STAT 6710
Mathematical Statistics I
Fall Semester 1999

Dr. Jürgen Symanzik
Utah State University
Department of Mathematics and Statistics
3900 Old Main Hill
Logan, UT 84322–3900

Tel.: (435) 797–0696
FAX: (435) 797–1822

e-mail: symanzik@sunfs.math.usu.edu
Contents

Acknowledgements 1

1 Axioms of Probability 1
  1.1 σ–Fields .................................................... 1
  1.2 Manipulating Probability .................................. 5
  1.3 Combinatorics and Counting ................................ 11
  1.4 Conditional Probability and Independence ............... 16

2 Random Variables 23
  2.1 Measurable Functions ....................................... 23
  2.2 Probability Distribution of a Random Variable ............ 27
  2.3 Discrete and Continuous Random Variables ................. 31
  2.4 Transformations of Random Variables ....................... 36

3 Moments and Generating Functions 42
  3.1 Expectation ................................................. 42
  3.2 Generating Functions ....................................... 51
  3.3 Complex–Valued Random Variables and Characteristic Functions .... 57
  3.4 Probability Generating Functions .......................... 68
  3.5 Moment Inequalities ....................................... 69

4 Random Vectors 72
  4.1 Joint, Marginal, and Conditional Distributions ............. 72
  4.2 Independent Random Variables ............................. 77
  4.3 Functions of Random Vectors ............................... 82
  4.4 Order Statistics ............................................ 91
  4.5 Multivariate Expectation .................................. 93
  4.6 Multivariate Generating Functions ......................... 98
  4.7 Conditional Expectation ................................... 104
  4.8 Inequalities and Identities ................................. 106
5 Particular Distributions

5.1 Multivariate Normal Distributions ........................................ 113
5.2 Exponential Family of Distributions .................................... 120

6 Limit Theorems

6.1 Modes of Convergence ..................................................... 123
6.2 Weak Laws of Large Numbers .......................................... 136
6.3 Strong Laws of Large Numbers ......................................... 140
Acknowledgements

I would like to thank my students, Hanadi B. Eltahir, Rich Madsen, and Bill Morphet, who helped in typesetting these lecture notes using \LaTeX and for their suggestions how to improve some of the material presented in class.

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 Rohatgi (1976) and other sources listed below, form the basis of the script presented here.

The textbook required for this class is:


A Web page dedicated to this class is accessible at:


This course closely follows Rohatgi (1976) 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, January 18, 2000
1 Axioms of Probability

1.1 σ–Fields

Let \( \Omega \) be the sample space of all possible outcomes of a chance experiment. Let \( \omega \in \Omega \) (or \( x \in \Omega \)) be any outcome.

Example:
Count \# of heads in \( n \) coin tosses. \( \Omega = \{0, 1, 2, \ldots, n\} \).

Any subset \( A \) of \( \Omega \) is called an event.

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

Instead, we consider classes of subsets of \( \Omega \) called fields and σ–fields.

Definition 1.1.1:
A class \( L \) of subsets of \( \Omega \) is called a field if \( \Omega \in L \) and \( L \) is closed under complements and finite unions, i.e., \( L \) satisfies

(i) \( \Omega \in L \)

(ii) \( A \in L \implies A^C \in L \)

(iii) \( A, B \in L \implies A \cup B \in L \)

Since \( \Omega^C = \emptyset \), (i) and (ii) imply \( \emptyset \in L \). Therefore, (i): \( \emptyset \in L \) [can replace (i)].

Recall De Morgan’s Laws:
\[
\bigcup_{A \in \mathcal{A}} A = \left( \bigcap_{A \in \mathcal{A}} A^C \right)^C \quad \text{and} \quad \bigcap_{A \in \mathcal{A}} A = \left( \bigcup_{A \in \mathcal{A}} A^C \right)^C.
\]
Note:
So (ii), (iii) imply (iii): \( A, B \in L \implies A \cap B \in L \) [can replace (iii)].

Proof:

\[
A, B \in L \iff A^C, B^C \in L \iff (A^C \cup B^C) \in L \iff (A^C \cup B^C)^C \in L \implies \text{DM} \implies A \cap B \in L
\]

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

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). Then \( L \) is a field. But \( L \) is a \( \sigma \)-field iff (if and only if) \( \Omega \) is finite.

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

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

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 \( L \)”.

2
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) $\Omega \in L_\theta \quad \forall \theta \implies \Omega \in L$
(ii) Let $A \in L \implies A \in L_\theta \quad \forall \theta \implies A^C \in L_\theta \quad \forall \theta \implies A^C \in L$
(iv) Let $A_n \in L \quad \forall n \implies A_n \in L_\theta \quad \forall \theta \implies \bigcup_n A_n \in L_\theta \quad \forall \theta \implies \bigcup_n A_n \in L$

**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)
Since all unions are included, we have \( \sigma(b) = \{ \Omega, \emptyset, \{1\}, \{0,1\}, \{2,3\}, \{0,2,3\}, \{1,2,3\} \} \)

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 \( B \) or \( B_1 \). The sets of \( B \) are called **Borel sets**.

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

**Note:**

\( B \) 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: \((x, \infty) = \bigcup_{n=1}^{\infty} (x, x + n] \)

We now have a **measurable space** \((\Omega, L)\). We next define a **probability measure** \( P(\cdot) \) on \((\Omega, L)\) to obtain a **probability space** \((\Omega, L, P)\).

**Definition 1.1.7: Kolmogorov Axioms of Probability**

A **probability measure** \((p.m.)\), \( P \), on \((\Omega, L)\) is a set function \( P : L \to \mathbb{R} \) satisfying

(i) \( 0 \leq P(A) \forall A \in L \)

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

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

**Note:**

\( \bigcup_{n=1}^{\infty} A_n \in L \) holds automatically if \( L \) is a \( \sigma \)-field but it is needed as a precondition in the case that \( L \) is just a field. Property (iii) is called **countable additivity**.
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) $A_n \equiv \emptyset$ \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 $n$.

$P(\emptyset) = P\left(\bigcup_{n=1}^{\infty} A_n\right)$ \text{Def 1.1.7}(iii) $= \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} P(\emptyset)$

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

(ii) $A_1 \equiv A, A_2 \equiv A^C, A_n \equiv \emptyset$ \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.

\[
1 = P(\Omega) = P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \tag{Def 1.1.7}(iii)
\]
\[ P(A) = P(A_1) + P(A_2) + \sum_{n=3}^{\infty} P(A_n) \]

**Thm 1.2.1 (i)**

\[ P(A_1) + P(A_2) = P(A) + P(A^C) \]

\[ \Rightarrow P(A^C) = 1 - P(A) \quad \forall A \in L. \]

(iii) By Th. 1.2.1 (ii) \( P(A) = 1 - P(A^C) \Rightarrow P(A) \leq 1 \quad \forall A \in L \) since \( P(A^C) \geq 0 \) 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 \( (A \cap B^C), (A \cap B), (B \cap A^C), \)

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

By Def 1.1.7 (iii)

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

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

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

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

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

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

By Def 1.1.7 (iii)

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

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

\[ \Rightarrow P(A) \leq P(B) \text{ since } P(B \cap A^C) \geq 0 \text{ by Def 1.1.7 (i)} \]

**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)

**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 induction base:
For $n = 1$, Th. 1.2.3 right side evaluates to $P(A_1) \leq P(A_1)$, which is true.

For $n = 2$, Th. 1.2.3 right side evaluates to $P(A_1 \cup A_2) \leq P(A_1) + P(A_2)$.
$P(A_1 \cup A_2)^{Th.1.2.1(iv)} \Rightarrow P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2)$ since $P(A_1 \cap A_2) \geq 0$ by Def. 1.1.7 (i).

This establishes the induction base for the right side of Th. 1.2.3.

Right side induction step assumes Th. 1.2.3 right side is true for $n$ and shows that it is true for $n + 1$:

$$\begin{align*}
P\left( \bigcup_{i=1}^{n+1} A_i \right) &= P(\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) \\
&\stackrel{Th.1.2.1(iv)}{=} P\left( \bigcup_{i=1}^{n} A_i \right) + P(A_{n+1}) - P\left( \bigcup_{i=1}^{n} A_i \right) \cap A_{n+1}) \\
&\leq P\left( \bigcup_{i=1}^{n} A_i \right) + P(A_{n+1}) \\
&\leq \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) \\
&= \sum_{i=1}^{n+1} P(A_i)
\end{align*}$$

Left side induction base:
For $n = 1$, Th. 1.2.3 left side evaluates to $P(A_1) \leq P(A_1)$, which is true.

For $n = 2$, Th. 1.2.3 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 $n = 3$, Th. 1.2.3 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

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

\[
= P((A_1 \cup A_2) \cup A_3) \\
\overset{Th.1.2.1(iv)}{=} P(A_1 \cup A_2) + P(A_3) - P((A_1 \cup A_2) \cap A_3) \\
= P(A_1 \cup A_2) + P(A_3) - P((A_1 \cap A_3) \cup (A_2 \cap A_3)) \\
\overset{Th.1.2.1(iv)}{=} P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\
+ P((A_1 \cap A_3) \cap (A_2 \cap A_3)) \\
\geq P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\
= P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3)
\]

This establishes the induction base for the left side of Th. 1.2.3.

Left side induction step assumes Th. 1.2.3 left side is true for \( n \) and shows that it is true for \( n + 1 \):

\[
P(\bigcup_{i=1}^{n+1} A_i) = P(\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) \\
= P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) - P((\bigcup_{i=1}^{n} A_i) \cap A_{n+1}) \\
\overset{left \ I.B.}{\geq} \sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) + P(A_{n+1}) - P((\bigcup_{i=1}^{n} A_i) \cap A_{n+1}) \\
= \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) - P(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})) \\
\overset{Th.1.2.3 \ right \ side}{\geq} \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) - \sum_{i=1}^{n} P(A_i \cap A_{n+1}) \\
= \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n+1} P(A_i \cap A_j)
\]
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 $A = \bigcap_{n=1}^{\infty} A_n$.

Theorem 1.2.6:
If $\{A_n\}_{n=1}^{\infty}, A_n \in L$ and $A \in L$, then $\lim_{n \to \infty} P(A_n) = P(A)$ 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 $B_k = A_k - A_{k-1} = A_k \cap A_{k-1}^c \quad \forall k \geq 2$

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 $A_n = \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} B_i$
\[ P(A) = \sum_{k=1}^{\infty} P(B_k) \quad \text{By def. 1.1.7 (iii)} \]
\[ = \sum_{k=1}^{\infty} P(B_k) = \lim_{n \to \infty} \left[ \sum_{k=1}^{n} P(B_k) \right] \]
\[ \text{again by def. 1.1.7 (iii)} \]
\[ \lim_{n \to \infty} \left[ P(\bigcup_{k=1}^{n} B_k) \right] = \lim_{n \to \infty} \left[ P(\bigcup_{k=1}^{n} A_k) \right] = \lim_{n \to \infty} P(A_n) \]

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 = \bigcap_{n=1}^{\infty} A_n \quad \text{De Morgan} \] \( A_n^C \]

\[ P(A^C) = \lim_{n \to \infty} P(A_n^C) \]

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

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

**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 \]

\[ 0 \preceq P(\bigcup_{n=1}^{\infty} A_n) \quad \text{By Bonferroni’s inequality} \]
\[ \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} 0 = 0 \]

Therefore \( P(\bigcup_{n=1}^{\infty} A_n) = 0 \)

**Part (ii):**

Let \( \{ A_n \}_{n=1}^{\infty} \in L, \ P(A_n) = 1 \ \forall n \]

\[ \implies P(A_n^C) = 0 \ \forall n \quad \implies P(\bigcup_{n=1}^{\infty} A_n^C) = 0 \quad \text{De Morgan} \]
\[ \implies P(\bigcap_{n=1}^{\infty} A_n) = 1 \]
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

$$P(\omega_j) = \frac{1}{n} \quad \forall j = 1, \ldots, n.$$  

**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

$$n_1 \times n_2 \times n_3 \times \ldots \times n_k$$

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 $= n_2 + n_2 + \ldots + n_2 = n_1 \times n_2$.

**Induction Step:**
Suppose it is true for $(k - 1)$. We show that 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 $= (n_1 \times n_2 \times n_3 \times \ldots \times n_{k-1}) \times n_k$.  

11
Definition 1.3.3:
For positive integer \( n \), we define \textbf{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 \textbf{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:
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 \textbf{ordered} (i.e., \{1, 2\} different from \{2, 1\}); and
(iv) whether the draws are \textbf{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 \( \frac{(n+r-1)!}{r!(n-1)!} = \binom{n+r-1}{r} \).

Proof:

(i) \( n \) choices to select 1\textsuperscript{st}

\[ n - 1 \text{ choices to select 2\textsuperscript{nd}} \]

\[ : \]

\[ n - r + 1 \text{ choices to select } r\textsuperscript{th} \]

By Theorem 1.3.2, there are \( n \times (n-1) \times \ldots \times (n-r+1) = \frac{n \times (n-1) \times \ldots \times (n-r+1)!}{(n-r)!} = \frac{n!}{(n-r)!} \)

ways to do so.

12
**Corollary:**
The number of permutations of \( n \) objects is \( n! \).

(ii) \( n \) choices to select 1\(^{st} \)

\( n \) choices to select 2\(^{nd} \)

\vdots

\( n \) choices to select \( r \)\(^{th} \)

By Theorem 1.3.2, there are \( \underbrace{n \times n \times \ldots \times n} \) ways to do so.

\( r \) times

(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 \((n-1)\) markers 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.

**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:**

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

\( n = 1: (1 + x)^1 = \sum_{r=0}^{1} \binom{1}{r} x^r = \binom{1}{0} x^0 + \binom{1}{1} x^1 = 1 + x \)
Induction Step:
Suppose it is true for $k$. We show that it is true for $k + 1$.

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

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

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

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

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

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

(*) 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 a non-negative integer $n$, it holds:

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

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

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

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

Proof:
Use the Binomial Theorem:

(i) Let $x = 1$. Then

\[
2^n = (1 + 1)^n = \sum_{r=0}^{n} \binom{n}{r} 1^r = \sum_{r=0}^{n} \binom{n}{r}
\]
(ii) Let \( x = -1 \). Then

\[
0 = (1 + (-1))^n \overset{\text{Bin. Th.}}{=} \sum_{r=0}^{n} \binom{n}{r} (-1)^r
\]

(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

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

(iv) Substitute \( x = -1 \) in (iii) above, then

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

since for \( \sum a_i = 0 \) also \( \sum (-a_i) = 0 \). 

\[\blacksquare\]

Note:
A useful extension for the binomial coefficient for \( n < r \) is

\[
\binom{n}{r} = \frac{n \cdot (n-1) \cdot \ldots \cdot 0 \cdot \ldots \cdot (n-r+1)}{1 \cdot 2 \cdot \ldots \cdot r} = 0.
\]

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 

\[\blacksquare\]
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:**
This 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) \( P(A \cap H) \geq 0 \) \( \Rightarrow \) \( P_H(A) = \frac{P(A \cap H)}{P(H)} \geq 0 \) \( \forall A \in L \)

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

(iii) Let \( \{A_n\}_{n=1}^{\infty} \) be a sequence of disjoint sets. Then,

\[
P_H(\bigcup_{n=1}^{\infty} A_n) \overset{\text{Def. 1.4.1}}{=} \frac{P(\bigcup_{n=1}^{\infty} A_n \cap H)}{P(H)} = \frac{P(\bigcup_{n=1}^{\infty} (A_n \cap H))}{P(H)}
\]
\[
\begin{align*}
\text{Def.1.1.7(iii)} & \quad \sum_{n=1}^{\infty} \frac{P(A_n \cap H)}{P(H)} \\
= & \quad \sum_{n=1}^{\infty} \frac{P(A_n \cap H)}{P(H)} \\
\text{Def.1.4.1} & \quad \sum_{n=1}^{\infty} P_H(A_n)
\end{align*}
\]

Note:
What we have done is to move to a new sample space \( \mathcal{H} \) 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

**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 \( \implies \) \( A \cap H_j \) are disjoint

\[
A = A \cap \Omega \overset{\text{Def 1.4.4}}{=} A \cap (\bigcup_{j=1}^{\infty} H_j) = \bigcup_{j=1}^{\infty} (A \cap H_j)
\]

\( \implies P(A) = P(\bigcup_{j=1}^{\infty} (A \cap H_j)) \overset{\text{Def 1.1.7(iii)}}{=} \sum_{j=1}^{\infty} P(A \cap H_j) \]

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 \cap A)}{P(A)} \overset{\text{Th 1.4.5}}{=} \frac{P(H_j) \cdot P(A|H_j)}{\sum_{n=1}^{\infty} P(H_n)P(A|H_n)}.
\]

Definition 1.4.7:

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

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 $B^C$.

