The cdf of $X \sim$ exponential(β) is

$$F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-x/\beta}, & x > 0. \end{cases}$$

Memoryless Property: For $s > t \geq 0$,

$$P_X(X > s|X > t) = P_X(X > s - t).$$

The exponential distribution is the only continuous distribution that has this property.

Recall: From our previous result relating the gamma and Poisson distributions, we see that the exponential distribution with mean $\beta = 1/\lambda$ describes the time to the first event in a Poisson process with intensity parameter $\lambda > 0$.

4. Chi-squared. A random variable X is said to have a chi-squared distribution with p degrees of freedom if its pdf is given by

$$f_X(x|p) = \frac{1}{\Gamma(\frac{p}{2})2^{p/2}}x^{\frac{p}{2}-1}e^{-x/2}I(x > 0),$$

where $p > 0$. **Notation:** $X \sim \chi^2_p$. The χ^2_p distribution is a special case of the gamma(α, β) distribution when $\alpha = p/2$ and $\beta = 2$. Usually, p will be an integer. The χ^2 distribution arises often in applied statistics.

Mean/Variance: The relevant moments of $X \sim \chi^2_p$ are

$$\begin{aligned} E(X) &= p \\ \text{var}(X) &= 2p. \end{aligned}$$

MGF: The mgf of $X \sim \chi^2_p$ is

$$M_X(t) = \left(\frac{1}{1-2t}\right)^{p/2}, \qquad t < \frac{1}{2}.$$

5. Weibull. A random variable X is said to have a Weibull distribution if its pdf is given by

$$f_X(x|\gamma,\beta) = \frac{\gamma}{\beta}x^{\gamma-1}e^{-x^\gamma/\beta}I(x>0),$$

where $\gamma > 0$ and $\beta > 0$. **Notation:** $X \sim \text{Weibull}(\gamma,\beta)$. The Weibull distribution is used extensively in engineering applications. When $\gamma = 1$, the Weibull(γ,β) distribution reduces to the exponential(β) distribution.

Mean/Variance: The relevant moments of $X \sim \text{Weibull}(\gamma,\beta)$ are

$$\begin{aligned} E(X) &= \beta^{1/\gamma}\Gamma(1+1/\gamma) \\ \text{var}(X) &= \beta^{2/\gamma}\left[\Gamma(1+2/\gamma) - \Gamma^2(1+1/\gamma)\right]. \end{aligned}$$

The mgf of X exists only when $\gamma \geq 1$. Its form is not very useful.

6. Normal. A random variable X is said to have a normal (or Gaussian) distribution if its pdf is given by

$$f_X(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}I(x \in \mathbb{R}),$$

where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. **Notation:** $X \sim \mathcal{N}(\mu,\sigma^2)$.

Mean/Variance: The relevant moments of $X \sim \mathcal{N}(\mu,\sigma^2)$ are

$$\begin{aligned} E(X) &= \mu \\ \text{var}(X) &= \sigma^2. \end{aligned}$$

MGF: The mgf of $X \sim \mathcal{N}(\mu,\sigma^2)$ is

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2}.$$

Result: If $X \sim \mathcal{N}(\mu,\sigma^2)$, then

$$Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1).$$

Proof. We can derive the pdf of Z using a transformation; note that $g(x) = (x-\mu)/\sigma$ is one-to-one over $\mathbb{R}$, the support of X. The support of Z is also $\mathbb{R} = \{z : -\infty < z < \infty\}$. The inverse transformation $x = g^{-1}(z)$ is found as follows:

$$z = g(x) = \frac{x-\mu}{\sigma} \implies x = g^{-1}(z) = \sigma z + \mu.$$

The Jacobian is

$$\frac{d}{dz}g^{-1}(z) = \frac{d}{dz}(\sigma z + \mu) = \sigma.$$

Applying Theorem 2.1.5 directly, we have, for $z \in \mathbb{R}$,

$$\begin{aligned} f_Z(z) = f_X(g^{-1}(z))\left|\frac{d}{dz}g^{-1}(z)\right| &= \frac{1}{\sqrt{2\pi}\sigma}e^{-(\sigma z+\mu-\mu)^2/2\sigma^2} \times \sigma \\ &= \frac{1}{\sqrt{2\pi}}e^{-z^2/2}. \end{aligned}$$

This is the pdf of $Z \sim \mathcal{N}(0,1)$. $\square$

Remark: We can derive $E(Z) = 0$ and $\text{var}(Z) = 1$ directly (i.e., using expected value definitions). First,

$$E(Z) = \int_{\mathbb{R}} z\ \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz = -\frac{1}{\sqrt{2\pi}}\left(e^{-z^2/2}\Big|_{-\infty}^{\infty}\right) = -\frac{1}{\sqrt{2\pi}}(0-0) = 0.$$

Second,

$$E(Z^2) = \int_{\mathbb{R}} \underbrace{z^2\ \frac{1}{\sqrt{2\pi}}e^{-z^2/2}}_{=\ g(z),\ \text{say}}\, dz$$

Note that $g(z)$ is an even function; i.e., $g(z) = g(-z)$, for all $z \in \mathbb{R}$. This means that $g(z)$ is symmetric about $z = 0$. Therefore, the last integral

$$\int_{\mathbb{R}} z^2\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz = \int_0^{\infty} z^2\frac{2}{\sqrt{2\pi}}e^{-z^2/2}dz.$$

Now, let $u = z^2 \Longrightarrow du = 2zdz$. The last integral equals

$$\begin{aligned} \int_0^{\infty} z^2\frac{2}{\sqrt{2\pi}}e^{-z^2/2}dz &= \frac{2}{\sqrt{2\pi}}\int_0^{\infty} ue^{-u/2}du\ \frac{1}{2z} \\ &= \frac{2}{\sqrt{2\pi}}\int_0^{\infty} ue^{-u/2}\ \frac{1}{2\sqrt{u}}du \\ &= \frac{1}{\sqrt{2\pi}}\int_0^{\infty} u^{\frac{3}{2}-1}e^{-u/2}du. \end{aligned}$$

Recognizing $u^{\frac{3}{2}-1}e^{-u/2}$ as the kernel of a gamma distribution with shape parameter $\alpha = 3/2$ and scale parameter $\beta = 2$ (and because we are integrating over $\mathbb{R}^+$), the last expression

$$\begin{aligned} \frac{1}{\sqrt{2\pi}}\int_0^{\infty} u^{\frac{3}{2}-1}e^{-u/2}du &= \frac{1}{\sqrt{2\pi}}\Gamma\left(\frac{3}{2}\right)2^{3/2} \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{2}\Gamma\left(\frac{1}{2}\right)2^{3/2} = 1, \end{aligned}$$

because $\Gamma(1/2) = \sqrt{\pi}$. We have shown that $E(Z^2) = 1$. However, because $E(Z) = 0$, it follows that $\text{var}(Z) = E(Z^2) = 1$ as well.

Note: Because

$$Z \sim \mathcal{N}(0,1) \quad \Longrightarrow \quad X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2),$$

it follows immediately that

$$E(X) = E(\sigma Z + \mu) = \sigma E(Z) + \mu = \mu$$

and

$$\text{var}(X) = \text{var}(\sigma Z + \mu) = \sigma^2 \text{var}(Z) = \sigma^2.$$

Remaining issues:

- Showing that $f_Z(z)$ integrates to 1 over $\mathbb{R}$ is an interesting integration exercise using polar coordinates; see pp 103-104 (CB).
- We can show directly that if $Z \sim \mathcal{N}(0,1)$, then the mgf of Z is given by

$$M_Z(t) = E(e^{tZ}) = \int_{\mathbb{R}} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = e^{t^2/2}.$$

- With $M_Z(t) = e^{t^2/2}$ and $X = \sigma Z + \mu$, we can use Theorem 2.3.15 to show that

$$M_X(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t} e^{(\sigma t)^2/2} = e^{\mu t + \sigma^2 t^2/2}.$$

7. Beta. A random variable X is said to have a beta distribution if its pdf is given by

$$f_X(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I(0<x<1),$$

where $\alpha > 0$ and $\beta > 0$. **Notation:** $X \sim \text{beta}(\alpha, \beta)$. Note that the support of X is $\mathcal{X} = \{x : 0 < x < 1\}$; this is different than our other "named" distributions. The beta distribution is useful in modeling proportions (or probabilities).

Mean/Variance: The relevant moments of $X \sim \text{beta}(\alpha, \beta)$ are

$$\begin{aligned} E(X) &= \frac{\alpha}{\alpha+\beta} \\ \text{var}(X) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \end{aligned}$$

The mgf of X exists, but its form is usually not very helpful.

Remark: The pdf of $X \sim \text{beta}(\alpha, \beta)$ is sometimes displayed as

$$f_X(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} I(0<x<1),$$

where

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx$$

is the **beta function**. Note that integrals of the form $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ can therefore be calculated quickly (similarly to our gamma integration result).

Example 3.3. If $X \sim \text{beta}(\alpha, \beta)$, derive $E(X^k)$, where $k > 0$.
Solution. By definition,

$$\begin{aligned} E(X^k) &= \int_0^1 x^k \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+k+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+k)}. \end{aligned}$$

The mean of X, for example, is

$$E(X) = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)} = \frac{\Gamma(\alpha+\beta)\ \alpha\Gamma(\alpha)}{\Gamma(\alpha)\ (\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta},$$

as claimed. To derive $\text{var}(X)$, calculate $E(X^2)$ and use the variance computing formula.

Remark: The pdf of $X \sim \text{beta}(\alpha, \beta)$ is very flexible; i.e., the pdf $f_X(x|\alpha, \beta)$ can assume many shapes over $\mathcal{X} = \{x : 0 < x < 1\}$. For example,

1. $\alpha = \beta \Longrightarrow f_X(x|\alpha, \beta)$ is symmetric about $x = 1/2$
   - $\alpha = \beta = 1 \Longrightarrow X \sim \mathcal{U}(0, 1)$
   - $\alpha = \beta = \frac{1}{2} \Longrightarrow f_X(x|\alpha, \beta) \propto x^{\frac{1}{2}-1}(1-x)^{\frac{1}{2}-1}$ is "bathtub-shaped"
2. $\alpha > \beta \Longrightarrow f_X(x|\alpha, \beta)$ is skewed left
3. $\alpha < \beta \Longrightarrow f_X(x|\alpha, \beta)$ is skewed right.

8. Cauchy. A random variable X is said to have a Cauchy distribution if its pdf is given by

$$f_X(x|\mu, \sigma) = \frac{1}{\pi\sigma\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]} I(x \in \mathbb{R}),$$

where $-\infty < \mu < \infty$ and $\sigma > 0$. **Notation:** $X \sim \text{Cauchy}(\mu, \sigma)$. The parameters μ and σ do not represent the mean and standard deviation, respectively.

Note: When $\mu = 0$ and $\sigma = 1$, we have $Z \sim \text{Cauchy}(0, 1)$, which is known as the "standard" Cauchy distribution. The pdf of Z is

$$f_Z(z) = \frac{1}{\pi(1+z^2)} I(z \in \mathbb{R})$$

and $X \sim \text{Cauchy}(\mu, \sigma)$ and $Z \sim \text{Cauchy}(0, 1)$ are related via $X = \sigma Z + \mu$, similar to what we observed in the $\mathcal{N}(\mu, \sigma^2)$ family. Note that

$$\int_{\mathbb{R}} f_Z(z)dz = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{1+z^2}dz = \frac{1}{\pi}\left(\arctan z \Big|_{-\infty}^{\infty}\right) = \frac{1}{\pi}\left[\frac{\pi}{2} - \left(-\frac{\pi}{2}\right)\right] = 1,$$

showing that $f_Z(z)$ is a valid pdf.

Remark: If $Z \sim \text{Cauchy}(0, 1)$, then $E(Z)$ does not exist.
Proof. It suffices to show that $E(|Z|) = +\infty$. Note that

$$E(|Z|) = \int_{\mathbb{R}} |z| f_Z(z)dz = \int_{\mathbb{R}} \underbrace{|z| \, \frac{1}{\pi(1+z^2)}}_{=\, g(z),\ \text{say}} dz.$$

Note that $g(z)$ is an even function; i.e., $g(z) = g(-z)$, for all $z \in \mathbb{R}$. This means that $g(z)$ is symmetric about $z = 0$. Therefore, the last integral

$$\begin{aligned}\int_{\mathbb{R}} |z| \, \frac{1}{\pi(1+z^2)}dz &= \frac{2}{\pi}\int_0^{\infty} \frac{z}{1+z^2}dz \\ &= \frac{2}{\pi}\left[\frac{\ln(1+z^2)}{2}\Big|_0^{\infty}\right] = +\infty. \ \square\end{aligned}$$

This result implies that none of Z's higher order moments exist; e.g., $E(Z^2)$, $E(Z^3)$, etc. Also, because $X \sim \text{Cauchy}(\mu, \sigma)$ and $Z \sim \text{Cauchy}(0, 1)$ are related via $X = \sigma Z + \mu$, none of X's moments exist either.

Other "named" continuous distributions: There are hundreds (thousands?) of other "named" continuous distributions. In many ways, this should not be surprising because it is easy to come up with a valid pdf. If $h(x)$ is a non-negative function with domain $\mathcal{D}$ and $\int_{\mathcal{D}} h(x)dx = K < \infty$, then $f_X(x) = h(x)/K$ is a valid pdf! Below I list some additional named continuous distributions; see CB for pdf and moment formulae.

- **Lognormal.** $X \sim \text{lognormal}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. This distribution arises according to the following transformation:

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \Longrightarrow \quad Y = g(X) = e^X \sim \text{lognormal}(\mu, \sigma^2).$$

  - positive support: $\mathcal{X} = \{x : x > 0\}$
  - popular competitor to the Weibull distribution in reliability/engineering applications.

- **LaPlace.** $X \sim \text{LaPlace}(\mu, \sigma)$, where $-\infty < \mu < \infty$ and $\sigma > 0$.

  - support over $\mathbb{R}$: $\mathcal{X} = \{x : -\infty < x < \infty\}$
  - pdf $f_X(x|\mu, \sigma)$ is symmetric about μ and has heavy tails, which makes it useful in robustness discussions

- sometimes also called the "double exponential distribution."

- **Inverted Gamma.** $X \sim \text{IG}(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$. This distribution arises according to the following transformation:

$$X \sim \text{gamma}(\alpha, \beta) \quad \Longrightarrow \quad Y = g(X) = \frac{1}{X} \sim \text{IG}(\alpha, \beta).$$

  - positive support: $\mathcal{X} = \{x : x > 0\}$
  - useful distribution to model variances in a Bayesian framework.

- **Pareto.** $X \sim \text{Pareto}(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$. This distribution arises according to the following transformation:

$$X \sim \text{exponential}(1/\beta) \quad \Longrightarrow \quad Y = g(X) = \alpha e^X \sim \text{Pareto}(\alpha, \beta).$$

  - support: $\mathcal{X} = \{x : x > \alpha\}$
  - useful in economics applications; e.g., income distributions, etc.

- **Logistic.** $X \sim \text{Logistic}(\mu, \beta)$, where $-\infty < \mu < \infty$ and $\sigma > 0$. This distribution arises according to the following transformation:

$$X \sim \text{exponential}(1) \quad \Longrightarrow \quad Y = g(X) = \mu + \beta \ln\left(e^X - 1\right) \sim \text{Logistic}(\mu, \beta).$$

  - support over $\mathbb{R}$: $\mathcal{X} = \{x : -\infty < x < \infty\}$
  - if you take the cdf of $X \sim \text{Logistic}(\mu, \beta)$ and write

$$\ln\left(\frac{F_X(x)}{1 - F_X(x)}\right),$$

    this is a linear function of x; this forms the basis for **logistic regression**.

Note: There are many more distributions that I will not list. Many "named" distributions arise in CB's exercises; look for these and do them. An excellent expository account of probability distributions (both discrete and continuous) and distributional relationships is given in the following paper:

- Leemis, L. and McQueston, J. (2008). Univariate distribution relationships. *American Statistician* **62**, 45-53.

3.4 Exponential Families

Definition: A family $\{f_X(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ of pdfs (or pmfs) is called an **exponential family** if its members can be expressed as

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right\},$$

for real functions $h(\cdot)$, $c(\cdot)$, $w_i(\cdot)$, and $t_i(\cdot)$, where

- $h(x) \geq 0$ cannot depend on $\boldsymbol{\theta}$
- $c(\boldsymbol{\theta}) \geq 0$ cannot depend on x
- $w_1(\boldsymbol{\theta}), w_2(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta})$ cannot depend on x
- $t_1(x), t_2(x), ..., t_k(x)$ cannot depend on $\boldsymbol{\theta}$.

Remark: Many families we know are exponential families; e.g., Poisson, normal, binomial, gamma, beta, etc. However, not all families can be "put into" this form; i.e., there are families that are not exponential families. We will see examples of some later.

Example 3.4. Suppose that $X \sim \text{Poisson}(\theta)$; i.e., the pmf of X is

$$f_X(x|\theta) = \begin{cases} \dfrac{\theta^x e^{-\theta}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise,} \end{cases}$$

where $\Theta = \{\theta : \theta > 0\}$. Write the support $\mathcal{X} = \{x : x = 0, 1, 2, ..., \}$ and the pmf as

$$\begin{aligned} f_X(x|\theta) &= \frac{\theta^x e^{-\theta}}{x!} I(x \in \mathcal{X}) \\ &= \frac{I(x \in \mathcal{X})}{x!}\, e^{-\theta}\, e^{x \ln \theta} \\ &= h(x)c(\theta) \exp\left\{w_1(\theta)t_1(x)\right\}, \end{aligned}$$

where $h(x) = I(x \in \mathcal{X})/x!$, $c(\theta) = e^{-\theta}$, $w_1(\theta) = \ln \theta$, and $t_1(x) = x$. Therefore, the Poisson(θ) family of pmfs is an exponential family with $k = 1$.

Example 3.5. Suppose that $X \sim \text{gamma}(\alpha, \beta)$; i.e., the pdf of X is

$$f_X(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} I(x > 0),$$

where $\Theta = \{\boldsymbol{\theta} = (\alpha, \beta)' : \alpha > 0, \beta > 0\}$. Note that the pdf of X can be written as

$$\begin{aligned} f_X(x|\alpha, \beta) &= \frac{I(x > 0)}{x} \frac{1}{\Gamma(\alpha)\beta^\alpha}\, e^{\alpha \ln x}\, e^{-x/\beta} \\ &= \frac{I(x > 0)}{x} \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp\left(\alpha \ln x - \frac{x}{\beta}\right) \\ &= h(x)c(\boldsymbol{\theta}) \exp\{w_1(\boldsymbol{\theta})t_1(x) + w_2(\boldsymbol{\theta})t_2(x)\}, \end{aligned}$$

where $h(x) = I(x > 0)/x$, $c(\boldsymbol{\theta}) = [\Gamma(\alpha)\beta^\alpha]^{-1}$, $w_1(\boldsymbol{\theta}) = \alpha$, $t_1(x) = \ln x$, $w_2(\boldsymbol{\theta}) = -1/\beta$, and $t_2(x) = x$. Therefore, the gamma(α, β) family of pdfs is an exponential family with $k = 2$.

Remark: As noted earlier, some families are not exponential families. For example, suppose that $X \sim \text{LaPlace}(\mu, \sigma)$; i.e., the pdf of X is

$$f_X(x|\mu, \sigma) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma} I(x \in \mathbb{R}),$$

where $\Theta = \{\boldsymbol{\theta} = (\mu, \sigma)' : -\infty < \mu < \infty,\ \sigma > 0\}$. It is not possible to put this pdf into exponential family form (the absolute value term $|x - \mu|$ messes things up). As another example, suppose $X \sim f_X(x|\theta)$, where

$$f_X(x|\theta) = e^{-(x-\theta)} I(x > \theta),$$

where $\Theta = \{\theta : -\infty < \theta < \infty\}$. The indicator function $I(x > \theta) \equiv I_{(\theta,\infty)}(x)$ can neither be "absorbed" into $h(x)$ nor into $c(\theta)$.

Important: Anytime you have a pdf/pmf $f_X(x|\boldsymbol{\theta})$ where the support $\mathcal{X}$ depends on an unknown parameter $\boldsymbol{\theta}$, it is not possible to put $f_X(x|\boldsymbol{\theta})$ into exponential family form.

Remark: In some instances, it is helpful to work with the exponential family in its canonical representation; see pp 114 (CB). We will not highlight this parameterization.

Important: Suppose that X has pdf in the exponential family; i.e., the pdf of X can be expressed as

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(x)\right\},$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)'$ and $d = \dim(\boldsymbol{\theta})$.

- When $d = k$, we call $\{f_X(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ a **full** exponential family.
- When $d < k$, we call $\{f_X(x|\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ a **curved** exponential family.

Example 3.6. Suppose that $X \sim \text{gamma}(\alpha, \beta)$; i.e., the pdf of X is

$$f_X(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} I(x > 0),$$

where $\Theta = \{\boldsymbol{\theta} = (\alpha, \beta)' : \alpha > 0, \beta > 0\}$. In Example 3.5, we showed that this was an exponential family with $d = k = 2$. Therefore, the gamma(α, β) family is a **full** exponential family. Now consider the gamma(α, β) **subfamily** where $\beta = 1/\alpha^2$, that is, $X \sim \text{gamma}(\alpha, 1/\alpha^2)$. The pdf of X is

$$\begin{aligned}
f_X(x|\alpha) &= \frac{1}{\Gamma(\alpha)\left(\frac{1}{\alpha^2}\right)^\alpha}\ x^{\alpha-1} e^{-x/(1/\alpha^2)} I(x > 0) \\
&= \frac{I(x > 0)}{x} \frac{\alpha^{2\alpha}}{\Gamma(\alpha)}\ e^{\alpha \ln x}\ e^{-\alpha^2 x} \\
&= \frac{I(x > 0)}{x} \frac{\alpha^{2\alpha}}{\Gamma(\alpha)} \exp\left(\alpha \ln x - \alpha^2 x\right) \\
&= h(x)c(\alpha) \exp\{w_1(\alpha)t_1(x) + w_2(\alpha)t_2(x)\},
\end{aligned}$$

where $h(x) = I(x > 0)/x$, $c(\alpha) = \alpha^{2\alpha}/\Gamma(\alpha)$, $w_1(\alpha) = \alpha$, $t_1(x) = \ln x$, $w_2(\alpha) = -\alpha^2$, and $t_2(x) = x$. Therefore, the gamma$(\alpha, 1/\alpha^2)$ subfamily has $d = 1$ and $k = 2$. Because $d < k$, the gamma$(\alpha, 1/\alpha^2)$ family is a **curved** exponential family.

Remark: For the original gamma(α, β) family, the **parameter space** is

$$\Theta = \{\boldsymbol{\theta} = (\alpha, \beta)' : \alpha > 0, \beta > 0\}.$$

For the gamma$(\alpha, 1/\alpha^2)$ subfamily, the parameter space is

$$\Theta_0 = \left\{\boldsymbol{\theta} = (\alpha, \beta)' : \alpha > 0, \beta = \frac{1}{\alpha^2}\right\}.$$

Clearly $\Theta_0 \subset \Theta$. Also, note that Θ contains an open set in $\mathbb{R}^2$, but Θ_0 does not. These theoretical issues will be important in Chapter 6 when we study sufficient statistics (data reduction) and completeness.

3.5 Location and Scale Families

Result: Suppose that Z is a continuous random variable with cdf $F_Z(z)$ and pdf $f_Z(z)$. Define $X = \sigma Z + \mu$, where $-\infty < \mu < \infty$ and $\sigma > 0$. The pdf of X can be written in terms of the pdf of Z, specifically,

$$f_X(x|\mu, \sigma) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right).$$

Proof. The cdf of X is

$$\begin{aligned} F_X(x|\mu,\sigma) = P_X(X \leq x|\mu,\sigma) &= P_Z(\sigma Z + \mu \leq x) \\ &= P_Z\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

The pdf of X is therefore

$$\begin{aligned} f_X(x|\mu,\sigma) = \frac{d}{dx} F_X(x|\mu,\sigma) &= \frac{d}{dx} F_Z\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right), \end{aligned}$$

the last step following from the chain rule. That $f_X(x|\mu, \sigma)$ is a valid pdf is easy to show. Clearly $f_X(x|\mu, \sigma)$ is non-negative because $\sigma > 0$ and $f_Z(z)$ is a pdf. Also,

$$\int_{\mathbb{R}} f_X(x|\mu,\sigma)dx = \int_{\mathbb{R}} \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right) dx = \int_{\mathbb{R}} f_Z(z)dz = 1,$$

the last step following after making a $z = (x - \mu)/\sigma$ substitution. □

Remark: In the language of location-scale families, we call $f_Z(z)$ a **standard pdf**. With $f_Z(z)$ and the transformation $X = \sigma Z + \mu$, we can "generate" a family of probability distributions indexed by μ and σ.

Remark: What we have just proven is essentially the sufficiency part ($\Longleftarrow$) of Theorem 3.5.6 (pp 120 CB). The necessity part ($\Longrightarrow$) is also true, that is, if

$$f_X(x|\mu,\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right),$$

then there exists a random variable $Z \sim f_Z(z)$ such that $X = \sigma Z + \mu$.

Definition: The collection of pdfs

$$\{f_X(x|\mu) = f_Z(x-\mu);\ \mu \in \mathbb{R}\}$$

is called a **location family** generated by $f_Z(z)$. The parameter μ is called a **location parameter**. From our previous result (taking $\sigma = 1$), we have that

$$Z \sim f_Z(z) \implies X = Z + \mu \sim f_X(x|\mu) = f_Z(x-\mu).$$

Example 3.7. Suppose that $Z \sim$ exponential(1); i.e., the pdf of Z is

$$f_Z(z) = e^{-z} I(z > 0).$$

The pdf of $X = Z + \mu$ is therefore

$$\begin{aligned} f_X(x|\mu) = f_Z(x-\mu) &= e^{-(x-\mu)} I(x-\mu > 0) \\ &= e^{-(x-\mu)} I(x > \mu). \end{aligned}$$

This is called a **shifted exponential distribution** with location parameter μ. The pdf of any member of this family is obtained by taking $f_Z(z)$ and shifting it to the left or right depending on if $\mu < 0$ or $\mu > 0$.

Definition: The collection of pdfs

$$\left\{f_X(x|\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right);\ \sigma > 0\right\}$$

is called a **scale family** generated by $f_Z(z)$. The parameter σ is called a **scale parameter**. From our previous result (taking $\mu = 0$), we have that

$$Z \sim f_Z(z) \implies X = \sigma Z \sim f_X(x|\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right).$$

Example 3.8. Suppose that $X \sim \mathcal{N}(0, \sigma^2)$; i.e., the pdf of X is

$$f_X(x|\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} I(x \in \mathbb{R}),$$

where $\sigma^2 > 0$. Show that the $\mathcal{N}(0, \sigma^2)$ family is a scale family.
Solution. We have to identify the standard pdf $f_Z(z)$ and the scale parameter σ that makes

$$f_X(x|\sigma^2) = \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right).$$

Note that if we take the $\mathcal{N}(0,1)$ pdf; i.e.,

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} I(z \in \mathbb{R}),$$

then

$$\begin{aligned} \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right) &= \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x}{\sigma}\right)^2/2} I\left(\frac{x}{\sigma} \in \mathbb{R}\right) \\ &= \underbrace{\frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} I(x \in \mathbb{R})}_{= f_X(x|\sigma^2)}. \end{aligned}$$

Thus, the $\mathcal{N}(0, \sigma^2)$ family is a scale family with standard pdf $f_Z(z)$ and scale parameter σ.

Remark: In Example 3.8, we see that the scale parameter $\sigma > 0$ does, in fact, represent the standard deviation of X. However, in general, μ and σ do not necessarily represent the mean and standard deviation.

Definition: The collection of pdfs

$$\left\{ f_X(x|\mu,\sigma) = \frac{1}{\sigma} f_Z\left(\frac{x-\mu}{\sigma}\right);\ \mu \in \mathbb{R},\ \sigma > 0 \right\}$$

is called a **location-scale family** generated by $f_Z(z)$.

Example 3.9. Suppose that $X \sim \text{Cauchy}(\mu, \sigma)$; i.e., the pdf of X is

$$f_X(x|\mu,\sigma) = \frac{1}{\pi\sigma\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]} I(x \in \mathbb{R}),$$

where $-\infty < \mu < \infty$ and $\sigma > 0$. It is easy to see that the Cauchy(μ, σ) family is a location-scale family generated by the standard pdf

$$f_Z(z) = \frac{1}{\pi(1+z^2)} I(z \in \mathbb{R}).$$

However, note that μ and σ do not refer to the mean and standard deviation of $X \sim \text{Cauchy}(\mu, \sigma)$; recall that $E(X)$ does not even exist for $X \sim \text{Cauchy}(\mu, \sigma)$. In this family, μ satisfies

$$P_X(X \leq \mu) = \int_{-\infty}^{\mu} f_X(x|\mu,\sigma)dx = 0.5;$$

i.e., μ is the **median** of X. Also, $2\sigma = \text{IQR}(X)$, the **interquartile range** of X.

Theorem 3.5.7. Suppose that $Z \sim f_Z(z)$ and let $X = \sigma Z + \mu$. If $E(Z)$ and $\text{var}(Z)$ exist, then

$$E(X) = \sigma E(Z) + \mu \qquad \text{and} \qquad \text{var}(X) = \sigma^2 \text{var}(Z).$$

Proof. The expected value of X is

$$\begin{aligned} E(X) = E(\sigma Z + \mu) &= \int_{\mathbb{R}} (\sigma z + \mu) f_Z(z) dz \\ &= \sigma \int_{\mathbb{R}} z f_Z(z) dz + \int_{\mathbb{R}} \mu f_Z(z) dz \\ &= \sigma E(Z) + \mu. \end{aligned}$$

Showing $\text{var}(X) = \sigma^2 \text{var}(Z)$ involves slightly more work, but is just as straightforward. □

Special case: If $E(Z) = 0$ and $\text{var}(Z) = 1$, then $E(X) = \mu$ and $\text{var}(X) = \sigma^2$.

Calculating probabilities: Suppose that $Z \sim f_Z(z), F_Z(z)$ and define $X = \sigma Z + \mu$. We know that X has a location-scale pdf given by

$$f_X(x|\mu, \sigma) = \frac{1}{\sigma} f_Z\left(\frac{x - \mu}{\sigma}\right)$$

and

$$P_X(X \leq x|\mu, \sigma) = F_X(x|\mu, \sigma) = F_Z\left(\frac{x - \mu}{\sigma}\right).$$

Therefore, calculating probabilities of events of the form $\{X \leq x\}$ can be done by using the cdf of Z, regardless of the values of μ and σ. This fact is often exploited in introductory courses where $F_Z(z)$ is presented in tabular form; $Z \sim \mathcal{N}(0, 1)$, for example. Of course, with computing today (e.g., R, etc.), this clumsy method of calculation is no longer necessary.

3.6 Inequalities and Identities

Remark: This section is split into two sub-sections:

- Section 3.6.1. Probability Inequalities
- Section 3.6.2. Identities (read on your own).

We will discuss only two results; other results are left as exercises. Markov's Inequality is actually presented in the Miscellanea section; see pp 136 (CB).

Markov's Inequality: Suppose Y is a random variable with

- $P_Y(Y \geq 0) = 1$; i.e., Y is a lifetime random variable
- $P_Y(Y = 0) < 1$; i.e., Y is not degenerate at $y = 0$.

For any $r > 0$,

$$P_Y(Y \geq r) \leq \frac{E(Y)}{r}.$$

Proof. The expected value of Y is

$$\begin{aligned} E(Y) = \int_0^\infty y f_Y(y) dy &\geq \int_{\{y:y\geq r\}} y f_Y(y) dy \\ &\geq \int_{\{y:y\geq r\}} r f_Y(y) dy = r P_Y(Y \geq r). \ \square \end{aligned}$$

Chebyshev's Inequality: Suppose X is a random variable with $\text{var}(X) = \sigma^2 < \infty$. For any $k > 0$,

$$P_X(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof. Rewrite the event $\{|X - \mu| \geq k\sigma\} = \{(X - \mu)^2 \geq k^2\sigma^2\}$. This is justified because $|X - \mu|$, k, and σ are all non-negative. Therefore,

$$P_X(|X - \mu| \geq k\sigma) = P_X((X - \mu)^2 \geq k^2\sigma^2).$$

Now, apply Markov's Inequality to the RHS with $Y = (X - \mu)^2$ and $r = k^2\sigma^2$ to get

$$P_X((X - \mu)^2 \geq k^2\sigma^2) \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}. \ \square$$

Remarks:

- Chebyshev's Inequality can be equivalently stated as

$$P_X(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

- Both Markov and Chebyshev bounds can be very conservative (i.e., they can be very crude upper/lower bounds). For example, suppose that $Y \sim$ exponential(5). Using Markov's upper bound for $P_Y(Y > 15)$ gives

$$P_Y(Y > 15) \leq \frac{E(Y)}{15} = \frac{5}{15}.$$

  However, the actual probability; i.e., when calculated under the exponential(5) model assumption, is

$$P_Y(Y > 15) = \int_{15}^\infty \frac{1}{5} e^{-y/5} dy \approx 0.0498.$$

  That Markov and Chebyshev bounds are conservative in general should not be surprising. These bounds utilize very little information about the true underlying distribution.

4 Multiple Random Variables

Complementary reading: Chapter 4 (CB). Sections 4.1-4.7.

4.1 Joint and Marginal Distributions

Definition: Let $(S, \mathcal{B}, P)$ be a probability space for a random experiment. Suppose that $X^{-1}(B) \in \mathcal{B}$ and $Y^{-1}(B) \in \mathcal{B}$, for all $B \in \mathcal{B}(\mathbb{R})$; i.e., X and Y are both random variables on $(S, \mathcal{B}, P)$. We call (X, Y) a **bivariate random vector**.

- When viewed as a function, (X, Y) is a mapping from $(S, \mathcal{B}, P)$ to $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), P_{X,Y})$.
- Sets $B \in \mathcal{B}(\mathbb{R}^2)$ are called (two-dimensional) Borel sets. One can characterize $\mathcal{B}(\mathbb{R}^2)$ as the smallest σ-algebra generated by the collection of all half-open rectangles; i.e.,

$$\{(x_1, x_2) : -\infty < x_1 \leq a_1, -\infty < x_2 \leq a_2,\ a_1, a_2 \in \mathbb{R}\}.$$

- $P_{X,Y}$ is a probability measure induced by the random vector (X, Y).

Note: Generalizing this definition to n-dimensional random vectors $\mathbf{X} = (X_1, X_2, ..., X_n)$ is straightforward and we do this in Section 4.6. In this case, $\mathbf{X}$ is a mapping from $(S, \mathcal{B}, P)$ to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\mathbf{X}})$ with the property that

$$\mathbf{X}^{-1}(B) \equiv \{\omega \in S : \mathbf{X}(\omega) \in B\} \in \mathcal{B},$$

for all $B \in \mathcal{B}(\mathbb{R}^n)$. As in the univariate random variable case, the measurability condition $\mathbf{X}^{-1}(B) \in \mathcal{B}$, for all $B \in \mathcal{B}(\mathbb{R}^n)$, suggests that events of interest like $\{\mathbf{X} \in B\}$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\mathbf{X}})$ can be assigned a probability in the same way that $\{\omega \in S : \mathbf{X}(\omega) \in B\}$ can be assigned a probability on $(S, \mathcal{B}, P)$.

