## Contents

### Acknowledgements

1

### 1 Axioms of Probability

1.1 $\sigma$-Fields ................................................... 1
1.2 Manipulating Probability ........................................ 5
1.3 Combinatorics and Counting ................................. 12
1.4 Conditional Probability and Independence ............... 18

### 2 Random Variables

2.1 Measurable Functions ........................................... 25
2.2 Probability Distribution of a Random Variable .......... 29
2.3 Discrete and Continuous Random Variables .......... 33
2.4 Transformations of Random Variables .................... 39

### 3 Moments and Generating Functions

3.1 Expectation ..................................................... 46
3.2 Moment Generating Functions ............................... 54
3.3 Complex–Valued Random Variables and Characteristic Functions ................. 61
3.4 Probability Generating Functions .......................... 72
3.5 Moment Inequalities ........................................... 75

### 4 Random Vectors

4.1 Joint, Marginal, and Conditional Distributions .......... 78
4.2 Independent Random Variables ............................ 85
4.3 Functions of Random Vectors ............................... 90
4.4 Order Statistics ............................................... 97
4.5 Multivariate Expectation .................................... 100
4.6 Multivariate Generating Functions ....................... 106
4.7 Conditional Expectation ..................................... 112
4.8 Inequalities and Identities .................................. 116
5 Particular Distributions 122

5.1 Multivariate Normal Distributions . . . . . . . . . . . . . . . . . . . . . . . . . 122
5.2 Exponential Family of Distributions . . . . . . . . . . . . . . . . . . . . . . . . . 129

Index 132
Acknowledgements

I would like to thank my students, Hanadi B. Eltahir, Rich Madsen, and Bill Morphet, who helped during the Fall 1999 and Spring 2000 semesters in typesetting these lecture notes using \LaTeX and for their suggestions how to improve some of the material presented in class. Thanks are also due to more than 50 students who took Stat 6710/20 with me since the Fall 2000 semester for their valuable comments that helped to improve and correct these lecture notes.

In addition, I particularly would like to thank Mike Minnotte and Dan Coster, who previously taught this course at Utah State University, for providing me with their lecture notes and other materials related to this course. Their lecture notes, combined with additional material from Casella/Berger (2002), Rohatgi (1976) and other sources listed below, form the basis of the script presented here.

The primary textbook required for this class is:


A Web page dedicated to this class is accessible at:


This course closely follows Casella and Berger (2002) as described in the syllabus. Additional material originates from the lectures from Professors Hering, Trenkler, and Gather I have attended while studying at the Universität Dortmund, Germany, the collection of Masters and PhD Preliminary Exam questions from Iowa State University, Ames, Iowa, and the following textbooks:


Additional definitions, integrals, sums, etc. originate from the following formula collections:


Jürgen Symanzik, August 28, 2010.
1 Axioms of Probability

(Based on Casella/Berger, Sections 1.1, 1.2 & 1.3)

1.1 σ-Fields

Let Ω be the sample space of all possible outcomes of a chance experiment. Let ω ∈ Ω (or x ∈ Ω) be any outcome.

Example:
Count # of heads in n coin tosses. Ω = {0, 1, 2, . . . , n}.

Any subset A of Ω is called an event.

For each event A ⊆ Ω, we would like to assign a number (i.e., a probability). Unfortunately, we cannot always do this for every subset of Ω.

Instead, we consider classes of subsets of Ω called fields and σ-fields.

Definition 1.1.1:
A class L of subsets of Ω is called a field if Ω ∈ L and L is closed under complements and finite unions, i.e., L satisfies

(i) Ω ∈ L
(ii) A ∈ L ⇒ A^C ∈ L
(iii) A, B ∈ L ⇒ A ∪ B ∈ L

Since Ω^C = ∅, (i) and (ii) imply ∅ ∈ L. Therefore, (i)': ∅ ∈ L [can replace (i)].

Note: De Morgan’s Laws
For any class A of sets, and sets A ∈ A, it holds:

\[ \bigcup_{A \in A} A = \left( \bigcap_{A \in A} A^C \right)^C \quad \text{and} \quad \bigcap_{A \in A} A = \left( \bigcup_{A \in A} A^C \right)^C. \]

Note:
So (ii), (iii) imply (iii)': A, B ∈ L ⇒ A ∩ B ∈ L [can replace (iii)].
Proof:

\[ A, B \in L \implies A^C, B^C \in L \implies (A^C \cup B^C) \in L \implies (A^C \cup B^C)^C \in L \implies A \cap B \in L \]

\[ \Box \]

Definition 1.1.2:
A class \( L \) of subsets of \( \Omega \) is called a \( \sigma \)–field (Borel field, \( \sigma \)–algebra) if it is a field and closed under countable unions, i.e.,

(iv) \( \{ A_n \}_{n=1}^{\infty} \in L \implies \bigcup_{n=1}^{\infty} A_n \in L. \)

Note:
(iv) implies (iii) by taking \( A_n = \emptyset \) for \( n \geq 3. \)

Example 1.1.3:
For some \( \Omega \), let \( L \) contain all finite and all cofinite sets (\( A \) is cofinite if \( A^C \) is finite — for example, if \( \Omega = \mathbb{N}, A = \{ x \mid x \geq c \} \) is not finite but since \( A^C = \{ x \mid x < c \} \) is finite, \( A \) is cofinite). Then \( L \) is a field. But \( L \) is a \( \sigma \)–field iff (if and only if) \( \Omega \) is finite.

For example, let \( \Omega = \mathbb{Z} \). Take \( A_n = \{ n \} \), each finite, so \( A_n \in L \). But \( \bigcup_{n=1}^{\infty} A_n = \mathbb{Z}^+ \notin L \), since the set is not finite (it is infinite) and also not cofinite \( \left( \bigcup_{n=1}^{\infty} A_n \right)^C = \mathbb{Z}_0^- \) is infinite, too).

Question: Does this construction work for \( \Omega = \mathbb{Z}^+ \)?

If we take \( A_n = \{ n \} \), then \( \left( \bigcup_{n=1}^{\infty} A_n \right)^C = \emptyset \in L. \) But, if we take \( A_n = \{ 2n \} \), then \( \left( \bigcup_{n=1}^{\infty} A_n \right) = \{ 2, 4, 6, \ldots \} \notin L \) and \( \left( \bigcup_{n=1}^{\infty} A_n \right)^C = \{ 1, 3, 5, \ldots \} \notin L. \)

Note:
The largest \( \sigma \)–field in \( \Omega \) is the power set \( \mathcal{P}(\Omega) \) of all subsets of \( \Omega \). The smallest \( \sigma \)–field is \( L = \{ \emptyset, \Omega \}. \)

Terminology:
A set \( A \in L \) is said to be “measurable \( L \)”.

Note:
We often begin with a class of sets, say \( a \), which may not be a field or a \( \sigma \)–field.
Definition 1.1.4:
The \( \sigma \)-field generated by \( a \), \( \sigma(a) \), is the smallest \( \sigma \)-field containing \( a \), or the intersection of all \( \sigma \)-fields containing \( a \).

Note:
(i) Such \( \sigma \)-fields containing \( a \) always exist (e.g., \( \mathcal{P}(\Omega) \)), and (ii) the intersection of an arbitrary \# of \( \sigma \)-fields is always a \( \sigma \)-field.

Proof:
(ii) Suppose \( L = \bigcap_\theta L_\theta \). We have to show that conditions (i) and (ii) of Def. 1.1.1 and (iv) of Def. 1.1.2 are fulfilled:
(i) \( \Omega \in L_\theta \ \forall \theta \implies \Omega \in L \)
(ii) Let \( A \in L \ \implies A \in L_\theta \ \forall \theta \implies A^C \in L_\theta \ \forall \theta \implies A^C \in L \)
(iv) Let \( A_n \in L \ \forall n \implies A_n \in L_\theta \ \forall \theta \implies \bigcup_n A_n \in L \)

Example 1.1.5:
\( \Omega = \{0, 1, 2, 3\} \), \( a = \{\{0\}\} \), \( b = \{\{0\}, \{0, 1\}\} \).

What is \( \sigma(a) \)?
\( \sigma(a) \): must include \( \Omega, \emptyset, \{0\} \)
also: \( \{1, 2, 3\} \) by 1.1.1 (ii)
Since all unions are included, we have \( \sigma(a) = \{\Omega, \emptyset, \{0\}, \{1, 2, 3\}\} \)

What is \( \sigma(b) \)?
\( \sigma(b) \): must include \( \Omega, \emptyset, \{0\}, \{0, 1\} \)
also: \( \{1, 2, 3\}, \{2, 3\} \) by 1.1.1 (ii)
\( \{0, 2, 3\} \) by 1.1.1 (iii)
\( \{1\} \) by 1.1.1 (ii)
Since all unions are included, we have \( \sigma(b) = \{\Omega, \emptyset, \{0\}, \{1\}, \{0, 1\}, \{2, 3\}, \{0, 2, 3\}, \{1, 2, 3\}\} \)

Note:
If \( \Omega \) is finite or countable, we will usually use \( L = \mathcal{P}(\Omega) \). If \( |\Omega| = n < \infty \), then \( |L| = 2^n \).
If \( \Omega \) is uncountable, \( \mathcal{P}(\Omega) \) may be too large to be useful and we may have to use some smaller \( \sigma \)-field.
Definition 1.1.6:
If $\Omega = \mathbb{R}$, an important special case is the **Borel $\sigma$-field**, i.e., the $\sigma$-field generated from all half-open intervals of the form $(a, b]$, denoted $\cal{B}$ or $\mathcal{B}_1$. The sets of $\cal{B}$ are called **Borel sets**.

The Borel $\sigma$-field on $\mathbb{R}^d$ ($\mathcal{B}_d$) is the $\sigma$-field generated by $d$-dimensional rectangles of the form \{(x_1, x_2, \ldots, x_d) \mid a_i < x_i \leq b_i; i = 1, 2, \ldots, d\}. \hfill \blacksquare

**Note:**
$\mathcal{B}$ contains all points: $\{x\} = \bigcap_{n=1}^{\infty} (x - \frac{1}{n}, x]$

- closed intervals: $[x, y] = (x, y] + \{x\} = (x, y] \cup \{x\}$
- open intervals: $(x, y) = (x, y] - \{y\} = (x, y] \cap \{y\}^C$
- and semi-infinite intervals: $(x, \infty) = \bigcup_{n=1}^{\infty} (x, x + n] \hfill \blacksquare$

**Note:**
We now have a **measurable space** $(\Omega, \mathcal{L})$. We next define a **probability measure** $P(\cdot)$ on $(\Omega, \mathcal{L})$ to obtain a **probability space** $(\Omega, \mathcal{L}, P)$. \hfill \blacksquare

**Definition 1.1.7: Kolmogorov Axioms of Probability**
A **probability measure** ($\text{pm}$), $P$, on $(\Omega, \mathcal{L})$ is a set function $P : \mathcal{L} \rightarrow \mathbb{R}$ satisfying

(i) $0 \leq P(A) \quad \forall A \in \mathcal{L}$

(ii) $P(\Omega) = 1$

(iii) If $\{A_n\}_{n=1}^{\infty}$ are disjoint sets in $\mathcal{L}$ and $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$, then $P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n). \hfill \blacksquare$

**Note:**
$\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$ holds automatically if $\mathcal{L}$ is a $\sigma$-field but it is needed as a precondition in the case that $\mathcal{L}$ is just a field. Property (iii) is called **countable additivity**. \hfill \blacksquare
1.2 Manipulating Probability

**Theorem 1.2.1:**
For $P$ a pm on $(\Omega, L)$, it holds:

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

(ii) $P(A^C) = 1 - P(A)$ $\forall A \in L$

(iii) $P(A) \leq 1$ $\forall A \in L$

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $\forall A, B \in L$

(v) If $A \subseteq B$, then $P(A) \leq P(B)$.

**Proof:**

(i) Let $A_n = \emptyset$ $\forall n \Rightarrow \bigcup_{n=1}^{\infty} A_n = \emptyset \in L$.

$A_i \cap A_j = \emptyset \cap \emptyset = \emptyset$ $\forall i, j \Rightarrow A_n$ are disjoint $\forall n$.

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

\[= \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P(\emptyset)\]  

This can only hold if $P(\emptyset) = 0$.

(ii) Let $A_1 = A, A_2 = A^C, A_n = \emptyset$ $\forall n \geq 3$.

$\Omega = \bigcup_{n=1}^{\infty} A_n = A_1 \cup A_2 \cup \bigcup_{n=3}^{\infty} A_n = A_1 \cup A_2 \cup \emptyset$.

$A_1 \cap A_2 = A_1 \cap \emptyset = A_2 \cap \emptyset = \emptyset \Rightarrow A_1, A_2, \emptyset$ are disjoint.

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

\[= \sum_{n=1}^{\infty} P(A_n)\]  

\[= P(A_1) + P(A_2) + \sum_{n=3}^{\infty} P(A_n)\]  

\[= P(A_1) + P(A_2)\]  

\[= P(A) + P(A^C)\]

$\Rightarrow P(A^C) = 1 - P(A)$ $\forall A \in L$. 

(iii) By Th. 1.2.1 (ii) $P(A) = 1 - P(A^C)$

\[ \Rightarrow P(A) \leq 1 \ \forall A \in L \text{ since } P(A^C) \geq 0 \text{ by Def. 1.1.7 (i)}. \]

(iv) $A \cup B = (A \cap B^C) \cup (A \cap B) \cup (B \cap A^C)$. So, $(A \cup B)$ can be written as a union of disjoint sets $(A \cap B^C), (A \cap B), (B \cap A^C)$.

\[ \Rightarrow P(A \cup B) = P((A \cap B^C) \cup (A \cap B) \cup (B \cap A^C)) \]

\[ = P(A \cap B^C) + P(A \cap B) + P(B \cap A^C) \] \[ = P(A \cap B) + P(B \cap A^C) + P(A \cap B) - P(A \cap B) \]

\[ = P(A) + P(B) - P(A \cap B) \]

(v) $B = (B \cap A^C) \cup A$ where $(B \cap A^C)$ and $A$ are disjoint sets.

\[ \Rightarrow P(B) = P((B \cap A^C) \cup A) \]

\[ = P(B \cap A^C) + P(A) \]

\[ \Rightarrow P(A) = P(B) - P(B \cap A^C) \]

\[ \Rightarrow P(A) \leq P(B) \text{ since } P(B \cap A^C) \geq 0 \text{ by Def. 1.1.7 (i)}. \]
**Theorem 1.2.2: Principle of Inclusion–Exclusion**

Let $A_1, A_2, \ldots, A_n \in L$. Then

\[ P(\bigcup_{k=1}^{n} A_k) = \sum_{k=1}^{n} P(A_k) - \sum_{k_1 < k_2}^{n} P(A_{k_1} \cap A_{k_2}) + \sum_{k_1 < k_2 < k_3}^{n} P(A_{k_1} \cap A_{k_2} \cap A_{k_3}) - \ldots + (-1)^{n+1} P(\bigcap_{k=1}^{n} A_k) \]

**Proof:**

$n = 1$ is trivial

$n = 2$ is Theorem 1.2.1 (iv)

Use induction for higher $n$ (Homework).

**Note:**

A proof by **induction** consists of two steps:

First, we have to establish the **induction base**. For example, if we state that something holds for all non–negative integers, then we have to show that it holds for $n = 0$. Similarly, if we state that something holds for all integers, then we have to show that it holds for $n = 1$. Formally, it is sufficient to verify a claim for the smallest valid integer only. However, to get some feeling how the proof from $n$ to $n + 1$ might work, it is sometimes beneficial to verify a claim for 1, 2, or 3 as well.

In the second step, we have to establish the result in the **induction step**, showing that something holds for $n + 1$, using the fact that it holds for $n$ (alternatively, we can show that it holds for $n$, using the fact that it holds for $n - 1$).

**Theorem 1.2.3: Bonferroni’s Inequality**

Let $A_1, A_2, \ldots, A_n \in L$. Then

\[ \sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) \leq P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i) \]

**Proof:**

**Right side:** $P(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$

**Induction Base:**

For $n = 1$, the right side evaluates to $P(A_1) \leq P(A_1)$, which is true.

Formally, the next step is not required. However, it does not harm to verify the claim for $n = 2$ as well. For $n = 2$, the right side evaluates to $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$.

$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$ since $P(A_1 \cap A_2) \geq 0$ by Def. 1.1.7 (i).
This holds since $P(F) \forall n$ by Th. 1.2.1 (iv).

**Induction Step:**
We assume the right side is true for $n$ and show that it is true for $n + 1$:

\[
P^n_{n+1}(\bigcup_i A_i) = P((\bigcup_i A_i) \cup A_{n+1})
\]

\[
\text{Th. 1.2.1(iv)} \Rightarrow P((\bigcup_i A_i) + P(A_{n+1}) - P((\bigcup_i A_i) \cap A_{n+1})
\]

\[
\text{Def 1.1.7(i)} \leq P((\bigcup_i A_i) + P(A_{n+1})
\]

\[
\text{I.B.} \leq \sum_{i=1}^n P(A_i) + P(A_{n+1})
\]

\[
= \sum_{i=1}^{n+1} P(A_i)
\]

**Left side:** $\sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) \leq P((\bigcup_i A_i)

**Induction Base:**
For $n = 1$, the left side evaluates to $P(A_1) \leq P(A_1)$, which is true.

For $n = 2$, the left side evaluates to $P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1 \cup A_2)$, which is true by Th. 1.2.1 (iv).

For $n = 3$, the left side evaluates to $P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \leq P(A_1 \cup A_2 \cup A_3)$.

This holds since

\[
P(A_1 \cup A_2 \cup A_3)
\]

\[
= P((A_1 \cup A_2) \cup A_3)
\]

\[
\text{Th. 1.2.1(iv)} \Rightarrow P(A_1 \cup A_2) + P(A_3) - P((A_1 \cup A_2) \cap A_3)
\]

\[
= P(A_1 \cup A_2) + P(A_3) - P((A_1 \cap A_3) \cup (A_2 \cap A_3))
\]

\[
\text{Th. 1.2.1(iv)} \Rightarrow P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) - P(A_1 \cap A_3) - P(A_2 \cap A_3)
\]

\[
+ P((A_1 \cap A_3) \cap (A_2 \cap A_3))
\]

\[
= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)
\]

\[
\text{Def 1.1.7(i)} \geq P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3)
\]

This establishes the induction base for the left side.
Induction Step:
We assume the left side is true for \( n \) and show that it is true for \( n + 1 \):

\[
P(\bigcup_{i=1}^{n+1} A_i) = P(\bigcup_{i=1}^{n} A_i \cup A_{n+1})
\]

\[
= P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) - P(\bigcup_{i=1}^{n} A_i \cap A_{n+1})
\]

*left I.B.*

\[
\geq \sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) + P(A_{n+1}) - P(\bigcup_{i=1}^{n} A_i \cap A_{n+1})
\]

*Th.1.2.3 right side*

\[
\geq \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) - \sum_{i=1}^{n} P(A_i \cap A_{n+1})
\]

\[
= \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n+1} P(A_i \cap A_j)
\]

Theorem 1.2.4: Boole’s Inequality

Let \( A, B \in L \). Then

(i) \( P(A \cap B) \geq P(A) + P(B) - 1 \)

(ii) \( P(A \cap B) \geq 1 - P(A^C) - P(B^C) \)

Proof:

Homework

Definition 1.2.5: Continuity of sets

For a sequence of sets \( \{A_n\}_{n=1}^{\infty} \), \( A_n \in L \) and \( A \in L \), we say

(i) \( A_n \uparrow A \) if \( A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots \) and \( A = \bigcup_{n=1}^{\infty} A_n \).

(ii) \( A_n \downarrow A \) if \( A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots \) and \( A = \bigcap_{n=1}^{\infty} A_n \).
Example:
$$A_n = \left[ 1, 2 - \frac{1}{n} \right] \uparrow [1, 2)$$
$$B_n = \left[ 1 - \frac{1}{n}, 2 \right) \downarrow [1, 2)$$

Theorem 1.2.6:
If \( \{A_n\}_{n=1}^\infty \) and \( A \in L \), then
\[
\lim_{n \to \infty} P(A_n) = P(A) \text{ if 1.2.5 (i) or 1.2.5 (ii) holds.}
\]

Proof:
Part (i): Assume that 1.2.5 (i) holds.
Let \( B_1 = A_1 \) and \( B_k = A_k - A_{k-1} = A_k \cap A_{k-1}^C \ \forall \ k \geq 2 \)

By construction, \( B_i \cap B_j = \emptyset \) for \( i \neq j \)

It is \( A = \bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n \)

and also \( A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i \)

\[
P(A) = P(\bigcup_{k=1}^{\infty} B_k) \overset{\text{Def. 1.1.7 (ii)}}{=} \sum_{k=1}^{\infty} P(B_k) = \lim_{n \to \infty} \left[ \sum_{k=1}^{n} P(B_k) \right]
\]

\[
\overset{\text{Def. 1.1.7 (iii)}}{=} \lim_{n \to \infty} \left[ P\left( \bigcup_{k=1}^{n} B_k \right) \right] = \lim_{n \to \infty} [P(\bigcup_{k=1}^{n} A_k)] = \lim_{n \to \infty} P(A_n)
\]

The last step is possible since \( A_n = \bigcup_{k=1}^{n} A_k \)

Part (ii): Assume that 1.2.5 (ii) holds.

Then, \( A_1^C \subseteq A_2^C \subseteq A_3^C \subseteq \ldots \) and \( A^C = \left( \bigcap_{n=1}^{\infty} A_n \right)^C \overset{\text{De Morgan}}{=} \bigcup_{n=1}^{\infty} A_n^C \)

\[
P(A^C) \overset{\text{By Part (i)}}{=} \lim_{n \to \infty} P(A_n^C)
\]

So \( 1 - P(A^C) = 1 - \lim_{n \to \infty} P(A_n^C) \)

\[
\implies P(A) = \lim_{n \to \infty} (1 - P(A_n^C)) \overset{\text{Th. 1.2.1 (ii)}}{=} \lim_{n \to \infty} P(A_n)
\]
Theorem 1.2.7:

(i) Countable unions of probability 0 sets have probability 0.

(ii) Countable intersections of probability 1 sets have probability 1.

Proof:

Part (i):
Let $\{A_n\}_{n=1}^{\infty} \in L$, $P(A_n) = 0\ \forall n$

\[
0 \leq P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} 0 = 0
\]

Therefore $P\left(\bigcup_{n=1}^{\infty} A_n\right) = 0$.

Part (ii):
Let $\{A_n\}_{n=1}^{\infty} \in L$, $P(A_n) = 1\ \forall n$

\[
\begin{align*}
\text{Th. 1.2.1 (ii)} & \Rightarrow P\left(A_n^C\right) = 0\ \forall n \quad \text{Th. 1.2.7 (i)} \Rightarrow P\left(\bigcup_{n=1}^{\infty} A_n^C\right) = 0 \Rightarrow P\left(\left(\bigcup_{n=1}^{\infty} A_n^C\right)^C\right) = 1 \\
\text{De Morgan} & \Rightarrow P\left(\bigcap_{n=1}^{\infty} A_n\right) = 1
\end{align*}
\]
1.3 Combinatorics and Counting

For now, we restrict ourselves to sample spaces containing a finite number of points.

Let \( \Omega = \{ \omega_1, \ldots, \omega_n \} \) and \( L = \mathcal{P}(\Omega) \). For any \( A \in L \), \( P(A) = \sum_{\omega_j \in A} P(\omega_j) \).

**Definition 1.3.1:**
We say the elements of \( \Omega \) are **equally likely** (or occur with uniform probability) if \( P(\omega_j) = \frac{1}{n} \forall j = 1, \ldots, n \).

**Note:**
If this is true, \( P(A) = \frac{\text{number } \omega_j \text{ in } A}{\text{number } \omega_j \text{ in } \Omega} \).
Therefore, to calculate such probabilities, we just need to be able to count elements accurately.

**Theorem 1.3.2: Fundamental Theorem of Counting**

If we wish to select one element (\( a_1 \)) out of \( n_1 \) choices, a second element (\( a_2 \)) out of \( n_2 \) choices, and so on for a total of \( k \) elements, there are

\[ n_1 \times n_2 \times \ldots \times n_k \]

ways to do it.

**Proof:** (By Induction)

**Induction Base:**
\( k = 1 \): trivial
\( k = 2 \): \( n_1 \) ways to choose \( a_1 \). For each, \( n_2 \) ways to choose \( a_2 \).
Total \# of ways = \( n_2 + n_2 + \ldots + n_2 = n_1 \times n_2 \), \( n_1 \) times

**Induction Step:**
Suppose it is true for \( k = (k - 1) + 1 \).

There are \( n_1 \times n_2 \times n_3 \times \ldots \times n_{k-1} \) ways to select one element (\( a_1 \)) out of \( n_1 \) choices, a second element (\( a_2 \)) out of \( n_2 \) choices, and so on, up to the \( (k - 1)^{th} \) element (\( a_{k-1} \)) out of \( n_{k-1} \) choices. For each of these \( n_1 \times n_2 \times n_3 \times \ldots \times n_{k-1} \) possible ways, we can select the \( k^{th} \) element (\( a_k \)) out of \( n_k \) choices. Thus, the total \# of ways = \( (n_1 \times n_2 \times n_3 \times \ldots \times n_{k-1}) \times n_k \).
**Definition 1.3.3:**
For positive integer \( n \), we define **n factorial** as \( n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 = n \times (n-1)! \) and \( 0! = 1 \).

**Definition 1.3.4:**
For nonnegative integers \( n \geq r \), we define the **binomial coefficient** (read as \( n \) choose \( r \)) as
\[
\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+1)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot r}.
\]

**Note:**
A useful extension for the binomial coefficient for \( n < r \) is
\[
\binom{n}{r} = \frac{n \cdot (n-1) \cdot \ldots \cdot 0 \cdot \ldots \cdot (n-r+1)}{1 \cdot 2 \cdot \ldots \cdot r} = 0.
\]

**Note:**
Most counting problems consist of drawing a fixed number of times from a set of elements (e.g., \{1, 2, 3, 4, 5, 6\}). To solve such problems, we need to know

(i) the size of the set, \( n \);

(ii) the size of the sample, \( r \);

(iii) whether the result will be **ordered** (i.e., is \{1, 2\} different from \{2, 1\}); and

(iv) whether the draws are **with replacement** (i.e, can results like \{1, 1\} occur?).

**Theorem 1.3.5:**
The number of ways to draw \( r \) elements from a set of \( n \), if

(i) **ordered, without replacement**, is \( \frac{n!}{(n-r)!} \);

(ii) **ordered, with replacement**, is \( n^r \);

(iii) **unordered, without replacement**, is \( \frac{n!}{r!(n-r)!} = \binom{n}{r} \);

(iv) **unordered, with replacement**, is \( \frac{(n+r-1)!}{r!(n-1)!} = \binom{n+r-1}{r} \).

13
Proof:

(i) $n$ choices to select 1st
   $n - 1$ choices to select 2nd
   
   ... 
   
   $n - r + 1$ choices to select $r^{th}$

   By Theorem 1.3.2, there are
   $\frac{n \times (n - 1) \times \ldots \times (n - r + 1) \times (n - r)!}{(n - r)!} = \frac{n!}{(n - r)!}$
   ways to do so.

Corollary:

The number of permutations of $n$ objects is $n!$.

(ii) $n$ choices to select 1st
    $n$ choices to select 2nd
    
    ... 
    
    $n$ choices to select $r^{th}$

    By Theorem 1.3.2, there are $n \times n \times \ldots \times n = n^r$ ways to do so.

(iii) We know from (i) above that there are $\frac{n!}{(n - r)!}$ ways to draw $r$ elements out of $n$ elements
    without replacement in the ordered case. However, for each unordered set of size $r$, there are $r!$ related
    ordered sets that consist of the same elements. Thus, there are
    $\frac{n!}{(n - r)!} \cdot \frac{1}{r!} = \binom{n}{r}$ ways to draw $r$ elements out of $n$ elements without replacement in the
    unordered case.

(iv) There is no immediate direct way to show this part. We have to come up with some extra motivation. We assume that there are $(n - 1)$ walls that separate the $n$ bins of possible outcomes and there are $r$ markers. If we shake everything, there are $(n - 1 + r)!$ permutations to arrange these $(n - 1)$ walls and $r$ markers according to the Corollary. Since the $r$ markers are indistinguishable and the $(n - 1)$ walls are also indistinguishable, we have to divide the number of permutations by $r!$ to get rid of identical permutations where only the markers are changed and by $(n - 1)!$ to get rid of identical permutations where only the walls are changed. Thus, there are $\frac{(n - 1 + r)!}{r!(n - 1)!} = \binom{n + r - 1}{r}$ ways to draw $r$ elements out of $n$ elements with replacement in the unordered case.

$\blacksquare$
Theorem 1.3.6: The Binomial Theorem

If \( n \) is a non-negative integer, then

\[
(1 + x)^n = \sum_{r=0}^{n} \binom{n}{r} x^r
\]

Proof: (By Induction)

Induction Base:

\( n = 0: \quad 1 = (1 + x)^0 = \sum_{r=0}^{0} \binom{0}{r} x^r = \binom{0}{0} x^0 = 1 \)

\( n = 1: \quad (1 + x)^1 = \sum_{r=0}^{1} \binom{1}{r} x^r = \binom{1}{0} x^0 + \binom{1}{1} x^1 = 1 + x \)

Induction Step:

Suppose it is true for \( k \). We show that it is true for \( k + 1 \).

\[
(1 + x)^{k+1} = (1 + x)^k (1 + x)
\]

\[
= \left( \sum_{r=0}^{k} \binom{k}{r} x^r \right) (1 + x)
\]

\[
= \sum_{r=0}^{k} \binom{k}{r} x^r + \sum_{r=0}^{k} \binom{k}{r} x^{r+1}
\]

\[
= \binom{k}{0} x^0 + \sum_{r=1}^{k} \left[ \binom{k}{r} + \binom{k}{r-1} \right] x^r + \binom{k}{k} x^{k+1}
\]

\[
= \binom{k+1}{0} x^0 + \sum_{r=1}^{k} \left[ \binom{k}{r} + \binom{k}{r-1} \right] x^r + \binom{k+1}{k+1} x^{k+1}
\]

\[
= \binom{k+1}{0} x^0 + \sum_{r=1}^{k} \binom{k+1}{r} x^r + \binom{k+1}{k+1} x^{k+1}
\]

\[
= \sum_{r=0}^{k+1} \binom{k+1}{r} x^r
\]

(*) Here we use Theorem 1.3.8 (i). Since the proof of Theorem 1.3.8 (i) only needs algebraic transformations without using the Binomial Theorem, part (i) of Theorem 1.3.8 can be applied here.
Corollary 1.3.7:
For integers $n$, it holds:

(i) $\binom{n}{0} + \binom{n}{1} + \ldots + \binom{n}{n} = 2^n$, $n \geq 0$

(ii) $\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \ldots + (-1)^n \binom{n}{n} = 0$, $n \geq 1$

(iii) $1 \cdot \binom{n}{1} + 2 \cdot \binom{n}{2} + 3 \cdot \binom{n}{3} + \ldots + n \cdot \binom{n}{n} = n2^{n-1}$, $n \geq 0$

(iv) $1 \cdot \binom{n}{1} - 2 \cdot \binom{n}{2} + 3 \cdot \binom{n}{3} + \ldots + (-1)^{n-1}n \cdot \binom{n}{n} = 0$, $n \geq 2$

Proof:
Use the Binomial Theorem:

(i) Let $x = 1$. Then for $n \geq 0$

$$2^n = (1 + 1)^n = \sum_{r=0}^{n} \binom{n}{r} 1^r = \sum_{r=0}^{n} \binom{n}{r}$$

(ii) Let $x = -1$. Then for $n \geq 1$

$$0 = 0^n = (1 + (-1))^n = \sum_{r=0}^{n} \binom{n}{r} (-1)^r$$

(iii) $\frac{d}{dx}(1 + x)^n = \frac{d}{dx} \sum_{r=0}^{n} \binom{n}{r} x^r$

$$\implies n(1 + x)^{n-1} = \sum_{r=1}^{n} r \cdot \binom{n}{r} x^{r-1}$$

Substitute $x = 1$, then for $n \geq 0$

$$n2^{n-1} = n(1 + 1)^{n-1} = \sum_{r=1}^{n} r \cdot \binom{n}{r}$$

(iv) Substitute $x = -1$ in (iii) above, then for $n \geq 2$

$$0 = n(1 + (-1))^{n-1} = \sum_{r=1}^{n} r \cdot \binom{n}{r} (-1)^{r-1}$$

Since for $\sum a_i = 0$ also $\sum (-a_i) = 0$, it also holds that

$$\sum_{r=1}^{n} r \cdot \binom{n}{r} (-1)^r = 0.$$
Theorem 1.3.8:
For non-negative integers, $n, m, r$, it holds:

(i) \[ \binom{n-1}{r} + \binom{n-1}{r-1} = \binom{n}{r} \]

(ii) \[ \binom{n}{0} \binom{m}{r} + \binom{n}{1} \binom{m}{r-1} + \ldots + \binom{n}{r} \binom{m}{0} = \binom{m+n}{r} \]

(iii) \[ \binom{0}{r} + \binom{1}{r} + \binom{2}{r} + \ldots + \binom{n}{r} = \binom{n+1}{r+1} \]

Proof:

Homework
### 1.4 Conditional Probability and Independence

So far, we have computed probability based only on the information that $\Omega$ is used for a probability space $(\Omega, L, P)$. Suppose, instead, we know that event $H \in L$ has happened. What statement should we then make about the chance of an event $A \in L$?