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

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 $\{A_1, \ldots, A_n\} \in L$, 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, 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

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

It is

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

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

Let $A \equiv \text{No student has a b-day matching the } 1\text{st student}$
Let $B \equiv \text{At least one of the other } (r-2)\text{ has b-day matching } 2\text{nd}$

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

Working backwards from the book

\[
\begin{align*}
p_2 &= \left( \frac{w_5 P_{2-1}}{(365)^{2-1}} \right) \left( 1 - \frac{2 - 1}{365} \right)^{r-2+1} \left[ 1 - \left( \frac{365 - 2}{365 - 2 + 1} \right)^{r-2} \right] \\
&= \frac{365}{365} \left( 1 - \frac{1}{365} \right)^{r-1} \left[ 1 - \left( \frac{363}{364} \right)^{r-2} \right] \\
&= \left( \frac{365 - 1}{365} \right)^{r-1} \left[ 1 - \left( \frac{363}{364} \right)^{r-2} \right]
\end{align*}
\]

(Same as what we found before)

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

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

Now:

\[
\begin{align*}
p_3 &= P(\text{ } A \cap B \cap C) \\
&= P(A) \cdot P(B|A) \cdot P(C|A \cap B) \\
&= \left( \frac{364}{365} \right)^{r-1} \left( \frac{363}{364} \right)^{r-2} \cdot [1 - P(\text{all } (r - 3) \text{ students have a b-day on the remaining 362 out of 364 days})] \\
&= \left( \frac{364}{365} \right)^{r-1} \left( \frac{363}{364} \right)^{r-2} \cdot \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \\
&= \left( \frac{364}{365} \right)^{r-1} \left( \frac{363}{364} \right)^{r-2} \cdot \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \\
&= \left( \frac{364^{r-1}}{364^r} \right) \left( \frac{363^{r-2}}{365^{r-1}} \right) \cdot \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right]
\end{align*}
\]
\[ = \left( \frac{364}{365} \right) \left( \frac{363^{r-2}}{365^{r-2}} \right) \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \]

Working backwards from the book:

\[ p_3 = \left( \frac{w_3 P_{3-1}}{(365)^3-1} \right) \left( 1 - \frac{3 - 1}{365} \right)^{r-3+1} \left[ 1 - \left( \frac{365 - 3}{365 - 3 + 1} \right)^{r-3} \right] \]
\[ = \left( \frac{w_3 P_2}{(365)^2} \right) \left( 1 - \frac{2}{365} \right)^{r-2} \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \]
\[ = \left( \frac{(365)(364)}{(365)^2} \right) \left( \frac{363}{365} \right)^{r-2} \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \]
\[ = \left( \frac{364}{365} \right) \left( \frac{363}{365} \right)^{r-2} \left[ 1 - \left( \frac{362}{363} \right)^{r-3} \right] \]

(Same as what we found before)

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

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 $L - B$ 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).

$\Omega = \{\text{All } 2^{50}\text{ possible sequences of } y/\text{n}\} \quad \text{— HUGE!}$

$L = \mathcal{P}(\Omega)$

$X : \Omega \to S = \{\text{All } 2^{50}\text{ possible sequences of } 1 (=y) \text{ and } 0 (=n)\}$

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

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

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, which takes only finite many values $x_1, \ldots, x_k$ is measurable iff $X^{-1}(x_i) \in L \ \forall x_i$.

Here, $X^{-1}(k) = \{\omega \in \Omega : \# \text{ 1’s in sequence } \omega = k\}$ is a subset of $\Omega$, so it is in $L = \mathcal{P}(\Omega)$. 
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 $X_n : \Omega \rightarrow \mathbb{R}$ be $X_n(\omega) = \text{"proportion of H's in 1st } n \text{ tosses"}$.
For each $n$, $X_n(\cdot)$ is simple (values $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, n\}$) 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 $L - B$. Now we have a sequence of rv’s $\{X_n\}_{n=1}^{\infty}$. We will show later that $P(\{\omega : X_n(\omega) \rightarrow \frac{1}{2}\}) = 1$, i.e., the Strong Law of Large Numbers (SLLN).
Some Technical Points about Measurable Functions

2.1.5: Suppose $(\Omega, L)$ and $(\mathcal{S}, \mathcal{B})$ are measure spaces and that a collection of sets $\mathcal{A}$ generates $\mathcal{B}$, i.e., $\sigma(\mathcal{A}) = \mathcal{B}$. Let $X : \Omega \to \mathcal{S}$. 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 $f : \mathbb{R}^k \to \mathbb{R}^k$ is a continuous function, then $f$ is measurable $\mathcal{B}^k - \mathcal{B}^k$.

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 $[-\infty, \infty] = \mathbb{R} \cup \{-\infty, \infty\}$.

We say $f : \Omega \to \mathbb{R}$ is measurable $L - \mathcal{B}$ if

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

2) $f^{-1}(-\infty), f^{-1}(\infty) \in L$ also.
2.1.10:
Suppose $f_1, f_2, \ldots$ is a sequence of real-valued measurable functions $(\Omega, L) \rightarrow (\mathbb{R}, \mathcal{B})$. Then it holds:

(i) $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, $\liminf_n f_n$, are measurable.

(ii) If $f = \lim_n f_n$ 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 $\{\omega : f_n(\omega) \rightarrow f(\omega)\} \in L$. 

26
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, $X^{-1}(B) \in L \quad \forall B \in B$.

**Proof:**
If $X$ induces a probability measure $Q$ on $(\mathbb{R}, B)$, then $Q$ must satisfy the Kolmogorov Axioms of probability.

Let $X : (\Omega, L) \to (S, B)$. $X$ is a rv $\Rightarrow X^{-1}(B) = \{ \omega : X(\omega) \in B \} = A \in L \quad \forall B \in B$.

1. $Q(B) = P(X^{-1}(B)) = P(\{ \omega : X(\omega) \in B \}) = P(A) \quad \forall B \in B$

2. $Q(\mathbb{R}) = P(X^{-1}(\mathbb{R})) = P(\Omega) = 1$

3. Let $\{B_n\}_{n=1}^{\infty} \in B, B_i \cap B_j = \emptyset \quad \forall i \neq j$. Then,

$$Q(\bigcup_{n=1}^{\infty} B_n) = P(X^{-1}(\bigcup_{n=1}^{\infty} B_n)) \overset{(*)}{=} P(\bigcup_{n=1}^{\infty} (X^{-1}(B_n))) \overset{\text{Def. 1.1.7(iii)}}{=} \sum_{n=1}^{\infty} P(X^{-1}(B_n)) = \sum_{n=1}^{\infty} Q(B_n)$$

$(*)$ holds since $X^{-1}(\cdot)$ 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:**

(i) Let $x_1 < x_2$

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

$$\Rightarrow F(x_1) = P(\{\omega : X(\omega) \leq x_1\}) \leq P(\{\omega : X(\omega) \leq x_2\}) = F(x_2)$$

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 $A_n = \{\omega : X(\omega) \in (x, x_n]\} \in \mathcal{L}$ 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

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

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

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

\[ \implies F(x) \] is right-continuous.

(iii) \( F(-n) \) = \( P(\{\omega : X(\omega) \leq -n\}) \)

\[ \implies F(-\infty) = \lim_{n \to \infty} F(-n) \]

\[ = \lim_{n \to \infty} P(\{\omega : X(\omega) \leq -n\}) \]

\[ = P(\lim_{n \to \infty} \{\omega : X(\omega) \leq -n\}) \]

\[ = P(\emptyset) \]

\[ = 0 \]

(iv) \( F(n) \) = \( P(\{\omega : X(\omega) \leq n\}) \)

\[ \implies F(\infty) = \lim_{n \to \infty} F(n) \]

\[ = \lim_{n \to \infty} P(\{\omega : X(\omega) \leq n\}) \]

\[ = P(\lim_{n \to \infty} \{\omega : X(\omega) \leq n\}) \]

\[ = P(\Omega) \]

\[ = 1 \]

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) \( X \) is **continuous** if \( F(x) \) is continuous in \( x \).

(ii) \( X \) 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 \( P_X(X \in A) = P_Y(Y \in A) \ \forall A \in \mathbb{L} \).

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 \( \text{Bin}(3, 0.5) \), 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):

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

by Def. 2.2.5

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

(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 \((\Omega, I, P)\). \( X \) 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 \( X : \Omega \to \mathbb{R} \) 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 \( x \in R \).
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 $B \in B$, $P(B) = \int_{B} f(t)dt$.

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

32
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

$$F(x) = a F_d(x) + (1 - a) F_c(x), \quad 0 \leq a \leq 1,$$

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 Wahrscheinlichkeitsrechung 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

$$F(x) = a_1 F_d(x) + a_2 F_c(x) + a_3 F_s(x), \quad a_1, a_2, a_3 \geq 0, a_1 + a_2 + a_3 = 1.$$

Here, $F_d(x)$ and $F_c(x)$ are discrete and continuous cdfs (as above). $F_s(x)$ is called a singular cdf. Singular means that $F_s(x)$ is continuous and its derivative $F'_s(x)$ 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  \( a F_d(x) + (1 - a) F_c(x) \), \( 0 \leq a \leq 1 \). How?

Since \( F(x) \) has only one jump at \( x = 0 \), it is reasonable to get started with a pmf \( p_0 = 1 \) and corresponding cdf
\[
F_d(x) = \begin{cases} 
0, & x < 0 \\
1, & x \geq 0 
\end{cases}
\]
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 \textbf{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 \( A \).
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 \( A \Rightarrow A_1 \Rightarrow A_2 \Rightarrow \cdots \Rightarrow A_{n-1} \Rightarrow A_n \Rightarrow B \) where one step directly follows from the previous step. However, there are different approaches to obtain the same result.

\( A \Rightarrow B \) is equivalent to \( \neg B \Rightarrow \neg A \) is equivalent to \( \neg A \lor B \):

\[
\begin{array}{c|c|c|c|c|c}
A & B & A \Rightarrow B & \neg A & \neg B & \neg B \Rightarrow \neg A & \neg A \lor B \\
\hline
1 & 1 & 1 & 0 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 & 1 \\
\end{array}
\]

\( A \Leftrightarrow B \) is equivalent to \((A \Rightarrow B) \land (B \Rightarrow A)\) is equivalent to \((\neg A \lor B) \land (A \lor \neg B)\):

\[
\begin{array}{c|c|c|c|c|c|c|c|c}
A & B & A \Leftrightarrow B & A \Rightarrow B & B \Rightarrow A & (A \Rightarrow B) \land (B \Rightarrow A) & \neg A \lor B & A \lor \neg B & (\neg A \lor B) \land (A \lor \neg B) \\
\hline
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\end{array}
\]

Negations of Quantifiers:

\[ \neg \forall x \in X : B(x) \text{ is equivalent to } \exists x \in X : \neg B(x) \]

\[ \neg \exists x \in X : B(x) \text{ is equivalent to } \forall x \in X : \neg B(x) \]

\[ \exists x \in X \forall y \in Y : B(x, y) \text{ implies } \forall y \in Y \exists x \in X : B(x, y) \]
2.4 Transformations of Random Variables

Let $X$ be a real-valued random variable on $(\Omega, \mathcal{L}, P)$, i.e., $X : (\Omega, \mathcal{L}) \rightarrow (\mathbb{R}, \mathcal{B})$. 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:**

\[
F_Y(y) = P(Y \leq y) = P(\omega : g(X(\omega)) \leq y) = P(\omega : X(\omega) \in B_y)\] where $B_y = g^{-1}(-\infty, y] \in \mathcal{B}$ since $g$ 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, $P_Y(Y \in C) = P_X(X \in g^{-1}(C)) \forall C \in \mathcal{B}$.

**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 $P(X = x) > 0 \forall x \in A$.

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

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

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

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 $\mathbb{R}$, 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 $X = X^{-1}(\{x : f_X(x) > 0\})$.

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 (strictly) **non-decreasing** if $x < y \Rightarrow f(x) (\leq) < f(y)$ $\forall x, y \in D$
- $f$ is (strictly) **non-increasing** if $x < y \Rightarrow f(x) (\geq) \geq f(y)$ $\forall x, y \in D$
- $f$ 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.

Then, from Rohatgi, page 9, Theorem 15:

$$\frac{d}{dy}g^{-1}(y) = \left( \frac{d}{dx}g(x) \bigg|_{x=g^{-1}(y)} \right)^{-1} > 0$$

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,

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}(F_X(g^{-1}(y))) \overset{\text{By Chain Rule}}{=} f_X(g^{-1}(y)) \cdot \frac{d}{dy}g^{-1}(y)$$

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.

Then, from Rohatgi, page 9, Theorem 15:

$$\frac{d}{dy}g^{-1}(y) = \left( \frac{d}{dx}g(x) \bigg|_{x=g^{-1}(y)} \right)^{-1} < 0$$

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,

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}(1 - F_X(g^{-1}(y))) \overset{\text{By Chain Rule}}{=} -f_X(g^{-1}(y)) \cdot \frac{d}{dy}g^{-1}(y) = f_X(g^{-1}(y)) \cdot \left( -\frac{d}{dy}g^{-1}(y) \right)$$

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
\[ f_Y(y) = \left. \frac{f(x)}{\frac{dg(x)}{dx}} \right|_{x=g^{-1}(y)}, \quad y \in g(\mathcal{X}) \]

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:

**Theorem 2.4.7:**
Let \( Y = g(X) \) where \( X \) is a rv with pdf \( f_X(x) \) on support \( \mathcal{X} \). Suppose there exists a partition \( A_0, A_1, \ldots, A_k \) of \( \mathcal{X} \) 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 \( \mathcal{Y} = g_i(A_i) = \{ y : y = g_i(x) \text{ for some } x \in A_i \} \) is the same for each \( i = 1, \ldots, k \), and

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

Then,
\[ f_Y(y) = \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \cdot \left| \frac{d}{dy} g_i^{-1}(y) \right| \cdot I_\mathcal{Y}(y) \]

Note:
Rohatgi, page 73, Theorem 4, removes condition (iii) by defining \( n = n(y) \) and \( x_1(y), \ldots, x_n(y) \).
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 $Y$.

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,

$$f_Y(y) = F'_Y(y)$$

$$= f_X(\sin^{-1}(y)) \frac{1}{\sqrt{1 - y^2}} + (-1) f_X(\pi - \sin^{-1}(y)) \frac{1}{\sqrt{1 - y^2}}$$

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

$$= \frac{1}{\sqrt{1 - y^2}} \left( \frac{2(\sin^{-1}(y))}{\pi^2} + \frac{2(\pi - \sin^{-1}(y))}{\pi^2} \right)$$

$$= \frac{1}{\pi^2 \sqrt{1 - y^2}} \cdot 2\pi$$

$$= \frac{2}{\pi \sqrt{1 - y^2}} \cdot I_{[0,1]}(y)$$

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

Thus, by use of Theorem 2.4.7, we get

$$f_Y(y) = \sum_{i=1}^{2} f_X(g_i^{-1}(y)) \cdot \left| \frac{dy}{dy}g_i^{-1}(y) \right| \cdot I_Y(y)$$

40
\[= \frac{2\sin^{-1}(y)}{\pi^2} \frac{1}{\sqrt{1 - y^2}} \cdot I_{(0,1)}(y) + \frac{2(\pi - \sin^{-1}(y))}{\pi^2} \frac{1}{\sqrt{1 - y^2}} \cdot I_{(0,1)}(y)\]
\[= \frac{2\pi}{\pi^2} \frac{1}{\sqrt{1 - y^2}} \cdot I_{(0,1)}(y)\]
\[= \frac{2}{\pi\sqrt{1 - y^2}} \cdot I_{(0,1)}(y)\]

Obviously, both results are identical. ■

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

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

For \(Y = F_X(X)\) and \(0 < y < 1\), we have
\[P(Y \leq y) = P(F_X(X) \leq y)\]
\[= P(F^{-1}_X(F_X(X)) \leq F^{-1}_X(y))\]
\[= P(X \leq F^{-1}_X(y))\]
\[= F_X(F^{-1}_X(y))\]
\[= y\]

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 (*) true? — In (a), if \(F_X\) is strictly increasing and continuous, it is certainly \(x = F^{-1}_X(F_X(x))\).

In (b), if \(F_X(x_1) = F_X(x_2)\) for \(x_1 < x < x_2\), it may be that \(F^{-1}_X(F_X(x)) \neq x\). But by definition, \(F^{-1}_X(F_X(x)) = x_1 \forall x \in [x_1, x_2]\). (*) 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. ■
3 Moments and Generating Functions

3.1 Expectation

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:
$X \sim$ 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:**

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

**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 \(k^{th}\) 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:**

\[
egin{align*}
Var(X) &= E((X - E(X))^2) \\
&= E(X^2) - 2XE(X) + (E(X))^2 \\
&= E(X^2) - 2E(X)E(X) + (E(X))^2 \\
&= E(X^2) - (E(X))^2
\end{align*}
\]
Theorem 3.1.7:
If $\text{Var}(X)$ exists and $a$ and $b$ are finite constants, then $\text{Var}(aX + b)$ exists and equals $a^2 \text{Var}(X)$.

Proof:
Existence & Numerical Result:
$\text{Var}(aX + b) = E \left( \left( (aX + b) - E(aX + b) \right)^2 \right)$ exists if $E \left( |(aX + b) - E(aX + b)|^2 \right)$ exists.