Example 4.1. Experiment: Toss two dice. Assume the model

$$\begin{aligned} S &= \{\omega = (\omega_1, \omega_2) :\ \omega_i \in \{1, 2, ..., 6\},\ i = 1, 2\} \\ \mathcal{B} &= 2^S \\ P &= \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/36, \text{ for all } \omega \in S. \end{aligned}$$

Define the random variables

$$\begin{aligned} X_1 &= \text{sum;} && \text{i.e., } X_1(\omega) = \omega_1 + \omega_2 \\ X_2 &= \text{absolute difference;} && \text{i.e., } X_2(\omega) = |\omega_1 - \omega_2| \end{aligned}$$

and let $\mathbf{X} = (X_1, X_2)$. The bivariate random vector

$$\mathbf{X}(\omega) = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}(\omega) = \begin{pmatrix} X_1(\omega) \\ X_2(\omega) \end{pmatrix}$$

is a vector of (univariate) random variables on $(S, \mathcal{B}, P)$. To show how probabilities are assigned on $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), P_{X_1,X_2})$, consider the (two-dimensional) Borel set $B = \{(5,3)\} \in \mathbb{R}^2$. Note that

$$\mathbf{X}^{-1}(\{(5,3)\}) = \{\omega \in S : \mathbf{X}(\omega) = (5,3)\} = \{(1,4),(4,1)\} \in \mathcal{B}.$$

This suggests that we can write

$$P_{X_1,X_2}(\mathbf{X} \in B) = \underbrace{P_{X_1,X_2}(X_1 = 5, X_2 = 3)}_{\text{calculated on } (\mathbb{R}^2,\mathcal{B}(\mathbb{R}^2))} = \underbrace{P(\{(1,4),(4,1)\})}_{\text{calculated on } (S,\mathcal{B})} = \frac{2}{36}.$$

Note: From now on, I will not emphasize probability as an induced measure (e.g., write P_X, $P_{X,Y}$, etc.), unless it is important to do so.

Definition: We call (X, Y) a **discrete random vector** if there exists a countable (support) set $\mathcal{A} \subset \mathbb{R}^2$ such that $P((X,Y) \in \mathcal{A}) = 1$. The **joint probability mass function** (pmf) of (X, Y) is a function $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ defined by

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

Analogous to the univariate case; i.e., as an extension of Theorem 1.6.5 (pp 36 CB), we have

(a) $f_{X,Y}(x,y) \geq 0$, for all $(x,y) \in \mathbb{R}^2$

(b) $\sum\sum_{(x,y)\in\mathcal{A}} f_{X,Y}(x,y) = 1$.

Also, for any $B \in \mathcal{B}(\mathbb{R}^2)$,

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

Definition: Suppose that (X, Y) is a discrete random vector with pmf $f_{X,Y}(x,y)$, support $\mathcal{A} \subset \mathbb{R}^2$, and suppose $g : \mathbb{R}^2 \to \mathbb{R}$. Then $g(X,Y)$ is a univariate random variable and its expected value is

$$E[g(X,Y)] = \sum\sum_{(x,y)\in\mathcal{A}} g(x,y) f_{X,Y}(x,y).$$

This definition is analogous to the definition of mathematical expectation for univariate discrete random variables. Existence issues are also identical; i.e., we need the sum above to converge absolutely (this concern only arises when $\mathcal{A}$ is countably infinite). If $\sum\sum_{(x,y)\in\mathcal{A}} |g(x,y)| f_{X,Y}(x,y)$ does not converge, then $E[g(X,Y)]$ does not exist.

Remark: The expectation properties summarized in Theorem 2.2.5 (pp 57 CB) for functions of univariate random variables also apply to functions of random vectors. Let a, b, and c be constants. For any functions $g_1(x,y)$ and $g_2(x,y)$ whose expectations exist,

(a) $E[ag_1(X,Y) + bg_2(X,Y) + c] = aE[g_1(X,Y)] + bE[g_2(X,Y)] + c$

(b) if $g_1(x,y) \geq 0$ for all x, y, then $E[g_1(X,Y)] \geq 0$

(c) if $g_1(x,y) \geq g_2(x,y)$ for all x, y, then $E[g_1(X,Y)] \geq E[g_2(X,Y)]$

(d) if $a \leq g_1(x,y) \leq b$ for all x, y, then $a \leq E[g_1(X,Y)] \leq b$.

These results are also true when (X,Y) is a continuous random vector (to be defined shortly).

Marginal Distributions (Discrete case): Suppose (X,Y) is a discrete random vector with pmf $f_{X,Y}(x,y)$. Suppose $B \in \mathcal{B}(\mathbb{R})$. Note that

$$\begin{aligned} P(X \in B) = P(X \in B, Y \in \mathbb{R}) &= P((X,Y) \in B \times \mathbb{R}) \\ &= \sum_{(x,y) \in B \times \mathbb{R}} \sum f_{X,Y}(x,y) \\ &= \sum_{x \in B} \underbrace{\sum_{y \in \mathbb{R}} f_{X,Y}(x,y)}_{= f_X(x)}. \end{aligned}$$

We call

$$f_X(x) = \sum_{y \in \mathbb{R}} f_{X,Y}(x,y)$$

the **marginal probability mass function** (pmf) of X. Similarly, we call

$$f_Y(y) = \sum_{x \in \mathbb{R}} f_{X,Y}(x,y)$$

the marginal pmf of Y. In other words, to find the marginal pmf of one random variable, you take the joint pmf and sum over the values of the other random variable.

Example 4.2. Suppose the joint distribution of (X,Y) is described via the following contingency table:

		y		
		0	**1**	**2**
x	**0**	0.1	0.2	0.2
	1	0.3	0.1	0.1

The entries in the table are the joint probabilities $f_{X,Y}(x,y)$. The support of (X,Y) is

$$\mathcal{A} = \{(0,0), (0,1), (0,2), (1,0), (1,1), (1,2)\}.$$

We use this joint pmf to calculate various quantities, illustrating many of the ideas we have seen so far. For example,

$$P(X = 1, Y = 1) = f_{X,Y}(1,1) = 0.1$$

and

$$\begin{aligned} E(XY) &= \sum_{x=0}^{1}\sum_{y=0}^{2} xy\ f_{X,Y}(x,y) \\ &= (0)(0)(0.1) + (0)(1)(0.2) + (0)(2)(0.2) + (1)(0)(0.3) + (1)(1)(0.1) + (1)(2)(0.1) \\ &= 0.3. \end{aligned}$$

The marginal pmfs of X and Y are, respectively,

$$\begin{aligned} f_X(x) &= 0.5I(x=0) + 0.5I(x=1) \\ f_Y(y) &= 0.4I(y=0) + 0.3I(y=1) + 0.3I(y=2). \end{aligned}$$

Note that

$$\begin{aligned} E(X) &= 0.5 \\ E(Y) &= 0.9. \end{aligned}$$

These can be calculated from the marginal distributions $f_X(x)$ and $f_Y(y)$, respectively, or from using the joint distribution, for example,

$$\begin{aligned} E(X) &= \sum_{x=0}^{1}\sum_{y=0}^{2} x\ f_{X,Y}(x,y) \\ &= (0)(0.1) + (0)(0.2) + (0)(0.2) + (1)(0.3) + (1)(0.1) + (1)(0.1) = 0.5. \end{aligned}$$

Definition: We call (X,Y) a **continuous random vector** if there exists a function $f_{X,Y}: \mathbb{R}^2 \to \mathbb{R}$ such that, for all $B \in \mathcal{B}(\mathbb{R}^2)$,

$$P((X,Y) \in B) = \int_B \int f_{X,Y}(x,y)dxdy.$$

We call $f_{X,Y}(x,y)$ a **joint probability density function** (pdf) of (X,Y). Analogous to the univariate case; i.e., as an extension of Theorem 1.6.5 (pp 36 CB), we have

(a) $f_{X,Y}(x,y) \geq 0$, for all $(x,y) \in \mathbb{R}^2$

(b) $\int\int_{\mathbb{R}^2} f_{X,Y}(x,y)dxdy = 1$.

Example 4.3. Suppose (X,Y) is a continuous random vector with joint pdf

$$f_{X,Y}(x,y) = cxy\ I(0 < y < x < 1).$$

(a) Find the constant c.
(b) Calculate $P(X - Y > \frac{1}{8})$.

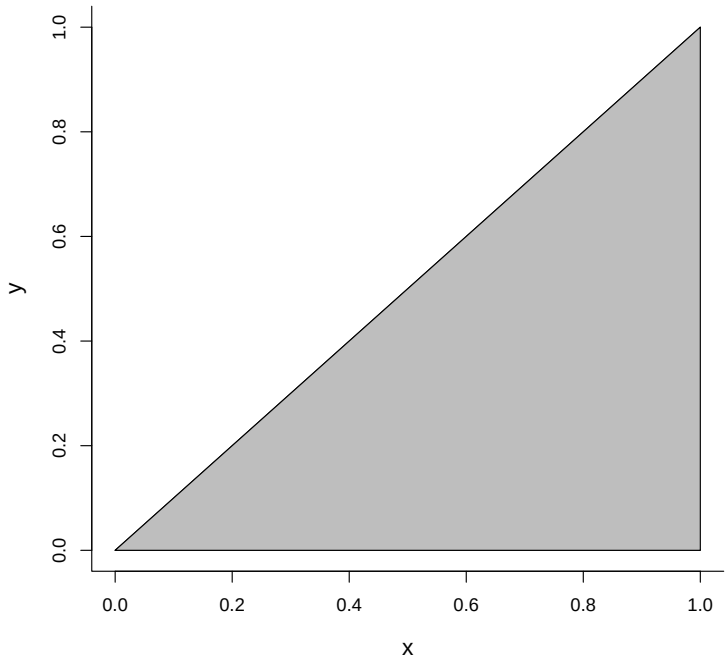


Figure 4.1: A graph of the support $\mathcal{A} = \{(x, y) \in \mathbb{R}^2 : 0 < y < x < 1\}$ in Example 4.3.

Solution. First, note that the two-dimensional support set identified in the indicator function $I(0 < y < x < 1)$ is

$$\mathcal{A} = \{(x, y) \in \mathbb{R}^2 : 0 < y < x < 1\}$$

which is depicted in Figure 4.1. The joint pdf $f_{X,Y}(x, y)$ is a three-dimensional function which is nonzero over this set (and is zero otherwise). To do part (a), we know that

$$\int\int_{\mathbb{R}^2} f_{X,Y}(x, y)dxdy = 1.$$

Therefore, the calculation

$$\int_{y=0}^{1}\int_{x=y}^{1} cxy \; dxdy = \frac{c}{8} \stackrel{\text{set}}{=} 1$$

shows that $c = 8$. The joint pdf of (X, Y) is therefore

$$f_{X,Y}(x, y) = 8xy \; I(0 < y < x < 1).$$

To calculate $P(X - Y > \frac{1}{8})$ in part (b), we simply integrate $f_{X,Y}(x, y)$ over the set

$$B = \left\{(x, y) \in \mathbb{R}^2 : 0 < y < x < 1, \; x - y > \frac{1}{8}\right\}.$$

The boundary of the set B is determined as follows:

$$x - y = \frac{1}{8} \implies y = x - \frac{1}{8}.$$

Therefore,

$$\begin{aligned} P\left(X - Y > \frac{1}{8}\right) &= \int\int_B f_{X,Y}(x,y)\, dxdy \\ &= \int_{y=0}^{\frac{7}{8}} \int_{x=y+\frac{1}{8}}^{1} 8xy\, dxdy \approx 0.698. \end{aligned}$$

This probability could also be calculated by interchanging the order of the integration (and adjusting the limits) as follows:

$$\int_{x=\frac{1}{8}}^{1} \int_{y=0}^{x-\frac{1}{8}} 8xy\, dxdy \approx 0.698.$$

Remark: Either way, we see that the limits on the double integral come directly from a well-constructed picture of the support and the region over which we are integrating. When working with joint distributions, not taking time to construct good pictures of the support and regions of integration usually (i.e., almost always) leads to the wrong answer.

Definition: Suppose (X, Y) is a continuous random vector with pdf $f_{X,Y}(x, y)$ and suppose $g : \mathbb{R}^2 \to \mathbb{R}$. Then $g(X, Y)$ is a univariate random variable and its expected value is

$$E[g(X,Y)] = \int\int_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y) dxdy.$$

This definition is analogous to the definition of mathematical expectation for univariate continuous random variables. Existence issues are also identical; i.e., we need the integral above to converge absolutely. If $\int\int_{\mathbb{R}^2} |g(x,y)| f_{X,Y}(x,y) dxdy$ is not finite, then $E[g(X,Y)]$ does not exist.

Example 4.4. Suppose (X, Y) is a continuous random vector with joint pdf

$$f_{X,Y}(x,y) = 8xy\, I(0 < y < x < 1),$$

as in Example 4.3 (see the support $\mathcal{A}$ in Figure 4.1). We have

$$\begin{aligned} E(X^2Y) &= \int\int_{\mathbb{R}^2} x^2 y\, f_{X,Y}(x,y) dxdy \\ &= \int_{y=0}^{1} \int_{x=y}^{1} x^2 y \times 8xy\, dxdy \\ &= \int_{y=0}^{1} \int_{x=y}^{1} 8x^3y^2\, dxdy = \frac{2}{7}. \end{aligned}$$

Marginal Distributions (Continuous case): Suppose (X, Y) is a continuous random vector with pdf $f_{X,Y}(x, y)$. Suppose $B \in \mathcal{B}(\mathbb{R})$. Note that

$$\begin{aligned} P(X \in B) = P(X \in B, Y \in \mathbb{R}) &= P((X,Y) \in B \times \mathbb{R}) \\ &= \int\int_{B \times \mathbb{R}} f_{X,Y}(x,y)dxdy \\ &= \int_B \underbrace{\int_{\mathbb{R}} f_{X,Y}(x,y)dy}_{= f_X(x)} dx. \end{aligned}$$

We call

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy$$

the **marginal probability density function** (pdf) of X. Similarly, we call

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx$$

the marginal pdf of Y.

Main point: To find the marginal pdf of one random variable, you take the joint pdf and integrate over the other variable.

Example 4.5. Suppose (X, Y) is a continuous random vector with joint pdf

$$f_{X,Y}(x,y) = 8xy\ I(0 < y < x < 1),$$

as in Example 4.3 (see the support $\mathcal{A}$ in Figure 4.1). The marginal pdf of X is, for $0 < x < 1$,

$$f_X(x) = \int_{y=0}^{x} 8xy\ dy = 4x^3.$$

The marginal pdf of Y is, for $0 < y < 1$,

$$f_Y(y) = \int_{x=y}^{1} 8xy\ dx = 4y(1-y^2).$$

Summarizing,

$$f_X(x) = 4x^3 I(0 < x < 1) \qquad \text{and} \qquad f_Y(y) = 4y(1-y^2)I(0 < y < 1).$$

These marginal pdfs are shown in Figure 4.2 (next page). Note that X has a beta distribution with parameters $\alpha = 4$ and $\beta = 1$; i.e., $X \sim \text{beta}(4, 1)$. The random variable Y does not have a "named" distribution but its pdf is clearly valid.

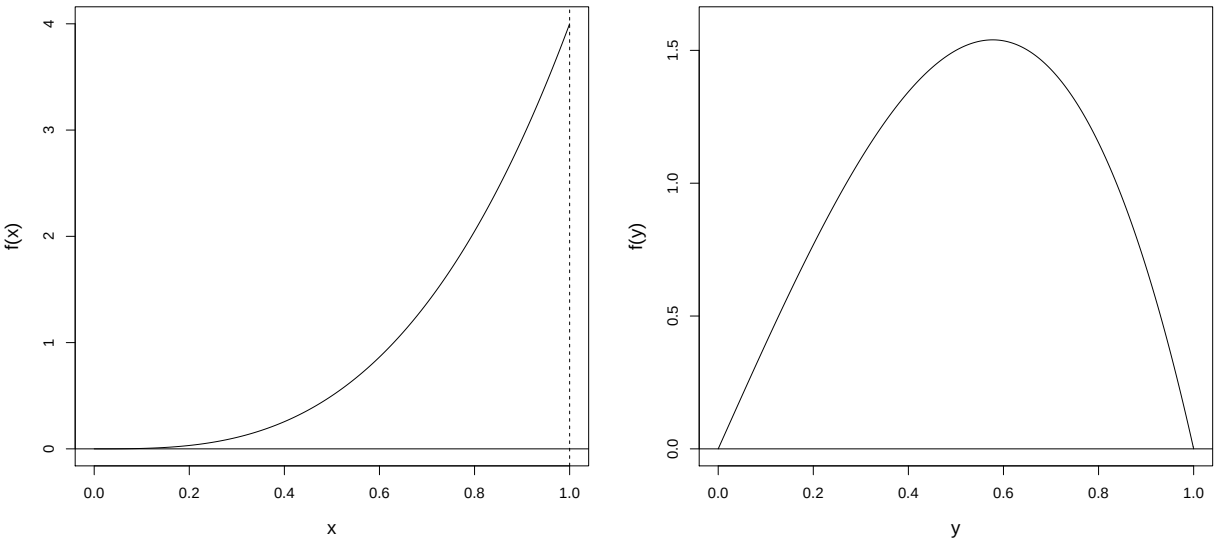


Figure 4.2: Marginal pdf of X (left) and marginal pdf of Y (right) in Example 4.5. Note that the marginal distribution of X is beta with parameters $\alpha = 4$ and $\beta = 1$; i.e., $X \sim \text{beta}(4, 1)$.

Extension: Suppose that in the last example, we wanted to calculate $P(Y > \frac{1}{2})$. We could do this in two ways:

1. Using the marginal distribution of Y,

$$P\left(Y > \frac{1}{2}\right) = \int_{y=\frac{1}{2}}^{1} f_Y(y)dy = \frac{9}{16}$$

2. Using the joint distribution of (X, Y),

$$P\left(Y > \frac{1}{2}\right) = \int_{y=\frac{1}{2}}^{1} \int_{x=y}^{1} f_{X,Y}(x, y)dxdy = \frac{9}{16}.$$

Note geometrically what we are doing in each case. In (1), we are calculating the **area** under $f_Y(y)$ over the set $B = \{y : \frac{1}{2} < y < 1\}$. In (2), we are calculating the **volume** under $f_{X,Y}(x, y)$ over the set $B = \{(x, y) : 0 < y < x < 1,\ \frac{1}{2} < y < 1\}$.

Example 4.6. Suppose (X, Y) is a continuous random vector with joint pdf

$$f_{X,Y}(x, y) = e^{-(x+y)} I(x > 0, y > 0).$$

In this problem, we find the distribution of

$$Z = g(X, Y) = \frac{X}{Y}$$

and calculate $E(Z)$. First, note that the two-dimensional support set identified in the indicator function $I(x > 0, y > 0)$ is

$$\mathcal{A} = \{(x, y) \in \mathbb{R}^2 : x > 0, y > 0\} = \mathbb{R}^+ \times \mathbb{R}^+.$$

The joint pdf $f_{X,Y}(x, y)$ is a three-dimensional function which is nonzero over this set (and is zero otherwise). Clearly, the random variable Z has positive support, say $\mathcal{Z} = \{z : z > 0\}$. We derive the cdf of Z first:

$$\begin{aligned} F_Z(z) = P(Z \leq z) &= P\left(\frac{X}{Y} \leq z\right) \\ &\overset{z>0}{=} \int_B \int f_{X,Y}(x, y)dxdy, \end{aligned}$$

where the set $B = \{(x, y) \in \mathbb{R}^2 : x > 0,\ y > 0,\ \frac{x}{y} \leq z\}$. The boundary of the set B is determined as follows:

$$\frac{x}{y} = z \implies y = \frac{x}{z}.$$

The double integral above becomes

$$\int_{x=0}^{\infty} \int_{y=x/z}^{\infty} e^{-(x+y)} dydx = \frac{z}{z+1}.$$

Therefore, the cdf of Z is

$$F_Z(z) = \begin{cases} 0, & z \leq 0 \\ \dfrac{z}{z+1}, & z > 0. \end{cases}$$

The pdf of Z is

$$\begin{aligned} f_Z(z) &= \frac{d}{dz} F_Z(z) \\ &= \frac{1}{(z+1)^2} I(z > 0) \end{aligned}$$

and is shown in Figure 4.3 (next page). Finally, note that

$$\begin{aligned} E(Z) = \int_{\mathbb{R}} z f_Z(z)dz &= \int_0^{\infty} \frac{z}{(z+1)^2}\, dz \\ &\overset{u=z+1}{=} \int_1^{\infty} \frac{u-1}{u^2}\, du \\ &= \left(\ln u + \frac{1}{u}\right)\Big|_{u=1}^{\infty} = +\infty; \end{aligned}$$

i.e., $E(Z)$ does not exist.

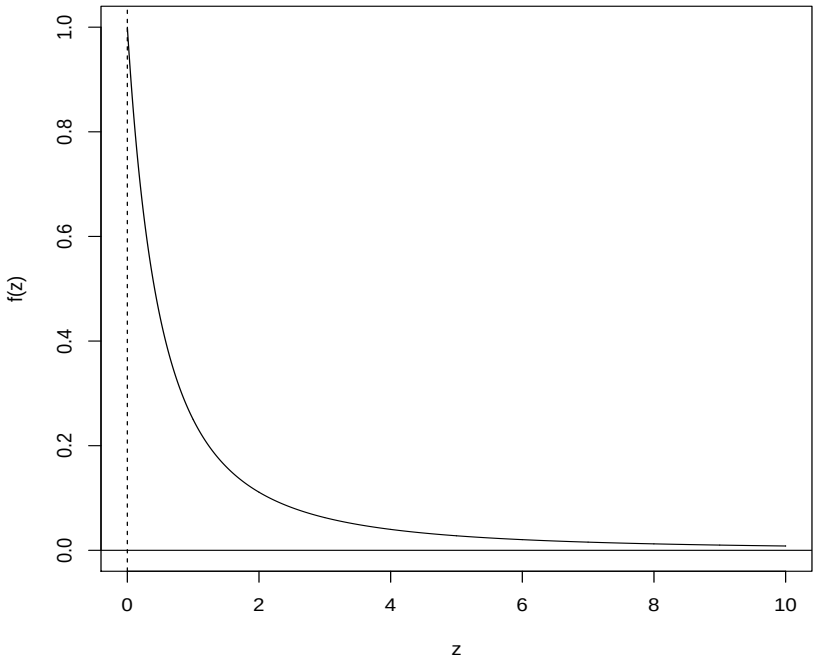


Figure 4.3: Pdf of Z in Example 4.6.

Definition: Suppose that (X, Y) is a random vector (discrete or continuous). The **joint cumulative distribution function** (cdf) of (X, Y) is

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y), \quad \text{for all } (x, y) \in \mathbb{R}^2.$$

As in the univariate case, a random vector's cdf completely determines its distribution. If (X, Y) is continuous with joint pdf $f_{X,Y}(x, y)$, then

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) dv du$$

and

$$\frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y} = f_{X,Y}(x, y).$$

These expressions summarize how $f_{X,Y}(x, y)$ and $F_{X,Y}(x, y)$ are related in bivariate settings.

Remark: The following material (on joint mgfs) is not covered in CB's §4.1 but is very useful. In addition, this material will be presented in other courses (e.g., STAT 714, etc.).

Definition: Suppose that $\mathbf{X} = (X_1, X_2)'$ is a bivariate random vector (discrete or continuous). The **joint moment generating function** (mgf) of X_1 and X_2 is

$$M_{\mathbf{X}}(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{X}}) = E(e^{t_1 X_1 + t_2 X_2}),$$

where $\mathbf{t} = (t_1, t_2)'$. For $M_{\mathbf{X}}(\mathbf{t})$ to exist, this expectation must be finite in an open neighborhood about $\mathbf{t} = \mathbf{0}$; i.e., $E(e^{t_1X_1+t_2X_2}) < \infty \; \forall t_1 \in (-h_1, h_1) \; \forall t_2 \in (-h_2, h_2) \; \exists h_1 > 0 \; \exists h_2 > 0$.

Notes:

1. We may also write
$$M_{\mathbf{X}}(\mathbf{t}) = M_{X_1,X_2}(t_1, t_2).$$

2. As with mgfs for univariate random variables, a random vector's mgf $M_{\mathbf{X}}(\mathbf{t})$ uniquely identifies the distribution of $\mathbf{X}$.

3. It is easy to see that
$$\begin{aligned} M_{X_1}(t_1) &= M_{X_1,X_2}(t_1, 0) \\ M_{X_2}(t_2) &= M_{X_1,X_2}(0, t_2). \end{aligned}$$
Therefore, the marginal mgfs are easily obtained from the joint mgf.

Example 4.7. Suppose $\mathbf{X} = (X_1, X_2)'$ is a continuous random vector with joint pdf
$$f_{X_1,X_2}(x_1, x_2) = e^{-x_2} I(0 < x_1 < x_2 < \infty).$$
The joint mgf of X_1 and X_2 is
$$\begin{aligned} M_{X_1,X_2}(t_1, t_2) &= E(e^{t_1X_1+t_2X_2}) \\ &= \int\int_{\mathbb{R}^2} e^{t_1x_1+t_2x_2} f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 \\ &= \int_{x_2=0}^{\infty} \int_{x_1=0}^{x_2} e^{t_1x_1+t_2x_2} e^{-x_2} \; dx_1 dx_2 \\ &= \frac{1}{(1-t_1-t_2)(1-t_2)}, \end{aligned}$$
provided that $t_1 + t_2 < 1$ and $t_2 < 1$. Therefore, the marginal mgf of X_1 is
$$M_{X_1}(t_1) = M_{X_1,X_2}(t_1, 0) = \frac{1}{1-t_1}, \quad \text{for } t_1 < 1,$$
and the marginal mgf of X_2 is
$$M_{X_2}(t_2) = M_{X_1,X_2}(0, t_2) = \left(\frac{1}{1-t_2}\right)^2, \quad \text{for } t_2 < 1.$$
Because mgfs are unique, we see that
$$\begin{aligned} X_1 &\sim \text{exponential}(1) \\ X_2 &\sim \text{gamma}(2, 1). \end{aligned}$$

Definition: Suppose that $\mathbf{X} = (X_1, X_2)'$ is a random vector. The **expected value** of $\mathbf{X}$ is

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix}_{2\times 1},$$

provided that both $E(X_1)$ and $E(X_2)$ exist. In the last example, we see that

$$E(\mathbf{X}) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

It is also possible to calculate $E(\mathbf{X})$ using the joint mgf:

$$E(\mathbf{X}) = \left.\frac{\partial M_{\mathbf{X}}(\mathbf{t})}{\partial \mathbf{t}}\right|_{\mathbf{t}=\mathbf{0}} = \left.\begin{pmatrix} \frac{\partial M_{\mathbf{X}}(\mathbf{t})}{\partial t_1} \\ \frac{\partial M_{\mathbf{X}}(\mathbf{t})}{\partial t_2} \end{pmatrix}\right|_{t_1=t_2=0};$$

i.e., $E(\mathbf{X})$ is the gradient of $M_{\mathbf{X}}(\mathbf{t})$ evaluated at $\mathbf{t} = \mathbf{0}$. Verify this with Example 4.7.

4.2 Conditional Distributions and Independence

Conditional Distributions (Discrete case): Suppose (X, Y) is a discrete random vector with pmf $f_{X,Y}(x, y)$. The **conditional probability mass function** (pmf) of Y given $X = x$ is

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

which is defined for values of x where $f_X(x) > 0$. In the discrete case, this definition follows directly from the definition of conditional probability; i.e.,

$$\begin{aligned} f_{Y|X}(y|x) &\equiv P(Y = y|X = x) \\ &= \frac{P(X = x, Y = y)}{P(X = x)} \\ &= \frac{f_{X,Y}(x, y)}{f_X(x)}. \end{aligned}$$

Interpretation: The function $f_{Y|X}(y|x)$ is a univariate pmf; it describes the distribution of Y (i.e., how Y varies) when X is **fixed** at the value x. Similarly, the conditional pmf of X given $Y = y$ is

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

which is defined for values of y where $f_Y(y) > 0$. This function is a univariate pmf and describes the distribution of X (i.e., how X varies) when Y is **fixed** at the value y.

Example 4.8. We revisit the joint pmf of (X, Y) in Example 4.2:

		y		
		0	**1**	**2**
x	**0**	0.1	0.2	0.2
	1	0.3	0.1	0.1

The conditional pmf of Y, when $X = x = 0$, is found as follows:

$$\begin{aligned} f_{Y|X}(0|0) &= \frac{f_{X,Y}(0,0)}{f_X(0)} = \frac{0.1}{0.5} = 0.2 \qquad (\text{for } y = 0) \\ f_{Y|X}(1|0) &= \frac{f_{X,Y}(0,1)}{f_X(0)} = \frac{0.2}{0.5} = 0.4 \qquad (\text{for } y = 1) \\ f_{Y|X}(2|0) &= \frac{f_{X,Y}(0,2)}{f_X(0)} = \frac{0.2}{0.5} = 0.4 \qquad (\text{for } y = 2) \end{aligned}$$

Therefore,

$$f_{Y|X}(y|0) = 0.2I(y = 0) + 0.4I(y = 1) + 0.4I(y = 2).$$

Note: Suppose $B \in \mathcal{B}(\mathbb{R})$. Conditional probabilities can be calculated using conditional pmfs as follows:

$$\begin{aligned} P(Y \in B|X = x) &= \sum_{y \in B} f_{Y|X}(y|x) \\ P(X \in B|Y = y) &= \sum_{x \in B} f_{X|Y}(x|y). \end{aligned}$$

For example, in Example 4.8,

$$P(Y \leq 1|X = 0) = \sum_{y=0}^{1} f_{Y|X}(y|0) = 0.2 + 0.4 = 0.6.$$

Conditional Distributions (Continuous case): Suppose (X, Y) is a continuous random vector with pdf $f_{X,Y}(x, y)$. The **conditional probability density function** (pdf) of Y given $X = x$ is

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

which is defined for values of x where $f_X(x) > 0$.

Interpretation: The function $f_{Y|X}(y|x)$ is a univariate pdf; it describes the distribution of Y (i.e., how Y varies) when X is **fixed** at the value x. Similarly, the conditional pdf of X given $Y = y$ is

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

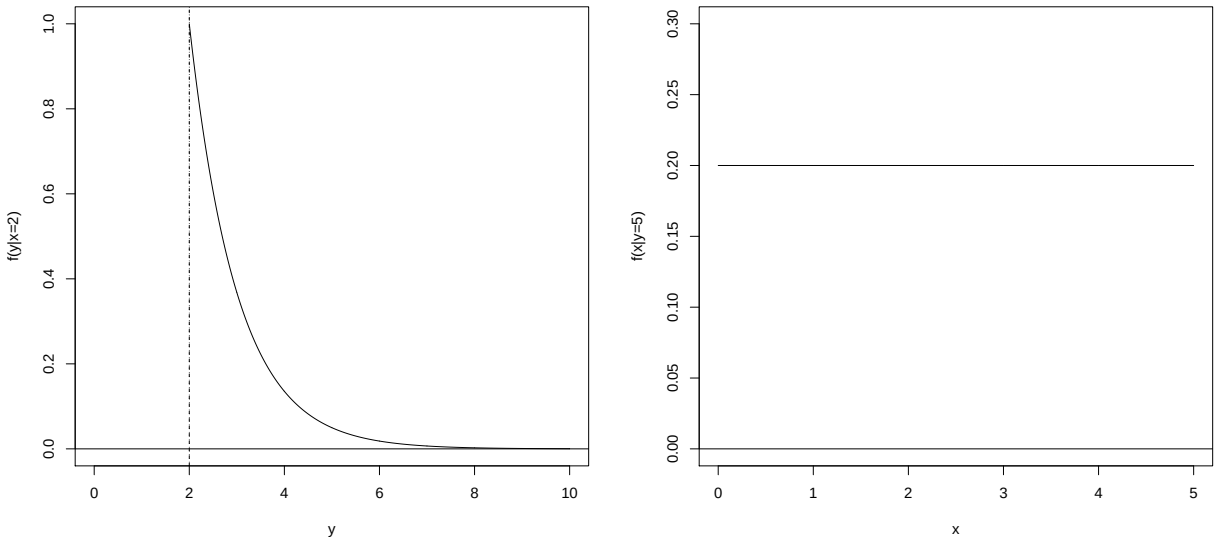


Figure 4.4: Example 4.9. Left: Conditional pdf of Y when $x = 2$. Right: Conditional pdf of X when $y = 5$. Note that the vertical axes are different in the two figures.

which is defined for values of y where $f_Y(y) > 0$. This function is a univariate pdf and describes the distribution of X (i.e., how X varies) when Y is **fixed** at the value y.

Example 4.9. In Example 4.7, we worked with the joint pdf

$$f_{X,Y}(x, y) = e^{-y} I(0 < x < y < \infty).$$

Recall that we showed (using mgfs) that $X \sim$ exponential(1) and $Y \sim$ gamma$(2, 1)$ so that the marginal pdfs are

$$\begin{aligned} f_X(x) &= e^{-x} I(x > 0) \\ f_Y(y) &= y e^{-y} I(y > 0). \end{aligned}$$

The conditional pdf of Y given $X = x$ is therefore

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x, y)}{f_X(x)} \\ &= \frac{e^{-y} I(0 < x < y < \infty)}{e^{-x} I(x > 0)} = e^{-(y-x)} I(y > x). \end{aligned}$$

This function describes the distribution of Y when X is fixed at $x > 0$. In Figure 4.4 (left), we display this conditional density when $x = 2$; i.e.,

$$f_{Y|X}(y|x = 2) = e^{-(y-2)} I(y > 2).$$

The conditional pdf of X given $Y = y$ is

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{e^{-y}I(0<x<y<\infty)}{ye^{-y}I(y>0)} = \frac{1}{y}I(0<x<y). \end{aligned}$$

This function describes the distribution of X when Y is fixed at $y > 0$. In Figure 4.4 (right), we display this conditional density when $y = 5$; i.e.,

$$f_{X|Y}(x|y=5) = \frac{1}{5}I(0<x<5).$$

Remark: Note that $Y|\{X = x\}$ has a shifted exponential distribution with location "parameter" x. Similarly, note that $X|\{Y = y\} \sim \mathcal{U}(0, y)$.

Note: Suppose $B \in \mathcal{B}(\mathbb{R})$. Conditional probabilities can be calculated using conditional pdfs as follows:

$$\begin{aligned} P(Y \in B|X = x) &= \int_B f_{Y|X}(y|x)dy \\ P(X \in B|Y = y) &= \int_B f_{X|Y}(x|y)dx. \end{aligned}$$

For example, in Example 4.9,

$$P(Y < 5|X = 2) = \int_{y=2}^{5} e^{-(y-2)}dy = 1 - e^{-3}$$

and

$$P(X > 3|Y = 5) = \int_{x=3}^{5} \frac{1}{5}\, dx = \frac{2}{5}.$$

Note: We now formally define conditional expectation (e.g., conditional means, conditional variances, and conditional mgfs).

Definition: Suppose (X, Y) is a continuous random vector. We define **conditional expectations** as follows:

$$\begin{aligned} E[g(Y)|X = x] &= \int_{\mathbb{R}} g(y)f_{Y|X}(y|x)dy \\ E[h(X)|Y = y] &= \int_{\mathbb{R}} h(x)f_{X|Y}(x|y)dx. \end{aligned}$$

Notes:

1. If (X, Y) is discrete, then integrals above are replaced by sums.
2. The same existence issues still remain; for example, for $E[g(Y)|X = x]$ to exist, we need $\int_{\mathbb{R}} |g(y)|f_{Y|X}(y|x)dy < \infty$.

Special case: If $g(Y) = Y$ and $h(X) = X$, then

$$\begin{aligned} E(Y|X=x) &= \int_{\mathbb{R}} y f_{Y|X}(y|x)dy \\ E(X|Y=y) &= \int_{\mathbb{R}} x f_{X|Y}(x|y)dx. \end{aligned}$$

These are called **conditional means**.