**Definition 1.4.1:**
Given $(\Omega, L, P)$ and $H \in L$, $P(H) > 0$, and $A \in L$, we define
\[
P(A|H) = \frac{P(A \cap H)}{P(H)} = P_H(A)
\]
and call this the **conditional probability of $A$ given $H$**.

**Note:**
$P(A|H)$ is undefined if $P(H) = 0$.

**Theorem 1.4.2:**
In the situation of Definition 1.4.1, $(\Omega, L, P_H)$ is a probability space.

**Proof:**
If $P_H$ is a probability measure, it must satisfy Def. 1.1.7.

(i) $P(H) > 0$ and by Def. 1.1.7 (i) $P(A \cap H) \geq 0 \implies P_H(A) = \frac{P(A \cap H)}{P(H)} \geq 0 \ \forall A \in L$

(ii) $P_H(\Omega) = \frac{P(H \cap \Omega)}{P(H)} = \frac{P(H)}{P(H)} = 1$

(iii) Let $\{A_n\}_{n=1}^\infty$ be a sequence of disjoint sets. Then,
\[
P_H(\bigcup_{n=1}^\infty A_n) \overset{Def.1.4.1}{=} \frac{P(\bigcup_{n=1}^\infty A_n \cap H)}{P(H)} = \frac{P(\bigcup_{n=1}^\infty (A_n \cap H))}{P(H)}
\]
\[
= \frac{\sum_{n=1}^\infty P(A_n \cap H)}{P(H)} \overset{Def.1.1.7(iii)}{=} \sum_{n=1}^\infty \frac{P(A_n \cap H)}{P(H)} = \sum_{n=1}^\infty P_H(A_n)
\]

\[
\]

18
Note:
What we have done is to move to a new sample space $H$ and a new $\sigma$–field $L_H = L \cap H$ of subsets $A \cap H$ for $A \in L$. We thus have a new measurable space $(\mathcal{H}, L_H)$ and a new probability space $(\mathcal{H}, L_H, P_H)$.

Note:
From Definition 1.4.1, if $A, B \in L$, $P(A) > 0$, and $P(B) > 0$, then

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B),$$

which generalizes to the following Theorem.

Theorem 1.4.3: Multiplication Rule
If $A_1, \ldots, A_n \in L$ and $P(\bigcap_{j=1}^{n-1} A_j) > 0$, then

$$P(\bigcap_{j=1}^{n} A_j) = P(A_1) \cdot P(A_2|A_1) \cdot P(A_3|A_1 \cap A_2) \cdot \ldots \cdot P(A_n|\bigcap_{j=1}^{n-1} A_j).$$

Proof:
Homework

Definition 1.4.4:
A collection of subsets $\{A_n\}_{n=1}^{\infty}$ of $\Omega$ form a partition of $\Omega$ if

(i) $\bigcup_{n=1}^{\infty} A_n = \Omega$, and

(ii) $A_i \cap A_j = \emptyset \ \forall i \neq j$, i.e., elements are pairwise disjoint.

Theorem 1.4.5: Law of Total Probability
If $\{H_j\}_{j=1}^{\infty}$ is a partition of $\Omega$, and $P(H_j) > 0 \ \forall j$, then, for $A \in L$,

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

Proof:
By the Note preceding Theorem 1.4.3, the summands on both sides are equal $\implies$ the right side of Th. 1.4.5 is true.
The left side proof:

$H_j$ are disjoint $\Rightarrow A \cap H_j$ are disjoint

$$A = A \cap \Omega \overset{\text{Def 1.4.4}}{=} A \cap \left( \bigcup_{j=1}^{\infty} H_j \right) = \bigcup_{j=1}^{\infty} (A \cap H_j)$$

$$\implies P(A) = P\left( \bigcup_{j=1}^{\infty} (A \cap H_j) \right) \overset{\text{Def 1.1.7(iii)}}{=} \sum_{j=1}^{\infty} P(A \cap H_j) \quad \blacksquare$$

**Theorem 1.4.6: Bayes’ Rule**

Let $\{H_j\}_{j=1}^{\infty}$ be a partition of $\Omega$, and $P(H_j) > 0 \ \forall j$. Let $A \in L$ and $P(A) > 0$. Then

$$P(H_j | A) = \frac{P(H_j)P(A | H_j)}{\sum_{n=1}^{\infty} P(H_n)P(A | H_n)} \quad \forall j.$$ 

**Proof:**

$$P(H_j \cap A) \overset{\text{Def 1.4.1}}{=} P(A) \cdot P(H_j | A) = P(H_j) \cdot P(A | H_j)$$

$$\implies P(H_j | A) = \frac{P(H_j)P(A | H_j)}{P(A)} \overset{\text{Th 1.4.5}}{=} \frac{P(H_j)P(A | H_j)}{\sum_{n=1}^{\infty} P(H_n)P(A | H_n)}. \quad \blacksquare$$

**Definition 1.4.7:**

For $A, B \in L$, $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$. \quad \blacksquare

**Note:**

- There are no restrictions on $P(A)$ or $P(B)$.
- If $A$ and $B$ are independent, then $P(A | B) = P(A)$ (given that $P(B) > 0$) and $P(B | A) = P(B)$ (given that $P(A) > 0$).
- If $A$ and $B$ are independent, then the following events are independent as well: $A$ and $B^C$; $A^C$ and $B$; $A^C$ and $B^C$. \quad \blacksquare

**Definition 1.4.8:**

Let $\mathcal{A}$ be a collection of $L$-sets. The events of $\mathcal{A}$ are pairwise independent iff for every distinct $A_1, A_2 \in \mathcal{A}$ it holds $P(A_1 \cap A_2) = P(A_1)P(A_2)$. \quad \blacksquare
Definition 1.4.9:
Let $\mathcal{A}$ be a collection of $L$–sets. The events of $\mathcal{A}$ are mutually independent (or completely independent) iff for every finite subcollection \( \{A_{i_1}, \ldots, A_{i_k}\}, A_{i_j} \in \mathcal{A} \), it holds
\[
P(\bigcap_{j=1}^{k} A_{i_j}) = \prod_{j=1}^{k} P(A_{i_j}).
\]

Note:
To check for mutually independence of $n$ events \( \{A_1, \ldots, A_n\} \in \mathcal{L} \), there are $2^n - n - 1$ relations (i.e., all subcollections of size 2 or more) to check.

Example 1.4.10:
Flip a fair coin twice. $\Omega = \{HH, HT, TH, TT\}$.
$A_1 = "H on 1st toss"
A_2 = "H on 2nd toss"
A_3 = "Exactly one $H"
Obviously, $P(A_1) = P(A_2) = P(A_3) = \frac{1}{2}$.
Question: Are $A_1, A_2$ and $A_3$ pairwise independent and also mutually independent?
$P(A_1 \cap A_2) = .25 = .5 \cdot .5 = P(A_1) \cdot P(A_2) \Rightarrow A_1, A_2$ are independent.
$P(A_1 \cap A_3) = .25 = .5 \cdot .5 = P(A_1) \cdot P(A_3) \Rightarrow A_1, A_3$ are independent.
$P(A_2 \cap A_3) = .25 = .5 \cdot .5 = P(A_2) \cdot P(A_3) \Rightarrow A_2, A_3$ are independent.
Thus, $A_1, A_2, A_3$ are pairwise independent.
$P(A_1 \cap A_2 \cap A_3) = 0 \neq .5 \cdot .5 \cdot .5 = P(A_1) \cdot P(A_2) \cdot P(A_3) \Rightarrow A_1, A_2, A_3$ are not mutually independent.

Example 1.4.11: (from Rohatgi, page 37, Example 5)
- $r$ students. 365 possible birthdays for each student that are equally likely.
- One student at a time is asked for his/her birthday.
- If one of the other students hears this birthday and it matches his/her birthday, this other student has to raise his/her hand — if at least one other student raises his/her hand, the procedure is over.
• We are interested in

\[ p_k = P(\text{procedure terminates at the } k\text{th student}) = P(\text{a hand is first risen when the } k\text{th student is asked for his/her birthday}) \]

• The textbook (Rohatgi) claims (without proof) that

\[ p_1 = 1 - \left( \frac{364}{365} \right)^{r-1} \]

and

\[ p_k = \left( \frac{365P_{k-1}}{(365)^{k-1}} \right) \left( \frac{1}{365} \right)^{r-k+1} \left[ 1 - \left( \frac{365 - k}{365 - k + 1} \right)^{r-k} \right], \quad k = 2, 3, \ldots, \]

where \(nP_r = n \cdot (n-1) \cdot \ldots \cdot (n-r+1)\).

Proof:
It is

\[ p_1 = P(\text{at least 1 other from the } (r-1) \text{ students has a birthday on this particular day.}) = 1 - P(\text{all } (r-1) \text{ students have a birthday on the remaining 364 out of 365 days}) = 1 - \left( \frac{364}{365} \right)^{r-1} \]

\[ p_2 = P(\text{no student has a birthday matching the first student and at least one of the other } (r-2) \text{ students has a b-day matching the second student}) \]

Let \( A \equiv \text{No student has a b-day matching the 1st student} \)

Let \( B \equiv \text{At least one of the other } (r-2) \text{ has b-day matching 2nd} \)

So \( p_2 = P(A \cap B) = P(A) \cdot P(B|A) = P(\text{no student has a matching b-day with the 1st student}) \times P(\text{at least one of the remaining students has a matching b-day with the second, given that no one matched the first.}) = (1-p_1)[1 - P(\text{all } (r-2) \text{ students have a b-day on the remaining 363 out of 364 days}) = \left( \frac{364}{365} \right)^{r-1} \left[ 1 - \left( \frac{363}{364} \right)^{r-2} \right] = \left( \frac{365 - 1}{365} \right)^{r-1} \left[ 1 - \left( \frac{363}{364} \right)^{r-2} \right] (\ast) \)
\[
\frac{365}{365} \left(1 - \frac{1}{365}\right)^{r-1} \left[1 - \left(\frac{363}{364}\right)^{r-2}\right]
\]

\[
= \frac{365 P_{2-1}}{(365)^2-1} \left(1 - \frac{2-1}{365}\right)^{r^2+1} \left[1 - \left(\frac{365 - 2}{365 - 2 + 1}\right)^{r-2}\right]
\]

Formally, we have to write this sequence of equalities in this order. However, it often might be easier to first work from both sides towards a particular result and combine partial results afterwards. Here, one might decide to stop at (\(\ast\)) with the “forward” direction of the equalities and first work “backwards” from the book, which makes things a lot simpler:

\[
p_2 = \frac{365 P_{2-1}}{(365)^2-1} \left(1 - \frac{2-1}{365}\right)^{r^2+1} \left[1 - \left(\frac{365 - 2}{365 - 2 + 1}\right)^{r-2}\right]
\]

\[
= \frac{365}{365} \left(1 - \frac{1}{365}\right)^{r-1} \left[1 - \left(\frac{363}{364}\right)^{r-2}\right]
\]

\[
= \left(\frac{365 - 1}{365}\right)^{r-1} \left[1 - \left(\frac{363}{364}\right)^{r-2}\right]
\]

We see that this is the same result as (\(\ast\)).

Now let us consider \(p_3\):

\[
p_3 = P(\text{No one has same b-day as first and no one same as second, and at least one of the remaining (}r - 3\text{) has a matching b-day with the 3rd student})
\]

Let A \equiv \text{No one has the same b-day as the first student}

Let B \equiv \text{No one has the same b-day as the second student}

Let C \equiv \text{At least one of the other (}r - 3\text{) has the same b-day as the third student}

Now:

\[
p_3 = P(A \cap B \cap C)
\]

\[
= P(A) \cdot P(B|A) \cdot P(C|A \cap B)
\]

\[
= \left(\frac{364}{365}\right)^{r-1} \left(\frac{363}{364}\right)^{r-2} \cdot \left[1 - P(\text{all (}r - 3\text{) students have a b-day on the remaining 362 out of 363 days})\right]
\]

\[
= \left(\frac{364}{365}\right)^{r-1} \left(\frac{363}{364}\right)^{r-2} \cdot \left[1 - \left(\frac{362}{363}\right)^{r-3}\right]
\]

\[
= \left(\frac{364}{365}\right)^{r-1} \cdot \left(\frac{363}{364}\right)^{r-2} \cdot \left[1 - \left(\frac{362}{363}\right)^{r-3}\right]
\]

23
\[
\begin{align*}
\left( \frac{364^{r-1}}{364^{r-2}} \right) \left( \frac{363^{r-2}}{365^{r-1}} \right) \cdot \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \\
= \left( \frac{364}{365} \right) \left( \frac{363^{r-2}}{365^{r-2}} \right) \cdot \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \\
= \left( \frac{(365)(364)}{(365)^2} \right) \left( \frac{363}{365} \right)^{r-2} \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \\
= \left( \frac{365 \cdot P_2}{(365)^2} \right) \left( 1 - \frac{2}{365} \right)^{r-2} \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \\
= \left( \frac{365 \cdot P_{3-1}}{(365)^3-1} \right) \left( 1 - \frac{3 - 1}{365} \right)^{r-3+1} \left[ 1 - \left( \frac{365 - 3}{365 - 3 + 1} \right)^{r-3} \right]
\end{align*}
\]

Once again, working “backwards” from the book should help to better understand these transformations.

For general \( p_k \) and restrictions on \( r \) and \( k \) see Homework. \[\blacksquare\]
2 Random Variables

(Based on Casella/Berger, Sections 1.4, 1.5, 1.6 & 2.1)

2.1 Measurable Functions

Definition 2.1.1:

- A random variable (rv) is a set function from $\Omega$ to $\mathbb{R}$.
- More formally: Let $(\Omega, L, P)$ be any probability space. Suppose $X : \Omega \to \mathbb{R}$ and that $X$ is a measurable function, then we call $X$ a random variable.
- More generally: If $X : \Omega \to \mathbb{R}^k$, we call $X$ a random vector, $X = (X_1(\omega), X_2(\omega), \ldots, X_k(\omega))$.

What does it mean to say that a function is measurable?

Definition 2.1.2:

Suppose $(\Omega, L)$ and $(S, B)$ are two measurable spaces and $X : \Omega \to S$ is a mapping from $\Omega$ to $S$. We say that $X$ is measurable if $X^{-1}(B) \in L$ for every set $B \in B$, where $X^{-1}(B) = \{ \omega \in \Omega : X(\omega) \in B \}$.

Example 2.1.3:

Record the opinion of 50 people: “yes” (y) or “no” (n).

$\Omega = \{ \text{All 250 possible sequences of y/n} \} — \text{HUGE!} \quad L = \mathcal{P}(\Omega)$

(i) Consider $X : \Omega \to S = \{ \text{All 250 possible sequences of 1 (= y) and 0 (= n)} \}$

$B = \mathcal{P}(S)$

$X$ is a random vector since each element in $S$ has a corresponding element in $\Omega$, for $B \in B$, $X^{-1}(B) \in L = \mathcal{P}(\Omega)$.

(ii) Consider $X : \Omega \to S = \{0, 1, 2, \ldots, 50\}$, where $X(\omega) = \# \text{ of y’s in } \omega$ is a more manageable random variable.

A simple function, i.e., a function that takes only finite many values $x_1, \ldots, x_k$, is measurable if $X^{-1}(x_i) \in L \ \forall x_i$.

Here, $X^{-1}(k) = \{ \omega \in \Omega : \# \text{ y’s in sequence } \omega = k \}$ is a subset of $\Omega$, so it is in $L = \mathcal{P}(\Omega)$. 

25
Example 2.1.4:
Let $\Omega = "\text{infinite fair coin tossing space}"$, i.e., infinite sequence of H’s and T’s.
Let $L_n$ be a $\sigma$-field for the 1st $n$ tosses.
Define $L = \sigma(\bigcup_{n=1}^{\infty} L_n)$.
Let $X_n : \Omega \to \mathbb{R}$ be $X_n(\omega) = "\text{proportion of H’s in 1st n tosses}"$.
For each $n$, $X_n(\cdot)$ is simple (values $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$) and $X_n^{-1}(\frac{k}{n}) \in L_n \ \forall k = 0, 1, \ldots, n$.
Therefore, $X_n^{-1}(\frac{k}{n}) \in L$.

So every random variable $X_n(\cdot)$ is measurable $L - \mathcal{B}$. Now we have a sequence of rv’s $\{X_n\}_{n=1}^{\infty}$.
We will show later that $P(\{\omega : X_n(\omega) \to \frac{1}{2}\}) = 1$, i.e., the Strong Law of Large Numbers (SLLN).

Some Technical Points about Measurable Functions

2.1.5:
Suppose $(\Omega, L)$ and $(S, \mathcal{B})$ are measure spaces and that a collection of sets $\mathcal{A}$ generates $\mathcal{B}$, i.e., $\sigma(\mathcal{A}) = \mathcal{B}$. Let $X : \Omega \to S$. If $X^{-1}(A) \in L \ \forall A \in \mathcal{A}$, then $X$ is measurable $L - \mathcal{B}$.

This means we only have to check measurability on a basis collection $\mathcal{A}$. The usage is: $\mathcal{B}$ on $\mathbb{R}$ is generated by $\{(-\infty, x] : x \in \mathbb{R}\}$.

2.1.6:
If $(\Omega, L), (\Omega', L')$, and $(\Omega'', L'')$ are measure spaces and $X : \Omega \to \Omega'$ and $Y : \Omega' \to \Omega''$ are measurable, then the composition $(YX) : \Omega \to \Omega''$ is measurable $L - L''$.

2.1.7:
If $f : \mathbb{R}^i \to \mathbb{R}^k$ is a continuous function, then $f$ is measurable $\mathcal{B}^i - \mathcal{B}^k$.

2.1.8:
If $f_j : \Omega \to \mathbb{R}, j = 1, \ldots k$ and $g : \mathbb{R}^k \to \mathbb{R}$ are measurable, then $g(f_1(\cdot), \ldots, f_k(\cdot))$ is measurable.

The usage is: $g$ could be sum, average, difference, product, (finite) maximums and minimums of $x_1, \ldots, x_k$, etc.
2.1.9:
Limits: Extend the real line to \([\mathbb{R} \cup \{-\infty, \infty\}] = \mathbb{IR}\).
We say \(f : \Omega \to \mathbb{R}\) is measurable \(L - B\) if

(i) \(f^{-1}(B) \in L, \forall B \in \mathcal{B}\), and

(ii) \(f^{-1}(-\infty), f^{-1}(\infty) \in L\) also.

\(\blacksquare\)

2.1.10:
Suppose \(f_1, f_2, \ldots\) is a sequence of real–valued measurable functions \((\Omega, L) \to (\mathbb{IR}, \mathcal{B})\). Then it holds:

(i) \(\sup_{n \to \infty} f_n\) (supremum), \(\inf_{n \to \infty} f_n\) (infimum), \(\lim_{n \to \infty} \sup f_n\) (limit superior), and \(\lim_{n \to \infty} \inf f_n\) (limit inferior) are measurable.

(ii) If \(f = \lim_{n \to \infty} f_n\) exists, then \(f\) is measurable.

(iii) The set \(\{\omega : f_n(\omega)\text{ converges}\} \in L\).

(iv) If \(f\) is any measurable function, the set \(\{\omega : f_n(\omega) \to f(\omega)\} \in L\).

\(\blacksquare\)

Example 2.1.11:

(i) Let
\[
    f_n(x) = \frac{1}{x^n}, \quad x > 1.
\]
It holds

- \(\sup_{n \to \infty} f_n(x) = \frac{1}{x}\)
- \(\inf_{n \to \infty} f_n(x) = 0\)
- \(\lim_{n \to \infty} \sup f_n(x) = 0\)
- \(\lim_{n \to \infty} \inf f_n(x) = 0\)
- \(\lim_{n \to \infty} f_n(x) = \lim_{n \to \infty} \sup f_n(x) = \lim_{n \to \infty} \inf f_n(x) = 0\)

(ii) Let
\[
    f_n(x) = \begin{cases} 
        x^3, & x \in [-n, n] \\
        0, & \text{otherwise}
    \end{cases}
\]
It holds
\[ \sup_{n \to \infty} f_n(x) = \begin{cases} x^3, & x \geq -1 \\ 0, & \text{otherwise} \end{cases} \]

\[ \inf_{n \to \infty} f_n(x) = \begin{cases} x^3, & x \leq 1 \\ 0, & \text{otherwise} \end{cases} \]

\[ \lim_{n \to \infty} \sup f_n(x) = x^3 \]

\[ \lim_{n \to \infty} \inf f_n(x) = x^3 \]

\[ \lim_{n \to \infty} f_n(x) = \lim_{n \to \infty} \sup f_n(x) = \lim_{n \to \infty} \inf f_n(x) = x^3 \]

(iii) Let

\[ f_n(x) = \begin{cases} (-1)^n x^3, & x \in [-n, n] \\ 0, & \text{otherwise} \end{cases} \]

It holds

\[ \sup_{n \to \infty} f_n(x) = |x|^3 \]

\[ \inf_{n \to \infty} f_n(x) = -|x|^3 \]

\[ \lim_{n \to \infty} \sup f_n(x) = |x|^3 \]

\[ \lim_{n \to \infty} \inf f_n(x) = -|x|^3 \]

\[ \lim_{n \to \infty} f_n(x) = \lim_{n \to \infty} \sup f_n(x) = \lim_{n \to \infty} \inf f_n(x) = 0 \text{ if } x = 0, \text{ but } \lim_{n \to \infty} f_n(x) \text{ does not exist for } x \neq 0 \text{ since } \lim_{n \to \infty} \sup f_n(x) \neq \lim_{n \to \infty} \inf f_n(x) \text{ for } x \neq 0 \]
2.2 Probability Distribution of a Random Variable

The definition of a random variable \( X : (\Omega, L) \to (S, B) \) makes no mention of \( P \). We now introduce a probability measure on \((S, B)\).

**Theorem 2.2.1:**
A random variable \( X \) on \((\Omega, L, P)\) induces a probability measure on a space \((\mathbb{R}, B, Q)\) with the probability distribution \( Q \) of \( X \) defined by
\[
Q(B) = P(X^{-1}(B)) = P(\{ \omega : X(\omega) \in B \}) \quad \forall B \in B.
\]

**Note:**
By the definition of a random variable, \( X^{-1}(B) \in L \quad \forall B \in B \). \( Q \) is called induced probability.

**Proof:**
If \( X \) induces a probability measure \( Q \) on \((\mathbb{R}, B)\), then \( Q \) must satisfy the Kolmogorov Axioms of probability.
\( X : (\Omega, L) \to (S, B) \). \( X \) is a rv \( \Rightarrow X^{-1}(B) = \{ \omega : X(\omega) \in B \} = A \in L \quad \forall B \in B \).

\( i \) \( Q(B) = P(X^{-1}(B)) = P(\{ \omega : X(\omega) \in B \}) = P(A) \quad \text{Def. 1.1.7(i)} \geq 0 \quad \forall B \in B \)

\( ii \) \( Q(\mathbb{R}) = P(X^{-1}(\mathbb{R})) \stackrel{X \text{rv}}{=} P(\Omega) \quad \text{Def. 1.1.7(ii)} = 1 \)

\( iii \) Let \( \{ B_n \}_{n=1}^{\infty} \in B, B_i \cap B_j = \emptyset \quad \forall i \neq j \). Then,
\[
Q(\bigcup_{n=1}^{\infty} B_n) = P(X^{-1}(\bigcup_{n=1}^{\infty} B_n)) \quad \text{(s) \ Stackrel{Def. 1.1.7(iii)}{=} P(\bigcup_{n=1}^{\infty} (X^{-1}(B_n))) = \sum_{n=1}^{\infty} P(X^{-1}(B_n)) = \sum_{n=1}^{\infty} Q(B_n)}
\]
\( \text{(s) holds since } X^{-1}(\cdot) \text{ commutes with unions/intersections and preserves disjointedness.} \)

**Definition 2.2.2:**
A real–valued function \( F \) on \((–\infty, \infty)\) that is non–decreasing, right–continuous, and satisfies
\[
F(–\infty) = 0, F(\infty) = 1
\]
is called a **cumulative distribution function** (cdf) on \( \mathbb{R} \).
Note:
No mention of probability space or measure $P$ in Definition 2.2.2 above.

Definition 2.2.3:
Let $P$ be a probability measure on $(\mathbb{R}, \mathcal{B})$. The cdf associated with $P$ is

$$F(x) = F_P(x) = P((\infty, x]) = P(\{\omega : X(\omega) \leq x\}) = P(X \leq x)$$

for a random variable $X$ defined on $(\mathbb{R}, \mathcal{B}, P)$.

Note:
$F(\cdot)$ defined as in Definition 2.2.3 above indeed is a cdf.

Proof (of Note):

(i) Let $x_1 < x_2$

$$\implies (-\infty, x_1] \subset (-\infty, x_2]\]$$

$$\implies F(x_1) = P(\{\omega : X(\omega) \leq x_1\}) \overset{Th.1.2.1(v)}{\leq} P(\{\omega : X(\omega) \leq x_2\}) = F(x_2)$$

Thus, since $x_1 < x_2$ and $F(x_1) \leq F(x_2)$, $F(\cdot)$ is non-decreasing.

(ii) Since $F$ is non-decreasing, it is sufficient to show that $F(\cdot)$ is right-continuous if for any sequence of numbers $x_n \to x+$ (which means that $x_n$ is approaching $x$ from the right) with $x_1 > x_2 > \ldots > x_n > \ldots > x : F(x_n) \to F(x)$.

Let $A_n = \{\omega : X(\omega) \in (x, x_n]\} \in L$ and $A_n \downarrow \emptyset$. None of the intervals $(x, x_n]$ contains $x$. As $x_n \to x+$, the number of points $\omega$ in $A_n$ diminishes until the set is empty. Formally,

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

By Theorem 1.2.6 it follows that

$$\lim_{n \to \infty} P(A_n) = P(\lim_{n \to \infty} A_n) = P(\emptyset) = 0.$$  

It is

$$P(A_n) = P(\{\omega : X(\omega) \leq x_n\}) - P(\{\omega : X(\omega) \leq x\}) = F(x_n) - F(x).$$

$$\implies (\lim_{n \to \infty} F(x_n)) - F(x) = \lim_{n \to \infty} (F(x_n) - F(x)) = \lim_{n \to \infty} P(A_n) = 0$$

$$\implies \lim_{n \to \infty} F(x_n) = F(x)$$

$$\implies F(x) \text{ is right-continuous.}$$
(iii) $F(-n) \overset{\text{Def.}}{=} P(\{\omega : X(\omega) \leq -n\})$

$$F(-\infty) = \lim_{n \to \infty} F(-n)$$
$$= \lim_{n \to \infty} P(\{\omega : X(\omega) \leq -n\})$$
$$= P(\lim_{n \to \infty} \{\omega : X(\omega) \leq -n\})$$
$$= P(\emptyset)$$
$$= 0$$

(iv) $F(n) \overset{\text{Def.}}{=} P(\{\omega : X(\omega) \leq n\})$

$$F(\infty) = \lim_{n \to \infty} F(n)$$
$$= \lim_{n \to \infty} P(\{\omega : X(\omega) \leq n\})$$
$$= P(\lim_{n \to \infty} \{\omega : X(\omega) \leq n\})$$
$$= P(\Omega)$$
$$= 1$$

Note that (iii) and (iv) implicitly use Theorem 1.2.6. In (iii), we use $A_n = (-\infty, -n)$ where $A_n \supset A_{n+1}$ and $A_n \downarrow \emptyset$. In (iv), we use $A_n = (-\infty, n)$ where $A_n \subset A_{n+1}$ and $A_n \uparrow \mathbb{R}$. □

Definition 2.2.4:
If a random variable $X : \Omega \to \mathbb{R}$ has induced a probability measure $P_X$ on $(\mathbb{R}, \mathcal{B})$ with cdf $F(x)$, we say

(i) rv $X$ is **continuous** if $F(x)$ is continuous in $x$.

(ii) rv $X$ is **discrete** if $F(x)$ is a step function in $x$.

□

Note:
There are rvs that are mixtures of continuous and discrete rvs. One such example is a truncated failure time distribution. We assume a continuous distribution (e.g., exponential) up to a given truncation point $x$ and assign the “remaining” probability to the truncation point. Thus, a single point has a probability $> 0$ and $F(x)$ jumps at the truncation point $x$. □
Definition 2.2.5:
Two random variables $X$ and $Y$ are **identically distributed** iff

$$P_X(X \in A) = P_Y(Y \in A) \forall A \in L.$$  

Note:
Def. 2.2.5 does not mean that $X(\omega) = Y(\omega) \forall \omega \in \Omega$. For example,

- $X = \# H$ in 3 coin tosses
- $Y = \# T$ in 3 coin tosses

$X, Y$ are both $Bin(3, 0.5)$, i.e., identically distributed, but for $\omega = (H, H, T)$, $X(\omega) = 2 \neq 1 = Y(\omega)$, i.e., $X \neq Y$.

Theorem 2.2.6:
The following two statements are equivalent:

(i) $X, Y$ are identically distributed.

(ii) $F_X(x) = F_Y(x) \forall x \in \mathbb{R}$.

Proof:
(i) $\Rightarrow$ (ii):

\[
F_X(x) = P_X((\infty, x]) = P\{\omega : X(\omega) \in (\infty, x]\} = P\{\omega : Y(\omega) \in (\infty, x]\}
\]

by Def. 2.2.5 = $P_Y((\infty, x]) = F_Y(x)$

(ii) $\Rightarrow$ (i):

Requires extra knowledge from measure theory.  

\[
\]
2.3 Discrete and Continuous Random Variables

We now extend Definition 2.2.4 to make our definitions a little bit more formal.

Definition 2.3.1:
Let $X$ be a real–valued random variable with cdf $F$ on $(\Omega, L, P)$. $X$ is discrete if there exists a countable set $E \subset \mathbb{R}$ such that $P(X \in E) = 1$, i.e., $P(\{\omega : X(\omega) \in E\}) = 1$. The points of $E$ which have positive probability are the jump points of the step function $F$, i.e., the cdf of $X$.

Define $p_i = P(\{\omega : X(\omega) = x_i, x_i \in E\}) = P_X(X = x_i) \ \forall i \geq 1$. Then, $p_i \geq 0$, $\sum_{i=1}^{\infty} p_i = 1$.

We call $\{p_i : p_i \geq 0\}$ the probability mass function (pmf) (also: probability frequency function) of $X$.

Note:
Given any set of numbers $\{p_n\}_{n=1}^{\infty}, p_n \geq 0 \ \forall n \geq 1, \sum_{n=1}^{\infty} p_n = 1$, $\{p_n\}_{n=1}^{\infty}$ is the pmf of some rv $X$.

Note:
The issue of continuous rv’s and probability density functions (pdfs) is more complicated. A rv $X : \Omega \rightarrow \mathbb{R}$ always has a cdf $F$. Whether there exists a function $f$ such that $f$ integrates to $F$ and $F'$ exists and equals $f$ (almost everywhere) depends on something stronger than just continuity.

Definition 2.3.2:
A real–valued function $F$ is continuous in $x_0 \in \mathbb{R}$ iff

$$\forall \epsilon > 0 \ \exists \delta > 0 \ \forall x : \ |x - x_0| < \delta \Rightarrow |F(x) - F(x_0)| < \epsilon.$$ 

$F$ is continuous iff $F$ is continuous in all $x \in \mathbb{R}$.
**Definition 2.3.3:**
A real-valued function $F$ defined on $[a, b]$ is **absolutely continuous** on $[a, b]$ iff

$$\forall \epsilon > 0 \ \exists \delta > 0 \ \forall \text{ finite subcollection of disjoint subintervals } [a_i, b_i], i = 1, \ldots, n :$$

$$\sum_{i=1}^{n} (b_i - a_i) < \delta \Rightarrow \sum_{i=1}^{n} |F(b_i) - F(a_i)| < \epsilon.$$

**Note:**
Absolute continuity implies continuity.

**Theorem 2.3.4:**

(i) If $F$ is absolutely continuous, then $F'$ exists almost everywhere.

(ii) A function $F$ is an indefinite integral iff it is absolutely continuous. Thus, every absolutely continuous function $F$ is the indefinite integral of its derivative $F'$.

**Definition 2.3.5:**
Let $X$ be a random variable on $(\Omega, L, P)$ with cdf $F$. We say $X$ is a **continuous rv** iff $F$ is absolutely continuous. In this case, there exists a non-negative integrable function $f$, the **probability density function (pdf)** of $X$, such that

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

From this it follows that, if $a, b \in \mathbb{R}, a < b$, then

$$P_X(a < X \leq b) = F(b) - F(a) = \int_{a}^{b} f(t)dt$$

exists and is well defined.

