STAT 712/713
MATHEMATICAL STATISTICS I/II

Fall 2015/Spring 2016

Lecture Notes

Joshua M. Tebbs
Department of Statistics
University of South Carolina

© by Joshua M. Tebbs
Contents

1 Probability Theory 1
   1.1 Set Theory ................................................. 1
   1.2 Basics of Probability Theory ................................. 6
   1.3 Conditional Probability and Independence .................... 15
   1.4 Random Variables ............................................ 19
   1.5 Distribution Functions ....................................... 22
   1.6 Density and Mass Functions .................................. 26

2 Transformations and Expectations 32
   2.1 Distributions of Functions of a Random Variable ............. 32
   2.2 Expected Values .............................................. 41
   2.3 Moments and Moment Generating Functions .................... 45

3 Common Families of Distributions 57
   3.1 Introduction ................................................... 57
   3.2 Discrete Distributions ........................................ 57
   3.3 Continuous Distributions ...................................... 63
   3.4 Exponential Families .......................................... 71
   3.5 Location and Scale Families .................................. 74
   3.6 Inequalities and Identities .................................. 77

4 Multiple Random Variables 78
   4.1 Joint and Marginal Distributions ............................... 78
   4.2 Conditional Distributions and Independence ................... 88
   4.3 Bivariate Transformations ..................................... 99
   4.4 Hierarchical Models and Mixture Distributions ............... 108
   4.5 Covariance and Correlation ................................... 113
   4.6 Multivariate Distributions .................................... 116
   4.7 Inequalities .................................................... 126

5 Properties of a Random Sample 128
   5.1 Basic Concepts of a Random Sample ............................ 128
   5.2 Sums of Random Variables from a Random Sample ............ 130
5.3 Sampling from the Normal Distribution ........................................ 137
5.4 Order Statistics ................................................................. 144
5.5 Convergence Concepts .......................................................... 151
  5.5.1 Convergence in probability .................................................. 152
  5.5.2 Almost sure convergence ..................................................... 158
  5.5.3 Convergence in distribution .................................................. 159
  5.5.4 Slutsky’s Theorem ............................................................ 163
  5.5.5 Delta Method ................................................................. 164
  5.5.6 Multivariate extensions ...................................................... 167

6 Principles of Data Reduction ......................................................... 171
  6.1 Introduction ................................................................. 171
  6.2 The Sufficiency Principle ....................................................... 172
    6.2.1 Sufficient statistics ......................................................... 172
    6.2.2 Minimal sufficient statistics .............................................. 181
    6.2.3 Ancillary statistics ........................................................ 183
    6.2.4 Sufficient, ancillary, and complete statistics ......................... 187

7 Point Estimation ................................................................. 195
  7.1 Introduction ................................................................. 195
  7.2 Methods of Finding Estimators ............................................... 196
    7.2.1 Method of moments ....................................................... 196
    7.2.2 Maximum likelihood estimation ......................................... 198
    7.2.3 Bayesian estimation ...................................................... 206
  7.3 Methods of Evaluating Estimators ........................................... 211
    7.3.1 Bias, variance, and MSE .................................................. 211
    7.3.2 Best unbiased estimators ................................................. 213
    7.3.3 Sufficiency and completeness .......................................... 220
  7.4 Appendix: CRLB Theory ....................................................... 226

8 Hypothesis Testing ............................................................... 232
  8.1 Introduction ................................................................. 232
  8.2 Methods of Finding Tests .................................................... 235
    8.2.1 Likelihood ratio tests .................................................... 235
# TABLE OF CONTENTS

8.2.2 Bayesian tests ................................................. 243
8.3 Methods of Evaluating Tests .................................... 245
  8.3.1 Error probabilities and the power function ................. 245
  8.3.2 Most powerful tests ....................................... 251
  8.3.3 Uniformly most powerful tests ............................ 257
  8.3.4 Probability values ....................................... 268

9 Interval Estimation .................................................. 271
  9.1 Introduction ................................................ 271
  9.2 Methods of Finding Interval Estimators ......................... 274
    9.2.1 Inverting a test statistic ............................. 274
    9.2.2 Pivotal quantities .................................... 277
    9.2.3 Pivoting the CDF ..................................... 281
    9.2.4 Bayesian intervals .................................. 285
  9.3 Methods of Evaluating Interval Estimators .................... 287

10 Asymptotic Evaluations ........................................... 290
  10.1 Introduction ............................................... 290
  10.2 Point Estimation .......................................... 290
  10.3 Hypothesis Testing ......................................... 300
    10.3.1 Wald tests ........................................ 301
    10.3.2 Score tests ....................................... 303
    10.3.3 Likelihood ratio tests ............................. 305
  10.4 Confidence Intervals ....................................... 310
    10.4.1 Wald intervals .................................... 311
    10.4.2 Score intervals ................................... 313
    10.4.3 Likelihood ratio intervals ......................... 315
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:

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

(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) $S = \{(HHH), (HHT), ..., (TTT)\}$ 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,
• $\mathbb{N} = \{1, 2, 3, \ldots\}$ 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

$$\omega \in A \implies \omega \in B.$$ 

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,

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

**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:**

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

(b) **Associativity:**

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

(c) **Distributive Laws:**

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

(d) **DeMorgan’s Laws:**

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$
Extension: Suppose that $A_1, A_2, \ldots, A_n$ is a finite sequence of sets, where each $A_i \subseteq S$.

- **Union:**
  \[
  \bigcup_{i=1}^{n} A_i = \{ \omega \in S : \omega \in A_i \text{ } \exists i \}\]

- **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 $A_1, A_2, \ldots$, 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 $A_1, A_2, \ldots$, are pairwise disjoint (or pairwise mutually exclusive) if
\[ A_i \cap A_j = \emptyset \quad \forall i \neq j. \]

**Definition:** Suppose $A_1, A_2, \ldots$, are subsets of $S$. We say that $A_1, A_2, \ldots$, 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 $i = 1, 2, \ldots$. Clearly, the $A_i$’s are pairwise disjoint and $\bigcup_{i=1}^{\infty} A_i = S$, so the sequence $A_1, A_2, \ldots$, 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:

$$B_1 = \bigcup_{k=1}^{\infty} A_k$$

$$B_2 = \bigcup_{k=2}^{\infty} A_k \quad \text{Note: } B_2 \subseteq B_1$$

$$B_3 = \bigcup_{k=3}^{\infty} A_k \quad \text{Note: } B_3 \subseteq B_2$$

$$\vdots$$

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,

$$\lim_{n \to \infty} B_n = \bigcap_{n=1}^{\infty} B_n$$

$$= \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \equiv \limsup_{n} A_n = \overline{\lim}_{n} A_n.$$

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

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

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

$$\omega \in \lim_{n} A_n \iff \omega \in A_n \text{ i.o. (infinitely often)}.$$
Now, let’s return to our arbitrary sequence of sets \( \{A_n\} \). Consider the following:

\[
C_1 = \bigcap_{k=1}^{\infty} A_k \\
C_2 = \bigcap_{k=2}^{\infty} A_k \quad \text{Note: } C_2 \supseteq C_1 \\
C_3 = \bigcap_{k=3}^{\infty} A_k \quad \text{Note: } C_3 \supseteq C_2 \\
\vdots
\]

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,

\[
\lim_{n \to \infty} C_n = \bigcup_{n=1}^{\infty} C_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k \equiv \lim inf_n A_n = \lim A_n.
\]

**Interpretation:** For a sequence of sets \( \{A_n\} \),

\[
\lim A_n = \{\omega \in S : \exists n \geq 1 \exists \omega \in A_k \forall k \geq n\}.
\]

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

\[
\omega \in \lim A_n \iff \omega \in A_n \text{ e.v. (eventually)}.
\]

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

\[
\lim A_n = \overline{\lim} A_n
\]

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

\[
\lim A_n \neq \overline{\lim} A_n,
\]

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 \( \lim A_n \) 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 $B$ denote a collection of subsets of $S$. We say that $B$ is a $\sigma$-algebra on $S$ if

(i) $\emptyset \in B$

(ii) $A \in B \implies A^c \in B$; i.e., $B$ is closed under complementation.

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

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

**Results:** Suppose that $B$ is a $\sigma$-algebra on $S$.

- Because $S = \emptyset^c$, we see that $S \in B$ (because $B$ is closed under complementation).
- If $A_1, A_2, \ldots, \in B$, then $A_1^c, A_2^c, \ldots, \in B$; see Property (ii);

\[
\implies \bigcup_{i=1}^{\infty} A_i^c \in B \quad \text{Property (iii)}
\]

\[
\implies \left( \bigcup_{i=1}^{\infty} A_i^c \right)^c = \bigcap_{i=1}^{\infty} A_i \in B \quad \text{DeMorgan’s Law; Property (ii)}
\]

Therefore, a $\sigma$-algebra $B$ is also closed under countable intersections.

**Examples:**

1. $B = \{\emptyset, S\}$. This is called the “trivial $\sigma$-algebra.”

2. $B = \{\emptyset, A, A^c, S\}$. This is the smallest $\sigma$-algebra that contains $A$; denoted by $\sigma(A)$. 

PAGE 6
3. $|S| = n$; i.e., the sample space is finite and contains $n$ outcomes. Define

$$B = 2^S$$

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

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

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

Note that $B = 2^S$ contains $2^3 = 8$ sets.

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

$$B_0 = \{\emptyset, S\}$$
$$B_1 = \{\emptyset, \{1\}, \{2, 3\}, S\}$$

are also $\sigma$-algebras on $S$. Note also that $B_0 \subset B_1 \subset B = 2^S$. We call $B_0$ and $B_1$ sub-$\sigma$-algebras of $S$.

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

$$B_1 = \{\emptyset, \{1\}, \{2, 3\}, S\},$$

we could assign a probability to the set $\{2, 3\}$ because it is measurable with respect to this $\sigma$-algebra. However, we could not assign a probability to the event $\{2\}$ using $B_1$; it is not $B_1$-measurable. Of course, $\{2\}$ is $B$-measurable.

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

- $\mathcal{A} \subset \sigma(\mathcal{A})$, and
- if $\Sigma(\mathcal{A})$ is another $\sigma$-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 $\sigma$-algebra on $\mathbb{R}$ is $\sigma(\mathcal{A})$. This $\sigma$-algebra is called the Borel $\sigma$-algebra on $\mathbb{R}$. Any set $B \in \sigma(\mathcal{A})$ is called a Borel set.
Remark: The Borel $\sigma$-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 $\sigma$-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 $\sigma$-algebra to avoid mathematical contradictions.

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

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

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

- $S = \{\text{H, T}\}$
- $\mathcal{B} = \{\emptyset, \{\text{H}\}, \{\text{T}\}, S\}$
- $P : \mathcal{B} \to [0,1]$, 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 $A \in \mathcal{B}$,
  \[ P(A) = \int_{A} \frac{1}{b-a} \, dx \]  ("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 \( \sigma \)-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 \( \omega_i \) 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
\[
P\left( \bigcup_{j=1}^{k} A_j \right) = \sum_{i=1}^{n} p_i I(\omega_i \in \bigcup_{j=1}^{k} A_j)
   = \sum_{i=1}^{n} p_i \sum_{j=1}^{k} I(\omega_i \in A_j)
   = \sum_{j=1}^{k} \sum_{i=1}^{n} p_i I(\omega_i \in A_j) = \sum_{j=1}^{k} P(A_j).
\]
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

\[ S = \{[2S, 2D, 2C, 3S], [2S, 2D, 2H, 2C, 3D], ..., [A_S, A_D, A_H, A_C, K_C]\}. \]

This is a finite sample space, and there are \(|S| = \binom{52}{5} = 2598960\) outcomes in \(S\). Assume an equiprobability model with \(B = 2^S\). Define \(A = \{3 \text{ of a kind}\}\).

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., \(S = \{\omega_1, \omega_2, \ldots\}\). For a \(\sigma\)-algebra on \(S\) (e.g., \(B = 2^S\), 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 \(S = \{1, 2, 3, \ldots\}, B = 2^S\), and

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

where \(0 < p < 1\). This is called a geometric probability measure. Note that

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

The Calculus of Probabilities: We now examine many results that follow from the Kolmogorov Axioms. In what follows, let \(S\) denote a sample space, \(B\) denote a \(\sigma\)-algebra on \(S\), and \(P\) denote a probability measure. All events (e.g., \(A, B, C\), etc.) are assumed to be measurable (i.e., \(A \in B\), 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 \( C_1, C_2, \ldots \), is any partition of \( S \).

(b) **Boole’s Inequality:** For any sequence \( A_1, A_2, \ldots, \)

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

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 \( A_1, A_2, \ldots, A_n \),

\[
P\left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} P(A_i) - \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \sum_{i_1 < i_2 < i_3} P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) - \cdots + (-1)^{n+1} P\left( \bigcap_{i=1}^{n} A_i \right). \]
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 $\omega \notin A_i \forall i = 1, 2, \ldots, n$, then LHS = RHS = 0 and the result holds. Otherwise, suppose that $\omega$ is in exactly $m > 0$ of the events $A_1, A_2, \ldots, A_n$. Clearly, $\omega \in \bigcup_{i=1}^{m} A_i$, so the probability associated with $\omega$ 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 $\omega$ is in exactly $\binom{m}{k}$ intersections of this type. Therefore, the probability associated with $\omega$ 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 $\omega$ was arbitrarily chosen, we are done. \(\Box\)

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 $n = 2, 3, \ldots$. 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. \(\Box\)
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 \(B_n = \bigcup_{i=1}^{n} A_i\). Clearly, \(\{B_n\}\) is an increasing sequence of sets and \(B_n \to \bigcup_{i=1}^{\infty} A_i\), as \(n \to \infty\). Note that

\[
B_j = B_{j-1} \cup A_j \implies P(B_j) \leq P(B_{j-1}) + P(A_j)
\]

\[
\implies P(A_j) \geq P(B_j) - P(B_{j-1}).
\]

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 \(\omega_j\) is the \(j\)th permutation of \(\{1, 2, ..., n\}\).
and \( N = n! \). Define the event \( A_i = \{ \text{ith man selects his own hat} \} \) 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:

\[ P(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n} \quad \forall i = 1, 2, \ldots, n \]

\[ P(A_{i_1} \cap A_{i_2}) = \frac{(n-2)!}{n!} \quad 1 \leq i_1 < i_2 \leq n \]

\[ P(A_{i_1} \cap A_{i_2} \cap A_{i_3}) = \frac{(n-3)!}{n!} \quad 1 \leq i_1 < i_2 < i_3 \leq n \]

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

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

\[= n \left( \frac{1}{n} \right) - \frac{n}{2} \frac{(n-2)!}{n!} + \frac{n}{3} \frac{(n-3)!}{n!} - \cdots + (-1)^{n+1} \frac{1}{n!}\]

\[= \sum_{k=1}^{n} (-1)^{k+1} \frac{n!}{k!(n-k)!} \frac{(n-k)!}{n!} = 1 - \sum_{k=0}^{n} \frac{(-1)^k}{k!}. \]

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 \(\sigma\)-algebra on \(B\).

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

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

Define

\[
\begin{align*}
A &= \{(\text{HH}), (\text{HT})\} \\
B &= \{(\text{HH}), (\text{HT}), (\text{TH})\}.
\end{align*}
\]

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
  \[
  \begin{align*}
  \mathcal{B}^* &= \{ B \cap C : C \in \mathcal{B} \} \\
  &= \{\emptyset, B, \{(\text{HH})\}, \{(\text{HT})\}, \{(\text{TH})\}, \{(\text{HH}), (\text{HT})\}, \{(\text{HH}), (\text{TH})\}, \{(\text{HT}), (\text{TH})\}\}
  \end{align*}
  \]
  and \(P_B\) is an equiprobability measure; i.e., \(P_B(\{\omega\}) = 1/3 \quad \forall \omega \in B\). 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 \( A_1, A_2, \ldots, \in \mathcal{B} \) are pairwise disjoint, then
   \[
P\left(\bigcup_{i=1}^{\infty} A_i \bigg| B\right) = \sum_{i=1}^{\infty} P(A_i|B).
   \]

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 \( A_1, A_2, \ldots, \)
   \[
P\left(\bigcup_{i=1}^{\infty} A_i \bigg| B\right) \leq \sum_{i=1}^{\infty} P(A_i|B).
   \]

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 \( B_1, B_2, \ldots, \in \mathcal{B} \) denote a partition of \( S \); i.e., \( B_i \cap B_j = \emptyset \ \forall i \neq j \) and \( \bigcup_{i=1}^{\infty} B_i = S \). 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

\[
P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)} \quad \Rightarrow \quad P(A \cap B_i) = P(A|B_i)P(B_i).
\]

**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

\[
D = \{\text{disease is present}\}
\]
\[
\bigstar = \{\text{test is positive}\}.
\]

We are given

\[
P(\bigstar|D) = 0.95 \quad \text{(sensitivity)}
\]
\[
P(\bigstar^c|D^c) = 0.99 \quad \text{(specificity)}
\]
\[
P(D) = 0.08 \quad \text{(prevalence)}
\]
The probability a randomly selected subject will test positively is

\[ P(\text{\textbullet}) = P(\text{\textbullet}|D)P(D) + P(\text{\textbullet}|D^c)P(D^c) \]

\[ = (0.95)(0.08) = (0.01)(0.92) \approx 0.0852. \]

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

\[ P(D|\text{\textbullet}) = \frac{P(D \cap \text{\textbullet})}{P(\text{\textbullet})} \]

\[ = \frac{P(\text{\textbullet}|D)P(D)}{P(\text{\textbullet}|D)P(D) + P(\text{\textbullet}|D^c)P(D^c)} \]

\[ = \frac{(0.95)(0.08)}{(0.95)(0.08) + (0.01)(0.92)} \approx 0.892. \]

**Note:** \( P(D|\text{\textbullet}) \) in this example is called the **positive predictive value** (PPV). As an exercise, calculate \( P(D^c|\text{\textbullet}^c) \), 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):

<table>
<thead>
<tr>
<th>Prior probability</th>
<th>Test result</th>
<th>Posterior probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(D) = 0.08 )</td>
<td>( \text{\textbullet} )</td>
<td>( P(D</td>
</tr>
<tr>
<td>( P(D) = 0.08 )</td>
<td>( \text{\textbullet}^c )</td>
<td>( P(D</td>
</tr>
</tbody>
</table>

**Bayes’ Rule:** Suppose \((S, B, P)\) is a model for a random experiment. Let \(B_1, B_2, ..., \in B\) 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,

\[ P(B_i) \rightarrow A \text{ occurs} \rightarrow P(B_i|A). \]

**Multiplication Rule:** For two events \(A, B \in \mathcal{B}\),

\[ P(A \cap B) = P(A|B)P(B) = P(B|A)P(A). \]

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

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

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):

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B).$$

**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**:

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3)$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3)$$

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 $A_i = \{i^{th} \text{ student is positive}\}$, 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 \( i = 1, 2, \ldots, 30 \). Therefore,
\[
P \left( \bigcup_{i=1}^{30} A_i \right) = 1 - P \left( \bigcap_{i=1}^{30} A_i^c \right) \\
= 1 - \prod_{i=1}^{30} P(A_i^c) \\
= 1 - (1 - p)^{30}.
\]

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 \( \omega \)’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 \( k = 0, 1, 2, \ldots, 30 \). 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
\[
S = \{(0,0,0,\ldots,0), (1,0,0,\ldots,0), (0,1,0,\ldots,0), \ldots, (1,1,1,\ldots,1)\}
\]
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
\[
\mathcal{X} = \{x : x = 0, 1, 2, \ldots, 30\}.
\]
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
\[
X^{-1}(B) \equiv \{\omega \in S : X(\omega) \in B\} \in \mathcal{B} \quad (1.2)
\]
for all \( B \in \mathcal{B}(\mathbb{R}) \), where \( \mathcal{B}(\mathbb{R}) \) is the Borel \( \sigma \)-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,

$$
X((0,0,0,...,0)) = 0 \\
X((1,0,0,...,0)) = 1 \\
X((1,1,1,...,1)) = 30.
$$

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

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

2. The measurability condition

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

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

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

Example 1.16. Consider a random experiment with

$S = \{1,2,3\} \quad B_1 = \{\emptyset,\{1\},\{2,3\},S\}$

$P = \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/3, \text{ for all } \omega \in S.$

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

$$
X^{-1}(B) = X^{-1}(\{0\}) = \{\omega \in S : X(\omega) = 0\} = \{1,2\} \notin B_1.
$$

Therefore, the function $X$ is not a random variable on $(S,B_1)$. It does not satisfy the measurability condition. Question: Is $X$ a random variable on $(S,B)$, where $B = 2^S$?

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

$$
S = \{\omega_1,\omega_2,...,\omega_n\} \quad B = 2^S \quad P = \text{a valid probability measure}.
$$

Here, we allow for both cases:

- $n < \infty \implies S \text{ finite}$
- “$n = \infty$” $\implies S \text{ countable (i.e., countably infinite).}
Suppose $X$ is a random variable on $(S, \mathcal{B}, P)$ with range $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$. We call $\mathcal{X}$ the **support** of the random variable $X$; we allow for both cases:

- $m < \infty \implies \text{“finite support”}$
- $m = \infty \implies \text{“countably infinite support.”}$

Define a new probability measure $P_X$ according to

$$P_X(X = x_i) = P(\{\omega \in S : X(\omega) = x_i\}).$$

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:

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

**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

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

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

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

The random variable $X$ satisfies

<table>
<thead>
<tr>
<th>$\omega$</th>
<th>$(HH)$</th>
<th>$(HT)$</th>
<th>$(TH)$</th>
<th>$(TT)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(\omega)$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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

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

We have the following **probability distribution** for the random variable $X$:
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 $x \in \mathcal{X} = \{x : x = 0, 1, 2, \ldots, 30\}$. 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 $B \in \mathcal{B}(\mathbb{R})$,

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

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

$$P_X(X \in B) = P(\{\omega \in S : X(\omega) \in B\}).$$

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
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.

**Theorem 1.5.3.** The function $F_X : \mathbb{R} \rightarrow [0, 1]$ 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.
Theorem 1.5.3. First, we have \( \lim_{n \to \infty} B_n = \lim_{n \to \infty} \{ X \leq x_n \} = \bigcup_{n=1}^{\infty} \{ X \leq x_n \} = \{ X < \infty \} \).

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. \( \square \)

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. 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

\[
\lim_{x \to \infty} F_X(x) = \lim_{x \to \infty} \left( 1 - e^{-x/\beta} \right) = 1 - \lim_{x \to \infty} e^{-x/\beta} = 1,
\]

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 \( x > 0 \),

\[
\frac{d}{dx} F_X(x) = \frac{d}{dx} \left( 1 - e^{-x/\beta} \right) = \frac{1}{\beta} e^{-x/\beta} > 0 \ \forall x > 0.
\]
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 $X \overset{d}{=} Y$.

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

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

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,

$$X \overset{d}{=} Y \nRightarrow X = Y.$$
For example, suppose that

\[ S = \{ (HH), (HT), (TH), (TT) \} \]
\[ B = 2^S \]
\[ P = \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/4, \text{ for all } \omega \in S. \]

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, \( X \overset{d}{=} Y \). However, \( 2 = X((HH)) \neq Y((HH)) = 0 \), 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

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

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

\[ X \text{ discrete } \iff F_X(x) \text{ is a step function} \]
\[ X \text{ continuous } \iff F_X(x) \text{ is continuous}. \]

**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

\[ f_X(x) = P_X(X = x), \text{ for all } x. \]

**Example 1.21.** Suppose that \( X \) is a random variable with pmf

\[ f_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \\ 0, & \text{otherwise,} \end{cases} \]

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

\[ P_X(X = 1) = f_X(1) = \frac{\lambda^1 e^{-\lambda}}{1!} = \lambda e^{-\lambda} \]
\[ P_X(X \leq 1) = F_X(1) = f_X(0) + f_X(1) = \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} = e^{-\lambda}(1 + \lambda). \]
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 $x \not\in X$, then $f_X(x) = P_X(X = x) = 0$.

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
\[
0 \leq P_X(X = x) \leq \lim_{\epsilon \to 0} [F_X(x) - F_X(x - \epsilon)] \\
= F_X(x) - \lim_{\epsilon \to 0} F_X(x - \epsilon) \\
= F_X(x) - F_X(x) = 0,
\]
because $F_X(x)$ is continuous. Therefore, we have shown that $0 \leq P_X(X = x) \leq 0$. Because $\epsilon$ was arbitrary, we are done. \(\square\)

Remark: This result highlights the salient difference between discrete and continuous random variables. Discrete random variables have positive probability assigned to support points $x \in X$. 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 \( x < 0 \),

\[
F_X(x) = \int_{-\infty}^{x} f_X(u) du = \int_{-\infty}^{x} \frac{1}{2} e^u du = \frac{1}{2} e^x \bigg|_{-\infty}^{x} = \frac{1}{2} (e^x - 0) = \frac{1}{2} e^x.
\]

Case 2: For \( x \geq 0 \),

\[
F_X(x) = \int_{-\infty}^{x} f_X(u) du = \int_{-\infty}^{0} f_X(u) du + \int_{0}^{x} f_X(u) du = \int_{-\infty}^{0} \frac{1}{2} e^u du + \int_{0}^{x} \frac{1}{2} e^{-u} du = \frac{1}{2} - \left( \frac{1}{2} e^{-u} \right)_{0}^{x} = 1 - \frac{1}{2} e^{-x}.
\]

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. This is an example of a LaPlace distribution.

(b) The desired probabilities are

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

PAGE 28
and

\[ P_X(-2 < X < 2) = P_X(-2 < X \leq 2) = P_X(X \leq 2) - P_X(X \leq -2) = F_X(2) - F_X(-2) = \left(1 - \frac{1}{2}e^{-2}\right) - \frac{1}{2}e^{-2} = 1 - e^{-2} \approx 0.8647. \]

**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,

\[ P_X(-2 < X \leq 2) = P_X(X \leq 2) - P_X(X \leq -2) = F_X(2) - F_X(-2). \]

In general, 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,

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

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

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

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.

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

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

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

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 \textbf{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

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

where the \textbf{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:

\[ \lim_{x \to -\infty} \int_{\mathbb{R}} f_X(u)I(u \leq x) du = \int_{\mathbb{R}} f_X(u) \lim_{x \to -\infty} I(u \leq x) du = 0. \]

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

\[ \lim_{x \to x_0^-} F_X(x) = \lim_{x \to x_0^-} \int_{-\infty}^{x} f_X(u) du = \lim_{x \to x_0^-} \int_{x \to -\infty} f_X(u)I(u \leq x) du \]

\[ = \int_{-\infty}^{x} f_X(u) \lim_{x \to x_0^-} I(u \leq x) du. \]

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

\[ \int_{-\infty}^{x} f_X(u) \lim_{x \to x_0^-} I(u \leq x) du = \int_{-\infty}^{x} f_X(u)I(u \leq x_0) du \]

\[ = \int_{-\infty}^{x_0} f_X(u) du = F_X(x_0). \]

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 $\mathcal{X}$. 

PAGE 32
Example 2.1. Suppose that $X$ is a discrete random variable with pmf

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

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” ($0 = \text{failure}; 1 = \text{success}$), 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 $\mathcal{X} = \{x : x = 0, 1, 2, \ldots, n\}$.

Question: What is the distribution of $Y = g(X) = n - X$?

Note that the support of $Y$ is given by

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

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

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

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

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

That is,

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

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

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

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

$$F_Y(y) = P_Y(Y \leq y) = P_X(g(X) \leq y) = \int_B f_X(x) dx,$$

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 $0 < x < 1 \implies 0 < y < \frac{1}{4}$. Therefore, the support of $Y$ is $\mathcal{Y} = \{y : 0 < y < \frac{1}{4}\}$.

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

**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

$$F_Y(y) = P_Y(Y \leq y) = P_X(X \leq x_1) + P_X(X \geq x_2)$$

$$= \int_0^{x_1} 1 \, dx + \int_{x_2}^{1} 1 \, dx,$$

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:

$$y = g(x) = x(1 - x) \implies -x^2 + x - y = 0.$$
The roots of this equation are

\[ x = \frac{-1 \pm \sqrt{(1)^2 - 4(-1)(-y)}}{2(-1)} \]

\[ = \frac{1}{2} \pm \frac{\sqrt{1 - 4y}}{2} \]

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

\[ F_Y(y) = P(Y \leq y) = \int_0^{\frac{1}{2} - \frac{\sqrt{1 - 4y}}{2}} 1 dx + \int_{\frac{1}{2} + \frac{\sqrt{1 - 4y}}{2}}^1 1 dx \]

\[ = \frac{1}{2} - \frac{\sqrt{1 - 4y}}{2} + 1 - \frac{1}{2} - \frac{\sqrt{1 - 4y}}{2} \]

\[ = 1 - \sqrt{1 - 4y}. \]

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

\[ f_Y(y) = \frac{d}{dy} F_Y(y) \]

\[ = \begin{cases} 
2(1 - 4y)^{-1/2}, & 0 < y < \frac{1}{4} \\
0, & \text{otherwise}.
\end{cases} \]

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
\[
F_Y(y) = P_Y(Y \leq y) = P_X(g(X) \leq y) = P_X(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).
\]
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
\[
f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \left[ \frac{d}{dy}g^{-1}(y) \right]_{>0}.
\]
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
\[
F_Y(y) = P_Y(Y \leq y) = P_X(g(X) \leq y) = P_X(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y)).
\]
Again, the penultimate equality results from noting that \( \{x : g(x) \leq y\} = \{x : x \geq g^{-1}(y)\} \).

We have shown that
\[
F_Y(y) = 1 - F_X(g^{-1}(y)), \quad \text{when } g \text{ is strictly decreasing.}
\]

Taking derivatives, the pdf of \( Y \) (where nonzero) is
\[
f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy} \left[1 - F_X(g^{-1}(y))\right] \\
= -f_X(g^{-1}(y)) \frac{d}{dy}g^{-1}(y) < 0.
\]

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

Find the pdf of
\[
Y = g(X) = -\beta \ln X,
\]
where \( \beta > 0 \).

**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 \( y > 0 \),
\[
f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy}g^{-1}(y) \right| \\
= 1 \times \frac{1}{\beta}e^{-y/\beta} = \frac{1}{\beta}e^{-y/\beta}.
\]
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 $X \sim U(0, 1) \implies Y = g(X) = -\beta \ln X \sim \text{exponential}(\beta)$.

**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

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

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

$$ \alpha \rightarrow \text{"shape parameter"} $$
$$ \beta \rightarrow \text{"scale parameter."} $$

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:

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

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 $y > 0$,

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

$$ = \frac{1}{\Gamma(\alpha)\beta^\alpha} \left( \frac{1}{y} \right)^{\alpha-1} e^{-1/\beta y} \times \frac{1}{y^2} $$

$$ = \frac{1}{\Gamma(\alpha)\beta^\alpha} \frac{1}{y^{\alpha+1}} e^{-1/\beta y}. $$
Summarizing, the pdf of $Y$ is

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

This is called the inverted gamma distribution (not surprisingly). We have shown that $X \sim \text{gamma}(\alpha, \beta) \implies Y = g(X) = \frac{1}{X} \sim \text{IG}(\alpha, \beta)$.

**Q:** What if $g$ is not one-to-one over $\mathcal{X}$?

**A:** We can always use the general result that

$$F_Y(y) = P_Y(Y \leq y) = P_X(g(X) \leq y) = \int_B f_X(x)dx,$$

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 $y > 0$,

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

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

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} [F_X(\sqrt{y}) - F_X(-\sqrt{y})]$$

$$= f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} - f_X(-\sqrt{y}) \left(-\frac{1}{2\sqrt{y}}\right)$$

$$= \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})].$$

**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-\(\chi^2\) 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 a continuous random variable, we have \( P_Y(Y = 0) = 0 \). We can therefore proceed assuming that \( y > 0 \). By the last result, we have, for \( y > 0 \),

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

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 \( \chi^2 \) distribution with \( \nu = 1 \) degree of freedom. We have shown that

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

We will use this fact repeatedly in this course.

**Interesting:** Recall that we defined the gamma function

\[
\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du.
\]

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

\[
f_Y(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}
= \frac{1}{\Gamma(1/2) 2^{1/2} y^{1/2}} e^{-y/2},
\]

which we recognize as a gamma(\( \alpha, \beta \)) pdf with parameters \( \alpha = 1/2 \) and \( \beta = 2 \). Therefore, the \( \chi^2_1 \) distribution is a special member of the gamma(\( \alpha, \beta \)) 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

\[
f_Y(y) = I(0 < y < 1) \quad \text{and} \quad F_Y(y) = \begin{cases} 
0, & y < 0 \\
y, & 0 \leq y \leq 1 \\
1, & y > 1.
\end{cases}
\]
Proof. Suppose that $F_X(x)$ is strictly increasing. Regardless of the support of $X$, the random variable $Y = F_X(X)$ has support $\mathcal{Y} = \{y : 0 \leq y \leq 1\}$. The cdf of $Y$, for $0 \leq y \leq 1$, is given by

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

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

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

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 $U(0, 1)$ 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

- Discrete case:
  $$E(X) = \sum_{x \in \mathcal{X}} x f_X(x)$$

- Continuous case:
  $$E(X) = \int_{\mathbb{R}} x f_X(x) dx$$

**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

$$f_X(x) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, ... \\ 0, & \text{otherwise}, \end{cases}$$
where \( \lambda > 0 \). The expected value of \( X \) is

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

because \( \sum_{y=0}^{\infty} \frac{\lambda^y}{y!} \) is the Mc Laurin series expansion of \( e^\lambda \). 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

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

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

\[
E(X) = \int_{\mathbb{R}} x f_X(x) dx = \int_{\alpha}^{\infty} x \frac{\beta \alpha}{x^{\beta+1}} dx = \beta \alpha \int_{\alpha}^{\infty} \frac{1}{x^\beta} dx
\]

\[
= \beta \alpha \left( \frac{1}{\beta-1} \frac{1}{x^{\beta-1}} \right) \bigg|_{x=\alpha}^{\infty} = \frac{\beta \alpha}{\beta-1},
\]

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

\[
E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) \quad \text{ (discrete case)}
\]

\[
E[g(X)] = \int_{\mathbb{R}} g(x) f_X(x) dx \quad \text{ (continuous case)}
\]

**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:

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

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

$$u = -\ln x \quad du = -\frac{1}{x} dx$$
$$dv = dx \quad v = x$$

Integration by parts shows that the last integral

$$\int_0^1 -\ln x \, dx = -x \ln x \bigg|_0^1 - \int_0^1 (-1) dx = (0 - 0) + 1 = 1.$$

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

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

Let

$$u = y \quad du = dy$$
$$dv = e^{-y} \quad v = -e^{-y}$$

Integration by parts shows that the last integral

$$\int_0^{\infty} ye^{-y} \, dy = -ye^{-y} \bigg|_0^{\infty} - \int_0^{\infty} -e^{-y} \, dy = (0 - 0) + 1 = 1.$$

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

$$E[ag_1(X) + bg_2(X) + c] = \int_{\mathbb{R}} [ag_1(x) + bg_2(x) + c]f_X(x)dx$$

$$= a \int_{\mathbb{R}} g_1(x)f_X(x)dx + b \int_{\mathbb{R}} g_2(x)f_X(x)dx + c \int_{\mathbb{R}} f_X(x)dx$$

$$= aE[g_1(X)] + bE[g_2(X)] + c.$$

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

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

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

4. $g(X) = tX$, where $t$ is a constant. The expectation
   \[ E[g(X)] = E(tX) \]
   is called the \textit{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 \textit{characteristic function} of $X$. In this case, the function $g : \mathbb{R} \rightarrow \mathbb{C}$.

### 2.3 Moments and Moment Generating Functions

**Definition:** Suppose that $X$ is a random variable. The \textit{kth (uncentered) moment} of $X$ is
\[ \mu'_k = E(X^k). \]
The \textit{kth 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 \textit{mean} of $X$.
- The 2nd central moment of $X$ is $\mu_2 = E[(X - \mu)^2]$. We call this the \textit{variance} of $X$ and usually denote this by $\sigma^2$ 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

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

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 $\mu$; 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 $\sigma$ 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

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

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

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

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 $E(X) = \text{var}(X) = \lambda$. 

PAGE 46
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

$$E(X^2) = \int_{\mathbb{R}} x^2 f_X(x) dx = \int_{\alpha}^{\infty} x^2 \frac{\beta \alpha^\beta}{x^{\beta + 1}} dx$$

$$= \frac{\beta \alpha^\beta}{\beta - 2} \left( \lim_{x \to \infty} \frac{1}{x^{\beta - 2}} - \frac{1}{x^{\beta - 2}} \right)_{x=\alpha}$$

$$= \frac{\beta \alpha^\beta}{\beta - 2} \left( \frac{1}{\alpha^{\beta - 2}} - \frac{1}{\lim_{x \to \infty} x^{\beta - 2}} \right) = \frac{\beta \alpha^2}{\beta - 2},$$

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

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

$$= \frac{\beta \alpha^2}{(\beta - 1)^2(\beta - 2)}.$$

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-Stieltjes integral, which is understood to mean

\[
M_X(t) = \sum_{x \in X} e^{tx} f_X(x) \quad \text{(discrete case)}
\]

\[
M_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x) dx \quad \text{(continuous case)}
\]

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

\[
f_X(x) = \begin{cases} 
\frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \ldots \\
0, & \text{otherwise,}
\end{cases}
\]

where \( \lambda > 0 \). The mgf of \( X \) is

\[
M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}.
\]

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 \( X \sim \text{b}(n,p) \); i.e., the pmf of \( X \) is

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

where \( 0 < p < 1 \). The mgf of \( X \) is

\[
M_X(t) = E(e^{tX}) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = (q + pe^t)^n,
\]

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 \( X \sim \text{exponential}(\beta) \); 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
\[
M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} \frac{1}{\beta} e^{-x/\beta} dx = \frac{1}{\beta} \int_0^\infty e^{-x}\left(\frac{1}{\beta} - t\right)dx
\]
\[
= \frac{1}{\beta} \left[-\frac{1}{\beta} - t e^{-x} (\frac{1}{\beta} - t) \right]_{x=0}^{\infty}
\]
\[
= \frac{1}{1 - \beta t} \left[1 - \lim_{x \to \infty} e^{-x} (\frac{1}{\beta} - t) \right]
\]
Note that
\[
\lim_{x \to \infty} e^{-x} (\frac{1}{\beta} - t) = 0 \quad \text{if} \quad \frac{1}{\beta} - t > 0
\]
\[
\lim_{x \to \infty} e^{-x} (\frac{1}{\beta} - t) = +\infty \quad \text{if} \quad \frac{1}{\beta} - t < 0.
\]
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 \( X \sim \text{gamma}(\alpha, \beta) \), then
\[
M_X(t) = \left(\frac{1}{1 - \beta t}\right)^{\alpha}, \quad t < \frac{1}{\beta}.
\]
When \( \alpha = 1 \), the gamma(\( \alpha, \beta \)) distribution reduces to the exponential(\( \beta \)) 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

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}\int_{\mathbb{R}} e^{tx}dF_X(x)$$

$$= \int_{\mathbb{R}} \frac{d}{dt} e^{tx}dF_X(x)$$

$$= \int_{\mathbb{R}} xe^{tx}dF_X(x) = E(Xe^{tX}).$$

Therefore,

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

Showing this for $k = 2, 3, \ldots$, is done similarly. $\square$

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

$$M_X(t) = M_X(0) + \frac{M_X^{(1)}(0)}{1!}(t - 0) + \frac{M_X^{(2)}(0)}{2!}(t - 0)^2 + \frac{M_X^{(3)}(0)}{3!}(t - 0)^3 + \cdots$$

$$= 1 + E(X)t + \frac{E(X^2)}{2}t^2 + \frac{E(X^3)}{6}t^3 + \frac{E(X^4)}{24}t^4 + \cdots$$

$$= \sum_{k=0}^{\infty} \frac{E(X^k)}{k!}t^k.$$

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. Writing $M_X(t)$ as an infinite sum (in its McLaurin series expansion) is something we will do later when we prove the Central Limit Theorem.

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

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

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

**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).

- $\xi = 0 \implies f_X(x)$ is symmetric about $\mu$
- $\xi > 0 \implies f_X(x)$ is skewed right
- $\xi < 0 \implies f_X(x)$ 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 $X$.

- $\kappa = 3 \implies$ mesokurtic; normal distribution (as a reference)
- $\kappa > 3 \implies$ leptokurtic; $f_X(x)$ has a more acute peak around $\mu$ and fatter tails
- $\kappa < 3 \implies$ platykurtic; $f_X(x)$ has a broader peak around $\mu$ 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 $k$th 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

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

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 ($\implies$) of Theorem 2.3.11 is easy. Proving the sufficiency ($\iff$) is not. Note that if $X$ is continuous,

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

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

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

The mgf of $Y$ is

$$M_Y(t) = E(e^{ty}) = E(e^{tcX}) = M_X(ct) = \left( \frac{1}{1 - \beta ct} \right)^\alpha,$$

which exists for $ct < 1/\beta \iff t < 1/\beta c$. We recognize $M_Y(t)$ as the mgf of a gamma distribution with shape parameter $\alpha$ and scale parameter $\beta c$. 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

$$M_Y(t) = E(e^{ty}) = E\left[e^{t(ax+b)}\right] = E(e^{bt}e^{atX}) = e^{bt}E(e^{atX}) = e^{bt}M_X(at). \quad \square$$
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 $\mu$ and variance $\sigma^2$, 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

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

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

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

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:

```
MGF exists \implies distribution is determined
\downarrow
moments are determined \implies ???
```

Q: Does an infinite set of moments uniquely determine a distribution?
A: Yes, if $X$ 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

$$ f_X(x) = \frac{1}{\sqrt{2\pi x}} e^{-(\ln x)^2/2} I(x > 0) $$

$$ f_Y(y) = f_X(y)[1 + \sin(2\pi \ln y)]. $$

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

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

\[
E(X^r) = e^{r^2/2}, \text{ for } r = 1, 2, 3, \ldots.
\]

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.,

\[
\lim_{n \to \infty} \left( 1 + \frac{b}{n} \right)^n = e^b,
\]

a result well known from calculus.
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

\[
MX_n(t) = E(e^{tX_n}) = (q_n + p_n e^t)^n = \left[ 1 + \frac{\lambda(e^t - 1)}{n} \right]^n,
\]

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

\[
\lim_{n \to \infty} MX_n(t) = e^{\lambda(e^t - 1)},
\]

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

Example 2.19. Suppose that \( Y_n \sim \text{gamma}(n, \beta) \), where \( \beta \) is free of \( n \), so that

\[
MY_n(t) = \left( \frac{1}{1-\beta t} \right)^n, \quad t < \frac{1}{\beta}.
\]

Find the limiting distribution of \( X_n = \frac{Y_n}{n} \).

Solution. The mgf of \( X_n \) is

\[
MX_n(t) = E(e^{tX_n}) = E\left[ e^{t\left(\frac{Y_n}{n}\right)}\right] = MY_n(t/n) = \left[ \frac{1}{1-\beta(t/n)} \right]^n = \left( 1 - \frac{\beta t}{n} \right)^{-n},
\]

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

\[
\lim_{n \to \infty} MX_n(t) = e^{\beta t}.
\]

The limiting mgf \( MX(t) = e^{\beta t} \) 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 \( X_n \overset{d}{\longrightarrow} X \), as \( n \to \infty \), where \( X \) has a degenerate distribution at \( \beta \). In Chapter 5, we will refer to this type of convergence as “convergence in probability” and will write \( X_n \overset{p}{\longrightarrow} \beta \), as \( n \to \infty \).
3 Common Families of Distributions


3.1 Introduction

Definition: A parametric model (or parametric family) is a set of distributions indexed by a finite-dimensional parameter \( \theta = (\theta_1, \theta_2, ..., \theta_d)' \), where \( d \geq 1 \). Unless otherwise stated, the parameter \( \theta \) is regarded as fixed (i.e., it is not random).

Example 3.1. Suppose \( X \sim \text{exponential}(\beta) \). 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 \( \theta = \beta \), 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 \( X \sim N(\mu, \sigma^2) \); 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 \( \theta = (\mu, \sigma^2)' \) is two-dimensional \((d = 2)\). One member of the \( N(\mu, \sigma^2) \) family arises when \( \mu = 0 \) and \( \sigma^2 = 1 \). This member is called the standard normal distribution and is denoted by \( N(0, 1) \).

Remark: A common format for a first-year sequence in probability and mathematical statistics (like STAT 712-713) is to accept a given family of distributions as being appropriate and then proceed to develop what is exclusively model-dependent 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 \( X \), is at most countable.
1. **Discrete Uniform.** A random variable is said to have a discrete uniform distribution if its pmf is given by
\[
f_X(x|N) = \begin{cases} \frac{1}{N}, & x = 1, 2, \ldots, N \\ 0, & \text{otherwise}, \end{cases}
\]
where \( N \in \mathbb{N} \). Note that this distribution puts the same weight \( 1/N \) on each outcome \( x \in \mathcal{X} = \{x : x = 1, 2, \ldots, N\} \). **Notation:** \( X \sim DU(1, N) \).

**Mean/Variance:** The relevant moments of \( X \sim DU(1, N) \) are
\[
E(X) = \frac{N + 1}{2}, \quad \text{and} \quad var(X) = \frac{(N + 1)(N - 1)}{12}.
\]

**MGF:** The mgf of \( X \sim DU(1, N) \) is
\[
M_X(t) = E(e^{tX}) = \sum_{x=1}^{N} \frac{e^{tx}}{N} = \frac{1}{N} e^t + \frac{1}{N} e^{2t} + \cdots + \frac{1}{N} e^{Nt}.
\]

**Generalization:** The discrete uniform distribution can be generalized easily. The pmf of \( X \sim DU(N_0, N_1) \) is
\[
f_X(x|N_0, N_1) = \begin{cases} \frac{1}{N_1-N_0+1}, & x = N_0, N_0+1, \ldots, N_1 \\ 0, & \text{otherwise}, \end{cases}
\]
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
\[
f_X(x|N, M, K) = \begin{cases} \frac{\binom{M}{x} \binom{N-M}{K-x}}{\binom{N}{K}}, & x = 0, 1, 2, \ldots, K \\ 0, & \text{otherwise}, \end{cases}
\]
where \( N, M, K \in \mathbb{N} \), \( M < N \), \( K < N \). The support \( \mathcal{X} = \{x : x = 0, 1, 2, \ldots, K\} \) 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.”

\[
N = \text{population size} \\
K = \text{sample size} \\
M = \text{number of Type I objects in the population}.
\]

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

\[
E(X) = \frac{KM}{N} \\
\text{var}(X) = \frac{KM}{N} \left( 1 - \frac{M}{N} \right) \left( \frac{N - K}{N - 1} \right).
\]

The term \( \left( \frac{N - K}{N - 1} \right) \) is called the “finite population correction factor.” The mgf of \( X \) 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 \( \text{hyper}(N, K, M) \) pmf also converges to the \( b(n, p) \) pmf under the same conditions; see Exercise 3.11 (pp 129 CB).

Implication: 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).

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: \( X \sim b(n, p) \).

Bernoulli trials: A Bernoulli trial is an experiment with two possible outcomes, where

- the outcomes can be thought of as “success” or “failure”
- \( p = \text{pr(“success”)} \) is the same for each trial.

Mean/Variance: The relevant moments of \( X \sim b(n, p) \) are

\[
E(X) = np \\
\text{var}(X) = np(1 - p).
\]

MGF: The mgf of \( X \sim b(n, p) \) is

\[
M_X(t) = (q + pe^t)^n, \quad \text{where} \quad q = 1 - p.
\]

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

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

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 $\mathcal{X} = \{x : x = 1, 2, 3, \ldots, \} = \mathbb{N}$. **Notation:** $X \sim \text{geom}(p)$.

**Mean/Variance:** The relevant moments of $X \sim \text{geom}(p)$ are

$$E(X) = \frac{1}{p}, \quad \text{var}(X) = \frac{q}{p^2}, \quad \text{where } q = 1 - p.$$  

**MGF:** The mgf of $X \sim \text{geom}(p)$ is

$$M_X(t) = E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx}(1-p)^{x-1}p = \frac{p}{q} \sum_{x=1}^{\infty} (qe^t)^x = \frac{p}{q} \left[ \sum_{x=0}^{\infty} (qe^t)^x - 1 \right] = \frac{p}{q} \left( \frac{1}{1-qe^t} - 1 \right) = \frac{pe^t}{1-qe^t},$$

for $qe^t < 1 \iff t < -\ln q$.

**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

$$f_X(x|r,p) = \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r}, & x = r, r+1, r+2, \ldots, \\ 0, & \text{otherwise}, \end{cases}$$

where $0 < p < 1$. **Conceptualization:** Suppose independent Bernoulli trials are performed. The random variable $X$ counts the number of trials needed to observe the $r$th success, where $r \geq 1$. The support of $X$ is $\mathcal{X} = \{x : x = r, r+1, r+2, \ldots, \}$. **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

$$E(X) = \frac{r}{p}, \quad \text{var}(X) = \frac{rq}{p^2}, \quad \text{where } q = 1 - p.$$
MGF: The mgf of $X \sim \text{nib}(r, p)$ is

$$M_X(t) = E(e^{tX}) = \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

$$= (pe^t)^r \sum_{x=r}^{\infty} \binom{x-1}{r-1} (qe^t)^{x-r} = \left( \frac{pe^t}{1-qe^t} \right)^r,$$

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

$$f'(w) = r(1 - w)^{-(r+1)}$$

$$f''(w) = r(r+1)(1 - w)^{-(r+2)}$$

$$\vdots$$

In general, $f^{(z)}(w) = r(r+1) \cdots (r + z - 1)(1 - w)^{-(r+z)}$, where $f^{(z)}(w)$ denotes the $z$th derivative of $f$ with respect to $w$. Note that

$$f^{(z)}(w) \bigg|_{w=0} = r(r+1) \cdots (r + z - 1).$$

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. \(\square\)

**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 $\mathcal{Y} = \{y : y = 0, 1, 2, \ldots\}$. Therefore, the pmf of $Y = g(X) = X - r$, for $y = 0, 1, 2, \ldots$, is given by

$$f_Y(y) = P_Y(Y = y) = P_X(X - r = y) = P_X(X = y + r) = \binom{y + r - 1}{y} p^r (1-p)^y.$$
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

\[
f_X(x|\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise}, \end{cases}
\]

where \( \lambda > 0 \). The support of \( X \) is \( X = \{x : x = 0, 1, 2, \ldots\} \). **Notation:** \( X \sim \text{Poisson}(\lambda) \).

**Mean/Variance:** The relevant moments of \( X \sim \text{Poisson}(\lambda) \) are

\[
E(X) = \lambda \\
\text{var}(X) = \lambda.
\]

**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 \( y = 0, 1, 2, \ldots \). That is, \( Y_r \) follows a negative binomial distribution, but where \( Y_r \) records the number of failures before the \( r \)th 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:** $X \sim U(a,b)$.

   **Mean/Variance:** The relevant moments of $X \sim U(a,b)$ are
   \[
   E(X) = \frac{a+b}{2}, \quad \text{var}(X) = \frac{(b-a)^2}{12}.
   \]

   **MGF:** The mgf of $X \sim U(a,b)$ is
   \[
   M_X(t) = \begin{cases} 
   e^{bt} - e^{at} & t \neq 0 \\
   \frac{1}{(b-a)t} & t = 0,
   \end{cases}
   \]

   **CDF:** The cdf of $X \sim U(a,b)$ is
   \[
   F_X(x) = \begin{cases} 
   0, & x \leq a \\
   \frac{x-a}{b-a}, & a < x < b \\
   1, & x \geq b.
   \end{cases}
   \]

   **Remark:** A special member of the $U(a,b)$ family arises when $a = 0$ and $b = 1$. It is called the “standard” uniform distribution; $X \sim U(0,1)$.

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
   \[
   \alpha \rightarrow \text{“shape parameter”} \\
   \beta \rightarrow \text{“scale parameter.”}
   \]

   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 $\alpha \in \mathbb{N} = \{1, 2, 3, \ldots\}$, then second (recursive) property implies $\Gamma(\alpha) = (\alpha - 1)!$

**Mean/Variance:** The relevant moments of $X \sim \text{gamma}(\alpha, \beta)$ are

$$E(X) = \alpha \beta $$

$$\text{var}(X) = \alpha \beta^2$$

**MGF:** The mgf of $X \sim \text{gamma}(\alpha, \beta)$ is

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

**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

$$F_W(w) = P_W(W \leq w) = 1 - P_W(W > w) = 1 - \text{pr}\{\text{fewer than } \alpha \text{ events in } [0, w]\}$$

$$= 1 - \sum_{j=0}^{\alpha-1} \frac{(\lambda w)^j e^{-\lambda w}}{j!}.$$

**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 $\lambda w$.

The pdf of $W$, for $w > 0$, is given by

$$f_W(w) = \frac{d}{dw} F_W(w)$$

$$= \lambda e^{-\lambda w} - e^{-\lambda w} \sum_{j=1}^{\alpha-1} \left[ \frac{\lambda (\lambda w)^j}{j!} - \frac{(\lambda w)^j}{j!} \right]_{\text{telescoping sum}}$$

$$= \lambda e^{-\lambda w} - e^{-\lambda w} \left[ \lambda - \frac{\lambda (\lambda w)^{\alpha-1}}{(\alpha - 1)!} \right]$$

$$= \frac{\lambda^\alpha}{\Gamma(\alpha)} w^{\alpha-1} e^{-\lambda w},$$

which is the pdf $W \sim \text{gamma}(\alpha, \beta)$, where $\beta = 1/\lambda$. □
Integration Trick: Because the gamma$(\alpha, \beta)$ pdf integrates to one, we have
\[
\int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx = 1 \implies \int_0^\infty x^{\alpha-1} e^{-x/\beta} dx = \Gamma(\alpha)\beta^\alpha.
\]
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: $X \sim \text{exponential}(\beta)$. The exponential$(\beta)$ distribution is a special case of the gamma$(\alpha, \beta)$ distribution when $\alpha = 1$.

Mean/Variance: The relevant moments of $X \sim \text{exponential}(\beta)$ are
\[
E(X) = \beta \\
\text{var}(X) = \beta^2.
\]

MGF: The mgf of $X \sim \text{exponential}(\beta)$ is
\[
M_X(t) = \frac{1}{1 - \beta t}, \quad t < \frac{1}{\beta}.
\]

CDF: The cdf of $X \sim \text{exponential}(\beta)$ 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
\[
f_X(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{p/2-1} e^{-x/2} I(x > 0),
\]
where $p > 0$. Notation: $X \sim \chi^2_p$. The $\chi^2_p$ distribution is a special case of the gamma$(\alpha, \beta)$ distribution when $\alpha = p/2$ and $\beta = 2$. Usually, $p$ will be an integer. The $\chi^2$ distribution arises often in applied statistics.
Mean/Variance: The relevant moments of $X \sim \chi^2_p$ are
\[
E(X) = p \\
\text{var}(X) = 2p.
\]

MGF: The mgf of $X \sim \chi^2_p$ is
\[
M_X(t) = \left(\frac{1}{1 - 2t}\right)^{p/2}, \quad t < \frac{1}{2}.
\]

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($\gamma, \beta$) distribution reduces to the exponential($\beta$) distribution.

Mean/Variance: The relevant moments of $X \sim \text{Weibull}(\gamma, \beta)$ are
\[
E(X) = \beta^{1/\gamma} \Gamma(1 + 1/\gamma) \\
\text{var}(X) = \beta^{2/\gamma} \left[ \Gamma(1 + 2/\gamma) - \Gamma^2(1 + 1/\gamma) \right].
\]
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
\[
E(X) = \mu \\
\text{var}(X) = \sigma^2.
\]

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:
\[
z = g(x) = \frac{x - \mu}{\sigma} \quad \Rightarrow \quad x = g^{-1}(z) = \sigma z + \mu.
\]
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 \( z \in \mathbb{R} \),
\[ f_Z(z) = f_X(g^{-1}(z)) \left| \frac{d}{dz}g^{-1}(z) \right| = \frac{1}{\sqrt{2\pi\sigma}} e^{-(\sigma z + \mu - \mu)^2/2\sigma^2} \times \sigma \]
\[ = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}. \]

This is the pdf of \( Z \sim \mathcal{N}(0, 1) \).

**Remark:** We can derive \( E(Z) = 0 \) and \( \text{var}(Z) = 1 \) directly (i.e., using expected value definitions). First,
\[ E(Z) = \int_{\mathbb{R}} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = - \frac{1}{\sqrt{2\pi}} \left( e^{-z^2/2} \right|_{-\infty}^{\infty} = - \frac{1}{\sqrt{2\pi}} (0 - 0) = 0. \]

Second,
\[ E(Z^2) = \int_{\mathbb{R}} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \]

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 \( u = z^2 \implies du = 2zdz \). The last integral equals
\[ \int_{0}^{\infty} z^2 \frac{2}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} u \frac{1}{2z} e^{-u/2} du = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} u e^{-u/2} \frac{1}{\sqrt{2u}} du \]
\[ = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} u^{3/2 - 1} e^{-u/2} du. \]

Recognizing \( u^{3/2 - 1} e^{-u/2} \) 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
\[ \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} u^{3/2 - 1} e^{-u/2} du = \frac{1}{\sqrt{2\pi}} \Gamma \left( \frac{3}{2} \right) 2^{3/2} \]
\[ = \frac{1}{\sqrt{2\pi}} \frac{1}{2} \Gamma \left( \frac{1}{2} \right) 2^{3/2} = 1, \]

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
\[ Z \sim \mathcal{N}(0,1) \implies X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2), \]
it follows immediately that
\[ E(X) = \sigma E(Z) + \mu = \mu \]
and
\[ \text{var}(X) = \sigma^2 \text{var}(Z) = \sigma^2. \]

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
\[
E(X) = \frac{\alpha}{\alpha + \beta},
\]
\[
\text{var}(X) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}.
\]
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,

\[
E(X^k) = \int_0^1 x^k \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1} dx
\]

\[
= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha + k - 1}(1 - x)^{\beta - 1} dx
\]

\[
= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \Gamma(\alpha + k + \beta)
\]

\[
= \frac{\Gamma(\alpha + \beta)\Gamma(\alpha)}{\Gamma(\alpha)\Gamma(\alpha + \beta + k)}
\]

The mean of \( X \), for example, is

\[
E(X) = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + 1)}{\Gamma(\alpha)\Gamma(\alpha + \beta + 1)} = \frac{\Gamma(\alpha + \beta)\alpha\Gamma(\alpha)}{\Gamma(\alpha)\Gamma(\alpha + \beta)} = \frac{\alpha}{\alpha + \beta},
\]
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 \( X = \{x : 0 < x < 1\} \). For example,

1. \( \alpha = \beta \implies f_X(x|\alpha, \beta) \) is symmetric about \( x = 1/2 \)
   - \( \alpha = \beta = 1 \implies X \sim U(0, 1) \)
   - \( \alpha = \beta = \frac{1}{2} \implies f_X(x|\alpha, \beta) \propto x^{\frac{1}{2} - 1}(1 - x)^{\frac{1}{2} - 1} \) is “bathtub-shaped”

2. \( \alpha > \beta \implies f_X(x|\alpha, \beta) \) is skewed left

3. \( \alpha < \beta \implies f_X(x|\alpha, \beta) \) 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 \( \mu \) and \( \sigma \) 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

\[
\int_{\mathbb{R}} f_Z(z) dz = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{1}{1 + z^2} dz = \frac{1}{\pi} \left[ \arctan z \right]_{-\infty}^{\infty} = \frac{1}{\pi} \left[ \frac{\pi}{2} - \left( -\frac{\pi}{2} \right) \right] = 1,
\]

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

$$E(|Z|) = \int_{\mathbb{R}} |z| f_Z(z) dz = \int_{\mathbb{R}} |z| \frac{1}{\pi (1 + z^2)} dz.$$

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

$$\int_{\mathbb{R}} |z| \frac{1}{\pi (1 + z^2)} dz = \frac{2}{\pi} \int_{0}^{\infty} \frac{z}{1 + z^2} dz
= \frac{2}{\pi} \left[ \ln(1 + z^2) \right]_{0}^{\infty} = +\infty.$$

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:

$$X \sim N(\mu, \sigma^2) \implies Y = g(X) = e^X \sim \text{lognormal}(\mu, \sigma^2).$$

- 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 $\mu$ and has heavy tails, which makes it useful in robustness discussions
- sometimes also called the “double exponential distribution.”

- Inverted Gamma. $X \sim IG(\alpha, \beta)$, where $\alpha > 0$ and $\beta > 0$. This distribution arises according to the following transformation:

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

- 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:

\[
X \sim \text{exponential}(1/\beta) \implies Y = g(X) = \alpha e^X \sim \text{Pareto}(\alpha, \beta).
\]

- 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:

\[
X \sim \text{exponential}(1) \implies Y = g(X) = \mu + \beta \ln(e^X - 1) \sim \text{Logistic}(\mu, \beta).
\]

- 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:


### 3.4 Exponential Families

**Definition:** A family \( \{ f_X(x|\theta); \theta \in \Theta \} \) of pdfs (or pmfs) is called an **exponential family** if its members can be expressed as

\[
f_X(x|\theta) = h(x)c(\theta)\exp \left\{ \sum_{i=1}^{k} w_i(\theta)t_i(x) \right\},
\]

for real functions \( h(\cdot), c(\cdot), w_i(\cdot), \) and \( t_i(\cdot) \), where

- \( h(x) \geq 0 \) cannot depend on \( \theta \)
- \( c(\theta) \geq 0 \) cannot depend on \( x \)
- \( w_1(\theta), w_2(\theta), ..., w_k(\theta) \) cannot depend on \( x \)
- \( t_1(x), t_2(x), ..., t_k(x) \) cannot depend on \( \theta \).

**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

$$f_X(x|\theta) = \begin{cases} \frac{\theta^x e^{-\theta}}{x!}, & x = 0, 1, 2, \\ 0, & \text{otherwise}, \end{cases}$$

where $\Theta = \{\theta : \theta > 0\}$. Write the support $\mathcal{X} = \{x : x = 0, 1, 2, \ldots\}$ and the pmf as

$$f_X(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} I(x \in \mathcal{X}) = \frac{I(x \in \mathcal{X})}{x!} e^{-\theta} e^{\theta x} = h(x)c(\theta) \exp \{w_1(\theta)t_1(x)\},$$

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($\theta$) 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 $\Theta = \{\theta = (\alpha, \beta) : \alpha > 0, \beta > 0\}$. Note that the pdf of $X$ can be written as

$$f_X(x|\alpha, \beta) = \frac{I(x > 0)}{x} \frac{1}{\Gamma(\alpha)\beta^\alpha} e^{\alpha \ln x} e^{-x/\beta} = \frac{I(x > 0)}{x} \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp \left( \alpha \ln x - \frac{x}{\beta} \right) = h(x)c(\theta) \exp \{w_1(\theta)t_1(x) + w_2(\theta)t_2(x)\},$$

where $h(x) = I(x > 0)/x$, $c(\theta) = [\Gamma(\alpha)\beta^\alpha]^{-1}$, $w_1(\theta) = \alpha$, $t_1(x) = \ln x$, $w_2(\theta) = -1/\beta$, and $t_2(x) = x$. Therefore, the gamma($\alpha, \beta$) 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 $\Theta = \{\theta = (\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$. 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 $f_X(x|\theta)$ where the support $\mathcal{X}$ depends on an unknown parameter $\theta$, it is not possible to put $f_X(x|\theta)$ 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

$$f_X(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\},$$

where $\theta = (\theta_1, \theta_2, ..., \theta_d)'$ and $d = \text{dim}(\theta)$.

- When $d = k$, we call $\{f_X(x|\theta); \theta \in \Theta\}$ a **full** exponential family.
- When $d < k$, we call $\{f_X(x|\theta); \theta \in \Theta\}$ a **curved** exponential family.

**Example 3.6.** Suppose that $X \sim $ 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 $\Theta = \{\theta = (\alpha, \beta)': \alpha > 0, \beta > 0\}$. In Example 3.5, we showed that this was an exponential family with $d = k = 2$. Therefore, the gamma($\alpha, \beta$) family is a **full** exponential family. Now consider the gamma($\alpha, \beta$) **subfamily** where $\beta = 1/\alpha^2$, that is, $X \sim $ gamma($\alpha, 1/\alpha^2$). The pdf of $X$ is

$$f_X(x|\alpha) = \frac{1}{\Gamma(\alpha)(1/\alpha^2)^\alpha}x^{\alpha-1}e^{-x/(1/\alpha^2)}I(x > 0)$$

$$= \frac{I(x > 0)}{x} \alpha^{2\alpha} \frac{\alpha^{\alpha x} e^{-\alpha^2 x}}{\Gamma(\alpha)}$$

$$= \frac{I(x > 0)}{x} \frac{\alpha^{2\alpha}}{\Gamma(\alpha)} \exp(\alpha \ln x - \alpha^2 x)$$

$$= h(x)c(\alpha)\exp\{w_1(\alpha)t_1(x) + w_2(\alpha)t_2(x)\},$$

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($\alpha, \beta$) family, the **parameter space** is

$$\Theta = \{\theta = (\alpha, \beta)': \alpha > 0, \beta > 0\}.$$  

For the gamma($\alpha, 1/\alpha^2$) subfamily, the parameter space is

$$\Theta_0 = \left\{\theta = (\alpha, \beta)': \alpha > 0, \beta = \frac{1}{\alpha^2}\right\}.$$  

Clearly $\Theta_0 \subset \Theta$. Also, note that $\Theta$ contains an open set in $\mathbb{R}^2$, but $\Theta_0$ 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

$$F_X(x|\mu, \sigma) = P_X(X \leq x|\mu, \sigma) = P_Z(\sigma Z + \mu \leq x) = P_Z\left(Z \leq \frac{x - \mu}{\sigma}\right) = F_Z\left(\frac{x - \mu}{\sigma}\right).$$

The pdf of $X$ is therefore

$$f_X(x|\mu, \sigma) = \frac{d}{dx} F_X(x|\mu, \sigma) = \frac{d}{dx} F_Z\left(\frac{x - \mu}{\sigma}\right) = \frac{1}{\sigma} f_Z\left(\frac{x - \mu}{\sigma}\right),$$

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 $\mu$ and $\sigma$.

Remark: What we have just proven is essentially the sufficiency part ($\Leftarrow$) of Theorem 3.5.6 (pp 120 CB). The necessity part ($\Rightarrow$) 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 $\mu$ 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 $Z \sim \text{exponential}(1)$; i.e., the pdf of $Z$ is
\[
f_Z(z) = e^{-z}I(z > 0).
\]

The pdf of $X = Z + \mu$ is therefore
\[
f_X(x|\mu) = f_Z(x - \mu) = e^{-(x-\mu)}I(x - \mu > 0) = e^{-(x-\mu)}I(x > \mu).
\]

This is called a \textbf{shifted exponential distribution} with location parameter $\mu$. 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 $\mu > 0$.

\textbf{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 \textbf{scale family} generated by $f_Z(z)$. The parameter $\sigma$ is called a \textbf{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.

\textbf{Solution.} We have to identify the standard pdf $f_Z(z)$ and the scale parameter $\sigma$ 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
\[
\frac{1}{\sigma}f_Z\left(\frac{x}{\sigma}\right) = \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}}e^{-\left(\frac{x}{\sigma}\right)^2/2}I\left(\frac{x}{\sigma} \in \mathbb{R}\right)
\]
\[
= \frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}I(x \in \mathbb{R}).
\]

Thus, the $\mathcal{N}(0, \sigma^2)$ family is a scale family with standard pdf $f_Z(z)$ and scale parameter $\sigma$.