Important: Conditional expectations are always functions of the variable on which you are conditioning. Furthermore, the use of notation for the conditioning variable is important in describing whether a conditional expectation is a fixed quantity or a random variable.

$$\begin{aligned} E(Y|X=x) &\longleftarrow \text{function of } x\text{; fixed} \\ E(Y|X) &\longleftarrow \text{function of } X\text{; random variable} \end{aligned}$$

$$\begin{aligned} E(X|Y=y) &\longleftarrow \text{function of } y\text{; fixed} \\ E(X|Y) &\longleftarrow \text{function of } Y\text{; random variable} \end{aligned}$$

Example 4.10. In Example 4.9, we worked with the joint pdf

$$f_{X,Y}(x,y) = e^{-y} I(0 < x < y < \infty)$$

and derived the conditional pdfs to be

$$\begin{aligned} f_{Y|X}(y|x) &= e^{-(y-x)} I(y > x) \\ f_{X|Y}(x|y) &= \frac{1}{y} I(0 < x < y). \end{aligned}$$

The conditional mean of Y given $X = x$ is

$$\begin{aligned} E(Y|X=x) &= \int_{\mathbb{R}} y e^{-(y-x)} I(y > x) dy \\ &= \int_{y=x}^{\infty} y e^{-(y-x)} dy \\ &\overset{u=y-x}{=} \int_0^{\infty} (u+x) e^{-u} du = E(U+x), \end{aligned}$$

where $U \sim$ exponential(1); in the last integral, note that $e^{-u} I(u > 0)$ is the pdf of $U \sim$ exponential(1). Therefore,

$$E(Y|X=x) = E(U+x) = E(U) + x = 1 + x.$$

The conditional mean of X given $Y = y$ is

$$E(X|Y=y) = \int_{\mathbb{R}} x \, \frac{1}{y} I(0 < x < y) dx = \frac{1}{y} \int_{x=0}^{y} x \, dx = \frac{1}{y} \left(\frac{x^2}{2} \Big|_{x=0}^{y} \right) = \frac{y}{2}.$$

This should not be surprising because $X|\{Y=y\} \sim \mathcal{U}(0,y)$.

Remark: We have just calculated $E(Y|X=x) = 1+x$ and $E(X|Y=y) = y/2$. These are fixed. The versions $E(Y|X) = 1+X$ and $E(X|Y) = Y/2$ are random. Because $E(Y|X)$ and $E(X|Y)$ are random variables, it makes sense to think about their distributions, their means, their variances, their moment generating functions, etc.

Definition: Suppose (X,Y) is a continuous random vector. For notational purposes, let

$$\begin{aligned} E(Y|X=x) &= \mu_{Y|X=x} \\ E(X|Y=y) &= \mu_{X|Y=y} \end{aligned}$$

denote the conditional means (viewed as fixed quantities; not random). The **conditional variance** of Y given $X=x$ is

$$\text{var}(Y|X=x) = E[(Y-\mu_{Y|X=x})^2|X=x] = \int_{\mathbb{R}} (y-\mu_{Y|X=x})^2 f_{Y|X}(y|x)dy.$$

Similarly, the conditional variance of X given $Y=y$ is

$$\text{var}(X|Y=y) = E[(X-\mu_{X|Y=y})^2|Y=y] = \int_{\mathbb{R}} (x-\mu_{X|Y=y})^2 f_{X|Y}(x|y)dx.$$

Note that $\text{var}(Y|X=x)$ is a function of x and $\text{var}(X|Y=y)$ is a function of y. If (X,Y) is discrete, then integrals are replaced by sums.

Computing Formulas (Conditional versions): Computing formulas for conditional variances are analogous to the unconditional versions:

$$\begin{aligned} \text{var}(Y|X=x) &= E(Y^2|X=x) - [E(Y|X=x)]^2 \\ \text{var}(X|Y=y) &= E(X^2|Y=y) - [E(X|Y=y)]^2. \end{aligned}$$

Exercise: With the conditional distributions in Example 4.10, show that $\text{var}(Y|X=x) = 1$ and $\text{var}(X|Y=y) = y^2/12$.

Remark: The following material (on conditional mgfs) is not covered in CB's §4.2 but is very useful. This material will be presented in other courses (e.g., STAT 714, etc.).

Definition: Suppose (X,Y) is a continuous random vector. The **conditional moment generating function** (mgf) of Y given $X=x$ is

$$M_{Y|X}(t) = E(e^{tY}|X=x) = \int_{\mathbb{R}} e^{ty} f_{Y|X}(y|x)dy.$$

Similarly, the conditional mgf of X given $Y=y$ is

$$M_{X|Y}(t) = E(e^{tX}|Y=y) = \int_{\mathbb{R}} e^{tx} f_{X|Y}(x|y)dx.$$

Notes:

1. If (X,Y) is discrete, then integrals above are replaced by sums.

2. As with unconditional mgfs, we need to require that the corresponding integrals (sums) above are finite for $t \in (-h, h)$ $\exists h > 0$. Otherwise, the mgfs do not exist.

3. Conditional mgfs enjoy all of the same properties that unconditional mgfs do (e.g., uniqueness, useful in generating moments−now, *conditional* moments).

Example 4.11. Suppose that (X, Y) is a continuous random vector with joint pdf

$$f_{X,Y}(x,y) = \frac{e^{-x/y}e^{-y}}{y} I(x > 0, y > 0).$$

In this example, we find the conditional mgf $M_{X|Y}(t)$. To do this, we need to find $f_{X|Y}(x|y)$, the conditional pdf of X given $Y = y$. Recall that

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

so we need to first find $f_Y(y)$, the marginal pdf of Y. For $y > 0$,

$$f_Y(y) = \int_{\mathbb{R}} \frac{e^{-x/y}e^{-y}}{y} I(x > 0, y > 0)dx = e^{-y} \underbrace{\int_{x=0}^{\infty} \frac{1}{y} e^{-x/y} dx}_{=1} = e^{-y}.$$

Therefore, the conditional pdf of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{\frac{e^{-x/y}e^{-y}}{y} I(x > 0, y > 0)}{e^{-y} I(y > 0)} = \frac{1}{y} e^{-x/y} I(x > 0).$$

That is, $X|\{Y = y\} \sim$ exponential(y). Finally, the conditional mgf of X given $Y = y$ is

$$\begin{aligned} M_{X|Y}(t) = E(e^{tX}|Y = y) &= \int_{\mathbb{R}} e^{tx} \frac{1}{y} e^{-x/y} I(x > 0) dx \\ &= \frac{1}{1 - yt}, \qquad \text{for } t < \frac{1}{y}. \end{aligned}$$

Now, let's illustrate how to use $M_{X|Y}(t)$ to "generate" conditional moments:

$$\begin{aligned} \frac{\partial}{\partial t} M_{X|Y}(t) = y(1 - yt)^{-2} &\implies E(X|Y = y) = \frac{\partial}{\partial t} M_{X|Y}(t)\Big|_{t=0} = y \\ \frac{\partial^2}{\partial t^2} M_{X|Y}(t) = 2y^2(1 - yt)^{-3} &\implies E(X^2|Y = y) = \frac{\partial^2}{\partial t^2} M_{X|Y}(t)\Big|_{t=0} = 2y^2. \end{aligned}$$

The conditional variance is therefore

$$\text{var}(X|Y = y) = E(X^2|Y = y) - [E(X|Y = y)]^2 = 2y^2 - y^2 = y^2.$$

These results are expected because $X|\{Y = y\} \sim$ exponential(y).

Definition: Let (X, Y) be a random vector (discrete or continuous) with joint pmf/pdf $f_{X,Y}(x, y)$. We say that X and Y are **independent** if

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

for all $x, y \in \mathbb{R}$. In other words, the joint pmf/pdf equals the product of the marginal pmfs/pdfs. The shorthand notation "$X \perp\!\!\!\perp Y$" means "X and Y are independent."

Observation: Suppose that $X \perp\!\!\!\perp Y$. The conditional pmf/pdf of Y given $X = x$ is

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

Therefore, if $X \perp\!\!\!\perp Y$, then for any $B \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned} P(Y \in B|X = x) &= \int_B f_{Y|X}(y|x)dy \\ &= \int_B f_Y(y)dy = P(Y \in B). \end{aligned}$$

In other words, knowledge that $X = x$ does not influence how we assign probability to the event $\{Y \in B\}$. Similarly, if $X \perp\!\!\!\perp Y$, then

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

Lemma 4.2.7. Suppose (X, Y) is a random vector with joint pmf/pdf $f_{X,Y}(x, y)$. The random variables X and Y are independent if and only if there exists functions $g(x)$ and $h(y)$ such that

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

for all $x, y \in \mathbb{R}$.

Remarks:

1. The usefulness of Lemma 4.2.7 is that the functions $g(x)$ and $h(y)$ can be any functions of x and y, respectively; they need not be valid pmfs/pdfs.

2. The factorization in Lemma 4.2.7 must hold for all $x, y \in \mathbb{R}$. This means that if $\mathcal{A}$, the support of (X, Y), involves a "constraint," then X and Y cannot be independent.

   - By "constraint," I mean something like this:

     $$\mathcal{A} = \{(x, y) \in \mathbb{R}^2 : 0 < x < y < \infty\}.$$

     Note that the corresponding indicator function $I(0 < x < y < \infty)$ cannot be absorbed into $g(x)$ or $h(y)$.

- Therefore, for Lemma 4.2.7 to be applicable, the support set $\mathcal{A}$ must be a Cartesian product of two sets, one that depends only on x and the other that depends only on y. For example,

$$\mathcal{A} = \{(x,y) \in \mathbb{R}^2 : 0 < x < 1,\ 0 < y < 1\} = \{x \in \mathbb{R} : x > 0\} \times \{y \in \mathbb{R} : y > 0\}.$$

  The corresponding indicator function $I(0 < x < 1, 0 < y < 1)$ in this case can be written as $I(0 < x < 1)I(0 < y < 1)$.

Proof of Lemma 4.2.7: Proving the necessity ($\Longrightarrow$) is straightforward. Suppose that $X \perp\!\!\!\perp Y$, and take $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. Because

$$f_{X,Y}(x,y) \stackrel{X \perp\!\!\!\perp Y}{=} f_X(x)f_Y(y) = g(x)h(y),$$

we have shown that there do exist functions $g(x)$ and $h(y)$ satisfying $f_{X,Y}(x,y) = g(x)h(y)$. Proving the sufficiency ($\Longleftarrow$) is done as follows. Suppose that the factorization holds; i.e., suppose that $f_{X,Y}(x,y) = g(x)h(y)$, for all $x, y \in \mathbb{R}$, for some functions $g(x)$ and $h(y)$. For illustration, suppose that (X,Y) is continuous. Let

$$\int_{\mathbb{R}} g(x)dx = c \qquad \text{and} \qquad \int_{\mathbb{R}} h(y)dy = d.$$

Note that

$$cd = \int_{\mathbb{R}} g(x)dx \int_{\mathbb{R}} h(y)dy = \int_{\mathbb{R}}\int_{\mathbb{R}} g(x)h(y)dxdy = \int_{\mathbb{R}^2}\int f_{X,Y}(x,y)dxdy = 1,$$

because the factorization $f_{X,Y}(x,y) = g(x)h(y)$ holds by assumption. Furthermore,

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy \ = \ \int_{\mathbb{R}} g(x)h(y)dy = dg(x).$$

An analogous argument shows that $f_Y(y) = ch(y)$. Therefore, for all $x, y \in \mathbb{R}$, we have

$$\begin{aligned} f_{X,Y}(x,y) &= g(x)h(y) \\ &= dg(x)ch(y) = f_X(x)f_Y(y), \end{aligned}$$

showing that $X \perp\!\!\!\perp Y$. For the discrete case, simply replace integrals with sums. □

Example 4.12. Suppose that (X,Y) is a continuous random vector with joint pdf

$$f_{X,Y}(x,y) = \frac{1}{384}x^2y^4e^{-y-x/2}I(x > 0, y > 0).$$

For all $x, y \in \mathbb{R}$, note that we can write

$$f_{X,Y}(x,y) = \underbrace{\frac{1}{384}x^2e^{-x/2}I(x > 0)}_{=\ g(x)} \ \times \ \underbrace{y^4e^{-y}I(y > 0)}_{=\ h(y)}.$$

By Lemma 4.2.7, we have that $X \perp\!\!\!\perp Y$.

Theorem 4.2.10. Suppose that X and Y are independent random variables.

(a) For all $A, B \in \mathcal{B}(\mathbb{R})$,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B),$$

that is, $\{X \in A\}$ and $\{Y \in B\}$ are independent events.

(b) If $g(x)$ is a function of x only and $h(y)$ is a function of y only, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)],$$

provided that all expectations exist.

Proof. We prove part (b) first because part (a) is a special case of part (b). Suppose (X, Y) is continuous. By definition,

$$\begin{aligned} E[g(X)h(Y)] &= \int\int_{\mathbb{R}^2} g(x)h(y)f_{X,Y}(x,y)dxdy \\ &\overset{X \perp\!\!\!\perp Y}{=} \int\int_{\mathbb{R}^2} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{\mathbb{R}} g(x)f_X(x)dx \int_{\mathbb{R}} h(y)f_Y(y)dy = E[g(X)]E[h(Y)]. \end{aligned}$$

If (X, Y) is discrete, simply replace integrals with sums. To prove part (a), suppose $A, B \in \mathcal{B}(\mathbb{R})$ and define

$$\begin{aligned} g(X) &= I(X \in A) \\ h(Y) &= I(Y \in B). \end{aligned}$$

Because the expectation of an indicator function is the probability of the set that it indicates (see next remark), we have

$$\begin{aligned} E[g(X)h(Y)] &= E[I(X \in A)I(Y \in B)] \\ &= E[I(X \in A, Y \in B)] \\ &= P(X \in A, Y \in B) \end{aligned}$$

and

$$\begin{aligned} E[g(X)]E[h(Y)] &= E[I(X \in A)]E[I(Y \in B)] \\ &= P(X \in A)P(Y \in B). \end{aligned}$$

Because A and B are arbitrary, the result follows. □

Remark: To see why expectations of indicator functions are probabilities, suppose that X is a random variable on $(S, \mathcal{B}, P)$ where, for all $\omega \in S$,

$$X(\omega) = I_A(\omega) \equiv \begin{cases} 1, & X(\omega) \in A \\ 0, & X(\omega) \notin A, \end{cases}$$

for $A \in \mathcal{B}(\mathbb{R})$. That is, X is a binary random variable and

$$E(X) = 1P_X(X \in A) + 0P_X(X \notin A) = P_X(X \in A).$$

Abusing notation, this is written simply as $P(A)$.

Theorem 4.2.12. Suppose that X and Y are independent random variables with marginal mgfs $M_X(t)$ and $M_Y(t)$, respectively. The mgf of $Z = X + Y$ is

$$M_Z(t) = M_X(t)M_Y(t).$$

That is, the mgf of the sum of two independent random variables is the product of the marginal mgfs.

Proof. The mgf of Z is

$$\begin{aligned} M_Z(t) = E(e^{tZ}) &= E\left[e^{t(X+Y)}\right] \\ &= E(e^{tX}e^{tY}) \\ &\overset{X \perp\!\!\!\perp Y}{=} E(e^{tX})E(e^{tY}) \\ &= M_X(t)M_Y(t). \ \square \end{aligned}$$

Remark: Theorem 4.2.12 is extremely useful. If $X \perp\!\!\!\perp Y$, then we can easily determine the distribution of the sum $Z = X + Y$ just by examining the mgf of Z.

Example 4.13. Suppose that $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$, and $X \perp\!\!\!\perp Y$. The mgf of $Z = X + Y$ is

$$\begin{aligned} M_Z(t) &= M_X(t)M_Y(t) \\ &= e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)}, \end{aligned}$$

which we recognize as the mgf of a Poisson distribution with mean $\lambda_1 + \lambda_2$. Because mgfs are unique, $Z = X + Y \sim \text{Poisson}(\lambda_1 + \lambda_2)$.

Remark: The following distributional results can also be established using the same argument as in Example 4.13. In each case, $X \perp\!\!\!\perp Y$.

1. $X \sim b(n_1, p)$, $Y \sim b(n_2, p) \implies Z = X + Y \sim b(n_1 + n_2, p)$
2. $X \sim \text{nib}(r_1, p)$, $Y \sim \text{nib}(r_2, p) \implies Z = X + Y \sim \text{nib}(r_1 + r_2, p)$
3. $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2) \implies Z = X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
4. $X \sim \text{gamma}(\alpha_1, \beta)$, $Y \sim \text{gamma}(\alpha_2, \beta) \implies Z = X + Y \sim \text{gamma}(\alpha_1 + \alpha_2, \beta)$

Note: We finish this section with an example, three results, and a remark.

Example 4.14. Suppose that $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$, and $X_1 \perp\!\!\!\perp X_2$. Find the conditional distribution of X_1 given $Z = X_1 + X_2 = n$, for $n \geq 1$.
Solution. The conditional pmf of X_1 given $Z = n$ is, for $x_1 = 0, 1, 2, ..., n$,

$$\begin{aligned}
f_{X_1|Z}(x_1|n) = \frac{f_{X_1,Z}(x_1,n)}{f_Z(n)} &= \frac{P(X_1 = x_1, Z = n)}{P(Z = n)} \\
&= \frac{P(X_1 = x_1, X_2 = n - x_1)}{P(Z = n)} \\
&\overset{X_1 \perp\!\!\!\perp X_2}{=} \frac{P(X_1 = x_1)P(X_2 = n - x_1)}{P(Z = n)} \\
&= \frac{\frac{\lambda_1^{x_1} e^{-\lambda_1}}{x_1!} \frac{\lambda_2^{n-x_1} e^{-\lambda_2}}{(n-x_1)!}}{\frac{(\lambda_1+\lambda_2)^n e^{-(\lambda_1+\lambda_2)}}{n!}} \\
&= \frac{n!}{x_1!(n-x_1)!} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{x_1} \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-x_1} \\
&= \binom{n}{x_1} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{x_1} \left(1 - \frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{n-x_1}.
\end{aligned}$$

That is, $X_1|\{X_1 + X_2 = n\} \sim b(n, p)$, where $p = \lambda_1/(\lambda_1 + \lambda_2)$.

Result: Suppose that (X, Y) is a random vector (discrete or continuous) with joint cdf $F_{X,Y}(x, y)$. Then X and Y are independent if and only if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y),$$

for all $x, y \in \mathbb{R}$, where $F_X(x)$ and $F_Y(y)$ are the marginal cdfs of X and Y, respectively.
Proof. Exercise.

Result: Suppose that (X, Y) is a random vector (discrete or continuous) with joint mgf $M_{X,Y}(t_1, t_2)$. Then X and Y are independent if and only if

$$M_{X,Y}(t_1, t_2) = M_X(t_1)M_Y(t_2),$$

for all values of $t_1, t_2 \in \mathbb{R}$ where these mgfs exist.
Proof. Exercise.

Result: If X and Y are independent then so are $U = g(X)$ and $V = h(Y)$. That is, functions of independent random variables are also independent.
Proof. We will prove this in the next section (Theorem 4.3.5).

Remark: The first result above (dealing with cdfs) might be a better characterization of independence than what we stated initially using pmfs/pdfs; i.e., that X and Y are independent if and only if

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

for all $x, y \in \mathbb{R}$. The reason for this is that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ may not hold on a set $A \subset \mathbb{R}^2$ where $P((X,Y) \in A) = 0$, yet, X and Y remain independent (see pp 156 CB). In this light, it might be better to say that X and Y are independent if and only

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

for "almost all" $x, y \in \mathbb{R}$, acknowledging that this may not be true on a set of measure zero.

4.3 Bivariate Transformations

Setting: Suppose (X,Y) is a random vector with joint pmf/pdf $f_{X,Y}(x,y)$ and support $\mathcal{A} \subseteq \mathbb{R}^2$. Define the random variables

$$\begin{aligned} U &= g_1(X,Y) \\ V &= g_2(X,Y), \end{aligned}$$

where $g_i : \mathbb{R}^2 \to \mathbb{R}$, for $i = 1, 2$. We would like to find the joint pmf/pdf of the random vector (U,V).

Note: From first principles, there is nothing to prevent us from deriving the cdf of (U,V). In the continuous case,

$$\begin{aligned} F_{U,V}(u,v) &= P(U \leq u, V \leq v) \\ &= P(g_1(X,Y) \leq u,\ g_2(X,Y) \leq v) \\ &= \int_B \int f_{X,Y}(x,y)dxdy, \end{aligned}$$

where the set $B = \{(x,y) \in \mathcal{A} : g_1(x,y) \leq u,\ g_2(x,y) \leq v\}$. With this, one could calculate the joint pdf by

$$f_{U,V}(u,v) = \frac{\partial^2 F_{U,V}(u,v)}{\partial u \partial v}.$$

Discrete case: If (X,Y) is discrete, we can calculate the joint pmf of (U,V) directly. By definition,

$$\begin{aligned} f_{U,V}(u,v) &= P(U = u, V = v) \\ &= P(g_1(X,Y) = u,\ g_2(X,Y) = v) \\ &= \sum_{(x,y) \in \mathcal{A}_{uv}} \sum f_{X,Y}(x,y), \end{aligned}$$

where the set $\mathcal{A}_{uv} = \{(x,y) \in \mathcal{A} : g_1(x,y) = u,\ g_2(x,y) = v\}$.

Notation: To match the notation of CB, we denote by

$$\begin{aligned} \mathcal{A} &= \text{support of } (X,Y) \\ \mathcal{B} &= \text{support of } (U,V). \end{aligned}$$

Note that

$$\mathcal{B} = \{(u,v) \in \mathbb{R}^2 : u = g_1(x,y),\ v = g_2(x,y),\ \text{for } (x,y) \in \mathcal{A}\}.$$

The vector-valued function $g : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying

$$\begin{pmatrix} U \\ V \end{pmatrix} = g\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} g_1(X,Y) \\ g_2(X,Y) \end{pmatrix}$$

is a mapping from $\mathcal{A}$ to $\mathcal{B}$; i.e., $g : \mathcal{A} \to \mathcal{B}$.

Example 4.15. Suppose that $X \sim b(n_1, p)$, $Y \sim b(n_2, p)$, and $X \perp\!\!\!\perp Y$. Define

$$\begin{aligned} U &= g_1(X,Y) = X + Y \\ V &= g_2(X,Y) = Y. \end{aligned}$$

(a) Find $f_{U,V}(u,v)$, the joint pmf of (U,V), using a bivariate transformation.
(b) Find $f_U(u)$, the marginal pmf of U.
Solution. The joint pmf of (X,Y) is

$$\begin{aligned} f_{X,Y}(x,y) &\overset{X \perp\!\!\!\perp Y}{=} f_X(x) f_Y(y) \\ &= \binom{n_1}{x} p^x (1-p)^{n_1 - x} \binom{n_2}{y} p^y (1-p)^{n_2 - y}, \end{aligned}$$

for values $(x,y) \in \mathcal{A}$, where

$$\mathcal{A} = \{(x,y) : x = 0, 1, 2, ..., n_1;\ y = 0, 1, 2, ..., n_2\}.$$

The support of (U,V) is

$$\begin{aligned} \mathcal{B} &= \{(u,v) \in \mathbb{R}^2 : u = x + y,\ v = y,\ \text{for } (x,y) \in \mathcal{A}\} \\ &= \{(u,v) \in \mathbb{R}^2 : u = 0, 1, 2, ..., n_1 + n_2,\ v = 0, 1, 2, ..., n_2,\ v \leq u\}; \end{aligned}$$

note that necessarily $v \leq u$ because $x \geq 0$. Now, the joint pmf of (U,V) equals

$$f_{U,V}(u,v) = \sum_{(x,y) \in \mathcal{A}_{uv}} \sum f_{X,Y}(x,y),$$

where the set $\mathcal{A}_{uv} = \{(x,y) \in \mathcal{A} : x + y = u,\ y = v\}$. In this case, the set $\mathcal{A}_{uv}$ consists of just one point, the singleton $\{(u - v, v)\}$. To see why this is true, note that the system

$$\begin{aligned} u &= g_1(x,y) = x + y \\ v &= g_2(x,y) = y \end{aligned}$$

has only one (unique) solution

$$\begin{aligned} x &= g_1^{-1}(u,v) = u - v \\ y &= g_2^{-1}(u,v) = v. \end{aligned}$$

Therefore, the joint pmf of (U, V), for values of $(u, v) \in \mathcal{B}$, is given by

$$\begin{aligned} f_{U,V}(u,v) &= \sum_{(x,y)\in\mathcal{A}_{uv}} \sum f_{X,Y}(x,y) \\ &= \binom{n_1}{u-v} p^{u-v}(1-p)^{n_1-(u-v)} \binom{n_2}{v} p^v (1-p)^{n_2-v} \\ &= \binom{n_1}{u-v}\binom{n_2}{v} p^u (1-p)^{n_1+n_2-u}. \end{aligned}$$

This completes part (a). To do part (b), the marginal pmf $f_U(u)$ is found by summing $f_{U,V}(u,v)$ over values of $v \in \mathcal{B}$, that is, $v = 0, 1, 2, ..., u$. For $u = 0, 1, 2, ..., n_1 + n_2$, we have

$$\begin{aligned} f_U(u) &= \sum_{v=0}^{u} \binom{n_1}{u-v}\binom{n_2}{v} p^u (1-p)^{n_1+n_2-u} \\ &= p^u (1-p)^{n_1+n_2-u} \sum_{v=0}^{u} \binom{n_1}{u-v}\binom{n_2}{v}. \end{aligned}$$

It can be shown that

$$\sum_{v=0}^{u} \binom{n_1}{u-v}\binom{n_2}{v} = \binom{n_1+n_2}{u};$$

this is known as **Vandermonde's Identity**. Therefore,

$$f_U(u) = \binom{n_1+n_2}{u} p^u (1-p)^{n_1+n_2-u},$$

for $u = 0, 1, 2, ..., n_1 + n_2$, showing that $U = X + Y \sim b(n_1 + n_2, p)$.

Remark: Had we only been interested in finding the distribution of $U = X + Y$ in this example, note that an mgf argument would have been much easier. The mgf of U is

$$M_U(t) \stackrel{X \perp\!\!\!\perp Y}{=} M_X(t) M_Y(t) = (q + pe^t)^{n_1} (q + pe^t)^{n_2} = (q + pe^t)^{n_1+n_2},$$

which we recognize as the $b(n_1 + n_2, p)$ mgf. The result follows because mgfs are unique.

Continuous case: Suppose (X, Y) is a continuous random vector with joint pdf $f_{X,Y}(x, y)$ and support $\mathcal{A} \subseteq \mathbb{R}^2$. Define

$$\begin{aligned} U &= g_1(X, Y) \\ V &= g_2(X, Y) \end{aligned}$$

so that

$$\begin{pmatrix} U \\ V \end{pmatrix} = g \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} g_1(X,Y) \\ g_2(X,Y) \end{pmatrix}$$

is a vector-valued mapping from $\mathcal{A}$ to $\mathcal{B} \subseteq \mathbb{R}^2$, where

$$\mathcal{B} = \{(u, v) \in \mathbb{R}^2 : u = g_1(x, y),\ v = g_2(x, y),\ \text{for } (x, y) \in \mathcal{A}\};$$

i.e., $g : \mathcal{A} \to \mathcal{B}$. In what follows, we will assume that g is a one-to-one transformation. That is, for each $(u, v) \in \mathcal{B}$, there is only one $(x, y) \in \mathcal{A}$ satisfying

$$\begin{aligned} u &= g_1(x, y) \\ v &= g_2(x, y). \end{aligned}$$

Because g is one-to-one, we can find the **inverse transformation**

$$\begin{aligned} x &= g_1^{-1}(u, v) \\ y &= g_2^{-1}(u, v). \end{aligned}$$

The **Jacobian** of the (inverse) transformation is defined as

$$J = \det \begin{vmatrix} \dfrac{\partial g_1^{-1}(u,v)}{\partial u} & \dfrac{\partial g_1^{-1}(u,v)}{\partial v} \\ \dfrac{\partial g_2^{-1}(u,v)}{\partial u} & \dfrac{\partial g_2^{-1}(u,v)}{\partial v} \end{vmatrix},$$

that is, J is the determinant of this 2×2 matrix of partial derivatives. We will assume that $J \neq 0$ over $\mathcal{B}$. By a theorem in analysis (the Change of Variables Theorem), we are able to conclude that the joint pdf of (U, V) is, for $(u, v) \in \mathcal{B}$,

$$f_{U,V}(u, v) = f_{X,Y}(g_1^{-1}(u, v), g_2^{-1}(u, v))|J|,$$

where $|J|$ denotes the absolute value of J. Of course, if $(u, v) \notin \mathcal{B}$, then $f_{U,V}(u, v) = 0$.

Discussion: Let $A \subseteq \mathcal{A}$ and $B = g(A) \subseteq \mathcal{B}$; i.e., $g(A)$ is the image of A under the mapping g. Because $g : \mathcal{A} \to \mathcal{B}$ is one-to-one, the events $\{(X, Y) \in A\}$ and $\{(U, V) \in B\}$ have the same probability; i.e.,

$$P((U, V) \in B) = P((X, Y) \in A) = \int_A \int f_{X,Y}(x, y)dxdy.$$

The Change of Variables Theorem from analysis says that

$$\int_A \int f_{X,Y}(x, y)dxdy = \int_B \int f_{X,Y}(g_1^{-1}(u, v), g_2^{-1}(u, v))|J|dudv.$$

Therefore, for any $B \subseteq \mathcal{B}$, we have

$$P((U, V) \in B) = \int_B \int f_{X,Y}(g_1^{-1}(u, v), g_2^{-1}(u, v))|J|dudv.$$

This implies that the joint pdf of (U, V), where positive, is

$$f_{U,V}(u, v) = f_{X,Y}(g_1^{-1}(u, v), g_2^{-1}(u, v))|J|.$$

Example 4.16. Suppose that $X \sim \text{gamma}(\alpha_1, \beta)$, $Y \sim \text{gamma}(\alpha_2, \beta)$, and $X \perp\!\!\!\perp Y$. Define

$$\begin{aligned} U &= g_1(X,Y) = X + Y \\ V &= g_2(X,Y) = \frac{X}{X+Y}. \end{aligned}$$

(a) Find $f_{U,V}(u,v)$, the joint pdf of (U,V), using a bivariate transformation.
(b) Find $f_U(u)$, the marginal pdf of U.
(c) Find $f_V(v)$, the marginal pdf of V.
Solution. First, note that the joint pdf of (X,Y) is

$$\begin{aligned} f_{X,Y}(x,y) &\overset{X \perp\!\!\!\perp Y}{=} f_X(x) f_Y(y) \\ &= \frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}} x^{\alpha_1 - 1} e^{-x/\beta} I(x > 0) \times \frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}} y^{\alpha_2 - 1} e^{-y/\beta} I(y > 0) \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} x^{\alpha_1 - 1} y^{\alpha_2 - 1} e^{-(x+y)/\beta} I(x > 0, y > 0), \end{aligned}$$

and the support of (X,Y) is

$$\mathcal{A} = \{(x,y) \in \mathbb{R}^2 : x > 0,\ y > 0\}.$$

The transformation above maps values of $(x,y) \in \mathcal{A}$ to

$$\mathcal{B} = \{(u,v) \in \mathbb{R}^2 : u > 0,\ 0 < v < 1\};$$

i.e., $\mathcal{B}$ is the support of (U,V). To verify the transformation is one-to-one, we show that $g(x,y) = g(x^*,y^*) \in \mathcal{B} \Longrightarrow x = x^*$ and $y = y^*$, where

$$g\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix} = \begin{pmatrix} x+y \\ \dfrac{x}{x+y} \end{pmatrix}.$$

Suppose $g(x,y) = g(x^*,y^*)$. This means that both of these equations hold:

$$x + y = x^* + y^* \qquad \text{and} \qquad \frac{x}{x+y} = \frac{x^*}{x^*+y^*}.$$

The two equations together imply that $x = x^*$. The first equation then implies $y = y^*$. Hence, the transformation $g : \mathcal{A} \to \mathcal{B}$ is one-to-one. The inverse transformation is found by solving

$$\begin{aligned} u &= x + y \\ v &= \frac{x}{x+y} \end{aligned}$$

for $x = g_1^{-1}(u,v)$ and $y = g_2^{-1}(u,v)$. Straightforward algebra shows that

$$\begin{pmatrix} x \\ y \end{pmatrix} = g^{-1}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} g_1^{-1}(u,v) \\ g_2^{-1}(u,v) \end{pmatrix} = \begin{pmatrix} uv \\ u(1-v) \end{pmatrix}.$$

The Jacobian is

$$J = \det \begin{vmatrix} \dfrac{\partial g_1^{-1}(u,v)}{\partial u} & \dfrac{\partial g_1^{-1}(u,v)}{\partial v} \\ \dfrac{\partial g_2^{-1}(u,v)}{\partial u} & \dfrac{\partial g_2^{-1}(u,v)}{\partial v} \end{vmatrix} = \det \begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix} = -uv - u(1-v) = -u,$$

which is nonzero over $\mathcal{B}$. Therefore, the joint pdf of (U, V) is, for $u > 0$ and $0 < v < 1$,

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(g_1^{-1}(u,v), g_2^{-1}(u,v))|J| \\ &= f_{X,Y}(uv, u(1-v))|-u| \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}(uv)^{\alpha_1-1}[u(1-v)]^{\alpha_2-1}e^{-[uv+u(1-v)]/\beta} \\ &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}u^{\alpha_1+\alpha_2-1}v^{\alpha_1-1}(1-v)^{\alpha_2-1}e^{-u/\beta}. \end{aligned}$$

This completes part (a). To find the marginal pdf of U in part (b), we integrate $f_{U,V}(u,v)$ over $0 < v < 1$; that is,

$$\begin{aligned} f_U(u) &\overset{u>0}{=} \int_{v=0}^{1} \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}u^{\alpha_1+\alpha_2-1}v^{\alpha_1-1}(1-v)^{\alpha_2-1}e^{-u/\beta}dv \\ &= \frac{u^{\alpha_1+\alpha_2-1}e^{-u/\beta}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \underbrace{\int_{v=0}^{1} v^{\alpha_1-1}(1-v)^{\alpha_2-1}dv}_{=\,B(\alpha_1,\alpha_2)} \\ &= \frac{u^{\alpha_1+\alpha_2-1}e^{-u/\beta}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}\frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)} \\ &= \frac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}}u^{\alpha_1+\alpha_2-1}e^{-u/\beta}I(u>0), \end{aligned}$$

which we recognize as a gamma pdf with shape parameter $\alpha_1 + \alpha_2$ and scale parameter β. That is, $U = X + Y \sim \text{gamma}(\alpha_1 + \alpha_2, \beta)$. This completes part (b).

Remark: Had we only been interested in finding the distribution of $U = X + Y$ in this example, note that an mgf argument would have been much easier. The mgf of U is

$$M_U(t) \overset{X \perp\!\!\!\perp Y}{=} M_X(t)M_Y(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha_1}\left(\frac{1}{1-\beta t}\right)^{\alpha_2} = \left(\frac{1}{1-\beta t}\right)^{\alpha_1+\alpha_2},$$

for $t < 1/\beta$, which we recognize as the gamma$(\alpha_1 + \alpha_2, \beta)$ mgf. The result follows because mgfs are unique.

Finally, in part (c), to find the marginal pdf of V, we integrate $f_{U,V}(u,v)$ over $u > 0$; that is,

$$\begin{aligned} f_V(v) &\overset{0<v<1}{=} \int_{u=0}^{\infty} \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}u^{\alpha_1+\alpha_2-1}v^{\alpha_1-1}(1-v)^{\alpha_2-1}e^{-u/\beta}du \\ &= \frac{v^{\alpha_1-1}(1-v)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \underbrace{\int_{u=0}^{\infty} u^{\alpha_1+\alpha_2-1}e^{-u/\beta}du}_{=\,\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}} \\ &= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}v^{\alpha_1-1}(1-v)^{\alpha_2-1}I(0<v<1), \end{aligned}$$

which we recognize as a beta pdf with parameters α_1 and α_2. That is, $V = X/(X+Y) \sim$ beta(α_1, α_2). This completes part (c).