**Theorem 2.3.6:**
Let $X$ be a continuous random variable with pdf $f$. Then it holds:

(i) For every Borel set $B \in \mathcal{B}, P(B) = \int_{B} f(t)dt$.

(ii) If $F$ is absolutely continuous and $f$ is continuous at $x$, then $F'(x) = \frac{dF(x)}{dx} = f(x)$.  

Proof:
Part (i): From Definition 2.3.5 above.
Part (ii): By Fundamental Theorem of Calculus.

Note:
As already stated in the Note following Definition 2.2.4, not every rv will fall into one of
these two (or if you prefer – three –, i.e., discrete, continuous/absolutely continuous) classes.
However, most rv which arise in practice will. We look at one example that is unlikely to
occur in practice in the next Homework assignment.

However, note that every cdf \( F \) can be written as
\[
F(x) = aF_d(x) + (1 - a)F_c(x), \quad 0 \leq a \leq 1,
\]
where \( F_d \) is the cdf of a discrete rv and \( F_c \) is a continuous (but not necessarily absolute
continuous) cdf.

Some authors, such as Marek Fisz Wahrscheinlichkeitsrechnung und mathematische Statistik,
VEB Deutscher Verlag der Wissenschaften, Berlin, 1989, are even more specific. There it is
stated that every cdf \( F \) can be written as
\[
F(x) = a_1F_d(x) + a_2F_c(x) + a_3F_s(x), \quad a_1, a_2, a_3 \geq 0, a_1 + a_2 + a_3 = 1.
\]
Here, \( F_d(x) \) and \( F_c(x) \) are discrete and absolute continuous cdfs. \( F_s(x) \) is called a singular cdf. Singular means that \( F_s(x) \) is continuous and its derivative \( F'(x) \) equals 0 almost
everywhere (i.e., everywhere but in those points that belong to a Borel–measurable set of
probability 0).

Question: Does “continuous” but “not absolutely continuous” mean “singular”? — We will
(hopefully) see later…

Example 2.3.7:
Consider
\[
F(x) = \begin{cases} 
0, & x < 0 \\
1/2, & x = 0 \\
1/2 + x/2, & 0 < x < 1 \\
1, & x \geq 1
\end{cases}
\]

We can write \( F(x) \) as \( aF_d(x) + (1 - a)F_c(x), 0 \leq a \leq 1 \). How?

Since \( F(x) \) has only one jump at \( x = 0 \), it is reasonable to get started with a pmf \( p_0 = 1 \) and

\[
F_d(x) = \begin{cases} 
0, & x < 0 \\
1, & x \geq 0
\end{cases}
\]
Since $F(x) = 0$ for $x < 0$ and $F(x) = 1$ for $x \geq 1$, it must clearly hold that $F_c(x) = 0$ for $x < 0$ and $F_c(x) = 1$ for $x \geq 1$. In addition $F(x)$ increases linearly in $0 < x < 1$. A good guess would be a pdf $f_c(x) = 1 \cdot I_{(0,1)}(x)$ and corresponding cdf

$$F_c(x) = \begin{cases} 
0, & x \leq 0 \\
 x, & 0 < x < 1 \\
1, & x \geq 1 
\end{cases}$$

Knowing that $F(0) = 1/2$, we have at least to multiply $F_d(x)$ by $1/2$. And, indeed, $F(x)$ can be written as

$$F(x) = \frac{1}{2}F_d(x) + \frac{1}{2}F_c(x).$$

**Definition 2.3.8:**

The two-valued function $I_A(x)$ is called **indicator function** and it is defined as follows:

$I_A(x) = 1$ if $x \in A$ and $I_A(x) = 0$ if $x \notin A$ for any set $A$. 

\[ \square \]
An Excursion into Logic

When proving theorems we only used direct methods so far. We used induction proofs to show that something holds for arbitrary $n$. To show that a statement $A$ implies a statement $B$, i.e., $A \Rightarrow B$, we used proofs of the type $A \Rightarrow A_1 \Rightarrow A_2 \Rightarrow \ldots \Rightarrow A_{n-1} \Rightarrow A_n \Rightarrow B$ where one step directly follows from the previous step. However, there are different approaches to obtain the same result.

Basic Operators:

Boolean Logic makes assertions on statements that can either be true (represented as 1) or false (represented as 0). Basic operators are “not” ($\neg$), “and” ($\land$), “or” ($\lor$), “implies” ($\Rightarrow$), “equivalent” ($\Leftrightarrow$), and “exclusive or” ($\oplus$).

These operators are defined as follows:

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$\neg A$</th>
<th>$\neg B$</th>
<th>$A \land B$</th>
<th>$A \lor B$</th>
<th>$A \Rightarrow B$</th>
<th>$A \Leftrightarrow B$</th>
<th>$A \oplus B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Implication: ($A$ implies $B$)

$A \Rightarrow B$ is equivalent to $\neg B \Rightarrow \neg A$ is equivalent to $\neg A \lor B$:

<table>
<thead>
<tr>
<th>$A$</th>
<th>$B$</th>
<th>$A \Rightarrow B$</th>
<th>$\neg A$</th>
<th>$\neg B$</th>
<th>$\neg B \Rightarrow \neg A$</th>
<th>$\neg A \lor B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Equivalence: \(A\) is equivalent to \(B\)

\[A \iff B \text{ is equivalent to } (A \Rightarrow B) \land (B \Rightarrow A)\]

\[\iff (\neg A \lor B) \land (A \lor \neg B)\]

<table>
<thead>
<tr>
<th>(A)</th>
<th>(B)</th>
<th>(A \Leftrightarrow B)</th>
<th>(A \Rightarrow B)</th>
<th>(B \Rightarrow A)</th>
<th>((A \Rightarrow B) \land (B \Rightarrow A))</th>
<th>(\neg A \lor B)</th>
<th>(A \lor \neg B)</th>
<th>((\neg A \lor B) \land (A \lor \neg B))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Negations of Quantifiers:

The quantifiers “for all” \((\forall)\) and “it exists” \((\exists)\) are used to indicate that a statement holds for all possible values or that there exists such a value that makes the statement true, respectively. When negating a statement with a quantifier, this means that we flip from one quantifier to the other with the remaining statement negated as well, i.e., \(\neg \forall \) becomes \(\exists\) and \(\neg \exists \) becomes \(\forall\).

\(\neg \forall x \in X : B(x)\) is equivalent to \(\exists x \in X : \neg B(x)\)

\(\neg \exists x \in X : B(x)\) is equivalent to \(\forall x \in X : \neg B(x)\)

\(\exists x \in X \forall y \in Y : B(x, y)\) implies \(\forall y \in Y \exists x \in X : B(x, y)\)
2.4 Transformations of Random Variables

Let $X$ be a real–valued random variable on $(\Omega, L, P)$, i.e., $X : (\Omega, L) \to (\mathbb{R}, \mathcal{B})$. Let $g$ be any Borel–measurable real–valued function on $\mathbb{R}$. Then, by statement 2.1.6, $Y = g(X)$ is a random variable.

**Theorem 2.4.1:**
Given a random variable $X$ with known induced distribution and a Borel–measurable function $g$, then the distribution of the random variable $Y = g(X)$ is determined.

**Proof:**

\[
F_Y(y) = P(Y \leq y) = P(\{\omega : g(X(\omega)) \leq y\}) = P(\{\omega : X(\omega) \in B_y\}) \quad \text{where } B_y = g^{-1}((-\infty, y]) \in \mathcal{B} \text{ since } g \text{ is Borel–measurable.}
\]

\[
= P(X^{-1}(B_y))
\]

**Note:**
From now on, we restrict ourselves to real–valued (vector–valued) functions that are Borel–measurable, i.e., measurable with respect to $(\mathbb{R}, \mathcal{B})$ or $(\mathbb{R}^k, \mathcal{B}^k)$.

More generally, $P_Y(Y \in C) = P_X(X \in g^{-1}(C)) \quad \forall C \in \mathcal{B}.

**Example 2.4.2:**
Suppose $X$ is a discrete random variable. Let $A$ be a countable set such that $P(X \in A) = 1$ and $P(X = x) > 0 \quad \forall x \in A$.

Let $Y = g(X)$. Obviously, the sample space of $Y$ is also countable. Then,

\[
P_Y(Y = y) = \sum_{x \in g^{-1}\{y\}} P_X(X = x) = \sum_{\{x : g(x) = y\}} P_X(X = x) \quad \forall y \in g(A).
\]

**Example 2.4.3:**
$X \sim U(-1, 1)$ so the pdf of $X$ is $f_X(x) = 1/2I_{[-1,1]}(x)$, which, according to Definition 2.3.8, reads as $f_X(x) = 1/2$ for $-1 \leq x \leq 1$ and 0 otherwise.

Let $Y = X^+ = \begin{cases} x, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$
Then,

\[
F_Y(y) = P_Y(Y \leq y) = \begin{cases} 
0, & y < 0 \\
1/2, & y = 0 \\
1/2 + y/2, & 0 < y < 1 \\
1, & y \geq 1 
\end{cases}
\]

This is the mixed discrete/continuous distribution from Example 2.3.7.

Note:
We need to put some conditions on \( g \) to ensure \( g(X) \) is continuous if \( X \) is continuous and avoid cases as in Example 2.4.3 above.

Definition 2.4.4:
For a random variable \( X \) from \((\Omega, L, P)\) to \((\mathbb{R}, \mathcal{B})\), the support of \( X \) (or \( P \)) is any set \( A \in L \) for which \( P(A) = 1 \). For a continuous random variable \( X \) with pdf \( f \), we can think of the support of \( X \) as \( X = \{ x : f_X(x) > 0 \} \).

Definition 2.4.5:
Let \( f \) be a real–valued function defined on \( D \subseteq \mathbb{R}, D \in \mathcal{B} \). We say:
- \( f \) is non–decreasing if \( x < y \implies f(x) \leq f(y) \ \forall x, y \in D \)
- \( f \) is strictly non–decreasing (or increasing) if \( x < y \implies f(x) < f(y) \ \forall x, y \in D \)
- \( f \) is non–increasing if \( x < y \implies f(x) \geq f(y) \ \forall x, y \in D \)
- \( f \) is strictly non–increasing (or decreasing) if \( x < y \implies f(x) > f(y) \ \forall x, y \in D \)
- \( f \) is monotonic on \( D \) if \( f \) is either increasing or decreasing and write \( f \uparrow \) or \( f \downarrow \).

Theorem 2.4.6:
Let \( X \) be a continuous rv with pdf \( f_X \) and support \( \mathcal{X} \). Let \( y = g(x) \) be differentiable for all \( x \) and either (i) \( g'(x) > 0 \) or (ii) \( g'(x) < 0 \) for all \( x \).

Then, \( Y = g(X) \) is also a continuous rv with pdf

\[
f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| \cdot I_{g(\mathcal{X})(y)}.
\]

Proof:
Part (i): \( g'(x) > 0 \ \forall x \in \mathcal{X} \)
So \( g \) is strictly increasing and continuous.
Therefore, \( x = g^{-1}(y) \) exists and it is also strictly increasing and also differentiable.
It holds that
\[
\frac{d}{dy} g^{-1}(y) = \left( \frac{d}{dx} g(x) \big|_{x=g^{-1}(y)} \right)^{-1} > 0.
\]
We get \( F_Y(y) = P_Y(Y \leq y) = P_Y(g(X) \leq y) = P_X(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \) for \( y \in g(\mathcal{X}) \) and, by differentiation,
\[
f_Y(y) = F_Y'(y) = \frac{d}{dy}(F_X(g^{-1}(y))) \quad \text{By Chain Rule} \]
\[
= f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y)
\]

**Part (ii):** \( g'(x) < 0 \ \forall x \in \mathcal{X} \)

So \( g \) is strictly decreasing and continuous.
Therefore, \( x = g^{-1}(y) \) exists and it is also strictly decreasing and also differentiable.

It holds that
\[
\frac{d}{dy} g^{-1}(y) = \left( \frac{d}{dx} g(x) \big|_{x=g^{-1}(y)} \right)^{-1} < 0.
\]
We get \( F_Y(y) = P_Y(Y \leq y) = P_Y(g(X) \leq y) = P_X(X \geq g^{-1}(y)) = 1 - P_X(X \leq g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \) for \( y \in g(\mathcal{X}) \) and, by differentiation,
\[
f_Y(y) = F_Y'(y) = \frac{d}{dy}(1-F_X(g^{-1}(y))) \quad \text{By Chain Rule} \]
\[
= -f_X(g^{-1}(y)) \cdot \frac{d}{dy} g^{-1}(y) = f_X(g^{-1}(y)) \cdot \left( -\frac{d}{dy} g^{-1}(y) \right)
\]

Since \( \frac{d}{dy} g^{-1}(y) < 0 \), the negative sign will cancel out, always giving us a positive value. Hence the need for the absolute value signs.

Combining parts (i) and (ii), we can therefore write
\[
f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{d}{dy} g^{-1}(y) \right| \cdot I_{g(\mathcal{X})}(y).
\]

**Note:**
In Theorem 2.4.6, we can also write
\[
f_Y(y) = \frac{f_X(x)}{\left| \frac{dg(x)}{dx} \right|_{x=g^{-1}(y)}}, \ y \in g(\mathcal{X})
\]

If \( g \) is monotonic over disjoint intervals, we can also get an expression for the pdf/cdf of \( Y = g(X) \) as stated in the following Theorem.

\[
\]

\[
\]

41
**Theorem 2.4.7:**  
Let $Y = g(X)$ where $X$ is a rv with pdf $f_X(x)$ on support $X$. Suppose there exists a partition $A_0, A_1, \ldots, A_k$ of $X$ such that $P(X \in A_0) = 0$ and $f_X(x)$ is continuous on each $A_i$. Suppose there exist functions $g_1(x), \ldots, g_k(x)$ defined on $A_1$ through $A_k$, respectively, satisfying

(i) $g(x) = g_i(x)$ $\forall \, x \in A_i$,

(ii) $g_i(x)$ is monotonic on $A_i$,

(iii) the set $Y = g_i(A_i) = \{y : y = g_i(x) \text{ for some } x \in A_i\}$ is the same for each $i = 1, \ldots, k$,

and

(iv) $g_i^{-1}(y)$ has a continuous derivative on $Y$ for each $i = 1, \ldots, k$.

Then,

$$f_Y(y) = \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \cdot \left| \frac{d}{dy} g_i^{-1}(y) \right| \cdot I_Y(y)$$

**Example 2.4.8:**

Let $X$ be a rv with pdf $f_X(x) = \frac{2x}{\pi^2} \cdot I_{(0,\pi)}(x)$.

Let $Y = \sin(X)$. What is $f_Y(y)$?

Since $\sin$ is not monotonic on $(0, \pi)$, Theorem 2.4.6 cannot be used to determine the pdf of $Y$.

**Example 2.4.8**

Two possible approaches:
Method 1: cdfs

For $0 < y < 1$ we have

\[ F_Y(y) = P_Y(Y \leq y) \]
\[ = P_X(\sin X \leq y) \]
\[ = P_X([0 \leq X \leq \sin^{-1}(y)] \text{ or } [\pi - \sin^{-1}(y) \leq X \leq \pi]) \]
\[ = F_X(\sin^{-1}(y)) + (1 - F_X(\pi - \sin^{-1}(y))) \]

since $[0 \leq X \leq \sin^{-1}(y)]$ and $[\pi - \sin^{-1}(y) \leq X \leq \pi]$ are disjoint sets. Then,

\[ f_Y(y) = F_Y'(y) \]
\[ = f_X(\sin^{-1}(y)) \frac{1}{\sqrt{1-y^2}} + (-1)f_X(\pi - \sin^{-1}(y)) \frac{-1}{\sqrt{1-y^2}} \]
\[ = \frac{1}{\sqrt{1-y^2}} \left( f_X(\sin^{-1}(y)) + f_X(\pi - \sin^{-1}(y)) \right) \]
\[ = \frac{1}{\sqrt{1-y^2}} \left( \frac{2(\sin^{-1}(y))}{\pi^2} + \frac{2(\pi - \sin^{-1}(y))}{\pi^2} \right) \]
\[ = \frac{1}{\pi^2 \sqrt{1-y^2}} 2\pi \]
\[ = \frac{2\pi}{\pi \sqrt{1-y^2}} \cdot I_{(0,1)}(y) \]

Method 2: Use of Theorem 2.4.7

Let $A_1 = (0, \frac{\pi}{2})$, $A_2 = (\frac{\pi}{2}, \pi)$, and $A_0 = \{\frac{\pi}{2}\}$.

Let $g_1^{-1}(y) = \sin^{-1}(y)$ and $g_2^{-1}(y) = \pi - \sin^{-1}(y)$.

It is $\frac{\partial}{\partial y}g_1^{-1}(y) = \frac{1}{\sqrt{1-y^2}} = -\frac{\partial}{\partial y}g_2^{-1}(y)$ and $\gamma = (0, 1)$.

Thus, by use of Theorem 2.4.7, we get

\[ f_Y(y) = \sum_{i=1}^{2} f_X(g_i^{-1}(y)) \cdot \left| \frac{\partial}{\partial y}g_i^{-1}(y) \right| \cdot I_{\gamma}(y) \]
\[ = \frac{2\sin^{-1}(y)}{\pi^2} \frac{1}{\sqrt{1-y^2}} \cdot I_{(0,1)}(y) + \frac{2(\pi - \sin^{-1}(y))}{\pi^2} \frac{1}{\sqrt{1-y^2}} \cdot I_{(0,1)}(y) \]
\[ = \frac{2\pi}{\pi^2 \sqrt{1-y^2}} \cdot I_{(0,1)}(y) \]
\[ = \frac{2}{\pi \sqrt{1-y^2}} \cdot I_{(0,1)}(y) \]

Obviously, both results are identical.

\[ \square \]
Theorem 2.4.9:
Let $X$ be a rv with a continuous cdf $F_X(x)$ and let $Y = F_X(X)$. Then, $Y \sim U(0, 1)$.

Proof:
We have to consider two possible cases:
(a) $F_X$ is strictly increasing, i.e., $F_X(x_1) < F_X(x_2)$ for $x_1 < x_2$, and
(b) $F_X$ is non-decreasing, i.e., there exists $x_1 < x_2$ and $F_X(x_1) = F_X(x_2)$. Assume that $x_1$ is the infimum and $x_2$ the supremum of those values for which $F_X(x_1) = F_X(x_2)$ holds.

In (a), $F_X^{-1}(y)$ is uniquely defined. In (b), we define $F_X^{-1}(y) = \inf\{x : F_X(x) \geq y\}$

Without loss of generality:
$F_X^{-1}(1) = +\infty$ if $F_X(x) < 1 \ \forall x \in \mathbb{R}$ and
$F_X^{-1}(0) = -\infty$ if $F_X(x) > 0 \ \forall x \in \mathbb{R}$.

For $Y = F_X(X)$ and $0 < y < 1$, we have

$P(Y \leq y) = P(F_X(X) \leq y) \\ \overset{F_X^{-1}}{=} P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) \\ \overset{(\ast)}{=} P(X \leq F_X^{-1}(y)) \\ = F_X(F_X^{-1}(y)) \\ = y$

At the endpoints, we have $P(Y \leq y) = 1$ if $y \geq 1$ and $P(Y \leq y) = 0$ if $y \leq 0$. 


But why is (\ast) true? — In (a), if $F_X$ is strictly increasing and continuous, it is certainly $x = F_X^{-1}(F_X(x))$.

In (b), if $F_X(x_1) = F_X(x_2)$ for $x_1 < x < x_2$, it may be that $F_X^{-1}(F_X(x)) \neq x$. But by definition, $F_X^{-1}(F_X(x)) = x_1 \ \forall x \in [x_1, x_2]$. (\ast) holds since on $[x_1, x_2]$, it is $P(X \leq x) = P(X \leq x_1) \ \forall x \in [x_1, x_2]$. The flat cdf denotes $F_X(x_2) - F_X(x_1) = P(x_1 < X \leq x_2) = 0$ by definition.

**Note:**
This proof also holds if there exist multiple intervals with $x_i < x_j$ and $F_X(x_i) = F_X(x_j)$, i.e., if the support of $X$ is split in more than just 2 disjoint intervals.
3 Moments and Generating Functions

3.1 Expectation

(Based on Casella/Berger, Sections 2.2 & 2.3, and Outside Material)

Definition 3.1.1:
Let $X$ be a real-valued rv with cdf $F_X$ and pdf $f_X$ if $X$ is continuous (or pmf $f_X$ and support $\mathcal{X}$ if $X$ is discrete). The **expected value (mean)** of a measurable function $g(\cdot)$ of $X$ is

$$E(g(X)) = \begin{cases} \int_{-\infty}^{\infty} g(x)f_X(x)dx, & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x)f_X(x), & \text{if } X \text{ is discrete} \end{cases}$$

if $E(|g(X)|) < \infty$; otherwise $E(g(X))$ is undefined, i.e., it does not exist.

Example:
$X \sim \text{Cauchy}$, $f_X(x) = \frac{1}{\pi(1+x^2)}$, $-\infty < x < \infty$:

$$E(|X|) = \frac{2}{\pi} \int_{0}^{\infty} \frac{x}{1+x^2}dx = \frac{1}{\pi} [\log(1+x^2)]_{0}^{\infty} = \infty$$

So, $E(X)$ does not exist for the Cauchy distribution.

Theorem 3.1.2:
If $E(X)$ exists and $a$ and $b$ are finite constants, then $E(aX + b)$ exists and equals $aE(X) + b$.

Proof:
Continuous case only:

Existence:

$$E(|aX + b|) = \int_{-\infty}^{\infty} |ax + b| f_X(x)dx$$

$$\leq \int_{-\infty}^{\infty} (|a| \cdot |x| + |b|) f_X(x)dx$$

$$= |a| \int_{-\infty}^{\infty} |x| f_X(x)dx + |b| \int_{-\infty}^{\infty} f_X(x)dx$$

$$= |a| E(|X|) + |b|$$

$$< \infty$$
Numerical Result:

\[ E(aX + b) = \int_{-\infty}^{\infty} (ax + b) f_X(x) \, dx \]
\[ = a \int_{-\infty}^{\infty} x f_X(x) \, dx + b \int_{-\infty}^{\infty} f_X(x) \, dx \]
\[ = aE(X) + b \]

Theorem 3.1.3:
If \( X \) is bounded (i.e., there exists a \( M, 0 < M < \infty \), such that \( P( |X| < M) = 1 \)), then \( E(X) \) exists.

Definition 3.1.4:
The \( k^{th} \) moment of \( X \), if it exists, is \( m_k = E(X^k) \).

The \( k^{th} \) absolute moment of \( X \), if it exists, is \( \beta_k = E(|X|^k) \).

The \( k^{th} \) central moment of \( X \), if it exists, is \( \mu_k = E((X - E(X))^k) \).

Definition 3.1.5:
The variance of \( X \), if it exists, is the second central moment of \( X \), i.e.,

\[ Var(X) = E((X - E(X))^2) \]

Theorem 3.1.6:
\[ Var(X) = E(X^2) - (E(X))^2 \]

Proof:

\[ Var(X) = E((X - E(X))^2) \]
\[ = E(X^2 - 2XE(X) + (E(X))^2) \]
\[ = E(X^2) - 2E(X)E(X) + (E(X))^2 \]
\[ = E(X^2) - (E(X))^2 \]

Theorem 3.1.7:
If \( Var(X) \) exists and \( a \) and \( b \) are finite constants, then \( Var(aX + b) \) exists and equals \( a^2Var(X) \).
Proof:

Existence & Numerical Result:

\[ \text{Var}(aX + b) = E((aX + b) - E(aX + b))^2 \] exists if \( E(\mid(aX + b) - E(aX + b)\mid^2) \) exists.

It holds that

\[
E(\mid(aX + b) - E(aX + b)\mid^2)
= E((aX + b) - E(aX + b))^2
= \text{Var}(aX + b)
\]

\[ T h . 3 . 1 . 6 \]

\[ E((aX + b)^2) - (E(aX + b))^2 \]

\[ T h . 3 . 1 . 2 \]

\[ E(a^2X^2 + 2abX + b^2) - (aE(X) + b)^2 \]

\[ T h . 3 . 1 . 2 \]

\[ a^2E(X^2) + 2abE(X) + b^2 - a^2(E(X))^2 - 2abE(X) - b^2 \]

\[ = a^2(E(X^2) - (E(X))^2) \]

\[ = a^2\text{Var}(X) \]

\[ < \infty \quad \text{since Var}(X) \text{ exists} \]

Theorem 3.1.8:

If the \( t \)th absolute moment of a rv \( X \) exists for some \( t > 0 \), then all absolute moments of order \( 0 < s < t \) exist.

Proof:

Continuous case only:

\[
E(\mid X \mid^s) = \int_{\mid x \mid \leq 1} \mid x \mid^s f_X(x)dx + \int_{\mid x \mid > 1} \mid x \mid^s f_X(x)dx
\leq \int_{\mid x \mid \leq 1} 1 \cdot f_X(x)dx + \int_{\mid x \mid > 1} \mid x \mid^t f_X(x)dx
\leq P(\mid X \mid \leq 1) + E(\mid X \mid^t)
\leq \infty
\]

Theorem 3.1.9:

If the \( t \)th absolute moment of a rv \( X \) exists for some \( t > 0 \), then

\[ \lim_{n \to \infty} n^t P(\mid X \mid > n) = 0. \]

Proof:

Continuous case only:
\[ \infty > \int_R |x|^t f_X(x)dx = \lim_{n \to \infty} \int_{|x| \leq n} |x|^t f_X(x)dx \]
\[ \implies \lim_{n \to \infty} \int_{|x| > n} |x|^t f_X(x)dx = 0 \]

But, \( \lim_{n \to \infty} \int_{|x| > n} |x|^t f_X(x)dx \geq \lim_{n \to \infty} n^t \int_{|x| > n} f_X(x)dx = \lim_{n \to \infty} n^t P(\mid X \mid > n) = 0 \)

\[ \text{Note:} \]
The inverse is not necessarily true, i.e., if \( \lim_{n \to \infty} n^t P(\mid X \mid > n) = 0 \), then the \( t^{th} \) moment of a rv \( X \) does not necessarily exist. We can only approach \( \theta \) up to some \( \delta > 0 \) as the following Theorem 3.1.10 indicates.

**Theorem 3.1.10:**
Let \( X \) be a rv with a distribution such that \( \lim_{n \to \infty} n^t P(\mid X \mid > n) = 0 \) for some \( t > 0 \). Then,
\[ E(\mid X \mid^s) < \infty \quad \forall \quad 0 < s < t. \]

**Note:**
To prove this Theorem, we need Lemma 3.1.11 and Corollary 3.1.12.

**Lemma 3.1.11:**
Let \( X \) be a non-negative rv with cdf \( F \). Then,
\[ E(X) = \int_0^\infty (1 - F_X(x))dx \]
(if either side exists).

**Proof:**
Continuous case only:

To prove that the left side implies that the right side is finite and both sides are identical, we assume that \( E(X) \) exists. It is
\[ E(X) = \int_0^\infty x f_X(x)dx = \lim_{n \to \infty} \int_0^n x f_X(x)dx \]

Replace the expression for the right side integral using integration by parts.

Let \( u = x \) and \( dv = f_X(x)dx \), then
\[ \int_0^n x f_X(x)dx = (xF_X(x)) \bigg|_0^n - \int_0^n F_X(x)dx \]
\[ = n F_X(n) - 0 F_X(0) - \int_0^n F_X(x)dx \]

49
\[ \begin{align*}
&= nF_X(n) - n + n - \int_0^n F_X(x) \, dx \\
&= nF_X(n) - n + \int_0^n [1 - F_X(x)] \, dx \\
&= n[F_X(n) - 1] + \int_0^n [1 - F_X(x)] \, dx \\
&= -n[1 - F_X(n)] + \int_0^n [1 - F_X(x)] \, dx \\
&= -nP(X > n) + \int_0^n [1 - F_X(x)] \, dx \\
&= -n^1P(|X| > n) + \int_0^n [1 - F_X(x)] \, dx \\
\Rightarrow E(X^1) &= \lim_{n \to \infty} \left( -n^1P(|X| > n) + \int_0^n [1 - F_X(x)] \, dx \right) \\
&= \int_0^\infty [1 - F_X(x)] \, dx
\end{align*} \]

Thus, the existence of \( E(X) \) implies that \( \int_0^\infty [1 - F_X(x)] \, dx \) is finite and that both sides are identical.

We still have to show the converse implication:

If \( \int_0^\infty [1 - F_X(x)] \, dx \) is finite, then \( E(X) \) exists, i.e., \( E(|X|) = E(X) < \infty \), and both sides are identical. It is

\[
\int_0^n x f_X(x) \, dx \overset{X \geq 0}{=} \int_0^n |x| \, f_X(x) \, dx = -n[1 - F_X(n)] + \int_0^n [1 - F_X(x)] \, dx
\]

as seen above.

Since \(-n[1 - F_X(n)] \leq 0\), we get

\[
\int_0^n |x| \, f_X(x) \, dx \leq \int_0^n [1 - F_X(x)] \, dx \leq \int_0^\infty [1 - F_X(x)] \, dx < \infty \quad \forall n
\]

Thus,

\[
\lim_{n \to \infty} \int_0^n |x| \, f_X(x) \, dx = \int_0^\infty |x| \, f_X(x) \, dx \leq \int_0^\infty [1 - F_X(x)] \, dx < \infty
\]

\( \Rightarrow E(X) \) exists and is identical to \( \int_0^\infty [1 - F_X(x)] \, dx \) as seen above. 

\( \blacksquare \)
Corollary 3.1.12:

\[ E(|X|^s) = s \int_0^\infty y^{s-1}P(|X| > y)dy \]

Proof:

\[ E(|X|^s) \overset{\text{Cor. 3.1.12}}{=} \int_0^\infty [1 - F_{|X|^s}(z)]dz = \int_0^\infty P(|X|^s > z)dz \]

Let \( z = y^s \). Then \( \frac{dz}{dy} = sy^{s-1} \) and \( dz = sy^{s-1}dy \). Therefore,

\[
\int_0^\infty P(|X|^s > z)dz = \int_0^\infty P(|X|^s > y^s)sy^{s-1}dy
\]

\[
= s \int_0^\infty y^{s-1}P(|X| > y^s)dy
\]

\[
= s \int_0^\infty y^{s-1}P(|X| > y)dy
\]

\[
\text{monotonic} \uparrow
\]

Proof (of Theorem 3.1.10):

For any given \( \epsilon > 0 \), choose \( N \) such that the tail probability \( P(|X| > n) < \frac{\epsilon}{n^t} \forall n \geq N \).

\[
E(|X|^s) \overset{\text{Cor. 3.1.12}}{=} s \int_0^\infty y^{s-1}P(|X| > y)dy
\]

\[
= s \int_0^N y^{s-1}P(|X| > y)dy + s \int_N^\infty y^{s-1}P(|X| > y)dy
\]

\[
\leq \int_0^N sy^{s-1}1dy + s \int_N^\infty y^{s-1}\epsilon y^{-t}dy
\]

\[
= y^s \mid_0^N + s \epsilon \int_N^\infty y^{s-1-1}dy
\]

\[
= N^s + s \epsilon \int_N^\infty y^{s-1-t}dy
\]

It is

\[
\int_N^\infty y^c dy = \begin{cases} \frac{1}{c+1}y^{c+1} |^\infty_N, & c \neq -1 \\ \ln y |^\infty_N, & c = -1 \end{cases}
\]

\[
= \begin{cases} \infty, & c \geq -1 \\ -\frac{1}{c+1}N^{c+1} < \infty, & c < -1 \end{cases}
\]

Thus, for \( E(|X|^s) < \infty \), it must hold that \( s - 1 - t < -1 \), or equivalently, \( s < t \). So \( E(|X|^s) < \infty \), i.e., it exists, for every \( s \) with \( 0 < s < t \) for a rv \( X \) with a distribution such that \( \lim_{n \to \infty} n^t P(|X| > n) = 0 \) for some \( t > 0 \).
Theorem 3.1.13.
Let $X$ be a rv such that
\[
\lim_{k \to \infty} \frac{P(|X| > \alpha k)}{P(|X| > k)} = 0 \quad \forall \alpha > 1.
\]
Then, all moments of $X$ exist.

Proof:
• For $\epsilon > 0$, we select some $k_0$ such that
\[
P(|X| > \alpha k) < \epsilon \quad \forall k \geq k_0.
\]
• Select $k_1$ such that $P(|X| > k) < \epsilon \quad \forall k \geq k_1$.
• Select $N = \max(k_0, k_1)$.
• If we have some fixed positive integer $r$:
\[
\frac{P(|X| > \alpha^r k)}{P(|X| > k)} = \frac{P(|X| > \alpha k)}{P(|X| > k)} \cdot \frac{P(|X| > \alpha^2 k)}{P(|X| > \alpha k)} \cdot \frac{P(|X| > \alpha^3 k)}{P(|X| > \alpha^2 k)} \cdots \frac{P(|X| > \alpha^r k)}{P(|X| > \alpha^{r-1} k)} = \frac{P(|X| > \alpha k)}{P(|X| > k)} \cdot \frac{P(|X| > \alpha \cdot (a k))}{P(|X| > 1 \cdot (a k))} \cdot \frac{P(|X| > \alpha \cdot (\alpha^2 k))}{P(|X| > 1 \cdot (\alpha^2 k))} \cdots \frac{P(|X| > \alpha \cdot (\alpha^{r-1} k))}{P(|X| > 1 \cdot (\alpha^{r-1} k))}
\]
• Note: Each of these $r$ terms on the right side is $< \epsilon$ by our original statement of selecting some $k_0$ such that $P(|X| > \alpha k) < \epsilon$ and since $\alpha > 1$ and therefore $\alpha^\alpha k \geq k_0$.
• Now we get for our entire expression that $\frac{P(|X| > \alpha^r k)}{P(|X| > k)} \leq \epsilon^r$ for $k \geq N$ (since in this case also $k \geq k_0$) and $\alpha > 1$.
• Overall, we have $P(|X| > \alpha^r k) \leq \epsilon^r P(|X| > k) \leq \epsilon^{r+1}$ for $k \geq N$ (since in this case also $k \geq k_1$).