\textbf{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, $\mu$ and $\sigma$ 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

\[
f_X(x|\mu, \sigma) = \frac{1}{\pi \sigma} \left[ 1 + \left( \frac{x-\mu}{\sigma} \right)^2 \right]^{-1} I(x \in \mathbb{R}),
\]

where \(-\infty < \mu < \infty \) and \( \sigma > 0 \). It is easy to see that the Cauchy(\( \mu, \sigma \)) 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 \( \mu \) and \( \sigma \) 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, \( \mu \) satisfies

\[
P_X(X \leq \mu) = \int_{-\infty}^{\mu} f_X(x|\mu, \sigma) dx = 0.5;
\]

i.e., \( \mu \) 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

\[
E(X) = \sigma E(Z) + \mu \quad \text{and} \quad \text{var}(X) = \sigma^2 \text{var}(Z).
\]

Proof. The expected value of \( X \) is

\[
E(X) = E(\sigma Z + \mu) = \int_{\mathbb{R}} (\sigma z + \mu) f_Z(z) dz
\]

\[
= \sigma \int_{\mathbb{R}} z f_Z(z) dz + \int_{\mathbb{R}} \mu f_Z(z) dz
\]

\[
= \sigma E(Z) + \mu.
\]

Showing \( \text{var}(X) = \sigma^2 \text{var}(Z) \) involves slightly more work, but is just as straightforward. \( \Box \)

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 \( \mu \) and \( \sigma \). 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

$$E(Y) = \int_{0}^{\infty} y f_Y(y) dy \geq \int_{\{y \geq r\}} y f_Y(y) dy \geq \int_{\{y \geq r\}} r f_Y(y) dy = r P_Y(Y \geq r). \quad \Box$$

**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 $\sigma$ 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

$$P_X((X - \mu)^2 \geq k^2 \sigma^2) \leq \frac{E[(X - \mu)^2]}{k^2 \sigma^2} = \frac{\sigma^2}{k^2 \sigma^2} = \frac{1}{k^2}. \quad \Box$$

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). This is should not be surprising because we are providing almost no information about the 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 \(\sigma\)-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 \(X = (X_1, X_2, ..., X_n)\) is straightforward and we do this in Section 4.6. In this case, \(X\) is a mapping from \((S, \mathcal{B}, P)\) to \((\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_X)\) with the property that
\[
X^{-1}(B) \equiv \{\omega \in S : X(\omega) \in B\} \in \mathcal{B},
\]
for all \(B \in \mathcal{B}(\mathbb{R}^n)\). As in the univariate random variable case, the measurability condition \(X^{-1}(B) \in \mathcal{B}\), for all \(B \in \mathcal{B}(\mathbb{R}^n)\), suggests that events of interest like \(\{X \in B\}\) on \((\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_X)\) can be assigned a probability in the same way that \(\{\omega \in S : X(\omega) \in B\}\) can be assigned a probability on \((S, \mathcal{B}, P)\).

Example 4.1. Experiment: Toss two dice. Assume the model
\[
\begin{align*}
S &= \{\omega = (\omega_1, \omega_2) : \omega_i \in \{1, 2, ..., 6\}, i = 1, 2\} \\
\mathcal{B} &= 2^S \\
P &= \text{equiprobability measure; i.e., } P(\{\omega\}) = 1/36, \text{ for all } \omega \in S.
\end{align*}
\]
Define the random variables
\[
\begin{align*}
X_1 &= \text{sum; i.e., } X_1(\omega) = \omega_1 + \omega_2 \\
X_2 &= \text{absolute difference; i.e., } X_2(\omega) = |\omega_1 - \omega_2|
\end{align*}
\]
and let \(X = (X_1, X_2)\). We can envision the bivariate random vector
\[
X(\omega) = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}(\omega) = \begin{pmatrix} X_1(\omega) \\ X_2(\omega) \end{pmatrix}
\]
as a vector of (univariate) random variables on \((S, \mathcal{B}, P)\).
Illustration: To show how probabilities are assigned on \((\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2), P_{X_1,X_2})\) in Example 4.1, consider the (two-dimensional) Borel set \(B = \{(5,3)\} \in \mathbb{R}^2\). Note that

\[
P_{X_1,X_2}(X \in B) = P_{X_1,X_2}(X_1 = 5, X_2 = 3) = P((1,4),(4,1)) = \frac{2}{36}.
\]

This suggests that we can write

\[
P_{X_1,X_2}(X \in B) = \underbrace{P_{X_1,X_2}(X_1 = 5, X_2 = 3)}_{\text{calculated on } (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))} = \underbrace{P((1,4),(4,1))}_{\text{calculated on } (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))} = \frac{2}{36}.
\]

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

\[
P(X \in B) = P((X, Y) \in B \times \mathbb{R}) = \sum_{(x, y) \in B \times \mathbb{R}} f_{X,Y}(x, y) = \sum_{x \in B} \sum_{y \in \mathbb{R}} f_{X,Y}(x, y).
\]

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:

<table>
<thead>
<tr>
<th></th>
<th>(y)</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(x)</td>
<td>0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The entries in the table are the joint probabilities \(f_{X,Y}(x, y)\). The support of \((X, Y)\) is

\[A = \{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2)\}.
\]

We use this joint pmf to calculate various quantities, illustrating many of the ideas we have seen so far. For example,

\[
P(X = 1, Y = 1) = f_{X,Y}(1, 1) = 0.1
\]

\[E(XY) = \sum_{x=0}^{1} \sum_{y=0}^{2} xy f_{X,Y}(x, y) = (0)(0)(0.1) + (0)(1)(0.2) + (0)(2)(0.2) + (1)(0)(0.3) + (1)(1)(0.1) + (1)(2)(0.1) = 0.3.
\]

The marginal pmfs of \(X\) and \(Y\) are, respectively,

\[
f_X(x) = 0.5I(x = 0) + 0.5I(x = 1)
f_Y(y) = 0.4I(y = 0) + 0.3I(y = 1) + 0.3I(y = 2).
\]
Note that
\[ E(X) = 0.5 \]
\[ E(Y) = 0.9. \]

These can be calculated from the marginal distributions \( f_X(x) \) and \( f_Y(y) \), respectively, or from using the joint distribution, for example,
\[
E(X) = \sum_{x=0}^{1} \sum_{y=0}^{2} x f_{X,Y}(x,y) \\
= (0)(0.1) + (0)(0.2) + (0)(0.2) + (1)(0.3) + (1)(0.1) + (1)(0.1) = 0.5.
\]

**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 \( B \in \mathcal{B}(\mathbb{R}^2) \),
\[
P((X, Y) \in B) = \int_B \int f_{X,Y}(x,y) dx dy.
\]

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

1. \( f_{X,Y}(x,y) \geq 0, \) for all \((x, y) \in \mathbb{R}^2\)
2. \( \int \int_{\mathbb{R}^2} f_{X,Y}(x,y) dx dy = 1. \)

**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 \( P(X - Y > \frac{1}{8}) \).

**Solution.** First, note that the two-dimensional support set identified in the indicator function \( I(0 < y < x < 1) \) is
\[
A = \{(x, y) \in \mathbb{R}^2 : 0 < y < x < 1\}
\]
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
\[
\int \int_{\mathbb{R}^2} f_{X,Y}(x,y) dx dy = 1.
\]

Therefore, the calculation
\[
\int_{y=0}^{1} \int_{x=y}^{1} cxy \ dx \ dy = \frac{c}{8} \text{ set} = 1
\]
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,

$$P\left(X - Y > \frac{1}{8}\right) = \int_B \int f_{X,Y}(x,y) \, dxdy = \int_{y=0}^1 \int_{x=y+\frac{1}{8}}^1 8xy \, dxdy \approx 0.698. $$

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 \(g: \mathbb{R}^2 \rightarrow \mathbb{R}\). Then \(g(X,Y)\) is a univariate random variable and its expected value is

\[
E[g(X,Y)] = \int_{\mathbb{R}^2} g(x,y)f_{X,Y}(x,y)dxdy.
\]

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 \(\int_{\mathbb{R}^2} |g(x,y)|f_{X,Y}(x,y)dxdy\) 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

\[
f_{X,Y}(x,y) = 8xy \mathbb{I}(0 < y < x < 1),
\]

as in Example 4.3 (see the support \(A\) in Figure 4.1). We have

\[
E(X^2Y) = \int_{\mathbb{R}^2} x^2y f_{X,Y}(x,y)dxdy
= \int_{y=0}^{1} \int_{x=y}^{1} x^2y \times 8xy \, dxdy
= \int_{y=0}^{1} \int_{x=y}^{1} 8x^3y^2 \, dxdy = \frac{2}{7}.
\]

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

\[
P(X \in B) = P(X \in B, Y \in \mathbb{R}) = P((X,Y) \in B \times \mathbb{R})
= \int_{B \times \mathbb{R}} f_{X,Y}(x,y)dxdy
= \int_{B} \int_{\mathbb{R}} f_{X,Y}(x,y)dy \, dx,
\]

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 \(Y\). In other words, 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

\[
f_{X,Y}(x,y) = 8xy \mathbb{I}(0 < y < x < 1),
\]

as in Example 4.3.
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., $X \sim \text{beta}(4, 1)$. as in Example 4.3 (see the support $A$ 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 $0 < y < 1$,

$$f_X(x) = \int_{x=y}^{1} 8xy \, dy = 4y(1 - y^2).$$

Summarizing,

$$f_X(x) = 4x^3I(0 < x < 1) \quad \text{and} \quad f_Y(y) = 4y(1 - y^2)I(0 < y < 1).$$

These marginal pdfs are shown in Figure 4.2. 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.

**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 $Y$,

$$P\left(Y > \frac{1}{2}\right) = \int_{y=\frac{1}{2}}^{1} f_Y(y) \, dy = \frac{9}{16}.$$
Using the joint distribution of \((X, Y)\),
\[
P\left(Y > \frac{1}{2}\right) = \int_{y=\frac{1}{2}}^{1} \int_{x=y}^{1} f_{X,Y}(x, y) \, dx \, dy = \frac{9}{16}.
\]

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 \(Z = \{z : z > 0\}\).

We derive the cdf of \(Z\) first:
\[
F_Z(z) = P(Z \leq z) = P\left(\frac{X}{Y} \leq z\right) = \int_{z}^{\infty} \int_{B} f_{X,Y}(x, y) \, dx \, dy,
\]

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
\[
\int_{x=0}^{\infty} \int_{y=x/z}^{\infty} e^{-(x+y)} \, dy \, dx = \frac{z}{z+1}.
\]

Therefore, the cdf of \(Z\) is
\[
F_Z(z) = \begin{cases} 
0, & z \leq 0 \\
\frac{z}{z+1}, & z > 0.
\end{cases}
\]

The pdf of \(Z\) is
\[
f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{1}{(z+1)^2} I(z > 0)
\]

and is shown in Figure 4.3.
Finally, note that
\[
E(Z) = \int_{\mathbb{R}} z f_Z(z) \, dz = \int_{0}^{\infty} \frac{z}{(z+1)^2} \, dz = \int_{1}^{\infty} \frac{u-1}{u^2} \, du = \left( \ln u + \frac{1}{u} \right)_{u=1}^{\infty} = +\infty;
\]
i.e., \( E(Z) \) does not exist.

**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 \( X = (X_1, X_2)' \) is a bivariate random vector (discrete or continuous). The joint moment generating function (mgf) of \( X_1 \) and \( X_2 \) is

\[
M_X(t) = E(e^{t'X}) = E(e^{t_1 X_1 + t_2 X_2}),
\]

where \( t = (t_1, t_2)' \). For \( M_X(t) \) to exist, this expectation must be finite in an open neighborhood about \( t = 0 \); i.e., \( E(e^{t_1 X_1 + t_2 X_2}) < \infty \) \( \forall t_1 \in (-h_1, h_1) \) \( \forall t_2 \in (-h_2, h_2) \) \( \exists h_1 > 0 \) \( \exists h_2 > 0 \).

Notes:

1. We may also write \( M_X(t) = M_{X_1,X_2}(t_1, t_2) \).
2. As with mgfs for univariate random variables, a random vector’s mgf \( M_X(t) \) uniquely identifies the distribution of \( X \).
3. It is easy to see that

\[
M_{X_1}(t_1) = M_{X_1,X_2}(t_1, 0) \quad M_{X_2}(t_2) = M_{X_1,X_2}(0, t_2).
\]

Therefore, the marginal mgfs are easily obtained from the joint mgf.

Example 4.7. Suppose \( X = (X_1, X_2)' \) 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

\[
M_{X_1,X_2}(t_1, t_2) = E(e^{t_1 X_1 + t_2 X_2}) = \int \int e^{t_1 x_1 + t_2 x_2} f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 = \int_{x_2=0}^{\infty} \int_{x_1=0}^{x_2} e^{t_1 x_1 + t_2 x_2} e^{-x_2} \, dx_1 dx_2 = \frac{1}{(1 - t_1)(1 - t_2)}.
\]

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

\[
X_1 \sim \text{exponential}(1) \quad X_2 \sim \text{gamma}(2, 1).
\]
Definition: Suppose that \( X = (X_1, X_2)' \) is a random vector. The expected value of \( X \) is

\[
E(X) = \begin{pmatrix} E(X_1) \\ E(X_2) \end{pmatrix},
\]

provided that both \( E(X_1) \) and \( E(X_2) \) exist. In the last example, we see that

\[
E(X) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.
\]

It is also possible to calculate \( E(X) \) using the joint mgf:

\[
E(X) = \frac{\partial M_X(t)}{\partial t} \bigg|_{t=0} = \begin{pmatrix} \frac{\partial M_X(t)}{\partial t_1} \\ \frac{\partial M_X(t)}{\partial t_2} \end{pmatrix} \bigg|_{t_1=t_2=0};
\]

i.e., \( E(X) \) is the gradient of \( M_X(t) \) evaluated at \( t = 0 \). 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.,

\[
f_{Y|X}(y | x) \equiv P(Y = y | X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f_{X,Y}(x, y)}{f_X(x)}.
\]

**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:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.2</td>
</tr>
<tr>
<td>1</td>
<td>0.3</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>
The conditional pmf of $Y$, when $X = x = 0$, is found as follows:

\[
\begin{align*}
  f_{Y \mid X}(0 \mid 0) &= \frac{f_{X,Y}(0,0)}{f_X(0)} = \frac{0.1}{0.5} = 0.2 \quad \text{(for } y = 0) \\
  f_{Y \mid X}(1 \mid 0) &= \frac{f_{X,Y}(0,1)}{f_X(0)} = \frac{0.2}{0.5} = 0.4 \quad \text{(for } y = 1) \\
  f_{Y \mid X}(2 \mid 0) &= \frac{f_{X,Y}(0,2)}{f_X(0)} = \frac{0.2}{0.5} = 0.4 \quad \text{(for } y = 2)
\end{align*}
\]

Therefore,

\[
f_{Y \mid X}(y \mid 0) = 0.2I(y = 0) + 0.4I(y = 1) + 0.4I(y = 2).
\]

Note: Suppose $B \in \mathcal{B}(\mathbb{R})$. Conditional probabilities can be calculated using conditional pmfs as follows:

\[
\begin{align*}
P(Y \in B \mid X = x) &= \sum_{y \in B} f_{Y \mid X}(y \mid x) \\
P(X \in B \mid Y = y) &= \sum_{x \in B} f_{X \mid Y}(x \mid y).
\end{align*}
\]

For example, in Example 4.8,

\[
P(Y \leq 1 \mid X = 0) = \sum_{y=0}^{1} f_{Y \mid X}(y \mid 0) = 0.2 + 0.4 = 0.6.
\]

**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

\[
f_{Y \mid X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)},
\]

which is defined for values of $x$ where $f_X(x) > 0$.

**Interpretation:** The function $f_{Y \mid X}(y \mid x)$ 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

\[
f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},
\]

which is defined for values of $y$ where $f_Y(y) > 0$. This function is a univariate 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 $X \sim \text{exponential}(1)$ and $Y \sim \text{gamma}(2,1)$ so that the marginal pdfs are

\[
\begin{align*}
f_X(x) &= e^{-x}I(x > 0) \\
f_Y(y) &= ye^{-y}I(y > 0).
\end{align*}
\]
The conditional pdf of $Y$ given $X = x$ is therefore

\[
f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{e^{-y}I(0 < x < y < \infty)}{e^{-x}I(x > 0)} = e^{-(y-x)}I(y > x).
\]

This function describes the distribution of $Y$ when $X$ is fixed at $x > 0$. In Figure 4.4, 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

\[
f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{e^{-y}I(0 < x < y < \infty)}{ye^{-y}I(y > 0)} = \frac{1}{y}I(0 < x < y).
\]

This function describes the distribution of $X$ when $Y$ is fixed at $y > 0$. In Figure 4.4, 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:

$$P(Y \in B \mid X = x) = \int_B f_{Y \mid X}(y \mid x)dy$$
$$P(X \in B \mid Y = y) = \int_B f_{X \mid Y}(x \mid y)dx.$$

For example, in Example 4.9,

$$P(Y < 5 \mid X = 2) = \int_{y=2}^{5} e^{-(y-2)}dy = 1 - e^{-3}$$

and

$$P(X > 3 \mid Y = 5) = \int_{x=3}^{5} \frac{1}{5} dx = \frac{2}{5}.$$

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:

$$E[g(Y) \mid X = x] = \int_{\mathbb{R}} g(y)f_{Y \mid X}(y \mid x)dy$$
$$E[h(X) \mid Y = y] = \int_{\mathbb{R}} h(x)f_{X \mid Y}(x \mid y)dx.$$  

Notes:

1. If $(X, Y)$ is discrete, then integrals above are replaced by sums.
2. The same existence issues still remain; for example, for $E[g(Y) \mid X = x]$ to exist, we need $\int_{\mathbb{R}} |g(y)|f_{Y \mid X}(y \mid x)dy < \infty$.

Special case: If $g(Y) = Y$ and $h(X) = X$, then

$$E(Y \mid X = x) = \int_{\mathbb{R}} yf_{Y \mid X}(y \mid x)dy$$
$$E(X \mid Y = y) = \int_{\mathbb{R}} xf_{X \mid Y}(x \mid y)dx.$$ 

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.

$$E(Y \mid X = x) \leftarrow \text{function of } x; \text{ fixed}$$
$$E(Y \mid X) \leftarrow \text{function of } X; \text{ random variable}$$

$$E(X \mid Y = y) \leftarrow \text{function of } y; \text{ fixed}$$
$$E(X \mid Y) \leftarrow \text{function of } Y; \text{ random variable}$$
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
\[
\begin{align*}
    f_{Y|X}(y|x) &= e^{-(y-x)}I(y > x) \\
    f_{X|Y}(x|y) &= \frac{1}{y}I(0 < x < y).
\end{align*}
\]
The conditional mean of \( Y \) given \( X = x \) is
\[
E(Y|X = x) = \int_{\mathbb{R}} y e^{-(y-x)} I(y > x) dy
\]
\[= \int_{y=x}^{\infty} y e^{-(y-x)} dy = \int_{0}^{\infty} (u+x) e^{-u} du = E(U + x),
\]
where \( U \sim \text{exponential}(1) \); in the last integral, note that \( e^{-u} I(u > 0) \) is the pdf of \( U \sim \text{exponential}(1) \). Therefore,
\[
E(Y|X = x) = E(U + x) = E(U) + x = 1 + x.
\]
The conditional mean of \( X \) given \( Y = y \) is
\[
E(X|Y = y) = \int_{\mathbb{R}} x \frac{1}{y} I(0 < x < y) dx
\]
\[= \frac{1}{y} \int_{x=0}^{y} x dx
\]
\[= \frac{1}{y} \left( \frac{x^2}{2} \right)_{x=0}^{y} = \frac{y}{2}.
\]
This should not be surprising because \( X|\{Y = y\} \sim U(0, y) \).

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
\[
E(Y|X = x) = \mu_{Y|X=x} \\
E(X|Y = y) = \mu_{X|Y=y}
\]
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:
\[
\text{var}(Y|X = x) = E(Y^2|X = x) - [E(Y|X = x)]^2
\]
\[
\text{var}(X|Y = y) = E(X^2|Y = y) - [E(X|Y = y)]^2.
\]

**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 $t \in (-h, h)$ for some $h > 0$. 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 \( y > 0 \),

\[
f_Y(y) = \int_{\mathbb{R}} \frac{e^{-x/y}e^{-y}}{y} I(x > 0, y > 0) dx = e^{-y} \int_{x=0}^{\infty} \frac{1}{y} e^{-x/y} dx = e^{-y}.
\]

Therefore, the conditional pdf of \( X \) given \( Y = y \) is

\[
f_{X|Y}(x|y) = \frac{e^{-x/y}e^{-y}}{y} I(x > 0, y > 0) = \frac{1}{y} e^{-x/y} I(x > 0).
\]

That is, \( X|\{Y = y\} \sim \text{exponential}(y) \). Finally, the conditional mgf of \( X \) given \( Y = y \) is

\[
M_{X|Y}(t) = E(e^{tX}|Y = y) = \int_{\mathbb{R}} e^{tx} \frac{1}{y} e^{-x/y} I(x > 0) dx = \frac{1}{1 - yt}, \quad \text{for } t < \frac{1}{y}.
\]

Now, let’s illustrate how to use \( M_{X|Y}(t) \) to “generate” conditional moments:

\[
\frac{\partial}{\partial t} M_{X|Y}(t) = y(1 - yt)^{-2} \quad \Rightarrow \quad E(X|Y = y) = \frac{\partial}{\partial t} M_{X|Y}(t) \bigg|_{t=0} = y
\]

\[
\frac{\partial^2}{\partial t^2} M_{X|Y}(t) = 2y^2(1 - yt)^{-3} \quad \Rightarrow \quad E(X^2|Y = y) = \frac{\partial^2}{\partial t^2} M_{X|Y}(t) \bigg|_{t=0} = 2y^2.
\]

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 \( X|\{Y = y\} \sim \text{exponential}(y) \).

**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

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

Therefore, if \( X \perp \perp Y \), then for any \( B \in \mathcal{B}(\mathbb{R}) \),

\[
P(Y \in B|X = x) = \int_B f_{Y|X}(y|x) dy = \int_B f_Y(y) dy = P(Y \in B).
\]
In other words, knowledge that \( X = x \) does not influence how we assign probability to the event \( \{ Y \in B \} \). Similarly, if \( X \perp Y \), 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 (\(\Rightarrow\)) is straightforward. Suppose that \(X \perp Y\), and take \( g(x) = f_X(x) \) and \( h(y) = f_Y(y) \). Because

\[
f_{X,Y}(x, y) \overset{X \perp Y}{=} f_X(x)f_Y(y) = g(x)h(y),
\]
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 (\(\Leftarrow\)) 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

\[
\int_{\mathbb{R}} g(x)dx = c \quad \text{and} \quad \int_{\mathbb{R}} h(y)dy = d.
\]

Note that

\[
cd = \int_{\mathbb{R}} g(x)dx \int_{\mathbb{R}} h(y)dy = \int_{\mathbb{R}^2} g(x)h(y)dxdy = \int_{\mathbb{R}^2} f_{X,Y}(x, y)dxdy = 1,
\]
because the factorization \( f_{X,Y}(x,y) = g(x)h(y) \) holds by assumption. Furthermore,

\[
 f_X(x) = \int f_{X,Y}(x,y)\,dy = \int g(x)h(y)\,dy = dg(x).
\]

An analogous argument shows that \( f_Y(y) = ch(y) \). Therefore, for all \( x,y \in \mathbb{R} \), we have

\[
 f_{X,Y}(x,y) = g(x)h(y) = dg(x)ch(y) = f_X(x)f_Y(y),
\]

showing that \( X \perp \perp Y \). For the discrete case, simply replace integrals with sums. \( \square \)

**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

\[
 f_{X,Y}(x,y) = \frac{1}{384}x^2e^{-x/2}I(x > 0) \times y^4e^{-y}I(y > 0).
\]

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,

\[
 E[g(X)h(Y)] = \int \int g(x)h(y)f_{X,Y}(x,y)\,dxdy
\]

\[
 X \perp \perp Y \implies \int \int g(x)h(y)f_X(x)f_Y(y)\,dxdy
\]

\[
 = \int g(x)f_X(x)dx \int h(y)f_Y(y)\,dy = E[g(X)]E[h(Y)].
\]
If \((X,Y)\) is discrete, simply replace integrals with sums. To prove part (a), suppose \(A,B \in \mathcal{B}(\mathbb{R})\) and define
\[
  g(X) = I(X \in A) \\
  h(Y) = I(Y \in B).
\]
Because the expectation of an indicator function is the probability of the set that it indicates (see next remark), we have
\[
  E[g(X)h(Y)] = E[I(X \in A)I(Y \in B)] = P(X \in A, Y \in B),
\]
and
\[
  E[g(X)]E[h(Y)] = E[I(X \in A)]E[I(Y \in B)] = P(X \in A)P(Y \in B).
\]
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
\[
  M_Z(t) = E(e^{tZ}) = E[e^{t(X+Y)}] = E(e^{tX}e^{tY}) \\
  \overset{\text{X} \perp \perp Y}{=} E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t). \square
\]

Remark: Theorem 4.2.12 is extremely useful. If \(X \perp Y\), 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 $X \perp Y$. The mgf of $Z = X + Y$ is 

$$M_Z(t) = M_X(t)M_Y(t) = e^{\lambda_1(e^t-1)}e^{\lambda_2(e^t-1)} = e^{(\lambda_1+\lambda_2)(e^t-1)},$$

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, $X \perp Y$.

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 $X_1 \perp X_2$. 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 $x_1 = 0, 1, 2, ..., n$,

$$f_{X_1|Z}(x_1|n) = \frac{f_{X_1,Z}(x_1,n)}{f_Z(n)} = \frac{P(X_1 = x_1, Z = n)}{P(Z = n)} = \frac{P(X_1 = x_1, X_2 = n - x_1)}{P(Z = n)}$$

$$= \frac{\lambda_1^{x_1}e^{-\lambda_1}\lambda_2^{n-x_1}e^{-\lambda_2}}{(n-x_1)!} \frac{x_1!(n-x_1)!}{(\lambda_1 + \lambda_2)^n(n-x_1)!} = \frac{n!}{x_1!(n-x_1)!} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{x_1} \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{n-x_1} \left( 1 - \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^{n-x_1}.$$

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 \(A \subseteq \mathbb{R}^2\) 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 \(A \subseteq \mathbb{R}^2\). Define the random variables
\[
U = g_1(X,Y)
\]
\[
V = g_2(X,Y),
\]
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,
\[
F_{U,V}(u,v) = P(U \leq u, V \leq v) = P(g_1(X,Y) \leq u, g_2(X,Y) \leq v) = \int_B \int f_{X,Y}(x,y)\,dx\,dy,
\]
where the set \(B = \{(x,y) \in A : g_1(x,y) \leq u, g_2(x,y) \leq v\}\). 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,

\[
 f_{U,V}(u, v) = P(U = u, V = v) = P(g_1(X, Y) = u, g_2(X, Y) = v) = \sum_{(x,y) \in A_{uv}} f_{X,Y}(x,y),
\]

where the set \(A_{uv} = \{(x, y) \in A : g_1(x,y) = u, g_2(x,y) = v\}\).

**Notation:** To match the notation of CB, we denote by

\[A = \text{support of } (X,Y)\]
\[B = \text{support of } (U,V)\]

Note that

\[B = \{(u,v) \in \mathbb{R}^2 : u = g_1(x,y), v = g_2(x,y), \text{ for } (x,y) \in A\}\].

The vector-valued function \(g : \mathbb{R}^2 \rightarrow \mathbb{R}^2\) 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 \(A\) to \(B\); i.e., \(g : A \rightarrow B\).

**Example 4.15.** Suppose that \(X \sim b(n_1, p)\), \(Y \sim b(n_2, p)\), and \(X \perp \perp Y\). Define

\[
 U = g_1(X,Y) = X + Y \\
 V = g_2(X,Y) = Y.
\]

(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

\[
 f_{X,Y}(x,y) \overset{X \perp \perp Y}{=} f_X(x)f_Y(y) = \frac{n_1}{x}p^x(1-p)^{n_1-x}\left(\begin{array}{c} n_2 \\ y \end{array}\right)p^y(1-p)^{n_2-y},
\]

for values \((x, y) \in A\), where

\[A = \{(x, y) : x = 0, 1, 2, ..., n_1; y = 0, 1, 2, ..., n_2\}\].

The support of \((U, V)\) is

\[B = \{(u,v) \in \mathbb{R}^2 : u = x + y, v = y, \text{ for } (x,y) \in A\} = \{(u,v) \in \mathbb{R}^2 : u = 0, 1, 2, ..., n_1 + n_2, v = 0, 1, 2, ..., n_2, v \leq u\};\]

note that necessarily \(v \leq u\) because \(x \geq 0\). Now, the joint pmf of \((U, V)\) equals

\[
 f_{U,V}(u, v) = \sum_{(x,y) \in A_{uv}} f_{X,Y}(x,y),
\]
where the set \( A_{uv} = \{(x, y) \in A : x + y = u, y = v\} \). In this case, the set \( A_{uv} \) consists of just one point, the singleton \( \{(u - v, v)\} \). To see why this is true, note that the system

\[
\begin{align*}
  u &= g_1(x, y) = x + y \\
  v &= g_2(x, y) = y
\end{align*}
\]

has only one (unique) solution

\[
\begin{align*}
  x &= g_1^{-1}(u, v) = u - v \\
  y &= g_2^{-1}(u, v) = v.
\end{align*}
\]

Therefore, the joint pmf of \((U, V)\), for values of \((u, v) \in B\), is given by

\[
\sum_{(x, y) \in A_{uv}} f_{X,Y}(x, y)
\]

\[
= \binom{n_1}{u - v} p^{u-v}(1-p)^{n_1-(u-v)} \binom{n_2}{v} p^v (1-p)^{n_2-v}
\]

\[
= \binom{n_1}{u-v} \binom{n_2}{v} p^u (1-p)^{n_1+n_2-u}.
\]

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 \( v \in B \), that is, \( v = 0, 1, 2, \ldots, u \). For \( u = 0, 1, 2, \ldots, n_1 + n_2 \), we have

\[
\sum_{v=0}^{u} \binom{n_1}{u-v} \binom{n_2}{v} p^u (1-p)^{n_1+n_2-u}
\]

\[
= p^u (1-p)^{n_1+n_2-u} \sum_{v=0}^{u} \binom{n_1}{u-v} \binom{n_2}{v}.
\]

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 \( u = 0, 1, 2, \ldots, n_1 + n_2 \), 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

\[
M_U(t) \overset{X+Y}{=} M_X(t)M_Y(t) = (q + pe^t)^{n_1}(q + pe^t)^{n_2} = (q + pe^t)^{n_1+n_2},
\]

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 \( A \subseteq \mathbb{R}^2 \). Define

\[
U = g_1(X, Y) \\
V = g_2(X, Y)
\]
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 \(A\) to \(B \subseteq \mathbb{R}^2\), where
\[
B = \{ (u,v) \in \mathbb{R}^2 : u = g_1(x,y), \ v = g_2(x,y), \text{ for } (x,y) \in A \};
\]
i.e., \(g : A \to B\). In what follows, we will assume that \(g\) is a one-to-one transformation. That is, for each \((u,v)\) \(\in B\), there is only one \((x,y)\) \(\in A\) satisfying
\[
\begin{align*}
u &= g_1(x,y) \\
v &= g_2(x,y).
\end{align*}
\]
Because \(g\) is one-to-one, we can find the inverse transformation
\[
\begin{align*}
x &= g_1^{-1}(u,v) \\
y &= g_2^{-1}(u,v).
\end{align*}
\]
The Jacobian of the (inverse) transformation is defined as
\[
J = \det \begin{vmatrix}
\frac{\partial g_1^{-1}(u,v)}{\partial u} & \frac{\partial g_1^{-1}(u,v)}{\partial v} \\
\frac{\partial g_2^{-1}(u,v)}{\partial u} & \frac{\partial g_2^{-1}(u,v)}{\partial v}
\end{vmatrix},
\]
that is, \(J\) is the determinant of this \(2 \times 2\) matrix of partial derivatives. We will assume that \(J \neq 0\) over \(B\). By a theorem in analysis (the Change of Variables Theorem), we are able to conclude that the joint pdf of \((U,V)\) is, for \((u,v) \in B\),
\[
f_{U,V}(u,v) = f_{X,Y}(g_1^{-1}(u,v), g_2^{-1}(u,v))|J|,
\]
where \(|J|\) denotes the absolute value of \(J\). Of course, if \((u,v) \notin B\), then \(f_{U,V}(u,v) = 0\).

**Discussion:** Let \(A \subseteq A\) and \(B = g(A) \subseteq B\); i.e., \(g(A)\) is the image of \(A\) under the mapping \(g\). Because \(g : A \to B\) is one-to-one, the events \(\{(X,Y) \in A\}\) and \(\{(U,V) \in B\}\) have the same probability; i.e.,
\[
P((U,V) \in B) = P((X,Y) \in A) = \int_A \int f_{X,Y}(x,y) dx dy.
\]
The Change of Variables Theorem from analysis says that
\[
\int_A \int f_{X,Y}(x,y) dx dy = \int_B \int f_{X,Y}(g_1^{-1}(u,v), g_2^{-1}(u,v))|J| dudv.
\]
Therefore, for any \(B \subseteq B\), 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 $X \perp Y$. Define

$U = g_1(X, Y) = X + Y$

$V = g_2(X, Y) = \frac{X}{X + Y}$.

(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

$$f_{X,Y}(x,y) = \frac{1}{\Gamma(\alpha_1)\beta^{\alpha_1}} x^{\alpha_1-1} e^{-x/\beta} I(x > 0) \times \frac{1}{\Gamma(\alpha_2)\beta^{\alpha_2}} y^{\alpha_2-1} e^{-y/\beta} I(y > 0)$$

and the support of $(X,Y)$ is

$A = \{(x,y) \in \mathbb{R}^2 : x > 0, y > 0\}$.

The transformation above maps values of $(x,y) \in A$ to $B = \{(u,v) \in \mathbb{R}^2 : u > 0, 0 < v < 1\}$; i.e., $B$ is the support of $(U,V)$. To verify the transformation is one-to-one, we show that $g(x,y) = g(x^*,y^*) \in B \implies x = x^*$ and $y = y^*$, where

$$g \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} g_1(x,y) \\ g_2(x,y) \end{pmatrix} = \begin{pmatrix} x + y \\ \frac{x}{x+y} \end{pmatrix}.$$ 

Suppose $g(x,y) = g(x^*,y^*)$. This means that both of these equations hold:

$$x + y = x^* + y^* \quad \text{and} \quad \frac{x}{x+y} = \frac{x^*}{x^* + y^*}.$$ 

The two equations together imply that $x = x^*$. The first equation then implies $y = y^*$. Hence, the transformation $g : A \to B$ is one-to-one. The inverse transformation is found by solving

$$u = x + y$$

$$v = \frac{x}{x+y}$$

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

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(u,v)}{\partial u} & \frac{\partial g_1^{-1}(u,v)}{\partial v} \\ \frac{\partial g_2^{-1}(u,v)}{\partial u} & \frac{\partial g_2^{-1}(u,v)}{\partial v} \end{vmatrix} = \det \begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix} = -uv - u(1-v) = -u.$$
which is nonzero over \( \mathcal{B} \). Therefore, the joint pdf of \((U, V)\) is, for \( u > 0 \) and \( 0 < v < 1 \),

\[
 f_{U,V}(u,v) = f_{X,Y}(g_1^{-1}(u,v), g_2^{-1}(u,v)) |J| \\
= f_{X,Y}(uw, u(1-v)) |J| \\
= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} (uv)^{\alpha_1-1}[u(1-v)]^{\alpha_2-1}e^{-[uv+u(1-v)]/\beta} \\
= \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1+\alpha_2-1}v^{\alpha_1-1}(1-v)^{\alpha_2-1}e^{-u/\beta}.
\]

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,

\[
f_U(u) \overset{u \geq 0}{=} \int_0^1 \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1+\alpha_2-1}v^{\alpha_1-1}(1-v)^{\alpha_2-1}e^{-u/\beta}dv \\
= \frac{u^{\alpha_1+\alpha_2-1}e^{-u/\beta}}{\Gamma(\alpha_1+\alpha_2)} \int_0^1 v^{\alpha_1-1}(1-v)^{\alpha_2-1}dv \\
= \frac{u^{\alpha_1+\alpha_2-1}e^{-u/\beta}}{\Gamma(\alpha_1+\alpha_2)} \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \\
= \frac{1}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1+\alpha_2-1}e^{-u/\beta}I(u > 0),
\]

which we recognize as a gamma pdf with shape parameter \( \alpha_1 + \alpha_2 \) and scale parameter \( \beta \). 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

\[
 M_U(t) \overset{X \perp Y}{=} M_X(t)M_Y(t) = \left( \frac{1}{1-\beta t} \right)^{\alpha_1} \left( \frac{1}{1-\beta t} \right)^{\alpha_2} = \left( \frac{1}{1-\beta t} \right)^{\alpha_1+\alpha_2},
\]

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,

\[
f_V(v) \overset{0 < v < 1}{=} \int_0^\infty \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1+\alpha_2-1}v^{\alpha_1-1}(1-v)^{\alpha_2-1}e^{-u/\beta}du \\
= \frac{v^{\alpha_1-1}(1-v)^{\alpha_2-1}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \int_0^\infty u^{\alpha_1+\alpha_2-1}e^{-u/\beta}du \\
= \frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} v^{\alpha_1-1}(1-v)^{\alpha_2-1}I(0 < v < 1),
\]

which we recognize as a beta pdf with parameters \( \alpha_1 \) and \( \alpha_2 \). That is, \( V = X/(X + Y) \sim \beta(\alpha_1, \alpha_2) \). This completes part (c).
Remark: In addition to deriving the marginal distributions in this example, that is,

\[ U = X + Y \sim \text{gamma}(\alpha_1 + \alpha_2, \beta) \]
\[ V = \frac{X}{X + Y} \sim \text{beta}(\alpha_1, \alpha_2), \]

note that \( U \) and \( V \) are also independent. We know this because we can write the joint pdf

\[
f_{U,V}(u,v) = \frac{1}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} u^{\alpha_1+\alpha_2-1}v^{\alpha_1-1}(1-v)^{\alpha_2-1}e^{-u/\beta}I(u > 0, 0 < v < 1) \]

\[
= \frac{u^{\alpha_1+\alpha_2-1}e^{-u/\beta}I(u > 0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}} \times v^{\alpha_1-1}(1-v)^{\alpha_2-1}I(0 < v < 1). \]

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 \((u,v) \in B\). 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 \( X \perp Y \). Find the distribution of \( X/Y \).

Solution. First, note that the joint pdf of \((X,Y)\) is

\[
f_{X,Y}(x,y) \overset{X \perp Y}{=} f_X(x)f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \frac{1}{\sqrt{2\pi}}e^{-y^2/2} = \frac{1}{2\pi}e^{-(x^2+y^2)/2}, \]

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 $g : A^* \rightarrow B^*$, say, is one-to-one. Consider adding $V = g_2(X,Y) = Y$ so that the transformation is

$$U = g_1(X,Y) = \frac{X}{Y}$$
$$V = g_2(X,Y) = Y.$$

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

$$B^* = \{(u, v) \in \mathbb{R}^2 : u \in \mathbb{R}, \ v \in \mathbb{R} - \{0\}\}.$$ 

The inverse transformation is found by solving $\frac{u}{v} = x, \ v = y$ 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

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(u, v)}{\partial u} & \frac{\partial g_1^{-1}(u, v)}{\partial v} \\ \frac{\partial g_2^{-1}(u, v)}{\partial u} & \frac{\partial g_2^{-1}(u, v)}{\partial v} \end{vmatrix} = \det \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v(1) - u(0) = v,$$

which is nonzero over $B^*$. Therefore, the joint pdf of $(U, V)$, for $(u, v) \in B^*$, is given by

$$f_{U,V}(u, v) = f_{X,Y}(g_1^{-1}(u, v), g_2^{-1}(u, v))|J| = f_{X,Y}(uv, v)|v|$$
$$= \frac{|v|}{2\pi} \exp \frac{-((uv)^2 + v^2)/2}{\sigma^2}$$
$$= \frac{|v|}{2\pi} \exp \frac{-v^2/2}{\sigma^2}.$$

**Remark:** If we let $B = \mathbb{R}^2$, then the joint pdf $f_{U,V}(u, v)$ over $B$ and the one over $B^*$ (as shown above) define the same probability distribution for $(U, V)$ because $P(B \setminus B^*) = 0$. Therefore, in what follows, we can work with $f_{U,V}(u, v)$ defined over $B$ 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

$$f_U(u) = \int_{-\infty}^{\infty} f_{U,V}(u, v) dv$$
$$= \int_{-\infty}^{\infty} \frac{|v|}{2\pi} \exp \frac{-v^2/2}{\sigma^2} \frac{1}{(1+u^2)} dv.$$ 

To do this integral, let $\sigma^2 = 1/(1 + u^2)$ and write

$$f_U(u) = \int_{-\infty}^{\infty} \frac{|v|}{2\pi} \exp \frac{-v^2/2}{\sigma^2} \frac{1}{(1+u^2)} dv = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{|v|}{2\pi} \exp \frac{-v^2/2\sigma^2}{ \frac{1}{\sigma^2}} dv = h(v),$$ say.
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
\[
\int_{-\infty}^{\infty} |v| e^{-v^2/2\sigma^2} dv = 2 \int_{0}^{\infty} ve^{-v^2/2\sigma^2} dv.
\]
Therefore, for \( -\infty < u < \infty \),
\[
f_U(u) = \frac{1}{2\pi} 2 \int_{0}^{\infty} ve^{-v^2/2\sigma^2} dv
= \frac{1}{\pi} \left( -\sigma^2 e^{-v^2/2\sigma^2} \right)_{v=0}^{\infty} = \frac{\sigma^2}{\pi} (1 - 0) = \frac{1}{\pi(1 + u^2)},
\]
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
\[
A_u = \{ x \in \mathbb{R} : g(x) \leq u \}
B_v = \{ y \in \mathbb{R} : h(y) \leq v \}.
\]
The joint cdf of \((U, V)\) is
\[
F_{U,V}(u, v) = P(U \leq u, V \leq v)
= P(X \in A_u, Y \in B_v)
= P(X \in A_u) P(Y \in B_v),
\]
the last step following from Theorem 4.2.10(a). Therefore, the joint pdf of \((U, V)\) is
\[
f_{U,V}(u, v) = \frac{\partial^2}{\partial u \partial v} F_{U,V}(u, v)
= \frac{\partial^2}{\partial u \partial v} P(X \in A_u) P(Y \in B_v)
= \frac{d}{du} P(X \in A_u) \frac{d}{dv} P(Y \in A_v).
\]
By Lemma 4.2.7, \( U \) and \( V \) are independent. \( \square \)
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

$$X|P \sim b(n, P)$$

$$P \sim \text{beta}(\alpha, \beta),$$

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

$$X|N \sim b(N, p)$$

$$N \sim \text{Poisson}(\lambda),$$

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
- $p = \text{pr(“offspring survives”)}$ is the same for each egg.

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

$$f_{X,P}(x, p) = f_{X|P}(x|p)f_P(p)$$

$$= \binom{n}{x} p^x (1 - p)^{n-x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1} (1 - p)^{\beta - 1}$$

$$= \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{x + \alpha - 1} (1 - p)^{n-x + \beta - 1}.$$ 

Therefore, the marginal pmf of $X$, for $x = 0, 1, 2, ..., n$, is given by

$$f_X(x) = \int_0^1 f_{X|P}(x|p)f_P(p)dp$$

$$= \int_0^1 \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{x + \alpha - 1} (1 - p)^{n-x + \beta - 1}dp$$

$$= \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{x + \alpha - 1} (1 - p)^{n-x + \beta - 1}dp$$

$$= \binom{n}{x} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x + \alpha)\Gamma(n - x + \beta)}{\Gamma(\alpha + \beta + n)}.$$ 

This is called the **beta-binomial distribution**. We write $X \sim \text{beta-binomial}(n, \alpha, \beta)$.
Q: If $X \sim \text{beta-binomial}(n, \alpha, \beta)$, what are $E(X)$ and $\text{var}(X)$?

A: From the definition,

$$E(X) = \sum_{x=0}^{n} x f_X(x)$$

$$= \sum_{x=0}^{n} x \left( \frac{n}{x} \right) \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x + \alpha)\Gamma(n - x + \beta)}{\Gamma(\alpha + \beta + n)}.$$ 

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:

- $E(X)$ → refers to the marginal distribution of $X$
- $E(X|Y)$ → refers to the conditional distribution of $X|Y$
- $E[E(X|Y)]$ → calculated using the marginal distribution of $Y$.

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

$$E(X) = \int \int_{\mathbb{R}^2} x f_{X,Y}(x,y) dxdy$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} x f_{X|Y}(x|y)f_Y(y) dxdy$$

$$= \int_{\mathbb{R}} \left[ \int_{\mathbb{R}} x f_{X|Y}(x|y) dx \right] f_Y(y) dy$$

$$= \int_{\mathbb{R}} E(X|Y = y)f_Y(y) dy = E[E(X|Y)].$$

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,

$$X|P \sim b(n, P)$$

$$P \sim \text{beta}(\alpha, \beta).$$

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

\[
X | P \sim b(n, P) \\
P \sim \text{beta}(\alpha, \beta).
\]

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

\[
X | N \sim b(N, p) \\
N \sim \text{Poisson}(\lambda),
\]

Casella and Berger (pp 163) show that, for \( x = 0, 1, 2, \ldots \),

\[
f_X(x) = \sum_{n=x}^{\infty} f_{X|N}(x|n)f_N(n)
\]

\[
= \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \frac{\lambda^n e^{-\lambda}}{n!}
\]

\[
= \frac{(\lambda p)^x e^{-\lambda p}}{x!}.
\]

In this example, the Poisson pmf \( f_N(n) \) is the mixing distribution and, marginally, \( X \sim \text{Poisson}(\lambda p) \). Note that \( \lambda p = E(X) = E[E(X|N)] \).

**Example 4.20.** **Non-central \( \chi^2 \) distribution.** Consider the hierarchy

\[
X | Y \sim \chi_{p+2}^2 \\
Y \sim \text{Poisson}(\lambda),
\]

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

\[
f_X(x) = \sum_{y=0}^{\infty} f_{X,Y}(x, y) = \sum_{y=0}^{\infty} f_{X|Y}(x|y)f_Y(y)
= \sum_{y=0}^{\infty} \frac{1}{\Gamma(\frac{p}{2} + y)2^{\frac{p}{2}+y}} x^{\frac{p}{2}+y-1} e^{-x/2} \left( \frac{\lambda y e^{-\lambda}}{y!} \right).
\]
This is called the non-central $\chi^2$ distribution with $p$ degrees of freedom and non-centrality parameter $\lambda > 0$, written $X \sim \chi^2_p(\lambda)$. The non-central $\chi^2$ distribution can be thought of as a mixture distribution; it is essentially an infinite weighted average of (central) $\chi^2$ densities where the mixing distribution is Poisson($\lambda$). If $\lambda = 0$, the non-central $\chi^2_p(\lambda)$ distribution reduces to our “usual” central $\chi^2_p$ distribution.

Note: To find $E(X)$, we could calculate

$$E(X) = \int_0^\infty xf_X(x)dx$$
$$= \int_0^\infty x \sum_{y=0}^\infty \frac{1}{\Gamma\left(\frac{p}{2} + y\right)2^{y/2+y-1}} x^{\frac{p}{2}+y-1} e^{-x/2} \left(\frac{\lambda^y e^{-\lambda}}{y!}\right) dx.$$  

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

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

Second, note that

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

Combining these two equations completes the proof. $\square$

**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

$$X|Y \sim \chi^2_{p+2Y}$$
$$Y \sim \text{Poisson}(\lambda).$$
Note that
\[E[\text{var}(X|Y)] = E[2(p + 2Y)] = 2p + 4\lambda\]
\[\text{var}[E(X|Y)] = \text{var}(p + 2Y) = 4\lambda.\]

Therefore,
\[\text{var}(X) = E[\text{var}(X|Y)] + \text{var}[E(X|Y)] = 2p + 4\lambda + 4\lambda = 2p + 8\lambda.\]

**Example 4.22.** Find the moment generating function of \(X \sim \chi^2_p(\lambda)\).

*Solution.* We again exploit the hierarchy
\[X|Y \sim \chi^2_{p+2Y}, \quad Y \sim \text{Poisson}(\lambda).\]

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
\[M_X(t) = E\left[\left(\frac{1}{1 - 2t}\right)^{\frac{p}{2} + Y}\right] = \sum_{y=0}^{\infty} \left(\frac{1}{1 - 2t}\right)^{\frac{p}{2} + y} \frac{\lambda^y e^{-\lambda}}{y!} = e^{-\lambda} \sum_{y=0}^{\infty} \left(\frac{\lambda}{1 - 2t}\right)^y = \frac{e^{-\lambda}}{1 - 2t} = \exp\left(\frac{2\lambda t}{1 - 2t}\right).\]

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

$$
E(X) = \mu_X \quad \text{var}(X) = \sigma_X^2 < \infty
$$

$$
E(Y) = \mu_Y \quad \text{var}(Y) = \sigma_Y^2 < \infty
$$

**Definitions:** The covariance of $X$ and $Y$ is

$$
cov(X, Y) \equiv \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)].
$$

The correlation of $X$ and $Y$ is

$$
corr(X, Y) \equiv \rho_{XY} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}.
$$

**Notes:**

1. Both $\sigma_{XY}$ and $\rho_{XY}$ are real numbers. Specifically,

   $$
   -\infty < \sigma_{XY} < \infty
   $$

   and

   $$
   -1 \leq \rho_{XY} \leq 1.
   $$

2. Both $\sigma_{XY}$ and $\rho_{XY}$ 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 $Y$,

$$
cov(X, Y) = E(XY) - E(X)E(Y).
$$

This is called the covariance computing formula.

**Proof.** From the definition,

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

$$
= E(XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y)
$$

$$
= E(XY) - \mu_X \mu_Y - \mu_X \mu_Y + \mu_X \mu_Y
$$

$$
= E(XY) - \mu_X \mu_Y. \quad \square
$$

**Discoveries:** It is easy to establish each of the following results:

1. $cov(X, Y) = cov(Y, X)$
2. $cov(X, X) = var(X)$
3. $cov(a, X) = 0$, for any constant $a \in \mathbb{R}$. 
Theorem 4.5.5. If $X \independent Y$, then $\text{cov}(X, Y) = 0$.

Proof. If $X \independent Y$, then $E(XY) = E(X)E(Y)$ by Theorem 4.2.10(b). □

Remark: The converse of Theorem 4.5.5 is not true in general. That is, $\text{cov}(X, Y) = 0$ $\nRightarrow$ $X \independent Y$.

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 $\xi$, 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:

- $a = b = 1$:
  $$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$$

- $a = 1, b = -1$:
  $$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y) - 2\text{cov}(X, Y)$$

- $X \independent Y, a = 1, b = \pm 1$:
  $$\text{var}(X \pm Y) = \text{var}(X) + \text{var}(Y).$$

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,

$$h(t) = \sigma_X^2t^2 + 2\text{cov}(X, Y)t + \sigma_Y^2,$$

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

$$[2\text{cov}(X, Y)]^2 - 4\sigma_X^2\sigma_Y^2 \leq 0 \iff [\text{cov}(X, Y)]^2 \leq \sigma_X^2\sigma_Y^2 \iff -\sigma_X\sigma_Y \leq \text{cov}(X, Y) \leq \sigma_X\sigma_Y.$$ (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

$$|\rho_{XY}| = 1 \iff |\text{cov}(X,Y)| = \sigma_X\sigma_Y \iff [\text{cov}(X,Y)]^2 = \sigma_X^2\sigma_Y^2 \iff \text{LHS of Equation (4.1)} = 0 \iff h(t) \text{ has a single root of multiplicity 2.}$$

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

$$[(X - \mu_X)t + (Y - \mu_Y)]^2 = 0 \text{ a.s.} \iff (X - \mu_X)t + (Y - \mu_Y) = 0 \text{ a.s.} \iff Y = -tX + \mu_X t + \mu_Y \text{ a.s.}$$

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 ($\rho_{XY} = -1$).

**Bivariate Normal Distribution**

**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:** $(X,Y) \sim \text{mvn}_2(\mu, \Sigma)$, where the mean vector $\mu$ and the variance-covariance matrix $\Sigma$ are

$$\mu = \left( \begin{array}{c} \mu_X \\ \mu_Y \end{array} \right)_{2 \times 1} \quad \text{and} \quad \Sigma = \left( \begin{array}{cc} \sigma_X^2 & \sigma_{XY} \\ \sigma_{YX} & \sigma_Y^2 \end{array} \right)_{2 \times 2},$$

respectively, and where $\sigma_{XY} = \sigma_{YX} = \rho\sigma_X\sigma_Y$. Note that $\Sigma$ 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 $(X,Y) \sim \text{mvn}_2(\mu, \Sigma)$, then the (joint) moment generating function is

$$M_{X,Y}(t) = \exp(t'\mu + t'\Sigma t/2),$$

where $t = (t_1, t_2)'$.  

---

PAGE 115
Facts: Suppose \((X,Y) \sim \text{mvn}_2(\mu, \Sigma)\).

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
   
   \[
   \beta_0 = \mu_Y - \beta_1 \mu_X,
   \beta_1 = \rho \left( \frac{\sigma_Y}{\sigma_X} \right).
   \]

   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 (\(\Rightarrow\)) 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:

\[
\mathbf{X} = (X_1, X_2, \ldots, X_n) \quad \text{← random vector}
\]

\[
\mathbf{x} = (x_1, x_2, \ldots, x_n) \quad \text{← realization of } \mathbf{X}.
\]

Mathematical definition: Suppose \((S, \mathcal{B}, P)\) is a probability space. We call \(\mathbf{X} : S \rightarrow \mathbb{R}^n\) 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 \((n\text{-dimensional})\) Borel sets. One can characterize \(\mathcal{B}(\mathbb{R}^n)\) as the smallest \(\sigma\)-algebra generated by the collection of all half-open hyper-rectangles; i.e.,

\[
\{(x_1, x_2, \ldots, x_n) : -\infty < x_1 \leq a_1, -\infty < x_2 \leq a_2, \ldots, -\infty < x_n \leq a_n, a_i \in \mathbb{R}\}.
\]
The range probability space is \((\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathbb{P}_X)\). We call \(\mathbb{P}_X\) the induced probability measure of \(X\). Similar to the univariate case, there is a one-to-one correspondence between \(\mathbb{P}_X\) and a random vector’s cumulative distribution function (cdf), which is defined as
\[
F_X(x) = \mathbb{P}_X(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n),
\]
for all \(x \in \mathbb{R}^n\). As before, we will eventually start writing \(\mathbb{P}\) for \(\mathbb{P}_X\).

**Definition:** We call a random vector \(X\) discrete if there exists a countable set \(\mathcal{A} \subset \mathbb{R}^n\) such that \(\mathbb{P}_X(X \in \mathcal{A}) = 1\). The joint probability mass function (pmf) of \(X\) is
\[
f_X(x) = \mathbb{P}_X(X_1 = x_1, X_2 = x_2, ..., X_n = x_n).
\]

For any \(B \in \mathcal{B}(\mathbb{R}^n)\),
\[
\mathbb{P}_X(X \in B) = \sum_{x \in B} f_X(x).
\]

**Definition:** The random vector \(X\) is continuous if there exists a function \(f_X : \mathbb{R}^n \rightarrow \mathbb{R}\) such that
\[
\mathbb{P}_X(X \in B) = \int_B f_X(x) \, dx,
\]
for all \(B \in \mathcal{B}(\mathbb{R}^n)\). We call \(f_X(x)\) the joint probability density function (pdf) of \(X\). In the continuous case, the (joint) cdf and the joint pdf are related through
\[
\frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_X(x) = f_X(x),
\]
for all \(x \in \mathbb{R}^n\), provided that this partial derivative exists.

**Mathematical Expectation:** Suppose \(X\) is a random vector and let \(g : \mathbb{R}^n \rightarrow \mathbb{R}\). Then \(g(X)\) is a random variable and its expected value is
\[
E[g(X)] = \sum_{x \in \mathcal{A}} g(x) f_X(x) \quad \text{(discrete case)}
\]
\[
E[g(X)] = \int_{\mathbb{R}^n} g(x) f_X(x) \, dx \quad \text{(continuous case)}.
\]
The usual existence issues arise; we need the sum (integral) above to converge absolutely. Otherwise, \(E[g(X)]\) does not exist.

**Marginal distributions:** Suppose \(X \sim f_X(x)\). If \(X\) is continuous, the marginal pdf of \(X_i\) is given by
\[
f_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} f_X(x) \, dx_{(-i)},
\]
where \(x_{(-i)} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)\). If \(X\) is discrete, the marginal pmf of \(X_i\) is
\[
f_{X_i}(x_i) = \sum_{x_{(-i)} \in \mathcal{A}} f_X(x).
\]
In other words, to find the marginal pdf (pmf) of \(X_i\), we integrate (sum) \(f_X(x)\) 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) \(f_X(x)\) 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) \( f_X(x) \) by the pdf (pmf) of the other variables. For example, suppose \( X = (X_1, X_2, X_3) \sim f_{X_1, X_2, X_3}(x_1, x_2, x_3) \). Then, for example,

\[
\begin{align*}
    f_{X_1 | X_2, X_3}(x_1 | x_2, x_3) &= \frac{f_{X_1, X_2, X_3}(x_1, x_2, x_3)}{f_{X_2, X_3}(x_2, x_3)}, \\
    f_{X_1, X_2 | X_3}(x_1, x_2 | x_3) &= \frac{f_{X_1, X_2, X_3}(x_1, x_2, x_3)}{f_{X_3}(x_3)}.
\end{align*}
\]

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 \( t = (t_1, t_2, ..., t_n)' \). The moment generating function of \( X = (X_1, X_2, ..., X_n)' \) is

\[
M_X(t) = E(e^{t'X}) = \int \cdots \int e^{t_1x_1 + t_2x_2 + \cdots + t_nx_n} dF_X(x).
\]

For \( M_X(t) \) to exist, we need \( E(e^{t'X}) < \infty \) for all \( t \) in an open neighborhood about \( 0 \). That is, \( \exists h_1, h_2, \ldots, h_n > 0 \) such that \( E(e^{t'X}) < \infty \) for all \( t_i \in (-h_i, h_i), i = 1, 2, ..., n \). Otherwise, we say that \( M_X(t) \) 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

\[
M_{X_i}(t_i) = M_X(0, ..., 0, t_i, 0, ..., 0),
\]

where \( t_i \) is in the \( i \)th position. The (joint) marginal mgf of \( (X_i, X_j)' \) is found by

\[
M_{X_i, X_j}(t_i, t_j) = M_X(0, ..., 0, t_i, 0, ..., 0, t_j, 0, ..., 0).
\]

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:

<table>
<thead>
<tr>
<th>Trial outcome</th>
<th>Category 1</th>
<th>Probability ( p_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \uparrow )</td>
<td>Category 2</td>
<td>Probability ( p_2 )</td>
</tr>
<tr>
<td>( \rightarrow )</td>
<td>Category 3</td>
<td>Probability ( p_3 )</td>
</tr>
<tr>
<td>( \downarrow )</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>:</td>
<td>Category ( n )</td>
<td>Probability ( p_n )</td>
</tr>
</tbody>
</table>

The probabilities \( p_1, p_2, \ldots, p_n \) do not change from trial to trial and \( \sum_{i=1}^{n} p_i = 1 \). Define the random variable

\[ X_i = \text{number of outcomes in Category } i \text{ (out of } m \text{ trials)}. \]

We call \( X = (X_1, X_2, \ldots, X_n)' \) a multinomial random vector. The joint pmf of \( X \) is

\[ f_X(x) = \frac{m!}{x_1!x_2! \cdots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}, \]

for values of \( x \in \mathcal{A} \), where

\[ \mathcal{A} = \left\{ (x_1, x_2, \ldots, x_n) : x_i = 0, 1, 2, \ldots, m; \sum_{i=1}^{n} x_i = m \right\}. \]

We write \( X \sim \text{mult}(m, p; \sum_{i=1}^{n} p_i = 1) \). The parameter \( p = (p_1, p_2, \ldots, p_n) \) 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 \( p_1, p_2, \ldots, p_n \),

\[ (p_1 + p_2 + \cdots + p_n)^m = \sum_{x \in \mathcal{A}} \frac{m!}{x_1!x_2! \cdots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n}. \]

This is a generalization of the binomial theorem. Clearly, the multinomial pmf sums to one.

**MGF:** The mgf of \( X \sim \text{mult}(m, p; \sum_{i=1}^{n} p_i = 1) \) is

\[ M_X(t) = (p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_n e^{t_n})^m, \]

where \( t = (t_1, t_2, \ldots, t_n)' \).

**Proof.** The mgf is

\[ M_X(t) = E(e^{t'X}) = E(e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}) \]

\[ = \sum_{x \in \mathcal{A}} e^{t_1 x_1 + t_2 x_2 + \cdots + t_n x_n} \frac{m!}{x_1!x_2! \cdots x_n!} p_1^{x_1} p_2^{x_2} \cdots p_n^{x_n} \]

\[ = \sum_{x \in \mathcal{A}} \frac{m!}{x_1!x_2! \cdots x_n!} (p_1 e^{t_1})^{x_1} (p_2 e^{t_2})^{x_2} \cdots (p_n e^{t_n})^{x_n}, \]

which is the multinomial expansion of \( (p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_n e^{t_n})^m \). \( \square \)
Result: If \( X \sim \text{mult}(m, p; \sum_{i=1}^{n} p_i = 1) \), then \( X_i \sim b(m, p_i) \), \( i = 1, 2, \ldots, n \). That is, the category counts \( X_1, X_2, \ldots, X_n \) have marginal binomial distributions.

Proof. The mgf of \( X_i \) is

\[
M_{X_i}(t_i) = M_X(0, \ldots, 0, t_i, 0, \ldots, 0) = (q_i + p_i e^{t_i})^m,
\]

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

\[
E(X_i) = mp_i \quad \text{var}(X_i) = mp_i(1 - p_i).
\]

Result: If \( X \sim \text{mult}(m, p; \sum_{i=1}^{n} p_i = 1) \), then \((X_i, X_j)' \sim \text{trinomial}(m, p_i, p_j, 1 - p_i - p_j)\).

Trinomial Framework

<table>
<thead>
<tr>
<th>Trial outcome</th>
<th>Category ( i )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Category ( j )</td>
<td>( p_i )</td>
</tr>
<tr>
<td></td>
<td>Neither</td>
<td>( p_j )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( 1 - p_i - p_j )</td>
</tr>
</tbody>
</table>

Proof. The mgf of \((X_i, X_j)'\) is

\[
M_{X_i,X_j}(t_i, t_j) = M_X(0, \ldots, 0, t_i, 0, \ldots, 0, t_j, 0, \ldots, 0) = (q_{ij} + p_i e^{t_i} + p_j e^{t_j})^m,
\]

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,

\[
E(X_iX_j) = \frac{\partial^2}{\partial t_i \partial t_j} M_{X_i,X_j}(t_i, t_j) \bigg|_{t_i=t_j=0} = \frac{\partial^2}{\partial t_i \partial t_j} (q_{ij} + p_i e^{t_i} + p_j e^{t_j})^m \bigg|_{t_i=t_j=0} = m(m-1)p_ip_j.
\]

Therefore, for \( i \neq j \),

\[
\text{cov}(X_i, X_j) = E(X_iX_j) - E(X_i)E(X_j) = m(m-1)p_ip_j - (mp_i)(mp_j) = -mp_ip_j.
\]

Summary: The mean and variance-covariance matrix of \( X \sim \text{mult}(m, p; \sum_{i=1}^{n} p_i = 1) \) are

\[
\mu = E(X) = \begin{pmatrix} mp_1 \\ mp_2 \\ \vdots \\ mp_n \end{pmatrix} = mp
\]
and

\[
\Sigma = \text{cov}(X) = \begin{pmatrix}
mp_1(1-p_1) & -mp_1p_2 & \cdots & -mp_1p_n \\
-mp_2p_1 & mp_2(1-p_2) & \cdots & -mp_2p_n \\
\vdots & \vdots & \ddots & \vdots \\
-mp_np_1 & -mp_np_2 & \cdots & mp_n(1-p_n)
\end{pmatrix}
\]

\[
= m[\text{diag}(p) - pp'],
\]

respectively. Note that \(\Sigma\) 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 \(X_1, X_2, \ldots, X_n\) are random variables. Let \(f_{X_i}(x_i)\) denote the marginal pdf (pmf) of \(X_i\). The random variables \(X_1, X_2, \ldots, X_n\) are **mutually independent** if

\[
f_X(x) = \prod_{i=1}^n f_{X_i}(x_i)
= f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n),
\]

for all \(x \in \mathbb{R}^n\); i.e., the joint pdf (pmf) \(f_X(x)\) 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 \(X_1, X_2, \ldots, X_n\) are mutually independent if and only if

\[
F_X(x) = \prod_{i=1}^n F_{X_i}(x_i)
= F_{X_1}(x_1)F_{X_2}(x_2)\cdots F_{X_n}(x_n),
\]

for all \(x \in \mathbb{R}^n\); i.e., the joint cdf \(F_X(x)\) factors into the product of the marginal cdfs.

**Result:** The random variables \(X_1, X_2, \ldots, X_n\) are mutually independent if and only if

\[
M_X(t) = \prod_{i=1}^n M_{X_i}(t_i)
= M_{X_1}(t_1)M_{X_2}(t_2)\cdots M_{X_n}(t_n),
\]

for all \(t_i \in \mathbb{R}\) where these mgfs exist; that is, the joint mgf \(M_X(t)\) factors into the product of the marginal mgfs.

**Theorem 4.6.6.** Suppose \(X_1, X_2, \ldots, X_n\) are mutually independent. Suppose \(g_1, g_2, \ldots, g_n\) are real functions; i.e., \(g_i : \mathbb{R} \to \mathbb{R}\), where \(g_i\) is a function of \(x_i\) only, \(i = 1, 2, \ldots, n\). 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 $X_1, X_2, \ldots, X_n$ are mutually independent random variables. Suppose the marginal mgf of $X_i$ is $M_{X_i}(t)$, for $i = 1, 2, \ldots, n$. 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 $X_1, X_2, \ldots, X_n$ 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 $X_1, X_2, \ldots, X_n$ 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 $X_1, X_2, \ldots, X_n$ are iid Poisson($\lambda$), where $\lambda > 0$. The mgf of the sum

$$Z = X_1 + X_2 + \cdots + X_n$$

is given by

$$M_Z(t) = [M_X(t)]^n = [e^{\lambda(1-e^{-t})}]^n = e^{n\lambda(1-e^{-t})},$$

which we recognize as the mgf of a Poisson distribution with mean $n\lambda$. Because mgfs are unique, we know that $Z \sim \text{Poisson}(n\lambda)$.

**Example 4.24.** Suppose $X_1, X_2, \ldots, X_n$ are mutually independent, where $X_i \sim \text{gamma}(\alpha_i, \beta)$, 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

$$M_Z(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

$$= \prod_{i=1}^{n} \left( \frac{1}{1-\beta t} \right)^{\alpha_i} \quad \text{for } t < \frac{1}{\beta}$$

$$= \left( \frac{1}{1-\beta t} \right)^{\sum_{i=1}^{n} \alpha_i},$$

which we recognize as the mgf of a gamma distribution with shape parameter $\sum_{i=1}^{n} \alpha_i$ and scale parameter $\beta$. Because mgfs are unique, we know that $Z \sim \text{gamma}(\sum_{i=1}^{n} \alpha_i, \beta)$. 
Remark: The result in Example 4.24 has important special cases:

1. \( \alpha_i = 1, \) for \( i = 1, 2, \ldots, n. \) In this case, \( X_1, X_2, \ldots, X_n \) are iid exponential(\( \beta \)) and

\[
Z = \sum_{i=1}^{n} X_i \sim \text{gamma}(n, \beta).
\]

2. \( \alpha_i = \beta / p_i, \) where \( p_i > 0, \) and \( \beta = 2. \) In this case, \( X_1, X_2, \ldots, X_n \) are mutually independent, \( X_i \sim \chi^2_{p_i}, \) and

\[
Z = \sum_{i=1}^{n} X_i \sim \text{gamma} \left( \frac{n}{p}, \frac{\beta}{2} \right) \equiv \chi^2_{p},
\]

where \( p = \sum_{i=1}^{n} p_i; \) i.e., “the degrees of freedom add.”