Remark: In addition to deriving the marginal distributions in this example, that is,

$$\begin{aligned} U &= X + Y \sim \text{gamma}(\alpha_1 + \alpha_2, \beta) \\ V &= \frac{X}{X+Y} \sim \text{beta}(\alpha_1, \alpha_2), \end{aligned}$$

note that U and V are also independent. We know this because we can write the joint pdf

$$\begin{aligned} f_{U,V}(u,v) &= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1+\alpha_2-1} v^{\alpha_1-1}(1-v)^{\alpha_2-1} e^{-u/\beta} I(u>0, 0<v<1) \\ &= \underbrace{\frac{u^{\alpha_1+\alpha_2-1}e^{-u/\beta}I(u>0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}}_{=\, g(u)} \times \underbrace{v^{\alpha_1-1}(1-v)^{\alpha_2-1}I(0<v<1)}_{=\, h(v)}. \end{aligned}$$

We have factored the joint pdf $f_{U,V}(u,v)$ into two expressions, one of which depends only on u and the other which depends only on v. From Lemma 4.2.7, we know that $U \perp\!\!\!\perp V$. We could have also concluded this by noting that $f_{U,V}(u,v) = f_U(u)f_V(v)$ for all $(u,v) \in \mathcal{B}$. Clearly, $g(u)$ and $h(v)$ above are proportional to $f_U(u)$ and $f_V(v)$, respectively.

Remark: We now illustrate the utility of the bivariate transformation technique in a situation where only one function of X and Y is of interest.

Example 4.17. Suppose that $X \sim \mathcal{N}(0,1)$, $Y \sim \mathcal{N}(0,1)$, and $X \perp\!\!\!\perp Y$. Find the distribution of X/Y.

Solution. First, note that the joint pdf of (X,Y) is

$$\begin{aligned} f_{X,Y}(x,y) &\overset{X \perp\!\!\!\perp Y}{=} f_X(x)f_Y(y) \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \\ &= \frac{1}{2\pi} e^{-(x^2+y^2)/2}, \end{aligned}$$

for all $(x,y) \in \mathbb{R}^2$; i.e., the support of (X,Y) is $\mathcal{A} = \mathbb{R}^2$. We initially have an obvious problem; the transformation

$$U = g_1(X,Y) = \frac{X}{Y},$$

by itself, is not one-to-one; e.g., $g_1(1,1) = g_1(2,2) = 1$. A second problem is that the transformation $U = X/Y$ is not defined when $y = 0$. We deal with the second problem first. We do this by redefining the joint pdf as

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)/2},$$

but only for those values of $(x,y) \in \mathcal{A}^*$, where

$$\mathcal{A}^* = \{(x,y) \in \mathbb{R}^2 : x \in \mathbb{R}, \ y \in \mathbb{R} - \{0\}\}.$$

The joint pdf $f_{X,Y}(x,y)$ over $\mathcal{A}$ and the one over $\mathcal{A}^*$ define the same probability distribution for (X,Y) because $P(\mathcal{A} \setminus \mathcal{A}^*) = 0$; see also pp 156 (CB). Now, to "make" the transformation one-to-one, we define a second variable $V = g_2(X,Y)$ and augment $g_1(X,Y)$ with it; i.e.,

$$\begin{pmatrix} U \\ V \end{pmatrix} = g\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} g_1(X,Y) \\ g_2(X,Y) \end{pmatrix}.$$

We want to choose $V = g_2(X,Y)$ to be something "easy" and also so that $g : \mathcal{A}^* \to \mathcal{B}^*$, say, is one-to-one. Consider adding $V = g_2(X,Y) = Y$ so that the transformation is

$$\begin{aligned} U &= g_1(X,Y) = \frac{X}{Y} \\ V &= g_2(X,Y) = Y. \end{aligned}$$

Clearly, g is one-to-one; i.e., $g(x,y) = g(x^*,y^*)$ implies that $x = x^*$ and $y = y^*$. The support of (U,V) is

$$\mathcal{B}^* = \{(u,v) \in \mathbb{R}^2 : u \in \mathbb{R},\ v \in \mathbb{R} - \{0\}\}.$$

The inverse transformation is found by solving

$$\begin{aligned} u &= \frac{x}{y} \\ v &= y \end{aligned}$$

for $x = g_1^{-1}(u,v)$ and $y = g_2^{-1}(u,v)$. Straightforward algebra shows that

$$\begin{pmatrix} x \\ y \end{pmatrix} = g^{-1}\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} g_1^{-1}(u,v) \\ g_2^{-1}(u,v) \end{pmatrix} = \begin{pmatrix} uv \\ v \end{pmatrix}.$$

The Jacobian is

$$J = \det\begin{vmatrix} \dfrac{\partial g_1^{-1}(u,v)}{\partial u} & \dfrac{\partial g_1^{-1}(u,v)}{\partial v} \\ \dfrac{\partial g_2^{-1}(u,v)}{\partial u} & \dfrac{\partial g_2^{-1}(u,v)}{\partial v} \end{vmatrix} = \det\begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v(1) - u(0) = v,$$

which is nonzero over $\mathcal{B}^*$. Therefore, the joint pdf of (U,V), for $(u,v) \in \mathcal{B}^*$, is given by

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(g_1^{-1}(u,v), g_2^{-1}(u,v))|J| \\ &= f_{X,Y}(uv,v)|v| \\ &= \frac{|v|}{2\pi} e^{-[(uv)^2+v^2]/2} \\ &= \frac{|v|}{2\pi} e^{-v^2/2\left(\frac{1}{1+u^2}\right)}. \end{aligned}$$

Remark: If we let $\mathcal{B} = \mathbb{R}^2$, then the joint pdf $f_{U,V}(u,v)$ over $\mathcal{B}$ and the one over $\mathcal{B}^*$ (as shown above) define the same probability distribution for (U,V) because $P(\mathcal{B} \setminus \mathcal{B}^*) = 0$. Therefore, in what follows, we can work with $f_{U,V}(u,v)$ defined over $\mathcal{B}$ instead.

Recall that our original goal was to find the distribution of $U = X/Y$. The pdf of U, for $-\infty < u < \infty$, is given by

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_{U,V}(u,v)dv \\ &= \int_{-\infty}^{\infty} \frac{|v|}{2\pi} e^{-v^2/2\left(\frac{1}{1+u^2}\right)} dv. \end{aligned}$$

To do this integral, let $\sigma^2 = 1/(1+u^2)$ and write

$$f_U(u) = \int_{-\infty}^{\infty} \frac{|v|}{2\pi} e^{-v^2/2\left(\frac{1}{1+u^2}\right)} dv = \frac{1}{2\pi} \int_{-\infty}^{\infty} \underbrace{|v| e^{-v^2/2\sigma^2}}_{=\ h(v),\ \text{say}} dv.$$

Note that $h(v)$ is an even function; i.e., $h(v) = h(-v)$, for all $v \in \mathbb{R}$. This means that $h(v)$ is symmetric about $v = 0$. Therefore, the last integral

$$\int_{-\infty}^{\infty} |v| e^{-v^2/2\sigma^2} dv = 2\int_0^{\infty} v e^{-v^2/2\sigma^2} dv.$$

Therefore, for $-\infty < u < \infty$,

$$\begin{aligned} f_U(u) &= \frac{1}{2\pi}\, 2\int_0^{\infty} v e^{-v^2/2\sigma^2} dv \\ &= \frac{1}{\pi}\left(-\sigma^2 e^{-v^2/2\sigma^2}\Big|_{v=0}^{\infty}\right) = \frac{\sigma^2}{\pi}(1-0) = \frac{1}{\pi(1+u^2)}, \end{aligned}$$

which we recognize as the pdf of $U \sim \text{Cauchy}(0,1)$. We have shown that the ratio of two independent standard normal random variables follows a Cauchy distribution (specifically, a "standard" Cauchy distribution).

Remark: Compare our solution to Example 4.17 with the solution provided by CB (pp 162). The authors augmented the $g_1(X,Y) = X/Y$ transformation with $g_2(X,Y) = |Y|$ instead of with $g_2(X,Y) = Y$ as we did. Their transformation is not one-to-one, so they "break up" $\mathcal{A}$ into disjoint regions over which, individually, the transformation is one-to-one; they then apply the transformation separately over these regions.

Theorem 4.3.5. Suppose that X and Y are independent random variables (discrete or continuous). The random variables $U = g(X)$ and $V = h(Y)$ are also independent.

Remark: Theorem 4.3.5 says that functions of independent random variables are themselves independent. In the statement above, it is assumed that g is a function of X only; similarly, h is a function of Y only.

Proof. Assume that X and Y are jointly continuous. For any $u, v \in \mathbb{R}$, define the sets

$$\begin{aligned} A_u &= \{x \in \mathbb{R} : g(x) \leq u\} \\ B_v &= \{y \in \mathbb{R} : h(y) \leq v\}. \end{aligned}$$

The joint cdf of (U, V) is

$$\begin{aligned} F_{U,V}(u,v) &= P(U \leq u, V \leq v) \\ &= P(X \in A_u, Y \in B_v) \\ &\overset{X \perp\!\!\!\perp Y}{=} P(X \in A_u)P(Y \in B_v), \end{aligned}$$

the last step following from Theorem 4.2.10(a). Therefore, the joint pdf of (U, V) is

$$\begin{aligned} f_{U,V}(u,v) &= \frac{\partial^2}{\partial u \partial v} F_{U,V}(u,v) \\ &= \frac{\partial^2}{\partial u \partial v} P(X \in A_u)P(Y \in B_v) \\ &= \underbrace{\frac{d}{du}P(X \in A_u)}_{\text{function of } u} \underbrace{\frac{d}{dv}P(Y \in A_v)}_{\text{function of } v}. \end{aligned}$$

By Lemma 4.2.7, U and V are independent. □

4.4 Hierarchical Models and Mixture Distributions

Example 4.18. Suppose $X \sim b(n, p)$. As a frame of reference, suppose that

$$X = \text{ number of germinating seeds per plot (out of } n \text{ seeds).}$$

A generalization of this model would allow p, the probability of "success," to have its own probability distribution. Suppose that

$$\begin{aligned} X|P &\sim b(n, P) \\ P &\sim \text{beta}(\alpha, \beta), \end{aligned}$$

where n is fixed and $\alpha, \beta > 0$. The model in the second layer $P \sim \text{beta}(\alpha, \beta)$ acknowledges that the probability of success varies across plots.

Example 4.19. Consider the hierarchy

$$\begin{aligned} X|N &\sim b(N, p) \\ N &\sim \text{Poisson}(\lambda), \end{aligned}$$

where p is fixed and $\lambda > 0$. As a frame of reference, suppose N is the number of eggs laid (random) and X is the number of surviving offspring. This model would be applicable if each offspring's survival status (yes/no) is independent and $p = \text{pr}(\text{"offspring survives"})$ is the same for each egg.

Remark: The models in Example 4.18 and 4.19 are called **hierarchical models**.

Example 4.18 (continued). In the binomial-beta hierarchy, we now find $f_X(x)$, the marginal distribution of X.
Solution. Note that $x \in \{0, 1, 2, ..., n\}$ and $0 < p < 1$. The joint distribution of X and P is

$$\begin{aligned} f_{X,P}(x,p) &= f_{X|P}(x|p)f_P(p) \\ &= \binom{n}{x}p^x(1-p)^{n-x}\,\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1} \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,p^{x+\alpha-1}(1-p)^{n-x+\beta-1}. \end{aligned}$$

Therefore, the marginal pmf of X, for $x = 0, 1, 2, ..., n$, is given by

$$\begin{aligned} f_X(x) &= \int_0^1 f_{X|P}(x|p)f_P(p)dp \\ &= \int_0^1 \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,p^{x+\alpha-1}(1-p)^{n-x+\beta-1}dp \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1}dp \\ &= \binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}. \end{aligned}$$

This is called the **beta-binomial distribution**. We write $X \sim$ beta-binomial(n, α, β).

Q: If $X \sim$ beta-binomial(n, α, β), what are $E(X)$ and var(X)?
A: From the definition,

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x f_X(x) \\ &= \sum_{x=0}^{n} x\binom{n}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}. \end{aligned}$$

Clearly, this is not a friendly calculation. Unfortunately, $E(X^2)$ and $E(e^{tX})$ are even less friendly.

Theorem 4.4.3. If X and Y are any two random variables, then

$$E(X) = E[E(X|Y)],$$

provided that all expectations exist.

Remark: The result in Theorem 4.4.3 is called the **iterated rule for expectations**. Before we prove this result, it is important to note that there are really three different expectations here:

$$\begin{aligned} E(X) &\longrightarrow \text{refers to the marginal distribution of } X \\ E(X|Y) &\longrightarrow \text{refers to the conditional distribution of } X|Y \\ E[E(X|Y)] &\longrightarrow \text{calculated using the marginal distribution of } Y. \end{aligned}$$

Recall that $E(X|Y)$ is a function of Y, say $g(Y)$.

Proof. Suppose (X, Y) is continuous with joint pdf $f_{X,Y}(x, y)$. The LHS is

$$\begin{aligned} E(X) &= \int\int_{\mathbb{R}^2} x f_{X,Y}(x,y)dxdy \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} x f_{X|Y}(x|y) f_Y(y)dxdy \\ &= \int_{\mathbb{R}} \underbrace{\left[\int_{\mathbb{R}} x f_{X|Y}(x|y)dx\right]}_{E(X|Y=y)} f_Y(y)dy \\ &= \int_{\mathbb{R}} E(X|Y=y) f_Y(y)dy = E[E(X|Y)]. \end{aligned}$$

The discrete case is proven by replacing integrals with sums. □

Illustration: Let's return to our binomial-beta hierarchy in Example 4.18, that is,

$$\begin{aligned} X|P &\sim b(n, P) \\ P &\sim \text{beta}(\alpha, \beta). \end{aligned}$$

The mean of X is

$$E(X) = E[E(X|P)] = E(nP) = nE(P) = n\left(\frac{\alpha}{\alpha+\beta}\right).$$

Remark: This example illustrates an important lesson when finding expected values. In some problems, it is difficult to calculate $E(X)$ directly (i.e., using the marginal distribution of X). By judicious use of conditioning, the calculation becomes much easier.

Definition: A random variable X has a **mixture distribution** if the distribution of X depends on a quantity that also has a distribution.

Remark: We can classify the beta-binomial distribution as a mixture distribution because it arises from the hierarchy

$$\begin{aligned} X|P &\sim b(n, P) \\ P &\sim \text{beta}(\alpha, \beta). \end{aligned}$$

In general, we can write the beta-binomial pmf as

$$f_X(x) = \int_0^1 f_{X|P}(x|p) f_P(p)dp;$$

i.e., $f_X(x)$ can be thought of as an "average" of values of $f_{X|P}(x|p)$. The pdf $f_P(p)$ is called a **mixing distribution**. In the Example 4.19 hierarchy

$$\begin{aligned} X|N &\sim b(N, p) \\ N &\sim \text{Poisson}(\lambda), \end{aligned}$$

Casella and Berger (pp 163) show that, for $x = 0, 1, 2, ...$,

$$\begin{aligned} f_X(x) &= \sum_{n=0}^{\infty} f_{X|N}(x|n) f_N(n) \\ &\overset{n \geqslant x}{=} \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \, \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \frac{(\lambda p)^x e^{-\lambda p}}{x!}. \end{aligned}$$

In this example, the Poisson pmf $f_N(n)$ is the mixing distribution and, marginally, $X \sim$ Poisson(λp). Note that $\lambda p = E(X) = E[E(X|N)]$.

Example 4.20. *Non-central χ^2 distribution.* Consider the hierarchy

$$\begin{aligned} X|Y &\sim \chi^2_{p+2Y} \\ Y &\sim \text{Poisson}(\lambda), \end{aligned}$$

where $p > 0$ and $\lambda > 0$. The conditional pdf of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{1}{\Gamma(\frac{p}{2}+y)2^{\frac{p}{2}+y}} x^{\frac{p}{2}+y-1} e^{-x/2} I(x > 0).$$

Therefore, the marginal pdf of X, for $x > 0$, is given by

$$\begin{aligned} f_X(x) = \sum_{y=0}^{\infty} f_{X,Y}(x,y) &= \sum_{y=0}^{\infty} f_{X|Y}(x|y) f_Y(y) \\ &= \sum_{y=0}^{\infty} \frac{1}{\Gamma(\frac{p}{2}+y)2^{\frac{p}{2}+y}} x^{\frac{p}{2}+y-1} e^{-x/2} \left(\frac{\lambda^y e^{-\lambda}}{y!} \right). \end{aligned}$$

This is called the **non-central χ^2 distribution** with p degrees of freedom and non-centrality parameter $\lambda > 0$, written $X \sim \chi^2_p(\lambda)$. The non-central χ^2 distribution can be thought of as a mixture distribution; it is essentially an infinite weighted average of (central) χ^2 densities where the mixing distribution is Poisson(λ). If $\lambda = 0$, the non-central $\chi^2_p(\lambda)$ distribution reduces to our "usual" central χ^2_p distribution.

Note: To find $E(X)$, we could calculate

$$E(X) = \int_0^{\infty} x f_X(x) dx = \int_0^{\infty} x \, \sum_{y=0}^{\infty} \frac{1}{\Gamma(\frac{p}{2}+y)2^{\frac{p}{2}+y}} x^{\frac{p}{2}+y-1} e^{-x/2} \left(\frac{\lambda^y e^{-\lambda}}{y!} \right) dx.$$

Alternatively, we could simply calculate

$$E(X) = E[E(X|Y)] = E(p + 2Y) = p + 2\lambda$$

using the iterated rule for expectations.

Theorem 4.4.7. If X and Y are any two random variables, then

$$\text{var}(X) = E[\text{var}(X|Y)] + \text{var}[E(X|Y)],$$

provided that all expectations exist.

Remark: The result in Theorem 4.4.7 is called the **iterated rule for variances**. It is also known (informally) as "Adam's Rule."

Proof. First, note that

$$\begin{aligned} E[\text{var}(X|Y)] &= E\{E(X^2|Y) - [E(X|Y)]^2\} \\ &= E[E(X^2|Y)] - E\{[E(X|Y)]^2\} \\ &= E(X^2) - E\{[E(X|Y)]^2\}. \end{aligned}$$

Second, note that

$$\begin{aligned} \text{var}[E(X|Y)] &= E\{[E(X|Y)]^2\} - \{E[E(X|Y)]\}^2 \\ &= E\{[E(X|Y)]^2\} - [E(X)]^2. \end{aligned}$$

Combining these two equations completes the proof. □

Example 4.21. Calculate $\text{var}(X)$ if $X \sim \chi^2_p(\lambda)$.
Solution. Use the fact that X is a mixture random variable arising from the hierarchy

$$\begin{aligned} X|Y &\sim \chi^2_{p+2Y} \\ Y &\sim \text{Poisson}(\lambda). \end{aligned}$$

Note that

$$\begin{aligned} E[\text{var}(X|Y)] &= E[2(p+2Y)] = 2p + 4\lambda \\ \text{var}[E(X|Y)] &= \text{var}(p+2Y) = 4\lambda. \end{aligned}$$

Therefore,

$$\begin{aligned} \text{var}(X) &= E[\text{var}(X|Y)] + \text{var}[E(X|Y)] \\ &= 2p + 4\lambda + 4\lambda \\ &= 2p + 8\lambda. \end{aligned}$$

Example 4.22. Find the moment generating function of $X \sim \chi^2_p(\lambda)$.
Solution. We again exploit the hierarchy

$$\begin{aligned} X|Y &\sim \chi^2_{p+2Y} \\ Y &\sim \text{Poisson}(\lambda). \end{aligned}$$

The mgf of X is given by

$$M_X(t) = E(e^{tX}) = E[E(e^{tX}|Y)].$$

Because $X|Y \sim \chi^2_{p+2Y}$, we know that

$$E(e^{tX}|Y) = \left(\frac{1}{1-2t}\right)^{\frac{p}{2}+Y}, \quad \text{for } t < \frac{1}{2}.$$

Note that this is the (conditional) mgf of X given Y. Therefore,

$$M_X(t) = E[E(e^{tX}|Y)] = E\left[\left(\frac{1}{1-2t}\right)^{\frac{p}{2}+Y}\right].$$

The last expectation is an expectation taken with respect to the marginal distribution of Y. Because $Y \sim \text{Poisson}(\lambda)$, we have

$$\begin{aligned} M_X(t) = E\left[\left(\frac{1}{1-2t}\right)^{\frac{p}{2}+Y}\right] &= \sum_{y=0}^{\infty}\left(\frac{1}{1-2t}\right)^{\frac{p}{2}+y}\frac{\lambda^y e^{-\lambda}}{y!} \\ &= e^{-\lambda}\left(\frac{1}{1-2t}\right)^{p/2}\underbrace{\sum_{y=0}^{\infty}\frac{\left(\frac{\lambda}{1-2t}\right)^y}{y!}}_{=\exp\left(\frac{\lambda}{1-2t}\right)} \\ &= \left(\frac{1}{1-2t}\right)^{p/2}\exp\left(\frac{2\lambda t}{1-2t}\right). \end{aligned}$$

This is the mgf of $X \sim \chi^2_p(\lambda)$, valid for $t < 1/2$.

Exercise: If $X \sim \mathcal{N}(\mu, 1)$, show that $Y = X^2 \sim \chi^2_1(\lambda)$, where $\lambda = \mu^2/2$. *Hint:* Derive the mgf of Y.

4.5 Covariance and Correlation

Setting: We have two random variables X and Y with finite means and variances. Denote by

$$\begin{aligned} E(X) &= \mu_X \qquad & \text{var}(X) &= \sigma^2_X < \infty \\ E(Y) &= \mu_Y \qquad & \text{var}(Y) &= \sigma^2_Y < \infty. \end{aligned}$$

Definitions: The **covariance** of X and Y is

$$\text{cov}(X, Y) \equiv \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)].$$

The **correlation** of X and Y is

$$\text{corr}(X, Y) \equiv \rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Notes:

1. Both σ_{XY} and ρ_{XY} are real numbers. Specifically,

$$-\infty < \sigma_{XY} < \infty$$

and

$$-1 \leq \rho_{XY} \leq 1.$$

2. Both σ_{XY} and ρ_{XY} describe the strength and direction of the **linear relationship** between X and Y. Values of $\rho_{XY} = \pm 1$ indicate a perfect linear relationship.

Theorem 4.5.3. For any two random variables X and Y,

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

This is called the **covariance computing formula**.
Proof. From the definition,

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

Discoveries: It is easy to establish each of the following results:

1. $\text{cov}(X, Y) = \text{cov}(Y, X)$
2. $\text{cov}(X, X) = \text{var}(X)$
3. $\text{cov}(a, X) = 0$, for any constant $a \in \mathbb{R}$.

Theorem 4.5.5. If $X \perp\!\!\!\perp Y$, then $\text{cov}(X, Y) = 0$.
Proof. If $X \perp\!\!\!\perp Y$, then $E(XY) = E(X)E(Y)$ by Theorem 4.2.10(b). $\square$

Remark: The converse of Theorem 4.5.5 is not true in general. That is,

$$\text{cov}(X, Y) = 0 \quad \not\Longrightarrow \quad X \perp\!\!\!\perp Y.$$

Counterexample: $X \sim \mathcal{N}(0, 1)$ and $Y = X^2$. Note that

$$E(XY) = E(X^3) = E[(X - 0)^3],$$

which is the numerator of ξ, the skewness of X. Because the normal pdf is symmetric, $\xi = 0$. Therefore, $E(XY) = E(X^3) = 0$. Also, $E(X) = 0$ and $E(Y) = 1$, so $\text{cov}(X, Y) = 0$. However, clearly X and Y are not independent. They are perfectly related (just not linearly).

Theorem 4.5.6. If X and Y are any two random variables and a and b are constants, then

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

Proof. Apply the variance computing formula $\text{var}(W) = E(W^2) - [E(W)]^2$, with $W = aX + bY$. □

Note: The following are commonly-seen special cases of Theorem 4.5.6:

- $a = b = 1$:
$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$
- $a = 1$, $b = -1$:
$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)$$
- $X \perp\!\!\!\perp Y$, $a = 1$, $b = \pm 1$:
$$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y).$$

Theorem 4.5.7. For any two random variables X and Y,
(a) $-1 \leq \rho_{XY} \leq 1$
(b) $|\rho_{XY}| = 1$ if and only if there exists constants $a, b \in \mathbb{R}$, $a \neq 0$, such that

$$P(Y = aX + b) = 1;$$

i.e., $Y = aX + b$ with probability 1 ("almost surely").
Proof. Define the function

$$h(t) = E\{[(X - \mu_X)t + (Y - \mu_Y)]^2\}.$$

Note first that $h(t) \geq 0$ for all $t \in \mathbb{R}$. Expanding the square and taking expectations,

$$h(t) = \sigma_X^2 t^2 + 2\text{cov}(X, Y)t + \sigma_Y^2,$$

a quadratic function of t. Because non-negative quadratic functions can have at most one real root, the discriminant of $h(t)$; i.e., $[2\text{cov}(X, Y)]^2 - 4\sigma_X^2\sigma_Y^2 \leq 0$. However, note that

$$\begin{aligned} [2\text{cov}(X, Y)]^2 - 4\sigma_X^2\sigma_Y^2 \leq 0 &\iff [\text{cov}(X, Y)]^2 \leq \sigma_X^2\sigma_Y^2 \\ &\iff -\sigma_X\sigma_Y \leq \text{cov}(X, Y) \leq \sigma_X\sigma_Y. \end{aligned} \tag{4.1}$$

Dividing through by $\sigma_X\sigma_Y$ gives

$$-1 \leq \frac{\text{cov}(X, Y)}{\sigma_X\sigma_Y} \leq 1,$$

establishing part (a). To prove part (b), note that

$$\begin{aligned} |\rho_{XY}| = 1 &\iff |\text{cov}(X, Y)| = \sigma_X\sigma_Y \\ &\iff [\text{cov}(X, Y)]^2 = \sigma_X^2\sigma_Y^2 \\ &\iff \text{LHS of Equation (4.1)} = 0 \\ &\iff h(t) \text{ has a single root of multiplicity 2.} \end{aligned}$$

However, because $[(X-\mu_X)t+(Y-\mu_Y)]^2$ is a non-negative random variable, its expectation $h(t)=0$ if and only if

$$\begin{aligned}[(X-\mu_X)t+(Y-\mu_Y)]^2=0 \text{ a.s.} &\iff (X-\mu_X)t+(Y-\mu_Y)=0 \text{ a.s.} \\ &\iff Y=-tX+\mu_X t+\mu_Y \text{ a.s.}\end{aligned}$$

We have shown $Y=aX+b$ almost surely for some $a,b\in\mathbb{R}$, $a\neq 0$. Thus, we are done. □

Interpretation: If $|\rho_{XY}|=1$, then the entire bivariate distribution of (X,Y) falls on a straight line with positive slope ($\rho_{XY}=1$) or negative slope ($\rho_{XY}=-1$).

<u>Bivariate Normal Distribution</u>

Definition: The random vector is said to have a **bivariate normal distribution** if the joint pdf of (X,Y) is

$$f_{X,Y}(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-Q/2},$$

for all $(x,y)\in\mathbb{R}^2$, where

$$Q=\frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2-2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)+\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right].$$

Notation: $(X,Y)\sim \text{mvn}_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$ are

$$\boldsymbol{\mu}=\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}_{2\times 1} \quad \text{and} \quad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_X^2 & \sigma_{XY}\\ \sigma_{YX} & \sigma_Y^2\end{pmatrix}_{2\times 2},$$

respectively, and where $\sigma_{XY}=\sigma_{YX}=\rho\sigma_X\sigma_Y$. Note that $\boldsymbol{\Sigma}$ is symmetric.

Remark: When we discuss the bivariate normal distribution, we will assume that the correlation $\rho\in(-1,1)$. If $\rho=\pm 1$, then (X,Y) does not have a pdf. This situation gives rise to what is known as a "less than full rank normal distribution." As we have just seen, $\rho=\pm 1$ means that all of the probability mass for (X,Y) is completely concentrated in a linear subspace of $\mathbb{R}^2$.

Note: If $(X,Y)\sim \text{mvn}_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$, then the (joint) moment generating function is

$$M_{X,Y}(\mathbf{t})=\exp(\mathbf{t}'\boldsymbol{\mu}+\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}/2),$$

where $\mathbf{t}=(t_1,t_2)'$.

Facts: Suppose $(X,Y)\sim \text{mvn}_2(\boldsymbol{\mu},\boldsymbol{\Sigma})$.

1. $X\sim\mathcal{N}(\mu_X,\sigma_X^2)$ and $Y\sim\mathcal{N}(\mu_Y,\sigma_Y^2)$. That is, bivariate normality implies univariate normality. This is easy to show using the joint mgf above.
    - The converse is not true. That is, univariate normality of X and Y does not necessarily imply that (X,Y) is bivariate normal; see Exercise 4.47 (pp 200 CB).

2. Conditional distributions are normal. Specifically,

$$Y|\{X=x\} \sim \mathcal{N}(\beta_0+\beta_1 x,\ \sigma_Y^2(1-\rho^2)),$$

where

$$\begin{aligned}\beta_0 &= \mu_Y - \beta_1\mu_X \\ \beta_1 &= \rho\left(\frac{\sigma_Y}{\sigma_X}\right).\end{aligned}$$

Note that the conditional mean $E(Y|X=x)=\beta_0+\beta_1 x$ is a linear function of x and the conditional variance $\text{var}(Y|X=x)=\sigma_Y^2(1-\rho^2)$ is free of x.

3. In the bivariate normal model,

$$\text{cov}(X,Y)=0 \iff X \perp\!\!\!\perp Y.$$

Recall that this is not true in general. To prove the necessity ($\Longrightarrow$) one can show that the joint mgf factors into the product of the marginal normal mgfs (when $\rho=0$). One could also show that when $\rho=0$, joint pdf $f_{X,Y}(x,y)=f_X(x)f_Y(y)$, where $f_X(x)$ and $f_Y(y)$ are the marginal pdfs of $X \sim \mathcal{N}(\mu_X,\sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y,\sigma_Y^2)$, respectively.

4.6 Multivariate Distributions

Remark: We now generalize many of our bivariate distribution definitions and results to $n \geq 2$ dimensions. We will use the following notation:

$$\begin{aligned}\mathbf{X} &= (X_1,X_2,...,X_n) \quad \longleftarrow \quad \text{random vector} \\ \mathbf{x} &= (x_1,x_2,...,x_n) \quad \longleftarrow \quad \text{realization of } \mathbf{X}.\end{aligned}$$

Mathematical definition: Suppose $(S,\mathcal{B},P)$ is a probability space. We call $\mathbf{X}: S \to \mathbb{R}^n$ a random vector if

$$\mathbf{X}^{-1}(B) \equiv \{\omega \in S : \mathbf{X}(\omega) \in B\} \in \mathcal{B},$$

for all $B \in \mathcal{B}(\mathbb{R}^n)$. Sets $B \in \mathcal{B}(\mathbb{R}^n)$ are called (n-dimensional) Borel sets. One can characterize $\mathcal{B}(\mathbb{R}^n)$ as the smallest σ-algebra generated by the collection of all half-open hyperrectangles; i.e.,

$$\{(x_1,x_2,...,x_n) : -\infty < x_1 \leq a_1, -\infty < x_2 \leq a_2, ..., -\infty < x_n \leq a_n,\ a_i \in \mathbb{R}\}.$$

The range probability space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_{\mathbf{X}})$. We call $P_{\mathbf{X}}$ the induced probability measure of $\mathbf{X}$. Similar to the univariate case, there is a one-to-one correspondence between $P_{\mathbf{X}}$ and a random vector's cumulative distribution function (cdf), which is defined as

$$F_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{X}}(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n),$$

for all $\mathbf{x} \in \mathbb{R}^n$. As before, we will eventually start writing P for $P_{\mathbf{X}}$.

Definition: We call a random vector $\mathbf{X}$ **discrete** if there exists a countable set $\mathcal{A} \subset \mathbb{R}^n$ such that $P_{\mathbf{X}}(\mathbf{X} \in \mathcal{A}) = 1$. The joint probability mass function (pmf) of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{X}}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).$$

For any $B \in \mathcal{B}(\mathbb{R}^n)$,

$$P_{\mathbf{X}}(\mathbf{X} \in B) = \sum_{\mathbf{x} \in B} f_{\mathbf{X}}(\mathbf{x}).$$

Definition: The random vector $\mathbf{X}$ is **continuous** if there exists a function $f_{\mathbf{X}} : \mathbb{R}^n \to \mathbb{R}$ such that

$$P_{\mathbf{X}}(\mathbf{X} \in B) = \int_B f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x},$$

for all $B \in \mathcal{B}(\mathbb{R}^n)$. We call $f_{\mathbf{X}}(\mathbf{x})$ the joint probability density function (pdf) of $\mathbf{X}$. In the continuous case, the (joint) cdf and the joint pdf are related through

$$\frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_{\mathbf{X}}(\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}),$$

for all $\mathbf{x} \in \mathbb{R}^n$, provided that this partial derivative exists.

Mathematical Expectation: Suppose $\mathbf{X}$ is a random vector and let $g : \mathbb{R}^n \to \mathbb{R}$. Then $g(\mathbf{X})$ is a random variable and its expected value is

$$\begin{aligned} E[g(\mathbf{X})] &= \sum_{\mathbf{x} \in \mathcal{A}} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) \qquad \text{(discrete case)} \\ E[g(\mathbf{X})] &= \int_{\mathbb{R}^n} g(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \qquad \text{(continuous case)}. \end{aligned}$$

The usual existence issues arise; we need the sum (integral) above to converge absolutely. Otherwise, $E[g(\mathbf{X})]$ does not exist.

Marginal distributions: Suppose $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x})$. If $\mathbf{X}$ is continuous, the marginal pdf of X_i is given by

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}_{(-i)},$$

where $\mathbf{x}_{(-i)} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$. If $\mathbf{X}$ is discrete, the marginal pmf of X_i is

$$f_{X_i}(x_i) = \sum_{\mathbf{x}_{(-i)} \in \mathcal{A}} f_{\mathbf{X}}(\mathbf{x}).$$

In other words, to find the marginal pdf (pmf) of X_i, we integrate (sum) $f_{\mathbf{X}}(\mathbf{x})$ over the other $n-1$ variables. The "bivariate marginal" pdf $f_{X_i,X_j}(x_i, x_j)$ of (X_i, X_j) can be found by integrating (summing) $f_{\mathbf{X}}(\mathbf{x})$ over the other $n-2$ variables, and so on.

Conditional distributions: To find the conditional pdf (pmf) of a subset of random variables, divide the joint pdf (pmf) $f_{\mathbf{X}}(\mathbf{x})$ by the pdf (pmf) of the other variables. For

example, suppose $\mathbf{X} = (X_1, X_2, X_3) \sim f_{X_1,X_2,X_3}(x_1, x_2, x_3)$. Then, for example,

$$\begin{aligned} f_{X_1|X_2,X_3}(x_1|x_2,x_3) &= \frac{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}{f_{X_2,X_3}(x_2,x_3)} \\ f_{X_1,X_2|X_3}(x_1,x_2|x_3) &= \frac{f_{X_1,X_2,X_3}(x_1,x_2,x_3)}{f_{X_3}(x_3)}. \end{aligned}$$

The first conditional distribution describes the univariate distribution of X_1 when $X_2 = x_2$ and $X_3 = x_3$. The second distribution describes the bivariate distribution of X_1 and X_2 when $X_3 = x_3$.

Remark: Example 4.6.1 (pp 178-180 CB) illustrates many of the multivariate distribution concepts we have discussed so far.

Moment generating functions: Set $\mathbf{t} = (t_1, t_2, ..., t_n)'$. The moment generating function of $\mathbf{X} = (X_1, X_2, ..., X_n)'$ is

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{X}}) &= E\left(e^{t_1X_1+t_2X_2+\cdots+t_nX_n}\right) \\ &= \int_{\mathbb{R}^n} e^{t_1x_1+t_2x_2+\cdots+t_nx_n} dF_{\mathbf{X}}(\mathbf{x}). \end{aligned}$$

For $M_{\mathbf{X}}(\mathbf{t})$ to exist, we need $E(e^{\mathbf{t}'\mathbf{X}}) < \infty$ for all $\mathbf{t}$ in an open neighborhood about $\mathbf{0}$. That is, $\exists h_1 \exists h_2 \cdots \exists h_n > 0$ such that $E(e^{\mathbf{t}'\mathbf{X}}) < \infty$ for all $t_i \in (-h_i, h_i)$, $i = 1, 2, ..., n$. Otherwise, we say that $M_{\mathbf{X}}(\mathbf{t})$ does not exist.