It holds that

$$
\begin{align*}
E \left( |(aX + b) - E(aX + b)|^2 \right) &= E \left( (aX + b) - E(aX + b) \right)^2 \\
&= \text{Var}(aX + b) \\
&\overset{\text{Th.3.1.6}}{=} E((aX + b)^2) - (E(aX + b))^2 \\
&\overset{\text{Th.3.1.2}}{=} a^2 E(X^2) + 2abE(X) + b^2 - (a E(X) + b)^2 \\
&\overset{\text{Th.3.1.2}}{=} a^2 E(X^2) + 2abE(X) + b^2 - a^2 (E(X))^2 - 2abE(X) - b^2 \\
&= a^2 (E(X^2) - (E(X))^2) \\
&\overset{\text{Th.3.1.6}}{=} a^2 \text{Var}(X) \\
&< \infty \text{ since } \text{Var}(X) \text{ exists}
\end{align*}
$$
Theorem 3.1.8:
If the $t^{th}$ moment of a rv $X$ exists, then all moments of order $0 < s < t$ exist.

Proof:
Continuous case only:

$$E(|X|^t) = \int_{|x| \leq 1} |x|^t f_X(x) dx + \int_{|x| > 1} |x|^t f_X(x) dx$$
$$\leq \int_{|x| \leq 1} 1 \cdot f_X(x) dx + \int_{|x| > 1} |x|^t f_X(x) dx$$
$$\leq P(|X| \leq 1) + E(|X|^t)$$
$$< \infty$$

Theorem 3.1.9:
If the $t^{th}$ moment of a rv $X$ exists, then

$$\lim_{n \to \infty} n^t P(|X| > n) = 0.$$ 

Proof:
Continuous case only:

$$\infty > \int_{\mathbb{R}} |x|^t f_X(x) dx = \lim_{n \to \infty} \int_{|x| \leq n} |x|^t f_X(x) dx$$
$$\Rightarrow \lim_{n \to \infty} \int_{|x| > n} |x|^t f_X(x) dx = 0$$

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

Note:
The inverse is not necessarily true, i.e., if $\lim_{n \to \infty} n^t P(|X| > n) = 0$, then the $t^{th}$ moment of a rv $X$ does not necessarily exist. We can only approach $t$ 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 $\lim_{n\to\infty} n^t P(|X| > n) = 0$ for some $t > 0$. Then,

$$E(|X|^s) < \infty \quad \forall \ 0 < s < t.$$ 

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

$$\int_0^n x f_X(x) \, dx = (xF(x)) \bigg|_0^n - \int_0^n F_X(x) \, dx$$

$$= nF_X(n) - 0F_X(0) - \int_0^n F_X(x) \, dx$$

$$= nF_X(n) - n + \int_0^n F_X(x) \, dx$$

$$= nF_X(n) - n + \int_0^n [1 - F_X(x)] \, dx$$

46
\[
\begin{align*}
= n[F_X(n) - 1] + \int_0^n [1 - F_X(x)]dx \\
= -n[1 - F_X(n)] + \int_0^n [1 - F_X(x)]dx \\
= -nP(X > n) + \int_0^n [1 - F_X(x)]dx \\
X \geq 0 \\
= -n^1P(|X| > n) + \int_0^n [1 - F_X(x)]dx \\
\implies E(X^1) = \lim_{n \to \infty} \left[-n^1P(|X| > n) + \int_0^n [1 - F_X(x)]dx \right] \\
= \text{Th. } 3.1.9 + \lim_{n \to \infty} \int_0^n [1 - F_X(x)]dx \\
= \int_0^\infty [1 - F_X(x)]dx
\end{align*}
\]

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

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

Thus,

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

\( \implies E(X) \) 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:

\[ E(|X|^s) \] (Lemma 3.1.11) \[ = \int_0^\infty [1 - F_X(z)]dz = \int_0^\infty P(|X| > z)dz. \]

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

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

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

monotonic \( \uparrow \)

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

Proof (of Theorem 3.1.10):

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

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

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

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

\[ = y^s \bigg|_0^N + se \int_N^\infty y^{s-1-\epsilon}dy \]

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

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

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

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 \( \lim_{n \to \infty} n^tP(|X| > n) = 0 \) 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
  $$\frac{P(|X| > \alpha k)}{P(|X| > k)} < \epsilon \quad \forall k \geq k_0.$$ 
- 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 $r$:
  $$\frac{P(|X| > \alpha^r k)}{P(|X| > k)} = \frac{P(|X| > \alpha k)}{P(|X| > k)} \cdot \frac{P(|X| > \alpha^2 k)}{P(|X| > \alpha k)} \cdot \frac{P(|X| > \alpha^3 k)}{P(|X| > \alpha^2 k)} \cdots \frac{P(|X| > \alpha^r k)}{P(|X| > \alpha^{r-1} k)}$$
  $$\Rightarrow \frac{P(|X| > \alpha^r k)}{P(|X| > k)} = \frac{P(|X| > \alpha k)}{P(|X| > k)} \cdot \frac{P(|X| > \alpha \cdot (\alpha k))}{P(|X| > \alpha k)} \cdot \frac{P(|X| > \alpha \cdot (\alpha^2 k))}{P(|X| > \alpha^2 k)} \cdots \frac{P(|X| > \alpha \cdot (\alpha^{r-1} k))}{P(|X| > \alpha^{r-1} k)}$$
- Note: Each of these $r$ terms on the right side is $> \epsilon$ by our original statement of selecting some $k_0$ such that $\frac{P(|X| > \alpha k)}{P(|X| > k)} < \epsilon \quad \forall k \geq k_0$ and since $\alpha > 1$ and therefore $\alpha^n k \geq k_0$.
- 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 $n$:
  $$\text{E}(|X|^n) = \int_0^\infty x^{n-1} P(|X| > x)dx = \int_0^N x^{n-1} P(|X| > x)dx + \int_N^\infty x^{n-1} P(|X| > x)dx$$
- We know that:
  $$\int_0^N x^{n-1} P(|X| > x)dx \leq \int_0^N x^{n-1} dx = x^n \bigg|_0^N = N^n \quad < \infty$$
but is
\[ n \int_{N}^{\infty} x^{n-1} P(|X| > x) dx < \infty \]?

- To check the second part, we use:
\[ \int_{N}^{\infty} x^{n-1} P(|X| > x) dx = \sum_{r=1}^{\infty} \int_{\alpha^{-1} N}^{\alpha^r N} x^{n-1} P(|X| > x) dx \]

- We know that:
\[ \int_{\alpha^{-1} N}^{\alpha^r N} x^{n-1} P(|X| > x) dx \leq \epsilon^r \int_{\alpha^{-1} N}^{\alpha^r N} x^{n-1} dx \]

This step is possible since \( \epsilon^r \geq P(|X| \geq \alpha^{-1} N) \geq P(|X| > x) \geq P(|X| \geq \alpha^r N) \) \( \forall x \in (\alpha^{-1} N, \alpha^r N) \) and \( N = \max(k_0, k_1) \).

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

- Now we go back to our original inequality:
\[ \int_{N}^{\infty} x^{n-1} P(|X| > x) dx \leq \sum_{r=1}^{\infty} \epsilon^r \int_{\alpha^{-1} N}^{\alpha^r N} x^{n-1} dx \leq \sum_{r=1}^{\infty} \epsilon^r (\alpha^r N)^n = N^n \sum_{r=1}^{\infty} (\epsilon \alpha^r)^n \]
\[ = \frac{N^n \epsilon \alpha^n}{1 - \epsilon \alpha^n} \text{ if } \epsilon \alpha^n < 1 \text{ or, equivalently, if } \epsilon < \frac{1}{\alpha^n} \]

- Since \( \frac{N^n \epsilon \alpha^n}{1 - \epsilon \alpha^n} \) is finite, all moments \( E(|X|^n) \) exist.

\[ \Box \]
3.2 Generating Functions

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 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

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

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.

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

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

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

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

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

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

$$\frac{d^n}{dt^n} M_X(t) = \left. \frac{d}{dt} \left( \frac{d^{n-1}}{dt^{n-1}} M_X(t) \right) \right|_{t=0}$$

$$= \left. \frac{d}{dt} \left( \int_{-\infty}^{\infty} x^{n-1} e^{tx} f_X(x) \, dx \right) \right|_{t=0}$$

$$= \int_{-\infty}^{\infty} \left( \frac{\partial}{\partial t} x^{n-1} e^{tx} f_X(x) \right) \, dx$$
\[
\begin{align*}
\quad & = \int_{-\infty}^{\infty} x^n e^{tx} f_X(x) \, dx \\
& = E(X^n e^{tX})
\end{align*}
\]

Evaluating this at \( t = 0 \), we get: \( \frac{d^n}{dt^n} M_X(t) \mid_{t=0} = E(X^n) \)

---

Example 3.2.3:
\( X \sim U(a, b) \); \( f_X(x) = \frac{1}{b-a} \cdot I_{[a,b]}(x) \).
Then,
\[
M_X(t) = \int_a^b \frac{e^{tx}}{b-a} \, dx = \frac{e^{tb} - e^{ta}}{t(b-a)}
\]

\[
M_X(0) = \frac{0}{0} \quad \text{L'Hôpital}
\]

\[
= \frac{be^{tb} - ae^{ta}}{b-a}
\]

\[
= 1
\]

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 \( t \in \mathbb{R} \)).

\[
M_X'(t) = \frac{(be^{tb} - ae^{ta})t(b-a) - (e^{tb} - e^{ta})(b-a)}{t^2(b-a)^2}
\]

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

\[
\implies E(X) = M_X'(0) = \frac{0}{0} \quad \text{L'Hôpital}
\]

\[
= \frac{be^{tb} - ae^{ta} + t^2 e^{tb} - ta^2 e^{ta} - be^{tb} + ae^{ta}}{2t(b-a)} \bigg|_{t=0}
\]

\[
= \frac{t^2 e^{tb} - ta^2 e^{ta}}{2t(b-a)} \bigg|_{t=0} = 0
\]

\[
= \frac{b^2 - a^2}{2(b-a)}
\]

\[
= \frac{b + a}{2}
\]

52
Note:

In the previous example, we made use of L'Hopital's rule. This rule gives conditions under which we can resolve indefinite expressions of the type \( \frac{\infty}{\infty} \) and \( \frac{0}{\infty} \).

(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 \( g'(x) \neq 0 \) \( \forall x \in (x_0 - \delta, x_0 + \delta) \setminus \{x_0\} \). Then, \( \lim_{x \to x_0} \frac{f'(x)}{g'(x)} = A \implies \lim_{x \to x_0} \frac{f(x)}{g(x)} = A \). The same holds for the cases \( \lim_{x \to x_0} f(x) = \lim_{x \to x_0} g(x) = \infty \) and \( x \to x_0^+ \) or \( x \to x_0^- \).

(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, \( \lim_{x \to \infty} \frac{f'(x)}{g'(x)} = A \implies \lim_{x \to \infty} \frac{f(x)}{g(x)} = A \).

(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 \( \frac{0 \cdot \infty}{0 \cdot \infty} \) 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 \( f(x, \theta), a(\theta), \) and \( b(\theta) \) are differentiable with respect to \( \theta \) (for all \( x \) and \( -\infty < a(\theta) < b(\theta) < \infty \), then

\[
\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) \, dx = f(b(\theta), \theta) \frac{d}{d\theta} b(\theta) - f(a(\theta), \theta) \frac{d}{d\theta} a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x, \theta) \, dx.
\]

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: Lebesgue’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

\[
\int f_n(x) \, dx \to \int f(x) \, dx.
\]

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

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

Theorem 3.2.6:

Let \( f_n(x, \theta_0) = \frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} \) 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 \( \left| \frac{\partial}{\partial \theta} f(x, \theta) \right|_{\theta = \theta_0} \leq g(x, \theta) \ \forall x \ \forall \theta_0 \) 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 \[ \left| \frac{\partial}{\partial t} e^{tx} f_X(x) \right|_{t=t'} \leq |x| e^{tx} f_X(x) \text{ for } |t' - t| \leq \delta_0. \]

Choose \( t, \delta_0 \) small enough such that \( t + \delta_0 \in (-h, h) \) and \( t - \delta_0 \in (-h, h) \). 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 \( h > 1 \). 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, \( \int_{-\infty}^{\infty} g(x, t) dx \leq M_X(t + \delta_0 + 1) < \infty \) and \( \int_{-\infty}^{0} g(x, t) dx \leq M_X(t - \delta_0 - 1) < \infty \) 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 r.v’s with pdfs

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

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 Homeworks.

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 Homeworks!

**Theorem 3.2.9:**
Let $X$ and $Y$ be 2 r.v’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 r.v’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:

$$M_Y(t) = E(e^{(aX+b)t})$$
$$= E(e^{aX}e^{bt})$$
$$= e^{bt}E(e^{Xat})$$
$$= e^{bt}M_X(at)$$

3.3 Complex-Valued Random Variables and Characteristic Functions

Recall the following facts regarding complex numbers:

$i^0 = +1; i = \sqrt{-1}; i^2 = -1; i^3 = -i; i^4 = +1$; 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:

$$z_1 \pm z_2 = (a_1 \pm a_2) + i(b_1 \pm b_2)$$
$$z_1 \cdot z_2 = r_1r_2e^{i(\phi_1 + \phi_2)} = r_1r_2(\cos(\phi_1 - \phi_2) + i\sin(\phi_1 - \phi_2))$$
$$\frac{z_1}{z_2} = \frac{r_1}{r_2}e^{i(\phi_1 - \phi_2)} = \frac{r_1}{r_2}(\cos(\phi_1 - \phi_2) + i\sin(\phi_1 - \phi_2))$$

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

$$\sqrt[n]{z} = \sqrt[n]{a + ib} = \sqrt[n]{r} \left(\cos\left(\frac{\phi + k\cdot360^\circ}{n}\right) + i\sin\left(\frac{\phi + k\cdot360^\circ}{n}\right)\right)$$
for $k = 0, 1, \ldots, (n - 1)$ and the main value for $k = 0$

$$\ln z = \ln(a + ib) = \ln(|z|) + i\phi \pm 2n\pi$$
where $\phi = \arctan \frac{b}{a}$ and the main value for $n = 0$
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 \( z \in \mathbb{R} \)
- \( \overline{z_1 \pm z_2} = \overline{z_1} \pm \overline{z_2} \)
- \( \overline{z_1 \cdot z_2} = \overline{z_1} \cdot \overline{z_2} \)
- \( \overline{\left( \frac{z_1}{z_2} \right)} = \frac{\overline{z_1}}{\overline{z_2}} \)
- \( z \cdot \overline{z} = a^2 + b^2 \)
- \( \text{Re}(z) = a = \frac{1}{2}(z + \overline{z}) \)
- \( \text{Im}(z) = b = \frac{1}{2i}(z - \overline{z}) \)
- \( |z| = \sqrt{a^2 + b^2} = \sqrt{z \cdot \overline{z}} \)

**Definition 3.3.1:**

Let \((\Omega, L, P)\) be a probability space and \( X \) and \( Y \) real-valued rv’s, i.e., \( X, Y : (\Omega, L) \rightarrow (\mathbb{R}, B) \)

(i) \( Z = X + iY : (\Omega, L) \rightarrow (\mathcal{C}, B_{\mathcal{C}}) \) is called a **complex-valued random variable** (\( \mathcal{C} \)-rv).

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

**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, \( \Phi_X(t) : \mathbb{R} \rightarrow \mathcal{C} \) 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) \( \Phi_X(t) = \sum_{x \in \mathcal{X}} e^{itx}P(X = x) = \sum_{x \in \mathcal{X}} \cos(tx)P(X = x) + i \sum_{x \in \mathcal{X}} \sin(tx)P(X = x)(x) \) 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) $|\Phi_X(t)| \leq 1 \quad \forall t \in \mathbb{R}$.

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

(iv) $\Phi_X$ is a positive definite function, i.e., $\forall n \in \mathbb{N} \quad \forall \alpha_1, \ldots, \alpha_n \in \mathbb{C} \forall t_1, \ldots, t_n \in \mathbb{R}:
\sum_{l=1}^{n} \sum_{j=1}^{n} \alpha_l \alpha_j^* \Phi_X(t_l - t_j) \geq 0$.

(v) $\Phi_X(t) = \Phi_X(-t)$.

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

(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 $\varepsilon > 0$.

(ii) $\exists a > 0 : P(-a < X < +a) > 1 - \frac{\varepsilon}{2}$ and $P(|X| > a) < \frac{\varepsilon}{2}$.

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

This third condition holds since $|e^{a} - 1| = 0$ and the exponential function is continuous. Therefore, if we select $(t' - t)$ and $x$ small enough, $|e^{(t'-t)x} - 1|$ will be $< \frac{\varepsilon}{2}$ for a given $\varepsilon$. 59
Let $t, t' \in \mathbb{R}$, $t < t'$, and $t' - t < \delta$. Then,

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

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

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

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

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

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

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

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

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

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

(A) holds due to Note (iii) that follows Definition 3.3.2.