Example 4.25. Suppose \( X_1, X_2, \ldots, X_n \) are mutually independent, where \( X_i \sim \mathcal{N}(\mu_i, \sigma^2_i), \) for \( i = 1, 2, \ldots, n. \) 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

\[
M_Z(t) = E(e^{tZ}) = E[e^{a_1 t X_1 + a_2 t X_2 + \cdots + a_n t X_n}]
\]

\[
= E(e^{\sum_{i=1}^{n} a_i t X_i}) = \prod_{i=1}^{n} M_{X_i}(a_i t)
\]

\[
= \prod_{i=1}^{n} \exp \left( \mu_i a_i t + \frac{(a_i t)^2 \sigma_i^2}{2} \right)
\]

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. \( \square \)

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 \( X_1, X_2, \ldots, X_n \) are random variables with (marginal) means \( E(X_i) \) and (marginal) variances \( \text{var}(X_i), \) for \( i = 1, 2, \ldots, n. \) 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
\[
E(Z) = E(a_1X_1 + a_2X_2 + \cdots + a_nX_n)
= a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) = \sum_{i=1}^{n} a_i E(X_i).
\]

The variance of $Z$ is
\[
\text{var}(Z) = \text{var}(a_1X_1 + a_2X_2 + \cdots + a_nX_n)
= \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + \sum_{i \neq j} a_i a_j \text{cov}(X_i, X_j)
= \sum_{i=1}^{n} a_i^2 \text{var}(X_i) + 2 \sum_{i<j} a_i a_j \text{cov}(X_i, X_j).
\]

**Theorem 4.6.11.** Suppose $X = (X_1, X_2, ..., X_n)$ is an $n$-dimensional random vector with joint pdf (pmf) $f_X(x)$. 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
\[
f_X(x) = g_1(x_1)g_2(x_2) \cdots g_n(x_n),
\]
for all $x \in \mathbb{R}^n$. 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 $X = (X_1, X_2, ..., X_n)$ is a continuous random vector with joint pdf $f_X(x)$ and support $\mathcal{A} \subseteq \mathbb{R}^n$. Define
\[
U_1 = g_1(X_1, X_2, ..., X_n)
U_2 = g_2(X_1, X_2, ..., X_n)
\vdots
U_n = g_n(X_1, X_2, ..., X_n).
\]
Assume that this is a one-to-one transformation from $\mathcal{A} = \{x \in \mathbb{R}^n : f_X(x) > 0\}$ to
\[
\mathcal{B} = \{u \in \mathbb{R}^n : u_i = g_i(x), \ i = 1, 2, ..., n; \ x \in \mathcal{A}\},
\]
the support of $U = (U_1, U_2, ..., U_n)$. Because the transformation is one-to-one (by assumption), the inverse transformation is
\[
x_1 = g_1^{-1}(u_1, u_2, ..., u_n)
x_2 = g_2^{-1}(u_1, u_2, ..., u_n)
\vdots
x_n = g_n^{-1}(u_1, u_2, ..., u_n).
\]
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 \( B \), 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 \( \mathcal{A} = \{ x \in \mathbb{R}^3 : 0 < x_1 < x_2 < x_3 < 1 \} \), the upper orthant of the unit cube in \( \mathbb{R}^3 \). Define

\[
U_1 = g_1(X_1, X_2, X_3) = \frac{X_1}{X_2} \\
U_2 = g_2(X_1, X_2, X_3) = \frac{X_2}{X_3} \\
U_3 = g_3(X_1, X_2, X_3) = X_3.
\]

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

\[
\begin{align*}
x_1 &= g_1^{-1}(u_1, u_2, u_3) = u_1u_2u_3 \\
x_2 &= g_2^{-1}(u_1, u_2, u_3) = u_2u_3 \\
x_3 &= g_3^{-1}(u_1, u_2, u_3) = u_3
\end{align*}
\]

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 \( B \). Therefore, the joint pdf of \( U = (U_1, U_2, U_3) \), for \( u \in B \), is given by
\[
f_U(u_1, u_2, u_3) = f_X(g_1^{-1}(u), g_2^{-1}(u), g_3^{-1}(u)) |J|
\]
\[
= 48(u_1u_2u_3)(u_2u_3)(u_3)
\]
\[
= 48u_1u_2^3u_3^5.
\]

Notice that we can write
\[
f_U(u_1, u_2, u_3) = 48u_1u_2^3u_3^5 I(0 < u_1 < 1, 0 < u_2 < 1, 0 < u_3 < 1)
\]
\[
= 2u_1I(0 < u_1 < 1) 4u_2^3I(0 < u_2 < 1) 6u_3^5I(0 < u_3 < 1)
\]
\[
= f_{U_1}(u_1)f_{U_2}(u_2)f_{U_3}(u_3).
\]

We see that

- \( U_1 \sim \text{beta}(2, 1) \), \( U_2 \sim \text{beta}(4, 1) \), and \( U_3 \sim \text{beta}(6, 1) \)
- \( U_1, U_2, \) 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**.
• From the covariance inequality, it follows immediately that $-1 \leq \rho_{XY} \leq 1$.

• Establishing this result is much easier using Cauchy-Schwarz 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 $\xi$ is between $x$ and $\mu$ (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,

$$E[g(X)] \geq E[g(\mu) + g'(\mu)(X - \mu)] = g(\mu) + g'(\mu)E(X - \mu) = g(E(X)). \quad \square$$

**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 $X_1, X_2, \ldots, X_n$ are called a random sample from the population $f_X(x)$ if

1. $X_1, X_2, \ldots, X_n$ are mutually independent
2. The marginal pdf (pmf) of each $X_i$ is the same function $f_X(x)$.

Alternatively, we say that

"$X_1, X_2, \ldots, X_n$ are iid from $f_X(x)$.”

The acronym “iid” is short for “independent and identically distributed.” 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 $X_1, X_2, \ldots, X_n$.

Result: If $X_1, X_2, \ldots, X_n$ are iid from $f_X(x)$, the joint pdf (pmf) of $X = (X_1, X_2, \ldots, X_n)$ is

$$f_X(x) = \prod_{i=1}^{n} f_X(x_i) = f_X(x_1)f_X(x_2) \cdots f_X(x_n).$$

This follows immediately from Definition 4.6.5 (CB, pp 182).

Example 5.1. Suppose that $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Here, the $N(\mu, \sigma^2)$ population distribution is

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

This is the (marginal) pdf of each $X_i$. The joint pdf of $X = (X_1, X_2, \ldots, X_n)$ is

$$f_X(x) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}I(x_i \in \mathbb{R})$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2}I(x \in \mathbb{R}^n).$$

Remark: This function, when viewed as a function of $x_1, x_2, \ldots, x_n$, describes probabilistically the (joint) random behavior of $X_1, X_2, \ldots, X_n$. Later on (when we start thinking about
estimation), we will begin to regard \( f_X(x) \) not as a function of \( x_1, x_2, \ldots, x_n \) but instead as a function of \( \theta = (\mu, \sigma^2)' \) 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 \( X = (X_1, X_2, \ldots, X_n) \) are greatly simplified. Suppose that in Example 5.1 we wanted to calculate the probability that each \( X_i \) exceeded the mean \( \mu \) by two standard deviations; i.e.,

\[
P(X_1 > \mu + 2\sigma, X_2 > \mu + 2\sigma, \ldots, X_n > \mu + 2\sigma).
\]

From first principles, this probability could be found by taking \( f_X(x) \) and integrating it over the set \( B = \{ x \in \mathbb{R}^n : x_1 > \mu + 2\sigma, x_2 > \mu + 2\sigma, \ldots, x_n > \mu + 2\sigma \} \), that is, by calculating

\[
\int_{\mu+2\sigma}^{\infty} \cdots \int_{\mu+2\sigma}^{\infty} \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^n e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2} dx_1 dx_2 \cdots dx_n.
\]

This is a tedious calculation to make directly using the joint distribution \( f_X(x) \). However, note that we can write

\[
P(X_1 > \mu + 2\sigma, X_2 > \mu + 2\sigma, \ldots, X_n > \mu + 2\sigma)
= P(X_1 > \mu + 2\sigma) P(X_2 > \mu + 2\sigma) \cdots P(X_n > \mu + 2\sigma) \quad \text{(mutually independent)}
= [P(X_1 > \mu + 2\sigma)]^n, \quad \text{(identically distributed)}
\]

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 \( X_1, X_2, \ldots, X_n \).

**Remark:** When we say “random sample,” we essentially mean that we are sampling from an infinite population. When sampling from a finite population, say, \( \{x_1, x_2, \ldots, x_N\} \), 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 \( x_1, x_2, \ldots, x_N \). The \( X_i \)'s are also mutually independent because the process of choosing each \( x_i \) is the same. In other words, \( X_1, X_2, \ldots, X_n \) 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

$$P(X_1 = x_1) = \frac{1}{N}$$

$$P(X_2 = x_1 | X_1 = x_1) = 0.$$ 

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 $X_1, X_2, \ldots, X_n$ as iid.

5.2 Sums of Random Variables from a Random Sample

Definition: Suppose $X_1, X_2, \ldots, X_n$ is a random sample. A statistic $T$ is a function of $X_1, X_2, \ldots, X_n$, that is,

$$T = T(X) = T(X_1, X_2, \ldots, X_n).$$

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: $T(X) = \bar{X}$, where

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

2. Sample variance: $T(X) = S^2$, where

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

3. Minimum order statistic: $T(X) = X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$

4. Sample range: $T(X) = X_{(n)} - X_{(1)}$.

Note: The definition of a statistic is very broad. For example, even something like

$$T(X) = \ln(S^2 + 4) - 12.08e^{-\tan(\sum_{i=1}^{n} |X_i|)}$$

satisfies the definition. In addition, if $\mu$ and $\sigma^2$ are unknown, then

• $\bar{X}$ is a statistic, but $\bar{X} - \mu$ is not.

• $S^2$ is a statistic, but $S^2/\sigma^2$ is not.
Definition: Suppose \( X_1, X_2, \ldots, X_n \) is an iid sample from \( f_X(x) \). Suppose that \( T = T(X) \) is a statistic. The probability distribution of \( T \) is called its sampling distribution.

Revelation: Because \( T = T(X) \), a function of \( X_1, X_2, \ldots, X_n \), 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,

\[
X_1, X_2, \ldots, X_n \sim f_X(x) \quad \leftarrow \text{population distribution}
\]

\[
T = T(X_1, X_2, \ldots, X_n) \sim f_T(t) \quad \leftarrow \text{sampling distribution of } T
\]

Common goals: For a statistic \( T = T(X) \), 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 \( X_1, X_2, \ldots, X_n \) are iid with \( E(X) = \mu \) and \( \text{var}(X) = \sigma^2 < \infty \). Then

\[
E \left( \sum_{i=1}^{n} X_i \right) = nE(X_1) = n\mu
\]

\[
\text{var} \left( \sum_{i=1}^{n} X_i \right) = n\text{var}(X_1) = n\sigma^2.
\]

Proof. Exercise. Compare this result with Lemma 5.2.5 (CB, pp 213), which is slightly more general.

Theorem 5.2.6. Suppose \( X_1, X_2, \ldots, X_n \) 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

\[
E(\overline{X}) = E \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n} E \left( \sum_{i=1}^{n} X_i \right) = \frac{1}{n} (n\mu) = \mu
\]

\[
\text{var}(\overline{X}) = \text{var} \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \text{var} \left( \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}.
\]

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,
\[
E(S^2) = E \left[ \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = \frac{1}{n-1} \left[ E \left( \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right) \right] = \frac{1}{n-1} \left[ E \left( \sum_{i=1}^{n} X_i^2 \right) - nE(\bar{X}^2) \right].
\]

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
\[
E(\bar{X}^2) = \text{var}(\bar{X}) + [E(\bar{X})]^2 = \frac{\sigma^2}{n} + \mu^2.
\]
Therefore,
\[
E(S^2) = \frac{1}{n-1} \left[ n(\sigma^2 + \mu^2) - n \left( \frac{\sigma^2}{n} + \mu^2 \right) \right] = \sigma^2.
\]

**Curiosity:** How would we find 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 \(\bar{X}\) is
\[
M_{\bar{X}}(t) = [M_X(t/n)]^n.
\]

**Proof.** The mgf of \(\bar{X}\) is
\[
M_{\bar{X}}(t) = E(e^{t\bar{X}}) = E[e^{\frac{t}{n}(X_1 + X_2 + \cdots + X_n)}] = E(e^{\frac{t}{n}X_1}e^{\frac{t}{n}X_2} \cdots e^{\frac{t}{n}X_n})
\]
\[
\text{indep} = E(e^{\frac{t}{n}X_1})E(e^{\frac{t}{n}X_2}) \cdots E(e^{\frac{t}{n}X_n})
\]
\[
= M_{X_1}(t/n)M_{X_2}(t/n) \cdots M_{X_n}(t/n)
\]
\[
\text{ident} = [M_X(t/n)]^n.
\]

**Remark:** This result is very useful. It allows us to quickly obtain the mgf of \(\bar{X}\) (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
\[
M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}},
\]
for all \(t \in \mathbb{R}\). Therefore,
\[
M_{\bar{X}}(t) = [e^{\mu (t/n) + \frac{\sigma^2 (t/n)^2}{2}}]^n
\]
\[
= e^{\mu n t + \frac{\sigma^2 n^2 t^2}{2}},
\]
which we recognize as the mgf of a \(\mathcal{N}(\mu, \sigma^2/n)\) distribution. Because mgfs are unique, we know that \(\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)\).

**Example 5.3.** Suppose that \(X_1, X_2, ..., X_n\) are iid \(\text{gamma}(\alpha, \beta)\). Recall that the (population) mgf of \(X \sim \text{gamma}(\alpha, \beta)\) is
\[
M_X(t) = \left( \frac{1}{1 - \beta t} \right)^{\alpha},
\]
\[
\text{for all } t < \frac{1}{\beta}.
\]
for \( t < 1/\beta \). Therefore,

\[
M_X(t) = \left\{ \left[ \frac{1}{1 - \beta(t/n)} \right]^n \right\}
\]

\[
= \left[ \frac{1}{1 - (\beta/n)t} \right]^{n\alpha},
\]

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 \( X \sim \text{gamma}(n\alpha, \beta/n) \).

**Special case:** \( X_1, X_2, \ldots, X_n \sim \text{iid exponential}(\beta) \Rightarrow \overline{X} \sim \text{gamma}(n\alpha, \beta/n) \).

**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

\[
f_Z(z) = \int_{-\infty}^{\infty} f_X(w)f_Y(z-w)dw.
\]

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:

\[
\begin{align*}
w &= g_1(x, y) = x \\
z &= g_2(x, y) = x + y \\
\Rightarrow & \quad x = g_1^{-1}(w, z) = w \\
y = g_2^{-1}(w, z) = z - w
\end{align*}
\]

The Jacobian of the (inverse) transformation is

\[
J = \text{det} \begin{vmatrix} \frac{\partial g_1^{-1}(w, z)}{\partial w} & \frac{\partial g_1^{-1}(w, z)}{\partial z} \\ \frac{\partial g_2^{-1}(w, z)}{\partial w} & \frac{\partial g_2^{-1}(w, z)}{\partial z} \end{vmatrix} = \text{det} \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1.
\]

Therefore, the joint pdf of \( (W, Z) \) is given by

\[
f_{W,Z}(w, z) = f_{X,Y}(w, z - w)^{X \perp Y} f_X(w)f_Y(z-w).
\]

Finally,

\[
f_Z(z) = \int_{-\infty}^{\infty} f_X(w)f_Y(z-w)dw
\]

as claimed. \( \square \)

**Example 5.4.** Suppose that \( X \sim \text{U}(0, 1) \), \( Y \sim \text{U}(0, 1) \), and \( X \perp Y \). Find the pdf of \( Z = X + Y \).

**Solution.** First, note that the support of \( Z \) is \( Z = \{ z : 0 < z < 2 \} \). 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

\[
f_Z(z) = \int_{-\infty}^{\infty} f_X(w)f_Y(z-w)dw
\]

\[
= \int_{\infty}^{1} I(0 < w < 1)I(0 < z - w < 1)dw
\]

\[
= \int_{0}^{1} I(0 < z - w < 1)dw.
\]
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

$$f_Z(z) = \begin{cases} 
  z, & 0 < z \leq 1 \\
  2 - z, & 1 < z < 2 \\
  0, & \text{otherwise}.
\end{cases}$$

This pdf is shown in Figure 5.1.

**Example 5.5.** Suppose that $X \sim \text{Cauchy}(0, \sigma_X)$, $Y \sim \text{Cauchy}(0, \sigma_Y)$, and $X \perp Y$. Find the pdf of $Z = X + Y$.

**Solution.** First, note that the support of $Z$ is $Z = \{z : -\infty < z < \infty\}$. The marginal pdfs of $X$ and $Y$ are

$$f_X(x) = \frac{1}{\pi \sigma_X [1 + (\frac{x}{\sigma_X})^2]} I(x \in \mathbb{R})$$

$$f_Y(y) = \frac{1}{\pi \sigma_Y [1 + (\frac{y}{\sigma_Y})^2]} I(y \in \mathbb{R}).$$
Using the convolution formula, the pdf of Z, for all \( z \in \mathbb{R} \), is

\[
f_Z(z) = \int_{\mathbb{R}} f_X(w)f_Y(z-w)dw
= \int_{\mathbb{R}} \frac{1}{\pi \sigma_X[1 + (\frac{w}{\sigma_X})^2]} \frac{1}{\pi \sigma_Y[1 + (\frac{z-w}{\sigma_Y})^2]} \, dw
= \frac{1}{\pi(\sigma_X + \sigma_Y)[1 + (\frac{z}{\sigma_X + \sigma_Y})^2]},
\]

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 \( X_1, X_2, \ldots, X_n \) are iid Cauchy(\( \mu, \sigma \)), where \(-\infty < \mu < \infty \) and \( \sigma > 0 \). What is the sampling distribution of \( \bar{X} \)?

**Solution.** The easiest way to answer this would be to use characteristic functions. The characteristic function of \( X \sim \text{Cauchy}(\mu, \sigma) \) is

\[
\psi_X(t) = E(e^{itX}) = e^{\mu it} - \sigma |t|,
\]

where \( i = \sqrt{-1} \). Therefore,

\[
\psi_{\bar{X}}(t) = [\psi_X(t/n)]^n = [e^{\mu it/n} - \sigma |t/n|]^n = e^{\mu it - \sigma |t|},
\]

which we recognize as the characteristic function of the Cauchy(\( \mu, \sigma \)) distribution. Therefore,

\[
X_1, X_2, \ldots, X_n \sim \text{iid Cauchy}(\mu, \sigma) \implies \bar{X} \sim \text{Cauchy}(\mu, \sigma).
\]

**Q:** Can we show this without using characteristic functions?

**A:** Yes, this could be done in three steps.

1. First, argue that

\[
Z_1, Z_2, \ldots, Z_n \sim \text{iid Cauchy}(0, 1) \implies \sum_{i=1}^{n} Z_i \sim \text{Cauchy}(0, n).
\]

This could be done using convolution for \( n = 2 \). Then use induction.

2. Next, argue that

\[
f_{\bar{Z}}(z) = nf_{\sum_{i=1}^{n} Z_i}(nz)
= \frac{1}{\pi(1 + z^2)} I(z \in \mathbb{R}),
\]

i.e., \( \bar{Z} \sim \text{Cauchy}(0, 1) \). 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 \( i = 1, 2, \ldots, n \), so that \( \bar{X} = \sigma \bar{Z} + \mu \) (a location-scale transformation). Therefore,

\[
f_{\bar{X}}(x) = \frac{1}{\sigma} f_{\bar{Z}}\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\pi \sigma \left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]} I(x \in \mathbb{R}),
\]

i.e., \( \bar{X} \sim \text{Cauchy}(\mu, \sigma) \).
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.,

$$ f_X(x | \theta) = h(x)c(\theta)\exp\left\{ \sum_{i=1}^{k} w_i(\theta)t_i(x) \right\}. $$

Define the statistics

- $T_1 = T_1(X) = \sum_{j=1}^{n} t_1(X_j)$
- $T_2 = T_2(X) = \sum_{j=1}^{n} t_2(X_j)$

and set $T = (T_1, T_2, ..., T_k)$, a $k$-dimensional statistic. If $f_X(x | \theta)$ is a full exponential family; i.e., if $d = \dim(\theta) = k$,

then

$$ f_T(t | \theta) = H(t)[c(\theta)]^n\exp\left\{ \sum_{i=1}^{k} w_i(\theta)t_i \right\}. $$

That is, $T$ has pdf (pmf) in the exponential family as well.

**Example 5.6.** Suppose $X_1, X_2, ..., X_n$ are iid gamma($\alpha, \beta$); i.e., the population pdf is

$$ f_X(x | \theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-x/\beta}I(x > 0) = \frac{I(x > 0)}{x} \frac{1}{\Gamma(\alpha)\beta^\alpha} \exp\left( \alpha \ln x - \frac{x}{\beta} \right) = h(x)c(\theta)\exp\{w_1(\theta)t_1(x) + w_2(\theta)t_2(x)\}, $$

where $\theta = (\alpha, \beta)'$, $h(x) = I(x > 0)/x$, $c(\theta) = [\Gamma(\alpha)\beta^\alpha]^{-1}$, $w_1(\theta) = \alpha$, $t_1(x) = \ln x$, $w_2(\theta) = -1/\beta$, and $t_2(x) = x$. Note that this is a full exponential family with $d = k = 2$. Theorem 5.2.11 says that $T = (T_1, T_2)$, where

- $T_1 = T_1(X) = \sum_{j=1}^{n} \ln X_j$
- $T_2 = T_2(X) = \sum_{j=1}^{n} X_j$

has a (joint) pdf that also falls in the exponential family; i.e., $f_T(t | \theta)$ can be written in the form

$$ f_T(t | \theta) = H(t)[c(\theta)]^n\exp\left\{ \sum_{i=1}^{2} w_i(\theta)t_i \right\}. $$
5.3 Sampling from the Normal Distribution

Remark: This section is dedicated to results that arise when $X_1, X_2, \ldots, X_n$ are normally distributed.

Theorem 5.3.1. Suppose that $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. Let $\bar{X}$ and $S^2$ denote the sample mean and the sample variance, respectively. Then

(a) $\bar{X} \perp \perp S^2$

(b) $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ ← 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

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

$$= \frac{1}{n-1} \sum_{i=2}^{n} (X_i - \bar{X})^2 + \frac{1}{n-1} (X_1 - \bar{X})^2.$$ 

Note also that

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0 \implies X_1 - \bar{X} = -\sum_{i=2}^{n} (X_i - \bar{X}).$$

Therefore, we can rewrite $S^2$ as

$$S^2 = \frac{1}{n-1} \sum_{i=2}^{n} (X_i - \bar{X})^2 + \frac{1}{n-1} \left[ -\sum_{i=2}^{n} (X_i - \bar{X}) \right]^2$$

$$= g(X_2 - \bar{X}, X_3 - \bar{X}, \ldots, X_n - \bar{X}), \text{ say.}$$

Because functions of independent random variables (vectors) are independent, it suffices to show that $\bar{X} \perp \perp (X_2 - \bar{X}, X_3 - \bar{X}, \ldots, X_n - \bar{X})$. Consider the $n$-variate transformation

$$y_1 = \bar{X} \quad \quad x_1 = y_1 - \sum_{i=2}^{n} y_i$$

$$y_2 = x_2 - \bar{X} \quad \quad x_2 = y_1 + y_2$$

$$y_3 = x_3 - \bar{X} \quad \quad x_3 = y_1 + y_3$$

$$\vdots$$

$$y_n = x_n - \bar{X} \quad \quad x_n = y_1 + y_n.$$ 

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 $Y = (Y_1, Y_2, ..., Y_n)$ is given by

$$f_Y(y) = f_X(y_1 - \sum_{i=2}^{n} y_i, y_1 + y_2, y_1 + y_3, ..., y_1 + y_n) \cdot |n|. $$

Going forward, we assume (without loss of generality) that $X_1, X_2, ..., X_n$ are iid $N(\mu, \sigma^2)$, with $\mu = 0$ and $\sigma^2 = 1$. Under this assumption,

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

$$= \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} x_i^2} I(x \in \mathbb{R}^n).$$

This simplifies the calculations and is not prohibitive because the $N(\mu, \sigma^2)$ family is a location-scale family (see CB, pp 216-217). From above, we therefore have, for all $y \in \mathbb{R}^n$,

$$f_Y(y) = \frac{n}{(2\pi)^{n/2}} \exp \left\{ \frac{1}{2} \left[ \left( y_1 - \sum_{i=2}^{n} y_i \right)^2 + \sum_{i=2}^{n} (y_1 + y_i)^2 \right] \right\}$$

$$= \frac{n}{(2\pi)^{n/2}} e^{-ny_1^2/2} \times \exp \left\{ \frac{1}{2} \left[ \sum_{i=2}^{n} y_i^2 + \left( \sum_{i=2}^{n} y_i \right)^2 \right] \right\}.$$  

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 $N(\mu, \sigma^2)$, $\mu$ and $\sigma^2$ arbitrary; the calculations would have just been far messier. We have proven part (a). To prove part (c), we first recall the following:

$$X_i \sim N(\mu, \sigma^2) \implies Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1) \implies Z_i^2 \sim \chi_1^2.$$  

Therefore, the random variables $Z_1^2, Z_2^2, ..., Z_n^2$ are iid $\chi_1^2$. 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

$$W_1 = \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^{n} \left( \frac{X_i - \overline{X} + \overline{X} - \mu}{\sigma} \right)^2$$

$$= \sum_{i=1}^{n} \left( \frac{X_i - \overline{X}}{\sigma} \right)^2 + 2 \sum_{i=1}^{n} \left( \frac{X_i - \overline{X}}{\sigma} \right) \left( \frac{\overline{X} - \mu}{\sigma} \right) + \sum_{i=1}^{n} \left( \frac{\overline{X} - \mu}{\sigma} \right)^2.$$  

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
\[
W_1 = \sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 + \frac{n}{\sigma^2} \left( \frac{\bar{X} - \mu}{\sigma} \right)^2.
\]

Now,
\[
W_3 = n \left( \frac{\bar{X} - \mu}{\sigma} \right)^2 = \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi^2_1,
\]
because \(\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)\), and
\[
W_2 = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{(n-1)S^2}{\sigma^2}.
\]
Furthermore, we know that \(W_2 \perp W_3\) because \(\bar{X} \perp S^2\) and functions of independent random variables are independent. The mgf of \(W_1 \sim \chi^2_n\) is, for \(t < 1/2\),
\[
\left( \frac{1}{1-2t} \right)^{n/2} = M_{W_1}(t) = E(e^{tW_1}) = E[e^{t(W_2+W_3)}] = E(e^{tW_2}e^{tW_3}) \overset{W_2 \perp W_3}{=} E(e^{tW_2})E(e^{tW_3}) = M_{W_2}(t)M_{W_3}(t) = M_{W_2}(t) \left( \frac{1}{1-2t} \right)^{1/2},
\]
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 \(\chi^2_{n-1}\) random variable. Because mgfs are unique,
\[
W_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.
\]

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 \(\mathcal{N}(\mu, \sigma^2)\),
\[
\sum_{i=1}^{n} \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_n
\]
and
\[
\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2_{n-1}.
\]
In the second result, one can interpret having to “estimate” \(\mu\) with \(\bar{X}\) as being responsible for “losing” a degree of freedom.
Remark: When \(X_1, X_2, \ldots, X_n\) 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:”

\[
E \left[ \frac{(n - 1)S^2}{\sigma^2} \right] = n - 1 \\
\text{var} \left[ \frac{(n - 1)S^2}{\sigma^2} \right] = 2(n - 1).
\]

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 \(X_1, X_2, \ldots, X_n\) are iid \(\mathcal{N}(\mu, \sigma^2)\).

**General result:** Suppose that \(X_1, X_2, \ldots, X_n\) 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.

**Remark:** The results in Lemma 5.5.3 (CB, pp 220) might be better suited for STAT 714, where independence issues are treated more generally. I will present a special case here.

**Lemma 5.5.3** (special case). Suppose that \(X_1, X_2, \ldots, X_n\) are independent random variables with \(X_j \sim \mathcal{N}(\mu_j, \sigma^2_j)\), for \(j = 1, 2, \ldots, n\). Define the linear combinations

\[
U = \sum_{j=1}^{n} a_j X_j = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n
\]

\[
V = \sum_{j=1}^{n} b_j X_j = b_1 X_1 + b_2 X_2 + \cdots + b_n X_n,
\]

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 \(U = a_1 X_1 + a_2 X_2\) and \(V = b_1 X_1 + b_2 X_2\), 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 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 $j$,

$$\text{cov}(U, V) = \sum_{j=1}^{n} a_j b_j = a^\prime b,$$

where $a = (a_1, a_2, ..., a_n)^\prime$ and $b = (b_1, b_2, ..., b_n)^\prime$. Therefore $U \perp V$ if and only if $a$ and $b$ 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 $U \perp V$. 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

$$f_T(t) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{p\pi} \Gamma\left(\frac{p}{2}\right)} \left(1 + \frac{t^2}{p}\right)^{(p+1)/2} I(t \in \mathbb{R}).$$

**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

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

The quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

where $S$ denotes the sample standard deviation of $X_1, X_2, ..., X_n$. To see why, note that

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\sigma}{S} \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sqrt{(n-1)S^2/\sigma^2} \sim \frac{\bar{X} - \mu}{\sqrt{(n-1)(\bar{X}^2 - \frac{n}{n-1})}} \sim \frac{\bar{X} - \mu}{\sqrt{(n-1)(\bar{X}^2 - \frac{n}{n-1})}} \sim \frac{\bar{X} - \mu}{\sqrt{(n-1)}}$$

Because $\bar{X} \perp S^2$, 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 $U \perp V$. The joint pdf of $(U, V)$ is

$$f_{U,V}(u, v) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \sqrt{\frac{1}{\Gamma\left(\frac{p}{2}\right)}} 2^{p/2} V^{p/2-1} e^{-v/2},$$

where $\mathcal{N}(0, 1)$ pdf and $\chi^2_p$ pdf.
for $-\infty < u < \infty$ and $v > 0$. Consider the bivariate transformation

$$T = g_1(U, V) = \frac{U}{\sqrt{V/p}}$$

$$W = g_2(U, V) = V.$$ 

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

$$u = g^{-1}_1(t, w) = t\sqrt{w/p}$$

$$v = g^{-1}_2(t, w) = w.$$ 

The Jacobian of the (inverse) transformation is

$$J = \det \begin{vmatrix} \frac{\partial g^{-1}_1(t, w)}{\partial t} & \frac{\partial g^{-1}_1(t, w)}{\partial w} \\ \frac{\partial g^{-1}_2(t, w)}{\partial t} & \frac{\partial g^{-1}_2(t, w)}{\partial w} \end{vmatrix} = \det \begin{vmatrix} \sqrt{w/p} & \frac{1}{\sqrt{p}} \frac{1}{2} w^{-1/2} \\ 0 & 1 \end{vmatrix} = \sqrt{\frac{w}{p}},$$

which never vanishes over $\mathcal{B}$. For $(t, w) \in \mathcal{B}$, the joint pdf of $(T, W)$ is

$$f_{T,W}(t, w) = f_{U,V}(g^{-1}_1(t, w), g^{-1}_2(t, w))|J|$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(t\sqrt{w/p})^2}{2}} \frac{1}{\Gamma\left(\frac{p}{2}\right)2^{p/2}} w^{p/2-1} e^{-w/2} \left| \sqrt{\frac{w}{p}} \right|$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(t\sqrt{w/p})^2}{2}} \frac{1}{\Gamma\left(\frac{p}{2}\right)2^{p/2}} w^{p+1/2-1} e^{-w/2}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{(t\sqrt{w/p})^2}{2}} \frac{1}{\Gamma\left(\frac{p}{2}\right)2^{p/2}} w^{p+1/2-1} e^{-w(1+\frac{2}{p})/2}.$$ 

Therefore, the marginal pdf of $T$ is

$$f_T(t) = \int_0^\infty f_{T,W}(t, w)dw$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{p}} \Gamma\left(\frac{p}{2}\right)2^{p/2} \int_0^\infty w^{p+1/2-1} e^{-w(1+\frac{2}{p})/2} \left.\frac{\text{gamma}(a,b) \text{ kernel}}{\text{gamma}(a,b)}\right. dw,$$

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 $t \in \mathbb{R}$,

$$f_T(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{p}} \Gamma\left(\frac{p}{2}\right)2^{p/2} \Gamma\left(\frac{p + 1}{2}\right) \left[2 \left(1 + \frac{t^2}{p}\right)^{-1}\right]^{(p+1)/2}$$

$$= \frac{\Gamma\left(\frac{p+1}{2}\right)}{\sqrt{p\pi} \Gamma\left(\frac{p}{2}\right) \left(1 + \frac{t^2}{p}\right)^{(p+1)/2}},$$

as claimed. □
Moments: If $T \sim t_p$, then

$$E(T) = 0, \quad \text{if } p > 1$$

$$\text{var}(T) = \frac{p}{p-2}, \quad \text{if } p > 2.$$  

To show that $E(T) = 0$ when $p > 1$, suppose $U \sim \mathcal{N}(0,1)$, $V \sim \chi^2_p$, and $U \perp V$. Write

$$E(T) = E\left(\frac{U}{\sqrt{V/p}}\right) = E(U)E\left(\frac{1}{\sqrt{V/p}}\right) = 0,$$

because $E(U) = 0$. We need to investigate the second expectation to see why the “$p > 1$” condition is needed. Recall that $V \sim \chi^2_d = \gamma(p/2, 2)$. Therefore,

$$E\left(\frac{1}{\sqrt{V/p}}\right) = \sqrt{p}E\left(\frac{1}{\sqrt{V}}\right) = \sqrt{p} \int_0^\infty \frac{1}{\Gamma(p/2)2^{p/2}} v^{p/2-1}e^{-v/2} dv$$

$$= \frac{\sqrt{p}}{\Gamma(p/2)2^{p/2}} \int_0^\infty v^{p/2-1}e^{-v/2} dv$$

$$= \frac{\sqrt{p}}{\Gamma(p/2)2^{p/2}} \Gamma\left(\frac{p-1}{2}\right) 2^{(p-1)/2}$$

$$= \frac{\sqrt{p}\Gamma\left(\frac{p-1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{p}{2}\right)},$$

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 $U \perp V$. The random variable

$$W = \frac{U}{V/q} \sim F_{p,q},$$

an $F$ distribution with (numerator) $p$ and (denominator) $q$ degrees of freedom. The pdf of $W$ is

$$f_W(w) = \frac{\Gamma(p+q/2)}{\Gamma(p/2)\Gamma(q/2)} \left(\frac{p}{q}\right)^{p/2} \frac{w^{p-1}}{[1 + (\frac{p}{q})w]^{(p+q)/2}} I(w > 0).$$

**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**

$$E(W) = \frac{q}{q-2}, \quad \text{if } q > 2$$

$$\text{var}(W) = 2 \left(\frac{q}{q-2}\right)^2 \frac{p + q - 2}{pq(q - 4)}, \quad \text{if } q > 4.$$  

**Proof.** Exercise.
**Application:** Suppose we have independent random samples

\[ X_1, X_2, ..., X_n \sim \text{iid } \mathcal{N}(\mu_X, \sigma_X^2) \]

\[ Y_1, Y_2, ..., Y_m \sim \text{iid } \mathcal{N}(\mu_Y, \sigma_Y^2). \]

We know

\[ \frac{(n - 1)S_X^2}{\sigma_X^2} \sim \chi^2_{n - 1} \quad \text{and} \quad \frac{(m - 1)S_Y^2}{\sigma_Y^2} \sim \chi^2_{m - 1}. \]

Also, these quantities are independent (because the samples are; therefore, \( S_X^2 \perp \perp S_Y^2 \)). Therefore,

\[ W = \frac{(n - 1)S_X^2}{\sigma_X^2} / (n - 1) = \left( \frac{S_X^2}{S_Y^2} \right) \frac{\sigma_Y^2}{\sigma_X^2} \sim F_{n - 1, m - 1}. \]

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

\[ Y = \frac{(\frac{p}{q})X}{1 + (\frac{p}{q})X} \sim \text{beta } \left( \frac{p}{2}, \frac{q}{2} \right). \]

### 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.,

\[ X_{(1)} = \min_{1 \leq i \leq n} X_i \]
\[ X_{(2)} = \text{second smallest } X_i \]
\[ \vdots \]
\[ X_{(n)} = \max_{1 \leq i \leq n} X_i. \]

**Remark:** Many statistics seen in practice are order statistics or functions of order statistics. For example,

- \( T(X) = X_{(\frac{n+1}{2})} \), the sample median (\( n \) odd)
• \( T(X) = X_{(n)} - X_{(1)} \), the sample range
• \( T(X) = \frac{1}{2}(X_{(1)} + X_{(n)}) \), the sample midrange.

Revelation: Because order statistics are statistics (i.e., they are functions of \( X_1, X_2, \ldots, X_n \)), 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 \( X_1, X_2, \ldots, X_n \) 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 \( P_0 = 0 \), \( P_1 = p_1 \), \( P_2 = p_1 + p_2 \), \ldots, \( P_i = p_1 + p_2 + \cdots + p_i \). 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

\[
P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} [P_i^k (1 - P_i)^{n-k} - P_{i-1}^k (1 - P_{i-1})^{n-k}].
\]

Proof. Let \( Y \) denote the number of \( X_j \)'s (among \( X_1, X_2, \ldots, X_n \)) 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,

\[
P(X_{(j)} \leq x_i) = P(Y \geq j) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1 - P_i)^{n-k}.
\]

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. \( \square \)
Example 5.7. Suppose $X_1, X_2, \ldots, X_n$ are iid Bernoulli($\theta$), 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,

$$E(X_{(j)}) = P(X_{(j)} = 1) = 1 - P(X_{(j)} = 0) = 1 - \sum_{k=j}^{n} \binom{n}{k} (1 - \theta)^k \theta^{n-k}.$$

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 $X_1, X_2, \ldots, X_{10}$ are iid Poisson with mean $\lambda = 2.2$. Calculate $E(X_{(j)})$ for $j = 1, 2, \ldots, 10$. Compare each with $E(X_j) = 2.2$.

Theorem 5.4.4. Suppose $X_1, X_2, \ldots, X_n$ 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 $j = 1, 2, \ldots, n$, 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

$$f_{X_{(1)}}(x) = n f_X(x) [1 - F_X(x)]^{n-1}.$$

- $j = n$. The pdf of $X_{(n)}$, the maximum order statistic, collapses to

$$f_{X_{(n)}}(x) = n f_X(x) [F_X(x)]^{n-1}.$$

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 (among $X_1, X_2, \ldots, X_n$) 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,

\[
F_{X_{(j)}}(x) = P(X_{(j)} \leq x) = P(Y \geq j)
\]

\[
= \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}.
\]

The pdf of $X_{(j)}$ is given by

\[
f_{X_{(j)}}(x) = \frac{d}{dx} F_{X_{(j)}}(x)
\]

\[
= \frac{d}{dx} \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}
\]

\[
= \sum_{k=j}^{n} \binom{n}{k} \frac{d}{dx} [F_X(x)]^k [1 - F_X(x)]^{n-k}
\]

\[
= \sum_{k=j}^{n} \binom{n}{k} \left\{ k[F_X(x)]^{k-1} f_X(x)[1 - F_X(x)]^{n-k} - [F_X(x)]^k(n-k)[1 - F_X(x)]^{n-k-1} f_X(x) \right\}
\]

\[
= \binom{n}{j} j[F_X(x)]^{j-1} f_X(x)[1 - F_X(x)]^{n-j} + a - b,
\]

where

\[
a = \sum_{k=j+1}^{n} \binom{n}{k} k[F_X(x)]^{k-1} f_X(x)[1 - F_X(x)]^{n-k}
\]

\[
b = \sum_{k=j}^{n} \binom{n}{k} [F_X(x)]^k(n-k)[1 - F_X(x)]^{n-k-1} f_X(x).
\]

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

\[
a = \sum_{k=j}^{n-1} \binom{n}{k+1} (k+1)[F_X(x)]^k f_X(x)[1 - F_X(x)]^{n-k-1}
\]

\[
b = \sum_{k=j}^{n-1} \binom{n}{k} (n-k)[F_X(x)]^k f_X(x)[1 - F_X(x)]^{n-k-1}.
\]

To establish that $a - b = 0$, simply note that

\[
\binom{n}{k+1} (k+1) = \frac{n!}{(k+1)!(n-k-1)!} (k+1)
\]

\[
= \frac{n!}{k!(n-k-1)!}
\]

\[
= \frac{n!}{k!(n-k)!} (n-k) = \binom{n}{k} (n-k).
\]
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 \(X_1, X_2, \ldots, X_n\) as a “trial,” and consider the trinomial distribution with the following categories:

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Cell probability</th>
<th># Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Less than (x)</td>
<td>(p_1 = P(X &lt; x) = F_X(x))</td>
<td>(j - 1)</td>
</tr>
<tr>
<td>2</td>
<td>Equal to (x)</td>
<td>(p_2 = P(X = x) = \text{“}f_X(x)\text{“})</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Greater than (x)</td>
<td>(p_3 = P(X &gt; x) = 1 - F_X(x))</td>
<td>(n - j)</td>
</tr>
</tbody>
</table>

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

\[
   \frac{n!}{(j-1)!(n-j)!} = \frac{n!}{(j-1)!!(n-j)!!} 
\]

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 \(X_1, X_2, \ldots, X_n\) are iid \(U(0,1)\). Find the pdf of \(X_{(j)}\), the \(j\)th order statistic.

**Solution.** Recall that the population pdf and cdf are
\[
   f_X(x) = I(0 < x < 1) \quad \text{and} \quad F_X(x) = \begin{cases} 
   0, & x \leq 0 \\
   x, & 0 < x < 1 \\
   1, & x \geq 1. 
\end{cases} 
\]

The pdf of \(X_{(j)}\) is, for \(0 < x < 1, \)
\[
   f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} x^{j-1}(1-x)^{n-j} 
   = \frac{\Gamma(n+1)}{\Gamma(j)\Gamma(n-j+1)} x^{j-1}(1-x)^{(n-j+1)-1}, 
\]
i.e., \(X_{(j)} \sim \text{beta}(j,n-j+1)\).

**Example 5.9.** Suppose that \(X_1, X_2, \ldots, X_n\) are iid exponential(\(\beta\)), where \(\beta > 0\). Recall that the exponential pdf and cdf are
\[
   f_X(x) = \frac{1}{\beta} e^{-x/\beta} I(x > 0) \quad \text{and} \quad F_X(x) = \begin{cases} 
   0, & x \leq 0 \\
   1 - e^{-x/\beta}, & x > 0. 
\end{cases} 
\]

The pdf of \(X_{(1)}\) is
\[
   f_{X_{(1)}}(x) = \frac{n}{\beta} e^{-x/\beta} [1 - F_X(x)]^{n-1} 
   = \frac{n}{\beta} e^{-nx/\beta} I(x > 0), 
\]
Figure 5.2: Order statistic distributions. $X_1, X_2, ..., X_{10} \sim \text{iid exponential}(\beta = 2)$. Left: Pdf of $X_{(1)}$. Right: Pdf of $X_{(10)}$. Note that the horizontal axes are different in the two figures.

that is, $X_{(1)} \sim \text{exponential}(\beta/n)$. The pdf of $X_{(n)}$ is

$$fx_{(n)}(x) = nf_X(x)[F_X(x)]^{n-1}$$

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

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

$$f_{X_{(i)},X_{(j)}}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!}[F_X(u)]^{i-1}f_X(u)[F_X(v) - F_X(u)]^{j-1-i}$$

$$\times f_X(v)[1 - F_X(v)]^{n-j},$$

for $-\infty < u < v < \infty$.

**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:

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Cell probability</th>
<th># Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Less than $u$</td>
<td>$p_1 = F_X(u)$</td>
<td>$i - 1$</td>
</tr>
<tr>
<td>2</td>
<td>Equal to $u$</td>
<td>$p_2 = &quot;f_X(u)&quot;$</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Between $u$ and $v$</td>
<td>$p_3 = F_X(v) - F_X(u)$</td>
<td>$j - 1 - i$</td>
</tr>
<tr>
<td>4</td>
<td>Equal to $v$</td>
<td>$p_4 = &quot;f_X(v)&quot;$</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Greater than $v$</td>
<td>$p_5 = 1 - F_X(v)$</td>
<td>$n - j$</td>
</tr>
</tbody>
</table>
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 \(X_1, X_2, \ldots, X_n\) are iid \(U(0, 1)\). 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
\[
f_{X_1, X_n}(u, v) = n(n - 1) f_X(u) [F_X(v) - F_X(u)]^{n-2} f_X(v),
\]
for \(-\infty < u < v < \infty\).

Now consider the bivariate transformation
\[
R = g_1(X_1, X_n) = X_n - X_1, \\
S = g_2(X_1, X_n) = X_n.
\]

Note that the support of \((R, S)\) is \(B = \{(r, s) : 0 < r < s < 1\}\). The transformation above is one-to-one, so the inverse transformation exists and is given by
\[
x_1 = g_1^{-1}(r, s) = s - r, \\
x_n = g_2^{-1}(r, s) = s.
\]

The Jacobian of the (inverse) transformation is
\[
J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(r, s)}{\partial r} & \frac{\partial g_1^{-1}(r, s)}{\partial s} \\ \frac{\partial g_2^{-1}(r, s)}{\partial r} & \frac{\partial g_2^{-1}(r, s)}{\partial s} \end{vmatrix} = \det \begin{vmatrix} -1 & 1 \\ 0 & 1 \end{vmatrix} = -1.
\]

Therefore, the joint pdf of \((R, S)\) is, for all \((r, s) \in B\),
\[
f_{R, S}(r, s) = f_{X_1, X_n}(g_1^{-1}(r, s), g_2^{-1}(r, s)) |J| = n(n - 1)[s - (s - r)]^{n-2} = n(n - 1)r^{n-2}I(0 < r < s < 1).
\]

The marginal pdf of \(R\) is therefore
\[
f_R(r) = \int_r^1 n(n - 1)r^{n-2}ds = n(n - 1)r^{n-2}(1 - r)I(0 < r < 1) = \frac{\Gamma(n + 1)}{\Gamma(n - 1)\Gamma(2)} r^{(n-1)-1}(1 - r)^{2-1}I(0 < r < 1),
\]
a beta pdf with parameters \(n - 1\) and 2, that is, \(R \sim \text{beta}(n - 1, 2)\).

**Result:** Suppose \(X_1, X_2, \ldots, X_n\) is an iid sample from a **continuous** distribution with pdf \(f_X(x)\). The joint distribution of the \(n\) order statistics is
\[
f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = n! f_X(x_1)f_X(x_2) \cdots f_X(x_n),
\]
for \(-\infty < x_1 < x_2 < \cdots < x_n < \infty\).
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 \( X_1, X_2, \ldots, X_n \) are iid exponential(\( \theta \)), 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 \( X_1, X_2, \ldots, X_n \) 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 \( \forall \epsilon > 0 \ \exists n_0(\epsilon) \geq 1 \ \exists |x_n - x| < \epsilon \ \forall n \geq n_0(\epsilon) \). 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 $X_1, X_2, \ldots$ converges in probability to a random variable $X$ and write $X_n \xrightarrow{p} X$ if for all $\epsilon > 0$,

$$
\lim_{n \to \infty} P(|X_n - X| \geq \epsilon) = 0,
$$

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

$$
\{|X_n - X| \geq \epsilon\} = \{\text{"X_n stays away from X"}\}
$$

gets small as $n$ gets large.

- In most statistical applications, the limiting random variable $X$ is a constant.

**Example 5.12.** Suppose $X_1, X_2, \ldots, X_n$ are iid exponential($\theta$), where $\theta > 0$. Show that $X_n \xrightarrow{p} \theta$, as $n \to \infty$.

**Solution.** Suppose $\epsilon > 0$. Recalling that $\overline{X}_n \sim \text{gamma}(n, \theta/n)$, it would suffice to show that

$$
P(|\overline{X}_n - \theta| < \epsilon) = P(-\epsilon < \overline{X}_n - \theta < \epsilon) = P(\theta - \epsilon < \overline{X}_n < \theta + \epsilon) = \int_{\theta-\epsilon}^{\theta+\epsilon} \frac{1}{\Gamma(n)(\frac{\theta}{n})^n} x^{n-1} e^{-nx/\theta} dx \to 1, \quad \text{as } n \to \infty.
$$

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,

$$
P(|\overline{X}_n - \theta| \geq \epsilon) = P((\overline{X}_n - \theta)^2 \geq \epsilon^2) \leq \frac{E[(\overline{X}_n - \theta)^2]}{\epsilon^2} = \frac{\text{var}(\overline{X}_n)}{\epsilon^2} = \frac{\theta^2}{n\epsilon^2} \to 0,
$$

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 $X_1, X_2, \ldots, X_n$ is an iid sequence of random variables with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. Let

$$
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
$$

denote the sample mean. Then $\bar{X}_n \xrightarrow{p} \mu$, as $n \to \infty$.

Proof. Suppose $\epsilon > 0$. By Markov’s Inequality,

$$
P(|\bar{X}_n - \mu| \geq \epsilon) = P((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{E((\bar{X}_n - \mu)^2)}{\epsilon^2} = \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.
$$

Remark: In the version of the WLLN stated in Theorem 5.5.2, we assumed finite variances; i.e., that $E(X_i^2) < \infty$. 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 $\bar{X}_n \xrightarrow{p} \mu$, 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., $E(X_i^4) < \infty$. However, $S^2 \xrightarrow{p} \sigma^2$ under the weaker assumption of $E(X_i^2) < \infty$.

Remark: When the limiting random variable is a constant, convergence in probability is sometimes referred to as “consistency” (or “weak consistency”). We might say, “$\bar{X}_n$ is a consistent estimator of $\mu$” and “$S^2$ is a consistent estimator of $\sigma^2$.”

Example 5.13. Suppose $X_1, X_2, \ldots, X_n$ are iid with continuous cdf $F_X$. Let

$$
\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)
$$

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,

$$
\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \xrightarrow{p} E[I(X_1 \leq x)] = P(X_1 \leq x) = F_X(x).
$$

That is, $\hat{F}_n(x) \xrightarrow{p} F_X(x)$ at each fixed $x \in \mathbb{R}$. In a sense, we can think of the edf $\hat{F}_n(x)$ as an “estimate” of the population cdf $F_X(x)$. Figure 5.3 depicts the edf $\hat{F}_n(x)$ calculated when $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(0,1)$ for different values of $n$. 

PAGE 153
Continuity: Suppose \( X_n \overset{p}{\to} X \) and let \( h : \mathbb{R} \to \mathbb{R} \) be a continuous function. Then \( h(X_n) \overset{p}{\to} h(X) \). In other words, convergence in probability is preserved under continuous mappings.

Proof. Suppose \( X_n \overset{p}{\to} X \). Suppose \( \epsilon > 0 \). Because \( h \) is continuous, \( \exists \delta(\epsilon) > 0 \) such that \( |x_n - x| < \delta(\epsilon) \implies |h(x_n) - h(x)| < \epsilon \) (this is the definition of continuity). Define the events

\[
A = \{ x : |x_n - x| < \delta(\epsilon) \} \\
B = \{ x : |h(x_n) - h(x)| < \epsilon \}
\]

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 \( X_n \overset{p}{\to} X \), 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. \( \Box \)
Remark: In Example 5.12, we showed that $X_n \xrightarrow{p} \theta$. This means that

$$X_n^2 \xrightarrow{p} \theta^2, \quad e^{X_n} \xrightarrow{p} e^\theta, \quad \text{and} \quad \sin X_n \xrightarrow{p} \sin \theta.$$  

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

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

denote the sample proportion. Because $E(X_1) = p < \infty$, it follows from the WLLN that $\hat{p} \xrightarrow{p} p$, as $n \to \infty$. By continuity,

$$\ln \left( \frac{\hat{p}}{1 - \hat{p}} \right) \xrightarrow{p} \ln \left( \frac{p}{1 - p} \right).$$

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) $X_nY_n \xrightarrow{p} XY$
(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,

$$P(|X_n + Y_n - (X + Y)| \geq \epsilon) \leq P(|X_n - X| + |Y_n - Y| \geq \epsilon) \leq P(|X_n - X| \geq \epsilon/2) + P(|Y_n - Y| \geq \epsilon/2),$$

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. $\square$
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 \( U(0, \theta) \), where \( \theta > 0 \). Show that \( X_{(n)} \xrightarrow{p} \theta \).

Solution. Recall that the \( U(0, \theta) \) pdf and cdf are
\[
\begin{align*}
    f_X(x) &= \frac{1}{\theta} I(0 < x < \theta) \quad \text{and} \quad F_X(x) = \begin{cases} 
    0, & x \leq 0 \\
    \frac{x}{\theta}, & 0 < x < \theta \\
    1, & x \geq \theta.
\end{cases}
\end{align*}
\]

The cdf of \( X_{(n)} \) is
\[
F_{X_{(n)}}(x) = P(X_{(n)} \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) = \prod P(X_i \leq x) = [F_X(x)]^n, \quad \text{for } 0 < x < \theta.
\]

Therefore,
\[
F_{X_{(n)}}(x) = \begin{cases} 
    0, & x \leq 0 \\
    \left(\frac{x}{\theta}\right)^n, & 0 < x < \theta \\
    1, & x \geq \theta.
\end{cases}
\]

Suppose \( \epsilon > 0 \). By direct calculation, we have
\[
P(|X_{(n)} - \theta| < \epsilon) = P(-\epsilon < X_{(n)} - \theta < \epsilon) = P(\theta - \epsilon < X_{(n)} < \theta + \epsilon) = F_{X_{(n)}}(\theta + \epsilon) - F_{X_{(n)}}(\theta - \epsilon) = 1 - \left(\frac{\theta - \epsilon}{\theta}\right)^n \to 1.
\]

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
\[
E[(X_n - c)^2] = \text{var}(X_n) + [E(X_n) - c]^2 = \text{var}(X_n) + \text{Bias}(X_n)^2.
\]

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

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

Show that $S_n^2 \overset{p}{\to} \sigma^2$, 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

$$S_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \left(\frac{n-1}{n}\right) \frac{1}{n-1} \sum_{i=1}^{n} (X_i - X_n)^2 = \left(\frac{n-1}{n}\right) S,$$

where $S$ is the “usual” sample variance (with denominator $n-1$). Therefore,

$$E(S_n^2) = E\left(\left(\frac{n-1}{n}\right) S^2\right) = \left(\frac{n-1}{n}\right) E(S^2) = \left(\frac{n-1}{n}\right) \sigma^2$$

$$\text{var}(S_n^2) = \text{var}\left(\left(\frac{n-1}{n}\right) S^2\right) = \left(\frac{n-1}{n}\right)^2 \text{var}(S^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2} \to 0.$$ 

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 $X_n \overset{p}{\to} X$ and $c_n \to c$, as $n \to \infty$. Then $c_n X_n \overset{p}{\to} c X$.

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

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right).$$

From the WLLN, we have

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \overset{p}{\to} E(X_1^2).$$

Also from the WLLN, we have $\bar{X}_n \overset{p}{\to} E(X_1)$, so $\bar{X}_n^2 \overset{p}{\to} [E(X_1)]^2$, 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

$$S^2 = c_n \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right) \overset{p}{\to} 1 \times \{E(X_1^2) - [E(X_1)]^2\} = \text{var}(X_1) = \sigma^2.$$ 

This shows that $S^2 \overset{p}{\to} \sigma^2$ (i.e., $S^2$ is a consistent estimator of $\sigma^2$) 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 \( M_{X_n}(t) \xrightarrow{} M_X(t) \) where the limiting random variable \( X \) had a distribution that was degenerate at the constant \( \beta \). 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 \( X_1, X_2, \ldots, \) **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., \( X_n(\omega) \nrightarrow X(\omega) \), 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 \( X_1, X_2, \ldots, X_n \) is an iid sequence of random variables with \( E(X_1) = \mu \) and \( \text{var}(X_1) = \sigma^2 < \infty \). Let
\[
\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i
\]
denote the sample mean. Then \( \bar{X}_n \xrightarrow{a.s.} \mu \), as \( n \to \infty \).
5.5.3 Convergence in distribution

**Definition:** We say that a sequence of random variables $X_1, X_2, \ldots$, 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 $Y_1, Y_2, \ldots, Y_n$ 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

$$F_{Y_{(n)}}(y) = P(Y_{(n)} \leq y) \overset{\text{id}}{=} [P(Y_1 \leq y)]^n = [F_Y(y)]^n = (1 - e^{-y})^n, \text{ for } y > 0.$$
The cdf of \( X_n = Y(n) - \ln n \) is

\[
F_{X_n}(x) = P(X_n \leq x) = P(Y(n) \leq x + \ln n) = F_Y(x + \ln n) = \left[1 - e^{-(x+\ln n)}\right]^n = \left(1 - \frac{e^{-x}}{n}\right)^n \to \exp(-e^{-x}) = F_X(x),
\]

as \( n \to \infty \). This is the cdf of a (standard) Gumbel random variable. Note that \( F_{X_n}(x) \) 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, \( X_n^2 \xrightarrow{d} \chi_1^2 \). Recall that \( X \sim \mathcal{N}(0,1) \implies X^2 \sim \chi_1^2 \).

**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 \( X_n \xrightarrow{d} X \iff \psi_{X_n}(t) \to \psi_X(t) \), 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,

\[
X_n \xrightarrow{d} X \iff M_{X_n}(t) \to M_X(t), \quad \text{for all } |t| < h \ (\exists h > 0),
\]

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 $X_1, X_2, \ldots$ is an iid sequence of random variables with $E(X_1) = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. Then

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } n \to \infty.$$ 

Remark: Note that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}}.$$ 

Showing this is simple algebra. In more applied courses, it is common to write things like

$$\overline{X}_n \sim \mathcal{N}(\mu, \sigma^2/n) \quad \text{or} \quad \sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$$

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 $X_1, X_2, \ldots, X_n$ are iid $\chi^2_1$ so that $E(X_1) = \mu = 1$ and $\text{var}(X_1) = \sigma^2 = 2$. The CLT says that

$$Z_n = \frac{\overline{X}_n - 1}{\sqrt{2/n}} = \frac{\sum_{i=1}^{n} X_i - n}{\sqrt{2n}} \xrightarrow{d} \mathcal{N}(0, 1), \text{ as } n \to \infty.$$ 

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 $i = 1, 2, \ldots, n$, 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

\[
Z_n = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} = \frac{1}{\sqrt{n}} \sum^n_{i=1} Y_i.
\]

Therefore,

\[
M_{Z_n}(t) = E(e^{tZ_n}) = E\left(e^{\frac{1}{\sqrt{n}} \sum^n_{i=1} Y_i}\right) = E\left(e^{\frac{1}{\sqrt{n}} Y_1} e^{\frac{1}{\sqrt{n}} Y_2} \cdots e^{\frac{1}{\sqrt{n}} Y_n}\right) \text{ indep} = E(e^{\frac{1}{\sqrt{n}} Y_1}) E(e^{\frac{1}{\sqrt{n}} Y_2}) \cdots E(e^{\frac{1}{\sqrt{n}} Y_n}) \text{ ident} = [E(e^{\frac{1}{\sqrt{n}} Y_1})]^n = [M_Y(t/\sqrt{n})]^n.
\]

Now write \( M_Y(t/\sqrt{n}) \) in its McLaurin series expansion:

\[
M_Y(t/\sqrt{n}) = \sum_{k=0}^{\infty} M_Y^{(k)}(0) \frac{\left(\frac{t}{\sqrt{n}} - 0\right)^k}{k!},
\]

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,

\[
M_Y^{(0)}(0) = M_Y(0) = 1
\]

\[
M_Y^{(1)}(0) = E(Y) = 0
\]

\[
M_Y^{(2)}(0) = E(Y^2) = 1.
\]

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

\[
M_{Z_n}(t) = [M_Y(t/\sqrt{n})]^n = \left[ 1 + \frac{t^2/2}{n} + R_Y(t/\sqrt{n}) \right]^n = \left[ 1 + \frac{b}{n} + \frac{g(n)}{n} \right]^n,
\]

\[
\text{PAGE 162}
\]
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:

\[
\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1) \iff \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0,\sigma^2).
\]

Statements like the second statement are commonly seen in asymptotic results. We can interpret this as follows. If we

1. center \( \bar{X}_n \) by subtracting \( \mu \)
2. scale \( \bar{X}_n - \mu \) up by multiplying by \( \sqrt{n}, \)

then \( \sqrt{n}(\bar{X}_n - \mu) \) converges to a bonafide distribution. The sequence \( n^{p}(\bar{X}_n - \mu) \) collapses if \( p < 1/2 \) and blows up if \( p > 1/2. \) The value \( p = 1/2 \) is “just right” to ensure \( n^{p}(\bar{X}_n - \mu) \) converges to a nondegenerate distribution with no probability “escaping off” to \( \pm \infty. \)

**Example 5.20.** Suppose \( X_1, X_2, \ldots, X_n \) 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

\[
\sqrt{n}(\bar{X}_n - p) \xrightarrow{d} N(0,p(1 - p)),
\]

as \( n \to \infty. \) For the Bernoulli population distribution, the \( X_i \)'s are zeros and ones, so \( \bar{X}_n \) is a sample proportion (i.e., the proportion of ones in the sample). More familiar notation for the sample proportion is \( \hat{p}, \) as presented in Example 5.14 (notes). This result restated is

\[
\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0,p(1 - p)) \iff \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} N(0,1).
\]

**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) \( Y_nX_n \xrightarrow{d} aX \)

(b) \( X_n + Y_n \xrightarrow{d} X + a. \)
Example 5.21. Suppose \(X_1, X_2, \ldots, X_n\) is an iid sample with \(E(X_1) = \mu\) and \(\text{var}(X_1) = \sigma^2 < \infty\). The CLT says

\[
\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1),
\]
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,

\[
\frac{\bar{X}_n - \mu}{S/\sqrt{n}} = \frac{\sigma}{S} \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1).
\]

\(\frac{\sigma}{S} \xrightarrow{p} 1\).

Note that replacing \(\sigma\) (an unknown parameter) with \(S\) (a consistent estimate of \(\sigma\)) does not affect the asymptotic distribution.

Exercise: Suppose \(X_1, X_2, \ldots, X_n\) are iid Bernoulli(\(p\)), where \(0 < p < 1\). Show that

\[
\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{d} \mathcal{N}(0, 1).
\]

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 \(\theta\) 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 \(\theta\):

\[
g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + \frac{g''(\xi_n)}{2}(X_n - \theta)^2,
\]

where \(\xi_n\) is between \(X_n\) and \(\theta\). Multiplying by \(\sqrt{n}\) and then rearranging, we have

\[
\sqrt{n}|g(X_n) - g(\theta)| = g'(\theta)\sqrt{n}(X_n - \theta) + \frac{\sqrt{n}g''(\xi_n)}{2}(X_n - \theta)^2 \xrightarrow{d} \mathcal{N}(0, \sigma^2).
\]

\(R_n\)
Now, $\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ by assumption so

$$g'(\theta) \sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 \sigma^2).$$

Therefore, if we can show $R_n \xrightarrow{p} 0$, then the result will follow from Slutsky’s Theorem. Note that

$$R_n = \frac{g''(\xi_n)}{2} (X_n - \theta) \sqrt{n}(X_n - \theta).$$

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

$$\lim_{n \to \infty} P(|X_n - \theta| \geq \epsilon) = \lim_{n \to \infty} P\left(\sqrt{n}|X_n - \theta| \geq \sqrt{n}\epsilon\right).$$

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

$$\lim_{n \to \infty} P(|X_n - \theta| \geq \epsilon) = \lim_{n \to \infty} P\left(\sqrt{n}|X_n - \theta| \geq \sqrt{n}\epsilon\right) = 0.$$

Therefore, $X_n \xrightarrow{p} \theta$. By continuity, $X_n - \theta \xrightarrow{p} 0$. Finally,

$$\sqrt{n}[g(X_n) - g(\theta)] = \frac{g'(\theta) \sqrt{n}(X_n - \theta)}{2} + \frac{\sqrt{n}g''(\xi_n)}{2} (X_n - \theta)^2 \xrightarrow{d} 0.$$

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

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)),$$

as $n \to \infty$. We now find the asymptotic distribution of the **log-odds**

$$g(\hat{p}) = \ln\left(\frac{\hat{p}}{1 - \hat{p}}\right),$$

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

$$\sqrt{n}\left[\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) - \ln\left(\frac{p}{1 - p}\right)\right] \xrightarrow{d} \mathcal{N}\left(0, \left[\frac{1}{p(1-p)}\right]^2 p(1-p)\right) \equiv \mathcal{N}\left(0, \frac{1}{p(1-p)}\right),$$

\hfill\text{PAGE 165}
as \( n \to \infty \). In other words,
\[
\ln \left( \frac{\hat{p}}{1 - \hat{p}} \right) \sim \mathcal{AN} \left( \ln \left( \frac{p}{1 - p} \right), \frac{1}{np(1-p)} \right), \text{ for large } n.
\]

**Example 5.23.** Suppose \( X_1, X_2, \ldots, X_n \) are iid Poisson(\( \theta \)), where \( \theta > 0 \), so that \( E(X_1) = \theta \) and \( \text{var}(X_1) = \theta \). The CLT says that
\[
\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \theta),
\]
as \( n \to \infty \). Find a function of \( X_n \), say \( g(X_n) \), whose asymptotic variance is free of \( \theta \).

**Solution.** The delta method says that
\[
\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 \theta),
\]
as \( n \to \infty \). If we set the large sample variance \([g'(\theta)]^2 \theta \) equal to a constant (free of \( \theta \)), we can identify the function \( g \) that satisfies this equation; that is,
\[
[g'(\theta)]^2 \theta = c_0 \implies [g'(\theta)]^2 = \frac{c_0}{\theta} \implies g'(\theta) = \frac{c_1}{\sqrt{\theta}},
\]
where \( c_1 = \sqrt{c_0} \) (also free of \( \theta \)). 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 \( \theta \). Taking \( c_1 = 1/2 \) and \( c_2 = 0 \) yields \( g(\theta) = \sqrt{\theta} \).

**Claim:** The function \( g(X_n) = \sqrt{X_n} \) has asymptotic variance that is free of \( \theta \).

**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
\[
\sqrt{n} \left( \sqrt{X_n} - \sqrt{\theta} \right) \xrightarrow{d} \mathcal{N} \left( 0, \left[ \frac{1}{2 \sqrt{\theta}} \right]^2 \theta \right) = \mathcal{N} \left( 0, \frac{1}{4} \right),
\]
as \( n \to \infty \). In other words,
\[
\sqrt{X_n} \sim \mathcal{AN} \left( \sqrt{\theta}, \frac{1}{4n} \right), \text{ for large } n.
\]

In this example, we see that a **square root transformation** “stabilizes” the asymptotic variance of \( \bar{X}_n \).

**Exercise:** Suppose \( X_1, X_2, \ldots, X_n \) are iid Bernoulli(\( p \)), where \( 0 < p < 1 \). Find a function of \( \hat{p} \), say \( g(\hat{p}) \), whose asymptotic variance is free of \( p \).

**Ans:** \( g(\hat{p}) = \arcsin \sqrt{\hat{p}} \).
Theorem 5.5.26 (Second-order Delta Method). Suppose \( X_n \) is a sequence of random variables satisfying
\[
\sqrt{n}(X_n - \theta) \xrightarrow{d} N(0, \sigma^2),
\]
as \( n \to \infty \). Suppose \( g : \mathbb{R} \to \mathbb{R} \) is twice differentiable at \( \theta \), \( g'(\theta) = 0 \), and \( g''(\theta) \neq 0 \). Then
\[
n[g(X_n) - g(\theta)] \xrightarrow{d} \frac{\sigma^2}{2} g''(\theta) \chi_1^2 \xrightarrow{d} \text{gamma}(1/2, \sigma^2 g''(\theta)).
\]

Example 5.24. Suppose \( X_1, X_2, ..., X_n \) are iid with \( E(X_1) = \mu \) and var\( (X_1) = \sigma^2 < \infty \). The CLT guarantees that
\[
\sqrt{n}(X_n - \mu) \xrightarrow{d} N(0, \sigma^2),
\]
as \( n \to \infty \). Consider \( g(X_n) = X_n^2 \). With \( g(\mu) = \mu^2 \), we have \( g'(\mu) = 2\mu \), which is nonzero except when \( \mu = 0 \). Therefore, provided that \( \mu \neq 0 \),
\[
\sqrt{n}(X_n^2 - \mu^2) \xrightarrow{d} N(0, 4\mu^2 \sigma^2),
\]
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 \( \mu = 0 \)),
\[
n(X_n^2 - \mu^2) = nX_n^2 \xrightarrow{d} \frac{\sigma^2}{2} \chi_1^2 \xrightarrow{d} \text{gamma}(1/2, 2\sigma^2).
\]

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 \( X_1, X_2, ..., X_n \) is a sequence of iid random vectors (of dimension \( k \)) with \( E(X_1) = \mu_{k \times 1} \) and \( \text{cov}(X_1) = \Sigma_{k \times k} \). Let \( \overline{X}_n = (\overline{X}_1, \overline{X}_2, ..., \overline{X}_k)' \) denote the vector of sample means. Then \( \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \text{mvn}(0, \Sigma) \).

Multivariate Delta Method: Suppose \( X_n \) is a sequence of random vectors (of dimension \( k \)) satisfying

(A1) \( \sqrt{n}(X_n - \mu) \xrightarrow{d} \text{mvn}(0, \Sigma) \)

(A2) \( g : \mathbb{R}^k \to \mathbb{R} \) is differentiable at \( \mu \) (and is not zero).