Remark: As we saw in the bivariate case, we can obtain marginal mgfs from a joint mgf. The marginal mgf of X_i is

$$M_{X_i}(t_i) = M_{\mathbf{X}}(0, ..., 0, t_i, 0, ..., 0),$$

where t_i is in the ith position. The (joint) marginal mgf of $(X_i, X_j)'$ is found by

$$M_{X_i,X_j}(t_i, t_j) = M_{\mathbf{X}}(0, ..., 0, t_i, 0, ..., 0, t_j, 0, ..., 0).$$

Recall that from the marginal mgf $M_{X_i}(t_i)$, we can calculate

$$E(X_i) = \frac{d}{dt_i} M_{X_i}(t_i)\bigg|_{t_i=0}.$$

A new result is that

$$E(X_iX_j) = \frac{\partial^2}{\partial t_i \partial t_j} M_{X_i,X_j}(t_i, t_j)\bigg|_{t_i=t_j=0}.$$

From these, we can calculate

$$\text{cov}(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j).$$

Multinomial Distribution

Experiment: Perform $m \geq 1$ independent trials. Each trial results in one (and only one) of n distinct category outcomes:

			Probability
		Category 1	p_1
	$\nearrow$	Category 2	p_2
Trial outcome	$\longrightarrow$	Category 3	p_3
	$\searrow$	$\vdots$	$\vdots$
		Category n	p_n

The probabilities $p_1, p_2, ..., p_n$ do not change from trial to trial and $\sum_{i=1}^n p_i = 1$. Define

$$X_i = \text{ number of outcomes in Category } i \text{ (out of } m \text{ trials)}.$$

We call $\mathbf{X} = (X_1, X_2, ..., X_n)'$ a **multinomial random vector**. The joint pmf of $\mathbf{X}$ is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{m!}{x_1!x_2!\cdots x_n!}\, p_1^{x_1}p_2^{x_2}\cdots p_n^{x_n},$$

for values of $\mathbf{x} \in \mathcal{A}$, where $\mathcal{A} = \{(x_1, x_2, ..., x_n) : x_i = 0, 1, 2, ..., m; \sum_{i=1}^n x_i = m\}$. We write $\mathbf{X} \sim \text{mult}(m, \mathbf{p}; \sum_{i=1}^n p_i = 1)$. The parameter $\mathbf{p} = (p_1, p_2, ..., p_n)$ is an n-dimensional vector. However, because $\sum_{i=1}^n p_i = 1$, only $n-1$ of these parameters are "free to vary."

Theorem 4.6.4 (Multinomial Theorem). Let m and n be positive integers, and consider the set $\mathcal{A}$ defined above. For any numbers $p_1, p_2, ..., p_n$,

$$(p_1 + p_2 + \cdots + p_n)^m = \sum_{\mathbf{x}\in\mathcal{A}} \frac{m!}{x_1!x_2!\cdots x_n!}\, p_1^{x_1}p_2^{x_2}\cdots p_n^{x_n}.$$

This is a generalization of the binomial theorem. Clearly, the multinomial pmf sums to one.

MGF: The mgf of $\mathbf{X} \sim \text{mult}(m, \mathbf{p}; \sum_{i=1}^n p_i = 1)$ is

$$M_{\mathbf{X}}(\mathbf{t}) = (p_1e^{t_1} + p_2e^{t_2} + \cdots + p_ne^{t_n})^m,$$

where $\mathbf{t} = (t_1, t_2, ..., t_n)'$.

Proof. The mgf is

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) = E(e^{\mathbf{t}'\mathbf{X}}) &= E\left(e^{t_1X_1+t_2X_2+\cdots+t_nX_n}\right) \\ &= \sum_{\mathbf{x}\in\mathcal{A}} e^{t_1x_1+t_2x_2+\cdots+t_nx_n}\, \frac{m!}{x_1!x_2!\cdots x_n!}\, p_1^{x_1}p_2^{x_2}\cdots p_n^{x_n} \\ &= \sum_{\mathbf{x}\in\mathcal{A}} \frac{m!}{x_1!x_2!\cdots x_n!}\, (p_1e^{t_1})^{x_1}(p_2e^{t_2})^{x_2}\cdots(p_ne^{t_n})^{x_n}, \end{aligned}$$

which is the multinomial expansion of $(p_1e^{t_1} + p_2e^{t_2} + \cdots + p_ne^{t_n})^m$. □

Result: If $\mathbf{X} \sim \text{mult}(m, \mathbf{p}; \sum_{i=1}^{n} p_i = 1)$, then $X_i \sim b(m, p_i)$, $i = 1, 2, ..., n$. That is, the category counts $X_1, X_2, ..., X_n$ have marginal binomial distributions.
Proof. The mgf of X_i is

$$M_{X_i}(t_i) = M_{\mathbf{X}}(0, ..., 0, t_i, 0, ..., 0) = (q_i + p_i e^{t_i})^m,$$

where $q_i = \sum_{j \neq i} p_j = 1 - p_i$. We recognize $M_{X_i}(t_i)$ as the mgf of $X_i \sim b(m, p_i)$. Note that this implies

$$\begin{aligned} E(X_i) &= mp_i \\ \text{var}(X_i) &= mp_i(1 - p_i). \end{aligned}$$

Result: If $\mathbf{X} \sim \text{mult}(m, \mathbf{p}; \sum_{i=1}^{n} p_i = 1)$, then $(X_i, X_j)' \sim \text{trinomial}(m, p_i, p_j, 1 - p_i - p_j)$.

Trinomial Framework

			Probability
		Category i	p_i
Trial outcome	$\longrightarrow$	Category j	p_j
		Neither	$1 - p_i - p_j$

Proof. The mgf of $(X_i, X_j)'$ is

$$M_{X_i, X_j}(t_i, t_j) = M_{\mathbf{X}}(0, ..., 0, t_i, 0, ..., 0, t_j, 0, ..., 0) = (q_{ij} + p_i e^{t_i} + p_j e^{t_j})^m,$$

where $q_{ij} = \sum_{k \neq i,j} p_k = 1 - p_i - p_j$. We recognize $M_{X_i, X_j}(t_i, t_j)$ as the mgf of a trinomial$(m, p_i, p_j, 1 - p_i - p_j)$ distribution. Also,

$$\begin{aligned} E(X_i X_j) &= \frac{\partial^2}{\partial t_i \partial t_j} M_{X_i, X_j}(t_i, t_j)\Big|_{t_i = t_j = 0} \\ &= \frac{\partial^2}{\partial t_i \partial t_j} (q_{ij} + p_i e^{t_i} + p_j e^{t_j})^m \Big|_{t_i = t_j = 0} = m(m-1)p_i p_j. \end{aligned}$$

Therefore, for $i \neq j$,

$$\begin{aligned} \text{cov}(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\ &= m(m-1)p_i p_j - (mp_i)(mp_j) \\ &= -mp_i p_j. \end{aligned}$$

Summary: The mean and variance-covariance matrix of $\mathbf{X} \sim \text{mult}(m, \mathbf{p}; \sum_{i=1}^{n} p_i = 1)$ are

$$\boldsymbol{\mu} = E(\mathbf{X}) = \begin{pmatrix} mp_1 \\ mp_2 \\ \vdots \\ mp_n \end{pmatrix} = m\mathbf{p}$$

and

$$\boldsymbol{\Sigma} = \text{cov}(\mathbf{X}) = \begin{pmatrix} mp_1(1-p_1) & -mp_1p_2 & \cdots & -mp_1p_n \\ -mp_2p_1 & mp_2(1-p_2) & \cdots & -mp_2p_n \\ \vdots & \vdots & \ddots & \vdots \\ -mp_np_1 & -mp_np_2 & \cdots & mp_n(1-p_n) \end{pmatrix} = m[\text{diag}(\mathbf{p}) - \mathbf{pp}'],$$

respectively. Note that $\boldsymbol{\Sigma}$ is symmetric.

Remark: We now generalize many of the definitions/results we presented for bivariate distributions (with $n = 2$) to n-variate distributions. Proving the general results are natural extensions of the $n = 2$ results (so we will avoid).

Definition: Suppose $X_1, X_2, ..., X_n$ are random variables. Let $f_{X_i}(x_i)$ denote the marginal pdf (pmf) of X_i. The random variables $X_1, X_2, ..., X_n$ are **mutually independent** if

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} f_{X_i}(x_i) \\ &= f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n), \end{aligned}$$

for all $\mathbf{x} \in \mathbb{R}^n$; i.e., the joint pdf (pmf) $f_{\mathbf{X}}(\mathbf{x})$ factors into the product of the marginal pdfs (pmfs). Mutual independence implies **pairwise independence**, which requires only that $f_{X_i,X_j}(x_i, x_j) = f_{X_i}(x_i)f_{X_j}(x_j)$ for each $i \neq j$. The opposite is not true.

Result: The random variables $X_1, X_2, ..., X_n$ are mutually independent if and only if

$$\begin{aligned} F_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} F_{X_i}(x_i) \\ &= F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n), \end{aligned}$$

for all $\mathbf{x} \in \mathbb{R}^n$; i.e., the joint cdf $F_{\mathbf{X}}(\mathbf{x})$ factors into the product of the marginal cdfs.

Result: The random variables $X_1, X_2, ..., X_n$ are mutually independent if and only if

$$\begin{aligned} M_{\mathbf{X}}(\mathbf{t}) &= \prod_{i=1}^{n} M_{X_i}(t_i) \\ &= M_{X_1}(t_1)M_{X_2}(t_2)\cdots M_{X_n}(t_n), \end{aligned}$$

for all $t_i \in \mathbb{R}$ where these mgfs exist; that is, the joint mgf $M_{\mathbf{X}}(\mathbf{t})$ factors into the product of the marginal mgfs.

Theorem 4.6.6. Suppose $X_1, X_2, ..., X_n$ are mutually independent. Suppose $g_1, g_2, ..., g_n$ are real functions; i.e., $g_i : \mathbb{R} \to \mathbb{R}$, where g_i is a function of x_i only, $i = 1, 2, ..., n$. Then

$$E\left[\prod_{i=1}^{n} g_i(X_i)\right] = \prod_{i=1}^{n} E[g_i(X_i)],$$

that is, the expectation of the product is the product of the marginal expectations.

Theorem 4.6.7. Suppose $X_1, X_2, ..., X_n$ are mutually independent random variables. Suppose the marginal mgf of X_i is $M_{X_i}(t)$, for $i = 1, 2, ..., n$. The mgf of the sum

$$Z = X_1 + X_2 + \cdots + X_n$$

is given by

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t),$$

that is, the mgf of the sum is the product of the marginal mgfs.

Special case: If, in addition to being mutually independent, the random variables $X_1, X_2, ..., X_n$ also have the same (identical) distribution, characterized by the common mgf $M_X(t)$, then

$$M_Z(t) = \prod_{i=1}^{n} M_X(t) = [M_X(t)]^n.$$

Random variables $X_1, X_2, ..., X_n$ that are mutually independent and have the same distribution are said to be "**iid**," which is an acronym for "independent and identically distributed."

Remark: Theorem 4.6.7 (and its special case) makes getting the distribution of the **sum** of mutually independent random variables very easy. The (unique) distribution identified by $M_Z(t)$ is the answer.

Example 4.23. Suppose that $X_1, X_2, ..., X_n$ are iid Poisson(λ), where $\lambda > 0$. The mgf of the sum

$$Z = X_1 + X_2 + \cdots + X_n$$

is given by

$$M_Z(t) = [M_X(t)]^n = [e^{\lambda(e^t-1)}]^n = e^{n\lambda(e^t-1)},$$

which we recognize as the mgf of a Poisson distribution with mean $n\lambda$. Because mgfs are unique, we know that $Z \sim$ Poisson($n\lambda$).

Example 4.24. Suppose $X_1, X_2, ..., X_n$ are mutually independent, where $X_i \sim$ gamma(α_i, β), where $\alpha_i, \beta > 0$. Note that the X_i's are not iid because they have different marginal distributions (i.e., the shape parameters are potentially different). The mgf of the sum

$$Z = X_1 + X_2 + \cdots + X_n$$

is given by

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \left(\frac{1}{1-\beta t}\right)^{\alpha_i} = \left(\frac{1}{1-\beta t}\right)^{\sum_{i=1}^{n}\alpha_i},$$

which we recognize as the mgf of a gamma distribution with shape parameter $\sum_{i=1}^{n}\alpha_i$ and scale parameter β. Because mgfs are unique, we know that $Z \sim$ gamma($\sum_{i=1}^{n}\alpha_i, \beta$).

Remark: The result in Example 4.24 has important special cases:

1. $\alpha_i = 1$, for $i = 1, 2, ..., n$. In this case, $X_1, X_2, ..., X_n$ are iid exponential(β) and
$$Z = \sum_{i=1}^{n} X_i \sim \text{gamma}(n, \beta).$$

2. $\alpha_i = p_i/2$, where $p_i > 0$, and $\beta = 2$. In this case, $X_1, X_2, ..., X_n$ are mutually independent, $X_i \sim \chi^2_{p_i}$, and
$$Z = \sum_{i=1}^{n} X_i \sim \text{gamma}\left(\frac{p}{2}, 2\right) \stackrel{d}{=} \chi^2_p,$$
where $p = \sum_{i=1}^{n} p_i$; i.e., "the degrees of freedom add."

Example 4.25. Suppose $X_1, X_2, ..., X_n$ are mutually independent, where $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, for $i = 1, 2, ..., n$. Consider the linear combination
$$Z = \sum_{i=1}^{n} a_i X_i = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n,$$
where $a_i \in \mathbb{R}$. Then
$$Z \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$
In other words, linear combinations of (mutually independent) normal random variables are also normally distributed.

Proof. The mgf of Z is
$$\begin{aligned} M_Z(t) = E(e^{tZ}) &= E[e^{t(a_1X_1 + a_2X_2 + \cdots + a_nX_n)}] \\ &= E(e^{a_1tX_1} e^{a_2tX_2} \cdots e^{a_ntX_n}) \\ &\stackrel{\text{indep}}{=} E(e^{a_1tX_1}) E(e^{a_2tX_2}) \cdots E(e^{a_ntX_n}) \\ &= M_{X_1}(a_1t) M_{X_2}(a_2t) \cdots M_{X_n}(a_nt) \\ &= \prod_{i=1}^{n} \exp[\mu_i a_i t + (a_i t)^2 \sigma_i^2/2] \\ &= \exp\left[\left(\sum_{i=1}^{n} a_i \mu_i\right) t + \left(\sum_{i=1}^{n} a_i^2 \sigma_i^2\right) t^2/2\right], \end{aligned}$$
which we recognize as the mgf of a normal distribution with mean $\sum_{i=1}^{n} a_i \mu_i$ and variance $\sum_{i=1}^{n} a_i^2 \sigma_i^2$. Because mgfs are unique, the result follows. □

Remark: In the last example, Z remains normally distributed even when the X_i's are not mutually independent (the only thing that potentially changes is the variance of Z).

Linear Combinations: Suppose $X_1, X_2, ..., X_n$ are random variables with (marginal) means $E(X_i)$ and (marginal) variances $\text{var}(X_i)$, for $i = 1, 2, ..., n$. Consider the linear combination

$$Z = \sum_{i=1}^{n} a_i X_i = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n.$$

The mean of Z is

$$\begin{aligned} E(Z) &= E(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) \\ &= a_1 E(X_1) + a_2 E(X_2) + \cdots + a_n E(X_n) = \sum_{i=1}^{n} a_i E(X_i). \end{aligned}$$

The variance of Z is

$$\begin{aligned} \text{var}(Z) &= \text{var}(a_1 X_1 + a_2 X_2 + \cdots + a_n X_n) \\ &= \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + \sum_{i \neq j} a_i a_j \text{cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + 2 \sum_{i<j} a_i a_j \text{cov}(X_i, X_j). \end{aligned}$$

Theorem 4.6.11. Suppose $\mathbf{X} = (X_1, X_2, ..., X_n)$ is an n-dimensional random vector with joint pdf (pmf) $f_{\mathbf{X}}(\mathbf{x})$. The random variables $X_1, X_2, ..., X_n$ are mutually independent if and only if there exists functions $g_1(x_1)$, $g_2(x_2)$, ..., $g_n(x_n)$ such that

$$f_{\mathbf{X}}(\mathbf{x}) = g_1(x_1) g_2(x_2) \cdots g_n(x_n),$$

for all $\mathbf{x} \in \mathbb{R}^n$. This is a generalization of Lemma 4.2.7 for n-dimensional random vectors.

Theorem 4.6.12. Suppose the random variables $X_1, X_2, ..., X_n$ are mutually independent. The random variables $U_1 = g_1(X_1)$, $U_2 = g_2(X_2)$, ..., $U_n = g_n(X_n)$ are also mutually independent. This is a generalization of Theorem 4.3.5 for n-dimensional random vectors.

Multivariate Transformations

Setting: Suppose $\mathbf{X} = (X_1, X_2, ..., X_n)$ is a continuous random vector with joint pdf $f_{\mathbf{X}}(\mathbf{x})$ and support $\mathcal{A} \subseteq \mathbb{R}^n$. Define

$$\begin{aligned} U_1 &= g_1(X_1, X_2, ..., X_n) \\ U_2 &= g_2(X_1, X_2, ..., X_n) \\ &\vdots \\ U_n &= g_n(X_1, X_2, ..., X_n). \end{aligned}$$

Assume that this is a one-to-one transformation from $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^n : f_{\mathbf{X}}(\mathbf{x}) > 0\}$ to

$$\mathcal{B} = \{\mathbf{u} \in \mathbb{R}^n : u_i = g_i(\mathbf{x}),\ i = 1, 2, ..., n;\ \mathbf{x} \in \mathcal{A}\},$$

the support of $\mathbf{U} = (U_1, U_2, ..., U_n)$. Because the transformation is one-to-one (by assumption), the inverse transformation is

$$\begin{aligned} x_1 &= g_1^{-1}(u_1, u_2, ..., u_n) \\ x_2 &= g_2^{-1}(u_1, u_2, ..., u_n) \\ &\vdots \\ x_n &= g_n^{-1}(u_1, u_2, ..., u_n). \end{aligned}$$

With $\mathbf{u} = (u_1, u_2, ..., u_n)$, the Jacobian of the inverse transformation is

$$J = \det \begin{pmatrix} \frac{\partial g_1^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_1^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial g_1^{-1}(\mathbf{u})}{\partial u_n} \\ \frac{\partial g_2^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_2^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial g_2^{-1}(\mathbf{u})}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_n^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_n^{-1}(\mathbf{u})}{\partial u_2} & \cdots & \frac{\partial g_n^{-1}(\mathbf{u})}{\partial u_n} \end{pmatrix};$$

i.e., J is the determinant of this $n \times n$ matrix of partial derivatives. Provided that $J \neq 0$ over $\mathcal{B}$, the pdf of $\mathbf{U} = (U_1, U_2, ..., U_n)$, where nonzero, is

$$f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{X}}(g_1^{-1}(\mathbf{u}), g_2^{-1}(\mathbf{u}), ..., g_n^{-1}(\mathbf{u}))|J|.$$

This generalizes our discussion on bivariate ($n = 2$) transformations in Section 4.3.

Example 4.26. Suppose X_1, X_2, and X_3 have the joint pdf

$$f_{\mathbf{X}}(x_1, x_2, x_3) = 48x_1x_2x_3\ I(0 < x_1 < x_2 < x_3 < 1).$$

Note that the support of $\mathbf{X} = (X_1, X_2, X_3)$ is $\mathcal{A} = \{\mathbf{x} \in \mathbb{R}^3 : 0 < x_1 < x_2 < x_3 < 1\}$, the upper orthant of the unit cube in $\mathbb{R}^3$. Define

$$\begin{aligned} U_1 &= g_1(X_1, X_2, X_3) = \frac{X_1}{X_2} \\ U_2 &= g_2(X_1, X_2, X_3) = \frac{X_2}{X_3} \\ U_3 &= g_3(X_1, X_2, X_3) = X_3. \end{aligned}$$

This defines a one-to-one transformation from $\mathcal{A}$ to

$$\mathcal{B} = \{\mathbf{u} \in \mathbb{R}^3 : 0 < u_1 < 1, 0 < u_2 < 1, 0 < u_3 < 1\}.$$

The inverse transformation is

$$\begin{aligned} x_1 &= g_1^{-1}(u_1, u_2, u_3) = u_1u_2u_3 \\ x_2 &= g_2^{-1}(u_1, u_2, u_3) = u_2u_3 \\ x_3 &= g_3^{-1}(u_1, u_2, u_3) = u_3 \end{aligned}$$

and the Jacobian is

$$J = \det \begin{pmatrix} \frac{\partial g_1^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_1^{-1}(\mathbf{u})}{\partial u_2} & \frac{\partial g_1^{-1}(\mathbf{u})}{\partial u_3} \\ \frac{\partial g_2^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_2^{-1}(\mathbf{u})}{\partial u_2} & \frac{\partial g_2^{-1}(\mathbf{u})}{\partial u_3} \\ \frac{\partial g_3^{-1}(\mathbf{u})}{\partial u_1} & \frac{\partial g_3^{-1}(\mathbf{u})}{\partial u_2} & \frac{\partial g_3^{-1}(\mathbf{u})}{\partial u_3} \end{pmatrix} = \det \begin{pmatrix} u_2u_3 & u_1u_3 & u_1u_2 \\ 0 & u_3 & u_2 \\ 0 & 0 & 1 \end{pmatrix} = u_2u_3^2,$$

which is never equal to zero over $\mathcal{B}$. Therefore, the joint pdf of $\mathbf{U} = (U_1, U_2, U_3)$, for $\mathbf{u} \in \mathcal{B}$, is given by

$$\begin{aligned} f_{\mathbf{U}}(u_1, u_2, u_3) &= f_{\mathbf{X}}(g_1^{-1}(\mathbf{u}), g_2^{-1}(\mathbf{u}), g_3^{-1}(\mathbf{u}))|J| \\ &= 48(u_1u_2u_3)(u_2u_3)(u_3) \times u_2u_3^2 \\ &= 48u_1u_2^3u_3^5. \end{aligned}$$

Notice that we can write

$$\begin{aligned} f_{\mathbf{U}}(u_1, u_2, u_3) &= 48u_1u_2^3u_3^5\ I(0 < u_1 < 1, 0 < u_2 < 1, 0 < u_3 < 1) \\ &= 2u_1I(0 < u_1 < 1)\ \ 4u_2^3I(0 < u_2 < 1)\ \ 6u_3^5I(0 < u_3 < 1) \\ &= f_{U_1}(u_1)f_{U_2}(u_2)f_{U_3}(u_3). \end{aligned}$$

We see that $U_1 \sim \text{beta}(2, 1)$, $U_2 \sim \text{beta}(4, 1)$, and $U_3 \sim \text{beta}(6, 1)$. Also, U_1, U_2, and U_3 are mutually independent.

4.7 Inequalities

Remark: This section is divided into two parts. Section 4.7.1 presents numerical inequalities; Section 4.7.2 presents functional inequalities. We highlight one of each.

Hölder's Inequality: Suppose X and Y are random variables, and let p and q be constants that satisfy

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$|E(XY)| \leq E(|XY|) \leq [E(|X|^p)]^{1/p}[E(|Y|^q)]^{1/q}.$$

Proof. See CB (pp 186-187).

Note: The most important special case of Hölder's Inequality arises when $p = q = 2$:

$$|E(XY)| \leq E(|XY|) \leq [E(X^2)]^{1/2}[E(Y^2)]^{1/2}.$$

This is called the **Cauchy-Schwarz Inequality**.

Application: In the Cauchy-Schwarz Inequality, if we replace X with $X - \mu_X$ and Y with $Y - \mu_Y$, we get

$$|E[(X-\mu_X)(Y-\mu_Y)]| \leq \{E[(X-\mu_X)^2]\}^{1/2}\{E[(Y-\mu_Y)^2]\}^{1/2}.$$

Squaring both sides, we get

$$[\text{cov}(X,Y)]^2 \leq \sigma_X^2\sigma_Y^2.$$

This is called the **covariance inequality**.

Note: From the covariance inequality, it follows immediately than $-1 \leq \rho_{XY} \leq 1$. Using Cauchy-Schwarz is far easier than how we proved it in Theorem 4.5.7.

Jensen's Inequality: Suppose X is a random variable and suppose $g : \mathbb{R} \to \mathbb{R}$ is a convex function. Then

$$E[g(X)] \geq g(E(X)).$$

Remark: See Definition 4.7.6 for a general definition of convexity. If g is twice differentiable, then g is convex if $g''(x) \geq 0$ for all x. If g is strictly convex, then the inequality is strict. In the proof below, I will assume that g is twice differentiable. This assumption is not needed; see CB (pp 190) for a more general proof.

Proof. Expand $g(x)$ in a Taylor series about $\mu = E(X)$ of order two; i.e.,

$$g(x) = g(\mu) + g'(\mu)(x-\mu) + \frac{g''(\xi)}{2}(x-\mu)^2,$$

where ξ is between x and μ (a consequence of the Mean Value Theorem). Note that

$$\frac{g''(\xi)}{2}(x-\mu)^2 \geq 0$$

because g is convex by assumption. Therefore,

$$g(x) \geq g(\mu) + g'(\mu)(x-\mu)$$

and, taking expectations,

$$E[g(X)] \geq E[g(\mu) + g'(\mu)(X-\mu)] = g(\mu) + g'(\mu)\underbrace{E(X-\mu)}_{=0} = g(E(X)). \ \square$$

Application: Suppose X is a random variable with finite second moment; i.e., $E(X^2) < \infty$. Note that $g(x) = x^2$ is a convex function because $g''(x) = 2 > 0$, for all x. Therefore, Jensen's Inequality says that

$$E[g(X)] = E(X^2) \geq [E(X)]^2 = g(E(X)).$$

Of course, we already know this because $\text{var}(X) = E(X^2) - [E(X)]^2 \geq 0$.

Note: If g is concave, then $-g$ is convex. (If g is twice differentiable, then this is obvious). Therefore, if g is concave, the inequality switches:

$$E[g(X)] \leq g(E(X)).$$

For example, consider $g(x) = \ln x$, which is concave because $g''(x) = -1/x^2 < 0$, for all x. Therefore, assuming that all expectations exist, we have

$$E[g(X)] = E(\ln X) \leq \ln(E(X)) = g(E(X)).$$

5 Properties of a Random Sample

Complementary reading: Chapter 5 (CB). Sections 5.1-5.5.

5.1 Basic Concepts of a Random Sample

Definition: The random variables $X_1, X_2, ..., X_n$ are called a **random sample** from the population $f_X(x)$ if

1. $X_1, X_2, ..., X_n$ are **mutually independent**
2. The marginal pdf (pmf) of each X_i is the **same** function $f_X(x)$.

Alternatively, we say that

$$\text{“}X_1, X_2, ..., X_n \text{ are iid from } f_X(x)\text{.”}$$

The acronym “iid” is short for “**i**ndependent and **i**dentically **d**istributed.” The function $f_X(x)$ is called the **population distribution** because it is the distribution that describes the population from which the X_i's are “drawn.”

Conceptualization: Consider an experiment that is repeated n times, independently and under identical conditions. Each time the experiment is performed, you observe an X_i whose distribution is described by $f_X(x)$. Performing the experiment n times yields $X_1, X_2, ..., X_n$.

Result: If $X_1, X_2, ..., X_n$ are iid from $f_X(x)$, the joint pdf (pmf) of $\mathbf{X} = (X_1, X_2, ..., X_n)$ is

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i) = f_X(x_1) f_X(x_2) \cdots f_X(x_n).$$

This follows immediately from Definition 4.6.5 (CB, pp 182).

Example 5.1. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Here, the $\mathcal{N}(\mu, \sigma^2)$ population distribution is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} I(x \in \mathbb{R}).$$

This is the (marginal) pdf of each X_i. The joint pdf of $\mathbf{X} = (X_1, X_2, ..., X_n)$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} I(x_i \in \mathbb{R}) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} I(\mathbf{x} \in \mathbb{R}^n). \end{aligned}$$

Remark: This function, when viewed as a function of $x_1, x_2, ..., x_n$, describes probabilistically the (joint) random behavior of $X_1, X_2, ..., X_n$. Later on (when we start thinking about estimation), we will begin to regard $f_{\mathbf{X}}(\mathbf{x})$ not as a function of $x_1, x_2, ..., x_n$ but instead as a function of $\boldsymbol{\theta} = (\mu, \sigma^2)'$ with the x_i's held fixed. This will give rise to what we call a **likelihood function**.

Discussion: Under an iid sampling model, calculations involving the joint distribution of $\mathbf{X} = (X_1, X_2, ..., X_n)$ are greatly simplified. Suppose that in Example 5.1 we wanted to calculate the probability that each X_i exceeded the mean μ by two standard deviations; i.e.,

$$P(X_1 > \mu + 2\sigma, X_2 > \mu + 2\sigma, ..., X_n > \mu + 2\sigma).$$

From first principles, this probability could be found by taking $f_{\mathbf{X}}(\mathbf{x})$ and integrating it over the set $B = \{\mathbf{x} \in \mathbb{R}^n : x_1 > \mu + 2\sigma, x_2 > \mu + 2\sigma, ..., x_n > \mu + 2\sigma\}$, that is, by calculating

$$\int_{\mu+2\sigma}^{\infty} \int_{\mu+2\sigma}^{\infty} \cdots \int_{\mu+2\sigma}^{\infty} \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} dx_1 dx_2 \cdots dx_n.$$

This is a tedious calculation to make directly using the joint distribution $f_{\mathbf{X}}(\mathbf{x})$. However, note that we can write

$$\begin{aligned} &P(X_1 > \mu + 2\sigma, X_2 > \mu + 2\sigma, ..., X_n > \mu + 2\sigma) \\ &\quad = P(X_1 > \mu + 2\sigma)P(X_2 > \mu + 2\sigma) \cdots P(X_n > \mu + 2\sigma) && \text{(mutually independent)} \\ &\quad = [P(X_1 > \mu + 2\sigma)]^n, && \text{(identically distributed)} \end{aligned}$$

which is much easier. We have reduced an n-fold integral calculation to a single integral calculation from one marginal distribution. Of course, we pay a price for this simplicity−we have to make strong assumptions about the stochastic behavior of $X_1, X_2, ..., X_n$.

Remark: When we say "random sample," we essentially mean that we are sampling from an infinite population. When sampling from a **finite** population, say, $\{x_1, x_2, ..., x_N\}$, where $N < \infty$ denotes the size of the population, we can sample in two ways:

1. SRSWR (simple random sample with replacement).
   - When sampling **with replacement**, the value x_i is "replaced" after it is selected (e.g., think of drawing numbered balls out of a hat; each time you draw a ball and observe it, you put it back in the hat).
   - In this case, each X_i has the same discrete uniform distribution, with probability $1/N$ attached to each of $x_1, x_2, ..., x_N$. The X_i's are also mutually independent because the process of choosing each x_i is the same. In other words, $X_1, X_2, ..., X_n$ remain iid.
   - This type of sampling model forms the basis for the statistical (re)-sampling technique known as *bootstrapping*.

2. SRSWOR (simple random sample without replacement).

   - When sampling **without replacement**, the value x_i is not replaced after it is selected (e.g., after you draw a ball and observe it, you do not put it back).
   - In this case, the X_i's are no longer mutually independent. To see why, note that

$$\begin{aligned} P(X_1 = x_1) &= \frac{1}{N} \\ P(X_2 = x_1 | X_1 = x_1) &= 0. \end{aligned}$$

     Therefore, X_1 and X_2 are not independent.
   - Interestingly, the X_i's remain identically distributed; see CB (pp 210).
   - Heuristically, if the population size N is "large," this type of sampling closely approximates sampling from an infinite population; see Example 5.1.3 (CB, pp 210-211).

Disclaimer: In this course, unless otherwise noted, we will regard $X_1, X_2, ..., X_n$ as iid.

5.2 Sums of Random Variables from a Random Sample

Definition: Suppose $X_1, X_2, ..., X_n$ is a random sample. A **statistic** T is a function of $X_1, X_2, ..., X_n$, that is,

$$T = T(\mathbf{X}) = T(X_1, X_2, ..., X_n).$$

The only restriction is that a statistic T cannot depend on unknown parameters.

Examples: Each of the following satisfies the definition of a statistic:

1. Sample mean: $T(\mathbf{X}) = \overline{X}$, where

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

2. Sample variance: $T(\mathbf{X}) = S^2$, where

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

3. Minimum order statistic: $T(\mathbf{X}) = X_{(1)} = \min\{X_1, X_2, ..., X_n\}$
4. Sample range: $T(\mathbf{X}) = X_{(n)} - X_{(1)}$.

Note: The definition of a statistic is very broad. For example, even something like

$$T(\mathbf{X}) = \ln(S^2 + 4) - 12.08e^{-\tan\left(\sum_{i=1}^{n} |X_i^3|\right)}$$

satisfies the definition. In addition, if μ and σ^2 are unknown, then

- $\overline{X}$ is a statistic, but $\overline{X} - \mu$ is not.
- S^2 is a statistic, but S^2/σ^2 is not.

Definition: Suppose $X_1, X_2, ..., X_n$ is an iid sample from $f_X(x)$. Suppose that $T = T(\mathbf{X})$ is a statistic. The probability distribution of T is called its **sampling distribution**.

Revelation: Because $T = T(\mathbf{X})$, a function of $X_1, X_2, ..., X_n$, the statistic T is itself a random variable (or a random vector if T is vector-valued). Therefore, T has its own distribution! This distribution is called the sampling distribution of T. In notation,

$$\begin{aligned} X_1, X_2, ..., X_n &\sim f_X(x) \longleftarrow \text{population distribution} \\ T = T(X_1, X_2, ..., X_n) &\sim f_T(t) \longleftarrow \text{sampling distribution of } T \end{aligned}$$

Common goals: For a statistic $T = T(\mathbf{X})$, we may want to find its pdf (pmf) $f_T(t)$, its cdf $F_T(t)$, or perhaps its mgf $M_T(t)$. These functions identify the distribution of T. We might also want to calculate $E(T)$ or $\text{var}(T)$. These quantities describe characteristics of T's distribution.

Lemma: Suppose $X_1, X_2, ..., X_n$ are iid with $E(X) = \mu$ and $\text{var}(X) = \sigma^2 < \infty$. Then

$$\begin{aligned} E\left(\sum_{i=1}^{n} X_i\right) &= nE(X_1) = n\mu \\ \text{var}\left(\sum_{i=1}^{n} X_i\right) &= n\text{var}(X_1) = n\sigma^2. \end{aligned}$$

Proof. Exercise. Compare this result with Lemma 5.2.5 (CB, pp 213), which is slightly more general.

Theorem 5.2.6. Suppose $X_1, X_2, ..., X_n$ are iid with $E(X) = \mu$ and $\text{var}(X) = \sigma^2 < \infty$. Then

(a) $E(\overline{X}) = \mu$

(b) $\text{var}(\overline{X}) = \sigma^2/n$

(c) $E(S^2) = \sigma^2$.

Proof. To prove (a) and (b), just use the last lemma. We have

$$\begin{aligned} E(\overline{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}(n\mu) = \mu \\ \text{var}(\overline{X}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}. \end{aligned}$$

To prove part (c), first note that

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

Therefore,

$$\begin{aligned} E(S^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] &= \frac{1}{n-1}\, E\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \\ &= \frac{1}{n-1}\left[E\left(\sum_{i=1}^{n} X_i^2\right) - nE(\overline{X}^2)\right]. \end{aligned}$$

Now,

$$E\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} E(X_i^2) = \sum_{i=1}^{n}\{\text{var}(X_i) + [E(X_i)]^2\} = \sum_{i=1}^{n}(\sigma^2 + \mu^2) = n(\sigma^2 + \mu^2)$$

and

$$E(\overline{X}^2) = \text{var}(\overline{X}) + [E(\overline{X})]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Therefore,

$$E(S^2) = \frac{1}{n-1}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] = \sigma^2. \ \square$$

Curiosity: How would we find $\text{var}(S^2)$? This is in general a much harder calculation.