For a fixed positive integer $n$:
\[
E(|X|^n) = n \int_0^\infty x^{n-1} P(|X| > x)dx = n \int_0^N x^{n-1} P(|X| > x)dx + n \int_N^\infty x^{n-1} P(|X| > x)dx.
\]
• We know that:
\[
n \int_0^N x^{n-1} P(|X| > x)dx \leq \int_0^N nx^{n-1}dx = x^n \bigg|_0^N = N^n < \infty
\]
but is
\[
n \int_N^\infty x^{n-1} P(|X| > x)dx < \infty ?
\]
• To check the second part, we use:
\[
\int_N^\infty x^{n-1} P(|X| > x)dx = \sum_{r=1}^\infty \alpha^r \int_{\alpha^{r-1} N}^{\alpha^r N} x^{n-1} P(|X| > x)dx
\]
• We know that:
\[
\alpha^r N \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1}P(\| X \| > x)dx \leq \epsilon^r \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1}dx
\]
This step is possible since \( \epsilon^r \geq P(\| X \| \geq \alpha^{r-1}N) \geq P(\| X \| > x) \geq P(\| X \| \geq \alpha^r N) \)
\( \forall x \in (\alpha^{r-1}N, \alpha^r N) \) and \( N = \max(k_0, k_1) \).

• Since \((\alpha^{r-1}N)^{n-1} \leq x^{n-1} \leq (\alpha^r N)^{n-1} \forall x \in (\alpha^{r-1}N, \alpha^r N)\), we get:
\[
\epsilon^r \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1}dx \leq \epsilon^r (\alpha^r N)^{n-1} \int_{\alpha^{r-1}N}^{\alpha^r N} 1dx \leq \epsilon^r (\alpha^r N)^{n-1} (\alpha^r N) = \epsilon^r (\alpha^r N)^n
\]

• Now we go back to our original inequality:
\[
\int_{N}^{\infty} x^{n-1}P(\| X \| > x)dx \leq \sum_{r=1}^{\infty} \epsilon^r \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1}dx \leq \sum_{r=1}^{\infty} \epsilon^r (\alpha^r N)^n = N^n \sum_{r=1}^{\infty} (\epsilon \cdot \alpha^n)^r
\]
\[
= \frac{N^n \epsilon \alpha^n}{1 - \epsilon \alpha^n} \text{ if } \epsilon \alpha^n < 1 \text{ or, equivalently, if } \epsilon < \frac{1}{\alpha^n}
\]
• Since \( \frac{N^n \epsilon \alpha^n}{1 - \epsilon \alpha^n} \) is finite, all moments \( E(\| X \|^n) \) exist.
3.2 Moment Generating Functions

(Based on Casella/Berger, Sections 2.3 & 2.4)

Definition 3.2.1:
Let $X$ be a rv with cdf $F_X$. The moment generating function (mgf) of $X$ is defined as

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

provided that this expectation exists for $t$ in an (open) interval around 0, i.e., for $-h < t < h$ for some $h > 0$.

Theorem 3.2.2:
If a rv $X$ has a mgf $M_X(t)$ that exists for $-h < t < h$ for some $h > 0$, then

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

Proof:
We assume that we can differentiate under the integral sign. If, and when, this really is true will be discussed later in this section.

$$\frac{d}{dt}M_X(t) = \frac{d}{dt}\int_{-\infty}^{\infty}e^{tx}f_X(x)dx$$

$$= \int_{-\infty}^{\infty}\frac{\partial}{\partial t}e^{tx}f_X(x)dx$$

$$= \int_{-\infty}^{\infty}xe^{tx}f_X(x)dx$$

$$= E(Xe^{tX})$$

Evaluating this at $t = 0$, we get: $\frac{d}{dt}M_X(t) \big|_{t=0} = E(X)$

By induction, we get for $n \geq 2$:

$$\frac{d^n}{dt^n}M_X(t) = \frac{d}{dt}\left(\frac{d^{n-1}}{dt^{n-1}}M_X(t)\right)$$

$$= \frac{d}{dt}\left(\int_{-\infty}^{\infty}x^{n-1}e^{tx}f_X(x)dx\right)$$

$$= \int_{-\infty}^{\infty}\frac{\partial}{\partial t}x^{n-1}e^{tx}f_X(x)dx$$

$$= \int_{-\infty}^{\infty}x^{n}e^{tx}f_X(x)dx$$

$$= E(X^n e^{tX})$$

Evaluating this at $t = 0$, we get: $\frac{d^n}{dt^n}M_X(t) \big|_{t=0} = E(X^n)$
Note:
We use the notation $\frac{\partial}{\partial t} f(x, t)$ for the partial derivative of $f$ with respect to $t$ and the notation $\frac{d}{dt} f(t)$ for the (ordinary) derivative of $f$ with respect to $t$. 

Example 3.2.3:
$X \sim U(a, b)$, where $a < b$; $f_X(x) = \frac{1}{b-a} \cdot I_{[a,b]}(x)$.

Then,

$$M_X(t) = \int_a^b \frac{e^{tx}}{b-a} dx = \frac{e^{tb} - e^{ta}}{t(b-a)}$$

$$M_X(0) = 0$$

L'Hospital:

$$\left. \frac{be^{tb} - ae^{ta}}{b-a} \right|_{t=0} = 1$$

So $M_X(0) = 1$ and since $\frac{e^{tb} - e^{ta}}{t(b-a)}$ is continuous, it also exists in an open interval around 0 (in fact, it exists for every $t \in \mathbb{R}$).

$$M'_X(t) = \frac{(be^{tb} - ae^{ta})t(b-a) - (e^{tb} - e^{ta})(b-a)}{t^2(b-a)^2}$$

$$= \frac{t(be^{tb} - ae^{ta}) - (e^{tb} - e^{ta})}{t^2(b-a)}$$

$$\implies E(X) = M'_X(0) = 0$$

L'Hospital:

$$\left. \frac{be^{tb} - ae^{ta} + tb^2 e^{tb} - ta^2 e^{ta} - be^{tb} + ae^{ta}}{2t(b-a)} \right|_{t=0}$$

$$= \left. \frac{tb^2 e^{tb} - ta^2 e^{ta}}{2t(b-a)} \right|_{t=0}$$

$$= \left. \frac{b^2 e^{tb} - a^2 e^{ta}}{2(b-a)} \right|_{t=0}$$

$$= \left. \frac{b^2 - a^2}{2(b-a)} \right|_{t=0}$$

$$= \frac{b + a}{2}$$

55
Note:
In the previous example, we made use of L’Hospital’s rule. This rule gives conditions under which we can resolve indefinite expressions of the type “\(\frac{\pm 0}{\pm 0}\)” and “\(\frac{\pm \infty}{\pm \infty}\).

(i) Let \(f\) and \(g\) be functions that are differentiable in an open interval around \(x_0\), say in \((x_0 - \delta, x_0 + \delta)\), but not necessarily differentiable in \(x_0\). Let \(f(x_0) = g(x_0) = 0\) and \(g'(x) \neq 0 \forall x \in (x_0 - \delta, x_0 + \delta) - \{x_0\}\). Then, \(\lim_{x \to x_0} \frac{f'(x)}{g'(x)} = A\) implies that also \(\lim_{x \to x_0} \frac{f(x)}{g(x)} = A\). The same holds for the cases \(\lim_{x \to x_0^+} f(x) = \lim_{x \to x_0^-} g(x) = \infty\) and \(x \rightarrow x_0^+\) or \(x \rightarrow x_0^-\).

(ii) Let \(f\) and \(g\) be functions that are differentiable for \(x > a\) (\(a > 0\)). Let \(\lim_{x \to \infty} f(x) = \lim_{x \to \infty} g(x) = 0\) and \(\lim_{x \to \infty} g'(x) \neq 0\). Then, \(\lim_{x \to \infty} \frac{f'(x)}{g'(x)} = A\) implies that also \(\lim_{x \to \infty} \frac{f(x)}{g(x)} = A\).

(iii) We can iterate this process as long as the required conditions are met and derivatives exist, e.g., if the first derivatives still result in an indefinite expression, we can look at the second derivatives, then at the third derivatives, and so on.

(iv) It is recommended to keep expressions as simple as possible. If we have identical factors in the numerator and denominator, we can exclude them from both and continue with the simpler functions.

(v) Indefinite expressions of the form “\(0 \cdot \infty\)” can be handled by rearranging them to “\(\frac{0}{\frac{1}{\infty}}\)” and \(\lim_{x \to -\infty} \frac{f(x)}{g(x)}\) can be handled by use of the rules for \(\lim_{x \to -\infty} \frac{f(-x)}{g(-x)}\).

Note:
The following Theorems provide us with rules that tell us when we can differentiate under the integral sign. Theorem 3.2.4 relates to finite integral bounds \(a(\theta)\) and \(b(\theta)\) and Theorems 3.2.5 and 3.2.6 to infinite bounds.

**Theorem 3.2.4: Leibnitz’s Rule**
If \(f(x, \theta), a(\theta), \text{ and } b(\theta)\) are differentiable with respect to \(\theta\) (for all \(x\)) and \(-\infty < a(\theta) < b(\theta) < \infty\), then
\[
\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) \, dx = \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x, \theta) \, dx + \frac{d}{d\theta} f(b(\theta), \theta) - \frac{d}{d\theta} f(a(\theta), \theta).
\]
The first 2 terms are vanishing if \(a(\theta)\) and \(b(\theta)\) are constant in \(\theta\).
Proof:
Uses the Fundamental Theorem of Calculus and the chain rule.

Theorem 3.2.5: Lebesque’s Dominated Convergence Theorem

Let $g$ be an integrable function such that $\int_{-\infty}^{\infty} g(x)dx < \infty$. If $|f_n| \leq g$ almost everywhere (i.e., except for a set of Borel–measure 0) and if $f_n \to f$ almost everywhere, then $f_n$ and $f$ are integrable and

$$\int_{-\infty}^{\infty} f_n(x)dx \to \int_{-\infty}^{\infty} f(x)dx.$$  

Note:
If $f$ is differentiable with respect to $\theta$, then

$$\frac{\partial}{\partial \theta} f(x, \theta) = \lim_{\delta \to 0} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta}$$

and

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta)dx = \int_{-\infty}^{\infty} \lim_{\delta \to 0} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta} dx$$

while

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)dx = \lim_{\delta \to 0} \int_{-\infty}^{\infty} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta} dx$$

Theorem 3.2.6:
Let $f_n(x, \theta_0) = \frac{f(x, \theta_0 + \delta_n) - f(x, \theta_0)}{\delta_n}$ for some $\theta_0$. Suppose there exists an integrable function $g(x)$ such that $\int_{-\infty}^{\infty} g(x)dx < \infty$ and $|f_n(x, \theta)| \leq g(x)$ $\forall x$, then

$$\left[ \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)dx \right]_{\theta=\theta_0} = \int_{-\infty}^{\infty} \left[ \frac{\partial}{\partial \theta} f(x, \theta) \right]_{\theta=\theta_0} dx.$$  

Usually, if $f$ is differentiable for all $\theta$, we write

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta)dx.$$  

Corollary 3.2.7:
Let $f(x, \theta)$ be differentiable for all $\theta$. Suppose there exists an integrable function $g(x, \theta)$ such that $\int_{-\infty}^{\infty} g(x, \theta)dx < \infty$ and $|\frac{\partial}{\partial \theta} f(x, \theta)|_{\theta=\theta_0} \leq g(x, \theta)$ $\forall x$ $\forall \theta_0$ in some $\epsilon$-neighborhood of $\theta$, then

$$\frac{d}{d\theta} \int_{-\infty}^{\infty} f(x, \theta)dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta)dx.$$
More on Moment Generating Functions

Consider
\[ \left| \frac{\partial}{\partial t} e^{tx} f_X(x) \right|_{t=t'} = |x| e^{t'x} f_X(x) \text{ for } |t' - t| \leq \delta_0. \]

Choose \( t, \delta_0 \) small enough such that \( t + \delta_0 \in (-h, h) \) and \( t - \delta_0 \in (-h, h) \), or, equivalently, \( |t + \delta_0| < h \) and \( |t - \delta_0| < h \).

Section 3.2: MGF

Then,
\[ \left| \frac{\partial}{\partial t} e^{tx} f_X(x) \right|_{t=t'} \leq g(x, t) \]

where
\[ g(x, t) = \begin{cases} 
|x| e^{(t+\delta_0)x} f_X(x), & x \geq 0 \\
|x| e^{(t-\delta_0)x} f_X(x), & x < 0 
\end{cases} \]

To verify \( \int g(x, t)dx < \infty \), we need to know \( f_X(x) \).

Suppose mgf \( M_X(t) \) exists for \( |t| \leq h^* \) for some \( h^* > 1 \), where \( h^*-1 \geq h \). Then \( |t+\delta_0+1| < h^* \) and \( |t - \delta_0 - 1| < h^* \). Since \( |x| \leq e^{\|x\|} \forall x \), we get
\[ g(x, t) \leq \begin{cases} 
e^{(t+\delta_0+1)x} f_X(x), & x \geq 0 \\
e^{(t-\delta_0-1)x} f_X(x), & x < 0 
\end{cases} \]

Then, \( \int_0^\infty g(x, t)dx \leq M_X(t + \delta_0 + 1) < \infty \) and \( \int_0^0 g(x, t)dx \leq M_X(t - \delta_0 - 1) < \infty \) and, therefore, \( \int_{-\infty}^\infty g(x)dx < \infty \).

Together with Corollary 3.2.7, this establishes that we can differentiate under the integral in the Proof of Theorem 3.2.2.

If \( h^* \leq 1 \), we may need to check more carefully to see if the condition holds.
Note:
If $M_X(t)$ exists for $t \in (-h, h)$, then we have an infinite collection of moments.

Does a collection of integer moments $\{m_k : k = 1, 2, 3, \ldots\}$ completely characterize the distribution, i.e., cdf, of $X$? — Unfortunately not, as Example 3.2.8 shows.

Example 3.2.8:
Let $X_1$ and $X_2$ be rv’s with pdfs

$$f_{X_1}(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp\left(-\frac{1}{2}(\log x)^2\right) \cdot I_{(0,\infty)}(x)$$

and

$$f_{X_2}(x) = f_{X_1}(x) \cdot (1 + \sin(2\pi \log x)) \cdot I_{(0,\infty)}(x)$$

It is $E(X_1^r) = E(X_2^r) = e^{r^2/2}$ for $r = 0, 1, 2, \ldots$ as you have to show in the Homework.

Two different pdfs/cdfs have the same moment sequence! What went wrong? In this example, $M_{X_1}(t)$ does not exist as shown in the Homework!

Theorem 3.2.9:
Let $X$ and $Y$ be 2 rv’s with cdf’s $F_X$ and $F_Y$ for which all moments exist.

(i) If $F_X$ and $F_Y$ have bounded support, then $F_X(u) = F_Y(u)$ $\forall u$ iff $E(X^r) = E(Y^r)$ for $r = 0, 1, 2, \ldots$.

(ii) If both mgf’s exist, i.e., $M_X(t) = M_Y(t)$ for $t$ in some neighborhood of 0, then $F_X(u) = F_Y(u)$ $\forall u$.

Note:
The existence of moments is not equivalent to the existence of a mgf as seen in Example 3.2.8 above and some of the Homework assignments.

Theorem 3.2.10:
Suppose rv’s $\{X_i\}_{i=1}^\infty$ have mgf’s $M_{X_i}(t)$ and that $\lim_{i \to \infty} M_{X_i}(t) = M_X(t)$ $\forall t \in (-h, h)$ for some $h > 0$ and that $M_X(t)$ itself is a mgf. Then, there exists a cdf $F_X$ whose moments are determined by $M_X(t)$ and for all continuity points $x$ of $F_X(x)$ it holds that $\lim_{i \to \infty} F_{X_i}(x) = F_X(x)$, i.e., the convergence of mgf’s implies the convergence of cdf’s.

Proof:
Uniqueness of Laplace transformations, etc.
Theorem 3.2.11:
For constants $a$ and $b$, the mgf of $Y = aX + b$ is

$$M_Y(t) = e^{bt} M_X(at),$$
given that $M_X(t)$ exists.

Proof:

$$M_Y(t) = E(e^{(aX+b)t})$$
$$= E(e^{aXt} e^{bt})$$
$$= e^{bt} E(e^{Xat})$$
$$= e^{bt} M_X(at)$$
3.3 Complex–Valued Random Variables and Characteristic Functions

(Based on Casella/Berger, Section 2.6, and Outside Material)

Recall the following facts regarding complex numbers:

\[ i^0 = +1; \quad i = \sqrt{-1}; \quad i^2 = -1; \quad i^3 = -i; \quad i^4 = +1; \quad \text{etc.} \]

in the planar Gaussian number plane it holds that \( i = (0, 1) \)

\[ z = a + ib = r(\cos \phi + i \sin \phi) \]
\[ r = |z| = \sqrt{a^2 + b^2} \]
\[ \tan \phi = \frac{b}{a} \]

Euler’s Relation: \( z = r(\cos \phi + i \sin \phi) = re^{i\phi} \)

Mathematical Operations on Complex Numbers:

\[ z_1 \pm z_2 = (a_1 \pm a_2) + i(b_1 \pm b_2) \]
\[ z_1 \cdot z_2 = r_1 r_2 e^{i(\phi_1 + \phi_2)} = r_1 r_2 (\cos(\phi_1 + \phi_2) + i \sin(\phi_1 + \phi_2)) \]
\[ \frac{z_1}{z_2} = \frac{r_1}{r_2} e^{i(\phi_1 - \phi_2)} = \frac{r_1}{r_2} (\cos(\phi_1 - \phi_2) + i \sin(\phi_1 - \phi_2)) \]

Moivre’s Theorem: \( z^n = (r(\cos \phi + i \sin \phi))^n = r^n (\cos(n\phi) + i \sin(n\phi)) \)

\[ \sqrt[n]{z} = \sqrt[n]{a + ib} = \sqrt[n]{r} \left( \cos\left(\frac{\phi + k \cdot 2\pi}{n}\right) + i \sin\left(\frac{\phi + k \cdot 2\pi}{n}\right) \right) \]
for \( k = 0, 1, \ldots, (n-1) \) and the main value is obtained for \( k = 0 \)

\[ \ln z = \ln(|z|) + i\phi \pm ik \cdot 2\pi \] where \( \phi = \arctan \frac{b}{a} \), \( k = 0, \pm 1, \pm 2, \ldots \), and the main value is obtained for \( k = 0 \)
Note:
Similar to real numbers, where we define $\sqrt{4} = 2$ while it holds that $2^2 = 4$ and $(-2)^2 = 4$, the $n^{th}$ root and also the logarithm of complex numbers have one main value. However, if we read $n^{th}$ root and logarithm as mappings where the inverse mappings (power and exponential function) yield the original values again, there exist additional solutions that produce the original values. For example, the main value of $\sqrt{-1}$ is $i$. However, it holds that $i^2 = -1$ and $(-i)^2 = (-1)^2 i^2 = i^2 = i$. So, all solutions to $\sqrt{-1}$ are $\{i, -i\}$.

Conjugate Complex Numbers:
For $z = a + ib$, we define the conjugate complex number $\overline{z} = a - ib$. It holds:

$\overline{\overline{z}} = z$

$z = \overline{z}$ iff $z \in IR$

$\overline{z_1 \pm z_2} = \overline{z_1} \pm \overline{z_2}$

$\overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2}$

$\overline{\left( \frac{z_1}{z_2} \right)} = \frac{\overline{z_1}}{\overline{z_2}}$

$z \cdot \overline{z} = a^2 + b^2$

$Re(z) = a = \frac{1}{2}(z + \overline{z})$

$Im(z) = b = \frac{1}{2i}(z - \overline{z})$

$|z| = \sqrt{a^2 + b^2} = \sqrt{z \cdot \overline{z}}$

Definition 3.3.1:
Let $(\Omega, L, P)$ be a probability space and $X$ and $Y$ real–valued rv’s, i.e., $X, Y : (\Omega, L) \rightarrow (IR, B)$

(i) $Z = X + iY : (\Omega, L) \rightarrow (C, B_C)$ is called a complex–valued random variable (C–rv).

(ii) If $E(X)$ and $E(Y)$ exist, then $E(Z)$ is defined as $E(Z) = E(X) + iE(Y) \in C$.

Note:
$E(Z)$ exists iff $E(|X|)$ and $E(|Y|)$ exist. It also holds that if $E(Z)$ exists, then

$|E(Z)| \leq E(|Z|)$

(see Homework).
Definition 3.3.2:  
Let $X$ be a real–valued rv on $(\Omega, L, P)$. Then, $\Phi_X(t) : \mathbb{R} \to \mathbb{C}$ with $\Phi_X(t) = E(e^{itX})$ is called the **characteristic function** of $X$.

Note:

(i) $\Phi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x) dx = \int_{-\infty}^{\infty} \cos(tx) f_X(x) dx + i \int_{-\infty}^{\infty} \sin(tx) f_X(x) dx$

if $X$ is continuous.

(ii) $\Phi_X(t) = \sum_{x \in \mathcal{X}} e^{itx} P(X = x) = \sum_{x \in \mathcal{X}} \cos(tx) P(X = x) + i \sum_{x \in \mathcal{X}} \sin(tx) P(X = x)$

if $X$ is discrete and $\mathcal{X}$ is the support of $X$.

(iii) $\Phi_X(t)$ exists for all real–valued rv's $X$ since $|e^{itx}| = 1$.

Theorem 3.3.3:  
Let $\Phi_X$ be the characteristic function of a real–valued rv $X$. Then it holds:

(i) $\Phi_X(0) = 1$.

(ii) $|\Phi_X(t)| \leq 1 \ \forall t \in \mathbb{R}$.

(iii) $\Phi_X$ is uniformly continuous, i.e., $\forall \epsilon > 0 \ \exists \delta > 0 \ \forall t_1, t_2 \in \mathbb{R} : |t_1 - t_2| < \delta \Rightarrow |\Phi(t_1) - \Phi(t_2)| < \epsilon$.

(iv) $\Phi_X$ is a positive definite function, i.e., $\forall n \in \mathbb{N} \ \forall \alpha_1, \ldots, \alpha_n \in \mathbb{C} \ \forall t_1, \ldots, t_n \in \mathbb{R} :$

$$\sum_{l=1}^{n} \sum_{j=1}^{n} \alpha_l \overline{\alpha_j} \Phi_X(t_l - t_j) \geq 0.$$

(v) $\Phi_X(t) = \overline{\Phi_X(-t)}$.

(vi) If $X$ is symmetric around 0, i.e., if $X$ has a pdf that is symmetric around 0, then $\Phi_X(t) \in \mathbb{R} \ \forall t \in \mathbb{R}$.

(vii) $\Phi_{aX+b}(t) = e^{ibt} \Phi_X(at)$.

Proof:

See Homework for parts (i), (ii), (iv), (v), (vi), and (vii).
Part (iii):
Known conditions:

(i) Let $\epsilon > 0$.

(ii) $\exists a > 0 : P(-a < X < +a) > 1 - \frac{\epsilon}{4}$ and $P(|X| \geq a) \leq \frac{\epsilon}{4}$

(iii) $\exists \delta > 0 : |e^{i(t'-t)x} - 1| < \frac{\epsilon}{2}$ $\forall x$ s.t. $|x| < a$ and $\forall (t' - t)$ s.t. $0 < (t' - t) < \delta$.

This third condition holds since $|e^{i0} - 1| = 0$ and the exponential function is continuous.

Therefore, if we select $(t' - t)$ and $x$ small enough, $|e^{i(t'-t)x} - 1|$ will be $< \frac{\epsilon}{2}$ for a given $\epsilon$.

Let $t, t' \in \mathbb{R}$, $t < t'$, and $t' - t < \delta$. Then,

$$|\Phi_X(t') - \Phi_X(t)| = \int_{-\infty}^{+\infty} e^{it'x} f_X(x)dx - \int_{-\infty}^{+\infty} e^{itx} f_X(x)dx$$

$$= |\int_{-\infty}^{+\infty} (e^{it'x} - e^{itx}) f_X(x)dx|$$

$$= |\int_{-\infty}^{-a} (e^{it'x} - e^{itx}) f_X(x)dx + \int_{-a}^{+a} (e^{it'x} - e^{itx}) f_X(x)dx + \int_{a}^{+\infty} (e^{it'x} - e^{itx}) f_X(x)dx|$$

$$\leq |\int_{-\infty}^{-a} (e^{it'x} - e^{itx}) f_X(x)dx| + |\int_{-a}^{+a} (e^{it'x} - e^{itx}) f_X(x)dx|$$

$$+ |\int_{a}^{+\infty} (e^{it'x} - e^{itx}) f_X(x)dx|$$

We now take a closer look at the first and third of these absolute integrals. It is:

$$|\int_{-\infty}^{-a} (e^{it'x} - e^{itx}) f_X(x)dx| = |\int_{-\infty}^{-a} e^{it'x} f_X(x)dx - \int_{-\infty}^{-a} e^{itx} f_X(x)dx|$$

$$\leq |\int_{-\infty}^{-a} e^{it'x} f_X(x)dx| + |\int_{-\infty}^{-a} e^{itx} f_X(x)dx|$$

$$\leq \int_{-\infty}^{-a} |e^{it'x} f_X(x)dx + \int_{-\infty}^{-a} |e^{itx} f_X(x)dx|$$

$$\overset{(A)}{=} \int_{-\infty}^{-a} 1 f_X(x)dx + \int_{-\infty}^{-a} 1 f_X(x)dx$$

$$= \int_{-\infty}^{-a} 2 f_X(x)dx.$$

(A) holds due to Note (iii) that follows Definition 3.3.2.
Similarly,
\[ | \int_{-a}^{+\infty} (e^{it'x} - e^{itx}) f_X(x) dx | \leq \int_{-a}^{+\infty} 2f_X(x) dx \]

Returning to the main part of the proof, we get
\[ | \Phi_X(t') - \Phi_X(t) | \leq \int_{-\infty}^{-a} 2f_X(x) dx + | \int_{-a}^{+\infty} (e^{it'x} - e^{itx}) f_X(x) dx | + \int_{+a}^{+\infty} 2f_X(x) dx \]

\[ = 2 \left( \int_{-\infty}^{-a} f_X(x) dx + \int_{+a}^{+\infty} f_X(x) dx \right) + \int_{-a}^{+a} (e^{it'x} - e^{itx}) f_X(x) dx | \]

\[ = 2P(| X | \geq a) + | \int_{-a}^{+a} (e^{it'x} - e^{itx}) f_X(x) dx | \]

Condition (ii) ≤ \[ \frac{\epsilon}{2} + \int_{-a}^{+a} | e^{itx} (e^{i(t'-t)x} - 1) | f_X(x) dx | \]

\[ \leq \frac{\epsilon}{2} + \int_{-a}^{+a} | e^{itx} | \cdot | (e^{i(t'-t)x} - 1) | f_X(x) dx \]

\[ \leq \frac{\epsilon}{2} + \int_{-a}^{+a} \frac{\epsilon}{2} f_X(x) dx \]

\[ \leq \frac{\epsilon}{2} + \int_{-\infty}^{+\infty} \frac{\epsilon}{2} f_X(x) dx \]

\[ = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \]

(B) holds due to Note (iii) that follows Definition 3.3.2 and due to condition (iii).