Then
\[
\sqrt{n}[g(X_n) - g(\mu)] \xrightarrow{d} \mathcal{N}\left(0, \frac{\partial g(\mu)}{\partial x} \Sigma \frac{\partial g(\mu)}{\partial x'}\right),
\]
where
\[
\frac{\partial g(\mu)}{\partial x} = \left(\frac{\partial g(x)}{\partial x_1}, \frac{\partial g(x)}{\partial x_2}, ..., \frac{\partial g(x)}{\partial x_k}\right)_{x=\mu}.
\]
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
\[
\left( \frac{\partial g(\mathbf{\mu})}{\partial \mathbf{x}} \right) \Sigma \left( \frac{\partial g(\mathbf{\mu})}{\partial \mathbf{x}'} \right) = \text{a scalar.}
\]

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
\[
\left( \frac{\partial g(\mathbf{x})}{\partial \mathbf{x}} \right) = \begin{pmatrix}
\frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_k} \\
\frac{\partial g_2(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_2(\mathbf{x})}{\partial x_k} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial g_p(\mathbf{x})}{\partial x_1} & \frac{\partial g_p(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_p(\mathbf{x})}{\partial x_k}
\end{pmatrix}_{p \times k},
\]
in which case \( g(\mathbf{\mu}) \) is \( p \times 1 \) and
\[
\left( \frac{\partial g(\mathbf{\mu})}{\partial \mathbf{x}} \right) \Sigma \left( \frac{\partial g(\mathbf{\mu})}{\partial \mathbf{x}'} \right) = p \times p \text{ matrix.}
\]

Example 5.25. Suppose \( \mathbf{X} = (X_1, X_2)' \) is a continuous random vector with joint pdf
\[
f_{\mathbf{X}}(\mathbf{x}) = e^{-x_1^2}I(0 < x_1 < x_2 < \infty).
\]

In Example 4.7 (notes, pp 87), we showed that
\[
X_1 \sim \text{exponential}(1) \\
X_2 \sim \text{gamma}(2, 1).
\]

We have \( E(X_1) = 1, E(X_2) = 2, \text{var}(X_1) = 1, \) and \( \text{var}(X_2) = 2. \) Also,
\[
E(X_1X_2) = \int \int_{\mathbb{R}^2} x_1x_2f_{X_1,X_2}(x_1, x_2)dx_2dx_1 = 3
\]
so
\[
\text{cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = 3 - 2 = 1.
\]

Therefore, for the population described by \( f_{\mathbf{X}}(\mathbf{x}) \), we have
\[
\mathbf{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}_{2 \times 1} \quad \text{and} \quad \Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}_{2 \times 2}.
\]

Now suppose \( \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \) is an iid sample from \( f_{\mathbf{X}}(\mathbf{x}) \); i.e., \( \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \) are mutually independent and \( \mathbf{X}_j = (X_{1j}, X_{2j}) \sim f_{\mathbf{X}}(\mathbf{x}) \), for \( j = 1, 2, \ldots, n, \) Define
\[
\bar{X}_{1+} = \frac{1}{n} \sum_{j=1}^{n} X_{1j} \quad \text{and} \quad \bar{X}_{2+} = \frac{1}{n} \sum_{j=1}^{n} X_{2j}
\]
and denote by
\[
\bar{\mathbf{X}}_n = \begin{pmatrix} \bar{X}_{1+} \\ \bar{X}_{2+} \end{pmatrix}.
\]
the vector of sample means. The (multivariate) CLT says that
\[ \sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \text{mvn}_2(0, \Sigma), \]
as \( n \to \infty \), where \( \mu \) and \( \Sigma \) are given above. In other words,
\[ \bar{X}_n \sim \mathcal{A}N_2 \left( \left( \frac{1}{2} \right), \left( \frac{1}{n} \ 1/n \ 1/2n \right) \right), \quad \text{for large } n. \]

**Q:** Find the large sample distribution of
\[ R = g(\bar{X}_n) = \frac{\bar{X}_{1+}}{\bar{X}_{2+}}, \]
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
\[ \frac{\partial g(\mu)}{\partial x} = \left( \frac{1}{\mu_2} - \frac{\mu_1}{\mu_2^2} \right) = \left( \frac{1}{2} \ -1/4 \right). \]
The multivariate delta method says that
\[ \sqrt{n}[g(\bar{X}_n) - g(\mu)] = \sqrt{n} \left( R - \frac{1}{2} \right) \xrightarrow{d} \mathcal{N}(0, \sigma_R^2), \]
as \( n \to \infty \), where
\[ \sigma_R^2 = \frac{\partial g(\mu)}{\partial x} \Sigma \frac{\partial g(\mu)}{\partial x'} = \left( \frac{1}{2} \ -1/4 \right) \left( \begin{array}{cc} 1 & 1 \\ 1 & 2 \end{array} \right) \left( \frac{1}{2} \ 1/4 \right) = \frac{1}{8}. \]
In other words,
\[ R \sim \mathcal{A}N \left( \frac{1}{2}, \frac{1}{8n} \right), \quad \text{for large } n. \]

**Example 5.26.** In medical settings, it is common to observe data in the form of 2 × 2 tables such as

<table>
<thead>
<tr>
<th></th>
<th>Cured</th>
<th>Not cured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>( X_{11} )</td>
<td>( X_{12} )</td>
</tr>
<tr>
<td>Group 2</td>
<td>( X_{21} )</td>
<td>( X_{22} )</td>
</tr>
</tbody>
</table>

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 $Y_k = (Y_{1k}, Y_{2k}, Y_{1k}, Y_{2k})$ and $Y_{ijk} = I(k$th individual is in cell $ij)$, for $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, ..., n$. In other words, $Y_1, Y_2, ..., Y_n$ are iid mult($1; p_{11}, p_{12}, p_{21}, p_{22}$) with

$$
\mu = E(Y_1) = \begin{pmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \end{pmatrix} = \mathbf{p}
$$

and

$$
\Sigma = \text{cov}(Y_1) = \begin{pmatrix} p_{11}(1-p_{11}) & -p_{11}p_{12} & -p_{11}p_{21} & -p_{11}p_{22} \\ -p_{12}p_{11} & p_{12}(1-p_{12}) & -p_{12}p_{21} & -p_{12}p_{22} \\ -p_{21}p_{11} & -p_{21}p_{12} & p_{21}(1-p_{21}) & -p_{21}p_{22} \\ -p_{22}p_{11} & -p_{22}p_{12} & -p_{22}p_{21} & p_{22}(1-p_{22}) \end{pmatrix}
$$

$$
= D(\mathbf{p}) - \mathbf{p}\mathbf{p}'
$$

where $D(\mathbf{p}) = \text{diag}(\mathbf{p})$. Define the vector of sample proportions as

$$
\hat{\mathbf{p}} = \frac{\mathbf{X}}{n} = \frac{1}{n} \sum_{k=1}^{n} Y_k.
$$

The (multivariate) CLT says that $\sqrt{n} (\hat{\mathbf{p}} - \mathbf{p}) \overset{d}{\to} \text{mvn}(0, \Sigma)$, as $n \to \infty$. This is a less-than-full-rank normal distribution (STAT 714) because $r(D(\mathbf{p}) - \mathbf{p}\mathbf{p}') = 3 < 4$.

**Q:** Find the large sample distribution of the log-odds ratio $g(\hat{\mathbf{p}}) = \ln \left( \frac{p_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}} \right)$, suitably centered and scaled.

**Solution:** With

$$
g(\mathbf{p}) = \ln \left( \frac{p_{11}p_{22}}{p_{12}p_{21}} \right),
$$

we have

$$
\frac{\partial g(\mathbf{p})}{\partial \mathbf{p}} = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} p_{11}^{-1} & 0 & 0 & 0 \\ 0 & p_{12}^{-1} & 0 & 0 \\ 0 & 0 & p_{21}^{-1} & 0 \\ 0 & 0 & 0 & p_{22}^{-1} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} D^{-1}(\mathbf{p}).
$$

By the multivariate delta method,

$$
\sqrt{n}[g(\hat{\mathbf{p}}) - g(\mathbf{p})] \overset{d}{\to} \mathcal{N}(0, \sigma^2),
$$

as $n \to \infty$, where

$$
\sigma^2 = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} D^{-1}(\mathbf{p})[D(\mathbf{p}) - \mathbf{p}\mathbf{p}']D^{-1}(\mathbf{p}) \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix}'.
$$

$$
= \frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}.
$$

In other words,

$$
\ln \left( \frac{\hat{p}_{11}\hat{p}_{22}}{\hat{p}_{12}\hat{p}_{21}} \right) \sim \mathcal{N} \left( \ln \left( \frac{p_{11}p_{22}}{p_{12}p_{21}} \right), \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{1}{p_{ij}} \right), \text{ for large } n.
$$
6 Principles of Data Reduction


6.1 Introduction

Recall: We begin by recalling the definition of a statistic. Suppose that $X_1, X_2, ..., X_n$ is an iid sample. A statistic $T = T(X) = T(X_1, X_2, ..., X_n)$ is a function of the sample $X = (X_1, X_2, ..., X_n)$. The only restriction is that $T$ cannot depend on unknown parameters. For example,

$$T(X) = \overline{X} \quad T(X) = X_{(n)} \quad T(X) = \left( \prod_{i=1}^{n} X_i \right)^{1/n} \quad T(X) = X.$$

Recall: We can think of $X$ and $T$ as functions:

- $(S, B, P)$: probability space for random experiment
- $(\mathbb{R}^n, B(\mathbb{R}^n), P_X) \to$ range space of $X$
  - Recall: $X : S \to \mathbb{R}^n$ is a random vector if
    $$X^{-1}(B) \equiv \{ \omega \in S : X(\omega) \in B \} \in B,$$
    for all $B \in B(\mathbb{R}^n)$
  - $P_X$: induced probability measure of $X$; one-to-one correspondence with $F_X(x)$
  - $\mathcal{X} =$ support of $X$; $\mathcal{X} \subseteq \mathbb{R}^n$

- $(\mathbb{R}, B(\mathbb{R}), P_T) \to$ range space of $T$
  - $T : \mathbb{R}^n \to \mathbb{R}$, if $T$ is a scalar statistic
  - $P_T$ describes the (sampling) distribution of $T$; Chapters 4-5 (CB)
  - $\mathcal{T} =$ support of $T$; $\mathcal{T} \subseteq \mathbb{R}$;
    $$\mathcal{T} = \{ t : t = T(x), \ x \in \mathcal{X} \}$$
    = image set of $\mathcal{X}$ under $T$.

Conceptualization: A statistic $T$ forms a partition of $\mathcal{X}$, the support of $X$. Specifically, $T$ partitions $\mathcal{X} \subseteq \mathbb{R}^n$ into sets

$$A_t = \{ x \in \mathcal{X} : T(x) = t \},$$

for $t \in \mathcal{T}$. The statistic $T$ summarizes the data $x$ in that one can report

$$T(x) = t \iff x \in A_t$$
instead of reporting \( x \) itself. This is the idea behind data reduction. We reduce the data \( x \) so that they can be more easily understood without losing the meaning associated with the set of observations.

**Example 6.1.** Suppose \( X_1, X_2, X_3 \) are iid Bernoulli(\( \theta \)), where \( 0 < \theta < 1 \). The support of \( X = (X_1, X_2, X_3) \) is

\[
\mathcal{X} = \{(0,0,0), (1,0,0), (0,1,0), (0,0,1), (1,0,0), (1,0,1), (0,1,1), (1,1,1)\}.
\]

Define the statistic

\[
T = T(X) = X_1 + X_2 + X_3.
\]

The statistic \( T \) partitions \( \mathcal{X} \subset \mathbb{R}^3 \) into the following sets:

\[
A_0 = \{x \in \mathcal{X}: T(x) = 0\} = \{(0,0,0)\}
\]

\[
A_1 = \{x \in \mathcal{X}: T(x) = 1\} = \{(1,0,0), (0,1,0), (0,0,1)\}
\]

\[
A_2 = \{x \in \mathcal{X}: T(x) = 2\} = \{(1,1,0), (1,0,1), (0,1,1)\}
\]

\[
A_3 = \{x \in \mathcal{X}: T(x) = 3\} = \{(1,1,1)\}.
\]

The image of \( \mathcal{X} \) under \( T \) is

\[
T = \{t : t = T(x), \ x \in \mathcal{X}\} = \{0, 1, 2, 3\},
\]

the support of \( T \). The statistic \( T \) summarizes the data in that it reports only the value \( T(x) = t \). It does not report which \( x \in \mathcal{X} \) produced \( T(x) = t \).

**Connection:** Data reduction plays an important role in statistical inference. Suppose \( X_1, X_2, ..., X_n \) is an iid sample from \( f_X(x|\theta) \), where \( \theta \in \Theta \). We would like to use the sample \( X \) to learn about which member (or members) of this family might be reasonable. We also do not want to be burdened by having to work with the entire sample \( X \). Therefore, we are interested in statistics \( T \) that reduce the data \( X \) (for convenience) while still not compromising our ability to learn about \( \theta \).

**Preview:** Chapter 6 (CB) discusses three methods of data reduction:

- Section 6.2: Sufficiency Principle
- Section 6.3: Likelihood Principle
- Section 6.4: Equivariance Principle

We will focus exclusively on Section 6.2.

### 6.2 The Sufficiency Principle

#### 6.2.1 Sufficient statistics

**Informal Definition:** A statistic \( T = T(X) \) is a sufficient statistic for a parameter \( \theta \) if it contains “all of the information” about \( \theta \) that is available in the sample. In other words, we do not lose any information about \( \theta \) by reducing the sample \( X \) to the statistic \( T \).
**Sufficiency Principle:** If $T = T(X)$ is a sufficient statistic for $\theta$, then any inference regarding $\theta$ should depend on $X$ only through the value of $T(X)$.

- In other words, if $x \in X$, $y \in X$, and $T(x) = T(y)$, then inference for $\theta$ should be the same whether $X = x$ or $X = y$ is observed.
- For example, in Example 6.1, suppose
  
  $x = (1, 0, 0)$
  $y = (0, 0, 1)$

  so that $t = T(x) = T(y) = 1$. The Sufficiency Principle says that inference for $\theta$ depends only on the value of $t = 1$ and not on whether $x$ or $y$ was observed.

**Definition 6.2.1/Theorem 6.2.2.** A statistic $T = T(X)$ is a sufficient statistic for $\theta$ if the conditional distribution of $X$ given $T$ is free of $\theta$; i.e., if the ratio

$$f_{X|T}(x|t) = \frac{f_{X}(x|\theta)}{f_{T}(t|\theta)}$$

is free of $\theta$, for all $x \in X$. In other words, after conditioning on $T$, we have removed all information about $\theta$ from the sample $X$.

**Discussion:** Note that in the discrete case, all distributions above can be interpreted as probabilities. From the definition of a conditional distribution,

$$f_{X|T}(x|t) = \frac{f_{X,T}(x,t|\theta)}{f_{T}(t|\theta)} = \frac{P_{\theta}(X = x, T = t)}{P_{\theta}(T = t)}.$$

Because $\{X = x\} \subset \{T = t\}$, we have

$$P_{\theta}(X = x, T = t) = P_{\theta}(X = x) = f_{X}(x|\theta).$$

Therefore,

$$f_{X|T}(x|t) = \frac{f_{X}(x|\theta)}{f_{T}(t|\theta)}$$

as claimed. If $T$ is continuous, then $f_{T}(t|\theta) \neq P_{\theta}(T = t)$ and $f_{X|T}(x|t)$ cannot be interpreted as a conditional probability. Fortunately, the criterion above; i.e.,

$$f_{X|T}(x|t) = \frac{f_{X}(x|\theta)}{f_{T}(t|\theta)}$$

being free of $\theta$, still applies in the continuous case (although a more rigorous explanation would be needed to see why).

**Example 6.2.** Suppose $X_1, X_2, \ldots, X_n$ are iid Poisson($\theta$), where $\theta > 0$. Use Definition 6.2.1/Theorem 6.2.2 to show that

$$T = T(X) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic.
Proof. The pmf of $X$, for $x_i = 0, 1, 2, \ldots$, is given by

$$f_X(x|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}}{\prod_{i=1}^{n} x_i!}.$$  

Recall that $T \sim \text{Poisson}(n\theta)$, shown by using mgfs. Therefore, the pmf of $T$, for $t = 0, 1, 2, \ldots$, is

$$f_T(t|\theta) = \frac{(n\theta)^{t} e^{-n\theta}}{t!}.$$  

With $t = \sum_{i=1}^{n} x_i$, the conditional distribution

$$f_{X|T}(x|t) = \frac{f_X(x|\theta)}{f_T(t|\theta)} = \frac{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}}{\prod_{i=1}^{n} x_i!} \cdot \frac{t!}{(n\theta)^{t} e^{-n\theta}} = \frac{t!}{n^{t} \prod_{i=1}^{n} x_i!},$$  

which is free of $\theta$. From the definition of sufficiency and from Theorem 6.2.2, we have shown that $T = T(X) = \sum_{i=1}^{n} X_i$ is a sufficient statistic. □

Example 6.3. Suppose that $X_1, X_2, \ldots, X_n$ is an iid sample from

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

an exponential distribution with mean $\theta > 0$. Show that

$$T = T(X) = X$$

is a sufficient statistic.

Proof. The pdf of $X$, for $x_i > 0$, is given by

$$f_X(x|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} x_i/\theta}.$$  

Recall that if $X_1, X_2, \ldots, X_n$ are iid exponential($\theta$), then

$$\overline{X} \sim \text{gamma}(n, \theta/n).$$

Therefore, the pdf of $T = T(X) = \overline{X}$, for $t > 0$, is

$$f_T(t|\theta) = \frac{1}{\Gamma(n) \left( \frac{\theta}{n} \right)^n} t^{n-1} e^{-nt/\theta}.$$  

With $t = \overline{x}$ (i.e., $nt = \sum_{i=1}^{n} x_i$), the conditional distribution

$$f_{X|T}(x|t) = \frac{f_X(x|\theta)}{f_T(t|\theta)} = \frac{1}{\theta^n} e^{-\sum_{i=1}^{n} x_i/\theta} = \frac{\Gamma(n)}{\Gamma(n) \left( \frac{\theta}{n} \right)^n t^{n-1} e^{-nt/\theta}} = \frac{\Gamma(n)}{n^{n} t^{n-1}},$$  

which is free of $\theta$. From the definition of sufficiency and from Theorem 6.2.2, we have shown that $T = T(X) = \overline{X}$ is a sufficient statistic. □
Example 6.4. Suppose $X_1, X_2, ..., X_n$ is an iid sample from a continuous distribution with pdf $f_X(x|\theta)$, where $\theta \in \Theta$. Show that $T = T(X) = (X(1), X(2), ..., X(n))$, the vector of order statistics, is always sufficient.

Proof. Recall from Section 5.4 (CB) that the joint distribution of the $n$ order statistics is
\[
  f_{X(1), X(2), ..., X(n)}(x_1, x_2, ..., x_n|\theta) = n! f_X(x_1|\theta) f_X(x_2|\theta) \cdots f_X(x_n|\theta) = n! f_X(x|\theta),
\]
for $-\infty < x_1 < x_2 < \cdots < x_n < \infty$. Therefore, the ratio
\[
  \frac{f_X(x|\theta)}{f_T(x|\theta)} = \frac{f_X(x|\theta)}{n! f_X(x|\theta)} = \frac{1}{n!},
\]
which is free of $\theta$. From the definition of sufficiency and from Theorem 6.2.2, we have shown that $T = T(X) = (X(1), X(2), ..., X(n))$ is a sufficient statistic. □

Discussion: Example 6.4 shows that (with continuous distributions), the order statistics are always sufficient.

- Of course, reducing the sample $X = (X_1, X_2, ..., X_n)$ to $T(X) = (X(1), X(2), ..., X(n))$ is not that much of a reduction. However, in some parametric families, it is not possible to reduce $X$ any further without losing information about $\theta$ (e.g., Cauchy, logistic, etc.); see pp 275 (CB).

- In some instances, it may be that the parametric form of $f_X(x|\theta)$ is not specified. With so little information provided about the population, we should not be surprised that the only available reduction of $X$ is to the order statistics.

Remark: The approach we have outlined to show that a statistic $T$ is sufficient appeals to Definition 6.2.1 and Theorem 6.2.2; i.e., we are using the definition of sufficiency directly by showing that the conditional distribution of $X$ given $T$ is free of $\theta$.

- If I ask you to show that $T$ is sufficient by appealing to the definition of sufficiency, this is the approach I want you to take.

- What if we need to find a sufficient statistic? Then the approach we have just outlined is not practical to implement (i.e., imagine trying different statistics $T$ and for each one attempting to show that $f_{X|T}(x|t)$ is free of $\theta$). This might involve a large amount of trial and error and you would have to derive the sampling distribution of $T$ each time (which for many statistics can be difficult or even intractable).

- The Factorization Theorem makes getting sufficient statistics much easier.

Theorem 6.2.6 (Factorization Theorem). A statistic $T = T(X)$ is sufficient for $\theta$ if and only if there exists functions $g(t|\theta)$ and $h(x)$ such that
\[
  f_X(x|\theta) = g(t|\theta) h(x),
\]
for all support points $x \in \mathcal{X}$ and for all $\theta \in \Theta$. 
Proof. We prove the result for the discrete case only; the continuous case is beyond the scope of this course.

Necessity (\(\implies\)): Suppose \(T\) is sufficient. It suffices to show there exists functions \(g(t|\theta)\) and \(h(x)\) such that the factorization holds. Because \(T\) is sufficient, we know
\[
f_{X|T}(x|t) = P(X = x|T(X) = t)
\]
is free of \(\theta\) (this is the definition of sufficiency). Therefore, take
\[
g(t|\theta) = P_\theta(T(X) = t) \\
h(x) = P(X = x|T(X) = t).
\]
Because \(\{X = x\} \subset \{T(X) = t\}\),
\[
f_X(x|\theta) = P_\theta(X = x) \\
= P_\theta(X = x, T(X) = t) \\
= P_\theta(T(X) = t) P(X = x|T(X) = t) = g(t|\theta)h(x).
\]

Sufficiency (\(\iff\)): Suppose the factorization holds. To establish that \(T = T(X)\) is sufficient, it suffices to show that
\[
f_{X|T}(x|t) = P(X = x|T(X) = t)
\]
is free of \(\theta\). Denoting \(T(x) = t\), we have
\[
f_{X|T}(x|t) = P(X = x|T(X) = t) = \frac{P_\theta(X = x, T(X) = t)}{P_\theta(T(X) = t)} \\
= \frac{P_\theta(X = x)}{P_\theta(T(X) = t)} I(T(X) = t) \\
= \frac{g(t|\theta)h(x)}{P_\theta(T(X) = t)} I(T(X) = t),
\]
because the factorization holds by assumption. Now write
\[
P_\theta(T(X) = t) = P_\theta(X \in A_t),
\]
where recall \(A_t = \{x \in \mathcal{X} : T(x) = t\}\) is a set over \((\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), P_X)\). Note that
\[
P_\theta(X \in A_t) = \sum_{x \in \mathcal{X} : T(x) = t} P_\theta(X = x) \\
= \sum_{x \in \mathcal{X} : T(x) = t} g(t|\theta)h(x) \\
= g(t|\theta) \sum_{x \in \mathcal{X} : T(x) = t} h(x).
\]
Therefore,
\[
f_{X|T}(x|t) = \frac{g(t|\theta)h(x) I(T(X) = t)}{g(t|\theta) \sum_{x \in \mathcal{X} : T(x) = t} h(x)} = \frac{h(x) I(T(X) = t)}{\sum_{x \in \mathcal{X} : T(x) = t} h(x)},
\]
which is free of \(\theta\). \(\square\)
Example 6.2 (continued). Suppose \( X_1, X_2, ..., X_n \) are iid Poisson(\( \theta \)), where \( \theta > 0 \). We have already shown that

\[
T = T(X) = \sum_{i=1}^{n} X_i
\]

is a sufficient statistic (using the definition of sufficiency). We now show this using the Factorization Theorem. For \( x_i = 0, 1, 2, ... \), the pmf of \( X \) is

\[
f_X(x|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}
= \frac{\theta^{\sum_{i=1}^{n} x_i} e^{-n\theta}}{\prod_{i=1}^{n} x_i!}
= \frac{\theta^t e^{-n\theta}}{\prod_{i=1}^{n} x_i!}
= g(t|\theta)
\]

\[
\frac{1}{\prod_{i=1}^{n} x_i!}
= h(x)
\]

where \( t = \sum_{i=1}^{n} x_i \). By the Factorization Theorem, \( T = T(X) = \sum_{i=1}^{n} X_i \) is sufficient.

Example 6.5. Suppose \( X_1, X_2, ..., X_n \) are iid \( U(0, \theta) \), where \( \theta > 0 \). Find a sufficient statistic.

Solution. The pdf of \( X \) is

\[
f_X(x|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 < x_i < \theta)
= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 < x_i < \theta)
= \frac{1}{\theta^n} I(x_{(n)} < \theta) \prod_{i=1}^{n} I(x_i > 0),
= g(t|\theta)
= h(x)
\]

where \( t = x_{(n)} \). By the Factorization Theorem, \( T = T(X) = X_{(n)} \) is sufficient.

Example 6.6. Suppose \( X_1, X_2, ..., X_n \) are iid gamma(\( \alpha, \beta \)), where \( \alpha > 0 \) and \( \beta > 0 \). Note that in this family, the parameter \( \theta = (\alpha, \beta) \) is two-dimensional. The pdf of \( X \) is

\[
f_X(x|\theta) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-x_i/\beta} I(x_i > 0)
= \left[ \frac{1}{\Gamma(\alpha)\beta^\alpha} \right]^{\frac{n}{\beta^\alpha}} \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\sum_{i=1}^{n} x_i/\beta} \prod_{i=1}^{n} I(x_i > 0)
= g(t_1, t_2|\theta)
= h(x)
\]

where \( t_1 = \prod_{i=1}^{n} x_i \) and \( t_2 = \sum_{i=1}^{n} x_i \). By the Factorization Theorem,

\[
T = T(X) = \begin{pmatrix} \sum_{i=1}^{n} X_i \\ X_i \end{pmatrix}
\]

is sufficient.
Remark: In previous examples, we have seen that the dimension of a sufficient statistic $T$ often equals the dimension of the parameter $\theta$:

- Example 6.2: Poisson($\theta$). $T = \sum_{i=1}^{n} X_i$; $\text{dim}(T) = \text{dim}(\theta) = 1$
- Example 6.3: exponential($\theta$). $T = X; \text{dim}(T) = \text{dim}(\theta) = 1$
- Example 6.5: $U(0, \theta)$. $T = X(n); \text{dim}(T) = \text{dim}(\theta) = 1$
- Example 6.6: gamma($\alpha, \beta$). $T = (\prod_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i); \text{dim}(T) = \text{dim}(\theta) = 2$.

Sometimes the dimension of a sufficient statistic is larger than that of the parameter. We have already seen this in Example 6.4 where $T(X) = (X(1), X(2), ..., X(n))$, the vector of order statistics, was sufficient; i.e., $\text{dim}(T) = n$. In some parametric families (e.g., Cauchy, etc.), this statistic is sufficient and no further reduction is possible.

Example 6.7. Suppose $X_1, X_2, ..., X_n$ are iid $U(\theta, \theta + 1)$, where $-\infty < \theta < \infty$. This is a one-parameter family; i.e., $\text{dim}(\theta) = 1$. The pdf of $X$ is

$$f_X(x|\theta) = \prod_{i=1}^{n} I(\theta < x_i < \theta + 1)$$

$$= \prod_{i=1}^{n} I(x_i > \theta) \prod_{i=1}^{n} I(x_i - 1 < \theta)$$

$$= \left\{ I(x(1) > \theta) I(x(n) - 1 < \theta) \right\} \prod_{i=1}^{n} I(x_i \in \mathbb{R}),$$

where $t_1 = x(1)$ and $t_2 = x(n)$. By the Factorization Theorem,

$$T = T(X) = \begin{pmatrix} X(1) \\ X(n) \end{pmatrix}$$

is sufficient. In this family, $2 = \text{dim}(T) > \text{dim}(\theta) = 1$.

Remark: Sufficiency also extends to non-iid situations.

Example 6.8. Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

for $i = 1, 2, ..., n$, where $\epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2)$ and the $x_i$’s are fixed constants (i.e., not random). In this model, it is easy to show that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2),$$

so that $\theta = (\beta_0, \beta_1, \sigma^2)$. Note that $Y_1, Y_2, ..., Y_n$ are independent random variables (functions of independent random variables are independent); however, $Y_1, Y_2, ..., Y_n$ are not identically distributed because $E(Y_i) = \beta_0 + \beta_1 x_i$ changes as $i$ does.
For \( y \in \mathbb{R}^n \), the pdf of \( Y \) is
\[
  f_Y(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\
  = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right\}.
\]

It is easy to show that
\[
  \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\beta_0 \sum_{i=1}^{n} y_i - 2\beta_1 \sum_{i=1}^{n} x_i y_i + n\beta_0^2 + 2\beta_0\beta_1 \sum_{i=1}^{n} x_i + \beta_1^2 \sum_{i=1}^{n} x_i^2,
\]
where \( t_1 = \sum_{i=1}^{n} y_i^2 \), \( t_2 = \sum_{i=1}^{n} y_i \), and \( t_3 = \sum_{i=1}^{n} x_i y_i \). Taking \( h(y) = 1 \), the Factorization Theorem shows that
\[
  T = T(Y) = \left(\begin{array}{c}
  \sum_{i=1}^{n} y_i^2 \\
  \sum_{i=1}^{n} y_i \\
  \sum_{i=1}^{n} x_i y_i
  \end{array}\right)
\]
is sufficient. Note that \( \text{dim}(T) = \text{dim}(\theta) = 3 \).

**Sufficient statistics in the exponential family:**

**Theorem 6.2.10.** Suppose \( X_1, X_2, \ldots, X_n \) are iid from the exponential family
\[
  f_X(x|\theta) = h(x)c(\theta) \exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\},
\]
where \( \theta = (\theta_1, \theta_2, \ldots, \theta_d) \), \( d \leq k \). Then
\[
  T = T(X) = \left(\begin{array}{c}
  \sum_{j=1}^{n} t_1(X_j) \\
  \sum_{j=1}^{n} t_2(X_j) \\
  \vdots \\
  \sum_{j=1}^{n} t_k(X_j)
  \end{array}\right)
\]
is sufficient.

**Proof.** Use the Factorization Theorem. The pdf of \( X \) is
\[
  f_X(x|\theta) = \prod_{j=1}^{n} h(x_j)c(\theta) \exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x_j)\right\} \\
  = \prod_{j=1}^{n} h(x_j) \left[c(\theta)\right]^n \exp\left\{\sum_{i=1}^{k} w_i(\theta) \sum_{j=1}^{n} t_i(x_j)\right\},
\]
where \( t_i = \sum_{j=1}^{n} t_i(x_j) \), for \( i = 1, 2, \ldots, k \). \( \square \)
Example 6.9. Suppose $X_1, X_2, ..., X_n$ are iid Bernoulli($\theta$), where $0 < \theta < 1$. For $x = 0, 1$, the pmf of $X$ is
\[
 f_X(x|\theta) = \theta^x(1-\theta)^{1-x} \\
 = (1-\theta)\left(\frac{\theta}{1-\theta}\right)^x \\
 = (1-\theta)\exp\left\{\ln\left(\frac{\theta}{1-\theta}\right) x\right\} \\
 = h(x)c(\theta)\exp\{w_1(\theta)t_1(x)\},
\]
where $h(x) = 1$, $c(\theta) = 1-\theta$, $w_1(\theta) = \ln\{\theta/(1-\theta)\}$, and $t_1(x) = x$. By Theorem 6.2.10,
\[
 T = T(X) = \sum_{j=1}^{n} t_1(X_j) = \sum_{j=1}^{n} X_j
\]
is sufficient.

Result: Suppose $X \sim f_X(x|\theta)$, where $\theta \in \Theta$, and suppose $T = T(X)$ is sufficient. If $r$ is a one-to-one function, then $r(T(X))$ is also sufficient.

Proof. Let $T^*(X) = r(T(X))$ so that $T(X) = r^{-1}(T^*(X))$. We have
\[
 f_X(x|\theta) = g(T(X)|\theta)h(x) \\
 = g(r^{-1}(T^*(X))|\theta)h(x) \\
 = g^*(T^*(X)|\theta)h(x),
\]
where $g^* = g \circ r^{-1}$; i.e., $g^*$ is the composition of $g$ and $r^{-1}$. By the Factorization Theorem, $T^*(X)$ is sufficient. $\square$

Applications:

- In Example 6.9, we showed that
\[
 T = \sum_{i=1}^{n} X_i
\]
is a sufficient statistic for the Bernoulli family. By the previous result, we know that $T^*_1(X) = \bar{X}$ and $T^*_2(X) = e^{\sum_{i=1}^{n} X_i}$ are also sufficient. Note that $r_1(t) = t/n$ and $r_2(t) = e^t$ are one-to-one functions over $T = \{t : t = 0, 1, 2, ..., n\}$.

- In the $\mathcal{N}(\mu, \sigma^2)$ family where both parameters are unknown, it is easy to show that
\[
 \mathbf{T} = \mathbf{T}(X) = \begin{pmatrix} \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i^2 \end{pmatrix}
\]
is sufficient (just apply the Factorization Theorem directly or use our result dealing with exponential families). Define the function
\[
 r(t) = r(t_1, t_2) = \left(\frac{t_1/n}{n-1(t_2 - t_1^2/n)}\right),
\]
and note that \( r(t) \) is one-to-one over \( T = \{(t_1, t_2) : -\infty < t_1 < \infty, \ t_2 \geq 0\} \). Therefore,

\[
r(T(X)) = r \left( \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2} \right) = \left( \frac{X}{S^2} \right)
\]

is also sufficient in the \( \mathcal{N}(\mu, \sigma^2) \) family.

**Remark:** In the \( \mathcal{N}(\mu, \sigma^2) \) family where both parameters are unknown, the statistic \( T(X) = (\bar{X}, S^2) \) is sufficient.

- In the \( \mathcal{N}(\mu, \sigma^2_0) \) subfamily where \( \sigma^2_0 \) is known, \( T(X) = \bar{X} \) is sufficient.
- In the \( \mathcal{N}(\mu_0, \sigma^2) \) subfamily where \( \mu_0 \) is known,

\[
T(X) = \sum_{i=1}^{n} (X_i - \mu_0)^2
\]

is sufficient. Interestingly, \( S^2 \) is not sufficient in this subfamily. It is easy to show that \( f_{X|S^2}(x|s^2) \) depends on \( \sigma^2 \).

### 6.2.2 Minimal sufficient statistics

**Example 6.10.** Suppose that \( X_1, X_2, \ldots, X_n \) are iid \( \mathcal{N}(\mu, \sigma^2_0) \), where \( -\infty < \mu < \infty \) and \( \sigma^2_0 \) is known. Each of the following statistics is sufficient:

\[
T_1(X) = \bar{X}, \quad T_2(X) = \left( X_1, \sum_{i=2}^{n} X_i \right), \quad T_3(X) = (X_{(1)}, X_{(2)}, \ldots, X_{(n)}), \quad T_4(X) = X.
\]

**Q:** How much data reduction is possible?

**Definition:** A sufficient statistic \( T = T(X) \) is called a **minimal sufficient statistic** if, for any other sufficient statistic \( T^*(X) \), \( T(x) \) is a function of \( T^*(x) \).

**Remark:** A minimal sufficient statistic is a sufficient statistic that offers the most data reduction. Note that "\( T(x) \) is a function of \( T^*(x) \)" means

\[
T^*(x) = T^*(y) \implies T(x) = T(y).
\]

Informally, if you know \( T^*(x) \), you can calculate \( T(x) \), but not necessarily vice versa.

**Remark:** You can also characterize minimality of a sufficient statistic using the partitioning concept described at the beginning of this chapter. Consider the collection of sufficient statistics. A minimal sufficient statistic \( T = T(X) \) admits the **coarsest** possible partition in the collection.
Consider the following table:

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Description</th>
<th>Partition of $\mathcal{X}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T(x)$</td>
<td>Minimal sufficient</td>
<td>$A_t$, $t = 1, 2, \ldots$</td>
</tr>
<tr>
<td>$T^*(x)$</td>
<td>Sufficient</td>
<td>$B_t$, $t = 1, 2, \ldots$</td>
</tr>
</tbody>
</table>

By “coarsest possible partition,” we mean that $\mathcal{X}$ (the support of $X$) cannot be split up further and still be a sufficient partition (i.e., a partition for a sufficient statistic). This means that $\{B_t, t = 1, 2, \ldots\}$ must be a sub-partition of $\{A_t, t = 1, 2, \ldots\}$; i.e., each $B_t$ set associated with $T^*(X)$ is a subset of some $A_t$ associated with $T(X)$.

**Theorem 6.2.13.** Suppose $X \sim f_X(x|\theta)$, where $\theta \in \Theta$. Suppose there is a function $T(x)$ such that for all $x, y \in \mathcal{X}$, 

$$\frac{f_X(x|\theta)}{f_X(y|\theta)} \text{ is free of } \theta \iff T(x) = T(y).$$

Then $T(X)$ is a minimal sufficient statistic.

**Example 6.10** (continued). Suppose $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2_0)$, where $-\infty < \mu < \infty$ and $\sigma^2_0$ is known. For $x \in \mathbb{R}^n$, the pdf of $X$ is 

$$f_X(x|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_i-\mu)^2/2\sigma_0^2} = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n e^{-\sum_{i=1}^n (x_i-\mu)^2/2\sigma_0^2}.$$ 

Now write 

$$\sum_{i=1}^n (x_i-\mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$$

and form the ratio 

$$\frac{f_X(x|\mu)}{f_X(y|\mu)} = \frac{\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\{-\sum_{i=1}^n (x_i-\mu)^2/2\sigma_0^2\}}{\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n \exp\{-\sum_{i=1}^n (y_i-\mu)^2/2\sigma_0^2\}} = \frac{\exp\{-[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2]/2\sigma_0^2\}}{\exp\{-[\sum_{i=1}^n (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2]/2\sigma_0^2\}}.$$ 

Clearly, this ratio is free of $\mu$ if and only if $X = \bar{y}$. By Theorem 6.2.13, we know that $T(X) = \bar{X}$ is a minimal sufficient statistic.

**Result:** Suppose $X \sim f_X(x|\theta)$, where $\theta \in \Theta$, and suppose $T = T(X)$ is minimal sufficient. If $r$ is a one-to-one function, then $r(T(X))$ is also minimal sufficient.

**Example 6.7** (continued). Suppose $X_1, X_2, \ldots, X_n$ are iid $\mathcal{U}(\theta, \theta + 1)$, where $-\infty < \theta < \infty$. We have already shown the pdf of $X$ is 

$$f_X(x|\theta) = I(x(1) > \theta)I(x(n) - 1 < \theta) \prod_{i=1}^n I(x_i \in \mathbb{R}).$$
Clearly, the ratio
\[
f_X(x|\theta) = \frac{I(x(1) > \theta)I(x(n) - 1 < \theta) \prod_{i=1}^{n} I(x_i \in \mathbb{R})}{I(y(1) > \theta)I(y(n) - 1 < \theta) \prod_{i=1}^{n} I(y_i \in \mathbb{R})}
\]
is free of \( \theta \) if and only if \((x(1), x(n)) = (y(1), y(n))\). By Theorem 6.2.13, we know that \( T(X) = (X(1), X(n)) \) is a minimal sufficient statistic. Note that in this family, the dimension of a minimal sufficient statistic does not match the dimension of the parameter. Note also that a one-to-one function of \( T(X) \) is
\[
\begin{pmatrix}
X(n) - X(1)

X(1) + X(n)/2
\end{pmatrix}
\]
which is also minimal sufficient.

6.2.3 Ancillary statistics

**Definition:** A statistic \( S = S(X) \) is an **ancillary statistic** if the distribution of \( S \) does not depend on the model parameter \( \theta \).

**Remark:** You might characterize an ancillary statistic as being “unrelated” to a sufficient statistic. After all, sufficient statistics contain all the information about the parameter \( \theta \) and ancillary statistics have distributions that are free of \( \theta \).

**Example 6.11.** Suppose that \( X_1, X_2, \ldots, X_n \) are iid \( \mathcal{N}(0, \sigma^2) \), where \( \sigma^2 > 0 \). Note that
\[
\bar{X} \sim \mathcal{N}(0, \sigma^2/n),
\]
so \( \bar{X} \) is not ancillary (its distribution depends on \( \sigma^2 \)). However, the statistic
\[
S(X) = \frac{\bar{X}}{S/\sqrt{n}} \sim t_{n-1}
\]
is ancillary because its distribution, \( t_{n-1} \), does not depend on \( \sigma^2 \). Also, it is easy to show that
\[
T(X) = \sum_{i=1}^{n} X_i^2
\]
is a (minimal) sufficient statistic for \( \sigma^2 \).

**Recap:**
- \( T(X) = \sum_{i=1}^{n} X_i^2 \) contains all the information about \( \sigma^2 \).
- The distribution of \( S(X) \) does not depend on \( \sigma^2 \).
- Might we conclude that \( T(X) \perp \perp S(X) \)?
- I used R to generate \( B = 1000 \) draws from the bivariate distribution of \( (T(X), S(X)) \), when \( n = 10 \) and \( \sigma^2 = 100 \); see Figure 6.1.
Remark: Finding ancillary statistics is easy when you are dealing with location or scale families.

**Location-invariance:** For any $c \in \mathbb{R}$, suppose the statistic $S(X)$ satisfies

$$S(x_1 + c, x_2 + c, ..., x_n + c) = S(x_1, x_2, ..., x_n)$$

for all $x \in X$. We say that $S(X)$ is a **location-invariant statistic**. In other words, the value of $S(x)$ is unaffected by location shifts.

**Result:** Suppose $X_1, X_2, ..., X_n$ are iid from

$$f_X(x | \mu) = f_Z(x - \mu),$$

a location family with standard pdf $f_Z(\cdot)$ and location parameter $-\infty < \mu < \infty$. If $S(X)$ is location invariant, then it is ancillary.

**Proof.** Define $W_i = X_i - \mu$, for $i = 1, 2, ..., n$. We perform an $n$-variate transformation to find the distribution of $W = (W_1, W_2, ..., W_n)$. The inverse transformation is described by

$$x_i = g_i^{-1}(w_1, w_2, ..., w_n) = w_i + \mu,$$
for $i = 1, 2, \ldots, n$. It is easy to see that the Jacobian of the inverse transformation is 1 and therefore

$$f_W(w) = f_X(w_1 + \mu, w_2 + \mu, \ldots, w_n + \mu)$$

$$= \prod_{i=1}^{n} f_X(w_i + \mu)$$

$$= \prod_{i=1}^{n} f_Z(w_i + \mu - \mu) = \prod_{i=1}^{n} f_Z(w_i),$$

which does not depend on $\mu$. Because $S(X)$ is location invariant, we know

$$S(X) = S(X_1, X_2, \ldots, X_n)$$

$$= S(W_1 + \mu, W_2 + \mu, \ldots, W_n + \mu)$$

$$= S(W_1, W_2, \ldots, W_n)$$

$$= S(W).$$

Because the distribution of $W$ does not depend on $\mu$, the distribution of the statistic $S(W)$ cannot depend on $\mu$ either. But $S(W) = S(X)$, so we are done. □

**Examples:** Each of the following is a location-invariant statistic (and hence is ancillary when sampling from a location family):

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

**Note:** Above $M$ denotes the sample median of $X_1, X_2, \ldots, X_n$.

**Example 6.12.** Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and $\sigma_0^2$ is known. Show that the sample variance $S^2$ is ancillary.

**Proof.** First note that

$$f_X(x|\mu) = \frac{1}{\sqrt{2\pi \sigma_0}} e^{- (x-\mu)^2 / 2\sigma_0^2} I(x \in \mathbb{R}) = f_Z(x - \mu),$$

where

$$f_Z(z) = \frac{1}{\sqrt{2\pi \sigma_0}} e^{- z^2 / 2\sigma_0^2} I(z \in \mathbb{R}),$$

the $N(0, \sigma_0^2)$ pdf. Therefore, the $N(\mu, \sigma_0^2)$ family is a location family. We now show that $S(X) = S^2$ is location invariant. Let $W_i = X_i + c$, for $i = 1, 2, \ldots, n$. Clearly, $W = \overline{X} + c$ and

$$S(W) = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \overline{W})^2$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} [(X_i + c) - (\overline{X} + c)]^2$$

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

This shows that $S(X) = S^2$ is location invariant and hence is ancillary.
Remark: The preceding argument only shows that the distribution of $S^2$ does not depend on $\mu$. However, in this example, it is easy to find the distribution of $S^2$ directly. Recall that

$$\frac{(n - 1)S^2}{\sigma_0^2} \sim \chi_{n-1}^2 \overset{d}{=} \text{gamma} \left( \frac{n - 1}{2}, 2 \right) \implies S^2 \sim \text{gamma} \left( \frac{n - 1}{2}, \frac{2\sigma_0^2}{n-1} \right),$$

which, of course, does not depend on $\mu$.

Scale-invariance: For any $d > 0$, suppose the statistic $S(X)$ satisfies

$$S(dx_1, dx_2, ..., dx_n) = S(x_1, x_2, ..., x_n)$$

for all $x \in X$. We say that $S(X)$ is a scale-invariant statistic. In other words, the value of $S(x)$ is unaffected by changes in scale.

Result: Suppose $X_1, X_2, ..., X_n$ are iid from

$$f_X(x|\sigma) = \frac{1}{\sigma} f_Z \left( \frac{x}{\sigma} \right),$$

a scale family with standard pdf $f_Z(\cdot)$ and scale parameter $\sigma^2 > 0$. If $S(X)$ is scale invariant, then it is ancillary.

Proof. Define $W_i = X_i/\sigma$, for $i = 1, 2, ..., n$. We perform an $n$-variate transformation to find the distribution of $W = (W_1, W_2, ..., W_n)$. The inverse transformation is described by

$$x_i = g_i^{-1}(w_1, w_2, ..., w_n) = \sigma w_i,$$

for $i = 1, 2, ..., n$. It is easy to see that the Jacobian of the inverse transformation is $\sigma^n$ and therefore

$$f_W(w) = f_X(\sigma w_1, \sigma w_2, ..., \sigma w_n) \times \sigma^n$$

$$= \sigma^n \prod_{i=1}^n f_X(\sigma w_i)$$

$$= \sigma^n \prod_{i=1}^n \frac{1}{\sigma} f_Z \left( \frac{\sigma w_i}{\sigma} \right) = \prod_{i=1}^n f_Z(w_i),$$

which does not depend on $\sigma$. Because $S(X)$ is scale invariant, we know

$$S(X) = S(X_1, X_2, ..., X_n)$$

$$= S(\sigma W_1, \sigma W_2, ..., \sigma W_n)$$

$$= S(W_1, W_2, ..., W_n)$$

$$= S(W).$$

Because the distribution of $W$ does not depend on $\sigma$, the distribution of the statistic $S(W)$ cannot depend on $\sigma$ either. But $S(W) = S(X)$, so we are done. □

Examples: Each of the following is a scale-invariant statistic (and hence is ancillary when sampling from a scale family):

$$S(X) = \frac{S}{\overline{X}}, \quad S(X) = \frac{X_{(n)}}{\overline{X}_{(1)}}, \quad S(X) = \frac{\sum_{i=1}^k X_i^2}{\sum_{i=1}^n X_i^2}. $$
Example 6.13. Suppose \( X_1, X_2, \ldots, X_n \) is an iid sample from

\[
f_X(x|\sigma) = \frac{1}{2\sigma} e^{-|x|/\sigma} I(x \in \mathbb{R}).
\]

Show that

\[
S(X) = \frac{\sum_{i=1}^{k} |X_i|}{\sum_{i=1}^{n} |X_i|}
\]

is ancillary.

Proof. First note that

\[
f_X(x|\sigma) = \frac{1}{\sigma} f_Z \left( \frac{x}{\sigma} \right),
\]

where

\[
f_Z(z) = \frac{1}{2} e^{-|z|} I(z \in \mathbb{R}),
\]

a standard LaPlace pdf. Therefore, \( \{f_X(x|\sigma), \sigma > 0\} \) is a scale family. We now show that \( S(X) \) is scale invariant. For \( d > 0 \), let \( W_i = dX_i \), for \( i = 1, 2, \ldots, n \). We have

\[
S(W) = \frac{\sum_{i=1}^{k} |W_i|}{\sum_{i=1}^{n} |W_i|}
\]

\[
= \frac{\sum_{i=1}^{k} |dX_i|}{\sum_{i=1}^{n} |dX_i|}
\]

\[
= \frac{d \sum_{i=1}^{k} |X_i|}{d \sum_{i=1}^{n} |X_i|} = S(X).
\]

This shows that \( S(X) \) is scale invariant and hence is ancillary.

Remark: The preceding argument only shows that the distribution of \( S(X) \) does not depend on \( \sigma \). It can be shown that

\[
S(X) = \frac{\sum_{i=1}^{k} |X_i|}{\sum_{i=1}^{n} |X_i|} \sim \text{beta}(k, n),
\]

which, of course, does not depend on \( \sigma \).

Remark: It is straightforward to extend our previous discussions to location-scale families.

6.2.4 Sufficient, ancillary, and complete statistics

Definition: Let \( \{f_T(t|\theta); \theta \in \Theta\} \) be a family of pdfs (or pmfs) for a statistic \( T = T(X) \). We say that this family is a complete family if the following condition holds:

\[
E_\theta[g(T)] = 0 \quad \forall \theta \in \Theta \implies P_\theta(g(T) = 0) = 1 \quad \forall \theta \in \Theta;
\]

i.e., \( g(T) = 0 \) almost surely for all \( \theta \in \Theta \). We call \( T = T(X) \) a complete statistic.

Remark: This condition basically says that the only function of \( T \) that is an unbiased estimator of zero is the function that is zero itself (with probability 1).
Example 6.14. Suppose $X_1, X_2, \ldots, X_n$ are iid Bernoulli($\theta$), where $0 < \theta < 1$. Show that

$$T = T(X) = \sum_{i=1}^{n} X_i$$

is a complete statistic.

**Proof.** We know that $T \sim b(n, \theta)$, so it suffices to show that this family of distributions is a complete family. Suppose $E_{\theta}[g(T)] = 0$ $\forall \theta \in (0, 1)$. It suffices to show that $P_{\theta}(g(T) = 0) = 1$ for all $\theta \in (0, 1)$. Note that

$$0 = E_{\theta}[g(T)] = \sum_{t=0}^{n} g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = (1-\theta)^n \sum_{t=0}^{n} g(t) \binom{n}{t} r^t,$$

where $r = \theta/(1-\theta)$. Because $(1-\theta)^n$ is never zero, it must be that

$$\sum_{t=0}^{n} g(t) \binom{n}{t} r^t = 0.$$

The LHS of this equation is a polynomial (in $r$) of degree $n$. The only way this polynomial can be zero for all $\theta \in (0, 1)$; i.e., for all $r > 0$, is for the coefficients

$$g(t) \binom{n}{t} = 0,$$

for $t = 0, 1, 2, \ldots, n$. Because $\binom{n}{t} \neq 0$, this can only happen when $g(t) = 0$, for $t = 0, 1, 2, \ldots, n$. We have shown that $P_{\theta}(g(T) = 0) = 1$ for all $\theta$. Therefore, $T(X)$ is complete. □

**Remark:** To show that a statistic $T = T(X)$ is not complete, all we have to do is find one nonzero function $g(T)$ that satisfies $E_{\theta}[g(T)] = 0$, for all $\theta$.

Example 6.15. Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\theta, \theta^2)$, where $\theta \in \Theta = (-\infty, 0) \cup (0, \infty)$. Putting

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

into exponential family form shows that

$$T = T(X) = \left( \sum_{i=1}^{n} X_i \right) / \left( \sum_{i=1}^{n} X_i^2 \right)$$

is sufficient. However, $T$ is not complete. Consider

$$g(T) = 2 \left( \sum_{i=1}^{n} X_i \right)^2 - (n+1) \sum_{i=1}^{n} X_i^2.$$
It is straightforward to show that
\[ E_\theta[g(T)] = E_\theta \left[ 2 \left( \sum_{i=1}^{n} X_i \right)^2 - (n + 1) \sum_{i=1}^{n} X_i^2 \right] = 0. \]

We have found a nonzero function \( g(T) \) that has zero expectation. Therefore \( T \) cannot be complete.

**Theorem 6.2.24 (Basu’s Theorem).** Suppose \( T \) is a sufficient statistic. If \( T \) is also complete, then \( T \) is independent of every ancillary statistic \( S \).

**Proof.** Suppose \( S \) is ancillary. Let \( \phi \) and \( \psi \) be any functions. Using iterated expectation,
\[
E_\theta[\phi(S)\psi(T)] = E_\theta \{ E[\phi(S)\psi(T)|T] \} = E_\theta \{ \psi(T)E[\phi(S)|T] \},
\]
the last step following because once we condition on \( T = t \); i.e., we write \( E[\phi(S)\psi(t)|T = t] \), then \( \psi(t) \) is constant. Now, consider the quantity
\[
E_\theta \{ E[\phi(S)|T] \} = E_\theta[\phi(S)] = k,
\]
where \( k \) is a constant free of \( \theta \) (because \( S \) is ancillary by assumption). Define
\[
g(T) = E[\phi(S)|T] - k.
\]
From our argument above, we have
\[
E_\theta[g(T)] = E_\theta \{ E[\phi(S)|T] - k \} = E_\theta[\phi(S)] - k = 0
\]
for all \( \theta \). However, because \( T \) is complete by assumption, we know that
\[
g(T) \xrightarrow{a.s.} 0 \forall \theta \implies E[\phi(S)|T] \xrightarrow{a.s.} k \forall \theta.
\]
Because \( T \) is sufficient by assumption, we know that \( E[\phi(S)|T] \) does not depend on \( \theta \) either. From Equation (6.1), we have
\[
E_\theta[\phi(S)\psi(T)] = E_\theta \{ \psi(T)E[\phi(S)|T] \} = kE_\theta[\psi(T)] = E_\theta[\phi(S)]E_\theta[\psi(T)].
\]
Because \( E_\theta[\phi(S)\psi(T)] = E_\theta[\phi(S)]E_\theta[\psi(T)] \) holds for all functions (\( \phi \) and \( \psi \) were arbitrarily chosen), then equality holds when
\[
\phi(S) = I(S \leq s) \quad \psi(T) = I(T \leq t),
\]
for \( s, t \in \mathbb{R} \). Using this choice of \( \phi \) and \( \psi \), the joint cdf of \((S,T)\)
\[
F_{T,S}(t,s) = P_\theta(S \leq s, T \leq t) = E_\theta[\phi(S)\psi(T)] = E_\theta[\phi(S)]E_\theta[\psi(T)] = P_\theta(S \leq s)P_\theta(T \leq t) = F_S(s)F_T(t).
\]
We have shown that the joint cdf of \((S,T)\) factors into the product of the marginal cdfs. Because \( s \) and \( t \) are arbitrary, we are done. \( \square \)
Example 6.16. Suppose that $X_1, X_2, \ldots, X_n$ are iid $U(0, \theta)$, where $\theta > 0$. Show that $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent.

Proof. We will show that

- $T(X) = X_{(n)}$ is complete and sufficient.
- $S(X) = X_{(1)}/X_{(n)}$ is ancillary.

The result will then follow from Basu’s Theorem. First, note that

$$f_{X}(x | \theta) = \frac{1}{\theta} I(0 < x < \theta)$$

$$= \frac{1}{\theta} f_{Z} \left( \frac{x}{\theta} \right),$$

where $f_{Z}(z) = I(0 < z < 1)$ is the standard uniform pdf. Therefore, the $U(0, \theta)$ family is a scale family. We now show that $S(X)$ is scale invariant. For $d > 0$, let $W_i = dX_i$, for $i = 1, 2, \ldots, n$. We have

$$S(W) = \frac{W_{(1)}}{W_{(n)}}$$

$$= \frac{dX_{(1)}}{dX_{(n)}}$$

$$= \frac{X_{(1)}}{X_{(n)}} = S(X).$$

This shows that $S(X)$ is scale invariant and hence is ancillary.

We have already shown that $T = T(X) = X_{(n)}$ is sufficient; see Example 6.5 (notes). We now show $T$ is complete. We first find the distribution of $T$. The pdf of $T$, the maximum order statistic, is given by

$$f_{T}(t) = n f_{X}(t) [F_{X}(t)]^{n-1}$$

$$= n \frac{1}{\theta} I(0 < t < \theta) \left( \frac{t}{\theta} \right)^{n-1}$$

$$= \frac{n t^{n-1}}{\theta^n} I(0 < t < \theta).$$

Suppose $E_{\theta}[g(T)] = 0$ for all $\theta > 0$. This implies that

$$\int_{0}^{\theta} g(t) \frac{n t^{n-1}}{\theta^n} dt = 0 \quad \forall \theta > 0 \quad \implies \quad \int_{0}^{\theta} g(t) t^{n-1} dt = 0 \quad \forall \theta > 0$$

$$\implies \quad \frac{d}{d\theta} \int_{0}^{\theta} g(t) t^{n-1} dt = 0 \quad \forall \theta > 0$$

$$\implies \quad g(\theta) \theta^{n-1} = 0 \quad \forall \theta > 0,$$

the last step following from the Fundamental Theorem of Calculus, provided that $g$ is Riemann-integrable. Because $\theta^{n-1} \neq 0$, it must be true that $g(\theta) = 0$ for all $\theta > 0$. We have
therefore shown that the only function \( g \) satisfying \( E_\theta[g(T)] = 0 \) for all \( \theta > 0 \) is the function that is itself zero; i.e., we have shown

\[
P_\theta(g(T) = 0) = 1, \text{ for all } \theta > 0.
\]

Therefore \( T = T(X) = X_{(n)} \) is complete. \( \square \)

**Remark:** Our completeness argument in Example 6.16 is not entirely convincing. We have basically established that

\[
E_\theta[g(T)] = 0 \quad \forall \theta > 0 \implies P_\theta(g(T) = 0) = 1 \quad \forall \theta > 0
\]

for the class of functions \( g \) which are Riemann-integrable. There are many functions \( g \) that are not Riemann-integrable. CB note that “this distinction is not of concern.” This is another way of saying that the authors do not want to present completeness from a more general point of view (for good reason; this would involve a heavy dose of measure theory).

**Extension:** Suppose that, in Example 6.16, I asked you to find

\[
E \left( \frac{X_{(1)}}{X_{(n)}} \right).
\]

At first glance, this appears to be an extremely challenging expectation to calculate. From first principles, we could find the joint distribution of \((X_{(1)}, X_{(n)})\) and then calculate the first moment of the ratio. Another approach is to use Basu’s Theorem. Note that

\[
E(X_{(1)}) = E \left( X_{(n)} \frac{X_{(1)}}{X_{(n)}} \right)
= E(X_{(n)})E \left( \frac{X_{(1)}}{X_{(n)}} \right),
\]

the last step following because \( X_{(n)} \) and \( X_{(1)}/X_{(n)} \) are independent. Therefore, we can calculate the desired expectation by instead calculating \( E(X_{(1)}) \) and \( E(X_{(n)}) \). These are easier to calculate:

\[
E(X_{(1)}) = \frac{\theta}{n + 1}
\]

\[
E(X_{(n)}) = \left( \frac{n}{n + 1} \right) \theta.
\]

Therefore, we have

\[
\frac{\theta}{n + 1} = \left( \frac{n}{n + 1} \right) \theta E \left( \frac{X_{(1)}}{X_{(n)}} \right) \implies E \left( \frac{X_{(1)}}{X_{(n)}} \right) = \frac{1}{n}.
\]

It makes sense that this expectation would not depend on \( \theta \); recall that \( S(X) = X_{(1)}/X_{(n)} \) is ancillary.

**Completeness in the exponential family:**

**Recall:** Suppose \( X_1, X_2, ..., X_n \) are iid from the exponential family

\[
f_X(x|\theta) = h(x)c(\theta)\exp \left\{ \sum_{i=1}^{k} w_i(\theta)t_i(x) \right\},
\]
where $\theta = (\theta_1, \theta_2, ..., \theta_d)$, $d \leq k$. In Theorem 6.2.10, we showed that

$$T = T(X) = \left( \begin{array}{c} \sum_{j=1}^{n} t_1(X_j) \\ \sum_{j=1}^{n} t_2(X_j) \\ \vdots \\ \sum_{j=1}^{n} t_k(X_j) \end{array} \right)$$

is a sufficient statistic.

**New result** (Theorem 6.2.25): In the exponential family, the statistic $T = T(X)$ is complete if the natural parameter space

$$\{ \eta = (\eta_1, \eta_2, ..., \eta_k) : \eta_k = w_i(\theta); \theta \in \Theta \}$$

contains an open set in $\mathbb{R}^k$. For the most part, this means:

- $T = T(X)$ is complete if $d = k$ (full exponential family)
- $T = T(X)$ is not complete if $d < k$ (curved exponential family).

**Example 6.17.** Suppose that $X_1, X_2, ..., X_n$ is an iid sample from a gamma($\alpha, 1/\alpha^2$) distribution. The pdf of $X$ is

$$f_X(x|\alpha) = \frac{1}{\Gamma(\alpha)} \left( \frac{1}{\alpha^2} \right)^\alpha x^{\alpha-1} e^{-x/(1/\alpha^2)} I(x > 0)$$

$$= \frac{1}{\Gamma(\alpha)} \frac{\alpha^{2\alpha}}{x} e^{\alpha \ln x} e^{-\alpha^2 x}$$

$$= \frac{1}{\Gamma(\alpha)} \frac{\alpha^{2\alpha}}{x} \exp(\alpha \ln x - \alpha^2 x)$$

$$= h(x)c(\alpha) \exp\{w_1(\alpha)t_1(x) + w_2(\alpha)t_2(x)\},$$

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 $t_2(x) = x$. Theorem 6.2.10 tells us that

$$T = T(X) = \left( \begin{array}{c} \sum_{i=1}^{n} \ln X_i \\ \sum_{i=1}^{n} X_i \end{array} \right)$$

is a sufficient statistic. However, Theorem 6.2.25 tells us that $T$ is not complete because $\{f_X(x|\alpha), \alpha > 0\}$ is an exponential family with $d = 1$ and $k = 2$. Note also that

$$\{ \eta = (\eta_1, \eta_2) : (\alpha, -\alpha^2); \alpha > 0 \}$$

is a half-parabola (which opens downward); this set does not contain an open set in $\mathbb{R}^2$. 

PAGE 192
Example 6.18. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Prove that $\overline{X} \perp \perp S^2$.

Proof. We use Basu’s Theorem, but we have to use it carefully. Fix $\sigma^2 = \sigma_0^2$ and consider first the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily. The pdf of $X \sim \mathcal{N}(\mu, \sigma_0^2)$ is

$$f_X(x|\mu) = \frac{1}{\sqrt{2\pi} \sigma_0} e^{-(x-\mu)^2/2\sigma_0^2} I(x \in \mathbb{R})$$

$$= \frac{I(x \in \mathbb{R})}{\sqrt{2\pi} \sigma_0} e^{-x^2/2\sigma_0^2} e^{-\mu^2/2\sigma_0^2} e^{(\mu/\sigma_0^2)x}$$

$$= h(x) c(\mu) \exp\{w_1(\mu) t_1(x)\}.$$  

Theorem 6.2.10 tells us that $T = T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic. Because $d = k = 1$ (remember, this is for the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily), Theorem 6.2.25 tells us that $T$ is complete. In Example 6.12 (notes), we have already showed that

- the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily is a location family
- $S(X) = S^2$ is location invariant and hence ancillary for this subfamily.

Therefore, by Basu’s Theorem, we have proven that, in the $\mathcal{N}(\mu, \sigma_0^2)$ subfamily,

$$\sum_{i=1}^n X_i \perp \perp S^2 \implies \overline{X} \perp \perp S^2,$$

the last implication being true because $\overline{X}$ is a function of $T = T(X) = \sum_{i=1}^n X_i$ and functions of independent statistics are independent. Finally, because we fixed $\sigma^2 = \sigma_0^2$ arbitrarily, this same argument holds for all $\sigma_0^2$ fixed. Therefore, this independence result holds for any choice of $\sigma^2$ and hence for the full $\mathcal{N}(\mu, \sigma^2)$ family. $\square$

Remark: It is important to see that in the preceding proof, we cannot work directly with the $\mathcal{N}(\mu, \sigma^2)$ family and claim that

- $T(X) = \sum_{i=1}^n X_i$ is complete and sufficient
- $S(X) = S^2$ is ancillary

for this family. In fact, neither statement is true in the full family.

Remark: Outside the exponential family, Basu’s Theorem can be useful in showing that a sufficient statistic $T(X)$ is not complete.

Basu’s Theorem (Contrapositive version): Suppose $T(X)$ is sufficient and $S(X)$ is ancillary. If $T(X)$ and $S(X)$ are not independent, then $T(X)$ is not complete.
Example 6.19. Suppose that $X_1, X_2, ..., X_n$ is an iid sample from

$$f_X(x|\theta) = \frac{1}{\pi[1 + (x - \theta)^2]} I(x \in \mathbb{R}),$$

where $-\infty < \theta < \infty$. It is easy to see that \{f_X(x|\theta) : -\infty < \theta < \infty\} is a location family; i.e., $f_X(x|\theta) = f_Z(x - \theta)$, where

$$f_Z(z) = \frac{1}{\pi(1 + z^2)} I(z \in \mathbb{R})$$

is the standard Cauchy pdf. We now prove the sample range $S(X) = X_{(n)} - X_{(1)}$ is location invariant. Let $W_i = X_i + c$, for $i = 1, 2, ..., n$, and note

$$S(W) = W_{(n)} - W_{(1)} = (X_{(n)} + c) - (X_{(1)} + c) = X_{(n)} - X_{(1)} = S(X).$$

This shows that $S(X)$ is ancillary in this family. Finally, we know from Example 6.4 (notes) that the order statistics

$$T = T(X) = (X_{(1)}, X_{(2)}, ..., X_{(n)})$$

are sufficient for this family (in fact, $T$ is minimal sufficient; see Exercise 6.9, CB, pp 301). However, clearly $S(X)$ and $T(X)$ are not independent; e.g., if you know $T(x)$, you can calculate $S(x)$. By Basu’s Theorem (the contrapositive version), we know that $T(X)$ cannot be complete.

**Theorem 6.2.28.** Suppose that $T(X)$ is sufficient. If $T(X)$ is complete, then $T(X)$ is minimal sufficient.

**Remark:** Example 6.19 shows that the converse to Theorem 6.2.28 is not true; i.e.,

$$T(X)$$

minimal sufficient $\not\Rightarrow$ $T(X)$ complete.

Example 6.7 provides another counterexample. We showed that if $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(\theta, \theta + 1)$, then $T = T(X) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. However, $T$ cannot be complete because $T$ and the sample range $X_{(n)} - X_{(1)}$ (which is location invariant and hence ancillary in this model) are not independent. This implies that there exists a nonzero function $g(T)$ that has zero expectation for all $\theta \in \mathbb{R}$. In fact, it is easy to show that

$$E_{\theta}(X_{(n)} - X_{(1)}) = \frac{n - 1}{n + 1}.$$

Therefore,

$$g(T) = X_{(n)} - X_{(1)} - \frac{n - 1}{n + 1}$$

satisfies $E_{\theta}[g(T)] = 0$ for all $\theta$. 
7 Point Estimation

Complementary reading: Chapter 7 (CB).

7.1 Introduction

Remark: We will approach “the point estimation problem” from the following point of view. We have a parametric model for \( X = (X_1, X_2, ..., X_n) \):

\[ X \sim f_X(x|\theta), \text{ where } \theta \in \Theta \subseteq \mathbb{R}^k, \]

and the model parameter \( \theta = (\theta_1, \theta_2, ..., \theta_k) \) is unknown. We will assume that \( \theta \) is fixed (except when we discuss Bayesian estimation). Possible goals include

1. Estimating \( \theta \)
2. Estimating a function of \( \theta \), say \( \tau(\theta) \), where \( \tau : \mathbb{R}^k \rightarrow \mathbb{R}^q, q \leq k \) (often, \( q = 1 \); i.e., \( \tau(\theta) \) is a scalar parameter).

Remark: For most of the situations we will encounter in this course, the random vector \( X \) will consist of \( X_1, X_2, ..., X_n \), an iid sample from the population \( f_X(x|\theta) \). However, our discussion is also relevant when the independence assumption is relaxed, the identically distributed assumption is relaxed, or both.

Definition: A point estimator

\[ W(X) = W(X_1, X_2, ..., X_n) \]

is any function of the sample \( X \). Therefore, any statistic is a point estimator. We call \( W(x) = W(x_1, x_2, ..., x_n) \) a point estimate. \( W(x) \) is a realization of \( W(X) \).

Preview: This chapter is split into two parts. In this first part (Section 7.2), we present different approaches of finding point estimators. These approaches are:

- Section 7.2.1: Method of Moments (MOM)
- Section 7.2.2: Maximum Likelihood Estimation (MLE)
- Section 7.2.3: Bayesian Estimation
- Section 7.2.4: EM Algorithm (we will skip).

The second part (Section 7.3) focuses on evaluating point estimators; e.g., which estimators are good/bad? What constitutes a “good” estimator? Is it possible to find the best one? For that matter, how should we even define “best?”
7.2 Methods of Finding Estimators

7.2.1 Method of moments

**Strategy:** Suppose $X \sim f_X(x; \theta)$, where $\theta = (\theta_1, \theta_2, ..., \theta_k) \in \Theta \subseteq \mathbb{R}^k$. The method of moments (MOM) approach says to equate the first $k$ sample moments to the first $k$ population moments and then to solve for $\theta$.