Theorem 5.2.7. Suppose $X_1, X_2, ..., X_n$ are iid with moment generating function (mgf) $M_X(t)$. The mgf of $\overline{X}$ is

$$M_{\overline{X}}(t) = [M_X(t/n)]^n.$$

Proof. The mgf of $\overline{X}$ is

$$\begin{aligned} M_{\overline{X}}(t) = E(e^{t\overline{X}}) = E[e^{\frac{t}{n}(X_1+X_2+\cdots+X_n)}] &= E(e^{\frac{t}{n}X_1}e^{\frac{t}{n}X_2}\cdots e^{\frac{t}{n}X_n}) \\ &\overset{\text{indep}}{=} E(e^{\frac{t}{n}X_1})E(e^{\frac{t}{n}X_2})\cdots E(e^{\frac{t}{n}X_n}) \\ &= M_{X_1}(t/n)M_{X_2}(t/n)\cdots M_{X_n}(t/n) \\ &\overset{\text{ident}}{=} [M_X(t/n)]^n. \ \square \end{aligned}$$

Remark: Theorem 5.2.7 is useful. It allows us to quickly obtain the mgf of $\overline{X}$ (and hence quickly identify its distribution), as the next two examples illustrate.

Example 5.2. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. Recall that the (population) mgf of $X \sim \mathcal{N}(\mu, \sigma^2)$ is

$$M_X(t) = e^{\mu t + \sigma^2 t^2/2},$$

for all $t \in \mathbb{R}$. Therefore,

$$\begin{aligned} M_{\overline{X}}(t) &= [e^{\mu(t/n) + \sigma^2(t/n)^2/2}]^n \\ &= e^{\mu t + (\sigma^2/n)t^2/2}, \end{aligned}$$

which we recognize as the mgf of a $\mathcal{N}(\mu, \sigma^2/n)$ distribution. Because mgfs are unique, we know that $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

Example 5.3. Suppose that $X_1, X_2, ..., X_n$ are iid gamma(α, β). Recall that the (population) mgf of $X \sim$ gamma(α, β) is

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha},$$

for $t < 1/\beta$. Therefore,

$$\begin{aligned} M_{\overline{X}}(t) &= \left\{\left[\frac{1}{1-\beta(t/n)}\right]^{\alpha}\right\}^n \\ &= \left[\frac{1}{1-(\beta/n)t}\right]^{n\alpha}, \end{aligned}$$

for $t < n/\beta$, which we recognize as the mgf of a gamma$(n\alpha, \beta/n)$ distribution. Because mgfs are unique, we know that $\overline{X} \sim$ gamma$(n\alpha, \beta/n)$.

Special case: $X_1, X_2, ..., X_n \sim$ iid exponential$(\beta) \implies \overline{X} \sim$ gamma$(n, \beta/n)$.

Remark: In cases where Theorem 5.2.7 is not useful (e.g., the population mgf does not exist, etc.), the convolution technique can be.

Theorem 5.2.9 (Convolution). If X and Y are independent continuous random variables with marginal pdfs $f_X(x)$ and $f_Y(y)$, respectively, then the pdf of $Z = X + Y$ is

$$f_Z(z) = \int_{\mathbb{R}} f_X(w) f_Y(z-w) dw.$$

The pdf $f_Z(z)$ is called the **convolution** of $f_X(x)$ and $f_Y(y)$.
Proof. Introduce $W = X$ and perform a bivariate transformation:

$$\begin{aligned} w &= g_1(x,y) = x \\ z &= g_2(x,y) = x + y \end{aligned} \quad \Longrightarrow \quad \begin{aligned} x &= g_1^{-1}(w,z) = w \\ y &= g_2^{-1}(w,z) = z - w \end{aligned}$$

The Jacobian of the (inverse) transformation is

$$J = \det \begin{vmatrix} \dfrac{\partial g_1^{-1}(w,z)}{\partial w} & \dfrac{\partial g_1^{-1}(w,z)}{\partial z} \\ \dfrac{\partial g_2^{-1}(w,z)}{\partial w} & \dfrac{\partial g_2^{-1}(w,z)}{\partial z} \end{vmatrix} = \det \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.$$

Therefore, the joint pdf of (W, Z) is given by

$$f_{W,Z}(w,z) = f_{X,Y}(w, z-w) \overset{X \perp\!\!\!\perp Y}{=} f_X(w) f_Y(z-w).$$

Finally,

$$f_Z(z) = \int_{\mathbb{R}} f_X(w) f_Y(z-w) dw$$

as claimed. □

Example 5.4. Suppose that $X \sim \mathcal{U}(0,1)$, $Y \sim \mathcal{U}(0,1)$, and $X \perp\!\!\!\perp Y$. Find the pdf of $Z = X + Y$.

Solution. First, note that the support of Z is $\mathcal{Z} = \{z : 0 < z < 2\}$. The marginal pdfs of X and Y are $f_X(x) = I(0 < x < 1)$ and $f_Y(y) = I(0 < y < 1)$, respectively. Using the convolution formula, the pdf of $Z = X + Y$ is

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} f_X(w) f_Y(z-w) dw \\ &= \int_{\mathbb{R}} I(0 < w < 1) I(0 < z - w < 1) dw \\ &= \int_0^1 I(0 < z - w < 1) dw. \end{aligned}$$

Note that $0 < z - w < 1 \iff z - 1 < w < z$ and also $0 < w < 1$. Therefore, if $0 < z \leq 1$, then

$$f_Z(z) = \int_0^z dw = z.$$

If $1 < z < 2$, then

$$f_Z(z) = \int_{z-1}^1 dw = 2 - z.$$

Therefore, the pdf of $Z = X + Y$ is given by

$$f_Z(z) = \begin{cases} z, & 0 < z \leq 1 \\ 2 - z, & 1 < z < 2 \\ 0, & \text{otherwise.} \end{cases}$$

The pdf $f_Z(z)$, shown in Figure 5.1 (next page), is a member of the **triangular family** of distributions (the name is not surprising).

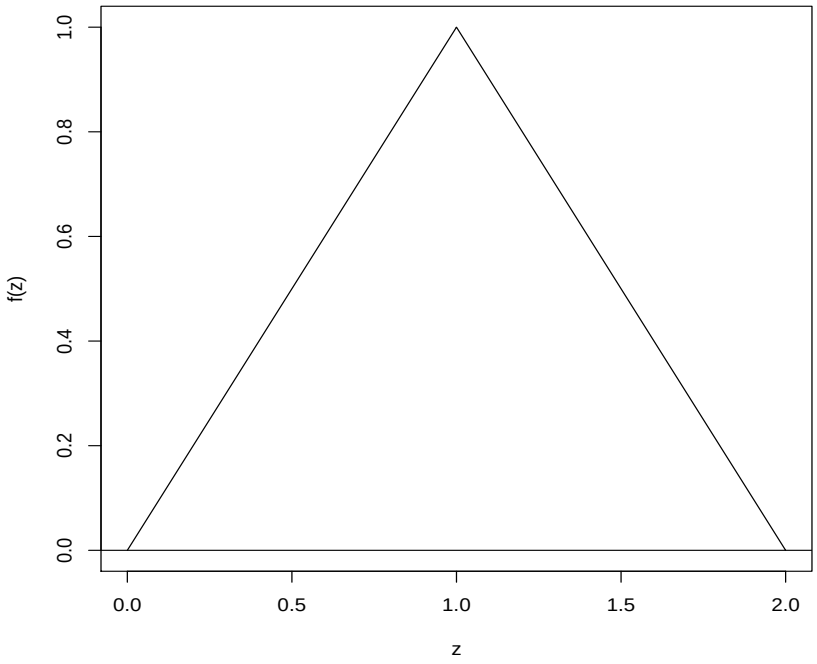


Figure 5.1: The pdf of Z in Example 5.4.

Example 5.5. Suppose that $X \sim \text{Cauchy}(0, \sigma_X)$, $Y \sim \text{Cauchy}(0, \sigma_Y)$, and $X \perp\!\!\!\perp Y$. Find the pdf of $Z = X + Y$.

Solution. First, note that the support of Z is $\mathcal{Z} = \{z : -\infty < z < \infty\}$. The marginal pdfs of X and Y are

$$\begin{aligned} f_X(x) &= \frac{1}{\pi\sigma_X[1 + (\frac{x}{\sigma_X})^2]}\, I(x \in \mathbb{R}) \\ f_Y(y) &= \frac{1}{\pi\sigma_Y[1 + (\frac{y}{\sigma_Y})^2]}\, I(y \in \mathbb{R}). \end{aligned}$$

Using the convolution formula, the pdf of Z, for all $z \in \mathbb{R}$, is

$$\begin{aligned} f_Z(z) &= \int_{\mathbb{R}} f_X(w) f_Y(z-w)dw \\ &= \int_{\mathbb{R}} \frac{1}{\pi\sigma_X[1 + (\frac{w}{\sigma_X})^2]} \frac{1}{\pi\sigma_Y[1 + (\frac{z-w}{\sigma_Y})^2]}\, dw \\ &= \frac{1}{\pi(\sigma_X + \sigma_Y)[1 + (\frac{z}{\sigma_X+\sigma_Y})^2]}, \end{aligned}$$

where the last step follows from using the method of partial fractions (see Exercise 5.7, CB, pp 256). Therefore, it follows that $Z \sim \text{Cauchy}(0, \sigma_X + \sigma_Y)$.

Extension: Suppose that $X_1, X_2, ..., X_n$ are iid Cauchy(μ, σ), where $-\infty < \mu < \infty$ and $\sigma > 0$. What is the sampling distribution of $\overline{X}$?
Solution. The easiest way to answer this would be to use **characteristic functions**. The characteristic function of $X \sim$ Cauchy(μ, σ) is

$$\psi_X(t) = E(e^{itX}) = e^{\mu it - \sigma|t|},$$

where $i = \sqrt{-1}$. Therefore,

$$\psi_{\overline{X}}(t) = [\psi_X(t/n)]^n = [e^{\mu i(t/n) - \sigma|t|/n}]^n = e^{\mu it - \sigma|t|},$$

which we recognize as the characteristic function of the Cauchy(μ, σ) distribution. Therefore,

$$X_1, X_2, ..., X_n \sim \text{iid Cauchy}(\mu, \sigma) \implies \overline{X} \sim \text{Cauchy}(\mu, \sigma).$$

Q: Can we show this without using characteristic functions?
A: Yes, this could be done in three steps.

1. First, argue that

$$Z_1, Z_2, ..., Z_n \sim \text{iid Cauchy}(0, 1) \implies \sum_{i=1}^{n} Z_i \sim \text{Cauchy}(0, n).$$

   This could be done using convolution for $n = 2$. Then use induction.

2. Next, argue that

$$\begin{aligned} f_{\overline{Z}}(z) &= n f_{\sum_{i=1}^n Z_i}(nz) \\ &= \frac{1}{\pi(1+z^2)} I(z \in \mathbb{R}), \end{aligned}$$

   i.e., $\overline{Z} \sim$ Cauchy$(0, 1)$. See Exercise 5.5 (CB, pp 256) to see why the first equality holds.

3. Finally, let $X_i = \sigma Z_i + \mu$, for each $i = 1, 2, ..., n$, so that $\overline{X} = \sigma\overline{Z} + \mu$ (a location-scale transformation). Therefore,

$$f_{\overline{X}}(x) = \frac{1}{\sigma} f_{\overline{Z}}\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\pi\sigma\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]} I(x \in \mathbb{R}),$$

   i.e., $\overline{X} \sim$ Cauchy(μ, σ).

Obviously, the characteristic function argument is much easier. However, to understand characteristic functions, you have to understand complex analysis. As in the words of CB, "you win some and you lose some."

Back to exponential families....

Theorem 5.2.11. Suppose $X_1, X_2, ..., X_n$ are iid with pdf (pmf) in the exponential family; i.e.,

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right\}.$$

Define the statistics

$$\begin{aligned} T_1 = T_1(\mathbf{X}) &= \sum_{j=1}^{n} t_1(X_j) \\ T_2 = T_2(\mathbf{X}) &= \sum_{j=1}^{n} t_2(X_j) \\ &\vdots \\ T_k = T_k(\mathbf{X}) &= \sum_{j=1}^{n} t_k(X_j) \end{aligned}$$

and set $\mathbf{T} = (T_1, T_2, ..., T_k)$, a k-dimensional statistic. If $f_X(x|\boldsymbol{\theta})$ is a full exponential family; i.e., if

$$d = \dim(\boldsymbol{\theta}) = k,$$

then

$$f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\theta}) = H(\mathbf{t})[c(\boldsymbol{\theta})]^n \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i\right\}.$$

That is, $\mathbf{T}$ has pdf (pmf) in the exponential family as well.

Example 5.6. Suppose $X_1, X_2, ..., X_n$ are iid gamma(α, β); i.e., the population pdf is

$$\begin{aligned} f_X(x|\boldsymbol{\theta}) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} I(x > 0) \\ &= \frac{I(x > 0)}{x} \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp\left(\alpha \ln x - \frac{x}{\beta}\right) \\ &= h(x)c(\boldsymbol{\theta}) \exp\{w_1(\boldsymbol{\theta})t_1(x) + w_2(\boldsymbol{\theta})t_2(x)\}, \end{aligned}$$

where $\boldsymbol{\theta} = (\alpha, \beta)'$, $h(x) = I(x > 0)/x$, $c(\boldsymbol{\theta}) = [\Gamma(\alpha)\beta^\alpha]^{-1}$, $w_1(\boldsymbol{\theta}) = \alpha$, $t_1(x) = \ln x$, $w_2(\boldsymbol{\theta}) = -1/\beta$, and $t_2(x) = x$. Note that this is a full exponential family with $d = k = 2$. Theorem 5.2.11 says that $\mathbf{T} = (T_1, T_2)$, where

$$T_1 = T_1(\mathbf{X}) = \sum_{j=1}^{n} \ln X_j \quad \text{and} \quad T_2 = T_2(\mathbf{X}) = \sum_{j=1}^{n} X_j,$$

has a (joint) pdf that also falls in the exponential family; i.e., $f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\theta})$ can be written in the form

$$f_{\mathbf{T}}(\mathbf{t}|\boldsymbol{\theta}) = H(\mathbf{t})[c(\boldsymbol{\theta})]^n \exp\left\{\sum_{i=1}^{2} w_i(\boldsymbol{\theta})t_i\right\}.$$

5.3 Sampling from the Normal Distribution

Remark: This section is dedicated to results that arise when $X_1, X_2, ..., X_n$ are normally distributed.

Theorem 5.3.1. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. Let $\overline{X}$ and S^2 denote the sample mean and the sample variance, respectively. Then

(a) $\overline{X} \perp\!\!\!\perp S^2$

(b) $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ $\longleftarrow$ we showed this in Example 5.2

(c) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Proof. We first prove part (a). Recall that

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\ &= \frac{1}{n-1}\sum_{i=2}^{n}(X_i - \overline{X})^2 + \frac{1}{n-1}(X_1 - \overline{X})^2. \end{aligned}$$

Note also that

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0 \quad \Longrightarrow \quad X_1 - \overline{X} = -\sum_{i=2}^{n}(X_i - \overline{X}).$$

Therefore, we can rewrite S^2 as

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=2}^{n}(X_i - \overline{X})^2 + \frac{1}{n-1}\left[-\sum_{i=2}^{n}(X_i - \overline{X})\right]^2 \\ &= g(X_2 - \overline{X}, X_3 - \overline{X}, ..., X_n - \overline{X}), \text{ say.} \end{aligned}$$

Because functions of independent random variables (vectors) are independent, it suffices to show that $\overline{X} \perp\!\!\!\perp (X_2 - \overline{X}, X_3 - \overline{X}, ..., X_n - \overline{X})$. Consider the n-variate transformation

$$\begin{array}{lcl} y_1 = \overline{x} & & x_1 = y_1 - \sum_{i=2}^{n} y_i \\ y_2 = x_2 - \overline{x} & & x_2 = y_1 + y_2 \\ y_3 = x_3 - \overline{x} & \Longrightarrow & x_3 = y_1 + y_3 \\ \quad\vdots & & \quad\vdots \\ y_n = x_n - \overline{x} & & x_n = y_1 + y_n. \end{array}$$

The Jacobian of the (inverse) transformation is

$$J = \det\begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix} = n.$$

Therefore, the pdf of $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ is given by

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\left(y_1 - \sum_{i=2}^{n} y_i,\ y_1 + y_2,\ y_1 + y_3, ...,\ y_1 + y_n\right)|n|.$$

Going forward, we assume (without loss of generality) that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, with $\mu = 0$ and $\sigma^2 = 1$. Under this assumption,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} I(x_i \in \mathbb{R}) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} I(\mathbf{x} \in \mathbb{R}^n). \end{aligned}$$

This simplifies the calculations and is not prohibitive because the $\mathcal{N}(\mu, \sigma^2)$ family is a location-scale family (see CB, pp 216-217). From above, we therefore have, for all $\mathbf{y} \in \mathbb{R}^n$,

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \frac{n}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}\left[\left(y_1 - \sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n}(y_1 + y_i)^2\right]\right\} \\ &= \underbrace{\frac{n}{(2\pi)^{n/2}} e^{-ny_1^2/2}}_{h_1(y_1)} \times \underbrace{\exp\left\{-\frac{1}{2}\left[\sum_{i=2}^{n} y_i^2 + \left(\sum_{i=2}^{n} y_i\right)^2\right]\right\}}_{h_2(y_2, y_3, ..., y_n)}. \end{aligned}$$

Because the joint pdf factors, by Theorem 4.6.11, it follows that $Y_1 \perp\!\!\!\perp (Y_2, Y_3, ..., Y_n)$, that is, $\overline{X} \perp\!\!\!\perp (X_2 - \overline{X}, X_3 - \overline{X}, ..., X_n - \overline{X})$. The same conclusion would have been reached had we allowed $X_1, X_2, ..., X_n$ to be iid $\mathcal{N}(\mu, \sigma^2)$, μ and σ^2 arbitrary; the calculations would have just been far messier. We have proven part (a). To prove part (c), we first recall the following:

$$X_i \sim \mathcal{N}(\mu, \sigma^2) \implies Z_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1) \implies Z_i^2 \sim \chi_1^2.$$

Therefore, the random variables $Z_1^2, Z_2^2, ..., Z_n^2$ are iid χ_1^2. Because the degrees of freedom add (see Example 4.24 in the notes),

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi_n^2.$$

Now, write

$$\begin{aligned} W_1 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X} + \overline{X} - \mu}{\sigma}\right)^2 \\ &= \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \underbrace{2\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma}\right)\left(\frac{\overline{X} - \mu}{\sigma}\right)}_{=\ 0} + \sum_{i=1}^{n} \left(\frac{\overline{X} - \mu}{\sigma}\right)^2. \end{aligned}$$

It is easy to show that the cross product term is zero because

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0.$$

Therefore, we have

$$W_1 = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \underbrace{\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2}_{=W_2} + \underbrace{n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2}_{=W_3}.$$

Now,

$$W_3 = n\left(\frac{\overline{X}-\mu}{\sigma}\right)^2 = \left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1,$$

because $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$, and

$$W_2 = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2 = \frac{(n-1)S^2}{\sigma^2}.$$

Furthermore, we know that $W_2 \perp\!\!\!\perp W_3$ because $\overline{X} \perp\!\!\!\perp S^2$ and functions of independent random variables are independent. The mgf of $W_1 \sim \chi^2_n$ is, for $t < 1/2$,

$$\begin{aligned}
\left(\frac{1}{1-2t}\right)^{n/2} = M_{W_1}(t) = E(e^{tW_1}) &= E[e^{t(W_2+W_3)}] \\
&= E(e^{tW_2}e^{tW_3}) \\
&\overset{W_2 \perp\!\!\!\perp W_3}{=} E(e^{tW_2})E(e^{tW_3}) \\
&= M_{W_2}(t)M_{W_3}(t) \\
&= M_{W_2}(t)\left(\frac{1}{1-2t}\right)^{1/2},
\end{aligned}$$

because $W_3 \sim \chi^2_1$. This shows that

$$M_{W_2}(t) = \left(\frac{1}{1-2t}\right)^{(n-1)/2},$$

which, when $t < 1/2$, we recognize as the mgf of a χ^2_{n-1} random variable. Because mgfs are unique,

$$W_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. \ \square$$

Remark: My proof of part (c) is different than the proof your authors provide on pp 219-220 (CB). They use mathematical induction (I like mine better).

Remark: When $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$,

$$\begin{aligned} \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 &\sim \chi^2_n \\ \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 &\sim \chi^2_{n-1}. \end{aligned}$$

In the second result, one can interpret having to "estimate" μ with $\overline{X}$ as being responsible for "losing" a degree of freedom.

Remark: When $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, we have shown that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Therefore, we get the following results "for free:"

$$\begin{aligned} E\left[\frac{(n-1)S^2}{\sigma^2}\right] &= n-1 \\ \text{var}\left[\frac{(n-1)S^2}{\sigma^2}\right] &= 2(n-1). \end{aligned}$$

The first result implies that $E(S^2) = \sigma^2$. Of course, this is nothing new. We proved $E(S^2) = \sigma^2$ in general; i.e., for any population distribution with finite variance. The second result implies

$$\frac{(n-1)^2}{\sigma^4}\,\text{var}(S^2) = 2(n-1) \implies \text{var}(S^2) = \frac{2\sigma^4}{n-1}.$$

This is a new result. However, it only applies when $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$.

General result: Suppose that $X_1, X_2, ..., X_n$ are iid with $E(X^4) < \infty$. Then

$$\text{var}(S^2) = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right],$$

where recall $\mu_4 = E[(X-\mu)^4]$ is the fourth central moment of X. See pp 257 (CB). As an exercise, show that this expression for $\text{var}(S^2)$ reduces to $2\sigma^4/(n-1)$ in the normal case.

Lemma 5.5.3 (special case). Suppose that $X_1, X_2, ..., X_n$ are independent random variables with $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, for $j = 1, 2, ..., n$. Define the linear combinations

$$\begin{aligned} U = \sum_{j=1}^{n} a_j X_j &= a_1X_1 + a_2X_2 + \cdots + a_nX_n \\ V = \sum_{j=1}^{n} b_j X_j &= b_1X_1 + b_2X_2 + \cdots + b_nX_n, \end{aligned}$$

where $a_j, b_j \in \mathbb{R}$ are fixed constants (i.e., not random). Then

$$U \perp\!\!\!\perp V \iff \text{cov}(U, V) = 0.$$

In other words, the linear combinations U and V are independent if and only if U and V are uncorrelated.

Remark: Your authors prove this result in a very special case, by assuming that $n = 2$ and X_1, X_2 are iid $\mathcal{N}(0,1)$, i.e., $\mu_1 = \mu_2 = 0$ and $\sigma_1^2 = \sigma_2^2 = 1$. They perform a bivariate transformation to obtain the joint pdf $f_{U,V}(u,v)$, where $U = a_1X_1 + a_2X_2$ and $V = b_1X_1 + b_2X_2$, and then show that

$$f_{U,V}(u,v) = f_U(u)f_V(v) \iff \text{cov}(U,V) = 0.$$

Result: Suppose U and V are linear combinations as defined on the last page. If $X_1, X_2, ..., X_n$ are independent random variables (not necessarily normal), then

$$\text{cov}(U,V) = \sum_{j=1}^{n} a_j b_j \sigma_j^2.$$

Establishing this result is not difficult and is merely an exercise in patience (use the covariance computing formula and then do lots of algebra). Interestingly, under the simplified assumption that $\sigma_j^2 = 1$ for all j,

$$\text{cov}(U,V) = \sum_{j=1}^{n} a_j b_j = \mathbf{a}'\mathbf{b},$$

where $\mathbf{a} = (a_1, a_2, ..., a_n)'$ and $\mathbf{b} = (b_1, b_2, ..., b_n)'$. Therefore $U \perp\!\!\!\perp V$ if and only if $\mathbf{a}$ and $\mathbf{b}$ are orthogonal vectors in $\mathbb{R}^n$.

Student's t distribution: Suppose that $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2_p$, and $U \perp\!\!\!\perp V$. The random variable

$$T = \frac{U}{\sqrt{V/p}} \sim t_p,$$

a t distribution with p degrees of freedom. The pdf of T is

$$f_T(t) = \frac{\Gamma(\frac{p+1}{2})}{\sqrt{p\pi}\,\Gamma(\frac{p}{2})} \frac{1}{(1+\frac{t^2}{p})^{(p+1)/2}}\, I(t \in \mathbb{R}).$$

Note: If $p = 1$, then $T \sim \text{Cauchy}(0,1)$.

Application: Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. We already know that

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

The quantity

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where S denotes the sample standard deviation of $X_1, X_2, ..., X_n$. To see why, note that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\sigma}{S}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) = \frac{\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}\Big/(n-1)}} \sim \frac{\text{“}\mathcal{N}(0,1)\text{”}}{\sqrt{\frac{\text{“}\chi^2_{n-1}\text{”}}{n-1}}}.$$

Because $\overline{X} \perp\!\!\!\perp S^2$, the numerator and denominator are independent. Therefore, $T \sim t_{n-1}$.

Derivation: Suppose $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2_p$, and $U \perp\!\!\!\perp V$. The joint pdf of (U,V) is

$$f_{U,V}(u,v) = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-u^2/2}}_{\mathcal{N}(0,1) \text{ pdf}} \underbrace{\frac{1}{\Gamma(\frac{p}{2})2^{p/2}} v^{\frac{p}{2}-1} e^{-v/2}}_{\chi^2_p \text{ pdf}},$$

for $-\infty < u < \infty$ and $v > 0$. Consider the bivariate transformation

$$\begin{aligned} T = g_1(U,V) &= \frac{U}{\sqrt{V/p}} \\ W = g_2(U,V) &= V. \end{aligned}$$

The support of (U,V) is the set $\mathcal{A} = \{(u,v) : -\infty < u < \infty, v > 0\}$. The support of (T,W) is $\mathcal{B} = \{(t,w) : -\infty < t < \infty, w > 0\}$. The transformation above is one-to-one, so the inverse transformation exists and is given by

$$\begin{aligned} u = g_1^{-1}(t,w) &= t\sqrt{w/p} \\ v = g_2^{-1}(t,w) &= w. \end{aligned}$$

The Jacobian of the (inverse) transformation is

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(t,w)}{\partial t} & \frac{\partial g_1^{-1}(t,w)}{\partial w} \\ \frac{\partial g_2^{-1}(t,w)}{\partial t} & \frac{\partial g_2^{-1}(t,w)}{\partial w} \end{vmatrix} = \det \begin{vmatrix} \sqrt{w/p} & \frac{t}{\sqrt{p}}\frac{1}{2}w^{-1/2} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{w}{p}},$$

which never vanishes over $\mathcal{B}$. For $(t,w) \in \mathcal{B}$, the joint pdf of (T,W) is

$$\begin{aligned} f_{T,W}(t,w) &= f_{U,V}(g_1^{-1}(t,w), g_2^{-1}(t,w))|J| \\ &= \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{w/p})^2/2} \frac{1}{\Gamma(\frac{p}{2})2^{p/2}} w^{\frac{p}{2}-1} e^{-w/2} \left|\sqrt{\frac{w}{p}}\right| \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{p}} e^{-(t\sqrt{w/p})^2/2} \frac{1}{\Gamma(\frac{p}{2})2^{p/2}} w^{\frac{p+1}{2}-1} e^{-w/2} \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{p}} \frac{1}{\Gamma(\frac{p}{2})2^{p/2}} w^{\frac{p+1}{2}-1} e^{-w(1+\frac{t^2}{p})/2}. \end{aligned}$$

Therefore, the marginal pdf of T is

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,W}(t,w)dw \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{p}} \frac{1}{\Gamma(\frac{p}{2})2^{p/2}} \int_0^\infty \underbrace{w^{\frac{p+1}{2}-1} e^{-w(1+\frac{t^2}{p})/2}}_{\text{gamma}(a,b) \text{ kernel}} dw, \end{aligned}$$

where $a = (p+1)/2$ and $b = 2\left(1 + \frac{t^2}{p}\right)^{-1}$. The gamma integral above equals

$$\Gamma(a)b^a = \Gamma\left(\frac{p+1}{2}\right)\left[2\left(1+\frac{t^2}{p}\right)^{-1}\right]^{(p+1)/2}.$$

Therefore, for all $t \in \mathbb{R}$,

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{p}} \frac{1}{\Gamma(\frac{p}{2}) 2^{p/2}} \Gamma\left(\frac{p+1}{2}\right) \left[2\left(1+\frac{t^2}{p}\right)^{-1}\right]^{(p+1)/2} \\ &= \frac{\Gamma(\frac{p+1}{2})}{\sqrt{p\pi}\,\Gamma(\frac{p}{2})} \frac{1}{(1+\frac{t^2}{p})^{(p+1)/2}}, \end{aligned}$$

as claimed. □

Moments: If $T \sim t_p$, then

$$\begin{aligned} E(T) &= 0, \quad \text{if } p > 1 \\ \text{var}(T) &= \frac{p}{p-2}, \quad \text{if } p > 2. \end{aligned}$$

To show that $E(T) = 0$ when $p > 1$, suppose $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2_p$, and $U \perp\!\!\!\perp V$. Write

$$E(T) = E\left(\frac{U}{\sqrt{V/p}}\right) \stackrel{U \perp\!\!\!\perp V}{=} E(U) E\left(\frac{1}{\sqrt{V/p}}\right) = 0,$$

because $E(U) = 0$. We need to investigate the second expectation to see why the "$p > 1$" condition is needed. Recall that $V \sim \chi^2_p \stackrel{d}{=} \text{gamma}(\frac{p}{2}, 2)$. Therefore,

$$\begin{aligned} E\left(\frac{1}{\sqrt{V/p}}\right) = \sqrt{p} E\left(\frac{1}{\sqrt{V}}\right) &= \sqrt{p} \int_0^\infty \frac{1}{\sqrt{v}} \frac{1}{\Gamma(\frac{p}{2}) 2^{p/2}} v^{\frac{p}{2}-1} e^{-v/2} dv \\ &= \frac{\sqrt{p}}{\Gamma(\frac{p}{2}) 2^{p/2}} \int_0^\infty v^{\frac{p-1}{2}-1} e^{-v/2} dv \\ &= \frac{\sqrt{p}}{\Gamma(\frac{p}{2}) 2^{p/2}} \Gamma\left(\frac{p-1}{2}\right) 2^{(p-1)/2} \\ &= \frac{\sqrt{p}\Gamma(\frac{p-1}{2})}{\sqrt{2}\Gamma(\frac{p}{2})}, \end{aligned}$$

which is finite. However, the penultimate equality holds only when $(p-1)/2 > 0$; i.e., when $p > 1$. Showing $\text{var}(T) = p/(p-2)$ when $p > 2$ is done similarly.

Snedecor's F distribution: Suppose that $U \sim \chi^2_p$, $V \sim \chi^2_q$, and $U \perp\!\!\!\perp V$. The random variable

$$W = \frac{U/p}{V/q} \sim F_{p,q},$$

an F distribution with (numerator) p and (denominator) q degrees of freedom. The pdf of W is

$$f_W(w) = \frac{\Gamma(\frac{p+q}{2})}{\Gamma(\frac{p}{2})\Gamma(\frac{q}{2})} \left(\frac{p}{q}\right)^{p/2} \frac{w^{\frac{p}{2}-1}}{[1+(\frac{p}{q})w]^{(p+q)/2}} I(w > 0).$$

Note: This pdf can be derived in the same way that the t pdf was derived. Apply a bivariate transformation; i.e., introduce $Z = U$ as a dummy variable, find $f_{W,Z}(w, z)$, and then integrate over z.

Moments: If $W \sim F_{p,q}$, then

$$\begin{aligned} E(W) &= \frac{q}{q-2}, \quad \text{if } q > 2 \\ \text{var}(W) &= 2\left(\frac{q}{q-2}\right)^2 \frac{p+q-2}{p(q-4)}, \quad \text{if } q > 4. \end{aligned}$$

Proof. Exercise.

Application: Suppose we have independent random samples

$$\begin{aligned} X_1, X_2, ..., X_n &\sim \text{iid } \mathcal{N}(\mu_X, \sigma_X^2) \\ Y_1, Y_2, ..., Y_m &\sim \text{iid } \mathcal{N}(\mu_Y, \sigma_Y^2). \end{aligned}$$

We know

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2 \quad \text{and} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi_{m-1}^2.$$

Also, these quantities are independent (because the samples are; therefore, $S_X^2 \perp\!\!\!\perp S_Y^2$). Therefore,

$$W = \frac{\frac{(n-1)S_X^2}{\sigma_X^2}\Big/(n-1)}{\frac{(m-1)S_Y^2}{\sigma_Y^2}\Big/(m-1)} = \left(\frac{S_X^2}{S_Y^2}\right)\frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n-1,m-1}.$$

Furthermore, if $\sigma_X^2 = \sigma_Y^2$ (perhaps an assumption under some H_0), then $S_X^2/S_Y^2 \sim F_{n-1,m-1}$. In this case,

$$E\left(\frac{S_X^2}{S_Y^2}\right) = \frac{m-1}{m-3} \approx 1.$$

Therefore, if $\sigma_X^2 = \sigma_Y^2$, we would expect the ratio of the sample variances to be close to 1 (especially if m is large).

Theorem 5.3.8.

(a) If $X \sim F_{p,q}$, then $Y = 1/X \sim F_{q,p}$.

(b) If $X \sim t_q$, then $Y = X^2 \sim F_{1,q}$.

(c) If $X \sim F_{p,q}$, then

$$Y = \frac{(\frac{p}{q})X}{1 + (\frac{p}{q})X} \sim \text{beta}\left(\frac{p}{2}, \frac{q}{2}\right).$$

Note: The first two results above are important in analysis of variance and regression. The result in part (c) is useful when estimating a binomial success probability.

5.4 Order Statistics

Definition: The **order statistics** of an iid sample $X_1, X_2, ..., X_n$ are the ordered values of the sample. They are denoted by $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$; i.e.,

$$\begin{aligned} X_{(1)} &= \min_{1\leq i\leq n} X_i \\ X_{(2)} &= \text{second smallest } X_i \\ &\vdots \\ X_{(n)} &= \max_{1\leq i\leq n} X_i. \end{aligned}$$

Remark: Many statistics seen in practice are order statistics or functions of order statistics. For example,

- $T(\mathbf{X}) = X_{(\frac{n+1}{2})}$, the sample median ($n$ odd)
- $T(\mathbf{X}) = X_{(n)} - X_{(1)}$, the sample range
- $T(\mathbf{X}) = \frac{1}{2}(X_{(1)} + X_{(n)})$, the sample midrange.

Revelation: Because order statistics are statistics (i.e., they are functions of $X_1, X_2, ..., X_n$), they have their own (sampling) distributions! This section is dedicated to studying these distributions.

Note: We will examine the discrete and continuous cases separately. In general, "ties" among order statistics are possible when the population distribution $f_X(x)$ is discrete; ties are not possible theoretically when $f_X(x)$ is continuous.

Theorem 5.4.3. Suppose $X_1, X_2, ..., X_n$ is an iid sample from a **discrete** distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < \cdots < x_i < \cdots$ are the possible values of X listed in ascending order. Define $P_0 = 0$,

$$P_1 = p_1, \quad P_2 = p_1 + p_2, ..., \quad P_i = p_1 + p_2 + \cdots + p_i, ...$$

Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the order statistics from the sample. Then

$$P(X_{(j)} \leq x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}$$

and

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}\right].$$

Proof. Let Y denote the number of X_j's (among $X_1, X_2, ..., X_n$) that are less than or equal to x_i. The key point to realize is that

$$\{X_{(j)} \leq x_i\} = \{Y \geq j\}.$$

Furthermore, $Y \sim b(n, P_i)$. To see why, consider each X_j as a "trial:"

- if $X_j \leq x_i$, then call this a "success"
- if $X_j > x_i$, then call this a "failure."