Similarly,

$$\left| \int_{a}^{\infty} (e^{it'x} - e^{itx}) f_X(x) dx \right| \leq \int_{a}^{\infty} 2 f_X(x) dx$$

Returning to the main part of the proof, we get

$$\left| \Phi_X(t') - \Phi_X(t) \right| \leq \int_{-\infty}^{-a} 2 f_X(x) dx + \left| \int_{-a}^{a} (e^{it'x} - e^{itx}) f_X(x) dx \right| + \int_{a}^{\infty} 2 f_X(x) dx$$

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

$$= 2 P(|X| > a) + \left| \int_{-a}^{a} (e^{it'x} - e^{itx}) f_X(x) dx \right|$$
\begin{align*}
\text{Condition } (ii) & \leq \frac{\epsilon}{4} + \left| \int_{-a}^{+a} (e^{it'x} - e^{tx}) f_X(x) \, dx \right| \\
& = \frac{\epsilon}{2} + \left| \int_{-a}^{+a} e^{itx} (e^{i(t'-t)x} - 1) f_X(x) \, dx \right| \\
& \leq \frac{\epsilon}{2} + \int_{-a}^{+a} \left| e^{itx} \right| \left| (e^{i(t'-t)x} - 1) \right| 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
\end{align*}

(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) $X \sim U(a, b)$</td>
<td>$\frac{e^{ita} - e^{itb}}{(b-a)it}$</td>
</tr>
<tr>
<td>(v) $X \sim N(0,1)$</td>
<td>$\exp(-t^2/2)$</td>
</tr>
<tr>
<td>(vi) $X \sim N(\mu, \sigma^2)$</td>
<td>$e^{it\mu} \exp(-\sigma^2t^2/2)$</td>
</tr>
<tr>
<td>(vii) $X \sim \Gamma(p, q)$</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) $X \sim \chi_n^2$</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) $\Phi_X(t) = \sum_{k=0}^{1} e^{ikt} P(X = k) = e^{it0}(1 - p) + e^{it1}p = 1 + p(e^{it} - 1)$

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

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

(v) $X \sim N(0, 1)$ is symmetric around 0

$\Rightarrow \Phi_X(t)$ is real since there is no imaginary part according to Theorem 3.3.3 (vi)

$\Rightarrow \Phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(tx) e^{-\frac{x^2}{2}} dx$
Since \( E(X) \) exists, \( \Phi_X(t) \) is differentiable according to Theorem 3.3.5 and the following holds:

\[
\Phi_X(t) = \text{Re}(\Phi_X'(t))
\]

\[
= \text{Re} \left( \int_{-\infty}^{\infty} i x \cos(tx) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \right)
= \text{Re} \left( \int_{-\infty}^{\infty} i x \cos(tx) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx + \int_{-\infty}^{\infty} -x \sin(tx) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \right)
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left( \frac{-\sin(tx)}{u} \right) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \quad | \quad u' = -t \cos(tx) \land v = -e^{-\frac{x^2}{2}}
= \frac{1}{\sqrt{2\pi}} \sin(tx) e^{-\frac{x^2}{2}} \bigg|_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (-t \cos(tx))(-e^{-\frac{x^2}{2}}) \, dx
= -t \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(tx) e^{-\frac{x^2}{2}} \, dx
= -t \Phi_X(t)
\]

Thus, \( \Phi_X'(t) = -t \Phi_X(t) \). It follows that \( \frac{\Phi_X'(t)}{\Phi_X(t)} = -t \) 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 continuous, \( \Phi_X \) must take the value 0 before it can eventually take a negative value. However, since \( e^{-\frac{1}{2}t^2} > 0 \ \forall t \in \mathbb{R} \), \( \Phi_X \) cannot take 0 as a possible value and therefore cannot pass into the negative numbers. So, it must hold that \( \Phi_X(t) = e^{-\frac{1}{2}t^2} \ \forall t \in \mathbb{R} \).

(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)

\[
\Phi_X(t) = \int_{0}^{\infty} e^{itx}(p,q,x) \, dx
= \int_{0}^{\infty} \frac{q^p}{\Gamma(p)} x^{p-1} e^{-(q-it)x} \, dx
= \frac{q^p}{\Gamma(p)} (q-it)^{-p} \int_{0}^{\infty} ((q-it)x)^{p-1} e^{-(q-it)x} \, dx \quad | \quad u = (q-it)x, \ du = (q-it) \, dx
\]
\[
\frac{q^p}{\Gamma(p)}(q - it)^{-p} \int_0^\infty (u)^{p-1}e^{-u}du
\]

\[
=q^p(q - it)^{-p}
\]

\[
\frac{(q - it)^{-p}}{q^{-p}}
\]

\[
\left(\frac{q - it}{q}\right)^{-p}
\]

\[
\left(1 - \frac{it}{q}\right)^{-p}
\]

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

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

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

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

Example 3.3.13:
Since we know that \(m_1 = E(X)\) and \(m_2 = E(X^2)\) exist for \(X \sim B\in(1, p)\), we can determine these moments according to Theorem 3.3.5 using the characteristic function.

It is

\[
\Phi_X(t) = 1 + p(e^{it} - 1)
\]

\[
\Phi'_X(t) = pi e^{it}
\]

\[
\Phi''_X(0) = pi
\]

\[
\Rightarrow m_1 = \frac{\Phi'_X(0)}{i} = \frac{pi}{i} = p = E(X)
\]

\[
\Phi''_X(t) = p^2 e^{it}
\]

\[
\Phi''_X(0) = p^2
\]

\[
\Rightarrow m_2 = \frac{\Phi''_X(0)}{i^2} = \frac{p^2}{i^2} = p = E(X^2)
\]

\[
\Rightarrow \text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1 - p)
\]
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,
- $X \sim \text{Dirac}(c)$:

$$
\int_{-\infty}^{\infty} |\Phi_X(t)| \, dt = \int_{-\infty}^{\infty} |e^{itc}| \, dt = \int_{-\infty}^{\infty} 1 \, dt = x \bigg|_{-\infty}^{\infty}
$$

which is undefined.

- Also for $X \sim Bin(1,p)$:

$$
\int_{-\infty}^{\infty} |\Phi_X(t)| \, dt = \int_{-\infty}^{\infty} |1 + p(e^{it} - 1)| \, dt = \int_{-\infty}^{\infty} |pe^{it} - (p - 1)| \, dt \geq \int_{-\infty}^{\infty} |pe^{it}| \, dt - \int_{-\infty}^{\infty} |(p - 1)| \, dt

\geq p \int_{-\infty}^{\infty} 1 \, dt - (1 - p) \int_{-\infty}^{\infty} 1 \, dt

= (2p - 1) \int_{-\infty}^{\infty} 1 \, dt

= (2p - 1)x \bigg|_{-\infty}^{\infty}
$$

which is undefined for $p \neq 1/2$.

If $p = 1/2$, we have

$$
\int_{-\infty}^{\infty} |pe^{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, $X \sim N(0,1)$:

\[
\int_{-\infty}^{\infty} |\Phi_X(t)| \, dt = \int_{-\infty}^{\infty} \exp\left(-\frac{t^2}{2}\right) \, dt
\]

\[
= \sqrt{2\pi} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \, dt
\]

\[
= \sqrt{2\pi}
\]

\[
< \infty
\]
3.4 Probability Generating Functions

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 $|s| < 1$.

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

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:

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

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:

$$E(X) = \frac{d}{ds} G(s) \bigg|_{s=1}$$
Definition 3.4.5:
The \( k \)th factorial moment of \( X \) is defined as

\[
E[X(X - 1)(X - 2) \cdots (X - k + 1)]
\]
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 
\( E[X(X - 1)(X - 2) \cdots (X - k + 1)] \) exists, then it holds:

\[
E[X(X - 1)(X - 2) \cdots (X - k + 1)] = \frac{d^k}{ds^k} G(s)_{|s=1}
\]

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.

3.5 Moment Inequalities

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 \( h(X) = \mathbb{E}[X^r] \) and \( \epsilon = k^r \) where \( r > 0 \) and \( k > 0 \). If \( \mathbb{E}[|X|^r] \) exists, then it holds:

\[
P(|X| \geq k) \leq \frac{\mathbb{E}[|X|^r]}{k^r}
\]

Proof:
Since \( P(|X| \geq k) = P(|X|^r \geq k^r) \) for \( k > 0 \), it follows using Theorem 3.5.1:

\[
P(|X| \geq k) = P(|X|^r \geq k^r) \leq \frac{\mathbb{E}[|X|^r]}{k^r}
\]

Corollary 3.5.3: Chebychev’s Inequality
Let \( h(X) = (X - \mu)^2 \) and \( \epsilon = k^2 \sigma^2 \) where \( \mathbb{E}(X) = \mu \), \( \text{Var}(X) = \sigma^2 < \infty \), and \( k > 0 \). Then it holds:

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

Proof:
Since \( P(|X - \mu| > k\sigma) = P(|X - \mu|^2 > k^2\sigma^2) \) for \( k > 0 \), it follows using Theorem 3.5.1:

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

Theorem 3.5.4: Lyapunov Inequality
Let \( 0 < \beta_k = \mathbb{E}(|X|^k) < \infty \). 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., \( (\mathbb{E}(|X|^{k-1}))^{\frac{1}{k-1}} \leq (\mathbb{E}(|X|^k))^{\frac{1}{k}} \).

Proof:
Continuous case only:
Let \( Q(u,v) = \mathbb{E}((u \mid X |^{k-1} + v \mid X |^{k+1})^2) \). Obviously, \( Q(u,v) \geq 0 \) \( \forall u, v \in \mathbb{R} \). Also,

\[
Q(u,v) = \int_{-\infty}^{\infty} \left( u \mid x \mid^{k-1} + v \mid x \mid^{k+1} \right)^2 f_X(x)dx
\]

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

\[
= u^2 \beta_{k-1} + 2uv \beta_k + v^2 \beta_{k+1}
\]

\[
\geq 0 \ \forall u, v \in \mathbb{R}
\]
Using the fact that $Ax^2 + 2Bxy + Cy^2 \geq 0 \ \forall x, y \in \mathbb{R}$ iff $A > 0$ and $AC - B^2 > 0$ (see Rohatgi, page 6, Section P2.4), we get with $A = \beta_{k-1}, B = \beta_k,$ and $C = \beta_{k+1}$:

$$\beta_{k-1} \beta_{k+1} - \beta_k^2 \geq 0$$

$$\implies \beta_k^2 \leq \beta_{k-1} \beta_{k+1}$$

$$\implies \beta_k^2 \leq \beta_{k-1} \beta_{k+1}$$

This means that $\beta_1^2 \leq \beta_0 \beta_2, \beta_2^2 \leq \beta_1 \beta_3, \beta_3^2 \leq \beta_2 \beta_4,$ and so on. Multiplying these, we get:

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

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

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

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

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

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

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

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

$$\implies \beta_{k-1} \leq \beta_{k}^{k}$$
4 Random Vectors

4.1 Joint, Marginal, and Conditional Distributions

Definition 4.1.1:
The vector $\mathbf{X} = (X_1, \ldots, X_n)'$ on $(\Omega, L, P) \rightarrow \mathbb{R}^n$ defined by $\mathbf{X}(\omega) = (X_1(\omega), \ldots, X_n(\omega))'$, $\omega \in \Omega$, is an $n$-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 $n$-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 $\mathbf{X} = (X_1, \ldots, X_n)'$ is an n-rv on $(\Omega, L, P)$ since for any $I$, it holds:

$$X^{-1}(I) = \{\omega : (X_1(\omega), \ldots, X_n(\omega)) \in I\} = \{\omega : X_1(\omega) \leq a_1, \ldots, X_n(\omega) \leq a_n\} = \bigcap_{k=1}^{n} \{\omega : X_k(\omega) \leq a_k\}_{\in L}^{\in L}$$

Definition 4.1.2:
For an n-rv $\mathbf{X}$, a function $F$ defined by

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

is the joint cumulative distribution function of $\mathbf{X}$. 

Note:

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

(ii) $\lim_{\mathbf{x} \rightarrow \mathbf{x}} F(\mathbf{x}) = \lim_{x_1 \rightarrow \infty, \ldots, x_n \rightarrow \infty} F(\mathbf{x}) = 1$ and $\lim_{x_k \rightarrow -\infty} F(\mathbf{x}) = 0$ \forall $x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n \in \mathbb{R}$. 

72
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) \( \forall \mathbf{z} \in \mathbb{R}^n \ \forall \epsilon_i > 0, i = 1, \ldots, n \), the following inequality holds:

\[
F(z + \mathbf{\epsilon}) - \sum_{i=1}^{n} F(x_1 + \epsilon_1, \ldots, x_i - 1 + \epsilon_i - 1, x_{i+1}, \ldots, x_n + \epsilon_n)
+ \sum_{1 \leq i < j \leq n} F(x_1 + \epsilon_1, \ldots, x_i - 1 + \epsilon_i - 1, x_{i+1}, \ldots, x_j - 1 + \epsilon_j - 1, x_{j+1}, \ldots, x_n + \epsilon_n)
+ \ldots
+ (-1)^n F(z)
\geq 0
\]

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

\[
P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = P(X \leq x_2, Y \leq y_2) - P(X \leq x_1, Y \leq y_2) - P(X \leq x_2, Y \leq y_1) + P(X \leq x_1, Y \leq y_1) \geq 0
\]

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 \((x_i, y_i)\) 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 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** 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
\[
F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) \, dv \, du \quad \forall (x, y) \in \mathbb{R}^2
\]
where \(F\) is the joint cdf of \((X, Y)\). We call \(f\) the **joint probability density function** of \((X, Y)\).

Note:
If \(F\) is 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** of \(X\) and \(f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, dx\) is called the **marginal probability density function** of \(Y\).

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

(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** of \( (X_{i_1}, \ldots, X_{i_k}) \), \( 1 \leq k \leq n-1, 1 \leq i_1 < i_2 < \cdots < i_k \leq n \), is given by

\[
\lim_{x_i \to \infty, i \neq i_1, \ldots, i_k} F_X(x) = F_X(\infty, \ldots, \infty, x_{i_1}, \infty, \ldots, \infty, x_{i_2}, \infty, \ldots, \infty, x_{i_k}, \infty, \ldots, \infty).
\]

**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** of \( X \) given \( Y = y_j \) (for fixed \( j \)) is defined as

\[
p_{i|j} = P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{p_j}.
\]

**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** of a rv $X$ given that $Y = y$ is defined to be

$$F_{X|Y}(x \mid y) = \lim_{\epsilon \to 0^+} P(X \leq x \mid Y \in (y - \epsilon, y + \epsilon])$$

provided that this limit exists. If it does exist, the **conditional probability density function** of $X$ given that $Y = y$ is any non-negative function $f_{X|Y}(x \mid y)$ satisfying

$$F_{X|Y}(x \mid y) = \int_{-\infty}^{x} f_{X|Y}(t \mid y) dt \quad \forall x \in \mathbb{R}.$$

Note:
For fixed $y$, $f_{X|Y}(x \mid y) \geq 0$ and $\int_{-\infty}^{\infty} f_{X|Y}(x \mid y) dx = 1$. 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

$$F_{X|Y}(x \mid y) = \lim_{\epsilon \to 0^+} \frac{P(X \leq x, Y \in (y - \epsilon, y + \epsilon])}{P(Y \in (y - \epsilon, y + \epsilon])}$$

$$= \lim_{\epsilon \to 0^+} \left( \frac{\frac{1}{2} \int_{-\infty}^{x} \int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(u,v) dv \ du}{\frac{1}{2} \int_{y-\epsilon}^{y+\epsilon} f_Y(v) dv} \right)$$

$$= \int_{-\infty}^{x} \frac{f_{X,Y}(u,y) du}{f_Y(y)}$$

$$= \int_{-\infty}^{x} \frac{f_{X,Y}(u,y)}{f_Y(y)} du.$$

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

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
\[ f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx = \int_0^y 2 \, dx = 2y \text{ for } 0 < y < 1 \]

The conditional pdf’s \( f_{Y|X}(y | x) \) and \( f_{X|Y}(x | y) \) are calculated as follows:
\[ f_{Y|X}(y | x) = \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{2}{2(1 - x)} = \frac{1}{1 - x} \text{ for } x < y < 1 \text{ (where } 0 < x < 1) \]
and
\[ f_{X|Y}(x | y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{2}{2y} = \frac{1}{y} \text{ for } 0 < x < y \text{ (where } 0 < y < 1) \]

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

4.2 Independent Random Variables

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
\[ f_{\alpha}(x_1, x_2, x_3) = f_1(x_1) f_2(x_2) f_3(x_3) \cdot (1 + \alpha(2 F_1(x_1) - 1)(2 F_2(x_2) - 1)(2 F_3(x_3) - 1)). \]

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:

$$P(a < X \leq b, c < Y \leq d) = F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c)$$

$$= F_X(b)F_Y(d) - F_X(a)F_Y(d) - F_X(b)F_Y(c) + F_X(a)F_Y(c)$$

$$= (F_X(b) - F_X(a))(F_Y(d) - F_Y(c))$$

$$= P(a < X \leq b)P(c < Y \leq d)$$

Definition 4.2.4:
A collection of rv’s $X_1, \ldots, X_n$ with joint cdf $F_{X}(x)$ and marginal cdf’s $F_{X_i}(x_i)$ are mutually (or completely) independent iff

$$F_{X}(x) = \prod_{i=1}^{n} F_{X_i}(x_i) \quad \forall x \in \mathbb{R}^n.$$  

Note:
We often simply say that the rv’s $X_1, \ldots, X_n$ are independent when we really mean that they are mutually independent.
Theorem 4.2.5: **Factorization Theorem**

(i) A necessary and sufficient condition for discrete rv’s $X_1, \ldots, X_n$ to be independent is that

$$P(X = x) = P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i) \quad \forall x \in \mathcal{X}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is the countable support of $X$.