**Recall:** The $j$th sample moment (uncentered) is given by

$$m'_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j.$$

If $X_1, X_2, ..., X_n$ is an iid sample, the $j$th population moment (uncentered) is

$$\mu'_j = E(X^j).$$

**Intuition:** The first $k$ sample moments depend on the sample $X$. The first $k$ population moments will generally depend on $\theta = (\theta_1, \theta_2, ..., \theta_k)$. Therefore, the system of equations

$$\begin{align*}
    m'_1 &= E(X) \\
    m'_2 &= E(X^2) \\
    &\vdots \\
    m'_k &= E(X^k)
\end{align*}$$

can (at least in theory) be solved for $\theta_1, \theta_2, ..., \theta_k$. A solution to this system of equations is called a method of moments (MOM) estimator.

**Example 7.1.** Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(0, \theta)$, where $\theta > 0$. The first sample moment is

$$m'_1 = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}.$$ 

The first population moment is

$$\mu'_1 = E(X) = \frac{\theta}{2}.$$ 

We set these moments equal to each other; i.e.,

$$\overline{X} = \frac{\theta}{2}$$

and solve for $\theta$. The solution

$$\hat{\theta} = 2\overline{X}$$

is a method of moments estimator for $\theta$.

**Example 7.2.** Suppose that $X_1, X_2, ..., X_n$ are iid $\mathcal{U}(-\theta, \theta)$, where $\theta > 0$. For this population, $E(X) = 0$ so this will not help us. Moving to second moments, we have

$$m'_2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$
and

$$\mu' = E(X^2) = \var(X) = \frac{\theta^2}{3}.$$ 

Therefore, we can set

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \overset{\text{set}}{=} \frac{\theta^2}{3},$$

and solve for $\theta$. The solution

$$\hat{\theta} = + \sqrt{\frac{3}{n} \sum_{i=1}^{n} X_i^2}$$

is a method of moments estimator for $\theta$. We keep the positive solution because $\theta > 0$ (although, technically, the negative solution is still a MOM estimator).

**Example 7.3.** Suppose $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. The first two population moments are $E(X) = \mu$ and $E(X^2) = \var(X) + [E(X)]^2 = \sigma^2 + \mu^2$. Therefore, method of moments estimators for $\mu$ and $\sigma^2$ are found by solving

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \overset{\text{set}}{=} \sigma^2 + \mu^2.$$ 

We have $\hat{\mu} = \bar{X}$ and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$ 

Note that the method of moments estimator for $\sigma^2$ is not our “usual” sample variance (with denominator $n - 1$).

**Remarks:**

- I think of MOM estimation as a “quick and dirty” approach. All we are doing is matching moments. We are attempting to learn about a population $f_X(x|\theta)$ by using moments only.
- Sometimes MOM estimators have good finite-sample properties (e.g., unbiasedness, small variance, etc.). Sometimes they do not.
- MOM estimators generally do have desirable large-sample properties (e.g., large-sample normality, etc.) but are usually less (asymptotically) efficient than other estimators.
- MOM estimators can be nonsensical. In fact, sometimes MOM estimators fall outside the parameter space $\Theta$. For example, in linear models with random effects, variance components estimated via MOM can be negative.
7.2.2 Maximum likelihood estimation

**Note:** We first formally define a likelihood function; see also Section 6.3 (CB).

**Definition:** Suppose $X \sim f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$. Given that $X = x$ is observed, the function

$$L(\theta|x) = f_X(x|\theta)$$

is called the **likelihood function**.

**Note:** The likelihood function $L(\theta|x)$ is the same function as the joint pdf/pmf $f_X(x|\theta)$. The only difference is in how we interpret each one.

- The function $f_X(x|\theta)$ is a model that describes the random behavior of $X$ when $\theta$ is fixed.
- The function $L(\theta|x)$ is viewed as a function of $\theta$ with the data $X = x$ held fixed.

**Interpretation:** When $X$ is discrete,

$$L(\theta|x) = f_X(x|\theta) = P_\theta(X = x).$$

That is, when $X$ is discrete, we can interpret the likelihood function $L(\theta|x)$ literally as a joint probability.

- Suppose that $\theta_1$ and $\theta_2$ are two possible values of $\theta$. Suppose $X$ is discrete and

$$L(\theta_1|x) = P_{\theta_1}(X = x) > P_{\theta_2}(X = x) = L(\theta_2|x).$$

This suggests the sample $x$ is more likely to have occurred with $\theta = \theta_1$ rather than if $\theta = \theta_2$. Therefore, in the discrete case, we can interpret $L(\theta|x)$ as “the probability of the data $x$.”

- Of course, this interpretation of $L(\theta|x)$ is not appropriate when $X$ is continuous because $P_{\theta}(X = x) = 0$. However, this description is still used informally when describing the likelihood function with continuous data. An attempt to make this description mathematical is given on pp 290 (CB).

- Section 6.3 (CB) describes how the likelihood function $L(\theta|x)$ can be viewed as a **data reduction** device.

**Definition:** Any maximizer $\hat{\theta} = \hat{\theta}(x)$ of the likelihood function $L(\theta|x)$ is called a **maximum likelihood estimate**.

- With our previous interpretation, we can think of $\hat{\theta}$ as “the value of $\theta$ that maximizes the probability of the data $x$.”

We call $\hat{\theta}(X)$ a **maximum likelihood estimator** (MLE).
Remarks:

1. Finding the MLE $\hat{\theta}$ is essentially a maximization problem. The estimate $\hat{\theta}(x)$ must fall in the parameter space $\Theta$ because we are maximizing $L(\theta|x)$ over $\Theta$; i.e.,

$$\hat{\theta}(x) = \arg \max_{\theta \in \Theta} L(\theta|x).$$

There is no guarantee that an MLE $\hat{\theta}(x)$ will be unique (although it often is).

2. Under certain conditions (so-called “regularity conditions”), maximum likelihood estimators $\hat{\theta}(X)$ have very nice large-sample properties (Chapter 10, CB).

3. In most “real” problems, the likelihood function $L(\theta|x)$ must be maximized numerically to calculate $\hat{\theta}(x)$.

Example 7.4. Suppose $X_1, X_2, \ldots, X_n$ are iid $U[0, \theta]$, where $\theta > 0$. Find the MLE of $\theta$.

Solution. The likelihood function is

$$L(\theta|x) = \prod_{i=1}^{n} \frac{1}{\theta} I(0 \leq x_i \leq \theta)$$

$$= \frac{1}{\theta^n} I(x_{(n)} \leq \theta) \prod_{i=1}^{n} I(x_i \geq 0).$$

view this as a function of $\theta$ with $x$ fixed

Note that

- For $\theta \geq x_{(n)}$, $L(\theta|x) = 1/\theta^n$, which decreases as $\theta$ increases.
- For $\theta < x_{(n)}$, $L(\theta|x) = 0$.

Clearly, the MLE of $\theta$ is $\hat{\theta} = X_{(n)}$.

Remark: Note that in this example, we “closed the endpoints” on the support of $X$; i.e., the pdf of $X$ is

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

Mathematically, this model is no different than had we “opened the endpoints.” However, if we used open endpoints, note that

$$x_{(n)} < \arg \max_{\theta > 0} L(\theta|x) < x_{(n)} + \epsilon$$

for all $\epsilon > 0$, and therefore the maximizer of $L(\theta|x)$; i.e., the MLE, would not exist.

Curiosity: In this uniform example, we derived the MOM estimator to be $\hat{\theta} = 2X$ in Example 7.1. The MLE is $\hat{\theta} = X_{(n)}$. Which estimator is “better?”
Note: In general, when the likelihood function \( L(\theta|x) \) is a differentiable function of \( \theta \), we can use calculus to maximize \( L(\theta|x) \). If an MLE \( \hat{\theta} \) exists, it must satisfy
\[
\frac{\partial}{\partial \theta_j} L(\hat{\theta}|x) = 0, \quad j = 1, 2, \ldots, k.
\]
Of course, second-order conditions must be verified to ensure that \( \hat{\theta} \) is a maximizer (and not a minimizer or some other value).

Example 7.5. Suppose that \( X_1, X_2, \ldots, X_n \) are iid \( \mathcal{N}(\theta, 1) \), where \(-\infty < \theta < \infty\). The likelihood function is
\[
L(\theta|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}} = \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i-\theta)^2}.
\]
The derivative
\[
\frac{\partial}{\partial \theta} L(\theta|x) = \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i-\theta)^2} \sum_{i=1}^{n} (x_i - \theta) \stackrel{\text{set}}{=} 0
\]
\[
\Rightarrow \sum_{i=1}^{n} (x_i - \theta) = 0.
\]
Therefore, \( \hat{\theta} = \bar{x} \) is a first-order critical point of \( L(\theta|x) \). Is \( \hat{\theta} = \bar{x} \) a maximizer? I calculated
\[
\frac{\partial^2}{\partial \theta^2} L(\theta|x) = \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i-\theta)^2} \left\{ \left[ \sum_{i=1}^{n} (x_i - \theta) \right]^2 - n \right\}.
\]
Because
\[
\frac{\partial^2}{\partial \theta^2} L(\theta|x) \bigg|_{\theta=\bar{x}} = -n \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i-\bar{x})^2} < 0,
\]
the function \( L(\theta|x) \) is concave down when \( \theta = \bar{x} \); i.e., \( \hat{\theta} = \bar{x} \) maximizes \( L(\theta|x) \). Therefore,
\[
\hat{\theta} = \hat{\theta}(X) = \bar{x}
\]
is the MLE of \( \theta \).

Illustration: Under the \( \mathcal{N}(\theta, 1) \) model assumption, I graphed in Figure 7.1 the likelihood function \( L(\theta|x) \) after observing \( x_1 = 2.437, x_2 = 0.993, x_3 = 1.123, x_4 = 1.900, \) and \( x_5 = 3.794 \) (an iid sample of size \( n = 5 \)). The sample mean
\[
\bar{x} = 2.049
\]
is our ML estimate of \( \theta \) based on this sample \( x \).
Q: What if, in Example 7.5, we constrained the parameter space to be $\Theta_0 = \{\theta : \theta \geq 0\}$? What is the MLE over $\Theta_0$?
A: We simply maximize $L(\theta|x)$ over $\Theta_0$ instead. It is easy to see the restricted MLE is

$$\hat{\theta}^* = \hat{\theta}^*(X) = \begin{cases} \bar{X}, & \bar{X} \geq 0 \\ 0, & \bar{X} < 0. \end{cases}$$

Important: Suppose that $L(\theta|x)$ is a likelihood function. Then

$$\hat{\theta}(x) = \arg \max_{\theta \in \Theta} L(\theta|x) = \arg \max_{\theta \in \Theta} \ln L(\theta|x).$$

The function $\ln L(\theta|x)$ is called the log-likelihood function. Analytically, it is usually easier to work with $\ln L(\theta|x)$ than with the likelihood function directly. The equations

$$\frac{\partial}{\partial \theta_j} \ln L(\theta|x) = 0, \quad j = 1, 2, \ldots, k,$$

are called the score equations.
Example 7.6. Suppose \(X_1, X_2, \ldots, X_n\) are iid \(\mathcal{N}(\mu, \sigma^2)\), where \(-\infty < \mu < \infty\) and \(\sigma^2 > 0\); i.e., both parameters are unknown. Set \(\theta = (\mu, \sigma^2)\). The likelihood function is

\[
L(\theta | x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}
\]

\[
= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2}.
\]

The log-likelihood function is

\[
\ln L(\theta | x) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.
\]

The score equations are

\[
\frac{\partial}{\partial \mu} \ln L(\theta | x) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \quad \text{set} \quad 0
\]

\[
\frac{\partial}{\partial \sigma^2} \ln L(\theta | x) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \quad \text{set} \quad 0.
\]

Clearly \(\hat{\mu} = \bar{x}\) solves the first equation; inserting \(\hat{\mu} = \bar{x}\) into the second equation and solving for \(\hat{\sigma}^2\) gives \(\hat{\sigma}^2 = \frac{n}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2\). A first-order critical point is \((\bar{x}, \frac{n}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2)\).

Q: How can we verify this solution is a maximizer?

A: In general, for a \(k\)-dimensional maximization problem, we can calculate the Hessian matrix

\[
H = \frac{\partial^2}{\partial \theta \partial \theta} \ln L(\theta | x),
\]

a \(k \times k\) matrix of second-order partial derivatives, and show this matrix is negative definite when we evaluate it at the first-order critical point \(\hat{\theta}\). This is a sufficient condition. Recall a \(k \times k\) matrix \(H\) is negative definite if \(a'Ha < 0\) for all \(a \in \mathbb{R}^k, a \neq 0\).

For the \(\mathcal{N}(\mu, \sigma^2)\) example, I calculated

\[
H = \begin{pmatrix}
-\frac{n}{\sigma^2} & -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) \\
-\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) & \frac{n}{\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2
\end{pmatrix}.
\]

With \(a' = (a_1, a_2)\), it follows that

\[
a'Ha\big|_{\mu = \hat{\mu}, \sigma^2 = \hat{\sigma}^2} = -\frac{na_2^2}{\sigma^2} < 0.
\]

This shows that

\[
\hat{\theta}(X) = \begin{pmatrix}
\bar{X} \\
\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2
\end{pmatrix}
\]

is the MLE of \(\theta\) in the \(\mathcal{N}(\mu, \sigma^2)\) model.
Exercise: Find the MLEs of $\mu$ and $\sigma^2$ in the respective sub-families:

- $\mathcal{N}(\mu, \sigma_0^2)$, where $\sigma_0^2$ is known
- $\mathcal{N}(\mu_0, \sigma^2)$, where $\mu_0$ is known.

Example 7.7. ML estimation under parameter constraints. Suppose $X_1, X_2$ are independent random variables where

$$X_1 \sim b(n_1, p_1)$$
$$X_2 \sim b(n_2, p_2),$$

where $0 < p_1 < 1$ and $0 < p_2 < 1$. The likelihood function of $\theta = (p_1, p_2)$ is

$$L(\theta|x_1, x_2) = f_{X_1}(x_1|p_1)f_{X_2}(x_2|p_2)$$
$$= \binom{n_1}{x_1}p_1^{x_1}(1-p_1)^{n_1-x_1}\binom{n_2}{x_2}p_2^{x_2}(1-p_2)^{n_2-x_2}.$$

The log-likelihood function is

$$\ln L(\theta|x_1, x_2) = c + x_1 \ln p_1 + (n_1 - x_1) \ln(1-p_1) + x_2 \ln p_2 + (n_2 - x_2) \ln(1-p_2),$$

where $c = \ln \binom{n_1}{x_1} + \ln \binom{n_2}{x_2}$ is free of $\theta$. Over

$$\Theta = \{\theta = (p_1, p_2) : 0 < p_1 < 1, 0 < p_2 < 1\},$$

it is easy to show that $\ln L(\theta|x_1, x_2)$ is maximized at

$$\hat{\theta} = \hat{\theta}(X_1, X_2) = \left( \begin{array}{c} \widehat{p}_1 \\ \widehat{p}_2 \end{array} \right) = \left( \begin{array}{c} X_1 \\ n_1 \\ X_2 \\ n_2 \end{array} \right),$$

the vector of sample proportions. Because this is the maximizer over the entire parameter space $\Theta$, we call $\hat{\theta}$ the unrestricted MLE of $\theta$.

Q: How do we find the MLE of $\theta$ subject to the constraint that $p_1 = p_2$?
A: We would now like to maximize $\ln L(\theta|x_1, x_2)$ over

$$\Theta_0 = \{\theta = (p_1, p_2) : 0 < p_1 < 1, 0 < p_2 < 1, p_1 = p_2\}.$$
This system becomes
\[
\begin{align*}
\frac{x_1}{p_1} + \frac{n_1 - x_1}{1 - p_1} &= \lambda \\
\frac{x_2}{p_2} + \frac{n_2 - x_2}{1 - p_2} &= -\lambda \\
p_1 - p_2 &= 0.
\end{align*}
\]

Solving this system for \(p_1\) and \(p_2\), we get
\[
\hat{\theta}^* = \hat{\theta}^*(X_1, X_2) = \left( \frac{X_1 + X_2}{n_1 + n_2}, \frac{X_1 + X_2}{n_1 + n_2} \right).
\]

Because this is the maximizer over the subspace \(\Theta_0\), we call \(\hat{\theta}^*\) the restricted MLE; i.e., the MLE of \(\theta\) subject to the \(p_1 = p_2\) restriction.

**Discussion:** The parameter constraint \(p_1 = p_2\) might arise in a hypothesis test; e.g., \(H_0 : p_1 = p_2\) versus \(H_1 : p_1 \neq p_2\). If \(H_0\) is true, then we would expect \(\hat{\theta}\) and \(\hat{\theta}\) to be “close” and the ratio
\[
\lambda(x_1, x_2) = \frac{L(\hat{\theta}^*|x_1, x_2)}{L(\hat{\theta}|x_1, x_2)} \approx 1.
\]
The farther \(\hat{\theta}^*\) is from \(\hat{\theta}\), the smaller \(\lambda(x_1, x_2)\) becomes. Therefore, it would make sense to reject \(H_0\) when \(\lambda(x_1, x_2)\) is small. This is the idea behind likelihood ratio tests.

**Example 7.8. Logistic regression.** In practice, finding maximum likelihood estimates usually requires numerical methods. Suppose \(Y_1, Y_2, \ldots, Y_n\) are independent Bernoulli random variables; specifically, \(Y_i \sim Bernoulli(p_i)\), where
\[
\ln \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_i \iff p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}.
\]

In this model, the \(x_i\)’s are fixed constants. The likelihood function of \(\theta = (\beta_0, \beta_1)\) is
\[
L(\theta|y) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}
\]
\[
= \prod_{i=1}^{n} \left[ \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{y_i} \left[ 1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{1-y_i}.
\]

Taking logarithms and simplifying gives
\[
\ln L(\theta|y) = \sum_{i=1}^{n} \left[ y_i (\beta_0 + \beta_1 x_i) - \ln(1 + e^{\beta_0 + \beta_1 x_i}) \right].
\]

Closed-form expressions for the maximizers \(\hat{\beta}_0\) and \(\hat{\beta}_1\) do not exist except in very simple situations. Numerical methods are needed to maximize \(\ln L(\theta|y)\); e.g., iteratively re-weighted least squares (the default method in R’s `glm` function).
Theorem 7.2.10 (Invariance property of MLEs). Suppose \( \hat{\theta} \) is the MLE of \( \theta \). For any function \( \tau(\theta) \), the MLE of \( \tau(\theta) \) is \( \tau(\hat{\theta}) \).

Proof. For simplicity, suppose \( \theta \) is a scalar parameter and that \( \tau : \mathbb{R} \to \mathbb{R} \) is one-to-one (over \( \Theta \)). In this case,

\[
\eta = \tau(\theta) \iff \theta = \tau^{-1}(\eta).
\]

The likelihood function of interest is \( L^*(\eta) \). It suffices to show that \( L^*(\eta) \) is maximized when \( \eta = \tau(\hat{\theta}) \), where \( \hat{\theta} \) is the maximizer of \( L(\theta) \). For simplicity in notation, I drop emphasis of a likelihood function’s dependence on \( x \). Let \( \hat{\eta} \) be a maximizer of \( L^*(\eta) \). Then

\[
L^*(\hat{\eta}) = \sup_{\eta} L^*(\eta) = \sup_{\eta} L(\tau^{-1}(\eta)) = \sup_{\hat{\theta}} L(\hat{\theta}).
\]

Therefore, the maximizer \( \hat{\eta} \) satisfies \( \tau^{-1}(\hat{\eta}) = \hat{\theta} \). Because \( \tau \) is one-to-one, \( \hat{\eta} = \tau(\hat{\theta}) \). \( \square \)

Remark: Our proof assumes that \( \tau \) is a one-to-one function. However, Theorem 7.2.10 is true for any function; see pp 319-320 (CB).

Example 7.8 (continued). In the logistic regression model

\[
\ln \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_i \iff p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = \tau(\beta_0, \beta_1), \text{ say,}
\]

the MLE of \( p_i \) is

\[
\hat{p}_i = \tau(\hat{\beta}_0, \hat{\beta}_1) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}.
\]

Example 7.9. Suppose \( X_1, X_2, \ldots, X_n \) are iid exponential(\( \beta \)), where \( \beta > 0 \). The likelihood function is

\[
L(\beta | x) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum_{i=1}^{n} x_i/\beta}.
\]

The log-likelihood function is

\[
\ln L(\beta | x) = -n \ln \beta - \frac{\sum_{i=1}^{n} x_i}{\beta}
\]

The score equation becomes

\[
\frac{\partial}{\partial \beta} \ln L(\beta | x) = -n \beta^{-1} + \frac{\sum_{i=1}^{n} x_i}{\beta^2} \text{ set } = 0.
\]

Solving the score equation for \( \beta \) gives \( \hat{\beta} = \bar{x} \). It is easy to show that this value maximizes \( \ln L(\beta | x) \). Therefore,

\[
\hat{\beta} = \beta(X) = \bar{x}
\]

is the MLE of \( \beta \).
Applications of invariance: In Example 7.9, 

- $\bar{X}^2$ is the MLE of $\beta^2$
- $1/\bar{X}$ is the MLE of $1/\beta$
- For $t$ fixed, $e^{-t/\bar{X}}$ is the MLE of $S_X(t|\beta) = e^{-t/\beta}$, the survivor function of $X$ at $t$.

7.2.3 Bayesian estimation

Remark: Non-Bayesians think of inference in the following way:

Observe $X \sim f_X(x|\theta) \longrightarrow$ Use $x$ to make statement about $\theta$.

In this paradigm, the model parameter $\theta$ is fixed (and unknown). I have taken $\theta$ to be a scalar here for ease of exposition.

Bayesians do not consider the parameter $\theta$ to be fixed. They regard $\theta$ as random, having its own probability distribution. Therefore, Bayesians think of inference in this way:

Model $\theta \sim \pi(\theta) \longrightarrow$ Observe $X|\theta \sim f_X(x|\theta) \longrightarrow$ Update with $\pi(\theta|x)$.

The model for $\theta$ on the front end is called the prior distribution. The model on the back end is called the posterior distribution. The posterior distribution combines prior information (supplied through the prior model) and the observed data $x$. For a Bayesian, all inference flows from the posterior distribution.

Important: Here are the relevant probability distributions that arise in a Bayesian context. These are given “in order” as to how the Bayesian uses them. Continue to assume that $\theta$ is a scalar.

1. Prior distribution: $\theta \sim \pi(\theta)$. This distribution incorporates the information available about $\theta$ before any data are observed.

2. Conditional distribution: $X|\theta \sim f_X(x|\theta)$. This is the distribution of $X$, but now viewed conditionally on $\theta$:

$$f_X(x|\theta) = L(\theta|x) = \prod_{i=1}^{n} f_{X|\theta}(x_i|\theta).$$

Mathematically, the conditional distribution is the same as the likelihood function.

3. Joint distribution: This distribution describes how $X$ and $\theta$ vary jointly. From the definition of a conditional distribution,

$$f_{X,\theta}(x, \theta) = \underbrace{f_{X|\theta}(x|\theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\text{prior}}.$$
4. **Marginal distribution.** This describes how $X$ is distributed marginally. From the definition of a marginal distribution,

$$m_X(x) = \int_{\Theta} f_{X,\theta}(x, \theta) d\theta = \int_{\Theta} f_{X|\theta}(x|\theta) \pi(\theta) d\theta,$$

where $\Theta$ is the “support” of $\theta$ (remember, we are now treating $\theta$ as a random variable).

5. **Posterior distribution.** This is the Bayesian’s “updated” distribution of $\theta$, given that the data $X = x$ have been observed. From the definition of a conditional distribution,

$$\pi(\theta|x) = \frac{f_{X,\theta}(x, \theta)}{m_X(x)} = \frac{f_{X|\theta}(x|\theta) \pi(\theta)}{\int_{\Theta} f_{X|\theta}(x|\theta) \pi(\theta) d\theta}.$$

**Remark:** The process of starting with $\pi(\theta)$ and performing the necessary calculations to end up with $\pi(\theta|x)$ is informally known as “turning the Bayesian crank.” The distributions above can be viewed as steps in a “recipe” for posterior construction (i.e., start with the prior and the conditional, calculate the joint, calculate the marginal, calculate the posterior). We will see momentarily that not all steps are needed. In fact, in practice, computational techniques are used to essentially bypass Step 4 altogether. You can see that this might be desirable, especially if $\theta$ is a vector (and perhaps high-dimensional).

**Example 7.10.** Suppose that, conditional on $\theta$, $X_1, X_2, \ldots, X_n$ are iid Poisson($\theta$), where the prior distribution for $\theta \sim$ gamma($a, b$), $a, b$ known. We now turn the Bayesian crank.

1. **Prior distribution.**

$$\pi(\theta) = \frac{1}{\Gamma(a) b^a} \theta^{a-1} e^{-\theta/b} I(\theta > 0).$$

2. **Conditional distribution.** For $x_i = 0, 1, 2, \ldots$,

$$f_{X|\theta}(x|\theta) = \prod_{i=1}^n \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!}.$$

Recall that this is the same function as the likelihood function.

3. **Joint distribution.** For $x_i = 0, 1, 2, \ldots$, and $\theta > 0$,

$$f_{X,\theta}(x, \theta) = f_{X|\theta}(x|\theta) \pi(\theta) = \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} \frac{1}{\Gamma(a) b^a} \theta^{a-1} e^{-\theta/b} = \frac{1}{\prod_{i=1}^n x_i! \Gamma(a) b^a} \theta^{\sum_{i=1}^n x_i + a - 1} e^{-\theta/(n+1)}.$$
4. Marginal distribution. For \( x_i = 0, 1, 2, \ldots, \)

\[
m_X(x) = \int_\Omega f_{X, \theta}(x, \theta) d\theta
= \frac{1}{\prod_{i=1}^n x_i! \Gamma(a) b^a} \int_0^\infty \frac{\theta^{\sum_{i=1}^n x_i + a - 1} e^{-\theta/(n + \frac{1}{b})}}{\text{gamma}(a^*, b^*) \text{ kernel}} d\theta,
\]

where

\[
a^* = \sum_{i=1}^n x_i + a \quad \text{and} \quad b^* = \frac{1}{n + \frac{1}{b}}.
\]

Therefore,

\[
m_X(x) = \frac{1}{\prod_{i=1}^n x_i! \Gamma(a) b^a} \Gamma \left( \sum_{i=1}^n x_i + a \right) \left( \frac{1}{n + \frac{1}{b}} \right)^{\sum_{i=1}^n x_i + a}.
\]

5. Posterior distribution. For \( \theta > 0, \)

\[
\pi(\theta|x) = \frac{f_{X, \theta}(x, \theta)}{m_X(x)}
= \frac{1}{\prod_{i=1}^n x_i! \Gamma(a) b^a} \frac{\theta^{\sum_{i=1}^n x_i + a - 1} e^{-\theta/(n + \frac{1}{b})}}{\Gamma \left( \sum_{i=1}^n x_i + a \right) \left( \frac{1}{n + \frac{1}{b}} \right)^{\sum_{i=1}^n x_i + a}}
= \frac{1}{\Gamma \left( \sum_{i=1}^n x_i + a \right) \left( \frac{1}{n + \frac{1}{b}} \right)^{\sum_{i=1}^n x_i + a}} \theta^{\sum_{i=1}^n x_i + a - 1} e^{-\theta/(n + \frac{1}{b})},
\]

which we recognize as the gamma pdf with parameters

\[
a^* = \sum_{i=1}^n x_i + a
b^* = \frac{1}{n + \frac{1}{b}}.
\]

That is, the posterior distribution

\[
\theta|X = x \sim \text{gamma} \left( \sum_{i=1}^n x_i + a, \frac{1}{n + \frac{1}{b}} \right).
\]

**Remark:** Note that the shape and scale parameters of the posterior distribution \( \pi(\theta|x) \) depend on

- \( a \) and \( b \), the prior distribution parameters (i.e., the “hyperparameters”)
- the data \( x \) through the sufficient statistic \( t(x) = \sum_{i=1}^n x_i \).

In this sense, the posterior distribution combines information from the prior and the data.
Q: In general, which functional of \( \pi(\theta|x) \) should we use as a point estimator?

A: Answering this question technically would require us to discuss loss functions (see Section 7.3.4, CB). In practice, it is common to use one of

\[
\hat{\theta}_B = E(\theta|X = x) \quad \rightarrow \quad \text{posterior mean}
\]

\[
\hat{\theta}_B = \text{med}(\theta|X = x) \quad \rightarrow \quad \text{posterior median}
\]

\[
\hat{\theta}_B^* = \text{mode}(\theta|X = x) \quad \rightarrow \quad \text{posterior mode}
\]

Note that in Example 7.10 (the Poisson-gamma example), the posterior mean equals

\[
\hat{\theta}_B = E(\theta|X = x) = \frac{\sum_{i=1}^n x_i + a}{n + \frac{1}{b}}
\]

That is, the posterior mean is a weighted average of the sample mean \( \bar{x} \) and the prior mean \( ab \). Note also that as the sample size \( n \) increases, more weight is given to the data (through \( \bar{x} \)) and less weight is given to the prior (through the prior mean).

Remark: In Example 7.10, we wrote the joint distribution (in Step 3) as

\[
f_{X,\theta}(x, \theta) = f_{X|\theta}(x|\theta)\pi(\theta)
\]

\[
= \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i! \Gamma(a)b^a} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot \theta^{\alpha-1} e^{-\theta/\beta}
\]

\[
= \frac{1}{\prod_{i=1}^n x_i! \Gamma(a)b^a} \cdot \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot \theta^{\sum_{i=1}^n x_i + a - 1} e^{-\theta/(n+\frac{1}{b})-1}
\]

This does not depend on \( \theta \).

At this step, we can clearly identify the kernel of the posterior distribution. We can therefore skip calculating the marginal distribution \( m_X(x) \) in Step 4, because we know \( m_X(x) \) does not depend on \( \theta \). Because of this, it is common to write, in general,

\[
\pi(\theta|x) \propto f_{X|\theta}(x|\theta)\pi(\theta)
\]

\[
= L(\theta|x)\pi(\theta)
\]

The posterior distribution is proportional to the likelihood function times the prior distribution. A (classical) Bayesian analysis requires these two functions \( L(\theta|x) \) and \( \pi(\theta) \) only.

Remark: Suppose \( X|\theta \sim f_{X|\theta}(x|\theta) \). If \( T = T(X) \) is sufficient, we can write

\[
f_{X|\theta}(x|\theta) = g(t|\theta)h(x),
\]

by the Factorization Theorem. Therefore, the posterior distribution

\[
\pi(\theta|x) \propto f_{X|\theta}(x|\theta)\pi(\theta)
\]

\[
\propto g(t|\theta)\pi(\theta).
\]

This shows that the posterior distribution will depend on the data \( x \) through the value of the sufficient statistic \( t = T(x) \). We can therefore write the posterior distribution as depending on \( t \) only; i.e.,

\[
\pi(\theta|t) \propto f_{T|\theta}(t|\theta)\pi(\theta),
\]

and restrict attention to the (sampling) distribution of \( T = T(X) \) from the beginning.
Example 7.11. Suppose that $X_1, X_2, ..., X_n$ are iid Bernoulli($\theta$), where the prior distribution for $\theta \sim \text{beta}(a, b)$, $a, b$ known. We know that

$$T = T(X) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic for the Bernoulli family and that $T \sim b(n, \theta)$. Therefore, for $t = 0, 1, 2, ..., n$ and $0 < \theta < 1$, the posterior distribution

$$\pi(\theta|t) \propto f_{T|\theta}(t|\theta)\pi(\theta) = \binom{n}{t} \frac{\theta^t (1-\theta)^{n-t}}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

where $a^* = t + a$ and $b^* = n - t + b$. From here, we can immediately conclude that the posterior distribution

$$\theta|T = t \sim \text{beta}(t + a, n - t + b),$$

where $t = T(X) = \sum_{i=1}^{n} x_i$.

**Discussion:** In Examples 7.10 and 7.11, we observed the following occurrence:

- Example 7.10. $\theta \sim \text{gamma}$ (prior) $\rightarrow$ $\theta|X = x \sim \text{gamma}$ (posterior).
- Example 7.11. $\theta \sim \text{beta}$ (prior) $\rightarrow$ $\theta|T = t \sim \text{beta}$ (posterior).

**Definition:** Let $F = \{f_X(x|\theta) : \theta \in \Theta\}$ denote a class of pdfs or pmfs. A class $\Pi$ of prior distributions is said to be a **conjugate prior family** for $F$ if the posterior distribution also belongs to $\Pi$.

As we have already seen in Examples 7.10 and 7.11,

- The gamma family is conjugate for the Poisson family.
- The beta family is conjugate for the binomial family.

**Example 7.12.** Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

- If $\sigma^2$ is known, a conjugate prior for $\mu$ is $\mu \sim \mathcal{N}(\xi, \tau^2)$, $\xi, \tau^2$ known.

- If $\mu$ is known, a conjugate prior for $\sigma^2$ is $\sigma^2 \sim \text{IG}(a, b)$, $a, b$ known.
7.3 Methods of Evaluating Estimators

7.3.1 Bias, variance, and MSE

**Definition:** Suppose \( W = W(X) \) is a point estimator. We call \( W \) an **unbiased estimator** of \( \theta \) if
\[
E_{\theta}(W) = \theta \quad \text{for all } \theta \in \Theta.
\]
More generally, we call \( W \) an unbiased estimator of \( \tau(\theta) \) if
\[
E_{\theta}(W) = \tau(\theta) \quad \text{for all } \theta \in \Theta.
\]

**Definition:** The **mean-squared error (MSE)** of a point estimator \( W = W(X) \) is
\[
\text{MSE}_{\theta}(W) = E_{\theta}[(W - \theta)^2] = \text{var}_{\theta}(W) + [E_{\theta}(W) - \theta]^2 = \text{var}_{\theta}(W) + \text{Bias}_{\theta}^2(W),
\]
where \( \text{Bias}_{\theta}(W) = E_{\theta}(W) - \theta \) is the bias of \( W \) as an estimator of \( \theta \). Note that if \( W \) is an unbiased estimator of \( \theta \), then for all \( \theta \in \Theta \),
\[
E_{\theta}(W) = \theta \implies \text{Bias}_{\theta}(W) = E_{\theta}(W) - \theta = 0.
\]
In this case,
\[
\text{MSE}_{\theta}(W) = \text{var}_{\theta}(W).
\]

**Remark:** In general, the MSE incorporates two components:

- \( \text{var}_{\theta}(W) \); this measures **precision**
- \( \text{Bias}_{\theta}(W) \); this measures **accuracy**.

**Example 7.13.** Suppose \( X_1, X_2, \ldots, X_n \) are iid \( \mathcal{N}(\mu, \sigma^2) \), where \(-\infty < \mu < \infty \) and \( \sigma^2 > 0 \); i.e., both parameters unknown. Set \( \theta = (\mu, \sigma^2) \). Recall that our “usual” sample variance estimator is
\[
S^2 = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \bar{X})^2
\]
and for all \( \theta \),
\[
E_{\theta}(S^2) = \sigma^2 \quad \text{var}_{\theta}(S^2) = \frac{2\sigma^4}{n - 1}.
\]
Consider the “competing estimator:”
\[
S_b^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2,
\]
which recall is the MOM and MLE of \( \sigma^2 \).
Note that

\[ S_b^2 = \left( \frac{n-1}{n} \right) S^2 \implies E_\theta(S_b^2) = E_\theta \left[ \left( \frac{n-1}{n} \right) S^2 \right] = \left( \frac{n-1}{n} \right) E_\theta(S^2) = \left( \frac{n-1}{n} \right) \sigma^2. \]

That is, the estimator \( S_b^2 \) is biased; it underestimates \( \sigma^2 \) on average.

**Comparison:** Let’s compare \( S^2 \) and \( S_b^2 \) on the basis of MSE. Because \( S^2 \) is an unbiased estimator of \( \sigma^2 \),

\[ \text{MSE}_\theta(S^2) = \text{var}_\theta(S^2) = \frac{2\sigma^4}{n-1}. \]

The MSE of \( S_b^2 \) is

\[ \text{MSE}_\theta(S_b^2) = \text{var}_\theta(S_b^2) + \text{Bias}_\theta^2(S_b^2). \]

The variance of \( S_b^2 \) is

\[ \text{var}_\theta(S_b^2) = \text{var}_\theta \left[ \left( \frac{n-1}{n} \right) S^2 \right] = \left( \frac{n-1}{n} \right)^2 \text{var}_\theta(S^2) = \left( \frac{n-1}{n} \right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}. \]

The bias of \( S_b^2 \) is

\[ \text{E}_\theta(S_b^2 - \sigma^2) = \text{E}_\theta(S_b^2) - \sigma^2 = \left( \frac{n-1}{n} \right) \sigma^2 - \sigma^2. \]

Therefore,

\[ \text{MSE}_\theta(S_b^2) = \frac{2(n-1)\sigma^4}{n^2} + \left[ \left( \frac{n-1}{n} \right) \sigma^2 - \sigma^2 \right]^2 = \left( \frac{2n-1}{n^2} \right) \sigma^4. \]

Finally, to compare \( \text{MSE}_\theta(S^2) \) with \( \text{MSE}_\theta(S_b^2) \), we are left to compare the constants

\[ \frac{2}{n-1} \quad \text{and} \quad \frac{2n-1}{n^2}. \]

Note that the ratio

\[ \frac{2n-1}{n^2} = \frac{2n^2 - 3n + 1}{2n^2} < 1, \]

for all \( n \geq 2 \). Therefore,

\[ \text{MSE}_\theta(S_b^2) < \text{MSE}_\theta(S^2), \]

showing that \( S_b^2 \) is a “better” estimator than \( S^2 \) on the basis of MSE.
Discussion: In general, how should we compare two competing estimators $W_1$ and $W_2$?

- If both $W_1$ and $W_2$ are unbiased, we prefer the estimator with the smaller variance.
- If either $W_1$ or $W_2$ is biased (or perhaps both are biased), we prefer the estimator with the smaller MSE.

There is no guarantee that one estimator, say $W_1$, will always beat the other for all $\theta \in \Theta$ (i.e., for all values of $\theta$ in the parameter space). For example, it may be that $W_1$ has smaller MSE for some values of $\theta \in \Theta$, but larger MSE for other values.

Remark: In some situations, we might have a biased estimator, but we can calculate its bias. We can then “adjust” the (biased) estimator to make it unbiased. I like to call this “making biased estimators unbiased.” The following example illustrates this.

Example 7.14. Suppose that $X_1, X_2, \ldots, X_n$ are iid $U[0, \theta]$, where $\theta > 0$. We know (from Example 7.4) that the MLE of $\theta$ is $X_{(n)}$, the maximum order statistic. It is easy to show that

$$E_\theta(X_{(n)}) = \left( \frac{n}{n+1} \right) \theta.$$ 

The MLE is biased because $E_\theta(X_{(n)}) \neq \theta$. However, the estimator

$$\left( \frac{n + 1}{n} \right) X_{(n)},$$

an “adjusted version” of $X_{(n)}$, is unbiased.

Remark: In the previous example, we might compare the following estimators:

$$W_1 = W_1(X) = \left( \frac{n + 1}{n} \right) X_{(n)}$$

$$W_2 = W_2(X) = 2X.$$ 

The estimator $W_1$ is an unbiased version of the MLE. The estimator $W_2$ is the MOM (which is also unbiased). I have calculated

$$\text{var}_\theta(W_1) = \frac{\theta^2}{n(n + 2)} \quad \text{and} \quad \text{var}_\theta(W_2) = \frac{\theta^2}{3n}.$$ 

It is easy to see that $\text{var}_\theta(W_1) \leq \text{var}_\theta(W_2)$, for all $n \geq 2$. Therefore, $W_1$ is a “better” estimator on the basis of this variance comparison. Are you surprised?

Curiosity: Might there be another unbiased estimator, say $W_3 = W_3(X)$ that is “better” than both $W_1$ and $W_2$? If a better (unbiased) estimator does exist, how do we find it?

7.3.2 Best unbiased estimators

Goal: Consider the class of estimators

$$C_\tau = \{ W = W(X) : E_\theta(W) = \tau(\theta) \ \forall \theta \in \Theta \}.$$ 

That is, $C_\tau$ is the collection of all unbiased estimators of $\tau(\theta)$. Our goal is to find the (unbiased) estimator $W^* \in C_\tau$ that has the smallest variance.
Remark: On the surface, this task seems somewhat insurmountable because $C_\tau$ is a very large class. In Example 7.14, for example, both $W_1 = \left(\frac{n+1}{n}\right) X_{(n)}$ and $W_2 = 2 \bar{X}$ are unbiased estimators of $\theta$. However, so is the convex combination

$$W_a = W_a(X) = a \left(\frac{n+1}{n}\right) X_{(n)} + (1 - a) 2 \bar{X},$$

for all $a \in (0, 1)$.

Remark: It seems that our discussion of “best” estimators starts with the restriction that we will consider only those that are unbiased. If we did not make a restriction like this, then we would have to deal with too many estimators, many of which are nonsensical. For example, suppose $X_1, X_2, ..., X_n$ are iid Poisson($\theta$), where $\theta > 0$.

- The estimators $\bar{X}$ and $S^2$ emerge as candidate estimators because they are unbiased.

- However, suppose we widen our search to consider all possible estimators and then try to find the one with the smallest MSE. Consider the estimator $\hat{\theta} = 17$.
  - If $\theta = 17$, then $\hat{\theta}$ can never be beaten in terms of MSE; its MSE = 0.
  - If $\theta \neq 17$, then $\hat{\theta}$ may be a terrible estimator; its MSE = $(17 - \theta)^2$.

- We want to exclude nonsensical estimators like this. Our solution is to restrict attention to estimators that are unbiased.

Definition: An estimator $W^* = W^*(X)$ is a uniformly minimum variance unbiased estimator (UMVUE) of $\tau(\theta)$ if

1. $E_\theta(W^*) = \tau(\theta)$ for all $\theta \in \Theta$

2. $\text{var}_\theta(W^*) \leq \text{var}_\theta(W)$, for all $\theta \in \Theta$, where $W$ is any other unbiased estimator of $\tau(\theta)$.

Note: This definition is stated in full generality. Most of the time (but certainly not always), we will be interested in estimating $\theta$ itself; i.e., $\tau(\theta) = \theta$. Also, as the notation suggests, we assume that $\tau(\theta)$ is a scalar parameter and that estimators are also scalar.

Discussion/Preview: How do we find UMVUEs? We start by noting the following:

- UMVUEs may not exist.

- If a UMVUE does exist, it is unique (we’ll prove this later).

We present two approaches to find UMVUEs:

**Approach 1:** Determine a lower bound, say $B(\theta)$, on the variance of any unbiased estimator of $\tau(\theta)$. Then, if we can find an unbiased estimator $W^*$ whose variance attains this lower bound, that is,

$$\text{var}_\theta(W^*) = B(\theta),$$

for all $\theta \in \Theta$, then we know that $W^*$ is UMVUE.

**Approach 2:** Link the notion of being “best” with that of sufficiency and completeness.
Theorem 7.3.9 (Crâmer-Rao Inequality). Suppose $X \sim f_X(x|\theta)$, where

1. the support of $X$ is free of all unknown parameters
2. for any function $h(x)$ such that $E_\theta[h(X)] < \infty$ for all $\theta \in \Theta$, the interchange
   \[ \frac{d}{d\theta} \int_{\mathbb{R}^n} h(x) f_X(x|\theta) dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta} h(x) f_X(x|\theta) dx \]
   is justified; i.e., we can interchange the derivative and integral (derivative and sum if $X$ is discrete).

For any estimator $W(X)$ with $\text{var}_\theta[W(X)] < \infty$, the following inequality holds:

\[ \text{var}_\theta[W(X)] \geq \frac{\left\{ \frac{d}{d\theta} E_\theta[W(X)] \right\}^2}{E_\theta \left\{ \left[ \frac{\partial}{\partial\theta} \ln f_X(X|\theta) \right]^2 \right\}}. \]

The quantity on the RHS is called the Crâmer-Rao Lower Bound (CRLB) on the variance of the estimator $W(X)$.

Remark: Note that in the statement of the CRLB in Theorem 7.3.9, we haven’t said exactly what $W(X)$ is an estimator for. This is to preserve the generality of the result; Theorem 7.3.9 holds for any estimator with finite variance. However, given our desire to restrict attention to unbiased estimators, we will usually consider one of these cases:

- If $W(X)$ is an unbiased estimator of $\tau(\theta)$, then the numerator becomes
  \[ \left[ \frac{d}{d\theta} \tau(\theta) \right]^2 = [\tau'(\theta)]^2. \]
- If $W(X)$ is an unbiased estimator of $\tau(\theta) = \theta$, then the numerator equals 1.

Important special case (Corollary 7.3.10): When $X$ consists of $X_1, X_2, ..., X_n$ which are iid from the population $f_X(x|\theta)$, then the denominator in Theorem 7.3.9

\[ E_\theta \left\{ \left[ \frac{\partial}{\partial\theta} \ln f_X(X|\theta) \right]^2 \right\} = n E_\theta \left\{ \left[ \frac{\partial}{\partial\theta} \ln f_X(X|\theta) \right]^2 \right\}, \]

or, using other notation,

\[ I_n(\theta) = n I_1(\theta). \]

We call $I_n(\theta)$ the Fisher information based on the sample $X$. We call $I_1(\theta)$ the Fisher information based on one observation $X$.

Lemma 7.3.11 (Information Equality): Under fairly mild assumptions (which hold for exponential families, for example), the Fisher information based on one observation

\[ I_1(\theta) = E_\theta \left\{ \left[ \frac{\partial}{\partial\theta} \ln f_X(X|\theta) \right]^2 \right\} = -E_\theta \left[ \frac{\partial^2}{\partial\theta^2} \ln f_X(X|\theta) \right]. \]

The second expectation is often easier to calculate.
Preview: In Chapter 10, we will investigate the large-sample properties of MLEs. Under certain regularity conditions, we will show an MLE $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\rightarrow} N(0, \sigma^2_{\hat{\theta}}),$$

where the asymptotic variance

$$\sigma^2_{\hat{\theta}} = \frac{1}{I_1(\theta)}.$$ 

This is an extremely useful (large-sample) result; e.g., it makes getting large-sample CIs and performing large-sample tests straightforward. Furthermore, an analogous large-sample result holds for vector-valued MLEs. If $\bm{\hat{\theta}}$ is the MLE of a $k \times 1$ dimensional parameter $\bm{\theta}$, then

$$\sqrt{n}(\bm{\hat{\theta}} - \bm{\theta}) \overset{d}{\rightarrow} \text{mvn}_k(0, \Sigma),$$

where the asymptotic variance-covariance matrix (now, $k \times k$)

$$\Sigma = [I_1(\theta)]^{-1}$$ 

is the inverse of the $k \times k$ Fisher information matrix $I_1(\theta)$.

Example 7.15. Suppose $X_1, X_2, ..., X_n$ are iid Poisson($\theta$), where $\theta > 0$. Find the CRLB on the variance of unbiased estimators of $\tau(\theta) = \theta$.

Solution. We know that the CRLB is

$$\frac{1}{I_n(\theta)} = \frac{1}{nI_1(\theta)},$$

where

$$I_1(\theta) = E_\theta \left\{ \left( \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right)^2 \right\} = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right].$$

For $x = 0, 1, 2, ...$,

$$\ln f_X(x|\theta) = \ln \left( \frac{\theta^x e^{-\theta}}{x!} \right) = x \ln \theta - \theta - \ln x!.$$

Therefore,

$$\frac{\partial}{\partial \theta} \ln f_X(x|\theta) = \frac{x}{\theta} - 1$$

$$\frac{\partial^2}{\partial \theta^2} \ln f_X(x|\theta) = -\frac{x}{\theta^2}.$$ 

The Fisher information based on one observation is

$$I_1(\theta) = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right]$$

$$= -E_\theta \left( -\frac{X}{\theta^2} \right) = \frac{1}{\theta}.$$ 

Therefore, the CRLB on the variance of all unbiased estimators of $\tau(\theta) = \theta$ is

$$\text{CRLB} = \frac{1}{nI_1(\theta)} = \frac{\theta}{n}.$$
Observation: Because $W(X) = \bar{X}$ is an unbiased estimator of $\tau(\theta) = \theta$ in the Poisson($\theta$) model and because

$$\text{var}_\theta(\bar{X}) = \frac{\theta}{n},$$

we see that $\text{var}_\theta(\bar{X})$ does attain the CRLB. This means that $W(X) = \bar{X}$ is the UMVUE for $\tau(\theta) = \theta$.

**Example 7.16.** Suppose $X_1, X_2, \ldots, X_n$ are iid gamma($\alpha_0, \beta$), where $\alpha_0$ is known and $\beta > 0$. Find the CRLB on the variance of unbiased estimators of $\beta$.

**Solution.** We know that the CRLB is

$$\frac{1}{I_n(\beta)} = \frac{1}{n I_1(\beta)},$$

where

$$I_1(\beta) = E_\beta \left\{ \left[ \frac{\partial}{\partial \beta} \ln f_X(X|\beta) \right]^2 \right\} = -E_\beta \left[ \frac{\partial^2}{\partial \beta^2} \ln f_X(X|\beta) \right].$$

For $x > 0$,

$$\ln f_X(x|\beta) = \ln \left[ \frac{1}{\Gamma(\alpha_0) \beta^{\alpha_0}} x^{\alpha_0 - 1} e^{-x/\beta} \right] = -\ln \Gamma(\alpha_0) - \alpha_0 \ln \beta + (\alpha_0 - 1) \ln x - \frac{x}{\beta}.$$

Therefore,

$$\frac{\partial}{\partial \beta} \ln f_X(x|\beta) = -\frac{\alpha_0}{\beta} + \frac{x}{\beta^2}$$

and

$$\frac{\partial^2}{\partial \beta^2} \ln f_X(x|\beta) = \frac{\alpha_0}{\beta^2} - \frac{2x}{\beta^3}.$$

The Fisher information based on one observation is

$$I_1(\beta) = -E_\beta \left[ \frac{\partial^2}{\partial \beta^2} \ln f_X(X|\beta) \right] = -E_\beta \left( \frac{\alpha_0}{\beta^2} - \frac{2X}{\beta^3} \right) = \frac{\alpha_0}{\beta^2}.$$

Therefore, the CRLB on the variance of all unbiased estimators of $\beta$ is

$$\text{CRLB} = \frac{1}{n I_1(\beta)} = \frac{\beta^2}{n \alpha_0}.$$
and
\[ \text{var}_\beta[W(X)] = \text{var}_\beta \left( \frac{\bar{X}}{\alpha_0} \right) = \frac{\alpha_0 \beta^2}{n\alpha_0^2} = \frac{\beta^2}{n\alpha_0}. \]

We see that \( W(X) = \frac{\bar{X}}{\alpha_0} \) is an unbiased estimator for \( \beta \) and \( \text{var}_\beta(\bar{X}/\alpha_0) \) attains the CRLB. This means that \( W(X) = \frac{\bar{X}}{\alpha_0} \) is the UMVUE for \( \beta \).

**Discussion:** Instead of estimating \( \beta \) in Example 7.16, suppose that we were interested in estimating \( \tau(\beta) = 1/\beta \) instead.

1. Show that
\[ W(X) = \frac{n\alpha_0 - 1}{n\bar{X}} \]
is an unbiased estimator of \( \tau(\beta) = 1/\beta \).

2. Derive the CRLB for the variance of unbiased estimators of \( \tau(\beta) = 1/\beta \).

3. Calculate \( \text{var}_\beta[W(X)] \) and show that it is strictly larger than the CRLB (i.e., the variance does not attain the CRLB).

**Q:** Does this necessarily imply that \( W(X) \) cannot be the UMVUE of \( \tau(\beta) = 1/\beta \)?

**Remark:** In general, the CRLB offers a lower bound on the variance of any unbiased estimator of \( \tau(\theta) \). However, this lower bound may be unattainable. That is, the CRLB may be strictly smaller than the variance of any unbiased estimator. If this is the case, then our “CRLB approach” to finding an UMVUE will not be helpful.

**Corollary 7.3.15 (Attainment).** Suppose \( X_1, X_2, ..., X_n \) is an iid sample from \( f_X(x|\theta) \), where \( \theta \in \Theta \), a family that satisfies the regularity conditions stated for the Cramér-Rao Inequality. If \( W(X) \) is an unbiased estimator of \( \tau(\theta) \), then \( \text{var}_\theta[W(X)] \) attains the CRLB if and only if the score function
\[ S(\theta|x) = a(\theta)[W(x) - \tau(\theta)] \]
is a linear function of \( W(x) \).

**Recall:** The score function is given by
\[
S(\theta|x) = \frac{\partial}{\partial \theta} \ln L(\theta|x)
= \frac{\partial}{\partial \theta} \ln f_X(x|\theta).
\]

**Example 7.16 (continued).** Suppose \( X_1, X_2, ..., X_n \) are iid gamma(\( \alpha_0, \beta \)), where \( \alpha_0 \) is known and \( \beta > 0 \). The likelihood function is
\[
L(\beta|x) = \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha_0)\beta^{\alpha_0}} x_i^{\alpha_0 - 1} e^{-x_i/\beta}
= \left[ \frac{1}{\Gamma(\alpha_0)\beta^{\alpha_0}} \right]^n \left( \prod_{i=1}^{n} x_i \right)^{\alpha_0 - 1} e^{-\sum_{i=1}^{n} x_i/\beta}.
\]
The log-likelihood function is
\[
\ln L(\beta|x) = -n \ln \Gamma(\alpha_0) - n\alpha_0 \ln \beta + (\alpha_0 - 1) \sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} x_i}{\beta}.
\]

The score function is
\[
S(\beta|x) = \frac{\partial}{\partial \beta} \ln L(\beta|x) = -\frac{n\alpha_0}{\beta} + \frac{\sum_{i=1}^{n} x_i}{\beta^2} = \frac{n\alpha_0}{\beta^2} \left( \frac{\sum_{i=1}^{n} x_i}{n\alpha_0} - \beta \right) = a(\beta)[W(x) - \tau(\beta)],
\]
where
\[
W(x) = \frac{\sum_{i=1}^{n} x_i}{n\alpha_0} = \frac{x}{\alpha_0}.
\]

We have written the score function \( S(\beta|x) \) as a linear function of \( W(x) = \frac{x}{\alpha_0} \). Because \( W(X) = \frac{\bar{X}}{\alpha_0} \) is an unbiased estimator of \( \tau(\beta) = \beta \) (shown previously), the variance \( \text{var}_\theta[W(X)] \) attains the CRLB for the variance of unbiased estimators of \( \tau(\beta) = \beta \).

**Remark:** The attainment result is interesting, but I have found that its usefulness may be limited if you want to find the UMVUE. Even if we can write
\[
S(\theta|x) = a(\theta)[W(x) - \tau(\theta)]
\]
where \( E_\theta[W(X)] = \tau(\theta) \), the RHS might involve a function \( \tau(\theta) \) for which there is no desire to estimate. To illustrate this, suppose \( X_1, X_2, \ldots, X_n \) are iid beta(\( \theta, 1 \)), where \( \theta > 0 \). The score function is
\[
S(\theta|x) = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i = n \left[ \frac{\sum_{i=1}^{n} \ln x_i}{n} - \left( -\frac{1}{\theta} \right) \right] = a(\theta)[W(x) - \tau(\theta)].
\]

It turns out that
\[
E_\theta[W(X)] = \frac{1}{n} \sum_{i=1}^{n} \ln X_i = -\frac{1}{\theta}.
\]

We have shown that \( \text{var}_\theta[W(X)] \) attains the CRLB on the variance of unbiased estimators of \( \tau(\theta) = -1/\theta \), a parameter we likely have no desire to estimate.

**Unresolved issues:**

1. What if \( f_X(x|\theta) \) does not satisfy the regularity conditions needed for the Cramér-Rao Inequality to apply? For example, \( X \sim U(0, \theta) \).
2. What if the CRLB is unattainable? Can we still find the UMVUE?
7.3.3 Sufficiency and completeness

Remark: We now move to our “second approach” on how to find UMVUEs. This approach involves sufficiency and completeness—two topics we discussed in the last chapter. We can also address the unresolved issues on the previous page.

Theorem 7.3.17 (Rao-Blackwell). Let \( W = W(X) \) be an unbiased estimator of \( \tau(\theta) \). Let \( T = T(X) \) be a sufficient statistic for \( \theta \). Define

\[
\phi(T) = E(W|T).
\]

Then

1. \( E_\theta[\phi(T)] = \tau(\theta) \) for all \( \theta \in \Theta \)
2. \( \text{var}_\theta[\phi(T)] \leq \text{var}_\theta(W) \) for all \( \theta \in \Theta \).

That is, \( \phi(T) = E(W|T) \) is a uniformly better unbiased estimator than \( W \).

Proof. This result follows from the iterated rules for means and variances. First,

\[
E_\theta[\phi(T)] = E_\theta[E(W|T)] = E_\theta(W) = \tau(\theta).
\]

Second,

\[
\text{var}_\theta(W) = E_\theta[\text{var}(W|T)] + \text{var}_\theta[E(W|T)] \\
= E_\theta[\text{var}(W|T)] + \text{var}_\theta[\phi(T)] \\
\geq \text{var}_\theta[\phi(T)],
\]

because \( \text{var}(W|T) \geq 0 \) (a.s.) and hence \( E_\theta[\text{var}(W|T)] \geq 0 \). □

Implication: We can always “improve” the unbiased estimator \( W \) by conditioning on a sufficient statistic.

Remark: To use the Rao-Blackwell Theorem, some students think they have to

1. Find an unbiased estimator \( W \).
2. Find a sufficient statistic \( T \).
3. Derive the conditional distribution \( f_{W|T}(w|t) \).
4. Find the mean \( E(W|T) \) of this conditional distribution.

This is not the case at all! Because \( \phi(T) = E(W|T) \) is a function of the sufficient statistic \( T \), the Rao-Blackwell result simply convinces us that in our search for the UMVUE, we can restrict attention to those estimators that are functions of a sufficient statistic.

Q: In the proof of the Rao-Blackwell Theorem, where did we use the fact that \( T \) was sufficient?

A: Nowhere. Thus, it would seem that conditioning on any statistic, sufficient or not, will result in an improvement over the unbiased \( W \). However, there is a catch:

- If \( T \) is not sufficient, then there is no guarantee that \( \phi(T) = E(W|T) \) will be an estimator; i.e., it could depend on \( \theta \). See Example 7.3.18 (CB, pp 343).
Remark: To understand how we can use the Rao-Blackwell result in our quest to find a UMVUE, we need two additional results. One deals with uniqueness; the other describes an interesting characterization of a UMVUE itself.

**Theorem 7.3.19 (Uniqueness).** If $W$ is UMVUE for $\tau(\theta)$, then it is unique.

*Proof.* Suppose that $W'$ is also UMVUE. It suffices to show that $W = W'$ with probability one. Define

$$W^* = \frac{1}{2}(W + W').$$

Note that

$$E_{\theta}(W^*) = \frac{1}{2}[E_{\theta}(W) + E_{\theta}(W')] = \tau(\theta), \text{ for all } \theta \in \Theta,$$

showing that $W^*$ is an unbiased estimator of $\tau(\theta)$. The variance of $W^*$ is

$$\text{var}_{\theta}(W^*) = \text{var}_{\theta}\left[\frac{1}{2}(W + W')\right]$$

$$= \frac{1}{4}\text{var}_{\theta}(W) + \frac{1}{4}\text{var}_{\theta}(W') + \frac{1}{2}\text{cov}_{\theta}(W, W')$$

$$\leq \frac{1}{4}\text{var}_{\theta}(W) + \frac{1}{4}\text{var}_{\theta}(W') + \frac{1}{2}[\text{var}_{\theta}(W)\text{var}_{\theta}(W')]^{1/2}$$

$$= \text{var}_{\theta}(W),$$

where the inequality arises from the covariance inequality (CB, pp 188, application of Cauchy-Schwarz) and the final equality holds because both $W$ and $W'$ are UMVUE by assumption (so their variances must be equal). Therefore, we have shown that

1. $W^*$ is unbiased for $\tau(\theta)$
2. $\text{var}_{\theta}(W^*) \leq \text{var}_{\theta}(W)$.

Because $W$ is UMVUE (by assumption), the inequality in (2) can not be strict (or else it would contradict the fact that $W$ is UMVUE). Therefore, it must be true that

$$\text{var}_{\theta}(W^*) = \text{var}_{\theta}(W).$$

This implies that the inequality above (arising from the covariance inequality) is an equality; therefore,

$$\text{cov}_{\theta}(W, W') = [\text{var}_{\theta}(W)\text{var}_{\theta}(W')]^{1/2}.$$

Therefore,

$$\text{corr}_{\theta}(W, W') = \pm 1 \implies W' = a(\theta)W + b(\theta), \text{ with probability } 1,$$

by Theorem 4.5.7 (CB, pp 172), where $a(\theta)$ and $b(\theta)$ are constants. It therefore suffices to show that $a(\theta) = 1$ and $b(\theta) = 0$. Note that

$$\text{cov}_{\theta}(W, W') = \text{cov}_{\theta}[W, a(\theta)W + b(\theta)] = a(\theta)\text{cov}_{\theta}(W, W)$$

$$= a(\theta)\text{var}_{\theta}(W).$$
However, we have previously shown that 
\[ \text{cov}_\theta(W, W') = [\text{var}_\theta(W)\text{var}_\theta(W')]^{1/2} = [\text{var}_\theta(W)\text{var}_\theta(W')]^{1/2} \]
\[ = \text{var}_\theta(W). \]

This implies \( a(\theta) = 1. \) Finally, 
\[ E_\theta(W') = E_\theta[a(\theta)W + b(\theta)] = E_\theta[W + b(\theta)] = E_\theta(W) + b(\theta). \]

Because both \( W \) and \( W' \) are unbiased, this implies \( b(\theta) = 0. \)

Theorem 7.3.20. Suppose \( E_\theta(W) = \tau(\theta) \) for all \( \theta \in \Theta. \) \( W \) is UMVUE of \( \tau(\theta) \) if and only if \( W \) is uncorrelated with all unbiased estimators of \( 0. \)

Proof. Necessity (\( \implies \)) Suppose \( E_\theta(W) = \tau(\theta) \) for all \( \theta \in \Theta. \) Suppose \( W \) is UMVUE of \( \tau(\theta). \) Suppose \( E_\theta(U) = 0 \) for all \( \theta \in \Theta. \) It suffices to show \( \text{cov}_\theta(W, U) = 0 \) for all \( \theta \in \Theta. \)

Define \( \phi_a = W + aU, \)
where \( a \) is a constant. It is easy to see that \( \phi_a \) is an unbiased estimator of \( \tau(\theta); \) for all \( \theta \in \Theta, \)
\[ E_\theta(\phi_a) = E_\theta(W + aU) = E_\theta(W) + a E_\theta(U) = \tau(\theta). \]

Also,
\[ \text{var}_\theta(\phi_a) = \text{var}_\theta(W + aU) \]
\[ = \text{var}_\theta(W) + a^2 \text{var}_\theta(U) + 2a \text{cov}_\theta(W, U). \]

Key question: Can this be negative?

\textbf{• Case 1:} Suppose \( \exists \theta_0 \in \Theta \) such that \( \text{cov}_{\theta_0}(W, U) < 0. \) Then
\[ a^2 \text{var}_{\theta_0}(U) + 2a \text{cov}_{\theta_0}(W, U) < 0 \iff a^2 \text{var}_{\theta_0}(U) < -2a \text{cov}_{\theta_0}(W, U) \]
\[ \iff a^2 < -\frac{2 \text{cov}_{\theta_0}(W, U)}{\text{var}_{\theta_0}(U)}. \]

I can make this true by picking
\[ 0 < a < -\frac{2 \text{cov}_{\theta_0}(W, U)}{\text{var}_{\theta_0}(U)} \]
and therefore I have shown that
\[ \text{var}_{\theta_0}(\phi_a) < \text{var}_{\theta_0}(W). \]

However, this contradicts the assumption that \( W \) is UMVUE. Therefore, it must be true that \( \text{cov}_\theta(W, U) \geq 0. \)
\textbf{Case 2:} Suppose \( \exists \theta_0 \in \Theta \) such that \( \text{cov}_{\theta_0}(W, U) > 0 \). Then
\[
a^2 \text{var}_{\theta_0}(U) + 2a \text{cov}_{\theta_0}(W, U) < 0 \iff a^2 \text{var}_{\theta_0}(U) < -2a \text{cov}_{\theta_0}(W, U) \\
\iff a^2 < -\frac{2a \text{cov}_{\theta_0}(W, U)}{\text{var}_{\theta_0}(U)}.
\]
I can make this true by picking
\[
-\frac{2 \text{cov}_{\theta_0}(W, U)}{\text{var}_{\theta_0}(U)} < a < 0
\]
and therefore I have shown that
\[
\text{var}_{\theta_0}(\phi_a) < \text{var}_{\theta_0}(W).
\]
However, this again contradicts the assumption that \( W \) is UMVUE. Therefore, it must be true that \( \text{cov}_{\theta}(W, U) \leq 0 \).

Combining Case 1 and Case 2, we are forced to conclude that \( \text{cov}_{\theta}(W, U) = 0 \). This proves the necessity.

Sufficiency (\( \iff \)): Suppose \( E_{\theta}(W) = \tau(\theta) \) for all \( \theta \in \Theta \). Suppose \( \text{cov}_{\theta}(W, U) = 0 \) for all \( \theta \in \Theta \) where \( U \) is any unbiased estimator of zero; i.e., \( E_{\theta}(U) = 0 \) for all \( \theta \in \Theta \). Let \( W' \) be any other unbiased estimator of \( \tau(\theta) \). It suffices to show that \( \text{var}_{\theta}(W) \leq \text{var}_{\theta}(W') \). Write
\[
W' = W + (W' - W)
\]
and calculate
\[
\text{var}_{\theta}(W') = \text{var}_{\theta}(W) + \text{var}_{\theta}(W' - W) + 2\text{cov}_{\theta}(W, W' - W).
\]
However, \( \text{cov}_{\theta}(W, W' - W) = 0 \) because \( W' - W \) is an unbiased estimator of 0. Therefore,
\[
\text{var}_{\theta}(W') = \text{var}_{\theta}(W) + \underbrace{\text{var}_{\theta}(W' - W)}_{\geq 0} \geq \text{var}_{\theta}(W).
\]
This proves the sufficiency. \( \square \)

**Summary:** We are now ready to put Theorem 7.3.17 (Rao-Blackwell), Theorem 7.3.19 (UMVUE uniqueness) and Theorem 7.3.20 together. Suppose \( X \sim f_X(x|\theta) \), where \( \theta \in \Theta \). Our goal is to find the UMVUE of \( \tau(\theta) \).

- Theorem 7.3.17 (Rao-Blackwell) assures us that we can restrict attention to functions of sufficient statistics.

Therefore, suppose \( T \) is a sufficient statistic for \( \theta \). Suppose that \( \phi(T) \), a function of \( T \), is an unbiased estimator of \( \tau(\theta) \); i.e.,
\[
E_{\theta}[\phi(T)] = \tau(\theta), \quad \text{for all } \theta \in \Theta.
\]

- Theorem 7.3.20 assures us that \( \phi(T) \) is UMVUE if and only if \( \phi(T) \) is uncorrelated with all unbiased estimators of 0.

Add the assumption that \( T \) is a complete statistic. \textit{The only unbiased estimator of 0 in complete families is the zero function itself.} Because \( \text{cov}_{\theta}[\phi(T), 0] = 0 \) holds trivially, we have shown that \( \phi(T) \) is uncorrelated with “all” unbiased estimators of 0. Theorem 7.3.20 says that \( \phi(T) \) must be UMVUE; Theorem 7.3.19 guarantees that \( \phi(T) \) is unique.
Recipe for finding UMVUEs: Suppose we want to find the UMVUE for $\tau(\theta)$.

1. Start by finding a statistic $T$ that is both sufficient and complete.
2. Find a function of $T$, say $\phi(T)$, that satisfies
   \[ E_{\theta}[\phi(T)] = \tau(\theta), \quad \text{for all } \theta \in \Theta. \]

Then $\phi(T)$ is the UMVUE for $\tau(\theta)$. This is essentially what is summarized in Theorem 7.3.23 (CB, pp 347).

Example 7.17. Suppose $X_1, X_2, ..., X_n$ are iid Poisson($\theta$), where $\theta > 0$.

- We already know that $X$ is UMVUE for $\theta$; we proved this by showing that $X$ is unbiased and that $\text{var}_\theta(X)$ attains the CRLB on the variance of all unbiased estimators of $\theta$.
- We now show $X$ is UMVUE for $\theta$ by using sufficiency and completeness.