Theorem 3.3.4: Bochner’s Theorem
Let \( \Phi : \mathbb{R} \to \mathcal{C} \) be any function with properties (i), (ii), (iii), and (iv) from Theorem 3.3.3. Then there exists a real-valued rv \( X \) with \( \Phi_X = \Phi \).
Theorem 3.3.5:
Let $X$ be a real-valued rv and $E(X^k)$ exists for an integer $k$. Then, $\Phi_X$ is $k$ times differentiable and $\Phi_X^{(k)}(t) = i^k E(X^k e^{itX})$. In particular for $t = 0$, it is $\Phi_X^{(k)}(0) = i^k m_k$. ■

Theorem 3.3.6:
Let $X$ be a real-valued rv with characteristic function $\Phi_X$ and let $\Phi_X$ be $k$ times differentiable, where $k$ is an even integer. Then the $k^{th}$ moment of $X$, $m_k$, exists and it is $\Phi_X^{(k)}(0) = i^k m_k$. ■

Theorem 3.3.7: **Levy’s Theorem**
Let $X$ be a real-valued rv with cdf $F_X$ and characteristic function $\Phi_X$. Let $a, b \in \mathbb{R}$, $a < b$. If $P(X = a) = P(X = b) = 0$, i.e., $F_X$ is continuous in $a$ and $b$, then

$$ F(b) - F(a) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \Phi_X(t) dt. $$

Theorem 3.3.8:
Let $X$ and $Y$ be a real-valued rv with characteristic functions $\Phi_X$ and $\Phi_Y$. If $\Phi_X = \Phi_Y$, then $X$ and $Y$ are identically distributed. ■

Theorem 3.3.9:
Let $X$ be a real-valued rv with characteristic function $\Phi_X$ such that $\int_{-\infty}^{\infty} |\Phi_X(t)| dt < \infty$. Then $X$ has pdf

$$ f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \Phi_X(t) dt. $$

Theorem 3.3.10:
Let $X$ be a real-valued rv with mgf $M_X(t)$, i.e., the mgf exists. Then $\Phi_X(t) = M_X(it)$. ■

Theorem 3.3.11:
Suppose real-valued rv’s $\{X_i\}_{i=1}^{\infty}$ have cdf’s $\{F_{X_i}\}_{i=1}^{\infty}$ and characteristic functions $\{\Phi_{X_i}(t)\}_{i=1}^{\infty}$. If $\lim_{i \to \infty} \Phi_{X_i}(t) = \Phi_X(t)$ $\forall t \in (-h, h)$ for some $h > 0$ and $\Phi_X(t)$ is itself a characteristic function (of a rv $X$ with cdf $F_X$), then

$$ \lim_{i \to \infty} F_{X_i}(x) = F_X(x) $$

for all continuity points $x$ of $F_X(x)$, i.e., the convergence of characteristic functions implies the convergence of cdf’s. ■
Theorem 3.3.12.
Characteristic functions for some well-known distributions:

<table>
<thead>
<tr>
<th>Distribution</th>
<th>( \Phi_X(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i) ( X \sim \text{Dirac}(c) )</td>
<td>( e^{itc} )</td>
</tr>
<tr>
<td>(ii) ( X \sim \text{Bin}(1, p) )</td>
<td>( 1 + p(e^{it} - 1) )</td>
</tr>
<tr>
<td>(iii) ( X \sim \text{Poisson}(c) )</td>
<td>( \exp(c(e^{it} - 1)) )</td>
</tr>
<tr>
<td>(iv) ( X \sim \text{U}(a, b) )</td>
<td>( \frac{e^{ibt} - e^{ita}}{(b-a)it} )</td>
</tr>
<tr>
<td>(v) ( X \sim \text{N}(0, 1) )</td>
<td>( \exp(-\frac{t^2}{2}) )</td>
</tr>
<tr>
<td>(vi) ( X \sim \text{N}(\mu, \sigma^2) )</td>
<td>( \exp(-\sigma^2 \frac{t^2}{2}) )</td>
</tr>
<tr>
<td>(vii) ( X \sim \text{\Gamma}(p, q) )</td>
<td>( (1 - \frac{it}{q})^{-p} )</td>
</tr>
<tr>
<td>(viii) ( X \sim \text{Exp}(c) )</td>
<td>( (1 - \frac{it}{c})^{-1} )</td>
</tr>
<tr>
<td>(ix) ( X \sim \chi^2_n )</td>
<td>( (1 - 2it)^{-n/2} )</td>
</tr>
</tbody>
</table>

Proof:

(i) \( \Phi_X(t) = E(e^{itX}) = e^{itc}P(X = c) = e^{itc} \)

(ii) \( \Phi_X(t) = \sum_{k=0}^{1} \frac{e^{ikt} P(X = k)}{(1-p) + p(e^{it} - 1)} = 1 + p(e^{it} - 1) \)

(iii) \( \Phi_X(t) = \sum_{n \in \mathbb{N}_0} \frac{1}{n!} \cdot \frac{c^n}{n!} e^{-c} = e^{-c} \sum_{n=0}^{\infty} \frac{1}{n!} (c \cdot e^{it})^n = e^{-c} \cdot e^{-it} = e^{(e^{it} - 1)} \)

since \( \sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x \)

(iv) \( \Phi_X(t) = \frac{1}{b-a} \int_a^b e^{itx} dx = \frac{1}{b-a} \left[ \frac{e^{itx}}{it} \right]_a^b = \frac{e^{ibt} - e^{ita}}{(b-a)it} \)

(v) \( X \sim \text{N}(0, 1) \) is symmetric around 0

\[ \implies \Phi_X(t) \text{ is real since there is no imaginary part according to Theorem 3.3.3 (vi)} \]

\[ \implies \Phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{itx} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(tx)e^{-\frac{x^2}{2}} dx \]

Since \( E(X) \) exists, \( \Phi_X(t) \) is differentiable according to Theorem 3.3.5 and the following holds:
\[ \Phi'(t) = Re(\Phi'(t)) \]
\[ = Re \left( \int_{-\infty}^{\infty} i x e^{itx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right) \]
\[ = Re \left( \int_{-\infty}^{\infty} i x \cos(tx) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx + \int_{-\infty}^{\infty} -x \sin(tx) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \right) \]
\[ = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \sin(tx) e^{-\frac{x^2}{2}} \frac{u}{v'} dx \quad \text{where } u' = -t \cos(tx) \text{ and } v = e^{-\frac{x^2}{2}} \]
\[ = \frac{1}{\sqrt{2\pi}} \sin(tx) e^{-\frac{t^2}{2}} \int_{-\infty}^{\infty} (-t \cos(tx))(-e^{-\frac{x^2}{2}}) dx \]
\[ = -t \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(tx) e^{-\frac{x^2}{2}} dx \]
\[ = -t \Phi_X(t) \]

Thus, \( \Phi'(t) = -t \Phi_X(t). \) It follows that \( \frac{\Phi'(t)}{\Phi_X(t)} = -t \) and by integrating both sides, we get \( \ln |\Phi_X(t)| = -\frac{1}{2} t^2 + c \) with \( c \in \mathbb{R}. \)

For \( t = 0, \) we know that \( \Phi_X(0) = 1 \) by Theorem 3.3.3 (i) and \( \ln |\Phi_X(0)| = 0. \) It follows that \( 0 = 0 + c. \) Therefore, \( c = 0 \) and \( |\Phi_X(t)| = e^{-\frac{1}{2} t^2}. \)

If we take \( t = 0, \) then \( \Phi_X(0) = 1 \) by Theorem 3.3.3 (i). Since \( \Phi_X \) is uniformly continuous, \( \Phi_X \) must take the value 0 before it can eventually take a negative value. However, since \( e^{-\frac{1}{2} t^2} > 0 \quad \forall t \in \mathbb{R}, \) \( \Phi_X \) cannot take 0 as a possible value and therefore cannot pass into the negative numbers. So, it must hold that \( \Phi_X(t) = e^{-\frac{1}{2} t^2} \quad \forall t \in \mathbb{R}. \)

(vi) For \( \sigma > 0, \mu \in \mathbb{R}, \) we know that if \( X \sim N(0, 1), \) then \( \sigma X + \mu \sim N(\mu, \sigma^2). \) By Theorem 3.3.3 (vii) we have
\[ \Phi_{\sigma X + \mu}(t) = e^{it\mu} \Phi_X(\sigma t) = e^{it\mu} e^{-\frac{1}{2} \sigma^2 t^2}. \]

(vii)
\[ \Phi_X(t) = \int_{0}^{\infty} e^{itx} \gamma(p, q, x) dx \]
\[ = \int_{0}^{\infty} e^{itx} \frac{q^p}{\Gamma(p)} x^{p-1} e^{-qx} dx \]
\[ = \int_{0}^{\infty} \frac{q^p}{\Gamma(p)} x^{p-1} e^{-(q-it)x} dx \]
\[
\begin{align*}
&= \frac{q^p}{\Gamma(p)}(q - it)^{-p} \int_0^\infty ((q - it)x)^{p-1} e^{-(q-it)x} (q - it)dx \mid u = (q - it)x, \ du = (q - it)dx \\
&= \frac{q^p}{\Gamma(p)}(q - it)^{-p} \int_0^\infty (u)^{p-1} e^{-u} du \\
&= q^p(q - it)^{-p} \\
&= \left(\frac{q - it}{q}\right)^{-p} \\
&= \left(1 - \frac{it}{q}\right)^{-p}
\end{align*}
\]

(viii) Since an $\Exp(c)$ distribution is a $\Gamma(1, c)$ distribution, we get for $X \sim \Exp(c) = \Gamma(1, c)$:

\[
\Phi_X(t) = (1 - \frac{it}{c})^{-1}
\]

(ix) Since a $\chi^2_n$ distribution (for $n \in \mathbb{N}$) is a $\Gamma(\frac{n}{2}, \frac{1}{2})$ distribution, we get for $X \sim \chi^2_n = \Gamma(\frac{n}{2}, \frac{1}{2})$:

\[
\Phi_X(t) = (1 - \frac{it}{1/2})^{-n/2} = (1 - 2it)^{-n/2}
\]

Example 3.3.13:

Since we know that $m_1 = E(X)$ and $m_2 = E(X^2)$ exist for $X \sim Bin(1, p)$, we can determine these moments according to Theorem 3.3.5 using the characteristic function.

It is

\[
\begin{align*}
\Phi_X(t) &= 1 + p(e^{it} - 1) \\
\Phi'_X(t) &= pie^{it} \\
\Phi'_X(0) &= pi \\
\Rightarrow m_1 &= \frac{\Phi'_X(0)}{i} = \frac{pi}{i} = p = E(X) \\
\Phi''_X(t) &= pi^2 e^{it} \\
\Phi''_X(0) &= pi^2 \\
\Rightarrow m_2 &= \frac{\Phi''_X(0)}{i^2} = \frac{pi^2}{i^2} = p = E(X^2) \\
\Rightarrow Var(X) &= E(X^2) - (E(X))^2 = p - p^2 = p(1 - p)
\end{align*}
\]

\[\blacksquare\]
Note:
The restriction \( \int_{-\infty}^{\infty} | \Phi_X(t) | \, dt < \infty \) in Theorem 3.3.9 works in such a way that we don’t end up with a (non–existing) pdf if \( X \) is a discrete rv. For example,

- \( X \sim \text{Dirac}(c) \):
  \[
  \int_{-\infty}^{\infty} | \Phi_X(t) | \, dt = \int_{-\infty}^{\infty} | e^{itc} | \, dt = \int_{-\infty}^{\infty} 1 \, dt = t \bigg|_{-\infty}^{\infty}
  \]
  which is undefined.

- Also for \( X \sim \text{Bin}(1, p) \):
  \[
  \int_{-\infty}^{\infty} | \Phi_X(t) | \, dt = \int_{-\infty}^{\infty} | 1 + p(e^{it} - 1) | \, dt
  \geq \int_{-\infty}^{\infty} | p e^{it} - (p - 1) | \, dt
  \geq \int_{-\infty}^{\infty} p e^{it} | \, dt - \int_{-\infty}^{\infty} | (p - 1) | \, dt
  = p \int_{-\infty}^{\infty} 1 \, dt - (1 - p) \int_{-\infty}^{\infty} 1 \, dt
  = (2p - 1) \int_{-\infty}^{\infty} 1 \, dt
  = (2p - 1) t \bigg|_{-\infty}^{\infty}
  \]
  which is undefined for \( p \neq 1/2 \).
If \( p = 1/2 \), we have
  \[
  \int_{-\infty}^{\infty} | p e^{it} - (p - 1) | \, dt = 1/2 \int_{-\infty}^{\infty} | e^{it} + 1 | \, dt
  = 1/2 \int_{-\infty}^{\infty} | \cos t + i \sin t + 1 | \, dt
  = 1/2 \int_{-\infty}^{\infty} \sqrt{(\cos t + 1)^2 + (\sin t)^2} \, dt
  = 1/2 \int_{-\infty}^{\infty} \sqrt{\cos^2 t + 2 \cos t + 1 + \sin^2 t} \, dt
  = 1/2 \int_{-\infty}^{\infty} \sqrt{2 + 2 \cos t} \, dt
  \]
  which also does not exist.
• Otherwise, $X \sim N(0, 1)$:

$$\int_{-\infty}^{\infty} | \Phi_X(t) | \ dt = \int_{-\infty}^{\infty} \exp(-t^2/2) dt$$

$$= \sqrt{2\pi} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt$$

$$= \sqrt{2\pi}$$

$$< \infty$$
3.4 Probability Generating Functions

(Based on Casella/Berger, Section 2.6, and Outside Material)

Definition 3.4.1:
Let $X$ be a discrete rv which only takes non-negative integer values, i.e., $p_k = P(X = k)$, and $\sum_{k=0}^{\infty} p_k = 1$. Then, the **probability generating function** (pgf) of $X$ is defined as

$$G(s) = \sum_{k=0}^{\infty} p_k s^k.$$  

Theorem 3.4.2:
$G(s)$ converges for $|s| \leq 1$.

Proof:
$$|G(s)| \leq \sum_{k=0}^{\infty} |p_k s^k| \leq \sum_{k=0}^{\infty} |p_k| = 1$$  

Theorem 3.4.3:
Let $X$ be a discrete rv which only takes non-negative integer values and has pgf $G(s)$. Then it holds:

$$P(X = k) = \frac{1}{k!} \frac{d^k}{ds^k} G(s) |_{s=0}$$  

Theorem 3.4.4:
Let $X$ be a discrete rv which only takes non-negative integer values and has pgf $G(s)$. If $E(X)$ exists, then it holds:

$$E(X) = \frac{d}{ds} G(s) |_{s=1}$$  

Definition 3.4.5:
The $k^{th}$ **factorial moment** of $X$ is defined as

$$E[ X(X-1)(X-2) \ldots (X-k+1) ]$$  

if this expectation exists.
Theorem 3.4.6:
Let $X$ be a discrete rv which only takes non-negative integer values and has pgf $G(s)$. If $E[X(X-1)(X-2)\ldots(X-k+1)]$ exists, then it holds:

$$E[X(X-1)(X-2)\ldots(X-k+1)] = \frac{d^k}{ds^k} G(s) \mid_{s=1}$$

Proof:
Homework

Note:
Similar to the Cauchy distribution for the continuous case, there exist discrete distributions where the mean (or higher moments) do not exist. See Homework.

Example 3.4.7:
Let $X \sim \text{Poisson}(c)$ with

$$P(X = k) = p_k = e^{-c} \frac{c^k}{k!}, \quad k = 0, 1, 2, \ldots$$

It is

$$G(s) = \sum_{k=0}^{\infty} p_k s^k$$

$$= \sum_{k=0}^{\infty} e^{-c} \frac{c^k}{k!} s^k$$

$$= e^{-c} \sum_{k=0}^{\infty} \frac{(cs)^k}{k!}$$

$$= e^{-c} e^{cs}$$

$$= e^{-c(1-s)}, \quad |s| \leq 1.$$ 

It follows:

$$G'(s) = e^{-c} ce^{cs}$$

$$G''(s) = e^{-c} c^2 e^{cs}$$

$$\vdots$$

$$G^{(k)}(s) = e^{-c} c^k e^{cs}$$

From Theorem 3.4.3, we get:

$$P(X = k) = \frac{1}{k!} \frac{d^k}{ds^k} G(s) \mid_{s=0} = \frac{1}{k!} e^{-c} c^k e^{cs} \mid_{s=0} = \frac{1}{k!} e^{-c} c^k$$
From Theorem 3.4.4, we get:

\[ E(X) = \frac{d}{ds} G(s) \bigg|_{s=1} = e^{-c} e^{cs} \bigg|_{s=1} = c \]

From Theorem 3.4.6, we get:

\[ E(X^2) - E(X) = E(X(X-1)) = \frac{d^2}{ds^2} G(s) \bigg|_{s=1} = e^{-c} e^{cs} \bigg|_{s=1} = c^2 \]

It follows:

\[ Var(X) = (E(X^2) - E(X)) + E(X) - (E(X))^2 = c^2 + c - c^2 = c \]
3.5 Moment Inequalities

(Based on Casella/Berger, Sections 3.6, 3.8, and Outside Material)

Theorem 3.5.1:
Let \( h(X) \) be a non-negative Borel–measurable function of a rv \( X \). If \( E(h(X)) \) exists, then it holds:
\[
P(h(X) \geq \epsilon) \leq \frac{E(h(X))}{\epsilon} \quad \forall \epsilon > 0
\]

Proof:
Continuous case only:
\[
E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x) dx
= \int_A h(x) f_X(x) dx + \int_{A^C} h(x) f_X(x) dx \quad \text{where } A = \{x : h(x) \geq \epsilon\}
\geq \int_A h(x) f_X(x) dx
\geq \int_A \epsilon f_X(x) dx
= \epsilon P(h(X) \geq \epsilon) \quad \forall \epsilon > 0
\]
Therefore, \( P(h(X) \geq \epsilon) \leq \frac{E(h(X))}{\epsilon} \quad \forall \epsilon > 0. \)

Corollary 3.5.2: Markov’s Inequality
Let \( h(X) = |X|^r \) and \( \epsilon = k^r \) where \( r > 0 \) and \( k > 0 \). If \( E(|X|^r) \) exists, then it holds:
\[
P(|X| \geq k) \leq \frac{E(|X|^r)}{k^r}
\]

Proof:
Since \( P(|X| \geq k) = P(|X|^r \geq k^r) \) for \( k > 0 \), it follows using Theorem 3.5.1:
\[
P(|X| \geq k) = P(|X|^r \geq k^r) \leq \frac{E(|X|^r)}{k^r}
\]

Corollary 3.5.3: Chebychev’s Inequality
Let \( h(X) = (X - \mu)^2 \) and \( \epsilon = k^2 \sigma^2 \) where \( E(X) = \mu, \text{Var}(X) = \sigma^2 < \infty \), and \( k > 0 \). Then it holds:
\[
P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}
\]
Proof:
Since \( P(|X - \mu| \geq k\sigma) = P(|X - \mu|^2 \geq k^2\sigma^2) \) for \( k > 0 \), it follows using Theorem 3.5.1:

\[
P(|X - \mu| \geq k\sigma) = P(|X - \mu|^2 \geq k^2\sigma^2) \leq \frac{E(|X - \mu|^2)}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}
\]

Note:
For \( k = 2 \), it follows from Corollary 3.5.3 that

\[
P(|X - \mu| < 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75,
\]
no matter what the distribution of \( X \) is. Unfortunately, this is not very precise for many distributions, e.g., the Normal distribution, where it holds that \( P(|X - \mu| < 2\sigma) \approx 0.95 \).

Theorem 3.5.4: Lyapunov’s Inequality
Let \( 0 < \beta_n = E(|X|^n) < \infty \). For arbitrary \( k \) such that \( 2 \leq k \leq n \), it holds that

\[
(\beta_{k-1})\frac{1}{k-1} \leq (\beta_k)^{\frac{1}{k}},
\]
i.e., \( (E(|X|^{k-1}))^{\frac{1}{k-1}} \leq (E(|X|^k))^{\frac{1}{k}} \).

Proof:
Continuous case only:
Let \( Q(u,v) = E \left( (|X|^\frac{j-1}{2} + v|X|^\frac{j+1}{2})^2 \right) \) where \( 1 \leq j \leq k - 1 \).

Obviously, by construction, \( Q(u,v) \geq 0 \ \forall u,v \in \mathbb{R} \).

\[
Q(u,v) = \int_{-\infty}^{\infty} (u|X|^\frac{j-1}{2} + v|X|^\frac{j+1}{2})^2 f_X(x)dx
\]

\[
= u^2 \int_{-\infty}^{\infty} |x|^{j-1} f_X(x)dx + 2uv \int_{-\infty}^{\infty} |x|^j f_X(x)dx + v^2 \int_{-\infty}^{\infty} |x|^{j+1} f_X(x)dx
\]

\[
= u^2 \beta_{j-1} + 2uv\beta_j + v^2 \beta_{j+1}
\]

\[
\geq 0 \ \forall u,v \in \mathbb{R}
\]

Note that for a binary quadratic form

\[
QF(x,y) = (x \ y) \begin{pmatrix} A & B \\ B & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = Ax^2 + 2Bxy + Cy^2
\]
it holds that \( QF \) is positive semidefinite, i.e., \( QF(x,y) \geq 0 \ \forall x,y \in \mathbb{R} \), iff \( A > 0 \) and \( AC - B^2 \geq 0 \). Here, we have by construction that \( Q(u,v) \geq 0 \) with \( A = \beta_{j-1} > 0, B = \beta_j > 0, \) and \( C = \beta_{j+1} > 0 \). Therefore, it must hold:
$$AC - B^2 \geq 0$$

$$\implies \beta_{j-1}\beta_{j+1} - \beta_j^2 \geq 0$$

$$\implies \beta_j^2 \leq \beta_{j-1}\beta_{j+1}$$

$$\implies \beta_j^{2} \leq \beta_{j-1}\beta_{j+1}$$

This means that $\beta_2^2 \leq \beta_0\beta_2$, $\beta_3^2 \leq \beta_1\beta_3$, $\beta_2^2 \leq \beta_1\beta_2$, and so on until $\beta_{k-1}^2 \leq \beta_{k-2}\beta_k$.

Multiplying these $k - 1$ inequalities, we get:

$$\prod_{j=1}^{k-1} \beta_j^2 \leq \prod_{j=1}^{k-1} \beta_{j-1}\beta_{j+1}$$

$$= (\beta_0\beta_2)(\beta_1^2\beta_3)(\beta_2^3\beta_4) \ldots (\beta_{k-2}^k\beta_{k-1}^2)(\beta_{k-2}^k\beta_k^1)$$

$$= \beta_0\beta_{k-1}^k \prod_{j=1}^{k-2} \beta_j^{2j}$$

Dividing both sides by $\prod_{j=1}^{k-2} \beta_j^{2j}$, we get:

$$\beta_{k-1}^{2k-2} \leq \beta_0\beta_{k-1}^{k-2}$$

$$\implies \beta_{k-1}^{k} \leq \beta_{k-1}^{k-1}$$

$$\implies \beta_{k-1}^{1} \leq \beta_{k-1}^{1}$$

$$\implies \beta_{k-1}^{1} \leq \beta_{k-1}^{1}$$

$(\ast)$ holds since $\beta_0 = E(\mid X \mid 0) = E(1) = 1.$

**Note:**

- It follows from Theorem 3.5.4 that

$$E(\mid X \mid)^1 \leq (E(\mid X \mid 2))^{1/2} \leq (E(\mid X \mid 3))^{1/3} \leq \ldots \leq (E(\mid X \mid n))^{1/n}.$$  

- For $X \sim \text{Dirac}(c)$, $c > 0$, with $P(X = c) = 1$, it follows immediately from Theorem 3.3.12 (i) and Theorem 3.3.5 that $m_k = E(X^k) = c^k$. So,

$$E(\mid X \mid k) = E(X^k) = c^k$$

and

$$(E(\mid X \mid k))^{1/k} = (E(X^k))^{1/k} = (c^k)^{1/k} = c.$$  

Therefore, equality holds in Theorem 3.5.4.
4 Random Vectors

4.1 Joint, Marginal, and Conditional Distributions

(Based on Casella/Berger, Sections 4.1 & 4.2)

Definition 4.1.1:
The vector $X' = (X_1, \ldots, X_n)$ on $(\Omega, L, P) \to \mathbb{R}^n$ defined by $X(\omega) = (X_1(\omega), \ldots, X_n(\omega))^\prime$, $\omega \in \Omega$, is an \textit{n–dimensional random vector (n–rv)} if $X^{-1}(I) = \{\omega : X_1(\omega) \leq a_1, \ldots, X_n(\omega) \leq a_n\} \in L$ for all \textit{n–dimensional intervals $I = \{(x_1, \ldots, x_n) : -\infty < x_i \leq a_i, a_i \in \mathbb{R} \forall i = 1, \ldots, n}\}$. ■

Note:
It follows that if $X_1, \ldots, X_n$ are any n rv’s on $(\Omega, L, P)$, then $X = (X_1, \ldots, X_n)'$ is an n–rv on $(\Omega, L, P)$ since for any $I$, it holds:

$$X^{-1}(I) = \{\omega : (X_1(\omega), \ldots, X_n(\omega)) \in I\}$$

$$= \{\omega : X_1(\omega) \leq a_1, \ldots, X_n(\omega) \leq a_n\}$$

$$= \bigcap_{k=1}^{n} \{\omega : X_k(\omega) \leq a_k\}_{\in L}$$

■

Definition 4.1.2:
For an n–rv $X$, a function $F$ defined by

$$F(x) = P(X \leq x) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \forall x \in \mathbb{R}^n$$

is the \textit{joint cumulative distribution function (joint cdf)} of $X$. ■

Note:

(i) $F$ is non–decreasing and right–continuous in each of its arguments $x_i$.

(ii) $\lim_{\substack{x_i \to \infty \rightarrow \mathbb{R}^n \\forall i \neq k \\rightarrow -\infty \\forall x_k}} F(x) = 1$ and $\lim_{\substack{x_k \to -\infty \rightarrow \mathbb{R}^n \\forall x_k \neq k \\rightarrow \infty \\forall x_k}} F(x) = 0 \forall x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n \in \mathbb{R}^n$.

However, conditions (i) and (ii) together are not sufficient for $F$ to be a joint cdf. Instead we need the conditions from the next Theorem. ■
Theorem 4.1.3:
A function \( F(x) = F(x_1, \ldots, x_n) \) is the joint cdf of some \( n \)-rv \( X \) iff 

(i) \( F \) is non-decreasing and right-continuous with respect to each \( x_i \),

(ii) \( F(-\infty, x_2, \ldots, x_n) = F(x_1, -\infty, x_3, \ldots, x_n) = \ldots = F(x_1, \ldots, x_{n-1}, -\infty) = 0 \) and \( F(\infty, \ldots, \infty) = 1 \), and

(iii) \( \forall x \in \mathbb{R}^n \forall \epsilon_i > 0, i = 1, \ldots, n \), the following inequality holds:

\[
F(x + \epsilon) - \sum_{i=1}^{n} F(x_1 + \epsilon_1, \ldots, x_i-1 + \epsilon_{i-1}, x_i, x_{i+1} + \epsilon_i, \ldots, x_n + \epsilon_n) + \sum_{1 \leq i < j \leq n} F(x_1 + \epsilon_1, \ldots, x_{i-1} + \epsilon_{i-1}, x_i, x_{i+1} + \epsilon_i, \ldots, x_{j-1} + \epsilon_{j-1}, x_j, x_{j+1} + \epsilon_{j+1}, \ldots, x_n + \epsilon_n) \]

\[+\ldots\]

\[+\ (-1)^n F(x) \]

\[\geq 0\]

Note:
We won’t prove this Theorem but just see why we need condition (iii) for \( n = 2 \):

\[
P(x_1 < X \leq x_2, y_1 < Y \leq y_2) =
P(X \leq x_2, Y \leq y_2) - P(X \leq x_1, Y \leq y_2) - P(X \leq x_2, Y \leq y_1) + P(X \leq x_1, Y \leq y_1) \geq 0
\]
Note:
We will restrict ourselves to \( n = 2 \) for most of the next Definitions and Theorems but those can be easily generalized to \( n > 2 \). The term bivariate rv is often used to refer to a 2-rv and multivariate rv is used to refer to an \( n \)-rv, \( n \geq 2 \).

**Definition 4.1.4:**
A 2-rv \((X, Y)\) is **discrete** if there exists a countable collection \( \mathcal{X} \) of pairs \((x_i, y_j)\) that has probability 1. Let \( p_{ij} = P(X = x_i, Y = y_j) > 0 \; \forall (x_i, y_j) \in \mathcal{X} \). Then, \( \sum_{i,j} p_{ij} = 1 \) and \( \{p_{ij}\} \) is the joint probability mass function (joint pmf) of \((X, Y)\).

**Definition 4.1.5:**
Let \((X, Y)\) be a discrete 2-rv with joint pmf \( \{p_{ij}\} \). Define
\[
p_i = \sum_{j=1}^{\infty} p_{ij} = \sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = P(X = x_i) \]
and
\[
p_j = \sum_{i=1}^{\infty} p_{ij} = \sum_{i=1}^{\infty} P(X = x_i, Y = y_j) = P(Y = y_j).
\]
Then \( \{p_i\} \) is called the **marginal probability mass function** (marginal pmf) of \( X \) and \( \{p_j\} \) is called the **marginal probability mass function** of \( Y \).

**Definition 4.1.6:**
A 2-rv \((X, Y)\) is **continuous** if there exists a non-negative function \( f \) such that
\[
F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) \, dv \, du \; \forall (x, y) \in \mathbb{R}^2
\]
where \( F \) is the joint cdf of \((X, Y)\). We call \( f \) the joint probability density function (joint pdf) of \((X, Y)\).

**Note:**
If \( F \) is absolutely continuous at \((x, y)\), then
\[
\frac{\partial^2 F(x, y)}{\partial x \partial y} = f(x, y).
\]
Definition 4.1.7:
Let $(X, Y)$ be a continuous 2–rv with joint pdf $f$. Then $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$ is called the **marginal probability density function** (marginal pdf) of $X$ and $f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$ is called the **marginal probability density function** of $Y$.

**Note:**
(i) $\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} f(x, y) dy \right) dx = F(\infty, \infty) = 1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy$ and $f_X(x) \geq 0 \ \forall x \in \mathbb{R}$ and $f_Y(y) \geq 0 \ \forall y \in \mathbb{R}$.

(ii) Given a 2–rv $(X, Y)$ with joint cdf $F(x, y)$, how do we generate a marginal cdf $F_X(x) = P(X \leq x)$? — The answer is $P(X \leq x) = P(X \leq x, -\infty < Y < \infty) = F(x, \infty)$.

Definition 4.1.8:
If $F_{X}(x_1, \ldots, x_n) = F_{X}(x)$ is the joint cdf of an $n$–rv $X = (X_1, \ldots, X_n)$, then the **marginal cumulative distribution function** (marginal cdf) of $(X_{i_1}, \ldots, X_{i_k}), 1 \leq k \leq n - 1, 1 \leq i_1 < i_2 < \ldots < i_k \leq n$, is given by

$$
\lim_{x_{i_1, \ldots, i_k} \to \infty, i_1 \neq i_1, \ldots, i_k} F_{X}(x) = F_{X}(\infty, \ldots, \infty, x_{i_1}, \infty, \ldots, \infty, x_{i_2}, \infty, \ldots, \ldots, \infty, x_{i_k}, \infty, \ldots, \infty).
$$

**Note:**
In Definition 1.4.1, we defined conditional probability distributions in some probability space $(\Omega, L, P)$. This definition extends to conditional distributions of 2–rv’s $(X, Y)$.

Definition 4.1.9:
Let $(X, Y)$ be a discrete 2–rv. If $P(Y = y_j) = p_j > 0$, then the **conditional probability mass function** (conditional pmf) of $X$ given $Y = y_j$ (for fixed $j$) is defined as

$$
p_{i|j} = P(X = x_i | Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_j}.
$$
Note:
For a continuous 2–rv \((X, Y)\) with pdf \(f\), \(P(X \leq x \mid Y = y)\) is not defined. Let \(\epsilon > 0\) and suppose that \(P(y - \epsilon < Y \leq y + \epsilon) > 0\). For every \(x\) and every interval \((y - \epsilon, y + \epsilon]\), consider the conditional probability of \(X \leq x\) given \(Y \in (y - \epsilon, y + \epsilon]\). We have
\[
P(X \leq x \mid y - \epsilon < Y \leq y + \epsilon) = \frac{P(X \leq x, y - \epsilon < Y \leq y + \epsilon)}{P(y - \epsilon < Y \leq y + \epsilon)}
\]
which is well–defined if \(P(y - \epsilon < Y \leq y + \epsilon) > 0\) holds.

So, when does
\[
\lim_{\epsilon \to 0^+} P(X \leq x \mid Y \in (y - \epsilon, y + \epsilon])
\]
exist? See the next definition.

Definition 4.1.10:
The conditional cumulative distribution function (conditional cdf) of a rv \(X\) given that \(Y = y\) is defined to be
\[
F_{X\mid Y}(x \mid y) = \lim_{\epsilon \to 0^+} P(X \leq x \mid Y \in (y - \epsilon, y + \epsilon])
\]
provided that this limit exists. If it does exist, the conditional probability density function (conditional pdf) of \(X\) given that \(Y = y\) is any non–negative function \(f_{X\mid Y}(x \mid y)\) satisfying
\[
F_{X\mid Y}(x \mid y) = \int_{-\infty}^{x} f_{X\mid Y}(t \mid y)dt \quad \forall x \in \mathbb{R}.
\]

Note:
For fixed \(y\), \(f_{X\mid Y}(x \mid y) \geq 0\) and \(\int_{-\infty}^{\infty} f_{X\mid Y}(x \mid y)dx = 1\). So it is really a pdf.

Theorem 4.1.11:
Let \((X, Y)\) be a continuous 2–rv with joint pdf \(f_{X,Y}\). It holds that at every point \((x, y)\) where \(f\) is continuous and the marginal pdf \(f_{Y}(y) > 0\), we have
\[
F_{X\mid Y}(x \mid y) = \lim_{\epsilon \to 0^+} \frac{P(X \leq x, Y \in (y - \epsilon, y + \epsilon])}{P(Y \in (y - \epsilon, y + \epsilon])}
\]
\[
= \lim_{\epsilon \to 0^+} \left( \frac{\frac{1}{2\epsilon} \int_{-\infty}^{x} \int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(u,v)dv\,du}{\frac{1}{2\epsilon} \int_{y-\epsilon}^{y+\epsilon} f_{Y}(v)dv} \right)
\]
82
\[
\int_{-\infty}^{x} f_{X,Y}(u, y) du = \frac{\int_{-\infty}^{x} f_{X,Y}(u, y) du}{f_Y(y)} = \int_{-\infty}^{x} \frac{f_{X,Y}(u, y)}{f_Y(y)} du.
\]

Thus, \( f_{X|Y}(x \mid y) \) exists and equals \( \frac{f_{X,Y}(x, y)}{f_Y(y)} \), provided that \( f_Y(y) > 0 \). Furthermore, since

\[
\int_{-\infty}^{x} f_{X,Y}(u, y) du = f_Y(y) F_{X|Y}(x \mid y),
\]

we get the following marginal cdf of \( X \):

\[
F_X(x) = \int_{-\infty}^{\infty} \left( \int_{-\infty}^{x} f_{X,Y}(u, y) du \right) dy = \int_{-\infty}^{\infty} f_Y(y) F_{X|Y}(x \mid y) dy
\]

Example 4.1.12:

Consider

\[
f_{X,Y}(x, y) = \begin{cases}
2, & 0 < x < y < 1 \\
0, & \text{otherwise}
\end{cases}
\]

Example 4.1.12
We calculate the marginal pdf’s $f_X(x)$ and $f_Y(y)$ first:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy = \int_x^1 2 dy = 2(1 - x) \text{ for } 0 < x < 1$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx = \int_y^1 2 dx = 2y \text{ for } 0 < y < 1$$

The conditional pdf’s $f_{Y|X}(y \mid x)$ and $f_{X|Y}(x \mid y)$ are calculated as follows:

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{2}{2(1 - x)} = \frac{1}{1 - x} \text{ for } x < y < 1 \text{ (where } 0 < x < 1)$$

and

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{2}{2y} = \frac{1}{y} \text{ for } 0 < x < y \text{ (where } 0 < y < 1)$$

Thus, it holds that $Y \mid X = x \sim U(x,1)$ and $X \mid Y = y \sim U(0,y)$, i.e., both conditional pdf’s are related to uniform distributions.
4.2 Independent Random Variables

(Based on Casella/Berger, Sections 4.2 & 4.6)

Example 4.2.1: (from Rohatgi, page 119, Example 1)
Let $f_1, f_2, f_3$ be 3 pdf’s with cdf’s $F_1, F_2, F_3$ and let $|\alpha| \leq 1$. Define

$$f_\alpha(x_1, x_2, x_3) = f_1(x_1)f_2(x_2)f_3(x_3) \cdot (1 + \alpha(2F_1(x_1) - 1)(2F_2(x_2) - 1)(2F_3(x_3) - 1)).$$

We can show

(i) $f_\alpha$ is a pdf for all $\alpha \in [-1, 1]$.

(ii) $\{f_\alpha : -1 \leq \alpha \leq 1\}$ all have marginal pdf’s $f_1, f_2, f_3$.

See book for proof and further discussion — but when do the marginal distributions uniquely determine the joint distribution?