(ii) For an absolutely continuous n-rv $X = (X_1, \ldots, X_n)$, $X_1, \ldots, X_n$ are independent iff

$$f_X(x) = f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i),$$

where $f_X$ is the joint pdf and $f_{X_i}$ are the marginal pdfs of $X$.

**Proof:**

(i) **Discrete case:**

This is a generalization of Rohatgi, page 120, Theorem 1 to $n$-dimensional random vectors. The Theorem is based on Lemma 4.2.3 (see also Rohatgi, page 120, Lemma 1) which gives the cumulative probability of a completely bounded region in a two-dimensional support, $a < X \leq b$, $c < Y \leq d$. The extension of the Lemma to $n$-dimensional space gives the cumulative probability in a completely bounded region in an $n$-dimensional support (hyperspace).

Let $X$ be a random vector whose components are independent random variables of the discrete type with $P(X = b) > 0$. Lemma 4.2.3 extends to:

$$\lim_{\mathbb{B}^n} P(a < X \leq b) = \lim_{a_i \uparrow b_i \forall i \in \{1,\ldots,n\}} P(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, \ldots, a_n < X_n \leq b_n)$$

$$= P(X_1 = b_1, X_2 = b_2, \ldots, X_n = b_n)$$

$$= P(X = b)$$

$$\overset{\text{indep.}}{=} \prod_{i=1}^{n} P(X_i = b_i)$$

Before considering the converse factorization of the joint cdf of an $n$-rv, recall that independence allows each value in the support of a component to combine with each of all possible
combinations of the other component values: a 3-dimensional vector where the first component has support \( \{x_{1a}, x_{1b}, x_{1c}\} \), the second component has support \( \{x_{2a}, x_{2b}, x_{2c}\} \), and the third component has support \( \{x_{3a}, x_{3b}, x_{3c}\} \), has \( 3 \times 3 \times 3 = 27 \) points in its support. Due to independence, all these vectors can be arranged into three sets: the first set having \( x_{1a} \), the second set having \( x_{1b} \), and the third set having \( x_{1c} \), with each set having 9 combinations of \( x_2 \) and \( x_3 \).

\[
\Rightarrow F(x_1, x_2, x_3,)
= P(X_1 = x_{1a}) F(x_2, x_3) + P(X_1 = x_{1b}) F(x_2, x_3) + P(X_1 = x_{1c}) F(x_2, x_3)
= F(x_1)(P(X_2 = x_{2a}) F(x_3, ) + P(X_2 = x_{2b}) F(x_3, ) + P(X_2 = x_{2c}) F(x_3, )]
= F(x_1)[P(X_2 = x_{2a}) + P(X_2 = x_{2b}) + P(X_2 = x_{2c})] F(x_3, )
= F(x_1, F(x_2), F(x_3, )
\]

More generally, for \( n \) dimensions, let \( \mathbf{x} = (x_{i1}, x_{i2}, \ldots, x_{in}) \), \( B = \{ \mathbf{x} : x_{i1} \leq x_1, x_{i2} \leq x_2, \ldots, x_{in} \leq x_n \} \), \( B \in \mathcal{X} \). Then it holds:

\[
F_X(x) = \sum_{\mathbf{x} \in B} P(X = \mathbf{x})
= \sum_{\mathbf{x} \in B} P(X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_n = x_{in})
\]

\[
\text{indep.} \quad \sum_{\mathbf{x} \in B} P(X_1 = x_{i1}) P(X_2 = x_{i2}, \ldots, X_n = x_{in})
= \sum_{x_{i1} \leq x_1} \sum_{x_{i2} \leq x_2, \ldots, x_{in} \leq x_n} P(X_1 = x_{i1}) P(X_2 = x_{i2}, \ldots, X_n = x_{in})
= F_{X_1}(x_1) \sum_{x_{i2} \leq x_2} P(X_2 = x_{i2}, \ldots, X_n = x_{in})
\]

\[
\text{indep.} \quad \sum_{x_{i1} \leq x_1} F_{X_1}(x_1) P(X_2 = x_{i2}) \sum_{x_{i3} \leq x_3, \ldots, x_{in} \leq x_n} P(X_3 = x_{i3}, \ldots, X_n = x_{in})
= F_{X_1}(x_1) F_{X_2}(x_2) \sum_{x_{i3} \leq x_3, \ldots, x_{in} \leq x_n} P(X_3 = x_{i3}, \ldots, X_n = x_{in})
= \ldots
= F_{X_1}(x_1) F_{X_2}(x_2) \ldots F_{X_n}(x_n)
= \prod_{i=1}^{n} F_{X_i}(x_i)
\]

(ii) Continuous case: Homework

\[\blacksquare\]
Theorem 4.2.6:
$X_1, \ldots, X_n$ are independent iff $P(X_i \in A_i, \ i = 1, \ldots, n) = \prod_{i=1}^{n} P(X_i \in A_i) \ \forall$ Borel sets $A_i \in B$
[i.e., rv’s are independent iff all events involving these rv’s are independent].

Proof:
Lemma 4.2.3 and definition of Borel sets.

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.

Proof:
\[
F_{g_1(X_1), g_2(X_2), \ldots, g_n(X_n)}(h_1, h_2, \ldots, h_n) = P(g_1(X_1) \leq h_1, g_2(X_2) \leq h_2, \ldots, g_n(X_n) \leq h_n)
\]
\[
= \prod_{i=1}^{n} P(g_i(X_i) \leq h_i)
\]
\[
= \prod_{i=1}^{n} F_{g_i(X_i)}(h_i)
\]
\[
\text{(i) holds since } g_1^{-1}(-\infty, h_1] \in B, \ldots, g_n^{-1}(-\infty, h_n] \in B
\]

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 \( F_{\mathbf{X},\mathbf{Y}}(x,y) = F_{\mathbf{X}}(x)F_{\mathbf{Y}}(y) \quad \forall x, y \in \mathbb{R}^n \).

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
\[
F_{X|Y}(x \mid y) = F_X(x) \quad \forall x,
\]
and
\[
F_{Y|X}(y \mid x) = F_Y(y) \quad \forall y.
\]

4.3 Functions of Random Vectors

Theorem 4.3.1:
If \( X \) and \( Y \) are rv’s on \((\Omega, L, P) \rightarrow \mathbb{R}\), then

(i) \( X + Y \) 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

$$MAX_n = \max \{X_1, \ldots, X_n\} = X_{(n)}$$

by

$$MAX_n(\omega) = \max \{X_1(\omega), \ldots, X_n(\omega)\} \ \forall \omega \in \Omega$$

and

$$MIN_n = \min \{X_1, \ldots, X_n\} = X_{(1)} = -\max \{-X_1, \ldots, -X_n\}$$

by

$$MIN_n(\omega) = \min \{X_1(\omega), \ldots, X_n(\omega)\} \ \forall \omega \in \Omega.$$ 

Then,

(i) $MIN_n$ and $MAX_n$ are rv’s.

(ii) If $X_1, \ldots, X_n$ are independent, then

$$F_{MAX_n}(z) = P(MAX_n \leq z) = P(X_i \leq z \ \forall i = 1, \ldots, n) = \prod_{i=1}^{n} F_{X_i}(z)$$

and

$$F_{MIN_n}(z) = P(MIN_n \leq z) = 1 - P(X_i > z \ \forall i = 1, \ldots, n) = 1 - \prod_{i=1}^{n} (1 - F_{X_i}(z)).$$

(iii) If $\{X_i\}_{i=1}^{n}$ are iid rv’s with common cdf $F_X$, then

$$F_{MAX_n}(z) = F_X^n(z)$$

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 $F_X$. 

$\blacksquare$
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.
Stat 6710 Mathematical Statistics I
Fall Semester 1999

Lecture 28: Wednesday 11/3/1999

Scribe: Jürgen Symanzik, Rich Madsen

Note:
The previous transformations are special cases of the following Theorem:

Theorem 4.3.3:
If \( g : \mathbb{R}^m \to \mathbb{R}^m \) is a Borel-measurable function (i.e., \( \forall B \subseteq \mathbb{B}^m : g^{-1}(B) \subseteq \mathbb{B}^n \)) and if \( \mathbf{X} = (X_1, \ldots, X_n) \) is an \( n \)-rv, then \( g(\mathbf{X}) \) is an \( m \)-rv.

Proof:
If \( B \subseteq \mathbb{B}^m \), then \( \{ \omega : g(\mathbf{X}(\omega)) \in B \} = \{ \omega : \mathbf{X}(\omega) \in g^{-1}(B) \} \subseteq \mathbb{B}^n \).

Question: How do we handle more general transformations of \( \mathbf{X} \)?

Discrete Case:
Let \( \mathbf{X} = (X_1, \ldots, X_n) \) be a discrete \( n \)-rv and \( \mathcal{X} \subseteq \mathbb{R}^n \) be the countable support of \( \mathbf{X} \), i.e., \( P(\mathbf{X} \in \mathcal{X}) = 1 \) and \( P(\mathbf{X} = \mathbf{x}) > 0 \ \forall \mathbf{x} \in \mathcal{X} \).

Define \( u_i = g_i(x_1, \ldots, x_n), i = 1, \ldots, n \) to be 1-to-1-mappings of \( \mathcal{X} \) onto \( B \). Let \( \mathbf{u} = (u_1, \ldots, u_n)' \). Then

\[
P(\mathbf{U} = \mathbf{u}) = P(g_1(X) = u_1, \ldots, g_n(X) = u_n) = P(X_1 = h_1(u), \ldots, X_n = h_n(u)) \ \forall \mathbf{u} \in B
\]

where \( x_i = h_i(u), i = 1, \ldots, n \), is the inverse transformation (and \( P(\mathbf{U} = \mathbf{u}) = 0 \ \forall \mathbf{u} \notin B \)).

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 \( \sim Bin(n, p) \), \( 0 < p < 1 \). Let \( U = \frac{X}{Y+1} \) and \( V = Y + 1 \). Then \( X = UV \) and \( Y = V - 1 \). So the joint pmf of \( U, V \) is

\[
P(U = u, V = v) = \binom{n}{uv}(1 - p)^{n-uv} \binom{n}{v-1}(1 - p)^{n+1-v}
\]

\[
= \binom{n}{uv} \binom{n}{v-1} p^{uv+v-1}(1 - p)^{2n+1-uv-v}
\]

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

\[
\mathbf{U} = \left( \begin{array}{c}
U_1 \\
\vdots \\
U_n
\end{array} \right) = g(\mathbf{X}) = \left( \begin{array}{c}
g_1(X) \\
\vdots \\
g_n(X)
\end{array} \right),
\]

i.e., \( U_i = g_i(X) \), be a mapping from \( \mathbb{R}^n \) into \( \mathbb{R}^n \).

If \( B \in \mathbb{B}^n \), then

\[
P(\mathbf{U} \in B) = P(\mathbf{X} \in g^{-1}(B)) = \int \cdots \int_{g^{-1}(B)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int \cdots \int_{g^{-1}(B)} f_{\mathbf{X}}(\mathbf{x}) \prod_{i=1}^{n} dx_i
\]

where \( g^{-1}(B) = \{\mathbf{z} = (x_1, \ldots, x_n) \in \mathbb{R}^n : g(\mathbf{z}) \in B\} \).

Suppose we define \( B \) as the half-infinite n-dimensional interval

\[
B_\mathbf{u} = \{(u'_1, \ldots, u'_n) : -\infty < u'_i < u_i \quad \forall i = 1, \ldots, n\}
\]

for any \( \mathbf{u} \in \mathbb{R}^n \). Then the joint cdf of \( \mathbf{U} \) is

\[
G(\mathbf{u}) = P(\mathbf{U} \in B_\mathbf{u}) = P(g_1(\mathbf{X}) \leq u_1, \ldots, g_n(\mathbf{X}) \leq u_n) = \int \cdots \int_{g^{-1}(B_\mathbf{u})} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}.
\]

If \( G \) happens to be absolutely continuous, the joint pdf of \( \mathbf{U} \) will be given by \( f_{\mathbf{U}}(\mathbf{u}) = \frac{\partial^n G(\mathbf{u})}{\partial u_1 \partial u_2 \cdots \partial u_n} \) at every continuity point of \( f_{\mathbf{U}} \).

Under certain conditions, we can write \( f_{\mathbf{U}} \) in terms of the original pdf \( f_{\mathbf{X}} \) of \( \mathbf{X} \) as stated in the next Theorem:
Theorem 4.3.5: Multivariate Transformation

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a continuous n-rv with joint pdf $f_{\mathbf{X}}$.

(i) 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 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 $x_i = h_i(u) = h_i(u_1, \ldots, u_n), i = 1, \ldots, n$, over the range of the transformation $g$.

(ii) Assume both $g$ and $h$ are continuous.

(iii) Assume partial derivatives $\frac{\partial x_i}{\partial u_j} = \frac{\partial h_i(u)}{\partial u_j}$, $i, j = 1, \ldots, n$, exist and are continuous.

(iv) Assume that the Jacobian of the inverse transformation

$$
J = \left| \frac{\partial(x_1, \ldots, x_n)}{\partial(u_1, \ldots, u_n)} \right| = \left| \begin{array}{ccc}
\frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial x_n}{\partial u_1} & \cdots & \frac{\partial x_n}{\partial u_n}
\end{array} \right|
$$

is different from 0 for all $u$ in the range of $g$.

Then the n-rv $\mathbf{U} = g(\mathbf{X})$ has a joint absolutely continuous cdf with corresponding joint pdf

$$
f_{\mathbf{U}}(u) = |J| f_{\mathbf{X}}(h_1(u), \ldots, h_n(u)).
$$

Proof:

Let $u \in \mathbb{R}^n$ and

$$
B_u = \{(u'_1, \ldots, u'_n) : -\infty < u'_i < u_i \ \forall i = 1, \ldots, n\}.
$$

Then,

$$
G_{\mathbf{U}}(u) = \int \cdots \int_{g^{-1}(B_u)} f_{\mathbf{X}}(x) dx = \int \cdots \int_{B_u} f_{\mathbf{X}}(h_1(u), \ldots, h_n(u)) |J| du
$$

The result follows from differentiation of $G_{\mathbf{U}}$.

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 \( \mathcal{X} = \{ \mathbf{x} : f_X(\mathbf{x}) > 0 \} \) be the support of \( X \).

(iii) Suppose that for each \( \mathbf{u} \in B = \{ \mathbf{u} \in \mathbb{R}^n : \mathbf{u} = g(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X} \} \) there is a finite number \( k = k(\mathbf{u}) \) 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 \( h_l(\mathbf{u}) = \begin{pmatrix} h_{l1}(\mathbf{u}) \\
\vdots \\
h_{ln}(\mathbf{u})\end{pmatrix} \), \( \mathbf{u} \in B \), i.e., for each \( \mathbf{u} \in B \), \( h_l(\mathbf{u}) \) is the unique \( \mathbf{x} \in \mathcal{X}_l \) such that \( \mathbf{u} = g(\mathbf{x}) \).

(v) Assume partial derivatives \( \frac{\partial x_i}{\partial u_j} = \frac{\partial h_{li}(\mathbf{u})}{\partial u_j} \), \( l = 1, \ldots, k \), \( i = 1, \ldots, n \), exist and are continuous.

(vi) Assume the Jacobian of each of the inverse transformations
\[
J_l = \begin{vmatrix}
\frac{\partial x_1}{\partial u_1} & \cdots & \frac{\partial x_1}{\partial u_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial x_n}{\partial u_1} & \cdots & \frac{\partial x_n}{\partial u_n}
\end{vmatrix} = \begin{vmatrix}
\frac{\partial h_{11}}{\partial u_1} & \cdots & \frac{\partial h_{1n}}{\partial u_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial h_{ln}}{\partial u_1} & \cdots & \frac{\partial h_{ln}}{\partial u_n}
\end{vmatrix}, \quad l = 1, \ldots, k,
\]
is different from 0 for all \( \mathbf{u} \) in the range of \( g \).