The pmf of $X$ is
\[
f_X(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} I(x = 0, 1, 2, ...,)
= \frac{I(x = 0, 1, 2, ...,)}{x!} e^{-\theta} e^{\theta \ln x}
= h(x)c(\theta) \exp\{w_1(\theta)t_1(x)\}.
\]

Therefore $X$ has pmf in the exponential family. Theorem 6.2.10 says that
\[ T = T(X) = \sum_{i=1}^{n} X_i \]
is a sufficient statistic. Because $d = k = 1$ (i.e., a full family), Theorem 6.2.25 says that $T$ is complete. Now,
\[ E_{\theta}(T) = E_{\theta}\left(\sum_{i=1}^{n} X_i\right)
= \sum_{i=1}^{n} E_{\theta}(X_i) = n\theta. \]

Therefore,
\[ E_{\theta}\left(\frac{T}{n}\right) = E_{\theta}(X) = \theta. \]

Because $X$ is unbiased and is a function of $T$, a complete and sufficient statistic, we know that $X$ is the UMVUE.

Example 7.18. Suppose $X_1, X_2, ..., X_n$ are iid $U(0, \theta)$, where $\theta > 0$. We have previously shown that
\[ T = T(X) = X_{(n)} \]
is sufficient and complete (see Example 6.5 and Example 6.16, respectively, in the notes). It follows that
\[ E_\theta(T) = E_\theta(X_{(n)}) = \left( \frac{n}{n+1} \right) \theta \]
for all \( \theta > 0 \). Therefore,
\[ E_\theta \left[ \left( \frac{n+1}{n} \right) X_{(n)} \right] = \theta. \]
Because \((n+1)X_{(n)}/n\) is unbiased and is a function of \(X_{(n)}\), a complete and sufficient statistic, it must be the UMVUE.

**Example 7.19.** Suppose \(X_1, X_2, \ldots, X_n\) are iid gamma\((\alpha_0, \beta)\), where \(\alpha_0\) is known and \(\beta > 0\). Find the UMVUE of \(\tau(\beta) = 1/\beta\).

*Solution.* The pdf of \(X\) is
\[
 f_X(x|\beta) = \frac{1}{\Gamma(\alpha_0)\beta^{\alpha_0}} x^{\alpha_0-1} e^{-x/\beta} I(x > 0) \\
 = \frac{x^{\alpha_0-1}I(x > 0)}{\Gamma(\alpha_0)} \frac{1}{\beta^{\alpha_0}} e^{(-1/\beta)x} \\
 = h(x)e(\beta)\exp\{w_1(\beta)t_1(x)\}
\]
a one-parameter exponential family with \(d = k = 1\) (a full family). Theorem 6.2.10 and Theorem 6.2.25 assure that
\[ T = T(X) = \sum_{i=1}^{n} X_i \]
is a sufficient and complete statistic, respectively. In Example 7.16 (notes), we saw that
\[ \phi(T) = \frac{n\alpha_0 - 1}{T} \]
is an unbiased estimator of \(\tau(\beta) = 1/\beta\). Therefore, \(\phi(T)\) must be the UMVUE.

**Remark:** In Example 7.16, recall that the CRLB on the variance of unbiased estimators of \(\tau(\beta) = 1/\beta\) was unattainable.

**Example 7.20.** Suppose \(X_1, X_2, \ldots, X_n\) are iid Poisson\((\theta)\), where \(\theta > 0\). Find the UMVUE for
\[ \tau(\theta) = P_\theta(X = 0) = e^{-\theta}. \]

*Solution.* We use an approach known as “direct conditioning.” We start with
\[ T = T(X) = \sum_{i=1}^{n} X_i, \]
which is sufficient and complete. We know that the UMVUE therefore is a function of \(T\). Consider forming
\[ \phi(T) = E(W|T), \]
where \(W\) is any unbiased estimator of \(\tau(\theta) = e^{-\theta}\). We know that \(\phi(T)\) by this construction is the UMVUE; clearly \(\phi(T) = E(W|T)\) is a function of \(T\) and
\[ E_\theta[\phi(T)] = E_\theta[E(W|T)] = E_\theta(W) = e^{-\theta}. \]
How should we choose $W$? Any unbiased $W$ will “work,” so let’s keep our choice simple, say $W = W(X) = I(X_1 = 0)$.

Note that $E_\theta(W) = E_\theta[I(X_1 = 0)] = P_\theta(X_1 = 0) = e^{-\theta}$,
showing that $W$ is an unbiased estimator. Now, we just calculate $\phi(T) = E(W|T)$ directly. For $t$ fixed, we have

$$\phi(t) = E(W|T = t) = E[I(X_1 = 0)|T = t] = P(X_1 = 0|T = t) = \frac{P_\theta(X_1 = 0, T = t)}{P_\theta(T = t)} = \frac{P_\theta(X_1 = 0, \sum_{i=2}^n X_i = t)}{P_\theta(T = t)} \overset{\text{indep}}{=} \frac{P_\theta(X_1 = 0)P_\theta(\sum_{i=2}^n X_i = t)}{P_\theta(T = t)}.$$

Recall that $X_1 \sim \text{Poisson}(\theta)$, $\sum_{i=2}^n X_i \sim \text{Poisson}((n-1)\theta)$, and $T \sim \text{Poisson}(n\theta)$. Therefore,

$$\phi(t) = \frac{e^{-\theta}[(n-1)\theta]^t e^{-(n-1)\theta}}{t!} \approx \left(\frac{n-1}{n}\right)^t.$$

Therefore,

$$\phi(T) = \left(\frac{n-1}{n}\right)^T$$

is the UMVUE of $\tau(\theta) = e^{-\theta}$.

**Remark:** It is interesting to note that in this example

$$\phi(t) = \left(\frac{n-1}{n}\right)^t = \left[\left(\frac{n-1}{n}\right)^n\right]^\frac{t}{n} = \left[\left(1 - \frac{1}{n}\right)^n\right]^\frac{t}{n} \approx e^{-\frac{t}{n}},$$
for $n$ large. Recall that $e^{-X}$ is the MLE of $\tau(\theta) = e^{-\theta}$ by invariance.

**Remark:** The last subsection in CB (Section 7.3.4) is on loss-function optimality. This material will be covered in STAT 822.

### 7.4 Appendix: CRLB Theory

**Remark:** In this section, we provide the proofs that pertain to the CRLB approach to finding UMVUEs. These proofs are also relevant for later discussions on MLEs and their large-sample characteristics.
Remark: We start by reviewing the Cauchy-Schwarz Inequality. Essentially, the main Cramér-Rao inequality result (Theorem 7.3.9) follows as an application of this inequality.

Recall: Suppose \( X \) and \( Y \) are random variables. Then
\[
|E(XY)| \leq E(|XY|) \leq \sqrt{E(X^2)E(Y^2)}.
\]
This is called the **Cauchy-Schwarz Inequality**. In this inequality, if we replace \( X \) with \( X - \mu_X \) and \( Y \) with \( Y - \mu_Y \), we get
\[
|E[(X - \mu_X)(Y - \mu_Y)]| \leq \sqrt{E((X - \mu_X)^2)E((Y - \mu_Y)^2)}.
\]
Squaring both sides, we get
\[
[\text{cov}(X,Y)]^2 \leq \sigma_X^2 \sigma_Y^2.
\]
This is called the **covariance inequality**.

**Theorem 7.3.9** (Crâmer-Rao Inequality). Suppose \( X \sim f_X(x|\theta) \), where

1. the support of \( X \) is free of all unknown parameters
2. for any function \( h(x) \) such that \( E_{\theta}[h(X)] < \infty \) for all \( \theta \in \Theta \), the interchange

\[
\frac{d}{d\theta} \int_{\mathbb{R}^n} h(x)f_X(x|\theta)dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} h(x)f_X(x|\theta)dx
\]

is justified; i.e., we can interchange the derivative and integral (derivative and sum if \( X \) is discrete).

For any estimator \( W(X) \) with \( \text{var}_\theta[W(X)] < \infty \), the following inequality holds:
\[
\text{var}_\theta[W(X)] \geq \frac{\left\{ \frac{d}{d\theta} E_{\theta}[W(X)] \right\}^2}{E_{\theta}\left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}}.
\]

**Proof.** First we state and prove a lemma.

**Lemma.** Let
\[
S(\theta|X) = \frac{\partial}{\partial \theta} \ln f_X(X|\theta)
\]
denote the **score function**. The score function is a zero-mean random variable; that is,
\[
E_{\theta}[S(\theta|X)] = \frac{\partial}{\partial \theta} \ln f_X(X|\theta) = 0.
\]

**Proof of Lemma:** Note that
\[
E_{\theta}\left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} \ln f_X(x|\theta) f_X(x|\theta)dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} f_X(x|\theta) f_X(x|\theta)dx = \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} f_X(x|\theta) dx = \frac{d}{d\theta} \int_{\mathbb{R}^n} f_X(x|\theta) dx = 0.
\]

The interchange of derivative and integral above is justified based on the assumptions stated in Theorem 7.3.9. Therefore, the lemma is proven. \( \square \)
Note: Because the score function is a zero-mean random variable,

\[ \text{var}_\theta [S(\theta | X)] = E_\theta \{ [S(\theta | X)]^2 \}; \]

that is,

\[ \text{var}_\theta \left( \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right) = E_\theta \left\{ \left( \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right)^2 \right\}. \]

We now return to the CRLB proof. Consider

\[
\text{cov}_\theta \left[ W(X), \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = E_\theta \left[ W(X) \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] - E_\theta[W(X)] E_\theta \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = 0.
\]

Now, write the covariance inequality with

1. \( W(X) \) playing the role of “\( X \)”
2. \( S(\theta | X) = \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \) playing the role of “\( Y \).”

We get

\[
\left\{ \text{cov}_\theta \left[ W(X), \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] \right\}^2 \leq \text{var}_\theta[W(X)] \text{var}_\theta \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right],
\]

that is,

\[
\left\{ \frac{d}{d\theta} E_\theta[W(X)] \right\}^2 \leq \text{var}_\theta[W(X)] E_\theta \left\{ \left( \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right)^2 \right\}.
\]

Dividing both sides by \( E_\theta \left\{ \left( \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right)^2 \right\} \) gives the result. □

**Corollary 7.3.10** (Cramer-Rao Inequality—iid case). With the same regularity conditions stated in Theorem 7.3.9, in the iid case,

\[
\text{var}_\theta[W(X)] \geq \frac{\left\{ \frac{d}{d\theta} E_\theta[W(X)] \right\}^2}{n E_\theta \left\{ \left( \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right)^2 \right\}}.
\]
Proof. It suffices to show

\[ E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\} = nE_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}. \]

Because \( X_1, X_2, ..., X_n \) are iid,

\[
\text{LHS} = E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} f_X(X_i|\theta) \right]^2 \right\} = E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f_X(X_i|\theta) \right]^2 \right\} = E_\theta \left\{ \left[ \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta) \right]^2 \right\} = \sum_{i=1}^{n} E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta) \right]^2 \right\} + \sum_{i \neq j}^{n} E_\theta \left[ \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta) \frac{\partial}{\partial \theta} \ln f_X(X_j|\theta) \right] \text{indep} = \sum_{i=1}^{n} E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta) \right]^2 \right\} = nE_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}.
\]

Therefore, all cross product expectations are zero and thus

\[
\text{LHS} = \sum_{i=1}^{n} E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta) \right]^2 \right\} \overset{\text{ident}}{=} nE_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}.
\]

This proves the iid case. \( \square \)

Remark: Recall our notation:

\[
I_n(\theta) = E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\} \quad \text{and} \quad I_1(\theta) = E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}.
\]

In the iid case, we have just proven that \( I_n(\theta) = nI(\theta) \). Therefore, in the iid case,

- If \( W(X) \) is an unbiased estimator of \( \tau(\theta) \), then
  \[
  \text{CRLB} = \frac{\left[ \tau'(\theta) \right]^2}{nI_1(\theta)}.
  \]

- If \( W(X) \) is an unbiased estimator of \( \tau(\theta) = \theta \), then
  \[
  \text{CRLB} = \frac{1}{nI_1(\theta)}.
  \]
Lemma 7.3.11 (Information Equality). Under regularity conditions,

\[ I_1(\theta) = E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\} = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right]. \]

Proof. From the definition of mathematical expectation,

\[ E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right] = \int_{\mathbb{R}} \frac{\partial^2}{\partial \theta^2} \ln f_X(x|\theta) f_X(x|\theta) dx = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \left[ \frac{\partial f_X(x|\theta)}{f_X(x|\theta)} \right] f_X(x|\theta) dx \]

Note: A sum replaces the integral above if \( X \) is discrete. The derivative

\[
\frac{\partial}{\partial \theta} \left[ \frac{\partial f_X(x|\theta)}{f_X(x|\theta)} \right] = \frac{\partial^2 f_X(x|\theta) f_X(x|\theta) - \frac{\partial}{\partial \theta} f_X(x|\theta) \frac{\partial}{\partial \theta} f_X(x|\theta)}{[f_X(x|\theta)]^2} \\
= \frac{\partial^2 f_X(x|\theta)}{f_X(x|\theta)} - \frac{\partial^2}{\partial \theta^2} f_X(x|\theta)
\]

Therefore, the last integral becomes

\[
\int_{\mathbb{R}} \left\{ \frac{\partial^2}{\partial \theta^2} f_X(x|\theta) - \frac{\partial^2}{\partial \theta^2} f_X(x|\theta)^2 \right\} f_X(x|\theta) dx = \int_{\mathbb{R}} \left\{ \frac{\partial^2}{\partial \theta^2} f_X(x|\theta) - \frac{\partial^2}{\partial \theta^2} f_X(x|\theta)^2 \right\} f_X(x|\theta) dx \\
= \int_{\mathbb{R}} \frac{\partial^2}{\partial \theta^2} f_X(x|\theta) dx - \int_{\mathbb{R}} \frac{\partial^2}{\partial \theta^2} f_X(x|\theta)^2 dx \\
= \frac{\partial^2}{\partial \theta^2} \int_{\mathbb{R}} f_X(x|\theta) dx - \int_{\mathbb{R}} \left[ \frac{\partial}{\partial \theta} \ln f_X(x|\theta) \right] f_X(x|\theta) dx \\
= -E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}.
\]

We have shown

\[ E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right] = -E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}. \]

Multiplying both sides by \(-1\) gives the information equality. □

Remark: We now finish by proving the attainment result.

Corollary 7.3.15. Suppose \( X_1, X_2, ..., X_n \) is an iid sample from \( f_X(x|\theta) \), where \( \theta \in \Theta \), a family that satisfies the regularity conditions stated for the Cramér-Rao Inequality. If \( W(X) \) is an unbiased estimator of \( \tau(\theta) \), then \( \text{var}_\theta[W(X)] \) attains the CRLB if and only if the score function

\[ S(\theta|x) = a(\theta)[W(x) - \tau(\theta)] \]

is a linear function of \( W(x) \).
Proof. From the CRLB proof, recall that we had

1. $W(X)$ playing the role of “$X$”
2. $\frac{\partial}{\partial \theta} \ln f_X(X|\theta)$ playing the role of “$Y$”

in applying the covariance inequality, which yields

$$\text{var}_\theta[W(X)] \geq \frac{[\tau'(\theta)]^2}{E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\}} \overset{\text{iid}}{=} \frac{[\tau'(\theta)]^2}{E_\theta \left\{ \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f_X(X_i|\theta) \right\}^2}.$$

Now, in the covariance inequality, we have equality when the correlation of $W(X)$ and $\frac{\partial}{\partial \theta} \ln f_X(X|\theta)$ equals ±1, which in turn implies

$$c(X - \mu_X) = Y - \mu_Y \quad \text{a.s.,}$$

or restated,

$$c[W(X) - \tau(\theta)] = \frac{\partial}{\partial \theta} \ln f_X(X|\theta) - 0 \quad \text{a.s.}$$

This is an application of Theorem 4.5.7 (CB, pp 172); i.e., two random variables are perfectly correlated if and only if the random variables are perfectly linearly related. In these equations, $c$ is a constant. Also, I have written “−0” on the RHS of the last equation to emphasize that

$$E_\theta \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = E_\theta \left[ \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f_X(X_i|\theta) \right] = 0.$$

Also, $W(X)$ is an unbiased estimator of $\tau(\theta)$ by assumption. Therefore, we have

$$c[W(X) - \tau(\theta)] = \frac{\partial}{\partial \theta} \ln f_X(X|\theta)$$
$$= \frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f_X(X_i|\theta)$$
$$= \frac{\partial}{\partial \theta} \ln L(\theta|X)$$
$$= S(\theta|X),$$

where $S(\theta|X)$ is the score function. The constant $c$ cannot depend on $W(X)$ nor on $\frac{\partial}{\partial \theta} \ln f_X(X|\theta)$, but it can depend on $\theta$. To emphasize this, we write

$$S(\theta|X) = a(\theta)[W(X) - \tau(\theta)].$$

Thus, var$_\theta[W(X)]$ attains the CRLB when the score function $S(\theta|X)$ can be written as a linear function of the unbiased estimator $W(X)$. $\square$
8 Hypothesis Testing

Complementary reading: Chapter 8 (CB).

8.1 Introduction

**Setting:** We observe \( X = (X_1, X_2, ..., X_n) \sim f_X(x|\theta) \), where \( \theta \in \Theta \subseteq \mathbb{R}^k \). For example, \( X_1, X_2, ..., X_n \) might constitute a random sample (iid sample) from a population \( f_X(x|\theta) \). We regard \( \theta \) as fixed and unknown.

**Definition:** A statistical hypothesis is a statement about \( \theta \). This statement specifies a collection of distributions that \( X \) can possibly have. Two complementary hypotheses in a testing problem are the null hypothesis

\[
H_0 : \theta \in \Theta_0
\]

and the alternative hypothesis

\[
H_1 : \theta \in \Theta_0^c,
\]

where \( \Theta_0^c = \Theta \setminus \Theta_0 \). We call \( \Theta_0 \) the null parameter space and \( \Theta_0^c \) the alternative parameter space.

**Example 8.1.** Suppose \( X_1, X_2, ..., X_n \) are iid \( \mathcal{N}(\theta, \sigma_0^2) \), where \( -\infty < \theta < \infty \) and \( \sigma_0^2 \) is known. Consider testing

\[
H_0 : \theta = \theta_0 \\
\text{versus} \\
H_1 : \theta \neq \theta_0,
\]

where \( \theta_0 \) is a specified value of \( \theta \). The null parameter space \( \Theta_0 = \{\theta_0\} \), a singleton. The alternative parameter space \( \Theta_0^c = \mathbb{R} \setminus \{\theta_0\} \).

**Terminology:** In Example 8.1, we call \( H_0 : \theta = \theta_0 \) a simple (or sharp) hypothesis. Note that \( H_0 \) specifies exactly one distribution, namely,

\[
\mathcal{N}(\theta_0, \sigma_0^2).
\]

A simple hypothesis specifies a single distribution.

**Terminology:** In Example 8.1, suppose we wanted to test

\[
H_0 : \theta \leq \theta_0 \\
\text{versus} \\
H_1 : \theta > \theta_0.
\]

We call \( H_0 \) a composite (or compound) hypothesis. Note that \( H_0 \) specifies a family of distributions, namely,

\[
\{\mathcal{N}(\theta, \sigma_0^2) : \theta \leq \theta_0\}.
\]
Goal: In a statistical hypothesis testing problem, we decide between the two complementary hypotheses $H_0$ and $H_1$ on the basis of observing $X = x$. In essence, a hypothesis test is a specification of the test function
\[
\phi(x) = P(\text{Reject } H_0 | X = x).
\]

Terminology: Let $\mathcal{X}$ denote the support of $X$.

- The subset of $\mathcal{X}$ for which $H_0$ is rejected is called the rejection region, denoted by $R$.
- The subset of $\mathcal{X}$ for which $H_0$ is not rejected is called the acceptance region, denoted by $R^c$.

If
\[
\phi(x) = I(x \in R) = \left\{ \begin{array}{ll}
1, & x \in R \\
0, & x \in R^c,
\end{array} \right.
\]
the test is said to be non-randomized.

Example 8.2. Suppose $X \sim \text{b}(10, \theta)$, where $0 < \theta < 1$, and consider testing

\[
H_0 : \theta \geq 0.35 \quad \text{versus} \quad H_1 : \theta < 0.35.
\]

Here is an example of a randomized test function:
\[
\phi(x) = \left\{ \begin{array}{ll}
1, & x \leq 2 \\
\frac{1}{5}, & x = 3 \\
0, & x \geq 4.
\end{array} \right.
\]

Using this test function, we would reject $H_0$ if $x = 0, 1, \text{ or } 2$. If $x = 3$, we would reject $H_0$ with probability $1/5$. If $x \geq 4$, we would not reject $H_0$.

- If we observed $x = 3$, we could then subsequently generate $U \sim \mathcal{U}(0, 1)$.
  - If $u \leq 0.2$, then reject $H_0$.
  - If $u > 0.2$, then do not reject $H_0$.

Remark: In most problems, a test function $\phi$ depends on $X$ through a one-dimensional test statistic, say
\[
W = W(X) = W(X_1, X_2, ..., X_n).
\]

1. We would like to work with test statistics that are sensible and confer tests with nice statistical properties (does sufficiency play a role?)
2. We would like to find the sampling distribution of $W$ under $H_0$ and $H_1$. 
Example 8.3. Suppose $X_1, X_2, ..., X_n$ are iid $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Consider testing

\[
H_0 : \sigma^2 = 40 \\
\text{versus} \\
H_1 : \sigma^2 \neq 40.
\]

In this problem, both

\[
W_1 = W_1(X) = |S^2 - 40|
\]
\[
W_2 = W_2(X) = \frac{(n - 1)S^2}{40}
\]

are reasonable test statistics.

- Because $S^2$ is an unbiased estimator of $\sigma^2$, large values of $W_1$ (intuitively) are evidence against $H_0$. However, what is $W_1$’s sampling distribution?

- The advantage of working with $W_2$ is that we know its sampling distribution when $H_0$ is true; i.e., $W_2 \sim \chi^2_{n-1}$. It is also easy to calculate the sampling distribution of $W_2$ when $H_0$ is not true; i.e., for values of $\sigma^2 \neq 40$.

Example 8.4. McCann and Tebbs (2009) summarize a study examining perceived unmet need for dental health care for people with HIV infection. Baseline in-person interviews were conducted with 2,864 HIV infected individuals (aged 18 years and older) as part of the HIV Cost and Services Utilization Study. Define

\[
X_1 = \text{number of patients with private insurance} \\
X_2 = \text{number of patients with medicare and private insurance} \\
X_3 = \text{number of patients without insurance} \\
X_4 = \text{number of patients with medicare but no private insurance}.
\]

Set $X = (X_1, X_2, X_3, X_4)$ and model $X \sim \text{mult}(2864, p_1, p_2, p_3, p_4; \sum_{i=1}^{4} p_i = 1)$. Under this assumption, consider testing

\[
H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4} \\
\text{versus} \\
H_1 : H_0 \text{ not true}.
\]

Note that an observation like $x = (0, 0, 0, 2864)$ should lead to a rejection of $H_0$. An observation like $x = (716, 716, 716, 716)$ should not. What about

\[
x = (658, 839, 811, 556)\)

Can we find a reasonable one-dimensional test statistic?
8.2 Methods of Finding Tests

Preview: The authors present three methods of finding tests:

1. Likelihood ratio tests (LRTs)
2. Bayesian tests
3. Union-Intersection and Intersection-Union tests (UIT/IUT)

We will focus largely on LRTs. We will discuss Bayesian tests briefly.

8.2.1 Likelihood ratio tests

Recall: Suppose $X = (X_1, X_2, ..., X_n) \sim f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$. The likelihood function is

$$L(\theta|x) = f_X(x|\theta) \prod_{i=1}^{n} f_X(x_i|\theta),$$

where $f_X(x|\theta)$ is the common population distribution (in the iid case). Recall that $\Theta$ is the parameter space.

Definition: The likelihood ratio test (LRT) statistic for testing

$$H_0 : \theta \in \Theta_0$$

versus

$$H_1 : \theta \in \Theta \setminus \Theta_0$$

is defined by

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)}.$$

A LRT is a test that has a rejection region of the form

$$R = \{x \in \mathcal{X} : \lambda(x) \leq c\},$$

where $0 \leq c \leq 1$.

Intuition: The numerator of $\lambda(x)$ is the largest the likelihood function can be over the null parameter space $\Theta_0$. The denominator is the largest the likelihood function can be over the entire parameter space $\Theta$. Clearly,

$$0 \leq \lambda(x) \leq 1.$$ 

The form of the rejection region above says to “reject $H_0$ when $\lambda(x)$ is too small.” When $\lambda(x)$ is small, the data $x$ are not consistent with the collection of models under $H_0$. 
Connection with MLEs:

- The numerator of $\lambda(x)$ is
  \[ \sup_{\theta \in \Theta_0} L(\theta|x) = L(\hat{\theta}_0|x), \]
  where $\hat{\theta}_0$ is the MLE of $\theta$ subject to the constraint that $\theta \in \Theta_0$. That is, $\hat{\theta}_0$ is the value of $\theta$ that maximizes $L(\theta|x)$ over the null parameter space $\Theta_0$. We call $\theta_0$ the restricted MLE.

- The denominator of $\lambda(x)$ is
  \[ \sup_{\theta \in \Theta} L(\theta|x) = L(\hat{\theta}|x), \]
  where $\hat{\theta}$ is the MLE of $\theta$. That is, $\hat{\theta}$ is the value of $\theta$ that maximizes $L(\theta|x)$ over the entire parameter space $\Theta$. We call $\hat{\theta}$ the unrestricted MLE.

- For notational simplicity, we often write
  \[ \lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}. \]
  This notation is easier and emphasizes how the definition of $\lambda(x)$ is tied to maximum likelihood estimation.

**Special case:** When $H_0$ is a simple hypothesis; i.e.,
\[ H_0 : \theta = \theta_0, \]
the null parameter space is $\Theta_0 = \{\theta_0\}$, a singleton. Clearly, in this case,
\[ \sup_{\theta \in \Theta_0} L(\theta|x) = L(\hat{\theta}_0|x) = L(\theta_0|x). \]
That is, there is only one value of $\theta$ “allowed” under $H_0$. We are therefore maximizing the likelihood function $L(\theta|x)$ over a single point in $\Theta$.

**Large-sample intuition:** We will learn in Chapter 10 that (under suitable regularity conditions), an MLE
\[ \hat{\theta} \xrightarrow{p} \theta, \quad \text{as } n \to \infty, \]
i.e., “MLEs are consistent” (I have switched to the scalar case here only for convenience). In the light of this asymptotic result, consider each of the following cases:

- Suppose that $H_0$ is true; i.e., $\theta \in \Theta_0$. Then
  \[ \hat{\theta}_0 \xrightarrow{p} \theta \quad \in \Theta_0 \]
  \[ \hat{\theta} \xrightarrow{p} \theta \quad \in \Theta_0. \]
  The MLEs $\hat{\theta}_0$ and $\hat{\theta}$ are converging to the same quantity (in probability) so they should be close to each other in large samples. Therefore, we would expect
  \[ \lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)} \]
  to be “close” to 1.
• Suppose $H_0$ is not true; i.e., $\theta \in \Theta \setminus \Theta_0$. Then

$$\hat{\theta} \xrightarrow{p} \theta \in \Theta \setminus \Theta_0,$$

but $\hat{\theta}_0 \in \Theta_0$ because $\hat{\theta}_0$ is calculated by maximizing $L(\theta|x)$ over $\Theta_0$ (i.e., $\hat{\theta}_0$ can never “escape from” $\Theta_0$). Therefore, there is no guarantee that $\hat{\theta}_0$ and $\hat{\theta}$ will be close to each other in large samples, and, in fact, the ratio

$$\lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}$$

could be much smaller than 1.

• This is why (at least by appealing to large-sample intuition) it makes sense to reject $H_0$ when $\lambda(x)$ is small.

**Example 8.5.** Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and $\sigma_0^2 = 1$. Consider testing

$$H_0 : \mu = \mu_0$$
versus

$$H_1 : \mu \neq \mu_0.$$

The likelihood function is

$$L(\mu|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}}$$

$$= \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i-\mu)^2}.$$ 

The relevant parameter spaces are

$$\Theta_0 = \{\mu_0\}, \text{ a singleton}$$
$$\Theta = \{\mu : -\infty < \mu < \infty\}.$$

Clearly,

$$\sup_{\mu \in \Theta_0} L(\mu|x) = L(\mu_0|x)$$

$$= \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i-\mu_0)^2}.$$ 

Over the entire parameter space $\Theta$, the MLE is $\hat{\mu} = \bar{X}$; see Example 7.5 (notes, pp 200). Therefore,

$$\sup_{\mu \in \Theta} L(\mu|x) = L(\bar{x}|x)$$

$$= \left( \frac{1}{\sqrt{2\pi}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} (x_i-\bar{x})^2}.$$
The LRT statistic is
\[ \lambda(x) = \frac{L(\mu_0|x)}{L(\overline{x}|x)} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \mu_0)^2}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i - \overline{x})^2}} = e^{-\frac{1}{2}\left[\sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \overline{x})^2\right]} . \]

Recall the algebraic identity
\[ \sum_{i=1}^n (x_i - \mu_0)^2 = \sum_{i=1}^n (x_i - \overline{x})^2 + n(\overline{x} - \mu_0)^2 . \]

Therefore, \( \lambda(x) \) reduces to
\[ \lambda(x) = e^{-\frac{n}{2}(\overline{x} - \mu_0)^2} . \]

An LRT rejects \( H_0 \) when \( \lambda(x) \) is “too small,” say, \( \lambda(x) \leq c \).

**Goal:** Write the rejection rule
\[ \lambda(x) \leq c \]
as a statement in terms of an easily-identified statistic. Note that
\[
\lambda(x) = e^{-\frac{n}{2}(\overline{x} - \mu_0)^2} \leq c \iff -\frac{n}{2}(\overline{x} - \mu_0)^2 \leq \ln c \\
\iff (\overline{x} - \mu_0)^2 \geq -\frac{2\ln c}{n} \\
\iff |\overline{x} - \mu_0| \geq \sqrt{-\frac{2\ln c}{n}} = c', \text{ say.}
\]

Therefore, the LRT rejection region can be written as
\[ R = \{ x \in \mathcal{X} : \lambda(x) \leq c \} = \{ x \in \mathcal{X} : |\overline{x} - \mu_0| \geq c' \} . \]

Rejecting \( H_0 \) when \( \lambda(x) \) is “too small” is the same as rejecting \( H_0 \) when \( |\overline{x} - \mu_0| \) is “too large.” The latter decision rule makes sense intuitively. Note that we have written our LRT rejection region and the corresponding test function
\[ \phi(x) = I(x \in R) = I(|\overline{x} - \mu_0| \geq c') = \begin{cases} 1, & |\overline{x} - \mu_0| \geq c' \\ 0, & |\overline{x} - \mu_0| < c' \end{cases} \]
in terms of the one-dimensional statistic \( W(X) = \overline{X} \). Recall that \( W(X) = \overline{X} \) is a sufficient statistic for the \( \mathcal{N}(\mu, 1) \) family.

**Example 8.6.** Suppose \( X_1, X_2, ..., X_n \) are iid with population pdf
\[ f_X(x|\theta) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & x < \theta, \end{cases} \]
where \(-\infty < \theta < \infty\). Note that this is a location exponential population pdf; the location parameter is \( \theta \). Consider testing
\( H_0 : \theta \leq \theta_0 \)
versus
\( H_1 : \theta > \theta_0 \).

The likelihood function is
\[
L(\theta | x) = \prod_{i=1}^{n} e^{-(x_i - \theta)} I(x_i \geq \theta)
\]
\[
= e^{-\sum_{i=1}^{n} x_i + n\theta} I(x(1) \geq \theta) \prod_{i=1}^{n} I(x_i \in \mathbb{R})
\]
\[
= e^{n\theta} I(x(1) \geq \theta) \prod_{i=1}^{n} I(x_i \in \mathbb{R})
\]
\[
= e^{n\theta} g(x(1) | \theta) \prod_{i=1}^{n} h(x_i)
\]

Note that \( W(X) = X(1) \) is a sufficient statistic by the Factorization Theorem. The relevant parameter spaces are
\[
\Theta_0 = \{ \theta : -\infty < \theta \leq \theta_0 \}
\]
\[
\Theta = \{ \theta : -\infty < \theta < \infty \}.
\]

We need to find the unrestricted MLE
\[
\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta | x)
\]
and the restricted MLE
\[
\hat{\theta}_0 = \arg \max_{\theta \in \Theta_0} L(\theta | x).
\]

**Unrestricted MLE:** Note that

- When \( \theta \leq x(1) \), \( L(\theta | x) = e^{-\sum_{i=1}^{n} x_i + n\theta} \), which increases as \( \theta \) increases.
  - For graphing purposes, it is helpful to note that
    \[
    \frac{\partial^2}{\partial \theta^2} L(\theta | x) = n^2 e^{-\sum_{i=1}^{n} x_i + n\theta} > 0,
    \]
    i.e., \( L(\theta | x) \) is convex.
- When \( \theta > x(1) \), \( L(\theta | x) = 0 \).
- Therefore, \( L(\theta | x) \) is an increasing function when \( \theta \) is less than or equal to the minimum order statistic \( x(1) \); when \( \theta \) is larger than \( x(1) \), the likelihood function drops to zero.
- Clearly, the unrestricted MLE of \( \theta \) is \( \hat{\theta} = X(1) \) and hence the denominator of \( \lambda(x) \) is
  \[
  \sup_{\theta \in \Theta} L(\theta | x) = L(\hat{\theta} | x) = L(x(1) | x).
  \]
Restricted MLE: By “restricted,” we mean “subject to the constraint that the estimate fall in $\Theta_0 = \{ \theta : -\infty < \theta \leq \theta_0 \}.”

- **Case 1:** If $\theta_0 < x_{(1)}$, then the largest $L(\theta|x)$ can be is $L(\theta_0|x)$. Therefore, the restricted MLE is $\hat{\theta}_0 = \theta_0$.

- **Case 2:** If $\theta_0 \geq x_{(1)}$, then the restricted MLE $\hat{\theta}_0$ coincides with the unrestricted MLE $\hat{\theta} = X_{(1)}$.

Therefore,

$$\hat{\theta}_0 = \begin{cases} \theta_0, & \theta_0 < X_{(1)} \\ X_{(1)}, & \theta_0 \geq X_{(1)}. \end{cases}$$

The LRT statistic is

$$\lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)} = \begin{cases} \frac{L(\theta_0|x)}{L(x_{(1)}|x)}, & \theta_0 < x_{(1)} \\ \frac{L(x_{(1)}|x)}{L(x_{(1)}|x)} = 1, & \theta_0 \geq x_{(1)}. \end{cases}$$

That $\lambda(x) = 1$ when $\theta_0 \geq x_{(1)}$ makes perfect sense in testing

$$H_0 : \theta \leq \theta_0$$

versus

$$H_1 : \theta > \theta_0.$$

- If $x_{(1)} \leq \theta_0$, we certainly don’t want to reject $H_0$ and conclude that $\theta > \theta_0$.

- It is only when $x_{(1)} > \theta_0$ do we have evidence that $\theta$ might be larger than $\theta_0$. The larger $x_{(1)}$ is ($x_{(1)} > \theta_0$), the smaller $\lambda(x)$ becomes; see Figure 8.2.1 (CB, pp 377). That is,

$$\text{larger } x_{(1)} \iff \text{smaller } \lambda(x) \iff \text{more evidence against } H_0.$$

Not surprisingly, we can write our LRT rejection region in terms of $W(X) = X_{(1)}$. When $\theta_0 < x_{(1)}$, the LRT statistic

$$\lambda(x) = \frac{L(\theta_0|x)}{L(x_{(1)}|x)} = \frac{e^{-\sum_{i=1}^{n} x_i + n\theta_0}}{e^{-\sum_{i=1}^{n} x_i + nx_{(1)}}} = e^{-n(x_{(1)} - \theta_0)}.$$ 

Note that

$$\lambda(x) = e^{-n(x_{(1)} - \theta_0)} \leq c \iff -n(x_{(1)} - \theta_0) \leq \ln c \iff x_{(1)} \geq \theta_0 - \frac{\ln c}{n} = c', \text{ say.}$$
Therefore, the LRT rejection region can be written as
\[ R = \{ x \in \mathcal{X} : \lambda(x) \leq c \} = \{ x \in \mathcal{X} : x^{(1)} \geq c' \}. \]

Rejecting \( H_0 \) when \( \lambda(x) \) is “too small” is the same as rejecting \( H_0 \) when \( x^{(1)} \) is “too large.” As noted earlier, the latter decision rule makes sense intuitively. Note that we have written our LRT rejection region and the corresponding test function
\[ \phi(x) = I(x \in R) = I(x^{(1)} \geq c') = \begin{cases} 1, & x^{(1)} \geq c' \\ 0, & x^{(1)} < c' \end{cases} \]
in terms of the one-dimensional statistic \( W(X) = X^{(1)} \), which is sufficient for the location exponential family.

**Theorem 8.2.4.** Suppose \( T = T(X) \) is a sufficient statistic for \( \theta \). If \( \lambda^*(T(x)) = \lambda^*(t) \) is the LRT statistic based on \( T \) and if \( \lambda(x) \) is the LRT statistic based on \( X \), then \( \lambda^*(T(x)) = \lambda(x) \) for all \( x \in \mathcal{X} \).

**Proof.** Because \( T = T(X) \) is sufficient, we can write (by the Factorization Theorem)
\[ f_X(x|\theta) = g_T(t|\theta)h(x), \]
where \( g_T(t|\theta) \) is the pdf (pmf) of \( T \) and \( h(x) \) is free of \( \theta \). Therefore,
\[
\lambda(x) = \frac{\sup_{\theta \in \Theta} L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)} = \frac{\sup_{\theta \in \Theta_0} g_T(t|\theta)h(x)}{\sup_{\theta \in \Theta} g_T(t|\theta)}
= \frac{\sup_{\theta \in \Theta_0} g_T(t|\theta)}{\sup_{\theta \in \Theta} g_T(t|\theta)}
= \frac{\sup_{\theta \in \Theta_0} L^*(\theta|t)}{\sup_{\theta \in \Theta} L^*(\theta|t)},
\]
where \( L^*(\theta|t) \) is the likelihood function based on observing \( T = t \). □

**Implication:** If a sufficient statistic \( T \) exists, we can immediately restrict attention to its distribution when deriving an LRT.

**Example 8.7.** Suppose \( X_1, X_2, ..., X_n \) are iid exponential(\( \theta \)), where \( \theta > 0 \). Consider testing
\[
H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.
\]

(a) Show that the LRT statistic based on \( X = x \) is
\[
\lambda(x) = \left( \frac{e}{n\theta_0} \right)^n \left( \sum_{i=1}^n x_i \right)^n e^{-\sum_{i=1}^n x_i/\theta_0}.
\]
(b) Show that the LRT statistic based on $T = T(X) = \sum_{i=1}^{n} X_i$ is

$$\lambda^*(t) = \left(\frac{e}{\theta_0}\right)^n t^n e^{-t/\theta_0},$$

establishing that $\lambda^*(t) = \lambda(x)$, as stated in Theorem 8.2.4.

(c) Show that

$$\lambda^*(t) \leq c \iff t \leq c_1 \text{ or } t \geq c_2,$$

for some $c_1$ and $c_2$ satisfying $c_1 < c_2$.

**Example 8.8.** Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\theta = (\mu, \sigma^2)$. Consider testing

$$H_0 : \mu = \mu_0$$

versus

$$H_1 : \mu \neq \mu_0.$$

The null hypothesis $H_0$ above looks simple, but it is not. The relevant parameter spaces are

$$\Theta_0 = \{\theta = (\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$$

and

$$\Theta = \{\theta = (\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}.$$

In this problem, we call $\sigma^2$ a **nuisance parameter**, because it is not the parameter that is of interest in $H_0$ and $H_1$. The likelihood function is

$$L(\theta|x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \sigma^2} e^{-(x_i-\mu)^2/2\sigma^2}$$

$$= \left(\frac{1}{2\pi \sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i-\mu)^2}.$$

**Unrestricted MLE:** In Example 7.6 (notes, pp 202), we showed that

$$\hat{\theta} = \begin{pmatrix} \bar{X} \\ S_b^2 \end{pmatrix} = \begin{pmatrix} \bar{X} \\ 1/n \sum_{i=1}^{n} (X_i - \bar{X})^2 \end{pmatrix}$$

maximizes $L(\theta|x)$ over $\Theta$.

**Restricted MLE:** It is easy to show that

$$\hat{\theta}_0 = \begin{pmatrix} \mu_0 \\ 1/n \sum_{i=1}^{n} (X_i - \mu_0)^2 \end{pmatrix}$$

maximizes $L(\theta|x)$ over $\Theta_0$. 
(a) Show that
\[ \lambda(x) = \frac{L(\hat{\theta}_0|x)}{L(\theta|x)} = \left[ \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2} \right]^{n/2}. \]

(b) Show that
\[ \lambda(x) \leq c \iff \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq c'. \]
This demonstrates that the “one-sample t test” is a LRT under normality.

**Exercise:** In Example 7.7 (notes, pp 203), derive the LRT statistic to test
\[ H_0 : p_1 = p_2 \]
versus
\[ H_1 : p_1 \neq p_2. \]

**Exercise:** In Example 8.4 (notes, pp 234), show that the LRT statistic is
\[ \lambda(x) = \lambda(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} \left( \frac{2864}{4x_i} \right)^{x_i}. \]
Also, show that
\[ \lambda(x) \leq c \iff -2 \ln \lambda(x) \geq c'. \]
Under \( H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}, \) we will learn later that \(-2 \ln \lambda(X)\) is distributed approximately as \( \chi^2_3. \) This suggests a “large-sample” LRT, namely, to reject \( H_0 \) if \(-2 \ln \lambda(x)\) is “too large.” We can use the \( \chi^2_3 \) distribution to specify what “too large” actually means.

### 8.2.2 Bayesian tests

**Remark:** Hypothesis tests of the form
\[ H_0 : \theta \in \Theta_0 \]
versus
\[ H_1 : \theta \in \Theta_c, \]
where \( \Theta_c = \Theta \setminus \Theta_0, \) can also be carried out within the Bayesian paradigm, but they are performed differently. Recall that, for a Bayesian, all inference is carried out using the posterior distribution \( \pi(\theta|x). \)

**Realization:** The posterior distribution \( \pi(\theta|x) \) is a valid probability distribution. It is the distribution that describes the behavior of the random variable \( \theta \), updated after observing the data \( x \). In this light, the probabilities
\[ P(H_0 \text{ true}|x) = P(\theta \in \Theta_0|x) = \int_{\Theta_0} \pi(\theta|x)d\theta \]
\[ P(H_1 \text{ true}|x) = P(\theta \in \Theta_c|x) = \int_{\Theta_c} \pi(\theta|x)d\theta \]
make perfect sense and be calculated (or approximated) “exactly.” Note that these probabilities make no sense to the non-Bayesian. S/he regards \( \theta \) as fixed, so that \( \{ \theta \in \Theta_0 \} \) and \( \{ \theta \in \Theta_0^c \} \) are not random events. We do not assign probabilities to events that are not random.

**Example 8.9.** Suppose that \( X_1, X_2, ..., X_n \) are iid Poisson(\( \theta \)), where the prior distribution for \( \theta \sim \text{gamma}(a, b) \), \( a, b \) known. In Example 7.10 (notes, pp 207-208), we showed that the posterior distribution

\[
\theta | X = x \sim \text{gamma} \left( \sum_{i=1}^{n} x_i + a, \frac{1}{n + \frac{1}{b}} \right).
\]

As an application, consider the following data, which summarize the number of goals per game in the 2013-2014 English Premier League season:

<table>
<thead>
<tr>
<th>Goals</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>72</td>
</tr>
<tr>
<td>4</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>10+</td>
<td>0</td>
</tr>
</tbody>
</table>

There were \( n = 380 \) games total. I modeled the number of goals per game \( X \) as a Poisson random variable and assumed that \( X_1, X_2, ..., X_{380} \) are iid Poisson(\( \theta \)). Before the season started, I modeled the mean number of goals per game as \( \theta \sim \text{gamma}(1.5, 2) \), which is a fairly diffuse prior distribution. Based on the observed data, I used R to calculate

\[
\text{sum(goals)}
\]

[1] 1060

The posterior distribution is therefore

\[
\theta | X = x \sim \text{gamma} \left( 1060 + 1.5, \frac{1}{380 + \frac{1}{2}} \right),
\]

\[\pi(\theta|x) \overset{d}{=} \text{gamma}(1061.5, 0.002628).\]

I have depicted the prior distribution \( \pi(\theta) \) and the posterior distribution \( \pi(\theta|x) \) in Figure 8.1 (next page). Suppose that I wanted to test

\[
H_0 : \theta \geq 3 \\
\text{versus} \\
H_1 : \theta < 3
\]

on the basis of the assumed Bayesian model and the observed data \( x \). The probability that \( H_0 \) is true is

\[
P(\theta \geq 3|x) = \int_{3}^{\infty} \pi(\theta|x)d\theta \approx 0.008,
\]

which I calculated in R using

\[
\text{1-pgamma}(3, 1061.5, 1/0.002628)
\]

[1] 0.008019202
8.3 Methods of Evaluating Tests

Setting: Suppose \( X = (X_1, X_2, ..., X_n) \sim f_X(x|\theta) \), where \( \theta \in \Theta \subseteq \mathbb{R} \) and consider testing

\[
H_0 : \theta \in \Theta_0 \\
\text{versus} \\
H_1 : \theta \in \Theta_0^c,
\]

where \( \Theta_0^c = \Theta \setminus \Theta_0 \). I will henceforth assume that \( \theta \) is a scalar parameter (for simplicity only).

8.3.1 Error probabilities and the power function

Definition: For a test (with test function)

\[
\phi(x) = I(x \in R),
\]

we can make one of two mistakes:

1. Type I Error: Rejecting \( H_0 \) when \( H_0 \) is true
2. Type II Error: Not rejecting \( H_0 \) when \( H_1 \) is true.
Therefore, for any test that we perform, there are four possible scenarios, described in the following table:

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>Reject $H_0$</td>
<td>Do not reject $H_0$</td>
</tr>
<tr>
<td>$H_1$</td>
<td>Type I Error</td>
<td>Type II Error</td>
</tr>
</tbody>
</table>

Calculations:

1. Suppose $H_0 : \theta \in \Theta_0$ is true. For $\theta \in \Theta_0$,
   \[
P(\text{Type I Error}|\theta) = P_\theta(X \in R) = E_\theta[I(X \in R)] = E_\theta[\phi(X)].
   \]

2. Suppose $H_1 : \theta \in \Theta_0^c$ is true. For $\theta \in \Theta_0^c$,
   \[
P(\text{Type II Error}|\theta) = P_\theta(X \in R^c) = 1 - P_\theta(X \in R) = 1 - E_\theta[\phi(X)] = E_\theta[1 - \phi(X)].
   \]

It is very important to note that both of these probabilities depend on $\theta$. This is why we emphasize this in the notation.

**Definition:** The **power function** of a test $\phi(x)$ is the function of $\theta$ given by

\[
\beta(\theta) = P_\theta(X \in R) = E_\theta[\phi(X)].
\]

In other words, the power function gives the probability of rejecting $H_0$ for all $\theta \in \Theta$. Note that if $H_1$ is true, so that $\theta \in \Theta_0^c$,

\[
\beta(\theta) = P_\theta(X \in R) = 1 - P_\theta(X \in R^c) = 1 - P(\text{Type II Error}|\theta).
\]

**Example 8.10.** Suppose $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and $\sigma_0^2$ is known. Consider testing

\[
H_0 : \mu \leq \mu_0 \quad \text{versus} \quad H_1 : \mu > \mu_0.
\]

The LRT of $H_0$ versus $H_1$ uses the test function

\[
\phi(x) = \begin{cases} 
1, & \frac{\overline{x} - \mu_0}{\sigma_0/\sqrt{n}} \geq c \\
0, & \text{otherwise.}
\end{cases}
\]
The power function for this test is given by

\[
\beta(\mu) = P_\mu(X \in R)
\]

\[
= P_\mu\left(\frac{\bar{X} - \mu_0}{\sigma_0/\sqrt{n}} \geq c\right)
\]

\[
= P_\mu\left(\bar{X} \geq \frac{c\sigma_0}{\sqrt{n}} + \mu_0\right)
\]

\[
= P_\mu\left(\frac{\bar{X} - \mu}{\sigma_0/\sqrt{n}} \geq \frac{c\sigma_0}{\sigma_0/\sqrt{n}} + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right)
\]

\[
= 1 - F_Z\left(c + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right),
\]

where \(Z \sim \mathcal{N}(0, 1)\) and \(F_Z(\cdot)\) is the standard normal cdf.

**Exercise:** Determine \(n\) and \(c\) such that

\[
\sup_{\mu \leq \mu_0} \beta(\mu) = 0.10
\]

\[
\inf_{\mu \geq \mu_0 + \sigma_0} \beta(\mu) = 0.80.
\]

- The first requirement implies that \(P(\text{Type I Error} | \mu)\) will not exceed 0.10 for all \(\mu \leq \mu_0\) (\(H_0\) true).
- The second requirement implies that \(P(\text{Type II Error} | \mu)\) will not exceed 0.20 for all \(\mu \geq \mu_0 + \sigma_0\) (these are values of \(\mu\) that make \(H_1\) true).

**Solution.** Note that

\[
\frac{\partial}{\partial \mu} \beta(\mu) = \frac{\partial}{\partial \mu} \left[1 - F_Z\left(c + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right)\right]
\]

\[
= \frac{\sqrt{n}}{\sigma_0} f_Z\left(c + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}}\right) > 0;
\]

i.e., \(\beta(\mu)\) is an increasing function of \(\mu\). Therefore,

\[
\sup_{\mu \leq \mu_0} \beta(\mu) = \beta(\mu_0) = 1 - F_Z(c) \overset{\text{set}}{=} 0.10 \implies c = 1.28,
\]

the 0.90 quantile of the \(\mathcal{N}(0, 1)\) distribution. Also, because \(\beta(\mu)\) is increasing,

\[
\inf_{\mu \geq \mu_0 + \sigma_0} \beta(\mu) = \beta(\mu_0 + \sigma_0) = 1 - F_Z(1.28 - \sqrt{n}) \overset{\text{set}}{=} 0.80
\]

\[
\implies 1.28 - \sqrt{n} = -0.84
\]

\[
\implies n = 4.49,
\]

which would be rounded up to \(n = 5\). The resulting power function with \(c = 1.28\), \(n = 5\), \(\mu_0 = 1.5\) and \(\sigma_0 = 1\) is shown in Figure 8.2 (next page).
Figure 8.2: Power function $\beta(\mu)$ in Example 8.10 with $c = 1.28$, $n = 5$, $\mu_0 = 1.5$ and $\sigma_0 = 1$. Horizontal lines at 0.10 and 0.80 have been added.

**Definition:** A test $\phi(x)$ with power function $\beta(\theta)$ is a size $\alpha$ test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha.$$ 

The test $\phi(x)$ is a level $\alpha$ test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$ 

Note that if $\phi(x)$ is a size $\alpha$ test, then it is also level $\alpha$. The converse is not true. In other words,

$$\{\text{class of size } \alpha \text{ tests}\} \subset \{\text{class of level } \alpha \text{ tests}\}.$$ 

**Remark:** Often, it is unnecessary to differentiate between the two classes of tests. However, in testing problems involving discrete distributions (e.g., binomial, Poisson, etc.), it is generally not possible to construct a size $\alpha$ test for a specified value of $\alpha$; e.g., $\alpha = 0.05$. Thus (unless one randomizes), we may have to settle for a level $\alpha$ test.

**Important:** As the definition above indicates, the size of any test $\phi(x)$ is calculated by maximizing the power function over the null parameter space $\Theta_0$ identified in $H_0$. 

---

**PAGE 248**
Example 8.11. Suppose $X_1, X_2$ are iid Poisson($\theta$), where $\theta > 0$, and consider testing

$$H_0 : \theta \geq 3$$

versus

$$H_1 : \theta < 3.$$ 

We consider the two tests

$$\phi_1 = \phi_1(x_1, x_2) = I(x_1 = 0)$$

$$\phi_2 = \phi_2(x_1, x_2) = I(x_1 + x_2 \leq 1).$$

The power function for the first test is

$$\beta_1(\theta) = E_\theta[I(X_1 = 0)] = P_\theta(X_1 = 0) = e^{-\theta}.$$ 

Recall that $T = T(X_1, X_2) = X_1 + X_2 \sim \text{Poisson}(2\theta)$. The power function for the second test is

$$\beta_2(\theta) = E_\theta[I(X_1 + X_2 \leq 1)] = P_\theta(X_1 + X_2 \leq 1) = e^{-2\theta} + 2\theta e^{-2\theta}.$$ 

I have plotted both power functions in Figure 8.3 (next page).

Size calculations: The size of each test is calculated as follows. For the first test,

$$\alpha = \sup_{\theta \geq 3} \beta_1(\theta) = \beta_1(3) = e^{-3} \approx 0.049787.$$ 

For the second test,

$$\alpha = \sup_{\theta \geq 3} \beta_2(\theta) = \beta_2(3) = e^{-6} + 6e^{-6} \approx 0.017351.$$ 

Both $\phi_1$ and $\phi_2$ are level $\alpha = 0.05$ tests.

Example 8.12. Suppose $X_1, X_2, \ldots, X_n$ are iid from

$$f_X(x|\theta) = e^{-(x-\theta)}I(x \geq \theta),$$

where $-\infty < \theta < \infty$. In Example 8.6 (notes, pp 238-241), we considered testing

$$H_0 : \theta \leq \theta_0$$

versus

$$H_1 : \theta > \theta_0$$

and derived the LRT to take the form

$$\phi(x) = I(x_{(1)} \geq c').$$

Find the value of $c'$ that makes $\phi(x)$ a size $\alpha$ test.
Solution. The pdf of $X(1)$ is

$$f_{X(1)}(x|\theta) = ne^{-n(x-\theta)}I(x \geq \theta).$$

We set

$$\alpha = \sup_{\theta \leq \theta_0} E_\theta[\phi(X)] = \sup_{\theta \leq \theta_0} P_\theta(X(1) \geq c')$$

$$= \sup_{\theta \leq \theta_0} \int_{c'}^\infty ne^{-n(x-\theta)}dx$$

$$= \sup_{\theta \leq \theta_0} e^{-n(c'-\theta)}$$

$$= e^{-n(c'-\theta_0)}.$$

Therefore, $c' = \theta_0 - n^{-1} \ln \alpha$. A size $\alpha$ LRT uses

$$\phi(x) = I(x(1) \geq \theta_0 - n^{-1} \ln \alpha).$$

Exercise: Take $\theta_0 = 1$ and $\alpha = 0.05$. Find the smallest $n$ to guarantee $\beta(1.2) = 0.8$. 

Figure 8.3: Power functions $\beta_1(\theta)$ and $\beta_2(\theta)$ in Example 8.11.
8.3.2 Most powerful tests

**Definition:** Let \( C \) be a class of tests for testing

\[
H_0 : \theta \in \Theta_0 \\
\text{versus} \\
H_1 : \theta \in \Theta_0^c,
\]

where \( \Theta_0^c = \Theta \setminus \Theta_0 \). A test in \( C \) with power function \( \beta(\theta) \) is a **uniformly most powerful (UMP)** class \( C \) test if

\[
\beta(\theta) \geq \beta^*(\theta) \quad \text{for all } \theta \in \Theta_0^c,
\]

where \( \beta^*(\theta) \) is the power function of any other test in \( C \). The “uniformly” part in this definition refers to the fact that the power function \( \beta(\theta) \) is larger than (i.e., at least as large as) the power function of any other class \( C \) test for all \( \theta \in \Theta_0^c \).

**Important:** In this course, we will restrict attention to tests \( \phi(x) \) that are level \( \alpha \) tests. That is, we will take

\[
C = \{ \text{all level } \alpha \text{ tests} \}.
\]

This restriction is analogous to the restriction we made in the “optimal estimation problem” in Chapter 7. Recall that we restricted attention to unbiased estimators first; we then wanted to find the one with the smallest variance (uniformly, for all \( \theta \in \Theta \)). In the same spirit, we make the same type of restriction here by considering only those tests that are level \( \alpha \) tests. This is done so that we can avoid having to consider “silly tests,” e.g.,

\[
\phi(x) = 1 \quad \text{for all } x \in X.
\]

The power function for this test is \( \beta(\theta) = 1 \), for all \( \theta \in \Theta \). This test cannot be beaten in terms of power when \( H_1 \) is true! Unfortunately, it is not a very good test when \( H_0 \) is true.

**Recall:** A test \( \phi(x) \) with power function \( \beta(\theta) \) is a **level** \( \alpha \) test if

\[
\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.
\]

That is, \( P(\text{Type I Error}|\theta) \) can be **no larger** than \( \alpha \) for all \( \theta \in \Theta_0 \).

**Starting point:** We start by considering the simple-versus-simple test:

\[
H_0 : \theta = \theta_0 \\
\text{versus} \\
H_1 : \theta = \theta_1.
\]

Both \( H_0 \) and \( H_1 \) specify exactly one probability distribution.

**Remark:** This type of test is rarely of interest in practice. However, it is the “building block” situation for more interesting problems.
Theorem 8.3.12 (Neyman-Pearson Lemma). Consider testing

\[ H_0 : \theta = \theta_0 \]

versus

\[ H_1 : \theta = \theta_1 \]

and denote by \( f_X(x|\theta_0) \) and \( f_X(x|\theta_1) \) the pdfs (pmfs) of \( X = (X_1, X_2, ..., X_n) \) corresponding to \( \theta_0 \) and \( \theta_1 \), respectively. Consider the test function

\[
\phi(x) = \begin{cases} 
1, & \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > k \\
0, & \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} < k,
\end{cases}
\]

for \( k \geq 0 \), where

\[ \alpha = P_{\theta_0}(X \in R) = E_{\theta_0}[\phi(X)]. \]  

(8.1)

Sufficiency: Any test satisfying the definition of \( \phi(x) \) above and Equation (8.1) is a most powerful (MP) level \( \alpha \) test.

Remarks:

- The necessity part of the Neyman-Pearson (NP) Lemma is less important for our immediate purposes (see CB, pp 388).

- In a simple-versus-simple test, any MP level \( \alpha \) test is obviously also UMP level \( \alpha \). Recall that the “uniformly” part in UMP refers to all \( \theta \in \Theta_0 \). However, in a simple \( H_1 \), there is only one value of \( \theta \in \Theta_0 \). I choose to distinguish MP from UMP in this situation (whereas the authors of CB do not).

Example 8.13. Suppose that \( X_1, X_2, ..., X_n \) are iid beta(\( \theta, 1 \)), where \( \theta > 0 \); i.e., the population pdf is

\[ f_X(x|\theta) = \theta x^{\theta - 1}I(0 < x < 1). \]

Derive the MP level \( \alpha \) test for

\[ H_0 : \theta = 1 \]

versus

\[ H_1 : \theta = 2. \]

Solution. The pdf of \( X = (X_1, X_2, ..., X_n) \) is, for \( 0 < x_i < 1 \),

\[
f_X(x|\theta) \overset{\text{iid}}{=} \prod_{i=1}^{n} \theta x_i^{\theta - 1} \\
= \theta^n \left( \prod_{i=1}^{n} x_i \right)^{\theta - 1}.
\]
Form the ratio

\[
\frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} = \frac{2^n \left(\prod_{i=1}^{n} x_i\right)^{2-1}}{1^n \left(\prod_{i=1}^{n} x_i\right)^{1-1}} = 2^n \prod_{i=1}^{n} x_i.
\]

The NP Lemma says that the MP level \(\alpha\) test uses the rejection rule

\[
R = \left\{ x \in \mathcal{X} : 2^n \prod_{i=1}^{n} x_i > k \right\},
\]

where the constant \(k\) satisfies

\[
\alpha = P_{\theta=1}(X \in R) = P \left( 2^n \prod_{i=1}^{n} X_i > k \mid \theta = 1 \right).
\]

Instead of finding the constant \(k\) that satisfies this equation, we rewrite the rejection rule \(\{2^n \prod_{i=1}^{n} x_i > k\}\) in a way that makes our life easier. Note that

\[
2^n \prod_{i=1}^{n} x_i > k \iff \prod_{i=1}^{n} x_i > 2^{-n}k \iff \sum_{i=1}^{n} -\ln x_i < -\ln(2^{-n}k) = k', \text{ say.}
\]

We have rewritten the rejection rule \(\{2^n \prod_{i=1}^{n} x_i > k\}\) as \(\{\sum_{i=1}^{n} -\ln x_i < k'\}\). Therefore,

\[
\alpha = P \left( 2^n \prod_{i=1}^{n} X_i > k \mid \theta = 1 \right) = P \left( \sum_{i=1}^{n} -\ln X_i < k' \mid \theta = 1 \right).
\]

We have now changed the problem to choosing \(k'\) to solve this equation above.

**Q:** Why did we do this?

**A:** Because it is easier to find the distribution of \(\sum_{i=1}^{n} -\ln X_i\) when \(H_0 : \theta = 1\) is true.

Recall that

\[
X_i \overset{H_0}{\sim} \mathcal{U}(0, 1) \implies -\ln X_i \overset{H_0}{\sim} \text{exponential}(1)
\]

\[
\implies \sum_{i=1}^{n} -\ln X_i \overset{H_0}{\sim} \text{gamma}(n, 1).
\]

Therefore, to satisfy the equation above, we take \(k' = g_{n,1,1-\alpha}\), the (lower) \(\alpha\) quantile of a gamma\((n, 1)\) distribution. This notation for quantiles is consistent with how CB have defined them on pp 386.
Summary: The MP level $\alpha$ test of

$$H_0 : \theta = 1$$

\[\text{versus}\]

$$H_1 : \theta = 2$$

in Example 8.13 has the rejection region

$$R = \left\{ x \in \mathcal{X} : \sum_{i=1}^{n} -\ln x_i < g_{n,1,1-\alpha} \right\}.$$

Special case: If $n = 10$ and $\alpha = 0.05$, then $g_{10,1,0.95} \approx 5.425$.

$$> \texttt{qgamma}(0.05,10,1)$$

[1] 5.425406

Q: What is $\beta(2)$, the power of this MP test (when $n = 10$ and $\alpha = 0.05$)?

A: We calculate

$$\beta(2) = P \left( \sum_{i=1}^{10} -\ln X_i < 5.425 \mid \theta = 2 \right).$$

Recall that

$$X_i \overset{H_1}{\sim} \text{beta}(2,1) \implies -\ln X_i \overset{H_1}{\sim} \text{exponential}(1/2)$$

$$\implies \sum_{i=1}^{10} -\ln X_i \overset{H_1}{\sim} \text{gamma}(10, 1/2).$$

Therefore,

$$\beta(2) = \int_{0}^{5.425} \frac{1}{\Gamma(10) \left(\frac{1}{2}\right)^{10}} u^9 e^{-2u} du \approx 0.643.$$

$$> \texttt{pgamma}(5.425,10,1/0.5)$$

[1] 0.6429912

Exercise: With $\alpha = 0.05$, find the smallest $n$ that confers $\beta(2) = 0.90$. Ans: $n = 19$.

Proof of NP Lemma. We prove the sufficiency part only. Define the test function

$$\phi(x) = \begin{cases} 
  1, & \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} > k; \\
  0, & \frac{f_X(x|\theta_1)}{f_X(x|\theta_0)} < k;
\end{cases}$$

where $k \geq 0$ and

$$\alpha = P_{\theta_0}(X \in R) = E_{\theta_0}[\phi(X)];$$
i.e., $\phi(x)$ is a size $\alpha$ test. We want to show that $\phi(x)$ is MP level $\alpha$. Therefore, let $\phi^*(x)$ be the test function for any other level $\alpha$ test of $H_0$ versus $H_1$. Note that

$$
E_{\theta_0}[\phi(X)] = \alpha
$$

$$
E_{\theta_0}[\phi^*(X)] \leq \alpha.
$$

Thus,

$$
E_{\theta_0}[\phi(X) - \phi^*(X)] = E_{\theta_0}[\phi(X)] - E_{\theta_0}[\phi^*(X)] \geq 0.
$$

Define

$$
b(x) = [\phi(x) - \phi^*(x)][f_X(x|\theta_1) - k f_X(x|\theta_0)].
$$

We want to show that $b(x) \geq 0$, for all $x \in \mathcal{X}$.

- **Case 1:** Suppose $f_X(x|\theta_1) - k f_X(x|\theta_0) > 0$. Then, by definition, $\phi(x) = 1$. Because $0 \leq \phi^*(x) \leq 1$, we have

$$
b(x) = [\phi(x) - \phi^*(x)][f_X(x|\theta_1) - k f_X(x|\theta_0)] \geq 0.
$$

- **Case 2:** Suppose $f_X(x|\theta_1) - k f_X(x|\theta_0) < 0$. Then, by definition, $\phi(x) = 0$. Because $0 \leq \phi^*(x) \leq 1$, we have

$$
b(x) = [\phi(x) - \phi^*(x)][f_X(x|\theta_1) - k f_X(x|\theta_0)] \geq 0.
$$

- **Case 3:** Suppose $f_X(x|\theta_1) - k f_X(x|\theta_0) = 0$. It is then obvious that $b(x) = 0$.

We have shown that $b(x) = [\phi(x) - \phi^*(x)][f_X(x|\theta_1) - k f_X(x|\theta_0)] \geq 0$. Therefore,

$$
[\phi(x) - \phi^*(x)] f_X(x|\theta_1) - k[\phi(x) - \phi^*(x)] f_X(x|\theta_0) \geq 0
$$

$$
\iff [\phi(x) - \phi^*(x)] f_X(x|\theta_1) \geq k[\phi(x) - \phi^*(x)] f_X(x|\theta_0).
$$

Integrating both sides, we get

$$
\int_{\mathbb{R}^n} [\phi(x) - \phi^*(x)] f_X(x|\theta_1)dx \geq k \int_{\mathbb{R}^n} [\phi(x) - \phi^*(x)] f_X(x|\theta_0)dx,
$$

that is,

$$
E_{\theta_1}[\phi(X) - \phi^*(X)] \geq k E_{\theta_0}[\phi(X) - \phi^*(X)] \geq 0.
$$

Therefore, $E_{\theta_1}[\phi(X) - \phi^*(X)] \geq 0$ and hence $E_{\theta_1}[\phi(X)] \geq E_{\theta_1}[\phi^*(X)]$. This shows that $\phi(x)$ is more powerful than $\phi^*(x)$. Because $\phi^*(x)$ is an arbitrary level $\alpha$ test, we are done. \qed
Corollary 8.3.13 (NP Lemma with a sufficient statistic $T$). Consider testing

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta = \theta_1,$$

and suppose that $T = T(X)$ is a sufficient statistic. Denote by $g_T(t|\theta_0)$ and $g_T(t|\theta_1)$ the pdfs (pmfs) of $T$ corresponding to $\theta_0$ and $\theta_1$, respectively. Consider the test function

$$\phi(t) = \begin{cases} 1, & g_T(t|\theta_1) > k \\ 0, & g_T(t|\theta_1) < k, \end{cases}$$

for $k \geq 0$, where, with rejection region $S \subset T$,

$$\alpha = P_{\theta_0}(T \in S) = E_{\theta_0}[\phi(T)].$$

The test that satisfies these specifications is a MP level $\alpha$ test.

Proof. See CB (pp 390).

Implication: In search of a MP test, we can immediately restrict attention to those tests based on a sufficient statistic.

Example 8.14. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2_0)$, where $-\infty < \mu < \infty$ and $\sigma^2_0$ is known. Find the MP level $\alpha$ test for

$$H_0 : \mu = \mu_0$$

versus

$$H_1 : \mu = \mu_1,$$

where $\mu_1 < \mu_0$.