Definition 4.2.2:
Let $F_{X,Y}(x, y)$ be the joint cdf and $F_X(x)$ and $F_Y(y)$ be the marginal cdf’s of a 2-rv $(X, Y)$. $X$ and $Y$ are independent iff

$$F_{X,Y}(x, y) = F_X(x)F_Y(y) \quad \forall (x, y) \in \mathbb{R}^2.$$

Lemma 4.2.3:
If $X$ and $Y$ are independent, $a, b, c, d \in \mathbb{R}$, and $a < b$ and $c < d$, then

$$P(a < X \leq b, c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d).$$

Proof:

$$P(a < X \leq b, c < Y \leq d) = F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_X(a, c)$$

$$= F_X(b)F_Y(d) - F_X(a)F_Y(d) - F_X(b)F_Y(c) + F_X(a)F_Y(c)$$

$$= (F_X(b) - F_X(a))(F_Y(d) - F_Y(c))$$

$$= P(a < X \leq b)P(c < Y \leq d).$$

Definition 4.2.4:
A collection of rv’s $X_1, \ldots, X_n$ with joint cdf $F_{\mathbf{X}}(\mathbf{x})$ and marginal cdf’s $F_{X_i}(x_i)$ are mutually (or completely) independent iff

$$F_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} F_{X_i}(x_i) \quad \forall \mathbf{x} \in \mathbb{R}^n.$$
Note:

(i) We often simply say that the r.v’s $X_1, \ldots, X_n$ are independent when we really mean that they are mutually independent.

(ii) Instead of saying “a necessary and sufficient condition for discrete r.v’s $X_1, \ldots, X_n$ to be independent is ...” in the following Theorem, we could say that “discrete r.v’s $X_1, \ldots, X_n$ are independent iff ...”, i.e., “necessary and sufficient” means the same as “iff”.

A is necessary for $B$ means: $B \Rightarrow A$ (because $\neg A \Rightarrow \neg B$)

A is sufficient for $B$ means: $A \Rightarrow B$

\[\blacksquare\]

\textbf{Theorem 4.2.5: Factorization Theorem}

(i) A necessary and sufficient condition for discrete r.v’s $X_1, \ldots, X_n$ to be independent is that

\[P(\underline{X} = \underline{x}) = P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \quad \forall \underline{x} \in \mathcal{X} \quad (\ast)\]

where $\mathcal{X} \subset \mathbb{R}^n$ is the countable support of $\underline{X}$.

(ii) For an absolutely continuous n-rv $\underline{X} = (X_1, \ldots, X_n)$, $X_1, \ldots, X_n$ are independent iff

\[f_{\underline{X}}(\underline{x}) = f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i),\]

where $f_{\underline{X}}$ is the joint pdf and $f_{X_1}, \ldots, f_{X_n}$ are the marginal pdfs of $\underline{X}$.

\textbf{Proof:}

(i) Discrete case:

$\Rightarrow$": Let $\underline{X}$ be a random vector whose components are independent random variables of the discrete type with $P(\underline{X} = \underline{b}) > 0$.

Lemma 4.2.3 can be extended to:

\[P(\underline{a} < \underline{X} \leq \underline{b}) = P(a_1 < X_1 \leq b_1, \ldots, a_n < X_n \leq b_n)\]
\[= P(a_1 < X_1 \leq b_1) \cdot \ldots \cdot P(a_n < X_n \leq b_n) \quad (A)\]
Therefore,

\[
P(X = b) = \lim_{a \downarrow b} P(a < X \leq b)
\]

\[
= \lim_{a_1 \downarrow b_1 \forall i \in \{1, \ldots, n\}} P(a_1 < X_1 \leq b_1, \ldots, a_n < X_n \leq b_n)
\]

\[
= \lim_{a_1 \downarrow b_1 \forall i \in \{1, \ldots, n\}} P(a_1 < X_1 \leq b_1) \cdots P(a_n < X_n \leq b_n)
\]

\[
= P(X_1 = b_1) \cdots P(X_n = b_n)
\]

\[
= \prod_{i=1}^{n} P(X_i = b_i),
\]

i.e., (⋆) holds.

"⇐": For \( n \) dimensions, let \( \mathbf{x} = (x_1, x_2, \ldots, x_n) \), \( \mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \), and \( B = \{ \mathbf{x}_i : x_{i1} \leq x_1, x_{i2} \leq x_2, \ldots, x_{in} \leq x_n \} \), \( B \in \mathcal{X} \). We assume that (⋆) holds. Then it follows:

\[
F_X(\mathbf{x}) = \sum_{\mathbf{x}_i \in B} P(\mathbf{X} = \mathbf{x}_i)
\]

\[
= \sum_{\mathbf{x}_i \in B} P(X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_n = x_{in})
\]

\[
= \sum_{\mathbf{x}_i \in B} \prod_{j=1}^{n} P(X_j = x_{ij})
\]

\[
= \sum_{x_{11} \leq x_1} \sum_{x_{21} \leq x_2} \cdots \sum_{x_{n-1,1} \leq x_{n-1}} \sum_{x_{n1} \leq x_n} \prod_{j=1}^{n} P(X_j = x_{ij}) P(X_n = x_{in})
\]

\[
= \sum_{x_{11} \leq x_1} \sum_{x_{21} \leq x_2} \cdots \sum_{x_{n-1,1} \leq x_{n-1}} \left( \prod_{j=1}^{n-1} P(X_j = x_{ij}) \right) \left( \sum_{x_{in} \leq x_n} P(X_n = x_{in}) \right)
\]

\[
= \sum_{x_{11} \leq x_1} \sum_{x_{21} \leq x_2} \cdots \sum_{x_{n-1,1} \leq x_{n-1}} \left( \prod_{j=1}^{n-2} P(X_j = x_{ij}) \right) \left( \sum_{x_{in} \leq x_n} P(X_n = x_{in}) \right)
\]

\[
= \sum_{x_{11} \leq x_1} \sum_{x_{21} \leq x_2} \cdots \left( \prod_{j=1}^{n-2} P(X_j = x_{ij}) \right) \left( \sum_{x_{in} \leq x_n} P(X_n = x_{in}) \right)
\]

\[
\vdots
\]

87
\[
\prod_{j=1}^{n} \left( \sum_{x_{ij} \leq x_j} P(X_j = x_{ij}) \right) = \prod_{j=1}^{n} F_{X_j}(x_j),
\]

i.e., \(X_1, \ldots, X_n\) are mutually independent according to Definition 4.2.4.

(ii) Continuous case: Homework

\[\text{Theorem 4.2.6:}\]
\(X_1, \ldots, X_n\) are independent iff \(P(X_i \in A_i, i = 1, \ldots, n) = \prod_{i=1}^{n} P(X_i \in A_i) \forall \text{ Borel sets } A_i \in \mathcal{B}\)
(i.e., rv’s are independent iff all events involving these rv’s are independent).

\[\text{Proof:}\]
Lemma 4.2.3 and definition of Borel sets.

\[\text{Theorem 4.2.7:}\]
Let \(X_1, \ldots, X_n\) be independent rv’s and \(g_1, \ldots, g_n\) be Borel–measurable functions. Then \(g_1(X_1), g_2(X_2), \ldots, g_n(X_n)\) are independent.

\[\text{Proof:}\]
\[
F_{g_1(X_1), g_2(X_2), \ldots, g_n(X_n)}(h_1, h_2, \ldots, h_n) = P(g_1(X_1) \leq h_1, g_2(X_2) \leq h_2, \ldots, g_n(X_n) \leq h_n)
\]
\[\stackrel{(*)}{=} \prod_{i=1}^{n} P(g_i(X_i) \leq h_i) \quad \text{Th.4.2.6}\]
\[= \prod_{i=1}^{n} F_{g_i(X_i)}(h_i)
\]
\[\text{(*) holds since } g_i^{-1}((-\infty, h_i]) \in \mathcal{B}, \ldots, g_n^{-1}((-\infty, h_n]) \in \mathcal{B}\]

88
Theorem 4.2.8:
If \( X_1, \ldots, X_n \) are independent, then also every subcollection \( X_{i_1}, \ldots, X_{i_k}, \ k = 2, \ldots, n - 1, \ 1 \leq i_1 < i_2 \ldots < i_k \leq n, \) is independent.

Definition 4.2.9:
A set (or a sequence) of rv’s \( \{X_n\}_{n=1}^{\infty} \) is independent iff every finite subcollection is independent.

Note:
Recall that \( X \) and \( Y \) are identically distributed iff \( F_X(x) = F_Y(x) \ \forall x \in \mathbb{R} \) according to Definition 2.2.5 and Theorem 2.2.6.

Definition 4.2.10:
We say that \( \{X_n\}_{n=1}^{\infty} \) is a set (or a sequence) of independent identically distributed (iid) rv’s if \( \{X_n\}_{n=1}^{\infty} \) is independent and all \( X_n \) are identically distributed.

Note:
Recall that \( X \) and \( Y \) being identically distributed does not say that \( X = Y \) with probability 1. If this happens, we say that \( X \) and \( Y \) are equivalent rv’s.

Note:
We can also extend the definition of independence to 2 random vectors \( \mathbf{X}^{n \times 1} \) and \( \mathbf{Y}^{n \times 1} \): \( \mathbf{X} \) and \( \mathbf{Y} \) are independent iff \( F_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y}) = F_{\mathbf{X}}(\mathbf{x})F_{\mathbf{Y}}(\mathbf{y}) \ \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n \).

This does not mean that the components \( X_i \) of \( \mathbf{X} \) or the components \( Y_i \) of \( \mathbf{Y} \) are independent. However, it does mean that each pair of components \( (X_i, Y_i) \) are independent, any subcollections \( (X_{i_1}, \ldots, X_{i_k}) \) and \( (Y_{j_1}, \ldots, Y_{j_l}) \) are independent, and any Borel–measurable functions \( f(\mathbf{X}) \) and \( g(\mathbf{Y}) \) are independent.

Corollary 4.2.11: (to Factorization Theorem 4.2.5)
If \( X \) and \( Y \) are independent rv’s, then
\[
F_{X|Y}(x \mid y) = F_X(x) \ \forall x,
\]
and
\[
F_{Y|X}(y \mid x) = F_Y(y) \ \forall y.
\]
4.3 Functions of Random Vectors

(Based on Casella/Berger, Sections 4.3 & 4.6)

Theorem 4.3.1:

If $X$ and $Y$ are rv’s on $(\Omega, L, P) \to \mathbb{R}$, then

(i) $X \pm Y$ is a rv.

(ii) $XY$ is a rv.

(iii) If $\{\omega : Y(\omega) = 0\} = \emptyset$, then $\frac{X}{Y}$ is a rv.

Theorem 4.3.2:

Let $X_1, \ldots, X_n$ be rv’s on $(\Omega, L, P) \to \mathbb{R}$. Define

$\text{MAX}_n = \max\{X_1, \ldots, X_n\} = X_{(n)}$

by

$\text{MAX}_n(\omega) = \max\{X_1(\omega), \ldots, X_n(\omega)\} \ \forall \omega \in \Omega$

and

$\text{MIN}_n = \min\{X_1, \ldots, X_n\} = X_{(1)} = -\max\{-X_1, \ldots, -X_n\}$

by

$\text{MIN}_n(\omega) = \min\{X_1(\omega), \ldots, X_n(\omega)\} \ \forall \omega \in \Omega$.

Then,

(i) $\text{MIN}_n$ and $\text{MAX}_n$ are rv’s.

(ii) If $X_1, \ldots, X_n$ are independent, then

$F_{\text{MAX}_n}(z) = P(\text{MAX}_n \leq z) = P(X_i \leq z \ \forall i = 1, \ldots, n) = \prod_{i=1}^{n} F_{X_i}(z)$

and

$F_{\text{MIN}_n}(z) = P(\text{MIN}_n \leq z) = 1 - P(X_i > z \ \forall i = 1, \ldots, n) = 1 - \prod_{i=1}^{n} (1 - F_{X_i}(z))$.

(iii) If $\{X_i\}_{i=1}^{n}$ are iid rv’s with common cdf $F_X$, then

$F_{\text{MAX}_n}(z) = F_X^n(z)$
and
\[ F_{MIN_n}(z) = 1 - (1 - F_X(z))^n. \]

If \( F_X \) is absolutely continuous with pdf \( f_X \), then the pdfs of \( MAX_n \) and \( MIN_n \) are
\[ f_{MAX_n}(z) = n \cdot F_X^{n-1}(z) \cdot f_X(z) \]
and
\[ f_{MIN_n}(z) = n \cdot (1 - F_X(z))^{n-1} \cdot f_X(z) \]
for all continuity points of \( F_X \).

Note:
Using Theorem 4.3.2, it is easy to derive the joint cdf and pdf of \( MAX_n \) and \( MIN_n \) for iid rv’s \( \{X_1, \ldots, X_n\} \). For example, if the \( X_i \)’s are iid with cdf \( F_X \) and pdf \( f_X \), then the joint pdf of \( MAX_n \) and \( MIN_n \) is
\[ f_{MAX_n,MIN_n}(x, y) = \begin{cases} 0, & x \leq y \\ n(n-1) \cdot (F_X(x) - F_X(y))^{n-2} \cdot f_X(x)f_X(y), & x > y \end{cases} \]
However, note that \( MAX_n \) and \( MIN_n \) are not independent. See Rohatgi, page 129, Corollary, for more details.

Note:
The previous transformations are special cases of the following Theorem 4.3.3.

Theorem 4.3.3:
If \( g : \mathbb{R}^n \to \mathbb{R}^m \) is a Borel–measurable function (i.e., \( \forall B \in \mathcal{B}^m : g^{-1}(B) \in \mathcal{B}^n \)) and if \( \underline{X} = (X_1, \ldots, X_n) \) is an n–rv, then \( g(\underline{X}) \) is an m–rv.

Proof:
If \( B \in \mathcal{B}^m \), then \( \{\omega : g(\underline{X}(\omega)) \in B\} = \{\omega : \underline{X}(\omega) \in g^{-1}(B)\} \in \mathcal{B}^n \).

Question: How do we handle more general transformations of \( \underline{X} \)?

Discrete Case:
Let \( \underline{X} = (X_1, \ldots, X_n) \) be a discrete n–rv and \( \mathcal{X} \subset \mathbb{R}^n \) be the countable support of \( \underline{X} \), i.e.,
\[ P(\underline{X} \in \mathcal{X}) = 1 \] and \( P(\underline{X} = \underline{x}) > 0 \ \forall \underline{x} \in \mathcal{X} \).
Define \( u_i = g_i(x_1, \ldots, x_n), i = 1, \ldots, n \) to be 1-to-1-mappings of \( X \) onto \( B \). Let \( u = (u_1, \ldots, u_n)' \). Then
\[
P(U = u) = P(g_1(X) = u_1, \ldots, g_n(X) = u_n) = P(X_1 = h_1(u), \ldots, X_n = h_n(u)) \quad \forall u \in B
\]
where \( x_i = h_i(u), i = 1, \ldots, n, \) is the inverse transformation (and \( P(U = u) = 0 \quad \forall u \not\in B \)).

The joint marginal pmf of any subcollection of \( u_i \)'s is now obtained by summing over the other remaining \( u_j \)'s.

**Example 4.3.4:**
Let \( X, Y \) be iid \( \sim \text{Bin}(n, p), 0 < p < 1 \). Let \( U = X Y + 1 \) and \( V = Y + 1 \).

Then \( X = U \cdot (Y + 1) = UV \) and \( Y = V - 1 \). So the joint pmf of \( U, V \) is
\[
P(U = u, V = v) = {n \choose uv} p^{uv}(1 - p)^{n - uv} \left( \frac{n}{v - 1} \right) p^{v - 1}(1 - p)^{n + 1 - v}
\]
for \( v \in \{1, 2, \ldots, n + 1\} \) and \( uv \in \{0, 1, \ldots, n\} \).

**Continuous Case:**
Let \( \mathbf{X} = (X_1, \ldots, X_n) \) be a continuous n–rv with joint cdf \( F_{\mathbf{X}} \) and joint pdf \( f_{\mathbf{X}} \).

Let
\[
\mathbf{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} = g(\mathbf{X}) = \begin{pmatrix} g_1(\mathbf{X}) \\ \vdots \\ g_n(\mathbf{X}) \end{pmatrix},
\]
i.e., \( U_i = g_i(\mathbf{x}) \), be a mapping from \( \mathbb{R}^n \) into \( \mathbb{R}^n \).

If \( B \in \mathcal{B}^n \), then
\[
P(\mathbf{U} \in B) = P(\mathbf{X} \in g^{-1}(B)) = \int \cdots \int_{g^{-1}(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int \cdots \int_{g^{-1}(B)} f_{\mathbf{X}}(\mathbf{z}) \prod_{i=1}^n dx_i
\]
where \( g^{-1}(B) = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : g(\mathbf{x}) \in B\} \).

Suppose we define \( B \) as the half–infinite n–dimensional interval
\[
B_{\mathbf{u}} = \{(\tilde{u}_1, \ldots, \tilde{u}_n) : -\infty < \tilde{u}_i < u_i \quad \forall i = 1, \ldots, n\}
\]
for any \( \mathbf{u} \in \mathbb{R}^n \). Then the joint cdf of \( \mathbf{U} \) is
\[
G_U(\mathbf{u}) = P(\mathbf{U} \in B_\mathbf{u}) = P(g_1(X) \leq u_1, \ldots, g_n(X) \leq u_n) = \int \cdots \int_{g^{-1}(B_\mathbf{u})} f_X(\mathbf{x}) d\mathbf{x}.
\]
If \( G \) happens to be absolutely continuous, the joint pdf of \( \mathbf{U} \) will be given by
\[
f_U(\mathbf{u}) = \frac{\partial^n G(\mathbf{u})}{\partial u_1 \partial u_2 \cdots \partial u_n} \text{ at every continuity point of } f_U.
\]
Under certain conditions, we can write \( f_U \) in terms of the original pdf \( f_X \) of \( X \) as stated in the next Theorem:

**Theorem 4.3.5: Multivariate Transformation**

Let \( X = (X_1, \ldots, X_n) \) be a continuous n–rv with joint pdf \( f_X \).

(i) Let
\[
\mathbf{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} = g(\mathbf{X}) = \begin{pmatrix} g_1(X) \\ \vdots \\ g_n(X) \end{pmatrix},
\]
(i.e., \( U_i = g_i(X) \)) be a 1–to–1–mapping from \( \mathbb{R}^n \) into \( \mathbb{R}^n \), i.e., there exist inverses \( h_i \), \( i = 1, \ldots, n \), such that \( x_i = h_i(\mathbf{u}) = h_i(u_1, \ldots, u_n), i = 1, \ldots, n \), over the range of the transformation \( g \).

(ii) Assume both \( g \) and \( h \) are continuous.

(iii) Assume partial derivatives \( \frac{\partial x_i}{\partial u_j} = \frac{\partial h_i(\mathbf{u})}{\partial u_j}, i, j = 1, \ldots, n \), exist and are continuous.

(iv) Assume that the Jacobian of the inverse transformation
\[
J = \det \left( \frac{\partial (x_1, \ldots, x_n)}{\partial (u_1, \ldots, u_n)} \right) = \det \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \cdots & \frac{\partial x_n}{\partial u_n} \end{vmatrix}
\]
is different from 0 for all \( \mathbf{u} \) in the range of the transformation.

Then the n–rv \( \mathbf{U} = g(\mathbf{X}) \) has a joint absolutely continuous cdf with corresponding joint pdf
\[
f_U(\mathbf{u}) = |J| f_X(h_1(\mathbf{u}), \ldots, h_n(\mathbf{u})).
\]

**Proof:**

Let \( \mathbf{u} \in \mathbb{R}^n \) and
\[
B_\mathbf{u} = \{ (\tilde{u}_1, \ldots, \tilde{u}_n) : -\infty < \tilde{u}_i < u_i \ \forall i = 1, \ldots, n \}.
\]
Then,
\[ G_U(u) = \int \ldots \int_{g^{-1}(B_u)} f_X(x) d(x) \]
\[ = \int \ldots \int_{B_u} f_X(h_1(u), \ldots, h_n(u)) | J | d(u) \]

The result follows from differentiation of \( G_U \).

For additional steps of the proof see Rohatgi (page 135 and Theorem 17 on page 10) or a book on multivariate calculus.

**Theorem 4.3.6:**
Let \( X = (X_1, \ldots, X_n) \) be a continuous n–rv with joint pdf \( f_X \).

(i) Let \( U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} = g(X) = \begin{pmatrix} g_1(X) \\ \vdots \\ g_n(X) \end{pmatrix} \),

(i.e., \( U_i = g_i(X) \)) be a mapping from \( \mathbb{R}^n \) into \( \mathbb{R}^n \).

(ii) Let \( \mathcal{X} = \{ x : f_X(x) > 0 \} \) be the support of \( X \).

(iii) Suppose that for each \( u \in B = \{ u \in \mathbb{R}^n : u = g(x) \text{ for some } x \in \mathcal{X} \} \) there is a finite number \( k = k(u) \) of inverses.

(iv) Suppose we can partition \( \mathcal{X} \) into \( \mathcal{X}_0, \mathcal{X}_1, \ldots, \mathcal{X}_k \) s.t.

(a) \( P(X \in \mathcal{X}_0) = 0 \).

(b) \( U = g(X) \) is a 1–to–1–mapping from \( \mathcal{X}_l \) onto \( B \) for all \( l = 1, \ldots, k \), with inverse transformation \( h_l(u) = \begin{pmatrix} h_{l1}(u) \\ \vdots \\ h_{ln}(u) \end{pmatrix}, u \in B \), i.e., for each \( u \in B, h_l(u) \) is the unique \( x \in \mathcal{X}_l \) such that \( u = g(x) \).

(v) Assume partial derivatives \( \frac{\partial x_i}{\partial u_j} = \frac{\partial h_i(u)}{\partial u_j}, l = 1, \ldots, k, i, j = 1, \ldots, n \), exist and are continuous.

(vi) Assume the Jacobian of each of the inverse transformations
\[ J_l = \det \begin{pmatrix} \frac{\partial x_1}{\partial u_1} & \ldots & \frac{\partial x_1}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \ldots & \frac{\partial x_n}{\partial u_n} \end{pmatrix} = \det \begin{pmatrix} \frac{\partial h_{l1}}{\partial u_1} & \ldots & \frac{\partial h_{l1}}{\partial u_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{ln}}{\partial u_1} & \ldots & \frac{\partial h_{ln}}{\partial u_n} \end{pmatrix}, \quad l = 1, \ldots, k, \]
is different from 0 for all \( u \) in the range of the transformation.
Then the joint pdf of $U$ is given by

$$f_U(u) = \sum_{l=1}^{k} \left| J_l \right| f_X(h_{1l}(u), \ldots, h_{ln}(u)).$$

Example 4.3.7:
Let $X, Y$ be iid $\sim N(0, 1)$. Define

$$U = g_1(X, Y) = \begin{cases} X, & Y \neq 0 \\ 0, & Y = 0 \end{cases}$$

and

$$V = g_2(X, Y) = \left| Y \right| .$$

$\mathcal{X} = \mathbb{R}^2$, but $U, V$ are not 1–to–1 mappings from $\mathcal{X}$ onto $B$ since $(U, V)(x, y) = (U, V)(-x, -y)$, i.e., conditions do not apply for the use of Theorem 4.3.5. Let

$$\mathcal{X}_0 = \{(x, y) : y = 0\}$$

$$\mathcal{X}_1 = \{(x, y) : y > 0\}$$

$$\mathcal{X}_2 = \{(x, y) : y < 0\}$$

Then $P((X, Y) \in \mathcal{X}_0) = 0$.

Let $B = \{(u, v) : v > 0\} = g(\mathcal{X}_1) = g(\mathcal{X}_2)$.

Inverses:

$$B \rightarrow \mathcal{X}_1 : x = h_{11}(u, v) = uv$$
$$y = h_{12}(u, v) = v$$

$$B \rightarrow \mathcal{X}_2 : x = h_{21}(u, v) = -uv$$
$$y = h_{22}(u, v) = -v$$

$$J_1 = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v \implies |J_1| = |v|$$

$$J_2 = \begin{vmatrix} -v & -u \\ 0 & -1 \end{vmatrix} = v \implies |J_2| = |v|$$

$$f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-x^2/2} e^{-y^2/2}$$

$$f_{U,V}(u, v) = \left| v \right| \frac{1}{2\pi} e^{-(uv)^2/2} e^{-v^2/2} + \left| v \right| \frac{1}{2\pi} e^{-(uv)^2/2} e^{-(v)^2/2}$$

$$= \frac{v}{\pi} e^{-\left(u^2+1\right)v^2/2}, \quad -\infty < u < \infty, \quad 0 < v < \infty$$

95
Marginal:

$$f_U(u) = \int_{0}^{\infty} \frac{1}{\pi(u^2 + 1)} e^{-\left(\frac{u^2 + 1}{2}\right)} dv \quad | \quad z = \frac{(u^2 + 1)v^2}{2}, \quad \frac{dz}{dv} = (u^2 + 1)v$$

$$= \int_{0}^{\infty} \frac{1}{\pi(u^2 + 1)} e^{-z} dz$$

$$= \left. \frac{1}{\pi(u^2 + 1)} (-e^{-z}) \right|_{0}^{\infty}$$

$$= \frac{1}{\pi(1 + u^2)}, \quad -\infty < u < \infty$$

Thus, the ratio of two iid $N(0,1)$ rv’s is a rv that has a Cauchy distribution.
4.4 Order Statistics

(Based on Casella/Berger, Section 5.4)

Definition 4.4.1:
Let \((X_1, \ldots, X_n)\) be an \(n\)-rv. The \(k\)th order statistic \(X_{(k)}\) is the \(k\)th smallest of the \(X_i\)'s, i.e., \(X_{(1)} = \min\{X_1, \ldots, X_n\}\), \(X_{(2)} = \min\{\{X_1, \ldots, X_n\} \setminus X_{(1)}\}\), \ldots, \(X_{(n)} = \max\{X_1, \ldots, X_n\}\). It is \(X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}\) and \(\{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}\) is the set of order statistics for \((X_1, \ldots, X_n)\). ■

Note:
As shown in Theorem 4.3.2, \(X_{(1)}\) and \(X_{(n)}\) are rv's. This result will be extended in the following Theorem:

Theorem 4.4.2:
Let \((X_1, \ldots, X_n)\) be an \(n\)-rv. Then the \(k\)th order statistic \(X_{(k)}\), \(k = 1, \ldots, n\), is also a rv. ■

Theorem 4.4.3:
Let \(X_1, \ldots, X_n\) be continuous iid rv's with pdf \(f_X\). The joint pdf of \(X_{(1)}, \ldots, X_{(n)}\) is

\[
f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^{n} f_X(x_i), & x_1 \leq x_2 \leq \ldots \leq x_n \\ 0, & \text{otherwise} \end{cases}
\]

Proof:
For the case \(n = 3\), look at the following scenario how \(X_1, X_2,\) and \(X_3\) can be possibly ordered to yield \(X_{(1)} < X_{(2)} < X_{(3)}\). Columns represent \(X_{(1)}, X_{(2)},\) and \(X_{(3)}\). Rows represent \(X_1, X_2,\) and \(X_3\):

<table>
<thead>
<tr>
<th>(X_{(1)})</th>
<th>(X_{(2)})</th>
<th>(X_{(3)})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X_1)</td>
<td>1 0 0</td>
<td></td>
</tr>
<tr>
<td>(k = 1): (X_1 &lt; X_2 &lt; X_3) : (X_2)</td>
<td>0 1 0</td>
<td></td>
</tr>
<tr>
<td>(X_3)</td>
<td>0 0 1</td>
<td></td>
</tr>
<tr>
<td>(k = 2): (X_1 &lt; X_3 &lt; X_2) :</td>
<td>1 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 0 1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 1 0</td>
<td></td>
</tr>
<tr>
<td>(k = 3): (X_2 &lt; X_1 &lt; X_3) :</td>
<td>0 1 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 0 0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0 0 1</td>
<td></td>
</tr>
</tbody>
</table>
For \( n = 3 \), there are \( 3! = 6 \) possible arrangements.

For example, if \( k = 2 \), we have

\[
X_1 < X_3 < X_2
\]

with corresponding inverse

\[
X_1 = X_{(1)}, \quad X_2 = X_{(3)}, \quad X_3 = X_{(2)}
\]

and

\[
J_2 = \det \begin{pmatrix}
\frac{\partial x_1}{\partial x_{(1)}} & \frac{\partial x_1}{\partial x_{(2)}} & \frac{\partial x_1}{\partial x_{(3)}} \\
\frac{\partial x_2}{\partial x_{(1)}} & \frac{\partial x_2}{\partial x_{(2)}} & \frac{\partial x_2}{\partial x_{(3)}} \\
\frac{\partial x_3}{\partial x_{(1)}} & \frac{\partial x_3}{\partial x_{(2)}} & \frac{\partial x_3}{\partial x_{(3)}}
\end{pmatrix} = \det \begin{pmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{pmatrix}
\]

with \( |J_2| = 1 \).

In general, there are \( n! \) arrangements of \( X_1, \ldots, X_n \) for each \( (X_{(1)}, \ldots, X_{(n)}) \). This mapping is not 1–to–1. For each mapping, we have a \( n \times n \) matrix \( J_k \) that results from an \( n \times n \) identity matrix through the rearrangement of rows. Therefore, \( J_k = \pm 1 \) and \( |J_k| = 1 \). By Theorem 4.3.6, we get for \( x_1 \leq x_2 \leq \ldots \leq x_n \)

\[
f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = f_{X_{(1)},\ldots,X_{(n)}}(x_{(1)}, \ldots, x_{(n)})
\]

\[
= \sum_{k=1}^{n!} |J_k| f_{X_1,\ldots,X_n}(x_{(k_1)}, x_{(k_2)}, \ldots, x_{(k_n)})
\]

\[
= n! f_{X_1,\ldots,X_n}(x_{(k_1)}, x_{(k_2)}, \ldots, x_{(k_n)})
\]

\[
= n! \prod_{i=1}^{n} f_{X_i}(x_{(k_i)})
\]

\[
= n! \prod_{i=1}^{n} f_X(x_i)
\]

\( \blacksquare \)
Theorem 4.4.4: Let $X_1, \ldots, X_n$ be continuous iid rv’s with pdf $f_X$ and cdf $F_X$. Then the following holds:

(i) The marginal pdf of $X_{(k)}$, $k = 1, \ldots, n$, is

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}(F_X(x))^{k-1}(1-F_X(x))^{n-k}f_X(x).$$

(ii) The joint pdf of $X_{(j)}$ and $X_{(k)}$, $1 \leq j < k \leq n$, is

$$f_{X_{(j)},X_{(k)}}(x_j, x_k) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} \times$$

$$\left( F_X(x_j) \right)^{j-1} \left( F_X(x_k) - F_X(x_j) \right)^{k-j-1} (1-F_X(x_k))^{n-k} f_X(x_j) f_X(x_k)$$

if $x_j < x_k$ and 0 otherwise.

\[\blacksquare\]
4.5 Multivariate Expectation

(Based on Casella/Berger, Sections 4.2, 4.6 & 4.7)

In this section, we assume that $X' = (X_1, \ldots, X_n)$ is an n–rv and $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a Borel–measurable function.

**Definition 4.5.1:**

If $m = 1$, i.e., $g$ is univariate, we define the following:

(i) Let $X$ be discrete with joint pmf $p_{i_1, \ldots, i_n} = P(X_1 = x_{i_1}, \ldots, X_n = x_{i_n})$. If
$$\sum_{i_1, \ldots, i_n} p_{i_1, \ldots, i_n} \cdot |g(x_{i_1}, \ldots, x_{i_n})| < \infty,$$
we define $E(g(X)) = \sum_{i_1, \ldots, i_n} p_{i_1, \ldots, i_n} \cdot g(x_{i_1}, \ldots, x_{i_n})$ and this value exists.

(ii) Let $X$ be continuous with joint pdf $f_X(x)$. If
$$\int_{\mathbb{R}^n} |g(x)| f_X(x) \, dx < \infty,$$
we define $E(g(X)) = \int_{\mathbb{R}^n} g(x) f_X(x) \, dx$ and this value exists.

Note:
The above can be extended to vector–valued functions $g \ (n > 1)$ in the obvious way. For example, if $g$ is the identity mapping from $\mathbb{R}^n \rightarrow \mathbb{R}^n$, then
$$E(X) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$
provided that $E(|X_i|) < \infty \ \forall i = 1, \ldots, n$.

Similarly, provided that all expectations exist, we get for the variance–covariance matrix:
$$Var(X) = \Sigma_X = E((X - E(X))(X - E(X))')$$
with $(i, j)^{th}$ component
$$E((X_i - E(X_i))(X_j - E(X_j))) = Cov(X_i, X_j)$$
and with $(i, i)^{th}$ component
$$E((X_i - E(X_i))(X_i - E(X_i))) = Var(X_i) = \sigma_i^2.$$

The correlation $\rho_{ij}$ of $X_i$ and $X_j$ is defined as
$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j}.$$
Joint higher-order moments can be defined similarly when needed.

Note:
We are often interested in (weighted) sums of rv’s or products of rv’s and their expectations. This will be addressed in the next two Theorems.