Then the joint pdf of \( U \) is given by
\[
f_U(\mathbf{u}) = \sum_{l=1}^{k} | J_l | f_X(h_{l1}(\mathbf{u}), \ldots, h_{ln}(\mathbf{u})).
\]
Example 4.3.7:
Let \( X, Y \) be iid \( \sim N(0,1) \). Define

\[
U = g_1(X, Y) = \begin{cases} 
\frac{X}{\sqrt{2}}, & Y \neq 0 \\
0, & Y = 0
\end{cases}
\]

and

\[
V = g_2(X, Y) = |Y|.
\]

\( \mathcal{X} = \mathbb{R}^2 \), but \( U, V \) are not 1-to-1 mappings 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:

\[
B \rightarrow \mathcal{X}_1 : x = h_{11}(u, v) = uv \\
y = h_{12}(u, v) = v
\]

\[
B \rightarrow \mathcal{X}_2 : x = h_{21}(u, v) = -uv \\
y = h_{22}(u, v) = -v
\]

\[
J_1 = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} \Rightarrow |J_1| = |v| \\
J_2 = \begin{vmatrix} -v & -u \\ 0 & -1 \end{vmatrix} \Rightarrow |J_2| = |v|
\]

\[
f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-x^2/2}e^{-y^2/2}
\]

\[
f_{U,V}(u,v) = |v| \left( \frac{1}{2\pi} e^{-uv^2/2}e^{-v^2/2} + |v| \frac{1}{2\pi} e^{-(uv)^2/2}e^{-(-v)^2/2} \right)
\]

\[
= \frac{v}{\pi} e^{-\left(\frac{u^2+1}{2}\right)v^2}, \quad -\infty < u < \infty, \quad 0 < v < \infty
\]

Marginal:

\[
f_U(u) = \int_0^\infty \frac{v}{\pi} e^{-\left(\frac{u^2+1}{2}\right)v^2} dv \quad |z = \frac{(u^2+1)v^2}{2}, \quad \frac{dz}{dv} = (u^2+1)v \]

\[
= \int_0^\infty \frac{1}{\pi(u^2+1)} e^{-z} dz
\]

89
\[
\frac{1}{\pi(u^2 + 1)} \left( -e^{-\frac{z}{u}} \right) \bigg|_0^\infty \\
= \frac{1}{\pi(1 + u^2)} , \quad -\infty < u < \infty
\]

Thus, the quotient of two iid \( N(0, 1) \) rv's is a rv that has a Cauchy distribution. \( \blacksquare \)
4.4 Order Statistics

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 an 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 \(X_3\):

\[
\begin{array}{c|ccc}
X_1 < X_2 < X_3 & 1 & 0 & 0 \\
& 0 & 1 & 0 \\
& 0 & 0 & 1 \\
\end{array}
\]

\[
\begin{array}{c|ccc}
X_1 < X_3 < X_2 & 1 & 0 & 0 \\
& 0 & 0 & 1 \\
& 0 & 1 & 0 \\
\end{array}
\]
For \( n = 3 \), there are \( 3! = 6 \) possible arrangements. 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 \( J \) that results from an \( n \times n \) identity matrix through the rearrangement or rows. Therefore, \( |J| = 1 \). By Theorem 4.3.6, we get

\[
\begin{align*}
 f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) &= n! f_{X_1, \ldots, X_n}(x_{(k_1)}, x_{(k_2)}, \ldots, x_{(k_n)}) \\
 &= n! \prod_{i=1}^{n} f_{X_i}(x_{(k_i)}) \\
 &= n! \prod_{i=1}^{n} f_X(x_i)
\end{align*}
\]

**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

\[
f_{X_{(j)}, X_{(k)}}(x_j, x_k) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} \times
\]

\[
(F_X(x_j))^{j-1}(F_X(x_k) - F_X(x_j))^{k-j-1}(1 - F_X(x_k))^{n-k} f_X(x_j)f_X(x_k)
\]

if \( x_j < x_k \) and 0 otherwise.
4.5 Multivariate Expectation

In this section, we assume that $X = (X_1, \ldots, X_n)$ is an $n$-rv and $g : \mathbb{R}^n \to \mathbb{R}^n$ is a Borel-measurable function.

**Definition 4.5.1:**
If $n = 1$, 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
$$
\int_{\mathbb{R}^n} |g(x)| \cdot f_X(x) \, dx < \infty,
$$
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 $\mathbb{R}^n \to \mathbb{R}^n$, 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:
$$
Var(X) = \Sigma_X = E((X - E(X))(X - E(X))^T)
$$
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
$$
E((X_i - E(X_i))(X_i - E(X_i))) = Var(X_i) = \sigma_i^2.
$$
Joint higher-order moments can be defined similarly when needed.
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:

$$E(|S|) = \int_{\mathbb{R}^n} \left| \sum_{i=1}^{n} a_i x_i \right| f_{\mathcal{X}}(x) \, dx$$

$$\leq \int_{\mathbb{R}^n} \sum_{i=1}^{n} |a_i| \cdot |x_i| \cdot f_{\mathcal{X}}(x) \, dx$$

$$= \sum_{i=1}^{n} |a_i| \int_{\mathbb{R}} |x_i| \left( \int_{\mathbb{R}^{n-1}} f_{\mathcal{X}}(x) \, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n \right) \, dx_i$$

$$= \sum_{i=1}^{n} |a_i| \int_{\mathbb{R}} |x_i| f_{X_i}(x_i) \, dx_i$$

$$= \sum_{i=1}^{n} |a_i| E(|X_i|)$$

$$< \infty$$

It follows that $E(S) = \sum_{i=1}^{n} a_i E(X_i)$ by the same argument without using the absolute values $| \cdot |.$

**Note:**
If $X_i, i = 1, \ldots, n,$ are iid with $E(X_i) = \mu,$ then

$$E(\overline{X}) = E\left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \frac{1}{n} E(X_i) = \mu.$$
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,

\[
E(\prod_{i=1}^{n} g_i(x_i)) = \int_{\mathbb{R}^n} \prod_{i=1}^{n} g_i(x_i) f_X(x) \, dx
\]

\[
= \int_{\mathbb{R}^n} \prod_{i=1}^{n} (g_i(x_i) f_{X_i}(x_i)) \, dx_i
\]

\[
= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{n} g_i(x_i) \prod_{i=1}^{n} f_{X_i}(x_i) \prod_{i=1}^{n} dx_i
\]

\[
= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} g_1(x_1) f_{X_1}(x_1) \, dx_1 \int_{\mathbb{R}} g_2(x_2) f_{X_2}(x_2) \, dx_2 \cdots \int_{\mathbb{R}} g_n(x_n) f_{X_n}(x_n) \, dx_n
\]

\[
= \prod_{i=1}^{n} \int_{\mathbb{R}} g_i(x_i) f_{X_i}(x_i) \, dx_i
\]

\[
= \prod_{i=1}^{n} E(g_i(X_i))
\]

Corollary 4.5.4:
If \( X, Y \) are independent, then \( \text{Cov}(X, Y) = 0 \).

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

\[
E(g_1(X)g_2(Y)) = E(g_1(X)) \cdot E(g_2(Y)).
\]
“$\iff$”: From Theorem 4.2.6, we know that $X$ and $Y$ are independent iff $P(X \in A_1, Y \in A_2) = P(X \in A_1) \cdot P(Y \in A_2)$ $\forall$ Borel sets $A_1$ 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:

$$g_1(x) = I_{A_1}(x) = \begin{cases} 1, & x \in A_1 \\ 0, & \text{otherwise} \end{cases}$$

$$g_2(y) = I_{A_2}(y) = \begin{cases} 1, & y \in A_2 \\ 0, & \text{otherwise} \end{cases}$$

Then,

$$E(g_1(X)) = 0 \cdot P(X \in A_1^c) + 1 \cdot P(X \in A_1) = P(X \in A_1),$$

$$E(g_2(Y)) = 0 \cdot P(Y \in A_2^c) + 1 \cdot P(Y \in A_2) = P(Y \in A_2)$$

and

$$E(g_1(X) \cdot g_2(Y)) = P(X \in A_1, Y \in A_2).$$

$$\implies P(X \in A_1, Y \in A_2) = E(g_1(X) \cdot g_2(Y)) \iff E(g_1(X)) \cdot E(g_2(Y)) = P(X \in A_1) \cdot P(Y \in A_2)$$

$$\implies P(X \in A_1, Y \in A_2) = P(X \in A_1) \cdot P(Y \in A_2)$$

$$\implies X, Y \text{ independent by Theorem 4.2.6.}$$

**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, we get

$$\mu_{0 \ldots 0 \uparrow 1 \ldots 0 \uparrow 1 \ldots 0} = \mu_{rs} = Cov(X_r, X_s).$$
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:

- $0 \leq (a - b)^2 = a^2 - 2ab + b^2 \implies ab \leq \frac{a^2 + b^2}{2}$
- $0 \leq (a + b)^2 = a^2 + 2ab + b^2 \implies -ab \leq \frac{a^2 + b^2}{2}$

$\implies |ab| \leq \frac{a^2 + b^2}{2} \forall a, b \in \mathbb{R}$ (*)&

(i)

$$E(|XY|) = \int_{\mathbb{R}^2} |xy| f_{X,Y}(x,y) \, dx \, dy$$

Can be rewritten:

$$= \int_{\mathbb{R}^2} \frac{x^2 + y^2}{2} f_{X,Y}(x,y) \, dx \, dy$$

$$= \frac{1}{2} \int_{\mathbb{R}^2} x^2 f_{X,Y}(x,y) \, dy \, dx + \frac{1}{2} \int_{\mathbb{R}^2} y^2 f_{X,Y}(x,y) \, dx \, dy$$

$$= \frac{1}{2} \int_{\mathbb{R}} x^2 f_X(x) \, dx + \frac{1}{2} \int_{\mathbb{R}} y^2 f_Y(y) \, dy$$

$$= \frac{E(X^2) + E(Y^2)}{2}$$

$$< \infty$$

$\implies E(XY)$ exists

$\implies \text{Cov}(X, Y) = E(XY) - E(X)E(Y)$ exists

(ii) $0 \leq E((\alpha X + \beta Y)^2) = \alpha^2 E(X^2) + 2\alpha\beta E(XY) + \beta^2 E(Y^2) \forall \alpha, \beta \in \mathbb{R}$ (A)

If $E(X^2) = 0$, then $X$ has a degenerate 1–point Dirac 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 $\alpha = \frac{E(XY)}{E(X^2)}$, $\beta = 1$. 

97
\[ \frac{(E(XY))^2}{E(X^2)} - 2\frac{(E(XY))^2}{E(X^2)} + E(Y^2) \geq 0 \]
\[ \Rightarrow - (E(XY))^2 + E(Y^2)E(X^2) \geq 0 \]
\[ \Rightarrow (E(XY))^2 \leq E(X^2)E(Y^2) \]

(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 \( E((\alpha X + \beta Y)^2) = 0 \). This can only happen if \( P(\alpha X + \beta Y = 0) = 1 \).

Otherwise, 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:
\[ (E(XY))^2 = (E(X \cdot -\frac{\alpha X}{\beta}))^2 = (\frac{\alpha}{\beta})^2 (E(X^2))^2 = E(X^2)(\frac{\alpha}{\beta})^2 E(X^2) = E(X^2)E(Y^2) \]

\[ 4.6 \text{ Multivariate Generating Functions} \]

**Definition 4.6.1:**

Let \( \mathbf{X} = (X_1, \ldots, X_n) \) be an \( n \)-rv. We define the **multivariate moment generating function** (mmgf) of \( \mathbf{X} \) as
\[ M_{\mathbf{X}}(t) = E(e^{t\mathbf{X}}) = E\left( \exp \left( \sum_{i=1}^{n} t_i X_i \right) \right) \]
if this expectation exists for \( |t| = \sqrt{\sum_{i=1}^{n} t_i^2} < h \) for some \( h > 0 \).

**Definition 4.6.2:**

Let \( \mathbf{X} = (X_1, \ldots, X_n) \) be an \( n \)-rv. We define the **\( n \)-dimensional characteristic function** \( \Phi_{\mathbf{X}} : \mathbb{R}^n \to \mathcal{C} \) of \( \mathbf{X} \) as
\[ \Phi_{\mathbf{X}}(t) = E(e^{it\mathbf{X}}) = E\left( \exp \left( i \sum_{j=1}^{n} t_j X_j \right) \right) \]

**Note:**

(i) \( \Phi_{\mathbf{X}}(t) \) exists for any real-valued \( n \)-rv.

(ii) If \( M_{\mathbf{X}}(t) \) exists, then \( \Phi_{\mathbf{X}}(t) = M_{\mathbf{X}}(it) \).
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 \( \Phi_{X_i}(t_i) = \Phi_X(0, t_i, 0) \).

(iii) Joint moments of all orders (if they exist) can be obtained as
\[
m_{i_1 \ldots i_n} = \left. \frac{\partial^{i_1+i_2+\ldots+i_n}}{\partial t_1^{i_1} \partial t_2^{i_2} \ldots \partial t_n^{i_n}} M_X(t) \right|_{t=0} = E(X_1^{i_1} X_2^{i_2} \ldots X_n^{i_n})
\]
if the mmgf exists and
\[
m_{i_1 \ldots i_n} = \frac{1}{i_1! i_2! \ldots i_n!} \left. \frac{\partial^{i_1+i_2+\ldots+i_n}}{\partial t_1^{i_1} \partial t_2^{i_2} \ldots \partial t_n^{i_n}} \Phi_X(t) \right|_{t=0} = E(X_1^{i_1} X_2^{i_2} \ldots X_n^{i_n}).
\]

(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
\[
M_Y(t) = \prod_{i=1}^{n} M_{X_i}(a_i t)
\]
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
\[
\Phi_Y(t) = \prod_{j=1}^{n} \Phi_{X_j}(a_j t)
\]
(iii) If mgf's $M_{X_1}(t), \ldots, M_{X_n}(t)$ exist, then the mmgf of $\mathbf{X}$ is

$$M_{\mathbf{X}}(t) = \prod_{i=1}^{n} M_{X_i}(t_i)$$

on the common interval where all individual mgf's exist.

(iv) The n-dimensional characteristic function of $\mathbf{X}$ is

$$\Phi_{\mathbf{X}}(t) = \prod_{j=1}^{n} \Phi_{X_j}(t_j).$$

**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:

$$G_{X_i}(s) = E(s^{X_i})$$

$$G_Y(s) = E(s^Y)$$

$$= E(s^{\sum_{i=1}^{n} X_i})$$

$$\overset{\text{indep.}}{=} \prod_{i=1}^{n} E(s^{X_i})$$

$$= \prod_{i=1}^{n} G_{X_i}(s)$$

Version 2: (case $n = 2$ only)

$$G_Y(s) = P(Y = 0) + P(Y = 1)s + P(Y = 2)s^2 + \ldots$$

$$= P(X_1 = 0, X_2 = 0) +$$

$$P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1)) s +$$

$$P(X_1 = 2, X_2 = 0) + P(X_1 = 1, X_2 = 1) + P(X_1 = 0, X_2 = 2)) s^2 + \ldots$$

$$\overset{\text{indep.}}{=} P(X_1 = 0)P(X_2 = 0) +$$

$$P(X_1 = 1)P(X_2 = 0) + P(X_1 = 0)P(X_2 = 1)) s +$$

$$P(X_1 = 2)P(X_2 = 0) + P(X_1 = 1)P(X_2 = 1) + P(X_1 = 0)P(X_2 = 2)) s^2 + \ldots$$

$$= \left( P(X_1 = 0) + P(X_1 = 1)s + P(X_1 = 2)s^2 + \ldots \right) \cdot$$

$$\left( P(X_2 = 0) + P(X_2 = 1)s + P(X_2 = 2)s^2 + \ldots \right)$$

$$= G_{X_1}(s) \cdot G_{X_2}(s)$$

101
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:**

\[
P(S_N = k) = \sum_{n=0}^\infty P(S_N = k|N = n) \cdot P(N = n)
\]

\[
\implies G_{S_N}(s) = \sum_{k=0}^\infty \sum_{n=0}^\infty P(S_N = k|N = n) \cdot P(N = n) \cdot s^k
\]

\[
= \sum_{n=0}^\infty P(N = n) \sum_{k=0}^\infty P(S_N = k|N = n) \cdot s^k
\]

\[
= \sum_{n=0}^\infty P(N = n) \sum_{k=0}^\infty P(\sum_{i=1}^n X_i = k) \cdot s^k
\]

From Th. 4.6.5

\[
\sum_{n=0}^\infty P(N = n) \prod_{i=1}^n G_{X_i}(s)
\]

\[
\overset{\text{iid}}{=} \sum_{n=0}^\infty P(N = n) \cdot (G_X(s))^n
\]

\[
= G_N(G_X(s))
\]

**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?

\[
G_X(s) = G_N(s) = ps^2 + qs + r
\]

\[
G_{S_N}(s) = p(ps^2 + ps + r)^2 + q(ps^2 + ps + r) + r
\]
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
\[
\begin{align*}
G_X(S) &= E(S^X) = E(e^{S^X}) = E(e^{\ln S^X}) = E(e^{\ln S}\cdot X) = M_X(\ln S) \\
\implies M_X(S) &= G_X(e^S) \\
\implies M_{S_N}(t) &= G_{S_N}(e^t) = M_N(G_X(e^t)) = M_N(M_X(t)) = M_N(\ln M_X(t))
\end{align*}
\]

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

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 $f_Y(y) > 0$, $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, and $f_{X,Y}$ and $f_Y$ are continuous, then $f_{X|Y}(x \mid y)$ 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. Assume that $E(h(X))$ exists. Then the **conditional expectation** of $h(X)$ given $Y$, i.e., $E(h(X) \mid Y)$, is a rv that takes the value $E(h(X) \mid y)$. It is defined as

$$E(h(X) \mid y) = \begin{cases} 
\sum_{x \in \mathcal{A}} h(x)P(X = x \mid Y = y), & \text{if } (X, Y) \text{ is discrete and } P(Y = y) > 0 \\
\int_{-\infty}^{\infty} h(x)f_{X|Y}(x \mid y)\,dx, & \text{if } (X, Y) \text{ is continuous and } f_Y(y) > 0
\end{cases}$$

**Note:**

(i) The rv $E(h(X) \mid Y) = g(Y)$ is a function of $Y$ as a rv.