Note that Y simply counts the number of "successes" out of these n Bernoulli trials. The probability of a "success" on any one trial is

$$P(X_j \leq x_i) = p_1 + p_2 + \cdots + p_i = P_i.$$

Therefore,

$$\begin{aligned} P(X_{(j)} \leq x_i) &= P(Y \geq j) \\ &= \sum_{k=j}^{n} \binom{n}{k} P_i^k (1-P_i)^{n-k}. \end{aligned}$$

The expression for $P(X_{(j)} = x_i)$ is simply $P(X_{(j)} \leq x_i) - P(X_{(j)} \leq x_{i-1})$. The definition of $P_0 = 0$ takes care of the $i = 1$ case. □

Example 5.7. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(θ), where $0 < \theta < 1$. Find $E(X_{(j)})$.
Solution. Recall that the Bernoulli (population) pmf can be written as

$$f_X(x) = \begin{cases} p_1 = 1-\theta, & x = x_1 = 0 \\ p_2 = \theta, & x = x_2 = 1 \\ 0, & \text{otherwise.} \end{cases}$$

Using the notation defined in Theorem 5.4.3, we have $P_0 = 0$, $P_1 = p_1 = 1 - \theta$, and $P_2 = p_1 + p_2 = (1-\theta) + \theta = 1$. The random variable $X_{(j)}$ is binary (0-1). Therefore,

$$\begin{aligned} E(X_{(j)}) = P(X_{(j)} = 1) &= 1 - P(X_{(j)} = 0) \\ &= 1 - P(X_{(j)} \leq 0) \\ &= 1 - \sum_{k=j}^{n} \binom{n}{k} (1-\theta)^k \theta^{n-k}. \end{aligned}$$

For example, if $\theta = 0.2$ and $n = 25$, then $E(X_{(13)}) \approx 0.000369$. Note that $X_{(13)}$ is the sample median if $n = 25$.

Exercise: Suppose that $X_1, X_2, ..., X_{10}$ are iid Poisson with mean $\lambda = 2.2$. Calculate $E(X_{(j)})$ for $j = 1, 2, ..., 10$. Compare each with $E(X_j) = 2.2$.

Theorem 5.4.4. Suppose $X_1, X_2, ..., X_n$ is an iid sample from a **continuous** distribution with pdf $f_X(x)$ and cdf $F_X(x)$. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the order statistics. The pdf of $X_{(j)}$, for $j = 1, 2, ..., n$, is given by

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} [F_X(x)]^{j-1} f_X(x) [1 - F_X(x)]^{n-j}.$$

Two special cases:

- $j = 1$. The pdf of $X_{(1)}$, the **minimum** order statistic, collapses to
$$f_{X_{(1)}}(x) = nf_X(x)[1 - F_X(x)]^{n-1}.$$
- $j = n$. The pdf of $X_{(n)}$, the **maximum** order statistic, collapses to
$$f_{X_{(n)}}(x) = nf_X(x)[F_X(x)]^{n-1}.$$

The formulae for these special cases should be committed to memory.

Proof of Theorem 5.4.4. Let $F_{X_{(j)}}(x) = P(X_{(j)} \leq x)$ denote the cdf of $X_{(j)}$. Let Y denote the number of X_i's that are less than or equal to x. As in the discrete case, $\{X_{(j)} \leq x\} = \{Y \geq j\}$. Furthermore, $Y \sim b(n, F_X(x))$. To see why, consider each X_i as a "trial:"

- if $X_i \leq x$, then call this a "success"
- if $X_i > x$, then call this a "failure."

Note that Y simply counts the number of "successes" out of these n Bernoulli trials. The probability of a "success" on any one trial is $P(X_i \leq x) = F_X(x)$. Therefore,

$$\begin{aligned} F_{X_{(j)}}(x) = P(X_{(j)} \leq x) &= P(Y \geq j) \\ &= \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}. \end{aligned}$$

The pdf of $X_{(j)}$ is given by

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{d}{dx} F_{X_{(j)}}(x) \\ &= \frac{d}{dx} \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k} \\ &= \sum_{k=j}^{n} \binom{n}{k} \frac{d}{dx} [F_X(x)]^k [1 - F_X(x)]^{n-k} \\ &= \sum_{k=j}^{n} \binom{n}{k} \Big\{ k[F_X(x)]^{k-1} f_X(x)[1 - F_X(x)]^{n-k} - [F_X(x)]^k (n-k)[1 - F_X(x)]^{n-k-1} f_X(x) \Big\} \\ &= \binom{n}{j} j[F_X(x)]^{j-1} f_X(x)[1 - F_X(x)]^{n-j} + a - b, \end{aligned}$$

where

$$\begin{aligned} a &= \sum_{k=j+1}^{n} \binom{n}{k} k[F_X(x)]^{k-1} f_X(x)[1 - F_X(x)]^{n-k} \\ b &= \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k (n-k)[1 - F_X(x)]^{n-k-1} f_X(x). \end{aligned}$$

Note that the expression above

$$\binom{n}{j} j[F_X(x)]^{j-1} f_X(x)[1-F_X(x)]^{n-j} = \frac{n!}{(j-1)!(n-j)!}[F_X(x)]^{j-1} f_X(x)[1-F_X(x)]^{n-j}$$

is the desired result. Therefore, it suffices to show that $a - b = 0$. Re-index the a sum and re-write the b sum as

$$\begin{aligned} a &= \sum_{k=j}^{n-1} \binom{n}{k+1} (k+1)[F_X(x)]^k f_X(x)[1-F_X(x)]^{n-k-1} \\ b &= \sum_{k=j}^{n-1} \binom{n}{k} (n-k)[F_X(x)]^k f_X(x)[1-F_X(x)]^{n-k-1}. \end{aligned}$$

To establish that $a - b = 0$, simply note that

$$\begin{aligned} \binom{n}{k+1}(k+1) &= \frac{n!}{(k+1)!(n-k-1)!}(k+1) \\ &= \frac{n!}{k!(n-k-1)!} \\ &= \frac{n!}{k!(n-k)!}(n-k) = \binom{n}{k}(n-k). \ \square \end{aligned}$$

Conceptualization: There is an easy way to remember

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!}[F_X(x)]^{j-1} f_X(x)[1-F_X(x)]^{n-j}.$$

Think of each of $X_1, X_2, ..., X_n$ as a "trial," and consider the trinomial distribution with the following categories:

Category	Description	Cell probability	# Observations
1	Less than x	$p_1 = P(X < x) = F_X(x)$	$j-1$
2	Equal to x	$p_2 = P(X = x) =$ "$f_X(x)$"	1
3	Greater than x	$p_3 = P(X > x) = 1 - F_X(x)$	$n-j$

Therefore, we can remember the formula for $f_{X_{(j)}}(x)$ by linking it to the trinomial distribution (see Section 4.6 in the notes). The constant

$$\frac{n!}{(j-1)!(n-j)!} = \frac{n!}{(j-1)!1!(n-j)!}$$

is the corresponding trinomial coefficient; it counts the number of ways the X_i's can fall in the three distinct categories.

Example 5.8. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0,1)$. Find the pdf of $X_{(j)}$, the jth order statistic.

Solution. Recall that the population pdf and cdf are

$$f_X(x) = I(0 < x < 1) \qquad \text{and} \qquad F_X(x) = \begin{cases} 0, & x \leq 0 \\ x, & 0 < x < 1 \\ 1, & x \geq 1. \end{cases}$$

The pdf of $X_{(j)}$ is, for $0 < x < 1$,

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1)(1-x)^{n-j} \\ &= \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1}; \end{aligned}$$

i.e., $X_{(j)} \sim \text{beta}(j, n-j+1)$.

Example 5.9. Suppose that $X_1, X_2, ..., X_n$ are iid exponential(β), where $\beta > 0$. Recall that the exponential pdf and cdf are

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta} I(x > 0) \qquad \text{and} \qquad F_X(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-x/\beta}, & x > 0. \end{cases}$$

The pdf of $X_{(1)}$ is

$$\begin{aligned} f_{X_{(1)}}(x) &= n f_X(x)[1 - F_X(x)]^{n-1} \\ &\overset{x>0}{=} n \, \frac{1}{\beta} e^{-x/\beta} \, [1 - (1 - e^{-x/\beta})]^{n-1} \\ &= \frac{n}{\beta} e^{-nx/\beta} I(x > 0), \end{aligned}$$

that is, $X_{(1)} \sim \text{exponential}(\beta/n)$. The pdf of $X_{(n)}$ is

$$\begin{aligned} f_{X_{(n)}}(x) &= n f_X(x)[F_X(x)]^{n-1} \\ &= \frac{n}{\beta} e^{-x/\beta}(1 - e^{-x/\beta})^{n-1} I(x > 0). \end{aligned}$$

The pdfs of $X_{(1)}$ and $X_{(n)}$ are depicted in Figure 5.2 for $n = 10$ and $\beta = 2$.

Theorem 5.4.6. Suppose $X_1, X_2, ..., X_n$ is an iid sample from a **continuous** distribution with pdf $f_X(x)$ and cdf $F_X(x)$. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the order statistics. The joint pdf of $(X_{(i)}, X_{(j)})$, $i < j$, is given by

$$\begin{aligned} f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-1-i)!(n-j)!} &[F_X(u)]^{i-1} f_X(u)[F_X(v) - F_X(u)]^{j-1-i} \\ &\times f_X(v)[1 - F_X(v)]^{n-j}, \end{aligned}$$

for $-\infty < u < v < \infty$.

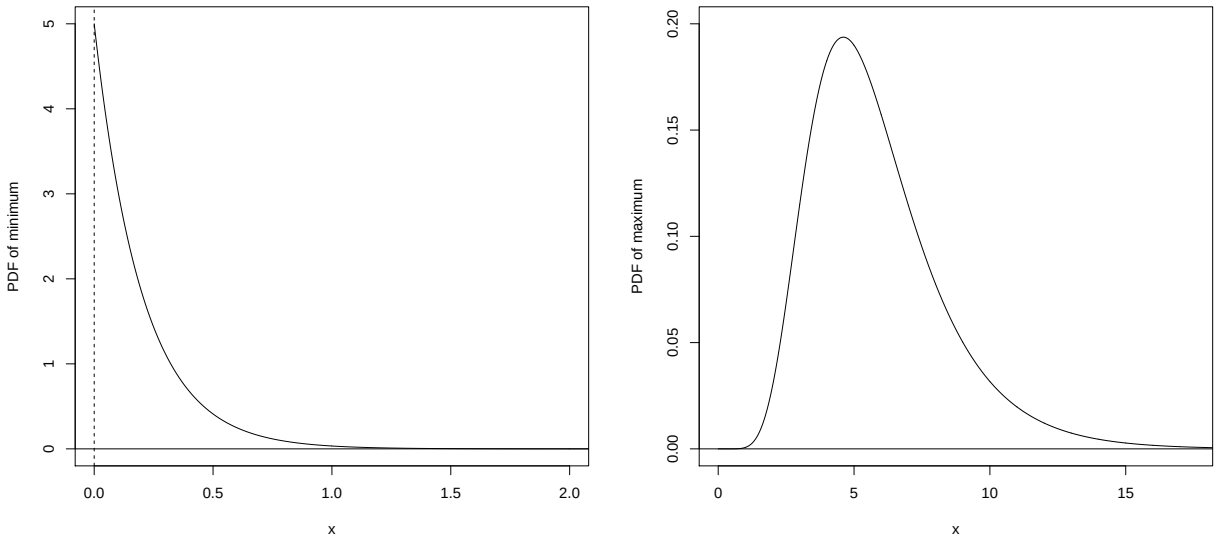


Figure 5.2: Order statistic distributions. $X_1, X_2, ..., X_{10} \sim$ iid exponential($\beta = 2$). Left: Pdf of $X_{(1)}$. Right: Pdf of $X_{(10)}$. Note that the horizontal axes are different in the two figures.

Remark: For a rigorous derivation of this result, see Exercise 5.26 (CB, pp 260). For a heuristic argument, we can again link the formula for $f_{X_{(i)},X_{(j)}}(u, v)$ to the multinomial distribution:

Category	Description	Cell probability	# Observations
1	Less than u	$p_1 = F_X(u)$	$i-1$
2	Equal to u	$p_2 =$ "$f_X(u)$"	1
3	Between u and v	$p_3 = F_X(v) - F_X(u)$	$j-1-i$
4	Equal to v	$p_4 =$ "$f_X(v)$"	1
5	Greater than v	$p_5 = 1 - F_X(v)$	$n-j$

Special case: The joint pdf of $(X_{(1)}, X_{(n)})$ is given by

$$f_{X_{(1)},X_{(n)}}(u, v) = n(n-1)f_X(u)[F_X(v) - F_X(u)]^{n-2}f_X(v),$$

for $-\infty < u < v < \infty$.

Example 5.10. Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, 1)$. Find the pdf of the sample range $R = X_{(n)} - X_{(1)}$.
Solution. The first step is to find the joint pdf of $X_{(1)}$ and $X_{(n)}$; this is given by

$$f_{X_{(1)},X_{(n)}}(u, v) = n(n-1)(v-u)^{n-2}I(0 < u < v < 1).$$

Now consider the bivariate transformation

$$\begin{aligned} R = g_1(X_{(1)}, X_{(n)}) &= X_{(n)} - X_{(1)} \\ S = g_2(X_{(1)}, X_{(n)}) &= X_{(n)}. \end{aligned}$$

Note that the support of (R, S) is $\mathcal{B} = \{(r, s) : 0 < r < s < 1\}$. The transformation above is one-to-one, so the inverse transformation exists and is given by

$$\begin{aligned} x_{(1)} = g_1^{-1}(r, s) &= s - r \\ x_{(n)} = g_2^{-1}(r, s) &= s. \end{aligned}$$

The Jacobian of the (inverse) transformation is

$$J = \det \begin{vmatrix} \dfrac{\partial g_1^{-1}(r,s)}{\partial r} & \dfrac{g_1^{-1}(r,s)}{\partial s} \\ \dfrac{\partial g_2^{-1}(r,s)}{\partial r} & \dfrac{\partial g_2^{-1}(r,s)}{\partial s} \end{vmatrix} = \det \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.$$

Therefore, the joint pdf of (R, S) is, for all $(r, s) \in \mathcal{B}$,

$$\begin{aligned} f_{R,S}(r, s) &= f_{X_{(1)},X_{(n)}}(g_1^{-1}(r, s), g_2^{-1}(r, s))|J| \\ &= n(n-1)[s - (s - r)]^{n-2} \\ &= n(n-1)r^{n-2} I(0 < r < s < 1). \end{aligned}$$

The marginal pdf of R is therefore

$$\begin{aligned} f_R(r) &= \int_r^1 n(n-1)r^{n-2} ds \\ &= n(n-1)r^{n-2}(1 - r) I(0 < r < 1) \\ &= \frac{\Gamma(n+1)}{\Gamma(n-1)\Gamma(2)} r^{(n-1)-1}(1 - r)^{2-1} I(0 < r < 1), \end{aligned}$$

a beta pdf with parameters $n - 1$ and 2, that is, $R \sim \text{beta}(n - 1, 2)$.

Result: Suppose $X_1, X_2, ..., X_n$ is an iid sample from a **continuous** distribution with pdf $f_X(x)$. The joint distribution of the n order statistics is

$$f_{X_{(1)},X_{(2)},...,X_{(n)}}(x_1, x_2, ..., x_n) = n! f_X(x_1) f_X(x_2) \cdots f_X(x_n),$$

for $-\infty < x_1 < x_2 < \cdots < x_n < \infty$.

Q: We know (by assumption) that $X_1, X_2, ..., X_n$ are iid. Are $X_{(1)}, X_{(2)}, ..., X_{(n)}$ also iid?
A: No. Clearly, $X_{(1)}, X_{(2)}, ..., X_{(n)}$ are dependent; i.e., look at the support! Also, the marginal distribution of each order statistic is different. Therefore, the order statistics are neither independent nor identically distributed.

5.5 Convergence Concepts

Remark: In some problems, exact distributional results may not be available. By "exact," we mean "finite sample;" i.e., results that are applicable for any fixed sample size n.

- In Example 5.10, the sample range $R \sim \text{beta}(n-1, 2)$. This is an exact result.

When exact results are not available, we may be able to gain insight by examining the stochastic behavior as the sample size n becomes infinitely large. These are called "large sample" or "asymptotic" results.

Q: Why bother? Large sample results are technically valid only under the assumption that $n \to \infty$. This is not realistic.
A: Because finite sample results are often not available (or they are intractable), and large sample results can offer a good **approximation** to them when n is "large."

Example 5.11. Suppose $X_1, X_2, ..., X_n$ are iid exponential(θ), where $\theta > 0$, and let

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

the sample mean based on n observations. An easy mgf argument (see Example 5.3 in the notes) shows that

$$\overline{X}_n \sim \text{gamma}(n, \theta/n).$$

This is an exact result; it is true for any finite sample size n. Later, we will use the Central Limit Theorem to show that

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \theta^2),$$

as $n \to \infty$. In other words,

$$\overline{X}_n \sim \mathcal{AN}(\theta, \theta^2/n)$$

for large n. The acronym "$\mathcal{AN}$" is read "approximately normal."

Remark: In Example 5.11, the exact distribution of $\overline{X}_n$ is available and is easy to derive. In other situations, it may not be. For example, what if $X_1, X_2, ..., X_n$ were iid beta? iid Bernoulli? iid lognormal?

Review: Suppose $(x_n)_{n=1}^{\infty}$ is a sequence of real numbers. We say that "x_n converges to x" and write

$$\lim_{n\to\infty} x_n = x \quad \text{or} \quad x_n \to x, \quad \text{as } n \to \infty,$$

if $\forall \epsilon > 0\ \exists n_0(\epsilon) \geq 1 \ni |x_n - x| < \epsilon\ \forall n \geq n_0(\epsilon)$. This means that every open neighborhood of x contains all but a finite number of the full sequence $(x_n)_{n=1}^{\infty}$.

Remark: The definition above is a statement about **non-stochastic** convergence. Non-stochastic means "not random;" i.e., the x's are real numbers (to us, fixed constants). On the other hand, stochastic convergence involves random variables. Interestingly, showing stochastic convergence often boils down to showing non-stochastic convergence.

5.5.1 Convergence in probability

Definition: We say that a sequence of random variables $X_1, X_2, ...$, **converges in probability** to a random variable X and write $X_n \xrightarrow{p} X$ if $\forall \epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0,$$

that is, $P(|X_n - X| \geq \epsilon) \to 0$, as $n \to \infty$. An equivalent definition is

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1;$$

i.e., $P(|X_n - X| < \epsilon) \to 1$, as $n \to \infty$.

- For $\epsilon > 0$, quantities like $P(|X_n - X| \geq \epsilon)$ and $P(|X_n - X| < \epsilon)$ are real numbers. Therefore, convergence in probability deals with the non-stochastic convergence of these sequences of real numbers.
- Informally, $X_n \xrightarrow{p} X$ means the probability of the event

  $$\{|X_n - X| \geq \epsilon\} = \{\text{“}X_n \text{ stays away from } X\text{”}\}$$

  gets small as n gets large.
- In most statistical applications, the limiting random variable X is a **constant**.

Example 5.12. Suppose $X_1, X_2, ..., X_n$ are iid exponential(θ), where $\theta > 0$. Show that $\overline{X}_n \xrightarrow{p} \theta$, as $n \to \infty$.

Solution. Suppose $\epsilon > 0$. Recalling that $\overline{X}_n \sim$ gamma$(n, \theta/n)$, it would suffice to show that

$$\begin{aligned} P(|\overline{X}_n - \theta| < \epsilon) &= P(-\epsilon < \overline{X}_n - \theta < \epsilon) \\ &= P(\theta - \epsilon < \overline{X}_n < \theta + \epsilon) \\ &= \int_{\theta-\epsilon}^{\theta+\epsilon} \underbrace{\frac{1}{\Gamma(n)(\frac{\theta}{n})^n} x^{n-1} e^{-nx/\theta}}_{\text{gamma}(n,\theta/n) \text{ pdf}} dx \quad \to \quad 1, \quad \text{as } n \to \infty. \end{aligned}$$

Unfortunately, it is not clear how to do this. Alternatively, we could try to show that $P(|\overline{X}_n - \theta| \geq \epsilon) \to 0$, as $n \to \infty$. This is easier to show. Recall that by Markov's Inequality,

$$\begin{aligned} P(|\overline{X}_n - \theta| \geq \epsilon) &= P((\overline{X}_n - \theta)^2 \geq \epsilon^2) \\ &\leq \frac{E[(\overline{X}_n - \theta)^2]}{\epsilon^2} = \frac{\text{var}(\overline{X}_n)}{\epsilon^2} = \frac{\theta^2}{n\epsilon^2} \to 0, \end{aligned}$$

as $n \to \infty$. We have used Markov's Inequality to bound $P(|\overline{X}_n - \theta| \geq \epsilon)$ above by a sequence that is converging to zero. Therefore, $P(|\overline{X}_n - \theta| \geq \epsilon) \to 0$ as well; i.e., $\overline{X}_n \xrightarrow{p} \theta$, as $n \to \infty$.

Note: Example 5.12 is a special case of a general result known as the **Weak Law of Large Numbers** (**WLLN**).

Theorem 5.5.2 (WLLN). Suppose that $X_1, X_2, ..., X_n$ is an iid sequence of random variables with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. Let

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

denote the sample mean. Then $\overline{X}_n \xrightarrow{p} \mu$, as $n \to \infty$.
Proof. Suppose $\epsilon > 0$. By Markov's Inequality,

$$\begin{aligned} P(|\overline{X}_n - \mu| \geq \epsilon) &= P((\overline{X}_n - \mu)^2 \geq \epsilon^2) \\ &\leq \frac{E[(\overline{X}_n - \mu)^2]}{\epsilon^2} = \frac{\text{var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0. \ \square \end{aligned}$$

Remark: In the version of the WLLN stated in Theorem 5.5.2, we assumed finite variances; i.e., that $E(X_1^2) < \infty$. Can we weaken this assumption? It turns out that the WLLN still holds for iid sequences as long as $E(|X_1|) < \infty$, i.e., the first moment is finite (this is called **Khintchine's WLLN**). Of course, the proof of this version is more difficult.

Remark: The WLLN guarantees that $\overline{X}_n \xrightarrow{p} \mu$, as $n \to \infty$. Does a similar result hold for S^2, the sample variance? That is, does $S^2 \xrightarrow{p} \sigma^2$, as $n \to \infty$?
A: Yes, in most cases. Suppose $\epsilon > 0$. From Markov's Inequality,

$$P(|S^2 - \sigma^2| \geq \epsilon) \leq \frac{E[(S^2 - \sigma^2)^2]}{\epsilon^2} = \frac{\text{var}(S^2)}{\epsilon^2}.$$

Therefore, a sufficient condition for $S^2 \xrightarrow{p} \sigma^2$ is that $\text{var}(S^2) \to 0$, as $n \to \infty$. Recall that

$$\text{var}(S^2) = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right],$$

where $\mu_4 = E[(X-\mu)^4]$ is the fourth central moment of X. Therefore, the sufficient condition $\text{var}(S^2) \to 0$ requires finite fourth moments; i.e., $E(X_1^4) < \infty$. However, $S^2 \xrightarrow{p} \sigma^2$ under the weaker assumption of $E(X_1^2) < \infty$.

Remark: When the limiting random variable is a constant, convergence in probability is sometimes referred to as "consistency" (or "weak consistency"). We might say, "$\overline{X}_n$ is a consistent estimator of μ" and "S^2 is a consistent estimator of σ^2."

Example 5.13. Suppose $X_1, X_2, ..., X_n$ are iid with continuous cdf F_X. Let

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

denote the **empirical distribution function** (**edf**). The edf is a non-decreasing step function that takes steps of size $1/n$ at each observed X_i. By the WLLN,

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) \xrightarrow{p} E[I(X_1 \leq x)] = P(X_1 \leq x) = F_X(x).$$

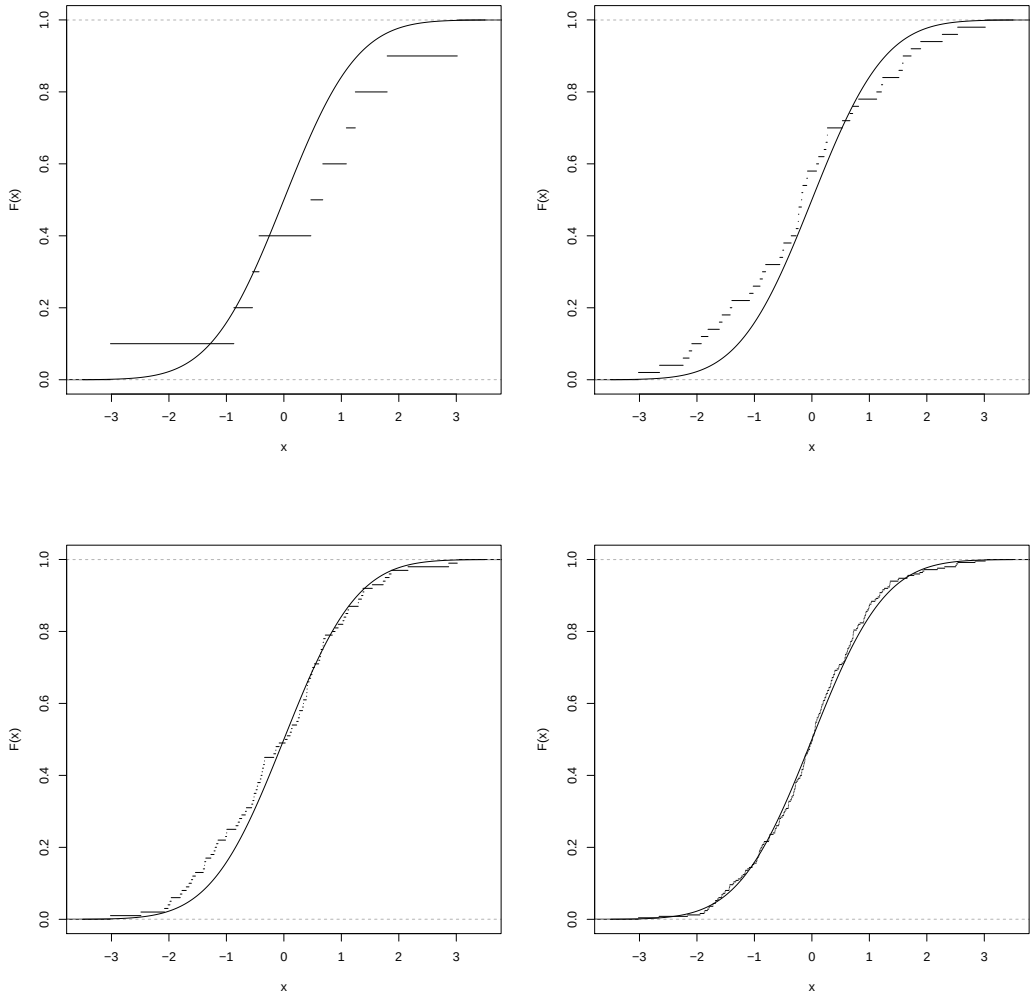


Figure 5.3: Empirical distribution functions $\widehat{F}_n(x)$ calculated from $X_1, X_2, ..., X_n$ iid $\mathcal{N}(0,1)$. Upper left: $n = 10$. Upper right: $n = 50$. Lower left: $n = 100$. Lower right: $n = 250$. The $\mathcal{N}(0,1)$ cdf is superimposed on each subfigure.

That is, $\widehat{F}_n(x) \xrightarrow{p} F_X(x)$ at each fixed $x \in \mathbb{R}$. In a sense, we can think of the edf $\widehat{F}_n(x)$ as an "estimate" of the population cdf $F_X(x)$. Figure 5.3 depicts the edf $\widehat{F}_n(x)$ calculated when $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(0,1)$ for different values of n.

Continuity: Suppose $X_n \xrightarrow{p} X$ and let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then $h(X_n) \xrightarrow{p} h(X)$. In other words, convergence in probability is preserved under continuous mappings.

Proof. Suppose $X_n \xrightarrow{p} X$. Suppose $\epsilon > 0$. Because h is continuous, $\exists \delta(\epsilon) > 0$ such that $|x_n - x| < \delta(\epsilon) \Longrightarrow |h(x_n) - h(x)| < \epsilon$ (this is the definition of continuity). Define the events

$$\begin{aligned} A &= \{x : |x_n - x| < \delta(\epsilon)\} \\ B &= \{x : |h(x_n) - h(x)| < \epsilon\} \end{aligned}$$

and note that $A \subseteq B$. Therefore, by monotonicity of P,

$$P(|X_n - X| < \delta(\epsilon)) = P(X_n \in A) \leq P(X_n \in B) = P(|h(X_n) - h(X)| < \epsilon).$$

However, because $X_n \xrightarrow{p} X$, this means that $P(|X_n - X| < \delta(\epsilon)) \to 1$, as $n \to \infty$. Clearly, $P(|h(X_n) - h(X)| < \epsilon) \to 1$ as well. Because $\epsilon > 0$ was arbitrary, we are done. □

Remark: In Example 5.12, we showed that $\overline{X}_n \xrightarrow{p} \theta$. This means that

$$\overline{X}_n^2 \xrightarrow{p} \theta^2, \quad e^{\overline{X}_n} \xrightarrow{p} e^{\theta}, \quad \text{and} \quad \sin \overline{X}_n \xrightarrow{p} \sin \theta.$$

Note that $h_1(x) = x^2$, $h_2(x) = e^x$, $h_3(x) = \sin x$ are each continuous functions on $\mathbb{R}^+$.

Example 5.14. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Let

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

denote the **sample proportion**. Because $E(X_1) = p < \infty$, it follows from the WLLN that $\widehat{p} \xrightarrow{p} p$, as $n \to \infty$. By continuity,

$$\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \xrightarrow{p} \ln\left(\frac{p}{1-p}\right).$$

The quantity $\ln[p/(1-p)]$ is the **log-odds** of p. Note that $h(x) = \ln[x/(1-x)]$ is a continuous function over $(0, 1)$.

Useful Results: Suppose $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then

(a) $cX_n \xrightarrow{p} cX$, for $c \neq 0$

(b) $X_n \pm Y_n \xrightarrow{p} X \pm Y$

(c) $X_n Y_n \xrightarrow{p} XY$

(d) $X_n/Y_n \xrightarrow{p} X/Y$, provided that $P(Y = 0) = 0$.

Proof. To prove part (a), suppose $X_n \xrightarrow{p} X$ and suppose $\epsilon > 0$. Note that

$$P(|cX_n - cX| \geq \epsilon) = P(|c||X_n - X| \geq \epsilon) = P(|X_n - X| \geq \epsilon/|c|).$$

However, $P(|X_n - X| \geq \epsilon/|c|) \to 0$ because $X_n \xrightarrow{p} X$, by assumption. Because $\epsilon > 0$ was arbitrary, part (a) holds. I will next prove the "+" version of part (b) and leave the remaining parts as exercises. Suppose $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$ and suppose $\epsilon > 0$. From the Triangle Inequality,

$$|X_n + Y_n - (X + Y)| = |X_n - X + Y_n - Y| \leq |X_n - X| + |Y_n - Y|.$$

Therefore,

$$\begin{aligned} P(|X_n + Y_n - (X+Y)| \geq \epsilon) &\leq P(|X_n - X| + |Y_n - Y| \geq \epsilon) \\ &\leq P(|X_n - X| \geq \epsilon/2) + P(|Y_n - Y| \geq \epsilon/2), \end{aligned}$$

the last step following from Boole's Inequality and monotonicity of P. However, because $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, we know $P(|X_n - X| \geq \epsilon/2) \to 0$ and $P(|Y_n - Y| \geq \epsilon/2) \to 0$. Because $\epsilon > 0$ was arbitrary, we are done. □

Remark: Convergence in probability generally can be established by using one of these approaches.

Approach 1: Appeal to the definition directly; that is, show

$$P(|X_n - X| \geq \epsilon) \to 0 \quad \text{or} \quad P(|X_n - X| < \epsilon) \to 1.$$

This approach is particularly useful when X_n is a sequence of order statistics (e.g., $X_{(1)}$, $X_{(n)}$, etc.) and the limiting random variable X is a constant.

Example 5.15. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. Show that $X_{(n)} \xrightarrow{p} \theta$.
Solution. Recall that the $\mathcal{U}(0, \theta)$ pdf and cdf are

$$f_X(x) = \frac{1}{\theta} I(0 < x < \theta) \qquad \text{and} \qquad F_X(x) = \begin{cases} 0, & x \leq 0 \\ \dfrac{x}{\theta}, & 0 < x < \theta \\ 1, & x \geq \theta. \end{cases}$$

The cdf of $X_{(n)}$ is

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \leq x) \\ &= P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) \\ &\overset{\text{indep}}{=} P(X_1 \leq x)P(X_2 \leq x) \cdots P(X_n \leq x) \\ &\overset{\text{ident}}{=} [P(X_1 \leq x)]^n \\ &= [F_X(x)]^n = \left(\frac{x}{\theta}\right)^n, \quad \text{for } 0 < x < \theta. \end{aligned}$$

Therefore,

$$F_{X_{(n)}}(x) = \begin{cases} 0, & x \leq 0 \\ \left(\dfrac{x}{\theta}\right)^n, & 0 < x < \theta \\ 1, & x \geq \theta. \end{cases}$$

Suppose $\epsilon > 0$. By direct calculation, we have

$$\begin{aligned} P(|X_{(n)} - \theta| < \epsilon) &= P(-\epsilon < X_{(n)} - \theta < \epsilon) \\ &= P(\theta - \epsilon < X_{(n)} < \theta + \epsilon) \\ &= \underbrace{F_{X_{(n)}}(\theta + \epsilon)}_{=1} - F_{X_{(n)}}(\theta - \epsilon) = 1 - \left(\frac{\theta - \epsilon}{\theta}\right)^n \to 1. \end{aligned}$$

Approach 2: When the limiting random variable is a constant, say c, use Markov's Inequality; i.e., for $r \geq 1$,

$$P(|X_n - c| \geq \epsilon) \leq \frac{E(|X_n - c|^r)}{\epsilon^r}$$

and show the RHS converges to 0 as $n \to \infty$. The most common case is $r = 2$, so that

$$\begin{aligned} E[(X_n - c)^2] &= \text{var}(X_n) + [E(X_n) - c]^2 \\ &= \text{var}(X_n) + [\text{Bias}(X_n)]^2. \end{aligned}$$

Therefore, it suffices to show that both $\text{var}(X_n)$ and $\text{Bias}(X_n)$ converge to 0.

Example 5.16. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Define

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

Show that $S_n^2 \xrightarrow{p} \sigma^2$, as $n \to \infty$.
Solution. It suffices to show that $\text{var}(S_n^2)$ and $\text{Bias}(S_n^2)$ converge to 0. First note that

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

where S^2 is the "usual" sample variance (with denominator $n-1$). Therefore,

$$\begin{aligned} E(S_n^2) &= E\left[\left(\frac{n-1}{n}\right)S^2\right] = \left(\frac{n-1}{n}\right)E(S^2) = \left(\frac{n-1}{n}\right)\sigma^2 \\ \text{var}(S_n^2) &= \text{var}\left[\left(\frac{n-1}{n}\right)S^2\right] = \left(\frac{n-1}{n}\right)^2 \text{var}(S^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2} \to 0. \end{aligned}$$

Also,

$$\text{Bias}(S_n^2) = E(S_n^2 - \sigma^2) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} \to 0.$$

Approach 3: Use continuity of convergence results in conjunction with the WLLN. This approach is widely used and often allows the weakest assumptions. The following lemma can be useful when making this type of argument.

LEMMA: Suppose $X_n \xrightarrow{p} X$ and $c_n \to c$, as $n \to \infty$. Then $c_n X_n \xrightarrow{p} cX$.
Proof. Exercise.

Example 5.17. Suppose $X_1, X_2, ..., X_n$ are iid with $E(X_1^4) < \infty$. Consider the "usual" sample variance

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

From the WLLN, we have

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} E(X_1^2).$$

Also from the WLLN, we have $\overline{X}_n \xrightarrow{p} E(X_1)$, so $\overline{X}_n^2 \xrightarrow{p} [E(X_1)]^2$, by continuity. Let $c_n = n/(n-1)$ and observe that $c_n \to 1$, as $n \to \infty$. From the previous lemma, we have

$$S^2 = c_n\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right) \xrightarrow{p} 1 \times \{E(X_1^2) - [E(X_1)]^2\} = \text{var}(X_1) = \sigma^2.$$

This shows that $S^2 \xrightarrow{p} \sigma^2$ (i.e., S^2 is a consistent estimator of σ^2) under finite fourth moment assumptions. This actually remains true for iid samples under finite second moments.