**Theorem 4.5.2:**
Let \( X_i, i = 1, \ldots, n \), be rv’s such that \( E(|X_i|) < \infty \). Let \( a_1, \ldots, a_n \in \mathbb{R} \) and define \( S = \sum_{i=1}^{n} a_i X_i \). Then it holds that \( E(|S|) < \infty \) and
\[
E(S) = \sum_{i=1}^{n} a_i E(X_i).
\]

**Proof:**
Continuous case only:
\[
E(|S|) = \int_{\mathbb{R}^n} \left| \sum_{i=1}^{n} a_i x_i \right| f_X(x) dx 
\]
\[
\leq \int_{\mathbb{R}^n} \sum_{i=1}^{n} |a_i| \cdot |x_i| f_X(x) dx 
\]
\[
= \sum_{i=1}^{n} |a_i| \int_{\mathbb{R}} |x_i| \left( \int_{\mathbb{R}^{n-1}} f_X(x) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n \right) dx_i 
\]
\[
= \sum_{i=1}^{n} |a_i| \int_{\mathbb{R}} |x_i| f_{X_i}(x_i) dx_i 
\]
\[
= \sum_{i=1}^{n} |a_i| E(|X_i|) 
\]
\[
< \infty
\]
It follows that \( E(S) = \sum_{i=1}^{n} a_i E(X_i) \) by the same argument without using the absolute values.

Note:
If \( X_i, i = 1, \ldots, n \), are iid with \( E(X_i) = \mu \), then
\[
E(\bar{X}) = E\left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \frac{1}{n} E(X_i) = \mu.
\]
Theorem 4.5.3:
Let $X_i, i = 1, \ldots, n$, be independent rv’s such that $E(|X_i|) < \infty$. Let $g_i, i = 1, \ldots, n$, be Borel–measurable functions. Then

$$E(\prod_{i=1}^{n} g_i(X_i)) = \prod_{i=1}^{n} E(g_i(X_i))$$

if all expectations exist.

Proof:
By Theorem 4.2.5, $f_X(x) = \prod_{i=1}^{n} f_{X_i}(x_i)$, and by Theorem 4.2.7, $g_i(X_i), i = 1, \ldots, n$, are also independent. Therefore,

$$E(\prod_{i=1}^{n} g_i(X_i)) = \int_{\mathbb{R}^n} \left( \prod_{i=1}^{n} g_i(x_i) \right) f_X(x) dx = \prod_{i=1}^{n} \int_{\mathbb{R}} g_i(x_i) f_{X_i}(x_i) dx_i$$

Th. 4.2.5

$$= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{n} g_i(x_i) \prod_{i=1}^{n} f_{X_i}(x_i) \prod_{i=1}^{n} dx_i$$

Th. 4.2.7

$$= \prod_{i=1}^{n} \int_{\mathbb{R}} g_i(x_i) f_{X_i}(x_i) dx_i$$

$$= \prod_{i=1}^{n} E(g_i(X_i))$$

$\blacksquare$

Corollary 4.5.4:
If $X, Y$ are independent, then $Cov(X, Y) = 0$.

$\blacksquare$

Theorem 4.5.5:
Two rv’s $X, Y$ are independent iff for all pairs of Borel–measurable functions $g_1$ and $g_2$ it holds that $E(g_1(X) \cdot g_2(Y)) = E(g_1(X)) \cdot E(g_2(Y))$ if all expectations exist.

Proof:
"$\Rightarrow$": It follows from Theorem 4.5.3 and the independence of $X$ and $Y$ that

$$E(g_1(X)g_2(Y)) = E(g_1(X)) \cdot E(g_2(Y)).$$

"$\Leftarrow$": From Theorem 4.2.6, we know that $X$ and $Y$ are independent iff

$$P(X \in A_1, Y \in A_2) = P(X \in A_1) P(Y \in A_2) \quad \forall \text{Borel sets } A_1 \text{ and } A_2.$$
How do we relate Theorem 4.2.6 to \( g_1 \) and \( g_2 \)? Let us define two Borel-measurable functions \( g_1 \) and \( g_2 \) as:

\[
\begin{align*}
g_1(x) &= I_{A_1}(x) = \begin{cases} 1, & x \in A_1 \\ 0, & \text{otherwise} \end{cases} \\
g_2(y) &= I_{A_2}(y) = \begin{cases} 1, & y \in A_2 \\ 0, & \text{otherwise} \end{cases}
\end{align*}
\]

Then,

\[
\begin{align*}
E(g_1(X)) &= 0 \cdot P(X \in A_1^c) + 1 \cdot P(X \in A_1) = P(X \in A_1), \\
E(g_2(Y)) &= 0 \cdot P(Y \in A_2^c) + 1 \cdot P(Y \in A_2) = P(Y \in A_2)
\end{align*}
\]

and

\[
E(g_1(X) \cdot g_2(Y)) = \begin{cases} 
= P(X \in A_1, Y \in A_2). & \text{given} \\
\end{cases}
\]

\[
\Rightarrow P(X \in A_1, Y \in A_2) = E(g_1(X) \cdot g_2(Y)) \\
= E(g_1(X)) \cdot E(g_2(Y)) \\
= P(X \in A_1)P(Y \in A_2)
\]

\[
\Rightarrow X, Y \text{ independent by Theorem 4.2.6.}
\]

**Definition 4.5.6:**

The \((i_1^{th}, i_2^{th}, \ldots, i_n^{th})\) **multi–way moment** of \( X = (X_1, \ldots, X_n) \) is defined as

\[
m_{i_1i_2\ldots i_n} = E(X_1^{i_1}X_2^{i_2}\ldots X_n^{i_n})
\]

if it exists.

The \((i_1^{th}, i_2^{th}, \ldots, i_n^{th})\) **multi–way central moment** of \( X = (X_1, \ldots, X_n) \) is defined as

\[
\mu_{i_1i_2\ldots i_n} = E(\prod_{j=1}^{n} (X_j - E(X_j))^{i_j})
\]

if it exists.

**Note:**

If we set \( i_r = i_s = 1 \) and \( i_j = 0 \ \forall j \neq r, s \) in Definition 4.5.6 for the multi–way central moment, we get

\[
\mu_{00\ldots010\ldots010\ldots0} = \mu_{rs} = Cov(X_r, X_s).
\]
**Theorem 4.5.7: Cauchy–Schwarz Inequality**

Let $X, Y$ be 2 rv's with finite variance. Then it holds:

(i) $\text{Cov}(X, Y)$ exists.

(ii) $(E(XY))^2 \leq E(X^2)E(Y^2)$.

(iii) $(E(XY))^2 = E(X^2)E(Y^2)$ iff there exists an $(\alpha, \beta) \in \mathbb{R}^2 - \{(0, 0)\}$ such that

$$P(\alpha X + \beta Y = 0) = 1.$$ 

**Proof:**

Assumptions: $\text{Var}(X), \text{Var}(Y) < \infty$. Then also $E(X^2), E(X), E(Y^2), E(Y) < \infty$.

Result used in proof:

$$0 \leq (a - b)^2 = a^2 - 2ab + b^2 \implies ab \leq \frac{a^2 + b^2}{2}$$

$$0 \leq (a + b)^2 = a^2 + 2ab + b^2 \implies -ab \leq \frac{a^2 + b^2}{2}$$

$$\implies |ab| \leq \frac{a^2 + b^2}{2} \forall a, b \in \mathbb{R} \quad (\ast)$$

(i)

$$E(|XY|) = \int_{\mathbb{R}^2} |xy|f_{X,Y}(x,y)dx \, dy \leq \int_{\mathbb{R}^2} \frac{x^2 + y^2}{2}f_{X,Y}(x,y)dx \, dy$$

$$= \int_{\mathbb{R}^2} \frac{x^2}{2}f_{X,Y}(x,y)dx \, dy + \int_{\mathbb{R}^2} \frac{y^2}{2}f_{X,Y}(x,y)dy \, dx$$

$$= \int_{\mathbb{R}^2} \frac{x^2}{2}f_X(x)dx + \int_{\mathbb{R}^2} \frac{y^2}{2}f_Y(y)dy$$

$$= \frac{E(X^2)}{2} + \frac{E(Y^2)}{2}$$

$$< \infty$$

$$\implies E(XY) \text{ exists}$$

$$\implies \text{Cov}(X, Y) = E(XY) - E(X)E(Y) \text{ exists}$$

(ii) $0 \leq E((\alpha X + \beta Y)^2) = \alpha^2 E(X^2) + 2\alpha\beta E(XY) + \beta^2 E(Y^2) \quad \forall \alpha, \beta \in \mathbb{R} \quad (A)$

If $E(X^2) = 0$, then $X$ has a degenerate 1–point Dirac(0) distribution and the inequality trivially is true. Therefore, we can assume that $E(X^2) > 0$. As $(A)$ is true for all $\alpha, \beta \in \mathbb{R}$, we can choose $\alpha = \frac{-E(XY)}{E(X^2)}, \beta = 1$.
\[ \Rightarrow \left( \frac{E(XY)^2}{E(X^2)} - 2 \frac{(E(XY))^2}{E(X^2)} + E(Y^2) \right) \geq 0 \]
\[ \Rightarrow -(E(XY))^2 + E(Y^2)E(X^2) \geq 0 \]
\[ \Rightarrow (E(XY))^2 \leq E(X^2)E(Y^2) \]

(iii) When are the left and right sides of the inequality in (ii) equal?

Assume that \( E(X^2) > 0 \). \((E(XY))^2 = E(X^2)E(Y^2)\) holds iff \( E((\alpha X + \beta Y)^2) = 0 \) based on (ii). It is therefore sufficient to show that \( E((\alpha X + \beta Y)^2) = 0 \) iff \( P(\alpha X + \beta Y = 0) = 1 \):

“\(\Rightarrow\)”:

Let \( Z = \alpha X + \beta Y \). Since \( E((\alpha X + \beta Y)^2) = E(Z^2) = Var(Z) + (E(Z))^2 = 0 \) and \( Var(Z) \geq 0 \) and \( (E(Z))^2 \geq 0 \), it follows that \( E(Z) = 0 \) and \( Var(Z) = 0 \).

This means that \( Z \) has a degenerate 1–point Dirac(0) distribution with \( P(Z = 0) = P(\alpha X + \beta Y = 0) = 1 \).

“\(\Leftarrow\)”:

If \( P(\alpha X + \beta Y = 0) = P(Y = -\frac{\alpha}{\beta}X) = 1 \) for some \((\alpha, \beta) \in \mathbb{R}^2 - \{(0,0)\} \), i.e., \( Y \) is linearly dependent on \( X \) with probability 1, this implies:

\[ (E(XY))^2 = (E(X \cdot \frac{-\alpha X}{\beta}))^2 = (\frac{\alpha}{\beta})^2 (E(X^2))^2 = E(X^2) (\frac{\alpha}{\beta})^2 E(X^2) = E(X^2) E(Y^2) \]
4.6 Multivariate Generating Functions

(Based on Casella/Berger, Sections 4.2 & 4.6)

Definition 4.6.1:
Let $X' = (X_1, \ldots, X_n)$ be an n–rv. We define the multivariate moment generating function (mmgf) of $X$ as

$$M_X(t) = E(e^{t'X}) = E \left( \exp \left( \sum_{i=1}^{n} t_i X_i \right) \right)$$

if this expectation exists for $|t| = \sqrt{\sum_{i=1}^{n} t_i^2} < h$ for some $h > 0$.

Definition 4.6.2:
Let $X' = (X_1, \ldots, X_n)$ be an n–rv. We define the $n$–dimensional characteristic function $\Phi_X$ of $X$ as

$$\Phi_X(t) = E(e^{it'X}) = E \left( \exp \left( i \sum_{j=1}^{n} t_j X_j \right) \right).$$

Note:
(i) $\Phi_X(t)$ exists for any real–valued n–rv.
(ii) If $M_X(t)$ exists, then $\Phi_X(t) = M_X(it)$.

Theorem 4.6.3:
(i) If $M_X(t)$ exists, it is unique and uniquely determines the joint distribution of $X$. $\Phi_X(t)$ is also unique and uniquely determines the joint distribution of $X$.
(ii) $M_X(t)$ (if it exists) and $\Phi_X(t)$ uniquely determine all marginal distributions of $X$, i.e., $M_{X_i}(t_i) = M_X(0, t_i, 0)$ and and $\Phi_{X_i}(t_i) = \Phi_X(0, t_i, 0)$.
(iii) Joint moments of all orders (if they exist) can be obtained as

$$m_{i_1 \ldots i_n} = \left. \frac{\partial^{i_1+i_2+\ldots+i_n}}{\partial t_1^{i_1} \partial t_2^{i_2} \ldots \partial t_n^{i_n}} M_X(t) \right|_{t=0} = E(X_{i_1}^{i_1} X_{i_2}^{i_2} \ldots X_{i_n}^{i_n})$$

if the mmgf exists and

$$m_{i_1 \ldots i_n} = \frac{1}{i_1+i_2+\ldots+i_n} \left. \frac{\partial^{i_1+i_2+\ldots+i_n}}{\partial t_1^{i_1} \partial t_2^{i_2} \ldots \partial t_n^{i_n}} \Phi_X(t) \right|_{t=0} = E(X_{i_1}^{i_1} X_{i_2}^{i_2} \ldots X_{i_n}^{i_n}).$$
(iv) $X_1, \ldots, X_n$ are independent rv’s iff

$$M_X(t_1, \ldots, t_n) = M_X(t_1, 0) \cdot M_X(0, t_2, 0) \cdot \ldots \cdot M_X(0, t_n) \quad \forall t_1, \ldots, t_n \in \mathbb{R},$$

given that $M_X(t)$ exists.

Similarly, $X_1, \ldots, X_n$ are independent rv’s iff

$$\Phi_X(t_1, \ldots, t_n) = \Phi_X(t_1, 0) \cdot \Phi_X(0, t_2, 0) \cdot \ldots \cdot \Phi_X(0, t_n) \quad \forall t_1, \ldots, t_n \in \mathbb{R}.$$

**Proof:**

Rohatgi, page 162: Theorem 7, Corollary, Theorem 8, and Theorem 9 (for mmgf and the case $n = 2$).

**Theorem 4.6.4:**

Let $X_1, \ldots, X_n$ be independent rv’s.

(i) If mgf’s $M_{X_1}(t), \ldots, M_{X_n}(t)$ exist, then the mgf of $Y = \sum_{i=1}^{n} a_i X_i$ is

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t) \quad [\text{Note: } t_i]$$

on the common interval where all individual mgf’s exist.

(ii) The characteristic function of $Y = \sum_{j=1}^{n} a_j X_j$ is

$$\Phi_Y(t) = \prod_{j=1}^{n} \Phi_{X_j}(a_j t) \quad [\text{Note: } t_j]$$

(iii) If mgf’s $M_{X_1}(t), \ldots, M_{X_n}(t)$ exist, then the mmgf of $X$ is

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(t_i) \quad [\text{Note: } t_i]$$

on the common interval where all individual mgf’s exist.

(iv) The n–dimensional characteristic function of $X$ is

$$\Phi_X(t) = \prod_{j=1}^{n} \Phi_{X_j}(t_j). \quad [\text{Note: } t_j]$$
Proof:
Homework (parts (ii) and (iv) only)

Theorem 4.6.5:
Let $X_1, \ldots, X_n$ be independent discrete rv’s on the non–negative integers with pgf’s $G_{X_1}(s), \ldots, G_{X_n}(s)$.
The pgf of $Y = \sum_{i=1}^{n} X_i$ is

$$G_Y(s) = \prod_{i=1}^{n} G_{X_i}(s).$$

Proof:

Version 1:

$$G_{X_i}(s) = \sum_{k=0}^{\infty} P(X_i = k)s^k = E(s^{X_i})$$
$$G_Y(s) = E(s^{Y}) = E(s^{\sum_{i=1}^{n} X_i}) = E(\prod_{i=1}^{n} s^{X_i})$$

Th. 4.5.3

$$= \prod_{i=1}^{n} E(s^{X_i}) = \prod_{i=1}^{n} G_{X_i}(s)$$

Version 2: (case $n = 2$ only)

$$G_Y(s) = P(Y = 0) + P(Y = 1)s + P(Y = 2)s^2 + \ldots$$
$$= P(X_1 = 0, X_2 = 0) + \left(P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1)\right)s + \left(P(X_1 = 2, X_2 = 0) + P(X_1 = 1, X_2 = 1) + P(X_1 = 0, X_2 = 2)\right)s^2 + \ldots$$

indcp.

$$= P(X_1 = 0)P(X_2 = 0) + (P(X_1 = 1)P(X_2 = 0) + P(X_1 = 0)P(X_2 = 1))s + (P(X_1 = 2)P(X_2 = 0) + P(X_1 = 1)P(X_2 = 1) + P(X_1 = 0)P(X_2 = 2))s^2 + \ldots$$

$$= \left(P(X_1 = 0) + P(X_1 = 1)s + P(X_1 = 2)s^2 + \ldots\right) \cdot \left(P(X_2 = 0) + P(X_2 = 1)s + P(X_2 = 2)s^2 + \ldots\right)$$

$$= G_{X_1}(s) \cdot G_{X_2}(s)$$
A generalized proof for $n \geq 3$ needs to be done by induction on $n$.

**Theorem 4.6.6:**

Let $X_1, \ldots, X_N$ be iid discrete rv’s on the non–negative integers with common pgf $G_X(s)$. Let $N$ be a discrete rv on the non–negative integers with pgf $G_N(s)$. Let $N$ be independent of the $X_i$’s. Define $S_N = \sum_{i=1}^{N} X_i$. The pgf of $S_N$ is

$$G_{S_N}(s) = G_N(G_X(s)).$$

**Proof:**

$$P(S_N = k) \overset{Th. 4.6.5}{=} \sum_{n=0}^{\infty} P(S_N = k|N = n) \cdot P(N = n)$$

$$\implies G_{S_N}(s) = \sum_{k=0}^{\infty} P(S_N = k) \cdot s^k$$

$$= \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} P(S_N = k|N = n) \cdot P(N = n) \cdot s^k$$

$$= \sum_{n=0}^{\infty} P(N = n) \sum_{k=0}^{\infty} P(S_N = k|N = n) \cdot s^k$$

$$= \sum_{n=0}^{\infty} P(N = n) \sum_{k=0}^{\infty} P(S_n = k) \cdot s^k$$

$$= \sum_{n=0}^{\infty} P(N = n) \cdot G_{S_n}(s)$$

$$\overset{Th. 4.6.5}{=} \sum_{n=0}^{\infty} P(N = n) \prod_{i=1}^{n} G_{X_i}(s)$$

$$\overset{iid}{=} \sum_{n=0}^{\infty} P(N = n) \cdot (G_X(s))^n$$

$$= G_N(G_X(s))$$

**Example 4.6.7:**

Starting with a single cell at time 0, after one time unit there is probability $p$ that the cell will have split (2 cells), probability $q$ that it will survive without splitting (1 cell), and probability $r$ that it will have died (0 cells). It holds that $p, q, r \geq 0$ and $p + q + r = 1$. Any surviving cells have the same probabilities of splitting or dying. What is the pgf for the # of cells at time 2?
Example 4.6.7

\[ G_X(s) = G_N(s) = ps^2 + qs + r = G_X^{(0)}(s) \]

\[ G_X^{(1)}(s) = 2ps + q \]

\[ G_X^{(2)}(s) = 2p \]

By Th. 3.4.3,

\[ P(X = k) = \frac{G^{(k)}(s)}{k!} \bigg|_{s=0} \]

Therefore,

\[
\begin{align*}
P(X = 0) &= \frac{G_X^{(0)}(0)}{0!} = r \\
P(X = 1) &= \frac{G_X^{(1)}(0)}{1!} = q \\
P(X = 2) &= \frac{G_X^{(2)}(0)}{2!} = p \\
\end{align*}
\]

\[ S_N = \sum_{i=1}^{N} X_i \]

\[ G_{S_N}(s) = p(ps^2 + qs + r)^2 + q(ps^2 + qs + r) + r = G_{S_N}^{(0)}(s) \]

\[ \frac{G_{S_N}^{(0)}(0)}{0!} = pr^2 + qr + r \]

etc.
Theorem 4.6.8:
Let $X_1, \ldots, X_N$ be iid rv’s with common mgf $M_X(t)$. Let $N$ be a discrete rv on the non-negative integers with mgf $M_N(t)$. Let $N$ be independent of the $X_i$’s. Define $S_N = \sum_{i=1}^N X_i$. The mgf of $S_N$ is

$$M_{S_N}(t) = M_N(\ln M_X(t)).$$

Proof:
Consider the case that the $X_i$’s are non-negative integers:

We know that

$$G_X(s) = E(s^X) = E(e^{\ln s^X}) = E(e^{(\ln s)\cdot X}) = M_X(\ln s) \quad (s)$$

$$\implies M_X(s) = G_X(e^s)$$

$$\implies M_{S_N}(t) = G_{S_N}(e^t) = G_N(G_X(e^t)) = G_N(M_X(t)) \overset{Th.4.6.6}{=} M_N(\ln M_X(t))$$

In the general case, i.e., if the $X_i$’s are not non-negative integers, we need results from Section 4.7 (conditional expectation) to prove this Theorem.

$\blacksquare$
4.7 Conditional Expectation

(Based on Casella/Berger, Section 4.4)

In Section 4.1, we established that the conditional pmf of \( X \) given \( Y = y_j \) (for \( P_Y(y_j) > 0 \)) is a pmf. For continuous rv’s \( X \) and \( Y \), when \( f_Y(y) > 0 \), \( f_{X|Y}(x | y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \), and \( f_{X,Y} \) and \( f_Y \) are continuous, then \( f_{X|Y}(x | y) \) is a pdf and it is the conditional pdf of \( X \) given \( Y = y \).

Definition 4.7.1:
Let \( X, Y \) be rv’s on \((\Omega, L, P)\). Let \( h \) be a Borel–measurable function. Then the conditional expectation of \( h(X) \) given \( Y \), i.e., \( E(h(X) | Y) \), is a rv that takes the value \( E(h(X) | y) \). It is defined as

\[
E(h(X) | y) = \begin{cases} 
\sum_{x \in \mathcal{X}} h(x)P(X = x | Y = y), & \text{if } (X, Y) \text{ is discrete and } P(Y = y) > 0 \\
\int_{-\infty}^{\infty} h(x)f_{X|Y}(x | y)dx, & \text{if } (X, Y) \text{ is continuous and } f_Y(y) > 0
\end{cases}
\]

Note:
Depending on the source, two different definitions of the conditional expectation exist:
(i) Casella and Berger (2002, p. 150), Miller and Miller (1999, p. 161), and Rohatgi and Saleh (2001, p. 165) do not require that \( E(h(X)) \) exists. (ii) Rohatgi (1976, p. 168) and Bickel and Doksum (2001, p. 479) require that \( E(h(X)) \) exists.

In case of the rv’s \( X \) and \( Y \) with joint pdf

\[ f_{X,Y}(x,y) = xe^{-x(y+1)}I_{[0,\infty)}(x)I_{[0,\infty)}(y), \]

it holds in case (i) that \( E(Y | X) = \frac{1}{X} \) even though \( E(Y) \) does not exist (see Rohatgi and Saleh 2001, p. 168, for details), whereas in case (ii), the conditional expectation does not exist!

Note:

(i) The rv \( E(h(X) | Y) = g(Y) \) is a function of \( Y \) as a rv.

(ii) The usual properties of expectations apply to the conditional expectation, given the conditional expectations exist:

(a) \( E(c | Y) = c \ \forall c \in \mathbb{R} \).
(b) \( E(aX + b | Y) = aE(X | Y) + b \ \forall a, b \in \mathbb{R} \).
(c) If \( g_1, g_2 \) are Borel–measurable functions and if \( E(g_1(X) \mid Y), E(g_2(X) \mid Y) \) exist, then 
\[
E(a_1 g_1(X) + a_2 g_2(X) \mid Y) = a_1 E(g_1(X) \mid Y) + a_2 E(g_2(X) \mid Y) \quad \forall a_1, a_2 \in \mathbb{R}.
\]

(d) If \( X \geq 0 \), i.e., \( P(X \geq 0) = 1 \), then \( E(X \mid Y) \geq 0 \).

(e) If \( X_1 \geq X_2 \), i.e., \( P(X_1 \geq X_2) = 1 \), then \( E(X_1 \mid Y) \geq E(X_2 \mid Y) \).

(iii) Moments are defined in the usual way. If \( E(X^r \mid Y) < \infty \) exists and is the \( r^{th} \) conditional moment of \( X \) given \( Y \).

Example 4.7.2:
Recall Example 4.1.12:
\[
f_{X,Y}(x,y) = \begin{cases} 
2, & 0 < x < y < 1 \\
0, & \text{otherwise}
\end{cases}
\]
The conditional pdf’s \( f_{Y\mid X}(y \mid x) \) and \( f_{X\mid Y}(x \mid y) \) have been calculated as:
\[
f_{Y\mid X}(y \mid x) = \frac{1}{1-x} \quad \text{for } x < y < 1 \quad \text{(where } 0 < x < 1 \text{)}
\]
and
\[
f_{X\mid Y}(x \mid y) = \frac{1}{y} \quad \text{for } 0 < x < y \quad \text{(where } 0 < y < 1 \text{)}.
\]
So,
\[
E(X \mid y) = \int_0^y \frac{x}{y} \, dx = \frac{y}{2}
\]
and
\[
E(Y \mid x) = \int_x^1 \frac{1}{1-x} \, y \, dy = \frac{1}{1-x} \left[ \frac{y^2}{2} \right]_x^1 = \frac{1}{2} \left( 1 - x^2 \right) = \frac{1}{2} + \frac{x}{2}.
\]
Therefore, we get the rv’s \( E(X \mid Y) = \frac{Y}{2} \) and \( E(Y \mid X) = \frac{1+X}{2} \).

Theorem 4.7.3:
If \( E(h(X)) \) exists, then
\[
E_Y(E_{X\mid Y}(h(X) \mid Y)) = E(h(X)).
\]

Proof:
Continuous case only:
\[
E_Y(E_{X\mid Y}(h(X) \mid Y)) = \int_{-\infty}^{\infty} E_{X\mid Y}(h(X) \mid y)f_Y(y) \, dy
\]
\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x)f_{X\mid Y}(x \mid y)f_Y(y) \, dx \, dy
\]

113
\[
= \int_{-\infty}^{\infty} h(x) \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx \\
= \int_{-\infty}^{\infty} h(x)f_X(x) dx \\
= E(h(X)) 
\]

**Theorem 4.7.4:**

If \( E(X^2) \) exists, then

\[
\text{Var}_Y(E(X \mid Y)) + E_Y(\text{Var}(X \mid Y)) = \text{Var}(X).
\]

**Proof:**

\[
\text{Var}_Y(E(X \mid Y)) + E_Y(\text{Var}(X \mid Y)) = E_Y((E(X \mid Y))^2) - (E_Y(E(X \mid Y)))^2 \\
+ E_Y(E(X^2 \mid Y) - (E(X \mid Y))^2)
\]

\[
\text{Th. 4.7.3} \quad E_Y((E(X \mid Y))^2) - (E(X))^2 + E(X^2) - E_Y((E(X \mid Y))^2)
\]

\[
= E(X^2) - (E(X))^2 \\
= \text{Var}(X)
\]

**Note:**

If \( E(X^2) \) exists, then \( \text{Var}(X) \geq \text{Var}_Y(E(X \mid Y)) \). \( \text{Var}(X) = \text{Var}_Y(E(X \mid Y)) \) iff \( X = g(Y) \).

The inequality directly follows from Theorem 4.7.4.

For equality, it is necessary that

\[
E_Y(\text{Var}(X \mid Y)) = E_Y(E((X - E(X \mid Y))^2 \mid Y)) = E_Y(E(X^2 \mid Y) - (E(X \mid Y))^2) = 0
\]

which holds if \( X = E(X \mid Y) = g(Y) \).

If \( X, Y \) are independent, \( F_{X \mid Y}(x \mid y) = F_X(x) \ \forall x \).

Thus, if \( E(h(X)) \) exists, then \( E(h(X) \mid Y) = E(h(X)) \).
**Proof:** (of Theorem 4.6.8)

\[ M_{SN}(t) \overset{Def.}{=} E(e^{tSN}) \]

\[
= E \left( \exp \left( t \sum_{i=1}^{N} X_i \right) \right)
\]

\[
\overset{\text{Th. 4.7.3}}{=} E_N \left( E_N \left( \sum_{i=1}^{N} X_i \mid N \right) \left( \exp \left( t \sum_{i=1}^{N} X_i \right) \mid N \right) \right)
\]

First consider

\[
E \left( \sum_{i=1}^{N} X_i \mid N \right) \left( \exp \left( t \sum_{i=1}^{N} X_i \right) \mid n \right) = E \left( \sum_{i=1}^{n} X_i \mid n \right) \left( \exp \left( t \sum_{i=1}^{n} X_i \right) \right)
\]

\[
\overset{\text{Th. 4.6.4(i)}}{=} \prod_{i=1}^{n} M_{X_i}(t)
\]

\[
X_i \overset{iid}{=} \prod_{i=1}^{n} M_{X}(t)
\]

\[
= (M_{X}(t))^n
\]

\[
\Rightarrow M_{SN}(t) = E_N \left( \prod_{i=1}^{N} M_{X}(t) \right)
\]

\[
= E_N((M_{X}(t))^N)
\]

\[
= E_N(\exp(N \ln M_{X}(t)))
\]

\[
\overset{(*)}{=} M_N(\ln M_{X}(t))
\]

\((*)\) holds since \( M_N(k) = E_N(\exp(N \cdot k)) \).
4.8 Inequalities and Identities

(Based on Casella/Berger, Section 4.7)

Lemma 4.8.1:
Let $a, b$ be positive numbers and $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ (i.e., $pq = p + q$ and $q = \frac{p}{p-1}$). Then it holds that

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality iff $a^p = b^q$.

Proof:
Fix $b$. Let

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

$$\implies g'(a) = a^{p-1} - b^q = 0$$

$$\implies b = a^{p-1}$$

$$\implies b^q = a^{(p-1)q} = a^p$$

$$g''(a) = (p - 1)a^{p-2} > 0$$

Since $g''(a) > 0$, this is really a minimum. The minimum value is obtained for $b = a^{p-1}$ and it is

$$\frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p = a^p(\frac{1}{p} + \frac{1}{q} - 1) = 0.$$ 

Since $g''(a) > 0$, the minimum is unique and $g(a) \geq 0$. Therefore,

$$g(a) + ab = \frac{1}{p}a^p + \frac{1}{q}b^q \geq ab.$$  

\[\blacksquare\]

Theorem 4.8.2: Hölder’s Inequality
Let $X, Y$ be 2 rv’s. Let $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ (i.e., $pq = p + q$ and $q = \frac{p}{p-1}$). Then it holds that

$$E(|XY|) \leq (E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}}.$$ 

Proof:
In Lemma 4.8.1, let

$$a = \frac{|X|}{(E(|X|^p))^{\frac{1}{p}}} > 0 \text{ and } b = \frac{|Y|}{(E(|Y|^q))^{\frac{1}{q}}} > 0.$$ 

116
Lemma 4.8.1 \( \frac{1}{p} \frac{X^p}{E(|X|^p)} + \frac{1}{q} \frac{Y^q}{E(|Y|^q)} \geq \frac{|XY|}{(E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}}} \)

Taking expectations on both sides of this inequality, we get

\[
1 = \frac{1}{p} + \frac{1}{q} \geq \frac{E(|XY|)}{(E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}}}
\]

The result follows immediately when multiplying both sides with \( (E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}} \).

Note:
Note that Theorem 4.5.7 (ii) (Cauchy–Schwarz Inequality) is a special case of Theorem 4.8.2 with \( p = q = 2 \). If \( X \sim \text{Dirac}(0) \) or \( Y \sim \text{Dirac}(0) \) and it therefore holds that \( E(|X|) = 0 \) or \( E(|Y|) = 0 \), the inequality trivially holds.

**Theorem 4.8.3: Minkowski’s Inequality**

Let \( X, Y \) be 2 rv’s. Then it holds for \( 1 \leq p < \infty \) that

\[
(E(|X + Y|^p))^{\frac{1}{p}} \leq (E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}}.
\]

**Proof:**

Assume \( p > 1 \) (the inequality trivially holds for \( p = 1 \)):

\[
E(|X + Y|^p) = E(|X + Y| \cdot |X + Y|^{p-1}) \\
\leq E(|X| + |Y| \cdot |X + Y|^{p-1}) \\
= E(|X| \cdot |X + Y|^{p-1}) + E(|Y| \cdot |X + Y|^{p-1}) \\
\overset{\text{Th. 4.8.2}}{\leq} \left( E(|X|^p) \right)^{\frac{1}{p}} \cdot \left( E(|X + Y|^{p-1}) \right)^{\frac{1}{p}} \\
+ \left( E(|Y|^p) \right)^{\frac{1}{p}} \cdot \left( E(|X + Y|^{p-1}) \right)^{\frac{1}{p}} \\
= \left( E(|X|^p) \right)^{\frac{1}{p}} + \left( E(|Y|^p) \right)^{\frac{1}{p}}
\]

Divide the left and right side of this inequality by \( (E(|X + Y|^p))^{\frac{1}{p}} \).

The result on the left side is \( (E(|X + Y|^p))^{1-\frac{1}{p}} = (E(|X + Y|^p))^{\frac{p}{p-1}} \), and the result on the right side is \( (E(|X|^p))^{\frac{1}{p}} + (E(|Y|^p))^{\frac{1}{p}} \). Therefore, Theorem 4.8.3 holds.