(ii) The usual properties of expectations apply to the conditional expectation:

(a) $E(c \mid Y) = c \ \forall c \in \mathbb{R}$.

(b) $E(aX + b \mid Y) = aE(X \mid Y) + b \ \forall a, b \in \mathbb{R}$.

(c) If $g_1, g_2$ are Borel-measurable functions and if $E(g_1(X)), E(g_2(X))$ exist, then $E(a_1g_1(X) + a_2g_2(X) \mid Y) = a_1E(g_1(X) \mid Y) + a_2E(g_2(X) \mid Y) \ \forall a_1, a_2 \in \mathbb{R}$.

(d) If $X \geq 0$ then $E(X \mid Y) \geq 0$.

(e) If $X_1 \geq X_2$ then $E(X_1 \mid Y) \geq E(X_2 \mid Y)$.

(iii) Moments are defined in the usual way. If $E(|X|) < \infty$, then $E(X^r \mid Y)$ exists and is the $r^{th}$ conditional moment of $X$ given $Y$. 

104
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:

\[ f_{Y \mid X}(y \mid x) = \frac{1}{1 - x} \text{ for } x < y < 1 \text{ (where } 0 < x < 1 \text{)} \]

and

\[ f_{X \mid Y}(x \mid y) = \frac{1}{y} \text{ for } 0 < x < y \text{ (where } 0 < y < 1 \text{)} \]

So,

\[ E(X \mid y) = \int_0^y \frac{x}{y} \, dx = \frac{y}{2} \]

and

\[ E(Y \mid x) = \int_x^1 \frac{1}{1 - x} \, y \, dy = \frac{1}{y^2} \left[ \frac{y^2}{1 - x} \right]_x^1 = \frac{1}{2} \frac{1}{1 - x} = \frac{1 + x}{2}. \]

Therefore, we get the rv's \( E(X \mid Y) = \frac{Y}{2} \) and \( E(Y \mid X) = \frac{1 + X}{2} \). \( \square \)

**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:

\[
E_Y(E(h(X) \mid Y)) = \int_{-\infty}^{\infty} E(h(x) \mid y) f_Y(y) \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x) f_{X \mid Y}(x \mid y) f_Y(y) \, dx \, dy = \int_{-\infty}^{\infty} h(x) \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy \, dx = \int_{-\infty}^{\infty} h(x) f_X(x) \, dx = E(h(X))
\]

\( \square \)

**Theorem 4.7.4:**
If \( E(X^2) \) exists, then

\[ Var_Y(E(X \mid Y)) + E_Y(Var(X \mid Y)) = Var(X). \]
Proof:

\[ Var_Y(E(X \mid Y)) + E_Y(Var(X \mid Y)) = E_Y((E(X \mid Y))^2) - (E_Y(E(X \mid Y)))^2 \]
\[ + E_Y(E(X^2 \mid Y) - (E(X \mid Y))^2) \]
\[ = E_Y((E(X \mid Y))^2 - (E(X))^2 + E(X^2) - E_Y((E(X \mid Y))^2)) \]
\[ = E(X^2) - (E(X))^2 \]
\[ = Var(X) \]

Note:
If \( E(X^2) \) exists, then \( Var(X) \geq Var_Y(E(X \mid Y)) \). \( Var(X) = Var_Y(E(X \mid Y)) \) iff \( X = g(Y) \). The inequality directly follows from Theorem 4.7.4.

For equality, it is necessary that \( E_Y(Var(X \mid Y)) = E_Y((X - E(X \mid Y))^2 \mid Y) = 0 \) 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 \( E(h(X) \mid Y) = E(h(X)) \).

4.8 Inequalities and Identities

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
\[ g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab \]
\[ \Rightarrow g'(a) = a^{p-1} - b \]
\[ = 0 \]
\[ \Rightarrow b = a^{p-1} \]
\[ \Rightarrow b^q = a^{(p-1)q} = a^p \]
\[ g''(a) = (p - 1)a^{p-2} > 0 \]
Since $g''(a) > 0$, this is really a minimum. The minimum value is obtained for $b = a^{p-1}$ and it is
\[
\frac{1}{p} a^p + \frac{1}{q} (a^{p-1})^q - a a^{p-1} = \frac{1}{p} a^p + \frac{1}{q} a^p - a a^p = a^p \left( \frac{1}{p} + \frac{1}{q} - 1 \right) = 0.
\]
Since $g''(a) > 0$, the minimum is unique and $g(a) \geq 0$. Therefore $g(a) + ab = \frac{1}{p} a^p + \frac{1}{q} b^q \geq ab$. ■

**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 $a = \frac{|X|}{(E(|X|)^p)^{\frac{1}{p}}}$ and $b = \frac{|Y|}{(E(|Y|)^q)^{\frac{1}{q}}}$. Then
\[
\frac{1}{p} E(|X|^p) + \frac{1}{q} E(|Y|^q) \geq \frac{|XY|}{(E(|X|)^p)^{\frac{1}{p}} (E(|Y|)^q)^{\frac{1}{q}}}
\]

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 $p = q = 2$. ■
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:

$$E(|X + Y|^p) = E(|X + Y| \cdot |X + Y|^{p-1}) \leq E(|X| + |Y| \cdot |X + Y|^{p-1}) = E(|X| \cdot |X + Y|^{p-1}) + E(|Y| \cdot |X + Y|^{p-1})$$

$$\leq (E(|X|^p))^{\frac{1}{p}} \cdot (E(|X + Y|^{p-1})^{\frac{1}{q}}) + (E(|Y|^p))^{\frac{1}{p}} \cdot (E(|X + Y|^{p-1})^{\frac{1}{q}}) \cdot (E(|X + Y|^p))^{\frac{1}{q}} \quad (A)$$

Divide the left and right side of this inequality by $(E(|X + Y|^p))^{\frac{1}{p}}$.

The result on the left side is $(E(|X + Y|^p))^{\frac{1}{p} - \frac{1}{q}} = (E(|X + Y|^q))^{\frac{1}{q}}$, 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 $q = \frac{p}{p-1}$. Therefore, $\frac{1}{p} + \frac{1}{q} = 1$ and $(p-1)q = p$.

$(*)$ holds since $\frac{1}{p} + \frac{1}{q} = 1$, a condition to meet Hölder’s Inequality.

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 $x_0 = E(X)$:

$$l(x) = ax + b$$

for some $a, b \in \mathbb{R}$

The function $g(x)$ is a convex function and falls above the tangent line $l(x)$

$$\implies g(x) \geq ax + b \quad \forall x \in \mathbb{R}$$

$$\implies E(g(X)) \geq E(aX + b) = aE(X) + b = \underbrace{E(E(X))}_{\text{tangent at point } g(E(X))}$$

Therefore, Theorem 4.8.5 holds.

Note:
Typical convex functions $g$ are:

(i) $g_1(x) = |x| \quad \implies E(|X|) \geq |E(X)|$

(ii) $g_2(x) = x^2 \quad \implies E(X^2) \geq (E(X))^2 \quad \implies Var(X) \geq 0$.

(iii) $g_3(x) = \frac{1}{x^p}$ for $x > 0, p > 0 \quad \implies E\left(\frac{1}{X^p}\right) \geq \frac{1}{(E(X))^p};$ for $p = 1$: $E(\frac{1}{X}) \geq \frac{1}{E(X)}$.

(iv) Other convex functions are $x^p$ for $x > 0, p \geq 1; \theta^p$ for $\theta > 1; -\ln(x)$ for $x > 0; \text{etc.}$

(v) Recall that if $g$ is convex and differentiable, then $g''(x) \geq 0 \quad \forall x$.

(vi) If the function $g$ is concave, the direction of the inequality in Jensen’s Inequality is reversed, i.e., $E(g(X)) \leq g(E(X))$.

Example 4.8.6:
Given the real numbers $a_1, a_2, \ldots, a_n > 0$, we define

\text{arithmetic mean} : \quad 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 : \( a_H = \frac{1}{\frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n}} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} \frac{1}{a_i}} \)

Let \( X \) be a rv that takes values \( a_1, a_2, \ldots, a_n > 0 \) with probability \( \frac{1}{n} \) each.

(i) \( a_A \geq a_G \):

\[
\ln(a_A) = \ln\left(\frac{1}{n} \sum_{i=1}^{n} a_i\right) \\
\geq \ln(\sum_{i=1}^{n} \frac{1}{n} a_i) \\
= \frac{1}{n} \sum_{i=1}^{n} \ln(a_i) \\
= \frac{1}{n} \ln(\prod_{i=1}^{n} a_i) \\
= \ln((\prod_{i=1}^{n} a_i)^{\frac{1}{n}}) \\
= \ln(a_G)
\]

Taking the anti–log of both sides gives \( a_A \geq a_G \).

(ii) \( a_A \geq a_H \):

\[
\frac{1}{a_A} = \frac{1}{\frac{1}{n} \sum_{i=1}^{n} a_i} \\
= \frac{1}{E(X)} \\
\leq E\left(\frac{1}{X}\right) \\
= \sum_{i=1}^{n} \frac{1}{n} \frac{1}{a_i} \\
= \frac{1}{n} \left( \frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n}\right) \\
= \frac{1}{a_H}
\]

Inverting both sides gives \( a_A \geq a_H \).
(iii) $a_G \geq a_H$:

$\ln(a_H) = \ln(a_G^{-1})$

$= \ln\left(\frac{1}{a_H}\right)$

$= \ln\left(\frac{1}{n} \left(\frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n}\right)\right)$

$= \ln(E\left(\frac{1}{X}\right))$

$\ln \text{ concave} \geq E(\ln\left(\frac{1}{X}\right))$

$= \sum_{i=1}^{n} \frac{1}{n} \ln\left(\frac{1}{a_i}\right)$

$= \frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{1}{a_i}\right)$

$= \frac{1}{n} \sum_{i=1}^{n} -\ln a_i$

$= -\frac{1}{n} \ln(\prod_{i=1}^{n} a_i)$

$= -\ln\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}}$

$= -\ln a_G$

Multiplying both sides with $-1$ gives $\ln a_H \leq \ln a_G$. Then taking the anti-log of both sides gives $a_H \leq a_G$.

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
5 Particular Distributions

5.1 Multivariate Normal Distributions

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 $\text{Var}(X) = \sigma^2$. If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ and $c_1, c_2 \in \mathbb{R}$, then $Y = c_1 X_1 + c_2 X_2 \sim N(c_1 \mu_1 + c_2 \mu_2, c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2)$.  

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
\[ X = \mu_1 + a_{11} Z_1 + a_{12} Z_2, \quad Y = \mu_2 + a_{21} Z_1 + a_{22} Z_2. \]

If we define
\[ A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad X = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, \]

then we can write
\[ X = AZ + \mu. \]

Note:
$E(X) = \mu_1 + a_{11}E(Z_1) + a_{12}E(Z_2) = \mu_1$ 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 $a_{11} = a_{12} = 0$.  

113
Theorem 5.1.3:
Define $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ as $g(\mathbf{x}) = C\mathbf{x} + \mathbf{d}$. If $\mathbf{X}$ is a bivariate Normal rv, then $g(\mathbf{X})$ also is a bivariate Normal rv.

**Proof:**
\[
g(\mathbf{X}) = C\mathbf{X} + \mathbf{d} = C(\mathbf{AZ} + \mu) + \mathbf{d} = \begin{pmatrix} C\mathbf{A} \\ \mu \end{pmatrix} \mathbf{Z} + \begin{pmatrix} C\mu + \mathbf{d} \\ \mathbf{d} \end{pmatrix}
\]
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
\[
\Sigma = AA' = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} = \begin{pmatrix} a_{11}^2 + a_{12}^2 & a_{11}a_{21} + a_{12}a_{22} \\ a_{11}a_{21} + a_{12}a_{22} & a_{21}^2 + a_{22}^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}.
\]

**Theorem 5.1.5:**
Assume that $\sigma_1 > 0, \sigma_2 > 0$ and $|\rho| < 1$. Then the joint pdf of $\mathbf{X} = (X, Y) = AZ + \mu$ (as defined in Definition 5.1.2) is
\[
f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi \sqrt{|\Sigma|}} \exp \left( -\frac{1}{2}(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu) \right)
\]
\[
= \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp \left( -\frac{1}{2(1 - \rho^2)} \left( \frac{x - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x - \mu_1}{\sigma_1} \right) \left( \frac{y - \mu_2}{\sigma_2} \right) + \left( \frac{y - \mu_2}{\sigma_2} \right)^2 \right)
\]

**Proof:**
The mapping $\mathbf{Z} \rightarrow \mathbf{X}$ is 1-to-1:
\[
\mathbf{X} = AZ + \mu
\]

114
\[ Z = A^{-1}(x - \mu) \quad \text{(requirement is that } A \text{ is invertible.}) \]
\[ J = |A^{-1}| = \frac{1}{|A|} \]
\[ |A| = \sqrt{|A|^2} = \sqrt{|A| \cdot |A^T|} = \sqrt{|AA^T|} = \sqrt{|\Sigma|} = \sqrt{\sigma_1^2 \sigma_2^2 - \rho^2 \sigma_1^2 \sigma_2^2} = \sigma_1 \sigma_2 \sqrt{1 - \rho^2} \]

We can use this result to get to the second line of the theorem:

\[ f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \]
\[ = \frac{1}{2\pi} e^{-\frac{1}{2} \langle z^T z \rangle} \]

As already stated, the mapping from \( Z \) to \( X \) is 1-to-1, so we can apply Theorem 4.3.5:

\[ f_X(x) = \frac{1}{2\pi \sqrt{|\Sigma|}} \exp(-\frac{1}{2} \langle x - \mu \rangle^T \Sigma^{-1} \langle x - \mu \rangle) \]
\[ \quad \left( \ast \right) \]
\[ \frac{1}{2\pi \sqrt{|\Sigma|}} \exp(-\frac{1}{2} \langle x - \mu \rangle^T \Sigma^{-1} \langle x - \mu \rangle) \]

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 following transformations:

\[ \sqrt{|\Sigma|} = \sigma_1 \sigma_2 \sqrt{1 - \rho^2} \]
\[ \Sigma^{-1} = \frac{1}{|\Sigma|} \begin{pmatrix} \sigma_2^2 & -\rho \sigma_1 \sigma_2 \\ -\rho \sigma_1 \sigma_2 & \sigma_1^2 \end{pmatrix} \]
\[ = \begin{pmatrix} \frac{1}{\sigma_1^2 (1 - \rho^2)} & -\rho \\ -\rho & \frac{1}{\sigma_2 (1 - \rho^2)} \end{pmatrix} \]

\[ \Rightarrow f_X(x) = \frac{1}{2\pi \sigma_1 \sigma_2 \sqrt{1 - \rho^2}} \exp \left( -\frac{1}{2(1 - \rho^2)} \left( \frac{x - \mu_1}{\sigma_1} \right)^2 - 2 \rho \left( \frac{x - \mu_1}{\sigma_1} \right) \left( \frac{y - \mu_2}{\sigma_2} \right) + \left( \frac{y - \mu_2}{\sigma_2} \right)^2 \right) \]

**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

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.
Theorem 5.1.8:
The mgf $M_X(t)$ of a bivariate Normal rv $X = (X, Y)$ is

$$M_X(t) = M_{X,Y}(t_1, t_2) = \exp(\mu t + \frac{1}{2} t^T \Sigma t) = \exp \left( \mu_1 t_1 + \mu_2 t_2 + \frac{1}{2} \left( \sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2 \rho \sigma_1 \sigma_2 t_1 t_2 \right) \right).$$

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 $X = (X, Y)$:

$$M_X(t) = E(\exp(tX))$$

$$= \int_{-\infty}^{\infty} \exp(tx) \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2}(x - \mu)^2 \right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} -2\sigma^2 tx + (x - \mu)^2 \right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} -2\sigma^2 tx + (x^2 - 2\mu x + \mu^2) \right) dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x^2 - 2(\mu + \sigma^2 t)x + (\mu + t\sigma^2)^2 - (\mu + t\sigma^2)^2 + \mu^2) \right) dx$$

$$= \exp \left( -\frac{1}{2\sigma^2} [(\mu + t\sigma^2)^2 + \mu^2] \right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} [x - (\mu + t\sigma^2)]^2 \right) dx$$