Solution. The sample mean $T = T(X) = \bar{X}$ is a sufficient statistic for the $\mathcal{N}(\mu, \sigma^2_0)$ family. Furthermore,

$$T \sim \mathcal{N} \left( \mu, \frac{\sigma^2_0}{n} \right) \implies g_T(t|\mu) = \frac{1}{\sqrt{2\pi\sigma^2_0/n}} e^{-\frac{n}{2\sigma^2_0} (t-\mu)^2},$$

for $t \in \mathbb{R}$. Form the ratio

$$\frac{g_T(t|\mu_1)}{g_T(t|\mu_0)} = \frac{1}{\sqrt{2\pi\sigma^2_0/n}} e^{-\frac{n}{2\sigma^2_0} (t-\mu_1)^2} \cdot \frac{1}{\sqrt{2\pi\sigma^2_0/n}} e^{-\frac{n}{2\sigma^2_0} (t-\mu_0)^2} = e^{-\frac{n}{2\sigma^2_0} [(t-\mu_1)^2 - (t-\mu_0)^2]}.$$

Corollary 8.3.13 says that the MP level $\alpha$ test rejects $H_0$ when

$$e^{-\frac{n}{2\sigma^2_0} [(t-\mu_1)^2 - (t-\mu_0)^2]} > k \iff t < \frac{2\sigma^2_0 n^{-1} \ln k - (\mu_1^2 - \mu_0^2)}{2(\mu_0 - \mu_1)} = k',$$

say.
Therefore, the MP level $\alpha$ test uses the rejection region
\[ S = \left\{ t \in T : \frac{g_T(t|\theta_1)}{g_T(t|\theta_0)} > k \right\} = \left\{ t \in T : t < k' \right\}, \]
where $k'$ satisfies
\[ \alpha = P_{\theta_0}(T < k') = P\left( Z < \frac{k' - \mu_0}{\sigma_0/\sqrt{n}} \right) \]
\[ \Rightarrow \frac{k' - \mu_0}{\sigma_0/\sqrt{n}} = -z_\alpha \]
\[ \Rightarrow k' = \mu_0 - z_\alpha \sigma_0/\sqrt{n}. \]
Therefore, the MP level $\alpha$ test rejects $H_0$ when $X < \mu_0 - z_\alpha \sigma_0/\sqrt{n}$.

8.3.3 Uniformly most powerful tests

Remark: So far, we have discussed “test related optimality” in the context of simple-versus-simple hypotheses. We now extend the idea of “most powerful” to more realistic situations involving composite hypotheses; e.g., $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

Definition: A family of pdfs (pmfs) $\{g_T(t|\theta) ; \theta \in \Theta\}$ for a univariate random variable $T$ has **monotone likelihood ratio (MLR)** if for all $\theta_2 > \theta_1$, the ratio
\[ \frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} \]
is a nondecreasing function of $t$ over the set $\{ t : g_T(t|\theta_1) > 0 \text{ or } g_T(t|\theta_2) > 0 \}$.

Example 8.15. Suppose $T \sim b(n, \theta)$, where $0 < \theta < 1$. The pmf of $T$ is
\[ g_T(t|\theta) = \binom{n}{t} \theta^t (1-\theta)^{n-t}, \]
for $t = 0, 1, 2, ..., n$. Suppose $\theta_2 > \theta_1$. Consider
\[ \frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} = \binom{n}{t} \theta_2^t (1-\theta_2)^{n-t} \binom{n}{t} \theta_1^t (1-\theta_1)^{n-t} = \left( \frac{1-\theta_2}{1-\theta_1} \right)^n = \left( \frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} \right)^t. \]
Note that $\left( \frac{1-\theta_2}{1-\theta_1} \right)^n > 0$ and is free of $t$. Also, because $\theta_2 > \theta_1$, both
\[ \frac{\theta_2}{\theta_1} > 1 \text{ and } \frac{1-\theta_1}{1-\theta_2} > 1. \]
Therefore,
\[ \frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} = c(\theta_1, \theta_2) a^t, \]
where $a > 1$. This is an increasing function of $t$ over $\{ t : t = 0, 1, 2, ..., n \}$. Therefore, the family $\{g_T(t|\theta) : 0 < \theta < 1\}$ has MLR.
Remark: Many common families of pdfs (pmfs) have MLR. For example, if
\[ T \sim g_T(t|\theta) = h(t)c(\theta)e^{w(\theta)t}, \]
i.e., \( T \) has pdf (pmf) in the one-parameter exponential family, then \( \{g_T(t|\theta); \theta \in \Theta\} \) has MLR if \( w(\theta) \) is a nondecreasing function of \( \theta \).

Proof. Exercise.

Q: Why is MLR useful?
A: It makes getting UMP tests easy.

Theorem 8.3.17 (Karlin-Rubin). Consider testing

\[
H_0 : \theta \leq \theta_0 \\
versus \\
H_1 : \theta > \theta_0.
\]

Suppose that \( T \) is sufficient. Suppose that \( \{g_T(t|\theta); \theta \in \Theta\} \) has MLR. The test that rejects \( H_0 \) iff \( T > t_0 \) is a UMP level \( \alpha \) test, where
\[
\alpha = P_{\theta_0}(T > t_0).
\]

Similarly, when testing

\[
H_0 : \theta \geq \theta_0 \\
versus \\
H_1 : \theta < \theta_0,
\]

the test that rejects \( H_0 \) iff \( T < t_0 \) is UMP level \( \alpha \), where \( \alpha = P_{\theta_0}(T < t_0) \).

Example 8.16. Suppose \( X_1, X_2, ..., X_n \) are iid Bernoulli(\( \theta \)), where \( 0 < \theta < 1 \), and consider testing

\[
H_0 : \theta \leq \theta_0 \\
versus \\
H_1 : \theta > \theta_0.
\]

We know that
\[ T = \sum_{i=1}^{n} X_i \]
is a sufficient statistic and \( T \sim b(n, \theta) \). In Example 8.15, we showed that the family \( \{g_T(t|\theta) : 0 < \theta < 1\} \) has MLR. Therefore, the Karlin-Rubin Theorem says that the UMP level \( \alpha \) test is
\[
\phi(t) = I(t > t_0),
\]
where \( t_0 \) solves
\[
\alpha = P_{\theta_0}(T > t_0) = \sum_{t=[t_0]+1}^{n} \binom{n}{t} \theta^t (1-\theta)^{n-t}.
\]
Figure 8.4: Power function $\beta(\theta)$ for the UMP level $\alpha = 0.0611$ test in Example 8.16 with $n = 30$ and $\theta_0 = 0.2$. A horizontal line at $\alpha = 0.0611$ has been added.

**Special case:** I took $n = 30$ and $\theta_0 = 0.2$. I used R to calculate the following:

<table>
<thead>
<tr>
<th>$t_0$</th>
<th>$P_{\theta_0}(T \geq [t_0] + 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$7 \leq t_0 &lt; 8$</td>
<td>$P(T \geq 8</td>
</tr>
<tr>
<td>$8 \leq t_0 &lt; 9$</td>
<td>$P(T \geq 9</td>
</tr>
<tr>
<td>$9 \leq t_0 &lt; 10$</td>
<td>$P(T \geq 10</td>
</tr>
<tr>
<td>$10 \leq t_0 &lt; 11$</td>
<td>$P(T \geq 11</td>
</tr>
<tr>
<td>$11 \leq t_0 &lt; 12$</td>
<td>$P(T \geq 12</td>
</tr>
</tbody>
</table>

Therefore, the UMP level $\alpha = 0.0611$ test of $H_0 : \theta \leq 0.2$ versus $H_1 : \theta > 0.2$ uses $I(t \geq 10)$. The UMP level $\alpha = 0.0256$ test uses $I(t \geq 11)$. Note that (without randomizing) it is not possible to write a UMP level $\alpha = 0.05$ test in this problem. For the level $\alpha = 0.0611$ test, the power function is

$$\beta(\theta) = P_{\theta}(T \geq 10) = \sum_{t=10}^{30} \binom{30}{t} \theta^t(1-\theta)^{30-t},$$

which is depicted in Figure 8.4 (above).
Example 8.17. Suppose that $X_1, X_2, \ldots, X_n$ are iid with population distribution
\[ f_X(x|\theta) = \theta e^{-\theta x} I(x > 0), \]
where $\theta > 0$. Note that this population distribution is an exponential distribution with mean $1/\theta$. Derive the UMP level $\alpha$ test for
\[ H_0 : \theta \geq \theta_0 \]
versus
\[ H_1 : \theta < \theta_0. \]

Solution. It is easy to show that
\[ T = \sum_{i=1}^{n} X_i \]
is a sufficient statistic and $T \sim \text{gamma}(n, 1/\theta)$. Suppose $\theta_2 > \theta_1$ and form the ratio
\[ \frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} = \frac{1}{\Gamma(n)} \frac{\Gamma(n-1, \frac{1}{\theta_2} t)}{\Gamma(n-1, \frac{1}{\theta_1} t)} = \left( \frac{\theta_2}{\theta_1} \right)^n e^{-t(\theta_2-\theta_1)}. \]
Because $\theta_2 - \theta_1 > 0$, we see that the ratio
\[ \frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} \]
is a decreasing function of $t$ over $\{t : t > 0\}$. However, the ratio is an increasing function of $t^* = -t$, and $T^* = T^*(X) = -\sum_{i=1}^{n} X_i$ is still a sufficient statistic (it is a one-to-one function of $T$). Therefore, we can apply the Karlin-Rubin Theorem using $T^* = -T$ instead. Specifically, the UMP level $\alpha$ test is
\[ \phi(t^*) = I(t^* < t_0), \]
where $t_0$ satisfies
\[ \alpha = E_{\theta_0}[\phi(T^*)] = P_{\theta_0}(T^* < t_0) = P_{\theta_0}(T > -t_0). \]

Because $T \sim \text{gamma}(n, 1/\theta)$, we take $-t_0 = g_{n,1/\theta_0,\alpha}$, the (upper) $\alpha$ quantile of a gamma$(n, 1/\theta_0)$ distribution. Therefore, the UMP level $\alpha$ test is $I(t > g_{n,1/\theta_0,\alpha})$; i.e., the UMP level $\alpha$ rejection region is
\[ R = \left\{ x \in X : \sum_{i=1}^{n} x_i > g_{n,1/\theta_0,\alpha} \right\}. \]

Using $\chi^2$ critical values: We can also write this rejection region in terms of a $\chi^2$ quantile. To see why, note that when $\theta = \theta_0$, the quantity $2\theta_0 T \sim \chi^2_{2n}$ so that
\[ \alpha = P_{\theta_0}(T > -t_0) = P_{\theta_0}(2\theta_0 T > -2\theta_0 t_0) \]
\[ \implies -2\theta_0 t_0 \text{ set } \chi^2_{2n,\alpha}. \]
Therefore, the UMP level $\alpha$ rejection region can be written as

$$R = \left\{ x \in \mathcal{X} : 2\theta_0 \sum_{i=1}^{n} x_i > \chi^2_{2n,\alpha} \right\} = \left\{ x \in \mathcal{X} : \sum_{i=1}^{n} x_i > \frac{\chi^2_{2n,\alpha}}{2\theta_0} \right\}.$$  

**Remark:** One advantage of writing the rejection region in this way is that it depends on a $\chi^2$ quantile, which, historically, may have been available in probability tables (i.e., in times before computers and R). Another small advantage is that we can express the power function $\beta(\theta)$ in terms of a $\chi^2$ cdf instead of a more general gamma cdf.

**Power function:** The power function of the UMP level $\alpha$ test is given by

$$\beta(\theta) = P_\theta(\mathbf{X} \in R) = P_\theta \left( T > \frac{\chi^2_{2n,\alpha}}{2\theta_0} \right) = P_\theta \left( 2\theta T > \frac{\theta \chi^2_{2n,\alpha}}{\theta_0} \right) = 1 - F_{\chi^2_{2n}} \left( \frac{\theta \chi^2_{2n,\alpha}}{\theta_0} \right),$$

where $F_{\chi^2_{2n}}(\cdot)$ is the $\chi^2_{2n}$ cdf. A graph of this power function, when $n = 10$, $\alpha = 0.10$, and $\theta_0 = 4$, is shown in Figure 8.5 (above).
Proof of Karlin-Rubin Theorem. We will prove this theorem in parts. The first part is a lemma.

**Lemma 1:** If \( g(x) \uparrow_{nd} x \) and \( h(x) \uparrow_{nd} x \), then

\[
\text{cov}[g(X), h(X)] \geq 0.
\]

**Proof.** Take \( X_1, X_2 \) to be iid with the same distribution as \( X \). Then

\[
E\{[h(X_1) - h(X_2)][g(X_1) - g(X_2)]\}
= E[h(X_1)g(X_1)] - E[h(X_2)g(X_1)] - E[h(X_1)g(X_2)] + E[h(X_2)g(X_2)]
= \text{cov}[g(X), h(X)]
\]

which equals \( 2\text{cov}[g(X), h(X)] \). Therefore,

\[
\text{cov}[g(X), h(X)] = \frac{1}{2} E\{[h(X_1) - h(X_2)][g(X_1) - g(X_2)]\}.
\]

However, note that

\[
[h(x_1) - h(x_2)][g(x_1) - g(x_2)] = \begin{cases} 
(\geq 0)(\geq 0), & x_1 > x_2 \\
0, & x_1 = x_2 \\
(\leq 0)(\leq 0), & x_1 < x_2
\end{cases}
\]

showing that \( [h(x_1) - h(x_2)][g(x_1) - g(x_2)] \geq 0 \), for all \( x_1, x_2 \in \mathbb{R} \). By Theorem 2.2.5 (CB, pp 57), \( E\{[h(X_1) - h(X_2)][g(X_1) - g(X_2)]\} \geq 0 \). \( \square \)

**Remark:** Our frame of reference going forward is testing \( H_0 : \theta \leq \theta_0 \) versus \( H_1 : \theta > \theta_0 \). Proving the other case stated in the Karlin-Rubin Theorem is analogous.

**Lemma 2.** Suppose the family \( \{g_T(t|\theta) : \theta \in \Theta\} \) has MLR. If \( \psi(t) \uparrow_{nd} t \), then \( E_\theta[\psi(T)] \uparrow_{nd} \theta \).

**Proof.** Suppose that \( \theta_2 > \theta_1 \). Because \( \{g_T(t|\theta) : \theta \in \Theta\} \) has MLR, we know that

\[
\frac{g_T(t|\theta_2)}{g_T(t|\theta_1)} \uparrow_{nd} t
\]

over the set \( \{t : g_T(t|\theta_1) > 0 \text{ or } g_T(t|\theta_2) > 0\} \). Therefore, by Lemma 1, we know

\[
\text{cov}_{\theta_1}[\psi(T), \frac{g_T(T|\theta_2)}{g_T(T|\theta_1)}] \geq 0 \implies E_{\theta_1}[\psi(T)] \frac{g_T(T|\theta_2)}{g_T(T|\theta_1)} \geq E_{\theta_1}[\psi(T)] E_{\theta_1}[\frac{g_T(T|\theta_2)}{g_T(T|\theta_1)}] = 1 \implies E_{\theta_2}[\psi(T)] \geq E_{\theta_1}[\psi(T)].
\]

Because \( \theta_1 \) and \( \theta_2 \) are arbitrary, the result follows. \( \square \)

**Lemma 3.** Under the same assumptions stated in Lemma 2,

\[
P_\theta(T > t_0) \uparrow_{nd} \theta
\]

for all \( t_0 \in \mathbb{R} \). In other words, the family \( \{g_T(t|\theta) : \theta \in \Theta\} \) is stochastically increasing in \( \theta \).
Proof. This is a special case of Lemma 2. Fix $t_0$. Take $\psi(t) = I(t > t_0)$. Clearly,
\[
\psi(t) = \begin{cases} 
1, & t > t_0 \\
0, & t \leq t_0 
\end{cases}
\]
is a nondecreasing function of $t$ (with $t_0$ fixed). From Lemma 2, we know that
\[
E_\theta[\psi(T)] = E_\theta[I(T > t_0)] = P_\theta(T > t_0) \uparrow_{\text{nd}} \theta.
\]
Because $t_0 \in \mathbb{R}$ was chosen arbitrarily, this result is true for all $t_0 \in \mathbb{R}$. 

**Implication:** In the statement of the Karlin-Rubin Theorem (for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$), we have now shown that the power function
\[
\beta(\theta) = P_\theta(T > t_0)
\]
is a nondecreasing function of $\theta$. This explains why $\alpha$ satisfies
\[
\alpha = P_{\theta_0}(T > t_0).
\]
Why? Because $P_\theta(T > t_0)$ is a nondecreasing function of $\theta$,
\[
\alpha = \sup_{H_0} \beta(\theta) = \sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = P_{\theta_0}(T > t_0).
\]
This shows that $\phi(t) = I(t > t_0)$ is a size $\alpha$ (and hence level $\alpha$) test function. Thus, all that remains is to show that this test is uniformly most powerful (i.e., most powerful $\forall \theta > \theta_0$). Remember that we are considering the test
\[
H_0 : \theta \leq \theta_0 \\
\text{versus} \\
H_1 : \theta > \theta_0.
\]

Let $\phi^*(x)$ be any other level $\alpha$ test of $H_0$ versus $H_1$. Fix $\theta_1 > \theta_0$ and consider the test of
\[
H^*_0 : \theta = \theta_0 \\
\text{versus} \\
H^*_1 : \theta = \theta_1
\]
instead. Note that
\[
E_{\theta_0}[\phi^*(X)] \leq \sup_{\theta \leq \theta_0} E_\theta[\phi^*(X)] \leq \alpha
\]
because $\phi^*(x)$ is a level $\alpha$ test of $H_0$ versus $H_1$. This also means that $\phi^*(x)$ is a level $\alpha$ test of $H^*_0$ versus $H^*_1$. However, Corollary 8.3.13 (Neyman Pearson with a sufficient statistic $T$) says that $\phi(t)$ is the most powerful (MP) level $\alpha$ test of $H^*_0$ versus $H^*_1$. This means that
\[
E_{\theta_1}[\phi^*(X)] \geq E_{\theta_1}[\phi^*(X)].
\]
Because $\theta_1 > \theta_0$ was chosen arbitrarily and because $\phi^*(x)$ was too, we have
\[
E_\theta[\phi^*(T)] \geq E_\theta[\phi^*(X)]
\]
for all $\theta > \theta_0$ and for any level $\alpha$ test $\phi^*(x)$ of $H_0$ versus $H_1$. Because $\phi(t)$ is a level $\alpha$ test of $H_0$ versus $H_1$ (shown above), we are done. 

\[\square\]
Note: In single parameter exponential families, we can find UMP tests for \( H_0 : \theta \leq \theta_0 \) versus \( H_1 : \theta > \theta_0 \) (or for \( H_0 : \theta \geq \theta_0 \) versus \( H_1 : \theta < \theta_0 \)). Unfortunately,

- once we get outside this setting (even with a one-sided \( H_1 \)), UMP tests do become scarce.
- with a two-sided \( H_1 \), that is \( H_1 : \theta \neq \theta_0 \), UMP tests do not exist.

In other words, the collection of problems for which a UMP test exists is somewhat small. In many ways, this should not be surprising. Requiring a test to outperform all other level \( \alpha \) tests for all \( \theta \) in the alternative space \( \Theta^c_0 \) is asking a lot. The “larger” \( \Theta^c_0 \) is, the harder it is to find a UMP test.

**Example 8.18.** Suppose \( X_1, X_2, \ldots, X_n \) are iid \( N(\mu, \sigma_0^2) \), where \(-\infty < \mu < \infty\) and \( \sigma_0^2 \) is known. Consider testing

\[
H_0 : \mu = \mu_0 \\
n\text{versus} \\
H_1 : \mu \neq \mu_0.
\]

There is no UMP test for this problem. A UMP test would exist if we could find a test whose power function “beats” the power function for all other level \( \alpha \) tests. For one-sided alternatives, it is possible to find one. However, a two-sided alternative space is too large. To illustrate, suppose we considered testing

\[
H'_0 : \mu \leq \mu_0 \\
n\text{versus} \\
H'_1 : \mu > \mu_0.
\]

The UMP level \( \alpha \) test for \( H'_0 \) versus \( H'_1 \) uses

\[
\phi'(x) = I \left( \bar{x} > \frac{z_\alpha \sigma_0}{\sqrt{n}} + \mu_0 \right)
\]

and has power function

\[
\beta'(\mu) = 1 - F_Z \left( z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} \right),
\]

where \( F_Z(\cdot) \) is the standard normal cdf. This is a size (and level) \( \alpha \) test for \( H'_0 \) versus \( H'_1 \). It is also a size (and level) test for \( H_0 \) versus \( H_1 \) because

\[
\sup_{\mu \in \Theta_0} \beta'(\mu) = \sup_{\mu = \mu_0} \beta'(\mu) = \beta'(\mu_0) = 1 - F_Z(z_\alpha) = \alpha.
\]

Now consider testing

\[
H''_0 : \mu \geq \mu_0 \\
n\text{versus} \\
H''_1 : \mu < \mu_0.
\]
The UMP level $\alpha$ test for $H'_0$ versus $H'_1$ uses
\[
\phi''(x) = I\left( x < -\frac{z_\alpha \sigma_0}{\sqrt{n}} + \mu_0 \right)
\]
and has power function
\[
\beta''(\mu) = F_Z\left( -z_\alpha + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} \right).
\]
This is a size (and level) $\alpha$ test for $H_0$ versus $H_1$ because
\[
\sup_{\mu \in \Theta_0} \beta''(\mu) = \sup_{\mu = \mu_0} \beta''(\mu) = \beta''(\mu_0) = F_Z(-z_\alpha) = \alpha.
\]
Therefore, we have concluded that
- $\phi'(x)$ is UMP level $\alpha$ when $\mu > \mu_0$
- $\phi''(x)$ is UMP level $\alpha$ when $\mu < \mu_0$.

However, $\phi'(x) \neq \phi''(x)$ for all $x \in \mathcal{X}$. Therefore, no UMP test can exist for $H_0$ versus $H_1$.

Q: How do we find an “optimal” test in situations like this (e.g., a two-sided $H_1$)?

A: We change what we mean by “optimal.”

**Definition:** Consider the test of
\[
H_0 : \theta \in \Theta_0 \\
versus \\
H_1 : \theta \in \Theta_0^c.
\]
A test with power function $\beta(\theta)$ is **unbiased** if $\beta(\theta') \geq \beta(\theta'')$ for all $\theta' \in \Theta_0^c$ and for all $\theta'' \in \Theta_0$. That is, the power is always larger in the alternative parameter space than it is in the null parameter space.

- Therefore, when no UMP test exists, we could further restrict attention to those tests that are level $\alpha$ and are unbiased. Conceptually, define
  \[
  \mathcal{C}^U = \{ \text{all level } \alpha \text{ tests that are unbiased} \}.
  \]
- The test in $\mathcal{C}^U$ that is UMP is called the **uniformly most powerful unbiased** (UMPU) test. The UMPU test has power function $\beta(\theta)$ that satisfies
  \[
  \beta(\theta) \geq \beta^*(\theta) \quad \text{for all } \theta \in \Theta_0^c,
  \]
  where $\beta^*(\theta)$ is the power function of any other (unbiased) test in $\mathcal{C}^U$.

**Example 8.18** (continued). Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\mu, \sigma_0^2)$, where $-\infty < \mu < \infty$ and $\sigma_0^2$ is known. Consider testing
\[
H_0 : \mu = \mu_0 \\
versus \\
H_1 : \mu \neq \mu_0.
\]
Figure 8.6: Pdf of $Z \sim \mathcal{N}(0, 1)$. The UMPU level $\alpha$ rejection region in Example 8.18 is shown shaded.

The UMPU level $\alpha$ test uses

$$
\phi(x) = \begin{cases} 
1, & \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} < -z_{\alpha/2} \quad \text{or} \quad \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} > z_{\alpha/2} \\
0, & \text{otherwise}
\end{cases}
$$

In other words, the UMPU level $\alpha$ rejection region is

$$
R = \{x \in \mathcal{X} : \phi(x) = 1\} = \{x \in \mathcal{X} : |z| > z_{\alpha/2}\},
$$

where

$$
z = \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}
$$

Note that, because $Z \overset{H_0}{\sim} \mathcal{N}(0, 1)$,

$$
P_{\mu_0}(X \in R) = P_{\mu_0}(|Z| > z_{\alpha/2}) = 1 - P_{\mu_0}(-z_{\alpha/2} < Z < z_{\alpha/2})
= 1 - [F_Z(z_{\alpha/2}) - F_Z(-z_{\alpha/2})]
= 1 - (1 - \alpha/2) + \alpha/2 = \alpha,
$$

which shows that $R$ is a size (and hence level) $\alpha$ rejection region.
Figure 8.7: Power function $\beta(\mu)$ of the UMPU level $\alpha = 0.05$ test in Example 8.18 with $n = 10$, $\mu_0 = 6$, and $\sigma_0^2 = 4$. Also shown are the power functions corresponding to the two UMP level $\alpha = 0.05$ tests with $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$.

The power function of the UMPU test $\phi(x)$ is

$$
\beta(\mu) = P_\mu(X \in R) = P_\mu(|Z| \geq z) = 1 - F_Z \left( z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} \right) + F_Z \left( -z_{\alpha/2} + \frac{\mu_0 - \mu}{\sigma_0/\sqrt{n}} \right).
$$

**Special case:** I took $n = 10$, $\alpha = 0.05$, $\mu_0 = 6$, and $\sigma_0^2 = 4$. The UMPU level $\alpha = 0.05$ power function $\beta(\mu)$ is shown in Figure 8.7 (above). For reference, I have also plotted in Figure 8.7 the UMP level $\alpha = 0.05$ power functions for the two one-sided tests (i.e., the tests with $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$, respectively).

- It is easy to see that the UMPU test is an unbiased test. Note that $\beta(\mu)$ is always larger in the alternative parameter space $\{\mu \in \mathbb{R} : \mu \neq \mu_0\}$ than it is when $\mu = \mu_0$.

- The UMPU test’s power function “loses” to each UMP test’s power function in the region where that UMP test is most powerful. This is the price one must pay for restricting attention to unbiased tests. The best unbiased test for a two-sided $H_1$ will not beat a one-sided UMP test. However, the UMPU test is clearly better than the UMP tests in each UMP test’s null parameter space.
8.3.4 Probability values

Definition: A p-value \( p(X) \) is a test statistic, satisfying \( 0 \leq p(x) \leq 1 \), for all \( x \in X \). Small values of \( p(x) \) are evidence against \( H_0 \). A p-value is said to be valid if

\[
P_\theta(p(X) \leq \alpha) \leq \alpha,
\]

for all \( \theta \in \Theta_0 \) and for all \( 0 \leq \alpha \leq 1 \).

Remark: Quoting your authors (CB, pp 397),

“If \( p(X) \) is a valid p-value, it is easy to construct a level \( \alpha \) test based on \( p(X) \).
The test that rejects \( H_0 \) if and only if \( p(X) \leq \alpha \) is a level \( \alpha \) test.”

It is easy to see why this is true. The validity requirement above guarantees that

\[
\phi(x) = I(p(x) \leq \alpha)
\]

is a level \( \alpha \) test function. Why? Note that

\[
\sup_{\theta \in \Theta_0} E_\theta[\phi(X)] = \sup_{\theta \in \Theta_0} P_\theta(p(X) \leq \alpha) \leq \alpha.
\]

Therefore, rejecting \( H_0 \) when \( p(x) \leq \alpha \) is a level \( \alpha \) decision rule.

Theorem 8.3.27. Let \( W = W(X) \) be a test statistic such that large values of \( W \) give evidence against \( H_0 \). For each \( x \in X \), define

\[
p(x) = \sup_{\theta \in \Theta_0} P_\theta(W(X) \geq w),
\]

where \( w = W(x) \). Then \( p(X) \) is a valid p-value. Note that the definition of \( p(x) \) for when small values of \( W \) give evidence against \( H_0 \) would be analogous.

Proof. Fix \( \theta \in \Theta_0 \). Let \( F_{-W}(w|\theta) \) denote the cdf of \( -W = -W(X) \). When the test rejects for large values of \( W \),

\[
p_\theta(x) \equiv P_\theta(W(X) \geq w) = P_\theta(-W(X) \leq -w) = F_{-W}(-w|\theta),
\]

where \( w = W(x) \). If \( -W(X) \) is a continuous random variable, then

\[
p_\theta(X) \overset{d}{=} F_{-W}(-W|\theta) \overset{d}{=} U(0,1),
\]

by the Probability Integral Transformation (Chapter 2). If \( -W(X) \) is discrete, then

\[
p_\theta(X) \overset{d}{=} F_{-W}(-W|\theta) \overset{\geq_{ST}}{=} U(0,1),
\]

where the notation \( X \overset{\geq_{ST}}{=} Y \) means “the distribution of \( X \) is stochastically larger than the distribution of \( Y \)” (see Exercise 2.10, CB, pp 77). Combining both cases, we have

\[
P_\theta(p_\theta(X) \leq \alpha) \leq \alpha,
\]
for all $0 \leq \alpha \leq 1$. Now, note that

$$p(x) \equiv \sup_{\theta \in \Theta_0} P_{\theta}(W(X) \geq w) \geq P_{\theta}(W(X) \geq w) \equiv p_\theta(x),$$

for all $x \in X$. Therefore,

$$P_\theta(p(X) \leq \alpha) \leq P_\theta(p_\theta(X) \leq \alpha) \leq \alpha.$$

Because we fixed $\theta \in \Theta_0$ arbitrarily, this result must hold for all $\theta \in \Theta_0$. We have shown that $p(X)$ is a valid p-value. □

**Example 8.19.** Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$; i.e., both parameters are unknown. Set $\theta = (\mu, \sigma^2)$. Consider testing

$$H_0 : \mu \leq \mu_0$$

versus

$$H_1 : \mu > \mu_0.$$

We have previously shown (see pp 242-243, notes) that large values of

$$W = W(X) = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

are evidence against $H_0$ (i.e., this is a “one-sample t test,” which is a LRT). The null parameter space is

$$\Theta_0 = \{\theta = (\mu, \sigma^2) : \mu \leq \mu_0, \sigma^2 > 0\}.$$

Therefore, with observed value $w = W(x)$, the p-value for the test is

$$p(x) = \sup_{\theta \in \Theta_0} P_{\theta}(W(X) \geq w) = \sup_{\theta \in \Theta_0} P_\theta\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq w\right) = \sup_{\theta \in \Theta_0} P_\theta\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} \geq w + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) = \sup_{\mu \leq \mu_0} P_{\mu_0}\left(T_{n-1} \geq w + \frac{\mu_0 - \mu}{S/\sqrt{n}}\right) = P(T_{n-1} \geq w),$$

where $T_{n-1}$ is a $t$ random variable with $n - 1$ degrees of freedom. The penultimate equality holds because the distribution of $(\bar{X} - \mu)/(S/\sqrt{n})$ does not depend on $\sigma^2$. The last equality holds because $(\mu_0 - \mu)/(S/\sqrt{n})$ is a nonnegative random variable.

**Remark:** In Example 8.19, calculating the supremum over $\Theta_0$ is relatively easy. In other problems, it might not be, especially when there are nuisance parameters. A very good discussion on this is given in Berger and Boos (1994). These authors propose another type of p-value by “suping” over subsets of $\Theta_0$ formed from calculating confidence intervals first (which can make the computation easier).

**Important:** If $H_0$ is simple, say $H_0 : \theta = \theta_0$, and if a p-value $p(x)$ satisfies

$$P_{\theta_0}(p(X) \leq \alpha) = \alpha,$$

for all $0 \leq \alpha \leq 1$, then $\phi(x) = I(p(x) \leq \alpha)$ is a size $\alpha$ test and $p(X) \overset{H_0}{\sim} \mathcal{U}(0, 1)$. 

---

PAGE 269
Example 8.20. Suppose \( X_1, X_2, \ldots, X_n \) are iid \( \mathcal{N}(0, 1) \). I used R to simulate \( B = 200 \) independent samples of this type, each with \( n = 30 \). With each random sample, I performed a \( t \) test for

\[
H_0 : \mu = 0 \\
\text{versus} \\
H_1 : \mu \neq 0
\]

and calculated the p-value for each test (note that \( H_0 \) is true). A (uniform) qq plot of the 200 p-values in Figure 8.8 shows agreement with the \( U(0, 1) \) distribution. Using \( \alpha = 0.05 \), there were 9 tests (out of 200) that incorrectly rejected \( H_0 \).

Remark: Your authors discuss calculating p-values using conditional distributions given a sufficient statistic on pp 399. This method of p-value calculation (through the use of conditioning) can be very helpful with small samples and discrete distributions. The most common application is Fisher’s Exact Test, a small-sample test for two independent binomial proportions; see Example 8.3.30 (CB, pp 399).
9 Interval Estimation

Complementary reading: Chapter 9 (CB).

9.1 Introduction

Setting: We observe $X = (X_1, X_2, ..., X_n) \sim f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$. Usually, $X_1, X_2, ..., X_n$ will constitute a random sample (iid sample) from a population $f_X(x|\theta)$. We regard $\theta$ as fixed and unknown.

Definition: An interval estimate of a real-valued parameter $\theta$ is any pair of functions

$$L(x) = L(x_1, x_2, ..., x_n) \quad U(x) = U(x_1, x_2, ..., x_n),$$

satisfying $L(x) \leq U(x)$ for all $x \in X$. When $X = x$ is observed, the inference

$$L(x) \leq \theta \leq U(x)$$

is made. The random version $[L(X), U(X)]$ is called an interval estimator.

Remark: In the definition above, a one-sided interval estimate is formed when one of the endpoints is $\pm \infty$. For example, if $L(x) = -\infty$, then the estimate is $(-\infty, U(x)]$. If $U(x) = \infty$, the estimate is $[L(x), \infty)$.

Definition: Suppose $[L(X), U(X)]$ is an interval estimator for $\theta$. The coverage probability of the interval is

$$P_{\theta}(L(X) \leq \theta \leq U(X)).$$

It is important to note the following:

- In the probability above, it is the endpoints $L(X)$ and $U(X)$ that are random; not $\theta$ (it is fixed).

- The coverage probability is regarded as a function of $\theta$. That is, the probability that $[L(X), U(X)]$ contains $\theta$ may be different for different values of $\theta \in \Theta$. This is usually true when $X$ is discrete.

Definition: The confidence coefficient of the interval estimator $[L(X), U(X)]$ is

$$\inf_{\theta \in \Theta} P_{\theta}(L(X) \leq \theta \leq U(X)).$$

An interval estimator with confidence coefficient equal to $1 - \alpha$ is called a $1 - \alpha$ confidence interval.

Remark: In some problems, it is possible that the estimator itself is not an interval. More generally, we use the term $1 - \alpha$ confidence set to allow for these types of estimators. The notation $C(X)$ is used more generally to denote a confidence set.
Example 9.1. Suppose that $X_1, X_2, \ldots, X_n$ are iid $U(0, \theta)$, where $\theta > 0$. We consider two interval estimators:

1. $(a X(n), b X(n))$, where $1 \leq a < b$
2. $(X(n) + c, X(n) + d)$, where $0 \leq c < d$.

The pdf of $X(n)$ is

$$f_{X(n)}(x) = n f_X(x)[F_X(x)]^{n-1} = n \left(\frac{1}{\theta}\right) \left(\frac{x}{\theta}\right)^{n-1} I(0 < x < \theta) = \frac{n x^{n-1}}{\theta^n} I(0 < x < \theta).$$

By transformation, the pdf of $T = \frac{X(n)}{\theta}$ is

$$f_T(t) = nt^{n-1} I(0 < t < 1);$$

i.e., $T \sim \text{beta}(n, 1)$. The coverage probability for the first interval is

$$P_{\theta}(a X(n) \leq \theta \leq b X(n)) = P_{\theta}\left(\frac{1}{b X(n)} \leq \frac{1}{\theta} \leq \frac{1}{a X(n)}\right) = P_{\theta}\left(\frac{1}{b} \leq \frac{X(n)}{\theta} \leq \frac{1}{a}\right) = \int_{1/b}^{1/a} nt^{n-1} dt = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n,$$

that is, the coverage probability is the same for all $\theta \in \Theta = \{\theta : \theta > 0\}$. The confidence coefficient of the interval $(a X(n), b X(n))$ is therefore

$$\inf_{\theta > 0} \left[\left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n\right] = \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n.$$}

On the other hand, the coverage probability for the second interval is

$$P_{\theta}(X(n) + c \leq \theta \leq X(n) + d) = P_{\theta}(c \leq \theta - X(n) \leq d) = P_{\theta}\left(\frac{c}{\theta} \leq 1 - \frac{X(n)}{\theta} \leq \frac{d}{\theta}\right) = P_{\theta}\left(1 - d \frac{\theta}{\theta} \leq \frac{X(n)}{\theta} \leq 1 - \frac{c}{\theta}\right) = \int_{1-d/\theta}^{1-c/\theta} nt^{n-1} dt = \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n,$$
which does depend on $\theta$. Interestingly, the confidence coefficient of $(X_{(n)} + c, X_{(n)} + d)$ is

$$\inf_{\theta > 0} \left[ \left(1 - \frac{c}{\theta}\right)^n - \left(1 - \frac{d}{\theta}\right)^n \right] = 0.$$

**Example 9.2.** Suppose that $X_1, X_2, \ldots, X_n$ are iid Bernoulli($p$), where $0 < p < 1$. A $1 - \alpha$ confidence interval commonly taught in undergraduate courses is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p}$ is the sample proportion, that is,

$$\hat{p} = \frac{Y}{n} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

where $Y = \sum_{i=1}^{n} X_i \sim b(n, p)$, and $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0, 1)$ distribution. In Chapter 10, we will learn that this is a large-sample “Wald-type” confidence interval. An expression for the coverage probability of this interval is

$$P_p \left( \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = E_p \left[ I \left( \frac{Y}{n} - z_{\alpha/2} \sqrt{\frac{Y(1-Y)}{n}} \leq p \leq \frac{Y}{n} + z_{\alpha/2} \sqrt{\frac{Y(1-Y)}{n}} \right) \right] = \sum_{y=0}^{n} I \left( \frac{y}{n} - z_{\alpha/2} \sqrt{\frac{Y(1-Y)}{n}} \leq p \leq \frac{y}{n} + z_{\alpha/2} \sqrt{\frac{Y(1-Y)}{n}} \right) \left( Y \right)_{b(n, p)}^{p^y(1-p)^{n-y}}.$$ 

**Special case:** I used R to graph this coverage probability function across values of $0 < p < 1$ when $n = 40$ and $\alpha = 0.05$; see Figure 9.1 (next page).

- The coverage probability rarely attains the nominal 0.95 level across $0 < p < 1$.
- The jagged nature of the coverage probability function (of $p$) arises from the discreteness of $Y \sim b(40, p)$.
- The confidence coefficient of the Wald interval (i.e., the infimum coverage probability across all $0 < p < 1$) is clearly 0.
- An excellent account of the performance of this confidence interval (and competing intervals) is given in Brown et al. (2001, *Statistical Science*).
- When $1 - \alpha = 0.95$, one competing interval mentioned in Brown et al. (2001) replaces $y$ with $y^* = y + 2$ and $n$ with $n^* = n + 4$. This “add two successes-add two failures” interval was proposed by Agresti and Coull (1998, *American Statistician*). Because this interval’s coverage probability is much closer to the nominal level across $0 < p < 1$ (and because it is so easy to compute), it has begun to usurp the Wald confidence interval in introductory level courses.
Figure 9.1: Coverage probability of the Wald confidence interval for a binomial proportion $p$ when $n = 40$ and $\alpha = 0.05$. A dotted horizontal line at $1 - \alpha = 0.95$ has been added.

### 9.2 Methods of Finding Interval Estimators

**Preview:** The authors present four methods to find interval estimators:

1. Test inversion (i.e., inverting a test statistic)
2. Using pivotal quantities
3. “Guaranteeing an interval” by pivoting a cdf
4. Bayesian credible intervals.

**Note:** Large-sample interval estimators will be discussed in Chapter 10.

#### 9.2.1 Inverting a test statistic

**Remark:** This method of interval construction is motivated by the strong duality between hypothesis testing and confidence intervals.
Motivation: Consider testing $H_0 : \theta = \theta_0$ using the \( (\text{non-randomized}) \) test function

\[ \phi(x) = I(x \in R_{\theta_0}) = \begin{cases} 1, & x \in R_{\theta_0} \\ 0, & x \in R'_{\theta_0}, \end{cases} \]

where

\[ P_{\theta_0}(X \in R_{\theta_0}) = E_{\theta_0}[\phi(X)] = \alpha; \]

i.e., \( \phi(x) \) is a size \( \alpha \) test. Note that we have used the notation \( R_{\theta_0} \) to emphasize that the rejection region \( R \) depends on the value of \( \theta_0 \). Let \( A_{\theta_0} = R'_{\theta_0} \) denote the “acceptance region” for the test, that is, \( A_{\theta_0} \) is the set of all \( x \in \mathcal{X} \) that do not lead to \( H_0 \) being rejected. For each \( x \in \mathcal{X} \), define

\[ C(x) = \{ \theta_0 : x \in A_{\theta_0} \}. \]

From this definition, clearly \( \theta_0 \in C(x) \iff x \in A_{\theta_0} \). Therefore,

\[ P_{\theta_0}(\theta_0 \in C(X)) = P_{\theta_0}(X \in A_{\theta_0}) = 1 - P_{\theta_0}(X \in R_{\theta_0}) = 1 - \alpha. \]

However, this same argument holds for all \( \theta_0 \in \Theta \); i.e., it holds regardless of the value of \( \theta \) under \( H_0 \). Therefore,

\[ C(X) = \{ \theta \in \Theta : X \in A_{\theta} \} \]

is a \( 1 - \alpha \) confidence set.

Example 9.3. Suppose \( X_1, X_2, \ldots, X_n \) are iid \( N(\mu, \sigma^2) \), where \(-\infty < \mu < \infty \) and \( \sigma^2 > 0 \); i.e., both parameters are unknown. A size \( \alpha \) likelihood ratio test (LRT) of \( H_0 : \mu = \mu_0 \) versus \( H_1 : \mu \neq \mu_0 \) uses the test function

\[ \phi(x) = \begin{cases} 1, & \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \geq t_{n-1, \alpha/2} \\ 0, & \text{otherwise.} \end{cases} \]

The “acceptance region” for this test is

\[ A_{\mu_0} = \left\{ x \in \mathcal{X} : \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} < t_{n-1, \alpha/2} \right\}, \]

where, note that

\[ P_{\mu_0}(X \in A_{\mu_0}) = P_{\mu_0}\left(\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} < t_{n-1, \alpha/2}\right) = P_{\mu_0}\left(-t_{n-1, \alpha/2} < \frac{\bar{x} - \mu_0}{s/\sqrt{n}} < t_{n-1, \alpha/2}\right) = 1 - \alpha. \]

Therefore, a \( 1 - \alpha \) confidence set for \( \mu \) is

\[ C(x) = \{ \mu \in \mathbb{R} : x \in A_{\mu} \} = \begin{cases} \mu : -t_{n-1, \alpha/2} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{n-1, \alpha/2} \\ \mu : -t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} < \bar{x} - \mu < t_{n-1, \alpha/2} \frac{s}{\sqrt{n}} \end{cases}. \]
The random version of this confidence set (interval) is written as
\[
\left( \bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right).
\]

**Remark:** As Example 9.3 suggests, when we invert a two-sided hypothesis test, we get a two-sided confidence interval. This will be true in most problems. Analogously, inverting one-sided tests generally leads to one-sided intervals.

**Example 9.4.** Suppose \(X_1, X_2, ..., X_n\) are iid exponential(\(\theta\)), where \(\theta > 0\). A uniformly most powerful (UMP) level \(\alpha\) test of \(H_0 : \theta = \theta_0\) versus \(H_1 : \theta > \theta_0\) uses the test function
\[
\phi(t) = \begin{cases} 
1, & t \geq \frac{\theta_0}{2} \chi^2_{2n,\alpha} \\
0, & \text{otherwise,}
\end{cases}
\]
where the sufficient statistic \(t = \sum_{i=1}^n x_i\). The “acceptance region” for this test is
\[
A_{\theta_0} = \left\{ x \in \mathcal{X} : t < \frac{\theta_0}{2} \chi^2_{2n,\alpha} \right\},
\]
where, note that
\[
P_{\theta_0}(X \in A_{\theta_0}) = P_{\theta_0} \left( T < \frac{\theta_0}{2} \chi^2_{2n,\alpha} \right) = P_{\theta_0} \left( \frac{2T}{\theta_0} < \chi^2_{2n,\alpha} \right) = 1 - \alpha,
\]
because \(2T/\theta_0 \overset{H_0}{\sim} \text{gamma}(n, 2) \overset{d}{=} \frac{2}{\chi^2_{2n}}\). Therefore, a \(1 - \alpha\) confidence set for \(\theta\) is
\[
C(x) = \{ \theta > 0 : x \in A_\theta \} = \left\{ \theta : t < \frac{\theta}{2} \chi^2_{2n,\alpha} \right\} = \left\{ \theta : \frac{2t}{\chi^2_{2n,\alpha}} < \theta \right\}.
\]
The random version of this confidence set is written as
\[
\left( \frac{2T}{\chi^2_{2n,\alpha}}, \infty \right),
\]
where \(T = \sum_{i=1}^n X_i\). This is a “one-sided” interval, as expected, because we have inverted a one-sided test.

**Remark:** The test inversion method makes direct use of the relationship between hypothesis tests and confidence intervals (sets). On pp 421, the authors of CB write,

“Both procedures look for consistency between sample statistics and population parameters. The hypothesis test fixes the parameter and asks what sample values (the acceptance region) are consistent with that fixed value. The confidence set fixes the sample value and asks what parameter values (the confidence interval) make this sample value most plausible.”

An illustrative figure (Figure 9.2.1, pp 421) displays this relationship in the \(\mathcal{N}(\mu, \sigma^2_0)\) case; i.e., writing a confidence interval for a normal mean \(\mu\) when \(\sigma^2_0\) is known.
### 9.2.2 Pivotal quantities

**Definition:** A random variable \( Q = Q(X, \theta) \) is a **pivotal quantity** (or **pivot**) if the distribution of \( Q \) does not depend on \( \theta \). That is, \( Q \) has the same distribution for all \( \theta \in \Theta \).

**Remark:** Finding pivots makes getting confidence intervals easy. If \( Q = Q(X, \theta) \) is a pivot, then we can set

\[
1 - \alpha = P_{\theta}(a \leq Q(X, \theta) \leq b),
\]

where \( a \) and \( b \) are quantiles of the distribution of \( Q \) that satisfy the equation. Because \( Q \) is a pivot, the probability on the RHS will be the same for all \( \theta \in \Theta \). Therefore, a \( 1 - \alpha \) confidence interval can be determined from this equation.

**Example 9.5.** Suppose that \( X_1, X_2, ..., X_n \) are iid \( U(0, \theta) \), where \( \theta > 0 \). In Example 9.1, we showed that

\[
Q = Q(X, \theta) = \frac{X_{(n)}}{\theta} \sim \text{beta}(n, 1).
\]

Because the distribution of \( Q \) is free of \( \theta \), we know that \( Q \) is a pivot. Let \( b_{n,1,1-\alpha/2} \) and \( b_{n,1,\alpha/2} \) denote the lower and upper \( \alpha/2 \) quantiles of a beta\((n, 1)\) distribution, respectively. We can then write

\[
1 - \alpha = P_{\theta}\left(b_{n,1,1-\alpha/2} \leq \frac{X_{(n)}}{\theta} \leq b_{n,1,\alpha/2}\right) = P_{\theta}\left(\frac{1}{b_{n,1,1-\alpha/2}} \geq \frac{\theta}{X_{(n)}} \geq \frac{1}{b_{n,1,\alpha/2}}\right) = P_{\theta}\left(b_{n,1,\alpha/2} \leq \theta \leq \frac{X_{(n)}}{b_{n,1,1-\alpha/2}}\right).
\]

This shows that

\[
\left(\frac{X_{(n)}}{b_{n,1,\alpha/2}}, \frac{X_{(n)}}{b_{n,1,1-\alpha/2}}\right)
\]

is a \( 1 - \alpha \) confidence interval for \( \theta \).

**Example 9.6.** Consider the simple linear regression model

\[
Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,
\]

where \( \epsilon_i \sim \text{iid } \mathcal{N}(0, \sigma^2) \) and the \( x_i \)'s are fixed constants (measured without error). Consider writing a confidence interval for

\[
\theta = E(Y|x_0) = \beta_0 + \beta_1 x_0,
\]

where \( x_0 \) is a specified value of \( x \). In a linear models course, you have shown that

\[
\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim \mathcal{N}\left(\theta, \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \right),
\]

where \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are the least-squares estimators of \( \beta_0 \) and \( \beta_1 \), respectively.
If \( \sigma^2 \) is known (completely unrealistic), we can use
\[
Q(Y, \theta) = \frac{\hat{\theta} - \theta}{\sqrt{\sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]}} \sim N(0, 1)
\]
as a pivot to write a confidence interval for \( \theta \).

More realistically, \( \sigma^2 \) is unknown and
\[
Q(Y, \theta) = \frac{\hat{\theta} - \theta}{\sqrt{\text{MSE} \left[ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]}} \sim t_{n-2},
\]
where MSE is the mean-squared error from the regression, is used as a pivot.

In the latter case, we can write
\[
1 - \alpha = P_{\beta, \sigma^2} \left( -t_{n-2, \alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{\text{MSE} \left[ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]}} \leq t_{n-2, \alpha/2} \right),
\]
for all \( \beta = (\beta_0, \beta_1)' \) and \( \sigma^2 \). It follows that
\[
\hat{\theta} \pm t_{n-2, \alpha/2} \sqrt{\text{MSE} \left[ \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right]}
\]
is a \( 1 - \alpha \) confidence interval for \( \theta \).

**Remark:** As Examples 9.5 and 9.6 illustrate, interval estimates are easily obtained after writing
\[
1 - \alpha = P_\theta(a \leq Q(X, \theta) \leq b),
\]
for constants \( a \) and \( b \) (quantiles of \( Q \)). More generally, \( \{\theta \in \Theta : Q(X, \theta) \in \mathcal{A}\} \) is a set estimate for \( \theta \), where \( \mathcal{A} \) satisfies
\[
1 - \alpha = P_\theta(Q(X, \theta) \in \mathcal{A}).
\]

For example, in Example 9.5, we could have written
\[
1 - \alpha = P_\theta \left( b_{n,1,1-\alpha} \leq \frac{X_{(n)}}{\theta} \leq 1 \right) = P_\theta \left( X_{(n)} \leq \theta \leq \frac{X_{(n)}}{b_{n,1,1-\alpha}} \right)
\]
and concluded that
\[
\left( \frac{X_{(n)}}{b_{n,1,1-\alpha}}, \frac{X_{(n)}}{b_{n,1,1-\alpha}} \right)
\]
is a \( 1 - \alpha \) confidence interval for \( \theta \). How does this interval compare with
\[
\left( \frac{X_{(n)}}{b_{n,1,\alpha/2}}, \frac{X_{(n)}}{b_{n,1,1-\alpha/2}} \right) ?
\]
Which one is “better?” For that matter, how should we define what “better” means?
Remark: The more general statement

\[ 1 - \alpha = P_\theta(Q(\mathbf{X}, \mathbf{\theta}) \in \mathcal{A}) \]

is especially useful when \( \mathbf{\theta} \) is a vector and the goal is to find a confidence set (i.e., confidence region) for \( \mathbf{\theta} \). In such cases, \( \mathcal{A} \) will generally be a subset of \( \mathbb{R}^k \) where \( k = \text{dim}(\mathbf{\theta}) \).

Example 9.7. Suppose \( X_1, X_2, \ldots, X_n \) are iid \( \mathcal{N}(\mu, \sigma^2) \), where \(-\infty < \mu < \infty \) and \( \sigma^2 > 0 \); i.e., both parameters are unknown. Set \( \mathbf{\theta} = (\mu, \sigma^2) \). We know that

\[ Q_1 = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \]

that is, \( Q_1 \) is a pivot. Therefore,

\[ 1 - \alpha = P_\theta\left( -t_{n-1,\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{n-1,\alpha/2} \right) \]
\[ = P_\theta\left( \bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right), \]

showing that

\[ C_1(\mathbf{X}) = \left( \bar{X} - t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2} \frac{S}{\sqrt{n}} \right) \]

is a \( 1 - \alpha \) confidence set for \( \mu \). Similarly, we know that

\[ Q_2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \]

that is, \( Q_2 \) is also a pivot. Therefore,

\[ 1 - \alpha = P_\theta\left( \frac{2}{\chi^2_{n-1,1-\alpha/2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{2}{\chi^2_{n-1,1-\alpha/2}} \right) \]
\[ = P_\theta\left( \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \right), \]

showing that

\[ C_2(\mathbf{X}) = \left( \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} \right) \]

is a \( 1 - \alpha \) confidence set for \( \sigma^2 \).

Extension: Suppose we wanted to write a confidence set (region) for \( \mathbf{\theta} = (\mu, \sigma^2) \) in \( \mathbb{R}^2 \). From the individual pivots, we know that \( C_1(\mathbf{X}) \) and \( C_2(\mathbf{X}) \) are each \( 1 - \alpha \) confidence sets.

Q: Is \( C_1(\mathbf{X}) \times C_2(\mathbf{X}) \), the Cartesian product of \( C_1(\mathbf{X}) \) and \( C_2(\mathbf{X}) \), a \( 1 - \alpha \) confidence region for \( \mathbf{\theta} \)?

A: No. By Bonferroni’s Inequality,

\[ P_\theta(\mathbf{\theta} \in C_1(\mathbf{X}) \times C_2(\mathbf{X})) \geq P_\theta(\mu \in C_1(\mathbf{X})) + P_\theta(\sigma^2 \in C_2(\mathbf{X})) - 1 \]
\[ = (1 - \alpha) + (1 - \alpha) - 1 \]
\[ = 1 - 2\alpha. \]

Therefore, \( C_1(\mathbf{X}) \times C_2(\mathbf{X}) \) is a \( 1 - 2\alpha \) confidence region for \( \mathbf{\theta} \).
Bonferroni adjustment: Adjust $C_1(X)$ and $C_2(X)$ individually so that the confidence coefficient of each is $1 - \alpha/2$. The adjusted set $C_1^*(X) \times C_2^*(X)$ is a $1 - \alpha$ confidence region for $\theta$. This region has coverage probability larger than or equal to $1 - \alpha$ for all $\theta$ (so it is “conservative”).

More interesting approach: Consider the quantity

$$Q = Q(X, \theta) = Q(X, \mu, \sigma^2) = \left(\frac{X - \mu}{\sigma/\sqrt{n}}\right)^2 + \frac{(n-1)S^2}{\sigma^2}.$$  

It is easy to show that $Q \sim \chi^2_n$, establishing that $Q$ is a pivot. Therefore, we can write

$$1 - \alpha = P_\theta(Q \leq \chi^2_{n,\alpha}) = P_\theta\left(\left(\frac{X - \mu}{\sigma/\sqrt{n}}\right)^2 + \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n,\alpha}\right),$$  

which shows that

$$C(X) = \{\theta = (\mu, \sigma^2) : Q(X, \mu, \sigma^2) \leq \chi^2_{n,\alpha}\}$$

is a $1 - \alpha$ confidence region (in $R^2$) for $\theta$. To see that this set looks like, note that the boundary is

$$Q(x, \mu, \sigma^2) = \chi^2_{n,\alpha} \iff \left(\frac{X - \mu}{\sigma/\sqrt{n}}\right)^2 + \frac{(n-1)S^2}{\sigma^2} = \chi^2_{n,\alpha} \iff (\mu - \bar{X})^2 = \frac{\chi^2_{n,\alpha}}{n} \left[\sigma^2 - \frac{(n-1)S^2}{\chi^2_{n,\alpha}}\right],$$  

which is a parabola in $\Theta = \{\theta = (\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$. The parabola has vertex at

$$\left(\bar{X}, \frac{(n-1)s^2}{\chi^2_{n,\alpha}}\right)$$

and it opens upward (because $\chi^2_{n,\alpha}/n > 0$). The confidence set is the interior of the parabola.

Discussion: Example 9.2.7 (CB, pp 427-428) provides tips on how to find pivots in location and scale (and location-scale) families.

<table>
<thead>
<tr>
<th>Family</th>
<th>Parameter</th>
<th>Pivot examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>$\mu$</td>
<td>$X - \mu, \bar{X} - \mu, X_{(n)} - \mu, X_{(1)} - \mu$</td>
</tr>
<tr>
<td>Scale</td>
<td>$\sigma$</td>
<td>$\bar{X}/\sigma, X_{(n)}/\sigma, X_{(1)}/\sigma$</td>
</tr>
</tbody>
</table>

In general, differences are pivotal in location family problems; ratios are pivotal for scale parameters.

Exercise: Suppose $X_1, X_2, ..., X_n$ are iid from

$$f_X(x|\mu) = f_Z(x - \mu),$$

where $-\infty < \mu < \infty$ and $f_Z(\cdot)$ is a standard pdf. Show that $Q(X, \mu) = \bar{X} - \mu$ is a pivotal quantity.
9.2.3 Pivoting the CDF

Example 9.8. Suppose $X_1, X_2, ..., X_n$ are iid with population pdf

$$f_X(x|\theta) = \begin{cases} e^{-(x-\theta)}, & x \geq \theta \\ 0, & x < \theta, \end{cases}$$

where $-\infty < \theta < \infty$. How can we obtain a confidence set for $\theta$? Note that, because $\{f_X(x|\theta) : -\infty < \theta < \infty\}$ is a location family, we could try working with

$$Q = Q(X, \theta) = \bar{X} - \theta.$$

From our recent discussion, we know that $Q$ is pivotal. In fact, it is easy to show that

$$X_i - \theta \sim \text{exponential}(1) \overset{d}{=} \text{gamma}(1, 1)$$

and hence

$$Q = \bar{X} - \theta = \frac{1}{n} \sum_{i=1}^{n} (X_i - \theta) \sim \text{gamma}(n, 1/n).$$

As expected, the distribution of $Q$ is free of $\theta$. Furthermore,

$$2nQ \sim \text{gamma}(n, 2) \overset{d}{=} \chi^2_{2n}.$$

Using $2nQ = 2n(\bar{X} - \theta)$ as a pivot, we can write

$$1 - \alpha = P_\theta(\chi^2_{2n,1-\alpha/2} \leq 2n(\bar{X} - \theta) \leq \chi^2_{2n,\alpha/2}) = P_\theta\left(\bar{X} - \frac{\chi^2_{2n,\alpha/2}}{2n} \leq \theta \leq \bar{X} - \frac{\chi^2_{2n,1-\alpha/2}}{2n}\right).$$

Therefore,

$$\left(\bar{X} - \frac{\chi^2_{2n,\alpha/2}}{2n}, \bar{X} - \frac{\chi^2_{2n,1-\alpha/2}}{2n}\right)$$

is a $1 - \alpha$ confidence set for $\theta$.

Criticism: Although this is a bonafide $1 - \alpha$ confidence set, it is not based on $T = T(X) = X_{(1)}$, a sufficient statistic for $\theta$. Let’s find a pivot based on $T$ instead. One example of such a pivot is $Q(T, \theta) = T - \theta \sim \text{exponential}(1/n)$. Another example is $Q(T, \theta) = F_T(T|\theta)$, the cdf of $T$, which is $\mathcal{U}(0, 1)$ by the Probability Integral Transformation.

CDF: It is easy to show that

$$F_T(t|\theta) = \begin{cases} 0, & t \leq \theta \\ 1 - e^{-n(t-\theta)}, & t > \theta. \end{cases}$$

Therefore, because $F_T(T|\theta) \sim \mathcal{U}(0, 1)$, we can write

$$1 - \alpha = P_\theta(\alpha/2 \leq F_T(T|\theta) \leq 1 - \alpha/2) = P_\theta(\alpha/2 \leq 1 - e^{-n(T-\theta)} \leq 1 - \alpha/2) = P_\theta\left(T + \frac{1}{n} \ln\left(\frac{\alpha}{2}\right) \leq \theta \leq T + \frac{1}{n} \ln\left(1 - \frac{\alpha}{2}\right)\right).$$
Figure 9.2: CDF of $T = X_{(1)}$ in Example 9.8, $F_T(t|\theta)$, plotted as a function of $\theta$ with $t$ fixed. The value of $t$ is 10.032, calculated based on an iid sample from $f_X(x|\theta)$ with $n = 5$. Dotted horizontal lines at $\alpha/2 = 0.025$ and $1 - \alpha/2 = 0.975$ have been added.

Therefore,

$$\left( T + \frac{1}{n} \ln \left( \frac{\alpha}{2} \right) , \; T + \frac{1}{n} \ln \left( 1 - \frac{\alpha}{2} \right) \right)$$

is a $1 - \alpha$ confidence set for $\theta$.

**Special case:** I used R to simulate an iid sample of size $n = 5$ from $f_X(x|\theta)$. The cdf of $T = X_{(1)}$ is plotted in Figure 9.2 as a function of $\theta$ with the observed value of $t = x_{(1)} = 10.032$ held fixed. A 0.95 confidence set is $(9.293, 10.026)$. The true value of $\theta$ is 10.

**Theorem 9.2.12.** Suppose $T$ is a statistic with a continuous cdf $F_T(t|\theta)$. Suppose $\alpha_1 + \alpha_2 = \alpha$. Suppose for all $t \in T$, the functions $\theta_L(t)$ and $\theta_U(t)$ are defined as follows:

- When $F_T(t|\theta)$ is a **decreasing** function of $\theta$,
  - $F_T(t|\theta_U(t)) = \alpha_1$
  - $F_T(t|\theta_L(t)) = 1 - \alpha_2$. 

• When $F_T(t|\theta)$ is an increasing function of $\theta$,
  
  - $F_T(t|\theta_U(t)) = 1 - \alpha_2$
  
  - $F_T(t|\theta_L(t)) = \alpha_1$.

Then the random interval $(\theta_L(T), \theta_U(T))$ is a $1 - \alpha_1 - \alpha_2$ confidence set for $\theta$.

Remark: Theorem 9.2.12 remains valid for any statistic $T$ with continuous cdf. In practice, we would likely want $T$ to be a sufficient statistic.

Remark: In practice, one often sets $\alpha_1 = \alpha_2 = \alpha/2$ so that $(\theta_L(T), \theta_U(T))$ is a $1 - \alpha$ confidence set. This is not necessarily the “optimal” approach, but it is reasonable in most situations. One sided confidence sets are obtained by letting either $\alpha_1$ or $\alpha_2$ equal 0.

Remark: Pivoting the cdf always “works” because (if $T$ is continuous), the cdf itself, when viewed as random, is a pivot. From the Probability Integral Transformation, we know that $F_T(T|\theta) \sim U(0,1)$. Therefore, when $F_T(t|\theta)$ is a decreasing function of $\theta$, we have

$$P_\theta(\theta_L(T) \leq \theta \leq \theta_U(T)) = P_\theta(\alpha_1 \leq F_T(T|\theta) \leq 1 - \alpha_2) = 1 - \alpha_1 - \alpha_2.$$ 

The case wherein $F_T(t|\theta)$ is an increasing function of $\theta$ is analogous.

Implementation: To pivot the cdf, it is not necessary that $F_T(t|\theta)$ be available in closed form (as in Example 9.8). All we really have to do is solve

$$\int_{-\infty}^{t_0} f_T(t|\theta_1^*(t_0)) dt \overset{\text{set}}{=} \alpha/2 \quad \text{and} \quad \int_{t_0}^{\infty} f_T(t|\theta_2^*(t_0)) dt \overset{\text{set}}{=} \alpha/2$$

(in the equal $\alpha_1 = \alpha_2 = \alpha/2$ case, say), based on the observed value $T = t_0$. We solve these equations for $\theta_1^*(t_0)$ and $\theta_2^*(t_0)$. One of these will be the lower limit $\theta_L(t_0)$ and the other will be the upper limit $\theta_U(t_0)$, depending on whether $F_T(t|\theta)$ is an increasing or decreasing function of $\theta$.

Remark: The discrete case (i.e., the statistic $T$ has a discrete distribution) is handled in the same way except that the integrals above are replaced by sums.

Example 9.9. Suppose $X_1, X_2, \ldots, X_n$ are iid Poisson($\theta$), where $\theta > 0$. We now pivot the cdf of $T = \sum_{i=1}^{n} X_i$, a sufficient statistic, to write a $1 - \alpha$ confidence set for $\theta$. Recall that $T = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\theta)$. If $T = t_0$ is observed, we set

$$P_\theta(T \leq t_0) = \sum_{k=0}^{t_0} \frac{(n\theta)^k e^{-n\theta}}{k!} \overset{\text{set}}{=} \alpha/2$$

$$P_\theta(T \geq t_0) = \sum_{k=t_0}^{\infty} \frac{(n\theta)^k e^{-n\theta}}{k!} \overset{\text{set}}{=} \alpha/2$$

and solve each equation for $\theta$. In practice, the solutions could be found by setting up a grid search over possible values of $\theta$ and then selecting the values that solve these equations (one solution will be the lower endpoint; the other solution will be the upper endpoint). In
this example, however, it is possible to get closed-form expressions for the confidence set endpoints. To see why, we need to recall the following result which “links” the Poisson and gamma distributions.

Result: If $X \sim \text{gamma}(a, b)$, $a \in \mathbb{N}$ (a positive integer), then

$$P(X \leq x) = P(Y \geq a),$$

where $Y \sim \text{Poisson}(x/b)$. This identity was stated in Example 3.3.1 (CB, pp 100-101).

Application: If we apply this result in Example 9.9 for the second equation to be solved, we have $a = t_0$, $x/b = n\theta$, and

$$\frac{\alpha}{2} = P_{\theta}(T \geq t_0) = P_{\theta}(X \leq bn\theta) = P_{\theta}\left(\frac{2X}{b} \leq 2n\theta\right) = P_{\theta}(\chi^2_{2t_0} \leq 2n\theta).$$

Therefore, we set

$$2n\theta = \chi^2_{2t_0,1-\alpha/2}$$
and solve for \( \theta \) (this will give the lower endpoint). A similar argument shows that the upper endpoint solves

\[ 2n\theta = \chi^2_{2(t_0+1),\alpha/2}. \]

Therefore, a \( 1 - \alpha \) confidence set for \( \theta \) is

\[ \left( \frac{1}{2n} \chi^2_{2t_0,1-\alpha/2}, \frac{1}{2n} \chi^2_{2(t_0+1),\alpha/2} \right). \]

**Remark:** When \( T \) is discrete, the coverage probability of \((\theta_L(T), \theta_U(T))\) found through pivoting the cdf will generally be a function of \( \theta \) and the interval itself will be conservative, that is,

\[ P_{\theta}(\theta_L(T) \leq \theta \leq \theta_U(T)) \geq 1 - \alpha, \]

for all \( \theta \in \Theta \). This is true because when \( T \) is discrete, the cdf \( F_T(T|\theta) \) is stochastically larger than a \( U(0,1) \) distribution. For example, consider Example 9.9 with \( n = 10 \) and \( 1 - \alpha = 0.90 \). Figure 9.3 shows that the coverage probability of a nominal 0.90 confidence interval is always at least 0.90 and can, in fact, be much larger than 0.90.

**Remark:** Pivoting a discrete cdf can be used to write confidence sets for parameters in other discrete distributions. For example, a \( 1 - \alpha \) confidence interval for a binomial probability \( p \) when using this technique is given by

\[ \left( \frac{1}{1 + \frac{n-x+1}{x} F_{2(n-x+1),2x,\alpha/2}}, \frac{x+1}{1 + \frac{x+1}{n-x} F_{2(x+1),2(n-x),\alpha/2}} \right), \]

where \( x \) is the realized value of \( X \sim b(n,p) \) and \( F_{a,b,\alpha/2} \) is the upper \( \alpha/2 \) quantile of an \( F \) distribution with degrees of freedom \( a \) and \( b \). This is known as the **Clopper-Pearson** confidence interval for \( p \) and it can (not surprisingly) be very conservative; see Brown et al. (2001, Statistical Science). The interval arises by first exploiting the relationship between the binomial and beta distributions (see CB, Exercise 2.40, pp 82) and then the relationship which “links” the beta and \( F \) distributions (see CB, Theorem 5.3.8, pp 225).

### 9.2.4 Bayesian intervals

**Recall:** In the Bayesian paradigm, all inference is carried out using the posterior distribution \( \pi(\theta|x) \). However, because the posterior \( \pi(\theta|x) \) is itself a legitimate probability distribution (for \( \theta \), updated after seeing \( x \)), we can calculate probabilities involving \( \theta \) directly by using this distribution.

**Definition:** For any set \( A \subseteq \mathbb{R} \), the **credible probability** associated with \( A \) is

\[ P(\theta \in A | X = x) = \int_A \pi(\theta|x) d\theta. \]

If the credible probability is \( 1 - \alpha \), we call \( A \) a \( 1 - \alpha \) **credible set**. If \( \pi(\theta|x) \) is discrete, we simply replace integrals with sums.
Note: Bayesian credible intervals are interpreted differently than (non-Bayesian) confidence intervals.

- **Confidence interval interpretation:** “If we were to perform the experiment over and over again, each time under identical conditions, and if we calculated a $1 - \alpha$ confidence interval each time the experiment was performed, then $100(1 - \alpha)$ percent of the intervals we calculated would contain the true value of $\theta$. Any specific interval we calculate represents one of these possible intervals.”

- **Credible interval interpretation:** “The probability our interval contains $\theta$ is $1 - \alpha$.”