In (A), we define \( q = \frac{p}{p-1} \), a condition to meet Hölder’s Inequality. Therefore, \( \frac{1}{p} + \frac{1}{q} = 1 \) and \( (p-1)q = p \).
Definition 4.8.4:
A function \( g(x) \) is **convex** if
\[
g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y) \quad \forall x, y \in \mathbb{R} \quad \forall 0 < \lambda < 1.
\]

Note:
(i) Geometrically, a convex function falls above all of its tangent lines. Also, a connecting line between any pairs of points \((x, g(x))\) and \((y, g(y))\) in the 2-dimensional plane always falls above the curve.
(ii) A function \( g(x) \) is **concave** iff \(-g(x)\) is convex.

Theorem 4.8.5: **Jensen’s Inequality**
Let \( X \) be a rv. If \( g(x) \) is a convex function, then
\[
E(g(X)) \geq g(E(X))
\]
given that both expectations exist.

Proof:
Construct a tangent line \( l(x) \) to \( g(x) \) at the (constant) point \( x_0 = E(X) \):
\[
l(x) = ax + b \text{ for some } a, b \in \mathbb{R}
\]
The function \( g(x) \) is a convex function and falls above the tangent line \( l(x) \)
\[
\Rightarrow g(x) \geq ax + b \quad \forall x \in \mathbb{R}
\]
\[
\Rightarrow E(g(X)) \geq E(ax + b) = aE(X) + b = l(E(X)) \quad \text{tangent point } g(E(X))
\]
Therefore, Theorem 4.8.5 holds.

Note:
Typical convex functions \( g \) are:
(i) \( g_1(x) = |x| \) \( \Rightarrow E(|X|) \geq |E(X)|. \)
(ii) \( g_2(x) = x^2 \) \( \Rightarrow E(X^2) \geq (E(X))^2 \) \( \Rightarrow Var(X) \geq 0. \)
(iii) \( g_3(x) = \frac{1}{x^p} \) for \( x > 0, p > 0 \) \( \Rightarrow E\left(\frac{1}{X^p}\right) \geq \frac{1}{(E(X))^p}; \text{ for } p = 1: E\left(\frac{1}{X}\right) \geq \frac{1}{E(X)} \)
(iv) Other convex functions are \( x^p \) for \( x > 0, p \geq 1; \theta^x \) for \( \theta > 1; -\ln(x) \) for \( x > 0; \) etc.
(v) Recall that if $g$ is convex and differentiable, then $g''(x) \geq 0 \ \forall x$.

(vi) If the function $g$ is concave, the direction of the inequality in Jensen’s Inequality is reversed, i.e., $E(g(X)) \leq g(E(X))$.

(vii) Does it hold that $E\left(\frac{X}{Y}\right)$ equals $\frac{E(X)}{E(Y)}$? The answer is no in most cases! Assuming that $X, Y$ are independent, it is

$$E\left(\frac{X}{Y}\right) \geq E(X) \cdot E\left(\frac{1}{Y}\right) \geq E(X) \cdot \frac{1}{E(Y)}.$$ 

Example 4.8.6:
Given the real numbers $a_1, a_2, \ldots, a_n > 0$, we define

**arithmetic mean** : $a_A = \frac{1}{n} (a_1 + a_2 + \ldots + a_n) = \frac{1}{n} \sum_{i=1}^{n} a_i$

**geometric mean** : $a_G = (a_1 \cdot a_2 \cdot \ldots \cdot a_n)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}}$

**harmonic mean** : $a_H = \frac{1}{\frac{1}{n} \left(\frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n}\right)} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{a_i}}$

Let $X$ be a rv that takes values $a_1, a_2, \ldots, a_n > 0$ with probability $\frac{1}{n}$ each.

(i) $a_A \geq a_G$:

$$\ln(a_A) = \ln\left(\frac{1}{n} \sum_{i=1}^{n} a_i\right) = \ln(E(X))$$

By ln concave

$$\geq E(\ln(X))$$

$$= \frac{1}{n} \sum_{i=1}^{n} \ln(a_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \ln(a_i)$$

$$= \frac{1}{n} \ln(\prod_{i=1}^{n} a_i)$$

$$= \ln\left(\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}}\right)$$

$$= \ln(a_G)$$

Taking the anti-log of both sides gives $a_A \geq a_G$. 

119
(ii) $a_A \geq a_H$:

\[
\frac{1}{a_A} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} a_i} = \frac{1}{E(X)}
\]

$1/X_{\text{convex}} \leq E(\frac{1}{X})$

\[
\leq \sum_{i=1}^{n} \frac{1}{n} \frac{1}{a_i} = \frac{1}{n} \left( \frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n} \right) = \frac{1}{a_H}
\]

Inverting both sides gives $a_A \geq a_H$.

(iii) $a_G \geq a_H$:

\[
-\ln(a_H) = \ln\left(\frac{1}{a_H} \right) = \ln\left(\frac{1}{\frac{1}{n} \sum_{i=1}^{n} a_i} \right) = \ln\left(\frac{1}{\frac{1}{n} \prod_{i=1}^{n} a_i} \right) = \ln\left(\frac{1}{E(\frac{1}{X})} \right)
\]

$\ln \text{concave} \geq E(\ln(\frac{1}{X}))$

\[
\geq \sum_{i=1}^{n} \frac{1}{n} \ln(\frac{1}{a_i}) = \frac{1}{n} \sum_{i=1}^{n} \ln(\frac{1}{a_i}) = \frac{1}{n} \sum_{i=1}^{n} -\ln a_i = -\frac{1}{n} \ln(\prod_{i=1}^{n} a_i) = -\ln(\prod_{i=1}^{n} a_i)^{\frac{1}{n}} = -\ln a_G
\]

Multiplying both sides with $-1$ gives $\ln a_H \leq \ln a_G$. Then taking the anti-log of both sides gives $a_H \leq a_G$. 

120
In summary, $a_H \leq a_G \leq a_A$. Note that it would have been sufficient to prove steps (i) and (iii) only to establish this result. However, step (ii) has been included to provide another example how to apply Theorem 4.8.5.

**Theorem 4.8.7: Covariance Inequality**

Let $X$ be a rv with finite mean $\mu$.

(i) If $g(x)$ is non-decreasing, then

$$E(g(X)(X - \mu)) \geq 0$$

if this expectation exists.

(ii) If $g(x)$ is non-decreasing and $h(x)$ is non-increasing, then

$$E(g(X)h(X)) \leq E(g(X))E(h(X))$$

if all expectations exist.

(iii) If $g(x)$ and $h(x)$ are both non-decreasing or if $g(x)$ and $h(x)$ are both non-increasing, then

$$E(g(X)h(X)) \geq E(g(X))E(h(X))$$

if all expectations exist.

**Proof:**

Homework

**Note:**

Theorem 4.8.7 is called Covariance Inequality because

- (ii) implies $\text{Cov}(g(X), h(X)) \leq 0$, and
- (iii) implies $\text{Cov}(g(X), h(X)) \geq 0$. 

121
5 Particular Distributions

5.1 Multivariate Normal Distributions

(Based on Casella/Berger, Exercises 4.45 through 4.50)

Definition 5.1.1:
A rv $X$ has a (univariate) Normal distribution, i.e., $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma > 0$, iff it has the pdf
$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$ $X$ has a standard Normal distribution iff $\mu = 0$ and $\sigma^2 = 1$, i.e., $X \sim N(0, 1)$.

Note:
If $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$ and $Var(X) = \sigma^2$.
If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, $X_1$ and $X_2$ independent, and $c_1, c_2 \in \mathbb{R}$, then
$$Y = c_1X_1 + c_2X_2 \sim N(c_1\mu_1 + c_2\mu_2, c_1^2\sigma_1^2 + c_2^2\sigma_2^2).$$

Definition 5.1.2:
A 2-rv $(X, Y)$ has a bivariate Normal distribution iff there exist constants $a_{11}, a_{12}, a_{21}, a_{22}, \mu_1, \mu_2 \in \mathbb{R}$ and iid $N(0, 1)$ rv’s $Z_1$ and $Z_2$ such that
$$X = \mu_1 + a_{11}Z_1 + a_{12}Z_2, \ Y = \mu_2 + a_{21}Z_1 + a_{22}Z_2.$$ If we define
$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ X = \begin{pmatrix} X \\ Y \end{pmatrix}, \ Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$
then we can write
$$X = AZ + \mu.$$ 

Note:
(i) Recall that for $X \sim N(\mu, \sigma^2)$, $X$ can be defined as $X = \sigma Z + \mu$, where $Z \sim N(0, 1)$.
(ii) $E(X) = \mu_1 + a_{11}E(Z_1) + a_{12}E(Z_2) = \mu_1$ and $E(Y) = \mu_2 + a_{21}E(Z_1) + a_{22}E(Z_2) = \mu_2$.
The marginal distributions are $X \sim N(\mu_1, a_{11}^2 + a_{12}^2)$ and $Y \sim N(\mu_2, a_{21}^2 + a_{22}^2)$. Thus, $X$ and $Y$ have (univariate) Normal marginal densities or degenerate marginal densities (which correspond to Dirac distributions) if $a_{11} = a_{12} = 0$. 

122
(iii) There exists another (equivalent) formulation of the previous definition using the joint pdf (see Rohatgi, page 227).

Example:
Let \( X \sim N(\mu_1, \sigma^2_1) \), \( Y \sim N(\mu_2, \sigma^2_2) \), \( X \) and \( Y \) independent.

What is the distribution of \( (X, Y) \)?

Since \( X \sim N(\mu_1, \sigma^2_1) \), it follows that \( X = \mu_1 + \sigma_1 Z_1 \), where \( Z_1 \sim N(0, 1) \).

Since \( Y \sim N(\mu_2, \sigma^2_2) \), it follows that \( Y = \mu_2 + \sigma_2 Z_2 \), where \( Z_2 \sim N(0, 1) \).

Therefore,
\[
\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}.
\]

Since \( X, Y \) are independent, it follows by Theorem 4.2.7 that \( Z_1 = \frac{X-\mu_1}{\sigma_1} \) and \( Z_2 = \frac{Y-\mu_2}{\sigma_2} \) are independent. Thus, \( Z_1, Z_2 \) are iid \( N(0, 1) \). It follows from Definition 5.1.2 that \( (X, Y) \) has a bivariate Normal distribution.

Theorem 5.1.3:
Define \( g : \mathbb{R}^2 \to \mathbb{R}^2 \) as \( g(x) = CX + d \) with \( C \in \mathbb{R}^{2 \times 2} \) a \( 2 \times 2 \) matrix and \( d \in \mathbb{R}^2 \) a 2-dimensional vector. If \( X \) is a bivariate Normal rv, then \( g(X) \) also is a bivariate Normal rv.

Proof:
\[
g(X) = CX + d = C(AZ + \mu) + d = (CA)Z + (C\mu + d) = \tilde{A}Z + \tilde{\mu} \]
which represents another bivariate Normal distribution.

Note:
\[
\rho \sigma_1 \sigma_2 = \text{Cov}(X, Y) = \text{Cov}(a_{11}Z_1 + a_{12}Z_2, a_{21}Z_1 + a_{22}Z_2) = a_{11}a_{21}\text{Cov}(Z_1, Z_1) + (a_{11}a_{22} + a_{12}a_{21})\text{Cov}(Z_1, Z_2) + a_{12}a_{22}\text{Cov}(Z_2, Z_2) = a_{11}a_{21} + a_{12}a_{22}
\]
since \( Z_1, Z_2 \) are iid \( N(0, 1) \) rv’s.
Definition 5.1.4: 
The variance–covariance matrix of \((X, Y)\) is

\[
\Sigma = AA' = \begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{pmatrix}
\begin{pmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{pmatrix}
= \begin{pmatrix}
a_1^2 + a_2^2 & a_1a_2 + a_2a_1 \\
a_1a_2 + a_2a_1 & a_1^2 + a_2^2
\end{pmatrix}
= \begin{pmatrix}
\sigma_1^2 & \rho\sigma_1\sigma_2 \\
\rho\sigma_1\sigma_2 & \sigma_2^2
\end{pmatrix}.
\]

Note:

\[
det(\Sigma) = |\Sigma| = \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 = \sigma_1^2\sigma_2^2(1 - \rho^2),
\]

\[
\sqrt{|\Sigma|} = \sigma_1\sigma_2\sqrt{1 - \rho^2}
\]

and

\[
\Sigma^{-1} = \frac{1}{|\Sigma|} \begin{pmatrix}
\sigma_2^2 & -\rho\sigma_1\sigma_2 \\
-\rho\sigma_1\sigma_2 & \sigma_1^2
\end{pmatrix} = \begin{pmatrix}
\frac{1}{\sigma_1(1-\rho^2)} & \frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)} \\
\frac{-\rho}{\sigma_1\sigma_2(1-\rho^2)} & \frac{1}{\sigma_2(1-\rho^2)}
\end{pmatrix}
\]

Theorem 5.1.5: 
Assume that \(\sigma_1 > 0, \sigma_2 > 0\) and \(|\rho| < 1\). Then the joint pdf of \(X = (X, Y) = AZ + \mu\) (as defined in Definition 5.1.2) is

\[
f_X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1 - \rho^2}}\exp\left(-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)\right)
\]

Proof:

Since \(\Sigma\) is positive definite and symmetric, \(A\) is invertible.

The mapping \(Z \rightarrow X\) is 1-to-1:

\[
X = AZ + \mu \\
\Rightarrow Z = A^{-1}(X - \mu) \\
|J| = |A^{-1}| = \frac{1}{|A|}
\]

\[
|A| = \sqrt{|A|^2} = \sqrt{|A| \cdot |AT|} = \sqrt{|AA^T|} = \sqrt{|\Sigma|} = \sqrt{\sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2} = \sigma_1\sigma_2\sqrt{1 - \rho^2}
\]

By Definition 5.1.2, it is for \(Z' = (Z_1, Z_2)\):

\[
f_Z(z) = \frac{1}{2\pi}e^{-\frac{z_1^2}{2}} \frac{1}{2\pi}e^{-\frac{z_2^2}{2}} = \frac{1}{2\pi}e^{-\frac{1}{2}(z^Tz)}
\]

124
We can apply Theorem 4.3.5 now:

\[
\begin{align*}
\hat{f}_X(x) &= \frac{1}{2\pi \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} (x - \mu)^T (A^{-1})^T A^{-1} (x - \mu) \right) \\
&= (\ast) \frac{1}{2\pi \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
\end{align*}
\]

This proves the 1st line of the Theorem. Step (\ast) holds since

\[
(A^{-1})^T A^{-1} = (A^T)^{-1} A^{-1} = (AA^T)^{-1} = \Sigma^{-1}.
\]

The second line of the Theorem is based on the transformations stated in the Note following Definition 5.1.4:

\[
\hat{f}_X(x) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp\left(-\frac{1}{2(1 - \rho^2)} \left( \left( \frac{x - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x - \mu_1}{\sigma_1} \right) \left( \frac{y - \mu_2}{\sigma_2} \right) + \left( \frac{y - \mu_2}{\sigma_2} \right)^2 \right) \right)
\]

\[ \blacksquare \]

Note:
In the situation of Theorem 5.1.5, we say that \((X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)\).

Theorem 5.1.6:
If \((X, Y)\) has a non–degenerate \(N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)\) distribution, then the conditional distribution of \(X\) given \(Y = y\) is

\[
N(\mu_1 + \rho \frac{\sigma_1}{\sigma_2} (y - \mu_2), \sigma_1^2 (1 - \rho^2)).
\]

Proof:
Homework \[ \blacksquare \]

Example 5.1.7:
Let rv’s \((X_1, Y_1)\) be \(N(0, 0, 1, 1, 0)\) with pdf \(f_1(x, y)\) and \((X_2, Y_2)\) be \(N(0, 0, 1, 1, \rho)\) with pdf \(f_2(x, y)\). Let \((X, Y)\) be the rv that corresponds to the pdf

\[
f_{X,Y}(x, y) = \frac{1}{2} f_1(x, y) + \frac{1}{2} f_2(x, y).
\]

\((X, Y)\) is a bivariate Normal rv iff \(\rho = 0\). However, the marginal distributions of \(X\) and \(Y\) are always \(N(0, 1)\) distributions. See also Rohatgi, page 229, Remark 2. \[ \blacksquare \]
Theorem 5.1.8:
The mgf $M_X(t)$ of a non–singular bivariate Normal rv $X' = (X, Y)$ is

$$M_X(t) = M_{X,Y}(t_1, t_2) = \exp(\mu'_1 t_1 + \mu'_2 t_2 + \frac{1}{2}(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2\rho\sigma_1\sigma_2 t_1 t_2)).$$

Proof:
The mgf of a univariate Normal rv $X \sim N(\mu, \sigma^2)$ will be used to develop the mgf of a bivariate Normal rv $X' = (X, Y)$:

$$M_X(t) = E(\exp(tX)) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

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

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

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

The pdf of $N(\mu + t\sigma^2, \sigma^2)$, that integrates to 1

$$= \exp\left(-\frac{1}{2\sigma^2}[-\mu^2 - 2\mu t\sigma^2 - t^2\sigma^4 + \mu^2]\right)$$

$$= \exp\left(-\frac{2\mu t\sigma^2 - t^2\sigma^4}{-2\sigma^2}\right)$$

$$= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$$

Bivariate Normal mgf:

$$M_{X,Y}(t_1, t_2) = E(\exp(t_1 X + t_2 Y))$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(t_1 x + t_2 y) f_{X,Y}(x, y) dx dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(t_1 x) \exp(t_2 y) f_X(x) f_{Y|X}(y | x) dy dx$$

$$= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \exp(t_2 y) f_{Y|X}(y | x) dy\right) \exp(t_1 x) f_X(x) dx$$
\( A \) follows from Theorem 5.1.6 since \( Y \mid X \sim N(\beta_X, \sigma^2_X(1 - \rho^2)) \). \( B \) follows when we apply our calculations of the mgf of a \( N(\mu, \sigma^2) \) distribution to a \( N(\beta_X, \sigma^2_X(1 - \rho^2)) \) distribution. \( C \) holds since the integral represents \( M_X(t + \frac{\bar{\sigma}_2}{\bar{\sigma}_1} t_2) \).

**Corollary 5.1.9:**
Let \((X, Y)\) be a bivariate Normal rv. \( X \) and \( Y \) are independent iff \( \rho = 0 \).

**Definition 5.1.10:**
Let \( Z \) be a \( k\)-rv of \( k \) iid \( N(0, 1) \) rv’s. Let \( A \in \mathbb{R}^{k \times k} \) be a \( k \times k \) matrix, and let \( \mu \in \mathbb{R}^k \) be a \( k \)-dimensional vector. Then \( X = AZ + \mu \) has a **multivariate Normal distribution** with mean vector \( \mu \) and variance–covariance matrix \( \Sigma = AA' \).

**Note:**
(i) If \( \Sigma \) is non–singular, \( X \) has the joint pdf
\[
f_X(z) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (z - \mu)' \Sigma^{-1} (z - \mu) \right).
\]
If \( \Sigma \) is singular, the joint pdf does exist but it cannot be written explicitly.
(ii) If \( \Sigma \) is singular, then \( \mathbf{X} - \mu \) takes values in a linear subspace of \( \mathbb{R}^k \) with probability 1.

(iii) If \( \Sigma \) is non–singular, then \( \mathbf{X} \) has mgf

\[
M_{\mathbf{X}}(t) = \exp(\mu' t + \frac{1}{2} t' \Sigma t).
\]

(iv) \( \mathbf{X} \) has characteristic function

\[
\Phi_{\mathbf{X}}(t) = \exp(i \mu' t - \frac{1}{2} t' \Sigma t)
\]

(no matter whether \( \Sigma \) is singular or non–singular).


---

**Theorem 5.1.11:**

The components \( X_1, \ldots, X_k \) of a normally distributed \( k \)-rv \( \mathbf{X} \) are independent iff \( \text{Cov}(X_i, X_j) = 0 \ \forall i, j = 1, \ldots, k, \ i \neq j \).

**Theorem 5.1.12:**

Let \( \mathbf{X}' = (X_1, \ldots, X_k) \). \( \mathbf{X} \) has a \( k \)-dimensional Normal distribution iff every linear function of \( \mathbf{X} \), i.e., \( \mathbf{X}'t = t_1 X_1 + t_2 X_2 + \ldots + t_k X_k \), has a univariate Normal distribution.

**Proof:**

The Note following Definition 5.1.1 states that any linear function of two Normal rv’s has a univariate Normal distribution. By induction on \( k \), we can show that every linear function of \( \mathbf{X} \), i.e., \( \mathbf{X}'t \), has a univariate Normal distribution.

Conversely, if \( \mathbf{X}'t \) has a univariate Normal distribution, we know from Theorem 5.1.8 that

\[
M_{\mathbf{X}'t}(s) = \exp \left( E(\mathbf{X}'t) \cdot s + \frac{1}{2} Var(\mathbf{X}'t) \cdot s^2 \right)
\]

\[
= \exp \left( \mu' ts + \frac{1}{2} t' \Sigma t s^2 \right)
\]

\[
\implies M_{\mathbf{X}'t}(1) = \exp \left( \mu' t + \frac{1}{2} t' \Sigma t \right)
\]

\[
= M_{\mathbf{X}}(t)
\]

By uniqueness of the mgf and Note (iii) that follows Definition 5.1.10, \( \mathbf{X} \) has a multivariate Normal distribution.
5.2 Exponential Family of Distributions

(Based on Casella/Berger, Section 3.4)

Definition 5.2.1:
Let \( \vartheta \) be an interval on the real line. Let \( \{ f(\cdot; \theta) : \theta \in \vartheta \} \) be a family of pdf’s (or pmf’s). We assume that the set \( \{ x : f(x; \theta) > 0 \} \) is independent of \( \theta \), where \( x = (x_1, \ldots, x_n) \).

We say that the family \( \{ f(\cdot; \theta) : \theta \in \vartheta \} \) is a **one–parameter exponential family** if there exist real–valued functions \( Q(\theta) \) and \( D(\theta) \) on \( \vartheta \) and Borel–measurable functions \( T(X) \) and \( S(X) \) on \( \mathbb{R}^n \) such that

\[
    f(x; \theta) = \exp(Q(\theta)T(x) + D(\theta) + S(x)).
\]

**Note:**
We can also write \( f(x; \theta) \) as

\[
    f(x; \eta) = h(x)c(\eta)\exp(\eta T(x))
\]

where \( h(x) = \exp(S(x)) \), \( \eta = Q(\theta) \), and \( c(\eta) = \exp(D(Q^{-1}(\eta))) \), and call this the **exponential family in canonical form** for a natural parameter \( \eta \).

Definition 5.2.2:
Let \( \vartheta \subseteq \mathbb{R}^k \) be a \( k \)--dimensional interval. Let \( \{ f(\cdot; \theta) : \theta \in \vartheta \} \) be a family of pdf’s (or pmf’s). We assume that the set \( \{ x : f(x; \theta) > 0 \} \) is independent of \( \theta \), where \( x = (x_1, \ldots, x_n) \).

We say that the family \( \{ f(\cdot; \theta) : \theta \in \vartheta \} \) is a **\( k \)--parameter exponential family** if there exist real–valued functions \( Q_1(\theta), \ldots, Q_k(\theta) \) and \( D(\theta) \) on \( \vartheta \) and Borel–measurable functions \( T_1(X), \ldots, T_k(X) \) and \( S(X) \) on \( \mathbb{R}^n \) such that

\[
    f(x; \theta) = \exp \left( \sum_{i=1}^{k} Q_i(\theta)T_i(x) + D(\theta) + S(x) \right).
\]

**Note:**
Similar to the Note following Definition 5.2.1, we can express the \( k \)--parameter **exponential family in canonical form** for a natural \( k \times 1 \) parameter vector \( \eta = (\eta_1, \ldots, \eta_k)' \) as

\[
    f(x; \eta) = h(x)c(\eta)\exp \left( \sum_{i=1}^{k} \eta_i T_i(x) \right),
\]
and define the **natural parameter space** as the set of points \( \eta \in W \subseteq \mathbb{R}^n \) for which the integral
\[
\int_{\mathbf{R}^n} \exp \left( \sum_{i=1}^{k} \eta_i T_i(x) \right) h(x) dx
\]
is finite.

**Example 5.2.3:**
Let \( X \sim N(\mu, \sigma^2) \) with both parameters \( \mu \) and \( \sigma^2 \) unknown. We have:
\[
f(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) = \exp \left( -\frac{1}{2\sigma^2} x^2 + \frac{\mu^2}{\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2) \right)
\]
\[
\theta = (\mu, \sigma^2)
\]
\[
\vartheta = \{ (\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0 \}
\]

Therefore,
\[
Q_1(\theta) = -\frac{1}{2\sigma^2}
\]
\[
T_1(x) = x^2
\]
\[
Q_2(\theta) = \frac{\mu}{\sigma^2}
\]
\[
T_2(x) = x
\]
\[
D(\theta) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2} \ln(2\pi\sigma^2)
\]
\[
S(x) = 0
\]
Thus, this is a 2–parameter exponential family.

**Canonical form:**
\[
f(x; \theta) = \exp(Q_1(\theta)T_1(x) + Q_2(\theta)T_2(x) + D(\theta) + S(x))
\]
\[
\implies h(x) = \exp(S(x)) = \exp(0) = 1
\]
\[
\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} Q_1(\theta) \\ Q_2(\theta) \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix}
\]
Therefore, when we solve \( \eta_1 = -\frac{1}{2\sigma^2}, \eta_2 = \frac{\mu}{\sigma^2} \) for \( \sigma^2, \mu, \) we get

\[
\sigma^2 = -\frac{1}{2\eta_1}, \quad \mu = \sigma^2 \eta_2 = -\frac{1}{2\eta_1} \eta_2 = -\frac{\eta_2}{2\eta_1}.
\]

Thus,

\[
C(\eta) = \exp(D(Q^{-1}(\eta))) = \exp\left(\frac{-\eta_2}{2\eta_1} - \frac{1}{2} \ln \left(\frac{2\pi}{-\frac{1}{2\eta_1}}\right)\right)
\]

\[
= \exp\left(\frac{\eta_2^2}{4\eta_1} - \frac{1}{2} \ln \left(\frac{\pi}{\eta_1}\right)\right)
\]

Therefore, \( f(x; \theta) \) can be reparametrized in canonical form as

\[
f(x; \eta) = 1 \cdot \exp\left(\frac{\eta_2^2}{4\eta_1} - \frac{1}{2} \ln \left(-\frac{\pi}{\eta_1}\right)\right) \exp\left(\eta_1 x^2 + \eta_2 x\right).
\]

The natural parameter space is

\[
\{(\eta_1, \eta_2) \mid \eta_1 < 0, \eta_2 \in \mathbb{R}\}
\]

because

\[
\int_{-\infty}^{\infty} \exp\left(\eta_1 x^2 + \eta_2 x\right) \cdot dx < \infty
\]

for \( \eta_1 < 0 \) (independent from \( \eta_2 \)), but

\[
\int_{-\infty}^{\infty} \exp\left(\eta_1 x^2 + \eta_2 x\right) \cdot dx \text{ undefined}
\]

for \( \eta_1 \geq 0 \) (independent from \( \eta_2 \)).

\[\text{To Be Continued ...}\]
Index

σ–Algebra, 2
σ–Field, 2
σ–Field, Borel, 4
σ–Field, Generated, 3
k–Parameter Exponential Family, 129
kth Absolute Moment, 47
kth Central Moment, 47
kth Factorial Moment, 72
kth Moment, 47
n–Dimensional Characteristic Function, 106
n–RV, 78
Absolutely Continuous, 34
Additivity, Countable, 4
Arithmetic Mean, 119
Bayes’ Rule, 20
Binomial Coefficient, 13
Binomial Theorem, 15
Bivariate Normal Distribution, 122
Bivariate RV, 80
Bochner’s Theorem, 65
Bonferroni’s Inequality, 7
Boole’s Inequality, 9
Borel σ–Field, 4
Borel Field, 2
Borel Sets, 4
Canonical Form, 129
Cauchy–Schwarz Inequality, 104
CDF, 29
CDF, Conditional, 82
CDF, Joint, 78
CDF, Marginal, 81
Characteristic Function, 63
Characteristic Function, n–Dimensional, 106
Chebychev’s Inequality, 75
Cofinite, 2
Completely Independent, 21, 85
Complex Numbers, 61
Complex Numbers, Conjugate, 62
Complex–Valued Random Variable, 62
Concave, 118
Conditional CDF, 82
Conditional Cumulative Distribution Function, 82
Conditional Expectation, 112
Conditional PDF, 82
Conditional PMF, 81
Conditional Probability, 18
Conditional Probability Density Function, 82
Conditional Probability Mass Function, 81
Conjugate Complex Numbers, 62
Continuity of Sets, 9
Continuous, 31, 33, 34
Continuous RV, 80
Continuous, Absolutely, 34
Convex, 118
Countable Additivity, 4
Counting, Fundamental Theorem of, 12
Covariance Inequality, 121
Cumulative Distribution Function, 29
Cumulative Distribution Function, Joint, 78
Decreasing, 40
Discrete, 31, 33
Discrete RV, 80
Equally Likely, 12
Equivalence, 38
Equivalent RV, 89
Euler’s Relation, 61
Event, 1
Expectation, Conditional, 112
Expectation, Multivariate, 100
Expected Value, 46
Exponential Family in Canonical Form, 129
Exponential Family, k–Parameter, 129
Exponential Family, One–Parameter, 129
Factorial, 13
Factorization Theorem, 86
Field, 1
Fundamental Theorem of Counting, 12
Geometric Mean, 119
Hölder’s Inequality, 116
Harmonic Mean, 119
Identically Distributed, 32
IID, 89
Implication, 37
Inclusion–Exclusion, Principle, 7
Increasing, 40
Independent, 20, 85, 86
Independent Identically Distributed, 89
Independent, Completely, 21, 85
Independent, Mutually, 21, 85
Independent, Pairwise, 20
Indicator Function, 36
Induced Probability, 29
Induction, 7
Induction Base, 7
Induction Step, 7
Inequality, Bonferroni’s, 7
Inequality, Boole’s, 9
Inequality, Cauchy–Schwarz, 104
Inequality, Chebychev’s, 75
Inequality, Covariance, 121
Inequality, Hölder’s, 116
Inequality, Jensen’s, 118
Inequality, Lyapunov’s, 76
Inequality, Markov’s, 75
Inequality, Minkowski’s, 117

Jensen’s Inequality, 118
Joint CDF, 78
Joint Cumulative Distribution Function, 78
Joint PDF, 80
Joint PMF, 80
Joint Probability Density Function, 80
Jump Points, 33

Kolmogorov Axioms of Probability, 4
L’Hospital’s Rule, 56
Law of Total Probability, 19
Lebesque’s Dominated Convergence Theorem, 57
Leibnitz’s Rule, 56
Levy’s Theorem, 66
Logic, 37
Lyapunov’s Inequality, 76

Marginal CDF, 81
Marginal Cumulative Distribution Function, 81
Marginal PDF, 81
Marginal PMF, 80
Marginal Probability Density Function, 81
Marginal Probability Mass Function, 80
Markov’s Inequality, 75
MAX, 90
Mean, 46
Mean, Arithmetic, 119
Mean, Geometric, 119
Mean, Harmonic, 119
Measurable, 2, 25
Measurable $L - \mathcal{B}$, 25
Measurable Functions, 26
Measurable Space, 4
MGF, 54
MIN, 90
Minkowski’s Inequality, 117
MMGF, 106
Moivre’s Theorem, 61
Moment Generating Function, 54
Moment Generating Function, Multivariate, 106
Moment, $k^{th}$, 47
Moment, $k^{th}$ Absolute, 47
Moment, $k^{th}$ Central, 47

Moment, $k^{th}$ Factorial, 72
Moment, Multi–Way, 103
Moment, Multi–Way Central, 103
Monotonic, 40
Multi–Way Central Moment, 103
Multi–Way Moment, 103
Multiplication Rule, 19
Multivariate Expectation, 100
Multivariate Moment Generating Function, 106
Multivariate Normal Distribution, 127
Multivariate RV, 80
Multivariate Transformation, 93
Mutually Independent, 21, 85

Natural Parameter Space, 130
Non–Decreasing, 40
Non–Decreasing, Strictly, 40
Non–Increasing, 40
Non–Increasing, Strictly, 40
Normal Distribution, 122
Normal Distribution, Bivariate, 122
Normal Distribution, Multivariate, 127
Normal Distribution, Standard, 122
Normal Distribution, Univariate, 122

One–Parameter Exponential Family, 129
Order Statistic, 97

Pairwise Independent, 20
Partition, 19
PDF, 34
PDF, Conditional, 82
PDF, Joint, 80
PDF, Marginal, 81
PMF, 72
PM, 4
PMF, 33
PMF, Conditional, 81
PMF, Joint, 80
PMF, Marginal, 80
Power Set, 2
Principle of Inclusion–Exclusion, 7
Probability Density Function, 34
Probability Generating Function, 72
Probability Mass Function, 33
Probability Measure, 4
Probability Space, 4
Probability, Conditional, 18
Probability, Induced, 29
Probability, Total, 19
Quantifier, 38
Random Variable, 25
Random Variable, Complex–Valued, 62