Approach 4: Show that $X_n \xrightarrow{d} c$; i.e., that X_n converges in distribution to a random variable whose distribution is **degenerate** at the constant c. This approach "works" because

$$X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$$

when the limiting random variable is a constant (it is not true otherwise).

Remark: We have already seen an illustration of this approach in Example 2.19 (pp 56 notes). We showed $M_{X_n}(t) \to M_X(t)$ where the limiting random variable X had a distribution that was degenerate at the constant β. That is, the cdf of X was

$$F_X(x) = \begin{cases} 0, & x < \beta \\ 1, & x \geq \beta. \end{cases}$$

Therefore, $X_n \xrightarrow{d} \beta$ and hence $X_n \xrightarrow{p} \beta$.

5.5.2 Almost sure convergence

Definition: Suppose that $(S, \mathcal{B}, P)$ is a probability space. We say that a sequence of random variables $X_1, X_2, ...$, **converges almost surely** to a random variable X and write $X_n \xrightarrow{a.s.} X$ if $\forall \epsilon > 0$,

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = P\left(\left\{\omega \in S : \lim_{n\to\infty} |X_n(\omega) - X(\omega)| < \epsilon\right\}\right) = 1.$$

Remark: The set

$$\left\{\omega \in S : \lim_{n\to\infty} |X_n(\omega) - X(\omega)| < \epsilon\right\}$$

is the set of all outcomes $\omega \in S$ where $X_n(\omega) \to X(\omega)$. Note that $X_n(\omega)$ is a sequence of real numbers for each $\omega \in S$. Therefore, if this sequence of real numbers $X_n(\omega)$ converges to $X(\omega)$, also a real number, for **almost all** $\omega \in S$, then $X_n \xrightarrow{a.s.} X$. By "almost all," we concede that there may exist a set $N \subset S$ where convergence does not occur; i.e., $X_n(\omega) \not\to X(\omega)$, for all $\omega \in N$. However, the set N has probability 0; i.e., $P(N) = 0$.

Remark: Almost sure convergence is a very strong form of convergence (often, much stronger than is needed). In fact, the following result holds:

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X.$$

That is, almost sure convergence implies convergence in probability. The converse is not true in general; see Example 5.5.8 (CB, pp 234).

Theorem 5.5.9 (SLLN). Suppose that $X_1, X_2, ..., X_n$ is an iid sequence of random variables with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. Let

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

denote the sample mean. Then $\overline{X}_n \xrightarrow{a.s.} \mu$, as $n \to \infty$.

Continuity: Suppose $X_n \xrightarrow{a.s} X$ and let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then $h(X_n) \xrightarrow{a.s.} h(X)$. In other words, almost sure convergence is preserved under continuous mappings.

Proof. Suppose $X_n \xrightarrow{a.s.} X$ and let $S_0 = \{\omega \in S : X_n(\omega) \to X(\omega)\}$. Because $X_n \xrightarrow{a.s.} X$, we know that $P(S_0) = 1$. Because h is continuous, $h(X_n(\omega)) \to h(X(\omega))$ for all $\omega \in S_0$. We have shown that $h(X_n(\omega)) \to h(X(\omega))$ for almost all $\omega \in S$. Thus, we are done. □

Conceptualization: Suppose $\widehat{\theta}_n$ is a sequence of **estimators** for an unknown parameter, say θ. We can think of updating the value of $\widehat{\theta}_n$ as data become available (e.g., $\overline{X}_1, \overline{X}_2, \overline{X}_3, ...$). One might wish that $\widehat{\theta}_n$ become "close" to θ when n is sufficiently large and then never "wander away" again after further data collection. Almost sure convergence guarantees this. Convergence in probability does not; it guarantees only that the *probability* $\widehat{\theta}_n$ "wanders away" becomes small. This may seem like an argument against convergence in probability. However, in practice (i.e., to establish many useful asymptotic results), convergence in probability is all we will ever need. In many ways, this should not be surprising. After all, statisticians tend to think in terms of probability rather than in terms of absolutes.

5.5.3 Convergence in distribution

Definition: We say that a sequence of random variables $X_1, X_2, ...$, **converges in distribution** to a random variable X and write $X_n \xrightarrow{d} X$ if the sequence of cdfs

$$F_{X_n}(x) \to F_X(x),$$

as $n \to \infty$, for all $x \in C_{F_X}$, the set of points $x \in \mathbb{R}$ where $F_X(\cdot)$ is continuous.

Remark: When we talk about convergence in distribution, we write $X_n \xrightarrow{d} X$. However, it is important to remember that it is not the random variables themselves that are converging. It is the cdfs $F_{X_n}(x)$ that are (pointwise at all continuity points of F_X). Mathematically, $\forall \epsilon > 0\ \forall x \in C_{F_X}\ \exists n_0(\epsilon, x) \geq 1 \ni |F_{X_n}(x) - F_X(x)| < \epsilon\ \forall n \geq n_0(\epsilon, x)$.

Example 5.18. Suppose $Y_1, Y_2, ..., Y_n$ are iid exponential random variables with mean $E(Y) = 1$; i.e., the pdf of Y is $f_Y(y) = e^{-y}I(y > 0)$. Define

$$X_n = Y_{(n)} - \ln n,$$

where $Y_{(n)} = \max_{1 \leq i \leq n} Y_i$, the maximum order statistic. Show that X_n converges in distribution and find the (limiting) distribution.

Solution. The cdf of Y is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y}, & y > 0. \end{cases}$$

The cdf of $Y_{(n)}$ is

$$\begin{aligned} F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) &\stackrel{\text{iid}}{=} [P(Y_1 \leq y)]^n \\ &= [F_Y(y)]^n \\ &= (1 - e^{-y})^n, \quad \text{for } y > 0. \end{aligned}$$

The cdf of $X_n = Y_{(n)} - \ln n$ is

$$\begin{aligned} F_{X_n}(x) = P(X_n \leq x) &= P(Y_{(n)} \leq x + \ln n) \\ &= F_{Y_{(n)}}(x + \ln n) \\ &= \left[1 - e^{-(x+\ln n)}\right]^n = \left(1 - \frac{e^{-x}}{n}\right)^n \to \exp(-e^{-x}) = F_X(x), \end{aligned}$$

as $n \to \infty$. This is the cdf of a (standard) Gumbel random variable. Note that $F_X(x)$ is continuous on $\mathbb{R}$ and also that $F_{X_n}(x) \to F_X(x)$ for all $x \in \mathbb{R}$.

Continuity: Suppose $X_n \xrightarrow{d} X$ and let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then $h(X_n) \xrightarrow{d} h(X)$. In other words, convergence in distribution is preserved under continuous mappings.

Example: Suppose $X_n \xrightarrow{d} X$, where $X \sim \mathcal{N}(0, 1)$. Because $h(x) = x^2$ is continuous, $X_n^2 \xrightarrow{d} \chi^2_1$. Recall that $X \sim \mathcal{N}(0, 1) \Longrightarrow X^2 \sim \chi^2_1$.

Remark: One of the most common approaches to showing $X_n \xrightarrow{d} X$ is to use moment generating functions. In a more advanced course, we might use **characteristic functions**. Lévy's Continuity Theorem on characteristic functions says that if $X_n \sim \psi_{X_n}(t)$, then

$$X_n \xrightarrow{d} X \iff \psi_{X_n}(t) \to \psi_X(t), \text{ for all } t \in \mathbb{R}.$$

This was actually stated in the *Miscellanea* section in Chapter 2 (see Theorem 2.6.1, CB, pp 84). The "mgf version" of this result, that is,

$$X_n \xrightarrow{d} X \iff M_{X_n}(t) \to M_X(t), \text{ for all } |t| < h \ (\exists h > 0),$$

was Theorem 2.3.12 (CB, pp 66). Of course, this result is applicable only when mgfs exist (characteristic functions, on the other hand, always exist).

Theorem 5.5.12. $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$.

Remark: The converse to Theorem 5.5.12 is not true in general. Suppose that $X_n \sim \mathcal{N}(0,1)$ for all n. Suppose that $X \sim \mathcal{N}(0,1)$. Clearly, $X_n \xrightarrow{d} X$. Why? The cdf of X_n is

$$F_{X_n}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = F_X(x), \quad \text{for each } n.$$

Trivially, $F_{X_n}(x) \to F_X(x)$, for all $x \in \mathbb{R}$. However, there is no guarantee that X_n will ever be "close" to X with high probability. For example, if $X_n \perp\!\!\!\perp X$, then $Y = X_n - X \sim \mathcal{N}(0,2)$. For $\epsilon > 0$, $P(|X_n - X| < \epsilon) = P(|Y| < \epsilon)$, a constant. This does not converge to 1.

Remark: The converse to Theorem 5.5.12 is true when the limiting random variable is a constant (see "Approach 4" in the convergence in probability subsection):

$$X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c.$$

This is precisely what is stated in Theorem 5.5.13 (CB, pp 236).

Theorem 5.5.14 (Central Limit Theorem, CLT). Suppose $X_1, X_2, ...,$ is an iid sequence of random variables with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. Then

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

Remark: Note that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}.$$

Showing this is simple algebra. In more applied courses, it is common to write things like

$$\overline{X}_n \sim \mathcal{AN}(\mu, \sigma^2/n) \quad \text{or} \quad \sum_{i=1}^n X_i \sim \mathcal{AN}(n\mu, n\sigma^2)$$

for large n. However, **do not ever write something like**

$$\overline{X}_n \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/n) \quad \text{or} \quad \sum_{i=1}^n X_i \xrightarrow{d} \mathcal{N}(n\mu, n\sigma^2).$$

These statements are not true, and, in fact, do not even make sense mathematically. Convergence in distribution is a statement about what happens when $n \to \infty$. The distribution in the limit can not depend on n, the quantity that is going off to infinity.

Example 5.19. Suppose $X_1, X_2, ..., X_n$ are iid χ^2_1 so that $E(X_1) = \mu = 1$ and $\text{var}(X_1) = \sigma^2 = 2$. The CLT says that

$$Z_n = \frac{\overline{X}_n - 1}{\sqrt{2/n}} = \frac{\sum_{i=1}^n X_i - n}{\sqrt{2n}} \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

Illustration: Suppose $n = 100$. We would expect the distribution of $\overline{X}_{100}$ to be well approximated by a $\mathcal{N}(1, 2/100)$ distribution, or, equivalently, the distribution of $\sum_{i=1}^{100} X_i$ to be well approximated by a $\mathcal{N}(100, 200)$ distribution.

Remark: To prove the CLT, we will assume that the mgf of X_i exists. This assumption is not necessary, but it does make the proof easier. A more general proof would involve characteristic functions (which always exist).

Proof of Theorem 5.5.14: Let

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

We will show $M_{Z_n}(t)$, the mgf of Z_n, converges pointwise to $M_Z(t) = e^{t^2/2}$, the mgf of $Z \sim \mathcal{N}(0, 1)$. Define

$$Y_i = \frac{X_i - \mu}{\sigma},$$

for $i = 1, 2, ..., n$, and let $M_Y(t)$ denote the common mgf of Y.

Notes:

- If the X_i's are iid, then so are the Y_i's.
- If the mgf of X_i exists for all $t \in (-h, h)$ $\exists h > 0$, then the mgf of Y_i exists for all $t \in (-\sigma h, \sigma h)$.
- Note that $E(Y_i) = 0$ and $\text{var}(Y_i) = 1$ by construction.

Simple algebra yields

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i.$$

Therefore,

$$\begin{aligned} M_{Z_n}(t) = E(e^{tZ_n}) &= E\big(e^{\frac{t}{\sqrt{n}} \sum_{i=1}^n Y_i}\big) \\ &= E\big(e^{\frac{t}{\sqrt{n}} Y_1} e^{\frac{t}{\sqrt{n}} Y_2} \cdots e^{\frac{t}{\sqrt{n}} Y_n}\big) \\ &\overset{\text{indep}}{=} E\big(e^{\frac{t}{\sqrt{n}} Y_1}\big) E\big(e^{\frac{t}{\sqrt{n}} Y_2}\big) \cdots E\big(e^{\frac{t}{\sqrt{n}} Y_n}\big) \\ &\overset{\text{ident}}{=} \big[E\big(e^{\frac{t}{\sqrt{n}} Y_1}\big)\big]^n \\ &= \big[M_Y(t/\sqrt{n})\big]^n. \end{aligned}$$

Now write $M_Y(t/\sqrt{n})$ in its McLaurin series expansion:

$$M_Y(t/\sqrt{n}) = \sum_{k=0}^{\infty} M_Y^{(k)}(0) \frac{\big(\frac{t}{\sqrt{n}} - 0\big)^k}{k!},$$

where

$$M_Y^{(k)}(0) = \frac{d^k}{dt^k} M_Y(t)\bigg|_{t=0}.$$

Because $M_Y(t)$ exists $\forall t \in (-\sigma h, \sigma h)$, this expansion is valid $\forall t \in (-\sqrt{n}\sigma h, \sqrt{n}\sigma h)$. Now,

$$\begin{aligned} M_Y^{(0)}(0) &= M_Y(0) = 1 \\ M_Y^{(1)}(0) &= E(Y) = 0 \\ M_Y^{(2)}(0) &= E(Y^2) = 1. \end{aligned}$$

Therefore, the expansion above becomes

$$M_Y(t/\sqrt{n}) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_Y(t/\sqrt{n}),$$

where the remainder term

$$R_Y(t/\sqrt{n}) = \sum_{k=3}^{\infty} M_Y^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!}.$$

To summarize, we have written

$$M_{Z_n}(t) = \left[M_Y(t/\sqrt{n})\right]^n = \left[1 + \frac{t^2/2}{n} + R_Y(t/\sqrt{n})\right]^n = \left[1 + \frac{b}{n} + \frac{g(n)}{n}\right]^{cn},$$

where $b = t^2/2$, $g(n) = nR_Y(t/\sqrt{n})$, and $c = 1$. It therefore suffices to show that $\lim_{n\to\infty} g(n) = \lim_{n\to\infty} nR_Y(t/\sqrt{n}) = 0$, for all $t \in \mathbb{R}$. An application of Taylor's Theorem (see Theorem 5.5.21, CB, pp 241) yields

$$\lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0$$

for $t \neq 0$ (fixed). Because t is fixed, we also have (for $t \neq 0$)

$$0 = \lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(t/\sqrt{n})^2} = \lim_{n\to\infty} \frac{R_Y(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n\to\infty} nR_Y(t/\sqrt{n}).$$

Also $\lim_{n\to\infty} nR_Y(t/\sqrt{n}) = 0$ when $t = 0$ because $R_Y(0) = 0$. We have shown that $\lim_{n\to\infty} nR_Y(t/\sqrt{n}) = 0$, for all $t \in \mathbb{R}$. Thus, we are done. □

Remark: We have the following important result:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \iff \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Statements like the second statement are commonly seen in asymptotic results. We can interpret this as follows. If we

1. center $\overline{X}_n$ by subtracting μ
2. scale $\overline{X}_n - \mu$ up by multiplying by $\sqrt{n}$,

then $\sqrt{n}(\overline{X}_n - \mu)$ converges to a bonafide distribution. The sequence $n^p(\overline{X}_n - \mu)$ collapses if $p < 1/2$ and blows up if $p > 1/2$. The value $p = 1/2$ is "just right" to ensure $n^p(\overline{X}_n - \mu)$ converges to a nondegenerate distribution with no probability "escaping off" to $\pm\infty$.

Example 5.20. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$, so that $E(X_1) = p$ and $\text{var}(X_1) = p(1-p)$. The CLT says that

$$\sqrt{n}(\overline{X}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)),$$

as $n \to \infty$. For the Bernoulli population distribution, the X_i's are zeros and ones, so $\overline{X}_n$ is a **sample proportion** (i.e., the proportion of ones in the sample). More familiar notation for the sample proportion is $\widehat{p}$, as presented in Example 5.14 (notes). This result restated is

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)) \iff \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

5.5.4 Slutsky's Theorem

Theorem 5.5.17 (Slutsky's Theorem). Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} a$, where a is a constant. Then

(a) $Y_n X_n \xrightarrow{d} aX$

(b) $X_n + Y_n \xrightarrow{d} X + a$.

Example 5.21. Suppose $X_1, X_2, ..., X_n$ is an iid sample with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. The CLT says

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. Let S^2 denote the sample variance. In Example 5.17 (notes), we showed that $S^2 \xrightarrow{p} \sigma^2$, as $n \to \infty$. Because $h(x) = \sigma/\sqrt{x}$ is continuous over $\mathbb{R}^+$,

$$\frac{\sigma}{S} \xrightarrow{p} 1.$$

Therefore, by Slutsky's Theorem,

$$\frac{\overline{X}_n - \mu}{S/\sqrt{n}} = \underbrace{\frac{\sigma}{S}}_{\xrightarrow{p} 1} \underbrace{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}_{\xrightarrow{d} \mathcal{N}(0,1)} \xrightarrow{d} \mathcal{N}(0,1).$$

Note that replacing σ (an unknown parameter) with S (a consistent estimate of σ) does not affect the asymptotic distribution.

Exercise: Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Show that

$$\frac{\widehat{p} - p}{\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

5.5.5 Delta Method

Theorem 5.5.24 (Delta Method). Suppose X_n is a sequence of random variables satisfying

$$\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

as $n \to \infty$. Suppose $g : \mathbb{R} \to \mathbb{R}$ is differentiable at θ and $g'(\theta) \neq 0$. Then

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{d} \mathcal{N}\left(0, [g'(\theta)]^2\sigma^2\right),$$

as $n \to \infty$. In other words,

$$g(X_n) \sim \mathcal{AN}\left(g(\theta), \frac{[g'(\theta)]^2\sigma^2}{n}\right), \text{ for large } n.$$

Proof. Write $g(X_n)$ in a (stochastic) Taylor series expansion about θ:

$$g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + \frac{g''(\xi_n)}{2}(X_n - \theta)^2,$$

where ξ_n is between X_n and θ. Multiplying by $\sqrt{n}$ and then rearranging, we have

$$\sqrt{n}[g(X_n) - g(\theta)] = g'(\theta)\underbrace{\sqrt{n}(X_n - \theta)}_{\xrightarrow{d}\mathcal{N}(0,\sigma^2)} + \underbrace{\frac{\sqrt{n}g''(\xi_n)}{2}(X_n - \theta)^2}_{=R_n}.$$

Now, $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by assumption so

$$g'(\theta)\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}\left(0, [g'(\theta)]^2\sigma^2\right).$$

Therefore, if we can show $R_n \xrightarrow{p} 0$, then the result will follow from Slutsky's Theorem. Note that

$$R_n = \frac{g''(\xi_n)}{2}(X_n - \theta)\underbrace{\sqrt{n}(X_n - \theta)}_{\xrightarrow{d}\mathcal{N}(0,\sigma^2)}.$$

Provided that $g''(\xi_n)$ converges to something finite, we can get $R_n \xrightarrow{p} 0$ if we can show $X_n - \theta \xrightarrow{p} 0$. Suppose $\epsilon > 0$. Consider

$$\lim_{n\to\infty} P(|X_n - \theta| \geq \epsilon) = \lim_{n\to\infty} P(\sqrt{n}|X_n - \theta| \geq \sqrt{n}\epsilon).$$

We know that $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by assumption, so

$$\sqrt{n}|X_n - \theta| = |\sqrt{n}(X_n - \theta)| \xrightarrow{d} |X|,$$

where $X \sim \mathcal{N}(0, \sigma^2)$, by continuity; note that $h(x) = |x|$ is continuous on $\mathbb{R}$. Because the distribution of $|X|$ does not have probability "escaping off" to $+\infty$, we have

$$\lim_{n\to\infty} P(|X_n - \theta| \geq \epsilon) = \lim_{n\to\infty} P(\underbrace{\sqrt{n}|X_n - \theta|}_{\xrightarrow{d}|X|} \geq \sqrt{n}\epsilon) = 0.$$

Therefore, $X_n \xrightarrow{p} \theta$. By continuity, $X_n - \theta \xrightarrow{p} 0$. Finally,

$$\sqrt{n}[g(X_n) - g(\theta)] = \underbrace{g'(\theta)\sqrt{n}(X_n - \theta)}_{\xrightarrow{d}\mathcal{N}(0,[g'(\theta)]^2\sigma^2)} + \underbrace{\frac{\sqrt{n}g''(\xi_n)}{2}(X_n - \theta)^2}_{\xrightarrow{p}0}.$$

Applying Slutsky's to the RHS gives the result. □

Example 5.22. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Recall that the CLT gives

$$\sqrt{n}(\widehat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)),$$

as $n \to \infty$. We now find the asymptotic distribution of the **log-odds**

$$g(\widehat{p}) = \ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right),$$

properly centered and scaled. Note that $g(p)$ is differentiable over $0 < p < 1$ and

$$g(p) = \ln\left(\frac{p}{1-p}\right) \implies g'(p) = \frac{1}{p(1-p)},$$

which never equals zero over $(0, 1)$. Therefore, the delta method applies and

$$\begin{aligned}\sqrt{n}\left[\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) - \ln\left(\frac{p}{1-p}\right)\right] &\xrightarrow{d} \mathcal{N}\left(0, \left[\frac{1}{p(1-p)}\right]^2 p(1-p)\right) \\ &\stackrel{d}{=} \mathcal{N}\left(0, \frac{1}{p(1-p)}\right),\end{aligned}$$

as $n \to \infty$. In other words,

$$\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \sim \mathcal{AN}\left(\ln\left(\frac{p}{1-p}\right), \frac{1}{np(1-p)}\right), \quad \text{for large } n.$$

Example 5.23. Suppose $X_1, X_2, ..., X_n$ are iid Poisson(θ), where $\theta > 0$, so that $E(X_1) = \theta$ and $\text{var}(X_1) = \theta$. The CLT says that

$$\sqrt{n}(\overline{X}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \theta),$$

as $n \to \infty$. Find a function of $\overline{X}_n$, say $g(\overline{X}_n)$, whose asymptotic variance is free of θ.
Solution. The delta method says that

$$\sqrt{n}[g(\overline{X}_n) - g(\theta)] \xrightarrow{d} \mathcal{N}\left(0, [g'(\theta)]^2\theta\right),$$

as $n \to \infty$. If we set the large sample variance $[g'(\theta)]^2\theta$ equal to a constant (free of θ), we can identify the function g that satisfies this equation; that is,

$$[g'(\theta)]^2\theta \stackrel{\text{set}}{=} c_0 \implies [g'(\theta)]^2 = \frac{c_0}{\theta} \implies g'(\theta) = \frac{c_1}{\sqrt{\theta}},$$

where $c_1 = \sqrt{c_0}$ (also free of θ). A solution to this first-order differential equation is

$$g(\theta) = \int \frac{c_1}{\sqrt{\theta}}\, d\theta \;=\; 2c_1\sqrt{\theta} + c_2,$$

where c_2 is a constant free of θ. Taking $c_1 = 1/2$ and $c_2 = 0$ yields $g(\theta) = \sqrt{\theta}$.

Claim: The function $g(\overline{X}_n) = \sqrt{\overline{X}_n}$ has asymptotic variance that is free of θ.
Proof. We have

$$g(\theta) = \sqrt{\theta} \implies g'(\theta) = \frac{1}{2\sqrt{\theta}},$$

which never equals zero (because $\theta > 0$). Therefore, the delta method says

$$\sqrt{n}\left(\sqrt{\overline{X}_n} - \sqrt{\theta}\right) \xrightarrow{d} \mathcal{N}\left(0, \left[\frac{1}{2\sqrt{\theta}}\right]^2 \theta\right) \;\overset{d}{=}\; \mathcal{N}\left(0, \frac{1}{4}\right),$$

as $n \to \infty$. In other words,

$$\sqrt{\overline{X}_n} \sim \mathcal{AN}\left(\sqrt{\theta}, \frac{1}{4n}\right), \quad \text{for large } n.$$

In this example, we see that a **square root transformation** "stabilizes" the asymptotic variance of $\overline{X}_n$.

Exercise: Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli(p), where $0 < p < 1$. Find a function of $\widehat{p}$, say $g(\widehat{p})$, whose asymptotic variance is free of p.
Ans: $g(\widehat{p}) = \arcsin\sqrt{\widehat{p}}$.

Theorem 5.5.26 (Second-order Delta Method). Suppose X_n is a sequence of random variables satisfying

$$\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

as $n \to \infty$. Suppose $g : \mathbb{R} \to \mathbb{R}$ is twice differentiable at θ, $g'(\theta) = 0$, and $g''(\theta) \neq 0$. Then

$$\begin{aligned} n[g(X_n) - g(\theta)] &\xrightarrow{d} \frac{\sigma^2}{2} g''(\theta)\chi^2_1 \\ &\overset{d}{=} \text{gamma}(1/2, \sigma^2 g''(\theta)). \end{aligned}$$

Example 5.24. Suppose $X_1, X_2, ..., X_n$ are iid with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. The CLT guarantees that

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

as $n \to \infty$. Consider $g(\overline{X}_n) = \overline{X}^2_n$. With $g(\mu) = \mu^2$, we have $g'(\mu) = 2\mu$, which is nonzero except when $\mu = 0$. Therefore, provided that $\mu \neq 0$,

$$\sqrt{n}(\overline{X}^2_n - \mu^2) \xrightarrow{d} \mathcal{N}(0, 4\mu^2\sigma^2),$$

by the (first-order) delta method. If $\mu = 0$, then the previous asymptotic distribution collapses. However, we can apply the second-order delta method. Note that $g''(\mu) = 2$ and therefore (when $\mu = 0$),

$$\begin{aligned} n(\overline{X}_n^2 - \mu^2) = n\overline{X}_n^2 &\xrightarrow{d} \sigma^2\chi_1^2 \\ &\stackrel{d}{=} \text{gamma}(1/2, 2\sigma^2). \end{aligned}$$

5.5.6 Multivariate extensions

Remark: We now briefly discuss asymptotic results for multivariate random vectors. All convergence concepts can be extended to handle sequences of random vectors.

Central Limit Theorem: Suppose $\mathbf{X}_1, \mathbf{X}_2, ...,$ is a sequence of iid random vectors (of dimension k) with $E(\mathbf{X}_1) = \boldsymbol{\mu}_{k\times 1}$ and $\text{cov}(\mathbf{X}_1) = \boldsymbol{\Sigma}_{k\times k}$. Let $\overline{\mathbf{X}}_n = (\overline{X}_{1+}, \overline{X}_{2+}, ..., \overline{X}_{k+})'$ denote the vector of sample means. Then $\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \text{mvn}_k(\mathbf{0}, \boldsymbol{\Sigma})$.

Multivariate Delta Method: Suppose $\mathbf{X}_n$ is a sequence of random vectors (of dimension k) satisfying

(A1) $\sqrt{n}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \text{mvn}_k(\mathbf{0}, \boldsymbol{\Sigma})$

(A2) $g : \mathbb{R}^k \to \mathbb{R}$ is differentiable at $\boldsymbol{\mu}$ (and is not zero).

Then

$$\sqrt{n}[g(\mathbf{X}_n) - g(\boldsymbol{\mu})] \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}} \boldsymbol{\Sigma} \frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}'}\right),$$

where

$$\frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}} = \left(\frac{\partial g(\mathbf{x})}{\partial x_1}, \frac{\partial g(\mathbf{x})}{\partial x_2}, ..., \frac{\partial g(\mathbf{x})}{\partial x_k}\right)\Bigg|_{\mathbf{x}=\boldsymbol{\mu}}.$$

Remark: The limiting distribution stated in the multivariate delta method is a univariate normal distribution. Note that $g : \mathbb{R}^k \to \mathbb{R}$, so $g(\mathbf{X}_n)$ is a scalar random variable. The quantity

$$\frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}} \boldsymbol{\Sigma} \frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}'} = \text{ a scalar.}$$

Remark: The multivariate delta method can be generalized further to allow for functions $g : \mathbb{R}^k \to \mathbb{R}^p$, where $p \leq k$; that is, g itself is vector valued. The only difference is that now

$$\frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_k} \\ \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_2(\mathbf{x})}{\partial x_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_p(\mathbf{x})}{\partial x_1} & \frac{\partial g_p(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_p(\mathbf{x})}{\partial x_k} \end{pmatrix}_{p\times k},$$

in which case $g(\boldsymbol{\mu})$ is $p \times 1$ and

$$\frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}} \boldsymbol{\Sigma} \frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}'} = \ p \times p \text{ matrix.}$$

Example 5.25. Suppose $\mathbf{X} = (X_1, X_2)'$ is a continuous random vector with joint pdf

$$f_{\mathbf{X}}(\mathbf{x}) = e^{-x_2} I(0 < x_1 < x_2 < \infty).$$

In Example 4.7 (notes), we showed that

$$\begin{aligned} X_1 &\sim \text{exponential}(1) \\ X_2 &\sim \text{gamma}(2, 1). \end{aligned}$$

We have $E(X_1) = 1$, $E(X_2) = 2$, $\text{var}(X_1) = 1$, and $\text{var}(X_2) = 2$. Also,

$$E(X_1 X_2) = \int_{\mathbb{R}^2} \int x_1 x_2 f_{X_1, X_2}(x_1, x_2) dx_2 dx_1 = 3$$

so

$$\text{cov}(X_1, X_2) = E(X_1 X_2) - E(X_1)E(X_2) = 3 - 2 = 1.$$

Therefore, for the population described by $f_{\mathbf{X}}(\mathbf{x})$, we have

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}_{2\times 1} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}_{2\times 2}.$$

Now suppose $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ is an iid sample from $f_{\mathbf{X}}(\mathbf{x})$; i.e., $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ are mutually independent and $\mathbf{X}_j = (X_{1j}, X_{2j}) \sim f_{\mathbf{X}}(\mathbf{x})$, for $j = 1, 2, ..., n$. Define

$$\overline{X}_{1+} = \frac{1}{n} \sum_{j=1}^{n} X_{1j} \quad \text{and} \quad \overline{X}_{2+} = \frac{1}{n} \sum_{j=1}^{n} X_{2j}$$

and denote by

$$\overline{\mathbf{X}}_n = \begin{pmatrix} \overline{X}_{1+} \\ \overline{X}_{2+} \end{pmatrix},$$

the vector of sample means. The (multivariate) CLT says that

$$\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \text{mvn}_2(\mathbf{0}, \boldsymbol{\Sigma}),$$

as $n \to \infty$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given above. In other words,

$$\overline{\mathbf{X}}_n \sim \mathcal{AN}_2 \left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \ \begin{pmatrix} 1/n & 1/n \\ 1/n & 2/n \end{pmatrix} \right),$$

for large n.

Q: Find the large sample distribution of

$$R = g(\overline{\mathbf{X}}_n) = \frac{\overline{X}_{1+}}{\overline{X}_{2+}},$$

suitably centered and scaled.
Solution: With $g(x_1, x_2) = x_1/x_2$, we have

$$\frac{\partial g(x_1, x_2)}{\partial x_1} = \frac{1}{x_2} \quad \text{and} \quad \frac{\partial g(x_1, x_2)}{\partial x_2} = -\frac{x_1}{x_2^2}$$

so that

$$\frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{1}{\mu_2} & -\frac{\mu_1}{\mu_2^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} \end{pmatrix}.$$

The multivariate delta method says that

$$\sqrt{n}[g(\overline{\mathbf{X}}_n) - g(\boldsymbol{\mu})] = \sqrt{n}\left(R - \frac{1}{2}\right) \xrightarrow{d} \mathcal{N}(0, \sigma_R^2),$$

as $n \to \infty$, where

$$\sigma_R^2 = \frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}} \boldsymbol{\Sigma} \frac{\partial g(\boldsymbol{\mu})}{\partial \mathbf{x}'} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{4} \end{pmatrix} = \frac{1}{8}.$$

In other words,

$$R \sim \mathcal{AN}\left(\frac{1}{2}, \frac{1}{8n}\right), \quad \text{for large } n.$$

Example 5.26. In medical settings, it is common to observe data in the form of 2×2 tables such as

	Cured	Not cured
Group 1	X_{11}	X_{12}
Group 2	X_{21}	X_{22}

Set $\mathbf{X} = (X_{11}, X_{12}, X_{21}, X_{22})$ and assume $\mathbf{X} \sim \text{mult}(n; p_{11}, p_{12}, p_{21}, p_{22})$. Note that

$$\mathbf{X} = \sum_{k=1}^{n} \mathbf{Y}_k,$$

where $\mathbf{Y}_k = (Y_{11k}, Y_{12k}, Y_{21k}, Y_{22k})$ and $Y_{ijk} = I(k\text{th individual is in cell } ij)$, for $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, ..., n$. In other words, $\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_n$ are iid $\text{mult}(1; p_{11}, p_{12}, p_{21}, p_{22})$ with

$$\boldsymbol{\mu} = E(\mathbf{Y}_1) = \begin{pmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \end{pmatrix} = \mathbf{p}$$

and

$$\begin{aligned}\boldsymbol{\Sigma} = \text{cov}(\mathbf{Y}_1) &= \begin{pmatrix} p_{11}(1-p_{11}) & -p_{11}p_{12} & -p_{11}p_{21} & -p_{11}p_{22} \\ -p_{12}p_{11} & p_{12}(1-p_{12}) & -p_{12}p_{21} & -p_{12}p_{22} \\ -p_{21}p_{11} & -p_{21}p_{12} & p_{21}(1-p_{21}) & -p_{21}p_{22} \\ -p_{22}p_{11} & -p_{22}p_{12} & -p_{22}p_{21} & p_{22}(1-p_{22}) \end{pmatrix} \\ &= \mathbf{D}(\mathbf{p}) - \mathbf{p}\mathbf{p}', \end{aligned}$$

where $\mathbf{D}(\mathbf{p}) = \text{diag}(\mathbf{p})$. Define the vector of sample proportions as

$$\widehat{\mathbf{p}} = \frac{\mathbf{X}}{n} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{Y}_k.$$

The (multivariate) CLT says that $\sqrt{n}(\widehat{\mathbf{p}} - \mathbf{p}) \xrightarrow{d} \text{mvn}(\mathbf{0}, \boldsymbol{\Sigma})$, as $n \to \infty$. This is a less-than-full-rank normal distribution (STAT 714) because $r(\mathbf{D}(\mathbf{p}) - \mathbf{p}\mathbf{p}') = 3 < 4$.

Q: Find the large sample distribution of the **log-odds ratio**

$$g(\widehat{\mathbf{p}}) = \ln\left(\frac{\widehat{p}_{11}\widehat{p}_{22}}{\widehat{p}_{12}\widehat{p}_{21}}\right),$$

suitably centered and scaled.
Solution: With

$$g(\mathbf{p}) = \ln\left(\frac{p_{11}p_{22}}{p_{12}p_{21}}\right),$$

we have

$$\frac{\partial g(\mathbf{p})}{\partial \mathbf{p}} = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} p_{11}^{-1} & 0 & 0 & 0 \\ 0 & p_{12}^{-1} & 0 & 0 \\ 0 & 0 & p_{21}^{-1} & 0 \\ 0 & 0 & 0 & p_{22}^{-1} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \mathbf{D}^{-1}(\mathbf{p}).$$

By the multivariate delta method,

$$\sqrt{n}[g(\widehat{\mathbf{p}}) - g(\mathbf{p})] \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

as $n \to \infty$, where

$$\begin{aligned}\sigma^2 &= \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \mathbf{D}^{-1}(\mathbf{p})[\mathbf{D}(\mathbf{p}) - \mathbf{p}\mathbf{p}']\mathbf{D}^{-1}(\mathbf{p}) \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix}' \\ &= \frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}. \end{aligned}$$

In other words,

$$\ln\left(\frac{\widehat{p}_{11}\widehat{p}_{22}}{\widehat{p}_{12}\widehat{p}_{21}}\right) \sim \mathcal{AN}\left(\ln\left(\frac{p_{11}p_{22}}{p_{12}p_{21}}\right), \frac{1}{n}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{1}{p_{ij}}\right),$$

for large n.