**Example 9.10.** Suppose that $X_1, X_2, ..., X_n$ are iid Poisson($\theta$), where the prior distribution for $\theta \sim$ gamma($a, b$), $a, b$ known. In Example 7.10 (notes, pp 207-208), we showed that the posterior distribution

$$\theta | X = x \sim \text{gamma} \left( \frac{1}{n + \frac{1}{b}}, \frac{\sum_{i=1}^{n} x_i + a}{n + \frac{1}{b}} \right).$$

In Example 8.9 (notes, pp 244-245), we used this Bayesian model setup with the number of goals per game in the 2013-2014 English Premier League season and calculated the posterior distribution for the mean number of goals $\theta$ to be

$$\theta | X = x \sim \text{gamma} \left( \frac{1060 + 1.5}{380 + \frac{1}{2}}, \frac{1}{380 + \frac{1}{2}} \right) \overset{d}{=} \text{gamma}(1061.5, 0.002628).$$

A 0.95 credible set for $\theta$ is $(2.62, 2.96)$.

```r
> qgamma(0.025,1061.5,1/0.002628)
[1] 2.624309
> qgamma(0.975,1061.5,1/0.002628)
[1] 2.959913
```

Q: Why did we select the “equal-tail” quantiles (0.025 and 0.975) in this example?
A: It’s easy!

**Note:** There are two types of Bayesian credible intervals commonly used: Equal-tail (ET) intervals and highest posterior density (HPD) intervals.

**Definition:** The set $\mathcal{A}$ is a **highest posterior density (HPD)** $1 - \alpha$ credible set if

$$\mathcal{A} = \{ \theta : \pi(\theta | x) \geq c \}$$

and the credible probability of $\mathcal{A}$ is $1 - \alpha$. ET and HPD intervals will coincide only when $\pi(\theta | x)$ is symmetric.

**Remark:** In practice, because Monte Carlo methods are often used to approximate posterior distributions, simple ET intervals are usually the preferred choice. HPD intervals can be far more difficult to construct and are rarely much better than ET intervals.
9.3 Methods of Evaluating Interval Estimators

Note: We will not cover all of the material in this subsection. We will have only a brief discussion of the relevant topics.

Evaluating estimators: When evaluating any interval estimator, there are two important criteria to consider:

1. **Coverage probability.** When the coverage probability is not equal to \(1 - \alpha\) for all \(\theta \in \Theta\) (as is usually the case in discrete distributions), we would like it to be as close as possible to the nominal \(1 - \alpha\) level.
   - Some intervals maintain a coverage probability \(\geq 1 - \alpha\) for all \(\theta \in \Theta\) but can be very conservative (as in Example 9.9).
   - Confidence intervals based on large-sample theory might confer a coverage probability \(\leq 1 - \alpha\) for some/all \(\theta \in \Theta\), even though they are designed to be nominal as \(n \to \infty\). Large-sample intervals are discussed in Chapter 10.

2. **Interval length.** Shorter intervals are more informative. Interval length (or expected interval length) depends on the interval’s underlying confidence coefficient.
   - It only makes sense to compare two interval estimators (on the basis of interval length) when the intervals have the same coverage probability (or confidence coefficient).

**Example 9.11.** Suppose \(X_1, X_2, \ldots, X_n\) are iid \(\mathcal{N}(\mu, \sigma^2)\), where \(-\infty < \mu < \infty\) and \(\sigma^2 > 0\); i.e., both parameters are unknown. Set \(\theta = (\mu, \sigma^2)\). A \(1 - \alpha\) confidence interval for \(\mu\) is

\[
C(X) = \left( \bar{X} + a \frac{S}{\sqrt{n}}, \bar{X} + b \frac{S}{\sqrt{n}} \right),
\]

where the constants \(a\) and \(b\) are quantiles from the \(t_{n-1}\) distribution satisfying

\[
1 - \alpha = P_\theta \left( \bar{X} + a \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + b \frac{S}{\sqrt{n}} \right).
\]

Which choice of \(a\) and \(b\) is “best?” More precisely, which choice minimizes the expected length? The length of this interval is

\[
L = (b - a) \frac{S}{\sqrt{n}},
\]

which, of course, is random. The expected length is

\[
E_\theta(L) = (b - a) \frac{E_\theta(S)}{\sqrt{n}} = (b - a) c(n) \sigma / \sqrt{n},
\]

where the constant

\[
c(n) = \frac{\sqrt{2} \Gamma \left( \frac{n}{2} \right)}{\sqrt{n - 1} \Gamma \left( \frac{n - 1}{2} \right)}.
\]

Note that the expected length is proportional to \(b - a\).
Theorem 9.3.2. Suppose $Q = Q(X, \theta)$ is a pivotal quantity and 

$$P_\theta(a \leq Q \leq b) = 1 - \alpha,$$

where $a$ and $b$ are constants. Let $f_Q(q)$ denote the pdf of $Q$. If

1. $\int_a^b f_Q(q) dq = 1 - \alpha$
2. $f_Q(a) = f_Q(b) > 0$
3. $f'_Q(a) > f'_Q(b),$

then $b - a$ is minimized relative to $Q$.

Remark: The version of Theorem 9.3.2 stated in CB (pp 441-442) is slightly different than the one I present above; the authors’ version requires that the pdf of $Q$ be unimodal (mine requires that it be differentiable).

Application: Consider Example 9.11 with

$$Q = Q(X, \theta) = \frac{X - \mu}{S/\sqrt{n}}$$

and

$$C(x) = \left( \bar{x} + a \frac{s}{\sqrt{n}}, \bar{x} + b \frac{s}{\sqrt{n}} \right).$$

If we choose $a = -t_{n-1, \alpha/2}$ and $b = t_{n-1, \alpha/2}$, then the conditions in Theorem 9.3.2 are satisfied. Therefore,

$$\left( \frac{\bar{x} - t_{n-1, \alpha/2}}{S/\sqrt{n}}, \frac{\bar{x} + t_{n-1, \alpha/2}}{S/\sqrt{n}} \right)$$

has the shortest expected length among all $1 - \alpha$ confidence intervals based on $Q$.

Proof of Theorem 9.3.2. Suppose $Q \sim f_Q(q)$, where

$$1 - \alpha = P_\theta(a \leq Q \leq b) = F_Q(b) - F_Q(a)$$

so that

$$F_Q(b) = 1 - \alpha + F_Q(a)$$

and

$$b = F_Q^{-1}[1 - \alpha + F_Q(a)] \equiv b(a),$$

say. The goal is to minimize $b - a = b(a) - a$. Taking derivatives, we have (by the Chain Rule)

$$\frac{d}{da} [b(a) - a] = \frac{d}{da} [F_Q^{-1}[1 - \alpha + F_Q(a)] - 1$$

$$= \frac{d}{da} [1 - \alpha + F_Q(a)] \frac{d}{d\eta} F_Q^{-1}(\eta) - 1,$$

where $\eta = 1 - \alpha + F_Q(a)$. However, note that by the inverse function theorem (from calculus),

$$\frac{d}{d\eta} F_Q^{-1}(\eta) = \frac{1}{F'_Q[F_Q^{-1}(\eta)]} = \frac{1}{F'_Q(b)} = \frac{1}{f_Q(b)}.$$
Therefore,
\[
\frac{d}{da} [b(a) - a] = \frac{f_Q(a)}{f_Q(b)} - 1 \overset{\text{set}}{=} 0 \implies f_Q(a) = f_Q(b).
\]

To finish the proof, all we need to show is that
\[
\frac{d^2}{da^2} [b(a) - a] > 0
\]
whenever \( f_Q(a) = f_Q(b) \) and \( f'_Q(a) > f'_Q(b) \). This will guarantee that the conditions stated in Theorem 9.3.2 lead to \( b - a \) being minimized. \( \square \)

**Remark:** The theorem we have just proven is applicable when an interval’s length (or expected length) is proportional to \( b - a \). This is often true when \( \theta \) is a location parameter and \( f_X(x|\theta) \) is a location family. When an interval’s length is not proportional to \( b - a \), then Theorem 9.3.2 is not directly applicable. However, we might be able to formulate a modified version of the theorem that is applicable.

**Example 9.12.** Suppose \( X_1, X_2, \ldots, X_n \) are iid exponential(\( \beta \)), where \( \beta > 0 \). A pivotal quantity based on \( T = T(X) = \sum_{i=1}^n X_i \), a sufficient statistic, is
\[
Q = Q(T, \beta) = \frac{2T}{\beta} \sim \chi^2_{2n}.
\]

Therefore, we can write
\[
1 - \alpha = P_\beta(a \leq Q \leq b) = P_\beta \left( a \leq \frac{2T}{\beta} \leq b \right)
= P_\beta \left( \frac{2T}{b} \leq \beta \leq \frac{2T}{a} \right),
\]
where \( a \) and \( b \) are quantiles from the \( \chi^2_{2n} \) distribution. In this example, the expected interval length is not proportional to \( b - a \). Instead, the expected length is
\[
E_\beta(L) = E_\beta \left( \frac{2T}{a} - \frac{2T}{b} \right) = \left( \frac{1}{a} - \frac{1}{b} \right) E_\beta(2T)
= \left( \frac{1}{a} - \frac{1}{b} \right) 2n\beta,
\]
which is proportional to
\[
\frac{1}{a} - \frac{1}{b}.
\]
Theorem 9.3.2 is therefore not applicable here. To modify the theorem (towards finding a shortest expected length confidence interval based on \( Q \)), we would have to minimize
\[
\frac{1}{a} - \frac{1}{b} = \frac{1}{a} - \frac{1}{b(a)}
\]
with respect to \( a \) subject to the constraint that
\[
\int_a^{b(a)} f_Q(q) dq = 1 - \alpha,
\]
where \( f_Q(q) \) is the pdf of \( Q \sim \chi^2_{2n} \). See CB (pp 444).
10 Asymptotic Evaluations

Complementary reading: Chapter 10 (CB).

10.1 Introduction

**Preview:** In this chapter, we revisit “large sample theory” and discuss three important topics in statistical inference:

1. **Point estimation** (Section 10.1)
   - Efficiency, consistency
   - Large sample properties of maximum likelihood estimators

2. **Hypothesis testing** (Section 10.3)
   - Wald, score, LRT
   - asymptotic distributions

3. **Confidence intervals** (Section 10.4)
   - Wald, score, LRT

Our previous inference discussions (i.e., in Chapters 7-9 CB) dealt with finite sample topics (i.e., unbiasedness, MSE, optimal estimators/tests, confidence intervals based on finite sample pivots, etc.). We now investigate **large sample inference**, a topic of utmost importance in statistical research.

10.2 Point Estimation

**Setting:** We observe \( X = (X_1, X_2, ..., X_n) \sim f_X(x|\theta) \), where \( \theta \in \Theta \subseteq \mathbb{R} \). Usually, \( X_1, X_2, ..., X_n \) will constitute a random sample (an iid sample) from a population \( f_X(x|\theta) \). We regard the scalar parameter \( \theta \) as fixed and unknown. Define

\[
W_n = W_n(X) = W_n(X_1, X_2, ..., X_n)
\]

to be a sequence of estimators. For example,

\[
W_1 = X_1,
W_2 = \frac{X_1 + X_2}{2},
W_3 = \frac{X_1 + X_2 + X_3}{3},
\]

and so on, so that in general,

\[
W_n = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]

Note that we emphasize the dependence of this sequence on the sample size \( n \).
Definition: A sequence of estimators $W_n$ is consistent for a parameter $\theta$ if

$$W_n \xrightarrow{p} \theta \text{ for all } \theta \in \Theta.$$  

That is, for all $\epsilon > 0$ and for all $\theta \in \Theta$,

$$\lim_{n \to \infty} P_\theta(|W_n - \theta| \geq \epsilon) = 0.$$ 

An equivalent definition is

$$\lim_{n \to \infty} P_\theta(|W_n - \theta| < \epsilon) = 1.$$ 

We call $W_n$ a consistent estimator of $\theta$. What makes consistency “different” from our usual definition of convergence in probability is that we require $W_n \xrightarrow{p} \theta$ for all $\theta \in \Theta$. In other words, convergence of $W_n$ must result for all members of the family $\{f_X(x|\theta) : \theta \in \Theta\}$.

Remark: From Markov’s Inequality, we know that for all $\epsilon > 0$,

$$P_\theta(|W_n - \theta| \geq \epsilon) \leq \frac{E_\theta[(W_n - \theta)^2]}{\epsilon^2}.$$ 

Therefore, a sufficient condition for $W_n$ to be consistent is

$$\frac{E_\theta[(W_n - \theta)^2]}{\epsilon^2} \to 0$$

for all $\theta \in \Theta$. However, note that

$$E_\theta[(W_n - \theta)^2] = \text{var}_\theta(W_n) + [E_\theta(W_n) - \theta]^2 = \text{var}_\theta(W_n) + [\text{Bias}_\theta(W_n)]^2.$$ 

This leads to the following theorem.

**Theorem 10.1.3.** If $W_n$ is a sequence of estimators of a parameter $\theta$ satisfying

1. $\text{var}_\theta(W_n) \to 0$, as $n \to \infty$, for all $\theta \in \Theta$
2. $\text{Bias}_\theta(W_n) \to 0$, as $n \to \infty$, for all $\theta \in \Theta$,

then $W_n$ is a consistent estimator of $\theta$.

**Weak Law of Large Numbers:** Suppose that $X_1, X_2, \ldots, X_n$ are iid with $E_\theta(X_1) = \mu$ and $\text{var}_\theta(X_1) = \sigma^2 < \infty$. Let

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

denote the sample mean. As an estimator of $\mu$, it is easy to see that the conditions of Theorem 10.1.3 are satisfied. Therefore, $\overline{X}_n$ is a consistent estimator of $E_\theta(X_1) = \mu$.

**Continuity:** Suppose $W_n$ is a consistent estimator of $\theta$. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a continuous function. Then

$$g(W_n) \xrightarrow{p} g(\theta) \text{ for all } \theta \in \Theta.$$ 

That is, $g(W_n)$ is a consistent estimator of $g(\theta)$.
**Consistency of MLEs:** Suppose $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta$. Let 

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta|x)$$

denote the maximum likelihood estimator (MLE) of $\theta$. Under “certain regularity conditions,” it follows that 

$$\hat{\theta} \xrightarrow{p} \theta \quad \text{for all } \theta \in \Theta,$$

as $n \to \infty$. That is, MLEs are consistent estimators.

**Remark:** Consistency also results for vector valued MLEs, say $\hat{\boldsymbol{\theta}}$, but we herein restrict attention to the scalar case.

**Sufficient conditions to prove consistency of MLEs:**

1. $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$.
2. The parameter $\theta$ is identifiable, that is, for $\theta_1, \theta_2 \in \Theta$,

$$f_X(x|\theta_1) = f_X(x|\theta_2) \implies \theta_1 = \theta_2.$$

In other words, different values of $\theta$ cannot produce the same probability distribution.

3. The family of pdfs $\{f_X(x|\theta) : \theta \in \Theta\}$ has common support $\mathcal{X}$. This means that the support does not depend on $\theta$. In addition, the pdf $f_X(x|\theta)$ is differentiable with respect to $\theta$.

4. The parameter space $\Theta$ contains an open set where the true value of $\theta$, say $\theta_0$, resides as an interior point.

**Remark:** Conditions 1-4 generally hold for exponential families that are of full rank.

**Example 10.1.** Suppose $X_1, X_2, \ldots, X_n$ are iid $\mathcal{N}(0, \theta)$, where $\theta > 0$. The MLE of $\theta$ is 

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

As an MLE, $\hat{\theta} \xrightarrow{p} \theta$, for all $\theta > 0$; i.e., $\hat{\theta}$ is a consistent estimator of $\theta$.

**Asymptotic normality of MLEs:** Suppose $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta$. Let $\hat{\theta}$ denote the MLE of $\theta$. Under “certain regularity conditions,” it follows that 

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

as $n \to \infty$, where the asymptotic variance 

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

Recall that $I_1(\theta)$, the **Fisher Information** based on one observation, is given by 

$$I_1(\theta) = E_\theta \left( \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right) = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right].$$
Remark: The four regularity conditions on the last page were sufficient conditions for consistency. For asymptotic normality, there are two additional sufficient conditions:

5. The pdf/pmf \( f_X(x|\theta) \) is three times differentiable with respect to \( \theta \), the third derivative is continuous in \( \theta \), and \( \int_{\mathbb{R}} f_X(x|\theta)dx \) can be differentiated three times under the integral sign.

6. There exists a function \( M(x) \) such that
\[
\left| \frac{\partial^3}{\partial \theta^3} \ln f_X(x|\theta) \right| \leq M(x)
\]
for all \( x \in \mathcal{X} \) for all \( \theta \in N_c(\theta_0) \) \( \exists c > 0 \) and \( E_{\theta_0}[M(X)] < \infty \).

Note: We now sketch a casual proof of the asymptotic normality result for MLEs. Let \( \theta_0 \) denote the true value of \( \theta \). Let \( S(\theta) = S(\theta|x) \) denote the score function; i.e.,
\[
S(\theta) = \frac{\partial}{\partial \theta} \ln f_X(x|\theta).
\]
Note that because \( \hat{\theta} \) is an MLE, it solves the score equation; i.e., \( S(\hat{\theta}) = 0 \). Therefore, we can write (via Taylor series expansion about \( \theta_0 \)),
\[
0 = S(\hat{\theta}) = S(\theta_0) + \frac{\partial S(\theta_0)}{\partial \theta}(\hat{\theta} - \theta_0) + \frac{1}{2} \frac{\partial^2 S(\hat{\theta}_*)}{\partial \theta^2}(\hat{\theta} - \theta_0)^2
\]
where \( \hat{\theta}_* \) is between \( \theta_0 \) and \( \hat{\theta} \). Therefore, we have
\[
0 = S(\theta_0) + (\hat{\theta} - \theta_0) \left[ \frac{\partial S(\theta_0)}{\partial \theta} + \frac{1}{2} \frac{\partial^2 S(\hat{\theta}_*)}{\partial \theta^2} (\hat{\theta} - \theta_0) \right].
\]
After simple algebra, we have
\[
\sqrt{n}(\hat{\theta} - \theta_0) = \frac{-\sqrt{n}S(\theta_0)}{\frac{\partial S(\theta_0)}{\partial \theta} + \frac{1}{2} \frac{\partial^2 S(\hat{\theta}_*)}{\partial \theta^2}(\hat{\theta} - \theta_0)} = \frac{-A}{B + C},
\]
where
\[
A = \sqrt{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta_0)
\]
\[
B = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f_X(X_i|\theta_0)
\]
\[
C = \frac{1}{2n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \ln f_X(X_i|\hat{\theta}_*)(\hat{\theta} - \theta_0).
\]
The first term
\[ A = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta_0) \xrightarrow{d} \mathcal{N}(0, I_1(\theta_0)). \]

Proof. For general \( \theta \), define
\[ Y_i = \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta), \]
for \( i = 1, 2, \ldots, n \). The \( Y_i \)'s are iid with mean
\[ E_\theta(Y) = E_\theta \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = \int_{\mathbb{R}} \frac{\partial}{\partial \theta} \ln f_X(x|\theta) f_X(x|\theta) dx = 0, \]
and variance
\[ \text{var}_\theta(Y) = \text{var}_\theta \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = E_\theta \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\} = I_1(\theta). \]

Therefore, the CLT says that
\[ A = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta) \xrightarrow{d} \mathcal{N}(0, I_1(\theta_0)), \]
as \( n \to \infty \). Note that when \( \theta = \theta_0 \), we have
\[ -A = -\sqrt{n}(\bar{Y} - 0) \xrightarrow{d} \mathcal{N}(0, I_1(\theta_0)), \]
because the \( \mathcal{N}(0, I_1(\theta)) \) limiting distribution above is symmetric about 0. \( \Box \)

The second term, by WLLN,
\[ B = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ln f_X(X_i|\theta_0) \xrightarrow{p} E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta_0) \right] = -I_1(\theta_0). \]

The third term
\[ C = \frac{1}{2n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \ln f_X(X_i|\theta_0) \frac{\hat{\theta} - \theta_0}{p} \to 0. \]

Proof (very casual). We have
\[ C = \frac{1}{2} (\hat{\theta} - \theta_0) \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \ln f_X(X_i|\hat{\theta}). \]

Note that \( \hat{\theta} - \theta_0 \xrightarrow{p} 0 \), because \( \hat{\theta} \) is consistent (i.e., \( \hat{\theta} \) converges in probability to \( \theta_0 \)). Therefore, it suffices to show that
\[ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \ln f_X(X_i|\hat{\theta}). \]
converges to something finite (in probability). Note that for \( n \) “large enough,” i.e., as soon as \( \hat{\theta}_n \in N_c(\theta_0) \) in Regularity Condition 6,

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^3}{\partial \theta^3} \ln f_X(X_i | \hat{\theta}_n) \leq \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\partial^3}{\partial \theta^3} \ln f_X(X_i | \hat{\theta}_n) \right| \leq \frac{1}{n} \sum_{i=1}^{n} M(X_i) \overset{p}{\rightarrow} E_{\theta_0}[M(X)] < \infty. \]

We have shown that \( C \overset{p}{\rightarrow} 0 \) and hence \( B + C \overset{p}{\rightarrow} -I_1(\theta_0) \). Finally, note that

\[
\frac{-A}{B+C} = \underbrace{-A}_{\overset{d}{\rightarrow} N(0,I_1(\theta_0))} \underbrace{\frac{1}{B+C}}_{\overset{p}{\rightarrow} \frac{1}{I_1(\theta_0)}} \overset{d}{\rightarrow} \mathcal{N} \left( 0, \frac{1}{I_1(\theta_0)} \right),
\]

by Slutsky’s Theorem. \( \square \)

**Remark:** We have shown that, under regularity conditions, an MLE \( \hat{\theta} \) satisfies

\[
\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\rightarrow} \mathcal{N}(0,v(\theta)),
\]

where

\[
v(\theta) = \frac{1}{I_1(\theta)}.
\]

Now recall the **Delta Method** from Chapter 5; i.e., if \( g : \mathbb{R} \rightarrow \mathbb{R} \) is differentiable at \( \theta \) and \( g'(\theta) \neq 0 \), then

\[
\sqrt{n}[g(\hat{\theta}) - g(\theta)] \overset{d}{\rightarrow} \mathcal{N} \left( 0, [g'(\theta)]^2 v(\theta) \right).
\]

Therefore, not only are MLEs asymptotically normal, but functions of MLEs are too.

**Example 10.1** (continued). Suppose \( X_1, X_2, \ldots, X_n \) are iid \( \mathcal{N}(0,\theta) \), where \( \theta > 0 \). The MLE of \( \theta \) is

\[
\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i^2.
\]

We know that \( \hat{\theta} \overset{p}{\rightarrow} \theta \), as \( n \rightarrow \infty \). We now derive the asymptotic distribution of \( \hat{\theta} \) (suitably centered and scaled). We know

\[
\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\rightarrow} \mathcal{N}(0,v(\theta)),
\]

where

\[
v(\theta) = \frac{1}{I_1(\theta)}.
\]

Therefore, all we need to do is calculate \( I_1(\theta) \). The pdf of \( X \) is, for all \( x \in \mathbb{R} \),

\[
f_X(x|\theta) = \frac{1}{\sqrt{2\pi \theta}} e^{-x^2/2\theta}.
\]

Therefore,

\[
\ln f_X(x|\theta) = -\frac{1}{2} \ln(2\pi \theta) - \frac{x^2}{2\theta}.
\]
The derivatives of \( \ln f_X(x|\theta) \) are

\[
\begin{align*}
\frac{\partial}{\partial \theta} \ln f_X(x|\theta) &= -\frac{1}{2\theta} + \frac{x^2}{2\theta^2}, \\
\frac{\partial^2}{\partial \theta^2} \ln f_X(x|\theta) &= \frac{1}{2\theta^2} - \frac{x^2}{\theta^3}.
\end{align*}
\]

Therefore,

\[
I_1(\theta) = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right] = E_\theta \left( \frac{X^2}{\theta^3} - \frac{1}{2\theta^2} \right) = \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{2\theta^2}
\]

and

\[
v(\theta) = \frac{1}{I_1(\theta)} = 2\theta^2.
\]

We have

\[
\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2).
\]

**Exercise:** Use the Delta Method to derive the large sample distributions of

\[
g_1(\hat{\theta}) = \hat{\theta}^2,
g_2(\hat{\theta}) = e^{\hat{\theta}}, \text{ and } g_3(\hat{\theta}) = \ln \hat{\theta},
\]

suitably centered and scaled.

**Important:** Suppose that an MLE \( \hat{\theta} \) (or any sequence of estimators) satisfies

\[
\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)).
\]

Suppose that \( \hat{v}(\theta) \) is a consistent estimator of \( v(\theta) \), that is,

\[
\hat{v}(\theta) \xrightarrow{p} v(\theta),
\]

for all \( \theta \in \Theta \) as \( n \to \infty \). We know that

\[
Z_n = \frac{\hat{\theta} - \theta}{\sqrt{\hat{v}(\theta)}} \xrightarrow{d} \mathcal{N}(0, 1).
\]

In addition,

\[
Z_n^* = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{v}(\theta)}{n}}} = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\hat{v}(\theta)}{n}}} \sqrt{\frac{\hat{v}(\theta)}{v(\theta)}} \xrightarrow{d} \mathcal{N}(0, 1),
\]

by Slutsky’s Theorem. Note that

\[
\sqrt{\frac{\hat{v}(\theta)}{v(\theta)}} \xrightarrow{p} 1
\]

because of continuity. This technique is widely used in large sample arguments.
Summary:

1. We start with a sequence of estimators (e.g., an MLE sequence, etc.) satisfying
   \[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)). \]

2. We find a consistent estimator of the asymptotic variance, say \( v(\hat{\theta}) \).

3. Slutsky’s Theorem and continuity of convergence are used to show that
   \[ Z^*_n = \frac{\hat{\theta} - \theta}{\sqrt{v(\hat{\theta})}} \xrightarrow{d} \mathcal{N}(0, 1). \]

One can then use \( Z^*_n \) to formulate large sample (Wald) hypothesis tests and confidence
intervals; see Sections 10.3 and 10.4, respectively.

**Example 10.1** (continued). Suppose \( X_1, X_2, ..., X_n \) are iid \( \mathcal{N}(0, \theta) \), where \( \theta > 0 \). The MLE of \( \theta \) is
\[ \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i^2. \]

We have shown that
\[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 2\theta^2) \iff \frac{\hat{\theta} - \theta}{\sqrt{2\theta^2/n}} \xrightarrow{d} \mathcal{N}(0, 1). \]

A consistent estimator of \( v(\theta) = 2\theta^2 \) is \( v(\hat{\theta}) = 2\hat{\theta}^2 \), by continuity. Therefore,
\[ Z^*_n = \frac{\hat{\theta} - \theta}{\sqrt{2\theta^2/n}} \xrightarrow{d} \mathcal{N}(0, 1), \]

by Slutsky’s Theorem.

**Definition:** Suppose we have two competing sequences of estimators (neither of which is
necessarily an MLE sequence) denoted by \( W_n \) and \( V_n \) that satisfy
\[ \sqrt{n}(W_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2_W), \]
\[ \sqrt{n}(V_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2_V). \]

Both estimators are consistent estimators of \( \theta \). Define the **asymptotic relative efficiency** (ARE) as
\[ \text{ARE}(W_n \text{ to } V_n) = \frac{\sigma^2_W}{\sigma^2_V}. \]
With this definition, the following interpretations are used:

1. If \( \text{ARE} < 1 \), then \( W_n \) is more efficient than \( V_n \).
2. If \( \text{ARE} = 1 \), then \( W_n \) is as efficient as \( V_n \).
3. If \( \text{ARE} > 1 \), then \( W_n \) is less efficient than \( V_n \).

The ARE is commonly used to compare the variances of two competing consistent estimators; the comparison is of course on the basis of each estimator’s large sample distribution.

**Remark:** Before we do an example illustrating ARE, let’s have a brief discussion about sample quantile estimators.

**Sample quantiles:** Suppose \( X_1, X_2, ..., X_n \) are iid with continuous cdf \( F \). Define

\[
\phi_p = F^{-1}(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}.
\]

We call \( \phi_p \) the \( p \)th quantile of the distribution of \( X \). Note that if \( F \) is strictly increasing, then \( F^{-1}(p) \) is well defined by

\[
\phi_p = F^{-1}(p) \iff F(\phi_p) = p.
\]

The simplest definition of the sample \( p \)th quantile is

\[
\hat{\phi}_p = \hat{F}_n^{-1}(p) = \begin{cases} X_{(np)}, & np \in \mathbb{Z}^+ \\ X_{([np]+1)}, & \text{otherwise} \end{cases}
\]

is the empirical distribution function (edf). The edf is a non-decreasing step function that takes steps of size \( 1/n \) at each observed \( X_i \). Therefore,

\[
\hat{\phi}_p \equiv \hat{F}_n^{-1}(p) = \left\{ \begin{array}{ll}
X_{(np)}, & np \in \mathbb{Z}^+ \\
X_{([np]+1)}, & \text{otherwise}
\end{array} \right.
\]

This is just a fancy way of saying that the sample \( p \)th quantile is one of the order statistics (note that other books may define this differently; e.g., by averaging order statistics, etc.). Whenever I teach STAT 823, I prove that

\[
\sqrt{n}(\hat{\phi}_p - \phi_p) \xrightarrow{d} \mathcal{N} \left( 0, \frac{p(1-p)}{f^2(\phi_p)} \right),
\]

where \( f \) is the population pdf of \( X \). For example, if \( p = 0.5 \), then \( \phi_p = \phi_{0.5} \) is the median of \( X \) and the sample median \( \hat{\phi}_{0.5} \) satisfies

\[
\sqrt{n}(\hat{\phi}_{0.5} - \phi_{0.5}) \xrightarrow{d} \mathcal{N} \left( 0, \frac{1}{4f^2(\phi_{0.5})} \right).
\]

**Example 10.2.** Suppose \( X_1, X_2, ..., X_n \) are iid \( \mathcal{N}(\mu, \sigma^2) \), where \( -\infty < \mu < \infty \) and \( \sigma^2 > 0 \); i.e., both parameters are unknown. Consider the following two estimators \( W_n = \bar{X}_n \) and \( V_n = \hat{\phi}_{0.5} \) as estimators of \( \mu \). Note that because the \( \mathcal{N}(\mu, \sigma^2) \) population distribution is symmetric, the population median \( \phi_{0.5} = \mu \) as well.
We know that
\[ \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2), \]
that is, this “limiting distribution” is the exact distribution of \( \sqrt{n}(\overline{X}_n - \mu) \) for each \( n \). From our previous discussion on sample quantiles, we know that
\[ \sqrt{n}(\hat{\phi}_{0.5} - \mu) \xrightarrow{d} N\left(0, \frac{1}{4f^2(\phi_{0.5})}\right), \]
where (under the normal assumption),
\[ \frac{1}{4f^2(\phi_{0.5})} = \frac{1}{4f^2(\mu)} = \frac{1}{4 \left( \frac{1}{\sqrt{2\pi\sigma}} \right)^2} = \frac{\pi}{2}\sigma^2. \]

Therefore, the asymptotic relative efficiency of the sample median \( \hat{\phi}_{0.5} \) when compared to the sample mean \( \overline{X}_n \) is
\[ \text{ARE}(\hat{\phi}_{0.5} \text{ to } \overline{X}_n) = \frac{\pi}{2\sigma^2} = \frac{\pi}{2} \approx 1.57. \]

**Interpretation:** The sample median \( \hat{\phi}_{0.5} \) would require 57 percent more observations to achieve the same level of (asymptotic) precision as \( \overline{X}_n \).

**Example 10.3.** Suppose \( X_1, X_2, \ldots, X_n \) are iid beta(\( \theta, 1 \)), where \( \theta > 0 \).

- Show that the MOM estimator of \( \theta \) is
  \[ \hat{\theta}_{\text{MOM}} = \frac{\overline{X}}{1 - \overline{X}} \]
  and that \( \hat{\theta}_{\text{MOM}} \) satisfies
  \[ \sqrt{n}(\hat{\theta}_{\text{MOM}} - \theta) \xrightarrow{d} N\left(0, \frac{\theta(\theta + 1)^2}{\theta + 2}\right). \]
  **Hint:** Use CLT and Delta Method.

- Show that the MLE of \( \theta \) is
  \[ \hat{\theta}_{\text{MLE}} = -\frac{n}{\sum_{i=1}^{n} \ln X_i} \]
  and that \( \hat{\theta}_{\text{MLE}} \) satisfies
  \[ \sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} N(0, \theta^2). \]
  **Hint:** Use large sample results for MLEs.

- Show that
  \[ \text{ARE}(\hat{\theta}_{\text{MOM}} \text{ to } \hat{\theta}_{\text{MLE}}) = \frac{(\theta + 1)^2}{\theta(\theta + 2)}. \]

- I graphed \( \text{ARE}(\hat{\theta}_{\text{MOM}} \text{ to } \hat{\theta}_{\text{MLE}}) \) as a function of \( \theta \) in Figure 10.1. Note that ARE is always greater than unity, meaning that the MOM estimator is not as efficient as the MLE.
10.3 Hypothesis Testing

Remark: In Chapter 8 (CB), we discussed methods to derive hypothesis tests and also optimality issues based on finite sample criteria. These discussions revealed that optimal tests (e.g., UMP tests) were available for just a small collection of problems (some of which were not realistic).

Preview: In this section, we present three large sample approaches to formulate hypothesis tests:

1. Wald (1943)
2. Score (1947, Rao); also known as “Lagrange multiplier tests”

These are known as the “large sample likelihood based tests.”
10.3.1 Wald tests

Recall: Suppose $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. As long as suitable regularity conditions hold, we know that an MLE $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

If $v(\theta)$ is a continuous function of $\theta$, then

$$v(\hat{\theta}) \xrightarrow{p} v(\theta),$$

for all $\theta$; i.e., $v(\hat{\theta})$ is a consistent estimator of $v(\theta)$, and

$$Z_n^* = \frac{\hat{\theta} - \theta}{\sqrt{\frac{v(\hat{\theta})}{n}}} = \frac{\hat{\theta} - \theta}{\sqrt{\frac{v(\theta)}{n}}} \xrightarrow{d} \mathcal{N}(0,1),$$

by Slutsky’s Theorem. This forms the basis for the Wald test.

Wald statistic: Suppose $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider testing

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta \neq \theta_0.$$

When $H_0$ is true, then

$$Z_n^W = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{v(\hat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Therefore,

$$R = \{x \in X : |z_n^W| \geq z_{\alpha/2}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0,1)$ distribution, is an approximate size $\alpha$ rejection region for testing $H_0$ versus $H_1$. One sided tests also use $Z_n^W$. The only thing that changes is the form of $R$.

Example 10.4. Suppose $X_1, X_2, \ldots, X_n$ are iid Bernoulli($p$), where $0 < p < 1$. Derive the Wald test of

$$H_0 : p = p_0$$

versus

$$H_1 : p \neq p_0.$$
Solution. We already know that the MLE of \( p \) is given by
\[
\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i,
\]
the so-called “sample proportion.” Because \( \hat{p} \) is an MLE, we know that
\[
\sqrt{n}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, v(p)),
\]
where
\[
v(p) = \frac{1}{I_1(p)}.
\]
We now calculate \( I_1(p) \). The pmf of \( X \) is, for \( x = 0, 1 \),
\[
f_X(x|p) = p^x(1-p)^{1-x}.
\]
Therefore,
\[
\ln f_X(x|p) = x \ln p + (1-x) \ln(1-p).
\]
The derivatives of \( \ln f_X(x|p) \) are
\[
\frac{\partial}{\partial p} \ln f_X(x|p) = \frac{x}{p} - \frac{1-x}{1-p},
\]
\[
\frac{\partial^2}{\partial p^2} \ln f_X(x|p) = -\frac{x}{p^2} + \frac{1-x}{(1-p)^2}.
\]
Therefore,
\[
I_1(p) = -E_p \left[ \frac{\partial^2}{\partial p^2} \ln f_X(X|p) \right] = E_p \left[ \frac{X}{p^2} + \frac{1-X}{(1-p)^2} \right] = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}
\]
and
\[
v(p) = \frac{1}{I_1(p)} = p(1-p).
\]
We have
\[
\sqrt{n}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p)).
\]
Because the asymptotic variance \( v(p) = p(1-p) \) is a continuous function of \( p \), it can be consistently estimated by \( v(\hat{p}) = \hat{p}(1-\hat{p}) \). The Wald statistic to test \( H_0 : p = p_0 \) versus \( H_1 : p \neq p_0 \) is given by
\[
Z^W_n = \frac{\hat{p} - p_0}{\sqrt{v(\hat{p})}} = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})}}.
\]
An approximate size \( \alpha \) rejection region is
\[
R = \{ x \in \mathcal{X} : |z^W_n| \geq z_{\alpha/2} \}.\]
10.3.2 Score tests

**Motivation:** Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Recall that the score function, when viewed as random, is

$$S(\theta|X) = \frac{\partial}{\partial \theta} \ln L(\theta|X) \quad \text{iid} = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f_X(X_i|\theta),$$

the sum of iid random variables. Recall that

$$E_{\theta} \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = 0$$

$$\text{var}_{\theta} \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right] = E_{\theta} \left\{ \left[ \frac{\partial}{\partial \theta} \ln f_X(X|\theta) \right]^2 \right\} = I_1(\theta).$$

Therefore, applying the CLT to the sum above, we have

$$\sqrt{n} \left( \frac{1}{n} S(\theta|X) - 0 \right) \xrightarrow{d} \mathcal{N}(0, I_1(\theta)),$$

which means

$$\frac{1}{n} S(\theta|X) \xrightarrow{\text{iid}} \frac{S(\theta|X)}{\sqrt{n I_1(\theta)}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where recall $I_n(\theta) = n I_1(\theta)$ is the Fisher information based on all $n$ iid observations. Therefore, the score function divided by the square root of the Fisher information (based on all $n$ observations) behaves asymptotically like a $\mathcal{N}(0, 1)$ random variable. This fact forms the basis for the score test.

**Score statistic:** Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider testing

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta \neq \theta_0.$$

When $H_0$ is true, then

$$Z_n^S = \frac{S(\theta_0|X)}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Therefore,

$$R = \{ x \in \mathcal{X} : |z_n^S| \geq z_{\alpha/2} \},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the $\mathcal{N}(0, 1)$ distribution, is an approximate size $\alpha$ rejection region for testing $H_0$ versus $H_1$. One sided tests also use $Z_n^S$. The only thing that changes is the form of $R$.
Example 10.5. Suppose $X_1, X_2, \ldots, X_n$ are iid Bernoulli($p$), where $0 < p < 1$. Derive the score test of

$$H_0 : p = p_0$$

versus

$$H_1 : p \neq p_0.$$ 

Solution. The likelihood function is given by

$$L(p|x) = \prod_{i=1}^{n} p^{x_i} (1 - p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i} (1 - p)^{n - \sum_{i=1}^{n} x_i}.$$ 

The log-likelihood function is

$$\ln L(p|x) = \sum_{i=1}^{n} x_i \ln p + \left( n - \sum_{i=1}^{n} x_i \right) \ln(1 - p).$$ 

The score function is

$$S(p|x) = \frac{\partial}{\partial p} \ln L(p|x) = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1 - p}.$$ 

Recall in Example 10.4, we calculated

$$I_1(p) = \frac{1}{p(1-p)}.$$ 

Therefore, the score statistic is

$$Z_n^S = \frac{S(p_0|X)}{\sqrt{I_n(p_0)}} = \frac{\sum_{i=1}^{n} X_i}{p_0} \frac{n}{1 - p_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}.$$ 

An approximate size $\alpha$ rejection region is

$$R = \{ x \in \mathcal{X} : |z_n^S| \geq z_{\alpha/2} \}.$$ 

Remark: It is insightful to compare

$$Z_n^W = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}} \quad \text{with} \quad Z_n^S = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}.$$ 

The two statistics differ only in how the standard error of $\hat{p}$ (as a point estimator of $p$) is calculated. The Wald statistic uses the estimated standard error. The score statistic uses the standard error calculated under the assumption that $H_0 : p = p_0$ is true (i.e., nothing is being estimated). This is an argument in favor of the score statistic.
10.3.3 Likelihood ratio tests

**Setting:** Suppose $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Consider testing

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta \neq \theta_0.$$ 

The likelihood ratio test (LRT) statistic is defined as

$$\lambda(x) = \sup_{\theta \in \Theta} \frac{L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x) = \frac{L(\theta_0|x)}{L(\theta_0)}.$$ 

Suppose the regularity conditions needed for MLEs to be consistent and asymptotically normal hold. When $H_0$ is true,

$$-2 \ln \lambda(X) \xrightarrow{d} \chi^2_1.$$ 

Because small values of $\lambda(x)$ are evidence against $H_0$, large values of $-2 \ln \lambda(x)$ are too. Therefore,

$$R = \{ x \in X : -2 \ln \lambda(x) \geq \chi^2_{1,\alpha} \},$$

where $\chi^2_{1,\alpha}$ is the upper $\alpha$ quantile of the $\chi^2_1$ distribution, is an approximate size $\alpha$ rejection region for testing $H_0$ versus $H_1$.

**Proof.** Our proof is casual. Suppose $H_0 : \theta = \theta_0$ is true. First, write $\ln L(\hat{\theta})$ in a Taylor series expansion about $\theta_0$, that is,

$$\ln L(\hat{\theta}) = \ln L(\theta_0) + (\hat{\theta} - \theta_0) \frac{\partial}{\partial \theta} \ln L(\theta_0) + \frac{1}{2} (\hat{\theta} - \theta_0)^2 \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}) \left( \right) = \ln L(\theta_0) + \sqrt{n} (\hat{\theta} - \theta_0) \frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \ln L(\theta_0) + \frac{n}{2} (\hat{\theta} - \theta_0)^2 \frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}), \quad (10.1)$$

where $\hat{\theta}$ is between $\hat{\theta}$ and $\theta_0$. Now write $\frac{\partial}{\partial \theta} \ln L(\theta_0)$ in a Taylor series expansion about $\hat{\theta}$, that is,

$$\frac{\partial}{\partial \theta} \ln L(\theta_0) = \frac{\partial}{\partial \theta} \ln L(\hat{\theta}) + (\theta_0 - \hat{\theta}) \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}),$$

where $\hat{\theta}$ is between $\theta_0$ and $\hat{\theta}$. Note that $\frac{\partial}{\partial \theta} \ln L(\hat{\theta}) = 0$ because $\hat{\theta}$ solves the score equation. From the last equation, we have

$$\frac{1}{\sqrt{n}} \frac{\partial}{\partial \theta} \ln L(\theta_0) = \sqrt{n} (\hat{\theta} - \theta_0) \left\{ -\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}) \right\}. \quad (10.2)$$

Combining Equations (10.1) and (10.2), we have

$$\ln L(\hat{\theta}) = \ln L(\theta_0) + \sqrt{n} (\hat{\theta} - \theta_0) \sqrt{n} (\hat{\theta} - \theta_0) \left\{ -\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}) \right\}$$

$$+ \frac{n}{2} (\hat{\theta} - \theta_0)^2 \frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}).$$
so that
\[ \ln L(\hat{\theta}) - \ln L(\theta_0) = n(\hat{\theta} - \theta_0)^2 \left\{ \frac{-1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\theta_*) \right\} + \frac{n}{2} (\hat{\theta} - \theta_0)^2 \left\{ \frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\theta_*) \right\} . \] (10.3)

Because \( \hat{\theta} \) is consistent (and because \( H_0 \) is true), we know that \( \hat{\theta} \xrightarrow{p} \theta_0 \), as \( n \to \infty \). Therefore, because \( \hat{\theta}_* \) and \( \hat{\theta}_{**} \) are both trapped between \( \hat{\theta} \) and \( \theta_0 \), both terms in the brackets, i.e.,
\[ \frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}_{**}) \quad \text{and} \quad \frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ln L(\hat{\theta}_*) \]
converge in probability to
\[ E_{\theta_0} \left[ \frac{\partial^2}{\partial \theta^2} \ln f_X(X|\theta) \right] = -I_1(\theta_0), \]
by the WLLN. Therefore, the RHS of Equation (10.3) will behave in the limit the same as
\[ \frac{n}{2} (\hat{\theta} - \theta_0)^2 I_1(\theta_0) = \frac{1}{2} \sqrt{n}(\hat{\theta} - \theta_0) \sqrt{n}(\hat{\theta} - \theta_0) I_1(\theta_0) \]
\[ = \frac{1}{2} \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{I_1(\theta_0)}} \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sqrt{I_1(\theta_0)}} \xrightarrow{d} \frac{1}{2} \chi^2, \]
by continuity. Therefore, when \( H_0 : \theta = \theta_0 \) is true,
\[-2 \ln \lambda(X) = -2[\ln L(\theta_0) - \ln L(\hat{\theta})] \xrightarrow{d} \chi^2. \]

Example 10.6. Suppose \( X_1, X_2, \ldots, X_n \) are iid Bernoulli(\( p \)), where \( 0 < p < 1 \). Derive the large sample LRT test of
\[ H_0 : p = p_0 \quad \text{versus} \quad H_1 : p \neq p_0. \]

Solution. The likelihood ratio statistic is
\[ \lambda(x) = \frac{L(p_0|x)}{L(\hat{p}|x)} = \frac{\sum_{i=1}^{n} x_i (1 - p_0)^{n - \sum_{i=1}^{n} x_i}}{\hat{p}^{\sum_{i=1}^{n} x_i} (1 - \hat{p})^{n - \sum_{i=1}^{n} x_i}} = \left( \frac{p_0}{\hat{p}} \right)^{\sum_{i=1}^{n} x_i} \left( \frac{1 - p_0}{1 - \hat{p}} \right)^{n - \sum_{i=1}^{n} x_i}. \]

Therefore,
\[ -2 \ln \lambda(X) = -2 \left[ \sum_{i=1}^{n} X_i \ln \left( \frac{p_0}{\hat{p}} \right) + \left( n - \sum_{i=1}^{n} X_i \right) \ln \left( \frac{1 - p_0}{1 - \hat{p}} \right) \right] \]
\[ = -2 \left[ n\hat{p} \ln \left( \frac{p_0}{\hat{p}} \right) + n(1 - \hat{p}) \ln \left( \frac{1 - p_0}{1 - \hat{p}} \right) \right]. \]

An approximate size \( \alpha \) rejection region is
\[ R = \{ x \in X : -2 \ln \lambda(x) \geq \chi^2_{1,\alpha} \}. \]
Monte Carlo Simulation: When $X_1, X_2, ..., X_n$ are iid Bernoulli($p$), where $0 < p < 1$, we have derived the Wald, score, and large sample LRT for testing $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Each test is a large sample test, so the size of each one is approximately equal to $\alpha$ when $n$ is large. We now perform a simulation to assess finite sample characteristics.

- Take $n = 20, n = 50, n = 100$
- Let $p_0 = 0.1$ and $p = 0.3$
- At each configuration of $n$ and $p_0$, we will
  - simulate $B = 10000$ Bernoulli($p_0$) samples (i.e., $H_0$ is true)
  - calculate $z^W_n$, $z^S_n$, and $-2 \ln \lambda(x)$ with each sample
  - record the percentage of times that $H_0$ is (incorrectly) rejected when $\alpha = 0.05$
  - this percentage is an estimate of the true size of the test (for a given configuration of $n$ and $p_0$).

Here are the results:

<table>
<thead>
<tr>
<th></th>
<th>Wald</th>
<th>Score</th>
<th>LRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 20$</td>
<td>0.1204</td>
<td>0.0441</td>
<td>0.1287</td>
</tr>
<tr>
<td>$p_0 = 0.1$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 50$</td>
<td>0.1189</td>
<td>0.0316</td>
<td>0.0627</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0.0716</td>
<td>0.0682</td>
<td>0.0456</td>
</tr>
<tr>
<td>$n = 20$</td>
<td>0.0538</td>
<td>0.0243</td>
<td>0.0538</td>
</tr>
<tr>
<td>$p_0 = 0.3$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$n = 50$</td>
<td>0.0646</td>
<td>0.0447</td>
<td>0.0447</td>
</tr>
<tr>
<td>$n = 100$</td>
<td>0.0506</td>
<td>0.0637</td>
<td>0.0506</td>
</tr>
</tbody>
</table>

Table 10.1: Monte Carlo simulation. Size estimates of nominal $\alpha = 0.05$ Wald, score, and LRTs for a binomial proportion $p$ when $n = 20, 50, 100$ and $p_0 = 0.1, 0.3$.

Important: Note that these sizes are really estimates of the true sizes (at each setting of $n$ and $p_0$). Therefore, we should acknowledge that these are estimates and report the margin of error associated with them.

- Because these are nominal size 0.05 tests, the margin of error associated with each “estimate,” assuming a 99 percent confidence level, is equal to

  $$B = 2.58 \sqrt{\frac{0.05(1 - 0.05)}{10000}} \approx 0.0056.$$ 

- Size estimates between 0.0444 and 0.0556 indicate that the test is operating at the nominal level. I have bolded the estimates in Table 10.1 that are within these bounds.
- Values <0.0444 suggest conservatism (the test rejects too often). Values >0.0556 suggest anti-conservatism (the test is not rejecting often enough).
Summary: Suppose $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. Assume that the regularity conditions needed for MLEs to be consistent and asymptotically normal (CAN) hold. We have presented three large sample procedures to test

$$H_0 : \theta = \theta_0$$

versus

$$H_1 : \theta \neq \theta_0.$$ 

- **Wald:**
  
  $$Z_n^W = \frac{\hat{\theta} - \theta_0}{\sqrt{v(\hat{\theta})/n}} = \frac{\hat{\theta} - \theta_0}{\sqrt{1/n I_1(\theta)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

- **Score:**
  
  $$Z_n^S = \frac{S(\theta_0|X)}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} \mathcal{N}(0, 1)$$

- **LRT:**
  
  $$-2 \ln \lambda(X) = -2[\ln L(\theta_0|X) - \ln L(\hat{\theta}|X)] \xrightarrow{d} \chi_1^2.$$ 

All convergence results are under $H_0 : \theta = \theta_0$.

- Note that $(Z_n^W)^2$, $(Z_n^S)^2$, and $-2 \ln \lambda(X)$ each converge in distribution to a $\chi_1^2$ distribution as $n \to \infty$.

- In terms of power (i.e., rejecting $H_0$ when $H_1$ is true), all three testing procedures are **asymptotically equivalent** when examining certain types of alternative sequences (i.e., Pitman sequences of alternatives). For these alternative sequences, $(Z_n^W)^2$, $(Z_n^S)^2$, and $-2 \ln \lambda(X)$ each converge to the same (noncentral) $\chi_1^2(\lambda)$ distribution. However, the powers may be quite different in finite samples.

**Remark:** The large sample LRT procedure can be easily generalized to multi-parameter hypotheses.

**Theorem 10.3.3.** Suppose $X_1, X_2, \ldots, X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^k$. Assume that the regularity conditions needed for MLEs to be CAN hold. Consider testing

$$H_0 : \theta \in \Theta_0$$

versus

$$H_1 : \theta \in \Theta_0^c$$

and define

$$\lambda(x) = \sup_{\theta \in \Theta_0} \frac{L(\theta|x)}{\sup_{\theta \in \Theta} L(\theta|x)} = \frac{L(\hat{\theta}_0|x)}{L(\hat{\theta}|x)}.$$
If $\theta \in \Theta_0$, then
\[ -2 \ln \lambda(X) = -2[\ln L(\hat{\theta}_0|X) - \ln L(\hat{\theta}|X)] \xrightarrow{d} \chi^2_{\nu}, \]
where $\nu = \dim(\Theta) - \dim(\Theta_0)$, the number of “free parameters” between $\Theta$ and $\Theta_0$.

**Implication:** Rejecting $H_0 : \theta \in \Theta_0$ when $\lambda(x)$ is small is equivalent to rejecting $H_0$ when $-2 \ln \lambda(x)$ is large. Therefore,
\[ R = \{ x \in X : -2 \ln \lambda(x) \geq \chi^2_{\nu,\alpha} \} \]
is an approximate size $\alpha$ rejection region. This means
\[ \lim_{n \to \infty} P_0(\text{Reject } H_0) = \alpha \text{ for all } \theta \in \Theta_0. \]

**Example 10.7.** McCann and Tebbs (2009) summarize a study examining perceived unmet need for dental health care for people with HIV infection. Baseline in-person interviews were conducted with 2,864 HIV infected individuals (aged 18 years and older) as part of the HIV Cost and Services Utilization Study. Define
\[
\begin{align*}
X_1 &= \text{number of patients with private insurance} \\
X_2 &= \text{number of patients with medicare and private insurance} \\
X_3 &= \text{number of patients without insurance} \\
X_4 &= \text{number of patients with medicare but no private insurance}.
\end{align*}
\]
Set $X = (X_1, X_2, X_3, X_4)$ and model $X \sim \text{mult}(2864, p_1, p_2, p_3, p_4; \sum_{i=1}^4 p_i = 1)$. Under this assumption, consider testing
\[
H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4} \quad \text{versus} \quad H_1 : H_0 \text{ not true.}
\]
The null parameter space is
\[ \Theta_0 = \{ \theta = (p_1, p_2, p_3, p_4) : p_1 = p_2 = p_3 = p_4 = 1/4 \}, \]
the singleton $(1/4, 1/4, 1/4, 1/4)$. The entire parameter space is
\[ \Theta = \left\{ \theta = (p_1, p_2, p_3, p_4) : 0 < p_1 < 1, 0 < p_2 < 1, 0 < p_3 < 1, 0 < p_4 < 1; \sum_{i=1}^4 p_i = 1 \right\}, \]
a simplex in $\mathbb{R}^4$. The number of “free parameters” is $\nu = \dim(\Theta) - \dim(\Theta_0) = 3 - 0 = 3$.
The observed data from the study are summarized by
\[ x = (658, 839, 811, 556). \]
The likelihood function is
\[ L(\theta|x) = L(p_1, p_2, p_3, p_4|x) = \frac{2864!}{x_1! x_2! x_3! x_4!} p_1^{x_1} p_2^{x_2} p_3^{x_3} p_4^{x_4}. \]
Maximizing $L(p_1, p_2, p_3, p_4 | x)$ over $\Theta$, noting that $p_4 = 1 - p_1 - p_2 - p_3$, gives the (unrestricted) maximum likelihood estimates

$$\hat{p}_1 = \frac{x_1}{2864}, \quad \hat{p}_2 = \frac{x_2}{2864}, \quad \hat{p}_3 = \frac{x_3}{2864}, \quad \hat{p}_4 = \frac{x_4}{2864}.$$ 

Therefore,

$$\lambda(x) = \lambda(x_1, x_2, x_3, x_4) = \frac{L(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})}{\hat{L}(\hat{p}_1, \hat{p}_2, \hat{p}_3, \hat{p}_4)} = \frac{\frac{2864!}{x_1! x_2! x_3! x_4!} \left( \frac{1}{4} \right)^x_1 \left( \frac{1}{4} \right)^x_2 \left( \frac{1}{4} \right)^x_3 \left( \frac{1}{4} \right)^x_4}{\frac{2864!}{x_1! x_2! x_3! x_4!} \left( \frac{x_1}{2864} \right)^x_1 \left( \frac{x_2}{2864} \right)^x_2 \left( \frac{x_3}{2864} \right)^x_3 \left( \frac{x_4}{2864} \right)^x_4} = \prod_{i=1}^{4} \left( \frac{2864}{4x_i} \right)^{x_i}.$$ 

The large sample LRT statistic is

$$-2 \ln \lambda(x) = -2 \sum_{i=1}^{4} x_i \ln \left( \frac{2864}{4x_i} \right) \approx 75.69.$$ 

An approximate size $\alpha = 0.05$ rejection region is

$$R = \{ x \in X : -2 \ln \lambda(x) \geq 7.81 \}.$$ 

Therefore, we have very strong evidence against $H_0$.

> qchisq(0.95, 3)

[1] 7.814728

### 10.4 Confidence Intervals

**Remark:** In Chapter 9 (CB), we discussed methods to derive confidence intervals based on exact (i.e., finite sample) distributions. We now present three large sample approaches:

1. Wald
2. Score
3. Likelihood ratio.

These are known as the “large sample likelihood based confidence intervals.”

**Definition:** Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. The random variable 

$$Q_n = Q_n(X, \theta)$$

is called a **large sample pivot** if its asymptotic distribution is free of all unknown parameters. If $Q_n$ is a large sample pivot and if

$$P_\theta(Q_n(X, \theta) \in A) \approx 1 - \alpha,$$

then $C(X) = \{ \theta : Q_n(X, \theta) \in A \}$ is called an **approximate** $1 - \alpha$ confidence set for $\theta$. 
10.4.1 Wald intervals

Recall: Suppose $X_1, X_2, ..., X_n$ are iid from $f_X(x|\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}$. As long as suitable regularity conditions hold, we know that an MLE $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, v(\theta)),$$

where

$$v(\theta) = \frac{1}{I_1(\theta)}.$$

If $v(\theta)$ is a continuous function of $\theta$, then $v(\hat{\theta}) \xrightarrow{p} v(\theta)$, for all $\theta$; i.e., $v(\hat{\theta})$ is a consistent estimator of $v(\theta)$, and

$$Q_n(X, \theta) = \frac{\hat{\theta} - \theta}{\sqrt{\frac{v(\hat{\theta})}{n}}},$$

by Slutsky’s Theorem. Therefore, $Q_n(X, \theta)$ is a large sample pivot and

$$1 - \alpha \approx P_{\theta}(-z_{\alpha/2} \leq Q_n(X, \theta) \leq z_{\alpha/2})$$

$$= P_{\theta}(-z_{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{\frac{v(\hat{\theta})}{n}}} \leq z_{\alpha/2})$$

$$= P_{\theta}(\hat{\theta} - z_{\alpha/2} \sqrt{\frac{v(\hat{\theta})}{n}} \leq \theta \leq \hat{\theta} + z_{\alpha/2} \sqrt{\frac{v(\hat{\theta})}{n}}).$$

Therefore,

$$\hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{v(\hat{\theta})}{n}}$$

is an approximate $1 - \alpha$ confidence interval for $\theta$.

Remark: We could have arrived at this same interval by inverting the large sample test of

$$H_0 : \theta = \theta_0$$
versus
$$H_1 : \theta \neq \theta_0$$

that uses the (Wald) test statistic

$$Z_n^W = \frac{\hat{\theta} - \theta_0}{\sqrt{\frac{v(\hat{\theta})}{n}}}$$

and rejection region

$$R = \{x \in \mathcal{X} : |Z_n^W| \geq z_{\alpha/2}\}.$$

This is why this type of large sample interval is called a Wald confidence interval (it is the interval that arises from inverting a large sample Wald test).
Extension: We can also write large sample Wald confidence intervals for functions of \( \theta \) using the Delta Method. Recall that if \( g : \mathbb{R} \to \mathbb{R} \) is differentiable at \( \theta \) and \( g'(\theta) \neq 0 \), then
\[
\sqrt{n}[g(\hat{\theta}) - g(\theta)] \xrightarrow{d} \mathcal{N}(0, [g'(\theta)]^2 v(\theta)).
\]
If \([g'(\theta)]^2 v(\theta)\) is a continuous function of \( \theta \), then we can find a consistent estimator for it, namely \([g'(\hat{\theta})]^2 v(\hat{\theta})\), because MLEs are consistent themselves and consistency is preserved under continuous mappings. Therefore,
\[
Q_n(X, \theta) = \frac{g(\hat{\theta}) - g(\theta)}{\sqrt{\frac{[g'(\hat{\theta})]^2 v(\hat{\theta})}{n}}} \xrightarrow{d} \mathcal{N}(0, 1),
\]
by Slutsky’s Theorem and
\[
g(\hat{\theta}) \pm z_{\alpha/2} \sqrt{\frac{[g'(\hat{\theta})]^2 v(\hat{\theta})}{n}}
\]
is an approximate \( 1 - \alpha \) confidence interval for \( g(\theta) \).

**Example 10.8.** Suppose \( X_1, X_2, \ldots, X_n \) are iid Bernoulli\((p)\), where \( 0 < p < 1 \).

(a) Derive a \( 1 - \alpha \) (large sample) Wald confidence interval for \( p \).

(b) Derive a \( 1 - \alpha \) (large sample) Wald confidence interval for
\[
g(p) = \ln\left(\frac{p}{1 - p}\right),
\]
the log odds of \( p \).

**Solution.** (a) We already know that the MLE of \( p \) is given by
\[
\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i.
\]
In Example 10.4, we showed that
\[
v(p) = \frac{1}{I_1(p)} = p(1 - p).
\]
Therefore,
\[
\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]
is an approximate \( 1 - \alpha \) Wald confidence interval for \( p \). The problems with this interval (i.e., in conferring the nominal coverage probability) are well known; see Brown et al. (2001).

(b) Note that \( g(p) = \ln[p/(1 - p)] \) is a differentiable function and
\[
g'(p) = \frac{1}{p(1 - p)} \neq 0.
\]
The Delta Method gives

\[
\sqrt{n} \left[ \ln \left( \frac{\hat{p}}{1 - \hat{p}} \right) - \ln \left( \frac{p}{1 - p} \right) \right] \xrightarrow{d} \mathcal{N} \left( 0, \left[ \frac{1}{p(1 - p)} \right]^2 p(1 - p) \right)
\]

Because the asymptotic variance \(1/p(1 - p)\) can be consistently estimated by \(1/\hat{p}(1 - \hat{p})\), we have

\[
\frac{\ln \left( \frac{\hat{p}}{1 - \hat{p}} \right) - \ln \left( \frac{p}{1 - p} \right)}{\sqrt{\frac{1}{n\hat{p}(1 - \hat{p})}}} \xrightarrow{d} \mathcal{N}(0, 1)
\]

by Slutsky’s Theorem, and

\[
\ln \left( \frac{\hat{p}}{1 - \hat{p}} \right) \pm \frac{z_{\alpha/2}}{\sqrt{\frac{1}{n\hat{p}(1 - \hat{p})}}}
\]

is an approximate \(1 - \alpha\) Wald confidence interval for \(g(p) = \ln[p/(1 - p)]\).

**Remarks:** As you can see, constructing (large sample) Wald confidence intervals is straightforward. We rely on the MLE being consistent and asymptotically normal (CAN) and also on being able to find a consistent estimator of the asymptotic variance of the MLE.

- More generally, if you have an estimator \(\hat{\theta}\) (not necessarily an MLE) that is asymptotically normal and if you can estimate its (large sample) variance consistently, you can do Wald inference. This general strategy for large sample inference is ubiquitous in statistical research.

- The problem, of course, is that because large sample standard errors must be estimated, the performance of Wald confidence intervals (and tests) can be poor in small samples. Brown et al. (2001) highlights this for the binomial proportion; however, this behavior is seen in other settings.

- I view Wald inference as a “fall back.” It is what to do when no other large sample inference procedures are available; i.e., “having something is better than nothing.” Of course, in very large sample settings (e.g., large scale Phase III clinical trials, public health studies with thousands of individuals, etc.), Wald inference is usually the default approach (probably because of its simplicity) and is generally satisfactory.

### 10.4.2 Score intervals

Recall: Suppose \(X_1, X_2, ..., X_n\) are iid from \(f_X(x|\theta)\), where \(\theta \in \Theta \subseteq \mathbb{R}\). We have shown previously that

\[
Q_n(X, \theta) = \frac{S(\theta|X)}{\sqrt{I_n(\theta)}} \xrightarrow{d} \mathcal{N}(0, 1),
\]

where \(I_n(\theta) = nI_1(\theta)\) is the Fisher information based on the sample.
Motivation: Score confidence intervals arise from inverting (large sample) score tests. Recall that in testing

\[ H_0 : \theta = \theta_0 \]

versus

\[ H_1 : \theta \neq \theta_0, \]

the score statistic

\[ Q_n(X, \theta_0) = \frac{S(\theta_0|X)}{\sqrt{I_n(\theta_0)}} \xrightarrow{d} N(0, 1) \]

when \( H_0 \) is true. Therefore,

\[ R = \{ x \in X : |Q_n(x, \theta_0)| \geq z_{\alpha/2} \} \]

is an approximate size \( \alpha \) rejection region for testing \( H_0 \) versus \( H_1 \). The acceptance region is

\[ A = R^c = \{ x \in X : |Q_n(x, \theta_0)| < z_{\alpha/2} \}. \]

From inverting this acceptance region, we can conclude that

\[ C(x) = \{ \theta : Q_n(x, \theta) < z_{\alpha/2} \} \]

is an approximate \( 1 - \alpha \) confidence set for \( \theta \). If \( C(x) \) is an interval, then we call it a score confidence interval.

Example 10.9. Suppose \( X_1, X_2, ..., X_n \) are iid Bernoulli\((p)\), where \( 0 < p < 1 \). Derive a \( 1 - \alpha \) (large sample) score confidence interval for \( p \).

Solution. From Example 10.5, we have

\[ Q_n(X, p) = \frac{S(p|X)}{\sqrt{I_n(p)}} = \frac{\sum_{i=1}^{n} X_i}{p} - \frac{n - \sum_{i=1}^{n} X_i}{1 - p} = \frac{\hat{p} - p}{\sqrt{\frac{n}{p(1 - p)}}}. \]

From our discussion above, the (random) set

\[ C(X) = \{ p : Q_n(X, p) < z_{\alpha/2} \} = \left\{ p : \frac{\hat{p} - p}{\sqrt{\frac{p(1 - p)}{n}}} < z_{\alpha/2} \right\} \]

forms the score interval for \( p \). After observing \( X = x \), this interval could be calculated numerically (e.g., using a grid search over values of \( p \) that satisfy this inequality). However, in the binomial case, we can get closed-form expressions for the endpoints. To see why, note that the boundary

\[ Q_n(x, p) = z_{\alpha/2} \iff (\hat{p} - p)^2 = z_{\alpha/2}^2 \frac{p(1 - p)}{n}. \]

After algebra, this equation becomes

\[ \left(1 + \frac{z_{\alpha/2}^2}{n}\right)p^2 - \left(2\hat{p} + \frac{z_{\alpha/2}^2}{n}\right)p + \hat{p}^2 = 0. \]
The LHS of the last equation is a quadratic function of \( p \). The roots of this equation, if they are real, delimit the score interval for \( p \). Using the quadratic formula, the lower and upper limits are

\[
 p_L = \frac{(2\bar{p} + z_{\alpha/2}^2/n) - \sqrt{(2\bar{p} + z_{\alpha/2}^2/n)^2 - 4(1 + z_{\alpha/2}^2/n)\bar{p}^2}}{2(1 + z_{\alpha/2}^2/n)}
\]

\[
 p_U = \frac{(2\bar{p} + z_{\alpha/2}^2/n) + \sqrt{(2\bar{p} + z_{\alpha/2}^2/n)^2 - 4(1 + z_{\alpha/2}^2/n)\bar{p}^2}}{2(1 + z_{\alpha/2}^2/n)}.
\]

respectively. Note that the score interval is much more complex than the Wald interval. However, the score interval (in this setting and elsewhere) typically confers very good coverage probability, that is, close to the nominal \( 1 - \alpha \) level, even for small samples. Therefore, although we have added complexity, the score interval is typically much better.

### 10.4.3 Likelihood ratio intervals

**Recall:** Suppose \( X_1, X_2, ..., X_n \) are iid from \( f_X(x|\theta) \), where \( \theta \in \Theta \subseteq \mathbb{R} \). Consider testing

\[
 H_0 : \theta = \theta_0 \\
 versus \\
 H_1 : \theta \neq \theta_0.
\]

The LRT statistic is

\[
 \lambda(x) = \frac{L(\theta_0|x)}{L(\hat{\theta}|x)}
\]

and

\[
 R = \{ x \in X : -2 \ln \lambda(x) \geq \chi^2_{1,\alpha} \}
\]

is an approximate size \( \alpha \) rejection region for testing \( H_0 \) versus \( H_1 \). Inverting the acceptance region,

\[
 C(x) = \left\{ \theta : -2 \ln \left[ \frac{L(\theta|x)}{L(\hat{\theta}|x)} \right] < \chi^2_{1,\alpha} \right\}
\]

is an approximate \( 1 - \alpha \) confidence set for \( \theta \). If \( C(x) \) is an interval, then we call it a **likelihood ratio confidence interval**.

**Example 10.10.** Suppose \( X_1, X_2, ..., X_n \) are iid Bernoulli(\( p \)), where \( 0 < p < 1 \). Derive a \( 1 - \alpha \) (large sample) likelihood ratio confidence interval for \( p \).

**Solution.** From Example 10.6, we have

\[
 -2 \ln \left[ \frac{L(p|x)}{L(\bar{p}|x)} \right] = -2 \left[ n\bar{p} \ln \left( \frac{p}{\bar{p}} \right) + n(1 - \bar{p}) \ln \left( \frac{1 - p}{1 - \bar{p}} \right) \right].
\]

Therefore, the confidence interval is

\[
 C(x) = \left\{ p : -2 \left[ n\bar{p} \ln \left( \frac{p}{\bar{p}} \right) + n(1 - \bar{p}) \ln \left( \frac{1 - p}{1 - \bar{p}} \right) \right] < \chi^2_{1,\alpha} \right\}.
\]

This interval must be calculated using numerical search methods.