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:

$$M_{X,Y}(t_1, t_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(t_1 x + t_2 y) f_{X,Y}(x, y) \, dx \, dy$$
\[
\begin{align*}
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(t_1 x) \exp(t_2 y) f_X(x) f_{Y \mid X}(y \mid x) \, dy \, dx \\
&= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} \exp(t_2 y) f_{Y \mid X}(y \mid x) \, dy \right) \exp(t_1 x) f_X(x) \, dx \\
&= \int_{-\infty}^{\infty} \left( \int_{-\infty}^{\infty} \frac{\exp(t_2 y)}{\sigma_2 \sqrt{1 - \rho^2} \sqrt{2\pi}} \exp\left( \frac{-(y - \beta X)^2}{2\sigma_2^2(1 - \rho^2)} \right) \, dy \right) \exp(t_1 x) f_X(x) \, dx \\
&\quad \mid \beta_X = \mu_2 + \frac{\sigma_2}{\sigma_1} (x - \mu_1)
\end{align*}
\]

\[ \int_{-\infty}^{\infty} \exp\left( \beta_X t_2 + \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (1 - \rho^2) t_2^2 \right) \exp(t_1 x) f_X(x) \, dx \]

\[ \int_{-\infty}^{\infty} \exp\left( \mu_2 t_2 + \frac{\sigma_2}{\sigma_1} (x - \mu_1) t_2 + \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (1 - \rho^2) t_2^2 + t_1 x \right) f_X(x) \, dx \]

\[ \int_{-\infty}^{\infty} \exp\left( \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (1 - \rho^2) t_2^2 + t_2 \mu_2 - \frac{\sigma_2}{\sigma_1} \mu_1 t_2 + \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (1 - \rho^2) t_2^2 + t_1 x \right) f_X(x) \, dx \]

\[ \exp\left( \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (1 - \rho^2) t_2^2 + t_2 \mu_2 - \frac{\sigma_2}{\sigma_1} \mu_1 t_2 \right) \int_{-\infty}^{\infty} \exp\left( t_1 + \frac{\sigma_2}{\sigma_1} t_2 x \right) f_X(x) \, dx \]

\[ \exp\left( \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (1 - \rho^2) t_2^2 + t_2 \mu_2 - \frac{\sigma_2}{\sigma_1} \mu_1 t_2 \right) \cdot \exp\left( \mu_1 (t_1 + \frac{\sigma_2}{\sigma_1} t_2) + \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (t_1 + \frac{\sigma_2}{\sigma_1} t_2)^2 \right) \]

\[ \exp\left( \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} (1 - \rho^2) t_2^2 - \frac{1}{2} \rho \frac{\sigma_2^2}{\sigma_1^2} t_2^2 + \mu_2 t_2 - \mu_1 \frac{\sigma_2}{\sigma_1} t_2 + \mu_1 t_1 + \mu_1 \frac{\sigma_2}{\sigma_1} t_2 + \frac{1}{2} \frac{\sigma_2^2}{\sigma_1^2} t_2^2 + \rho \frac{\sigma_1 \sigma_2}{\sigma_1^2} t_1 t_2 + \frac{1}{2} \rho^2 \frac{\sigma_2^2}{\sigma_1^2} t_2^2 \right) \]

\[ \exp\left( \mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_2 t_2^2 + \sigma_2 t_2^2 + 2 \rho \sigma_1 \sigma_2 t_1 t_2}{2} \right) \]

(A) follows from Theorem 5.1.6 since \(Y \mid X \sim N(\beta_X, \sigma_2^2(1 - \rho^2))\). (B) follows when we apply our calculations of the mgf of a \(N(\mu, \sigma^2)\) distribution to a \(N(\beta_X, \sigma_2^2(1 - \rho^2))\) distribution. (C) holds since the integral represents \(M_X(t_1 + \rho \frac{\sigma_2}{\sigma_1} t_2)\).

**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 \(\Sigma = AA'\).

118
Note:

(i) If $\Sigma$ is non-singular, $X$ has the joint pdf

\[ f_X(x) = \frac{1}{(2\pi)^{k/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right). \]

(ii) If $\Sigma$ is singular, then $X - \mu$ takes values in a linear subspace of $\mathbb{R}^k$ with probability 1.

(iii) If $\Sigma$ is non-singular, then $X$ has mgf

\[ M_X(t) = \exp(\mu' t + \frac{1}{2} t' \Sigma t). \]

---

Theorem 5.1.11:
The components $X_1, \ldots, X_k$ of a normally distributed $k$-rv $X$ are independent iff $Cov(X_i, X_j) = 0$ \forall i, j = 1, \ldots, k, i \neq j.

---

Theorem 5.1.12:
Let $X = (X_1, \ldots, X_k)'$. $X$ has a $k$-dimensional Normal distribution iff every linear function of $X$, i.e., $X' t = t_1 X_1 + t_2 X_2 + \ldots + t_k X_k$, 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 $X$, i.e., $X' t$, has a univariate Normal distribution.

Conversely, if $X' t$ has a univariate Normal distribution, we know from Theorem 5.1.8 that

\[ M_{X' t}(s) = \exp \left( E(X' t) \cdot s + \frac{1}{2} Var(X' t) \cdot s^2 \right) \]

\[ = \exp \left( \mu' t s + \frac{1}{2} t' \Sigma t s^2 \right) \]

\[ \Rightarrow M_{X' t}(1) = \exp \left( \mu' t + \frac{1}{2} t' \Sigma t \right) \]

\[ = M_X(t) \]

By uniqueness of the mgf and Note (iii) that follows Definition 5.1.10, $X$ has a multivariate Normal distribution.

---

119
5.2 Exponential Family of Distributions

Definition 5.2.1:
Let $\theta$ 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 $\eta = (\eta_1, \ldots, \eta_k)'$.

Example 5.2.3:
Let $X \sim N(\mu, \sigma^2)$ with both parameters $\mu$ and $\sigma^2$ unknown. We have:

$$f(x; \theta) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) = \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu^2}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right)$$

$\theta = (\mu, \sigma^2)$

$\vartheta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$
Therefore,

\[
\begin{align*}
Q_1(\bar{y}) &= -\frac{1}{2\sigma^2} \\
T_1(x) &= x^2 \\
Q_2(\bar{y}) &= \frac{\mu}{\sigma^2} \\
T_2(x) &= x \\
D(\bar{y}) &= -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \\
S(x) &= 0
\end{align*}
\]

Thus, this is a 2–parameter exponential family.
6 Limit Theorems

Motivation:
I found this slide from my Stat 250, Section 003, “Introductory Statistics” class (an undergraduate class I taught at George Mason University in Spring 1999):

What does this mean at a more theoretical level???
6.1 Modes of Convergence

Definition 6.1.1:
Let $X_1, \ldots, X_n$ be iid rv’s with common cdf $F_X(x)$. Let $T = T(X)$ be any statistic, i.e., a Borel-measurable function of $X$ that does not involve the population parameter(s) $\theta$, defined on the support $\mathcal{X}$ of $X$. The induced probability distribution of $T(X)$ is called the sampling distribution of $T(X)$.

Note:
(i) Commonly used statistics are:
Sample Mean: $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$
Sample Variance: $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$
Sample Median, Order Statistics, Min, Max, etc.

(ii) Recall that if $X_1, \ldots, X_n$ are iid and if $E(X)$ and $Var(X)$ exist, then $E(\overline{X}_n) = \mu = E(X)$, $E(S_n^2) = \sigma^2 = Var(X)$, and $Var(\overline{X}_n) = \frac{\sigma^2}{n}$.

(iii) Recall that if $X_1, \ldots, X_n$ are iid and if $X$ has mgf $M_X(t)$ or characteristic function $\Phi_X(t)$ then $M_{\overline{X}_n}(t) = (M_X(\frac{t}{n}))^n$ or $\Phi_{\overline{X}_n}(t) = (\Phi_X(\frac{t}{n}))^n$.

Note: Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of rv’s on some probability space $(\Omega, L, P)$. Is there any meaning behind the expression $\lim_{n \to \infty} X_n = X$? Not immediately under the usual definitions of limits. We first need to define modes of convergence for rv’s and probabilities.

Definition 6.1.2:
Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of rv’s with cdf’s $\{F_n\}_{n=1}^{\infty}$ and let $X$ be a rv with cdf $F$. If $F_n(x) \to F(x)$ at all continuity points of $F$, we say that $X_n$ converges in distribution to $X$ ($X_n \xrightarrow{d} X$) or $X_n$ converges in law to $X$ ($X_n \xrightarrow{L} X$), or $F_n$ converges weakly to $F$ ($F_n \xrightarrow{w} F$).

Example 6.1.3:
Let $X_n \sim N(0, \frac{1}{n})$. Then
$$F_n(x) = \int_{-\infty}^{x} \frac{\exp \left( -\frac{1}{2} n t^2 \right)}{\sqrt{\frac{2\pi}{n}}} \, dt$$

123
\[
\int_{-\infty}^{\sqrt{n}x} \frac{\exp(-\frac{1}{2}s^2)}{\sqrt{2\pi}} \, ds = \phi(\sqrt{n}x)
\]

\[
\implies F_n(x) \to \begin{cases} 
\phi(\infty) = 1, & \text{if } x > 0 \\
\phi(0) = \frac{1}{2}, & \text{if } x = 0 \\
\phi(-\infty) = 0, & \text{if } x < 0 
\end{cases}
\]

If \( F_X(x) = \begin{cases} 
1, & x \geq 0 \\
0, & x < 0
\end{cases} \) the only point of discontinuity is at \( x = 0 \). Everywhere else, \( \phi(\sqrt{n}x) = F_n(x) \to F_X(x) \).

So, \( X_n \xrightarrow{d} X \), where \( P(X = 0) = 1 \), or \( X_n \xrightarrow{d} 0 \) since the limiting rv here is degenerate, i.e., it has a Dirac(0) distribution.

Example 6.1.4:
In this example, the sequence \( \{F_n\}_{n=1}^{\infty} \) converges pointwise to something that is not a cdf:

Let \( X_n \sim \text{Dirac}(n) \), i.e., \( P(X_n = n) = 1 \). Then,

\[
F_n(x) = \begin{cases} 
0, & x < n \\
1, & x \geq n
\end{cases}
\]

It is \( F_n(x) \to 0 \quad \forall x \) which is not a cdf. Thus, there is no rv \( X \) such that \( X_n \xrightarrow{d} X \).

Example 6.1.5:
Let \( \{X_n\}_{n=1}^{\infty} \) be a sequence of rv’s such that \( P(X_n = 0) = 1 - \frac{1}{n} \) and \( P(X_n = n) = \frac{1}{n} \) and let \( X \sim \text{Dirac}(0) \), i.e., \( P(X = 0) = 1 \).

It is

\[
F_n(x) = \begin{cases} 
0, & x < 0 \\
1 - \frac{1}{n}, & 0 \leq x < n \\
1, & x \geq n
\end{cases}
\]

\[
F_X(x) = \begin{cases} 
0, & x < 0 \\
1, & x \geq 0
\end{cases}
\]

It holds that \( F_n \xrightarrow{w} F_X \) but

\[
E(X_n^k) = n^{k-1} \not\to E(X^k) = 0.
\]

Thus, convergence in distribution does not imply convergence of moments/means.
Note:
Convergence in distribution does not say that the $X_i$'s are close to each other or to $X$. It only
means that their cdf's are (eventually) close to some cdf $F$. The $X_i$'s do not even have to be defined
on the same probability space.

Example 6.1.6:
Let $X$ and $\{X_n\}_{n=1}^{\infty}$ be iid $N(0,1)$. Obviously, $X_n \xrightarrow{d} X$ but $\lim_{n \to \infty} X_n \neq X$.

Theorem 6.1.7:
Let $X$ and $\{X_n\}_{n=1}^{\infty}$ be discrete rv's with support $\mathcal{X}$ and $\{X'_n\}_{n=1}^{\infty}$, respectively. Define the count-
able set $A = \mathcal{X} \cup \bigcup_{n=1}^{\infty} X'_n = \{a_k : k = 1, 2, 3, \ldots\}$. Let $p_k = P(X = a_k)$ and $p_{nk} = P(X_n = a_k)$. Then
it holds that $p_{nk} \to p_k \ \forall k$ iff $X_n \xrightarrow{d} X$.

Theorem 6.1.8:
Let $X$ and $\{X_n\}_{n=1}^{\infty}$ be continuous rv's with pdf's $f$ and $\{f_n\}_{n=1}^{\infty}$, respectively. If $f_n(x) \to f(x)$ for
almost all $x$ as $n \to \infty$ then $X_n \xrightarrow{d} X$.

Theorem 6.1.9:
Let $X$ and $\{X_n\}_{n=1}^{\infty}$ be rv's such that $X_n \xrightarrow{d} X$. Let $c \in \mathbb{R}$ be a constant. Then it holds:

(i) $X_n + c \xrightarrow{d} X + c$.

(ii) $cX_n \xrightarrow{d} cX$.

(iii) If $a_n \to a$ and $b_n \to b$, then $a_n X_n + b_n \xrightarrow{d} aX + b$.

Proof:
Part (iii):
Suppose that $a > 0, a_n > 0$. Let $Y_n = a_n X_n + b_n$ and $Y = aX + b$. It is
\[ F_Y(y) = P(Y < y) = P(aX + b < y) = P(X < \frac{y-b}{a}) = F_X \left( \frac{y-b}{a} \right). \]
Likewise,
\[ F_{Y_n}(y) = F_{X_n} \left( \frac{y-b_n}{a_n} \right). \]
If $y$ is a continuity point of $F_Y$, $\frac{y-b}{a}$ is a continuity point of $F_X$. Since $a_n \to a, b_n \to b$ and
$F_{X_n}(x) \to F_X(x)$, it follows that $F_{Y_n}(y) \to F_Y(y)$ for every continuity point $y$ of $F_Y$. Thus,
$X_n \xrightarrow{d} X + b$. 

Definition 6.1.10:
Let \( \{X_n\}_{n=1}^{\infty} \) be a sequence of rv’s defined on a probability space \((\Omega, L, P)\). We say that \(X_n\) converges in probability to a rv \(X\) (\(X_n \xrightarrow{p} X\), \(\text{P}\lim_{n \to \infty} X_n = X\)) if
\[
\lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0 \quad \forall \varepsilon > 0.
\]

Note:
The following are equivalent:
\[
\lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0 \iff \lim_{n \to \infty} P(|X_n - X| \leq \varepsilon) = 1 \iff \lim_{n \to \infty} P(\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}) = 0.
\]

If \(X\) is degenerate, i.e., \(P(X = c) = 1\), we say that \(X_n\) is consistent for \(c\). For example, let \(X_n\) such that \(P(X_n = 0) = 1 - \frac{1}{n}\) and \(P(X_n = 1) = \frac{1}{n}\). Then
\[
P(|X_n| > \varepsilon) = \begin{cases} 
\frac{1}{n}, & 0 < \varepsilon < 1 \\
0, & \varepsilon \geq 1
\end{cases}
\]
Therefore, \(\lim_{n \to \infty} P(|X_n| > \varepsilon) = 0 \quad \forall \varepsilon > 0\). So \(X_n \xrightarrow{p} 0\), i.e., \(X_n\) is consistent for 0.

Theorem 6.1.11:
\[
\begin{align*}
\text{(i)} & \quad X_n \xrightarrow{p} X \iff X_n - X \xrightarrow{p} 0. \\
\text{(ii)} & \quad X_n \xrightarrow{p} X, X_n \xrightarrow{p} Y \implies P(X = Y) = 1. \\
\text{(iii)} & \quad X_n \xrightarrow{p} X, X_m \xrightarrow{p} X \implies X_n - X_m \xrightarrow{p} 0 \text{ as } n, m \to \infty. \\
\text{(iv)} & \quad X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_n \pm Y_n \xrightarrow{p} X \pm Y. \\
\text{(v)} & \quad X_n \xrightarrow{p} X, k \in \mathbb{R} \text{ a constant } \implies kX_n \xrightarrow{p} kX. \\
\text{(vi)} & \quad X_n \xrightarrow{p} k, k \in \mathbb{R} \text{ a constant } \implies X_n^r \xrightarrow{p} k^r \quad \forall r \in \mathbb{N}. \\
\text{(vii)} & \quad X_n \xrightarrow{p} a, Y_n \xrightarrow{p} b, a, b \in \mathbb{R} \implies X_nY_n \xrightarrow{p} ab. \\
\text{(viii)} & \quad X_n \xrightarrow{p} 1 \implies X_n^{-1} \xrightarrow{p} 1.
\end{align*}
\]
(ix) \( X_n \xrightarrow{p} a, Y_n \xrightarrow{p} b, a \in \mathbb{R}, b \in \mathbb{R} - \{0\} \implies \frac{X_n + Y_n}{n} \xrightarrow{p} \frac{a + b}{2}. \)

(x) \( X_n \xrightarrow{p} X, Y \) an arbitrary rv \( \implies X_nY \xrightarrow{p} XY. \)

(xi) \( X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \implies X_nY_n \xrightarrow{p} XY. \)

**Proof:**

See Rohatgi, page 244–245 for partial proofs.

---

**Theorem 6.1.12:**

Let \( X_n \xrightarrow{p} X \) and let \( g \) be a continuous function on \( \mathbb{R}. \) Then \( g(X_n) \xrightarrow{p} g(X). \)

**Proof:**

**Preconditions:**

1. \( X \) rv \( \implies \forall \varepsilon > 0 \ \exists k = k(\varepsilon) : P(|X| > k) < \frac{\varepsilon}{2} \)

2. \( g \) is continuous on \( \mathbb{R} \)

\( \implies \) \( g \) is also uniformly continuous (see Definition of u.c. in Theorem 3.3.3 (iii)) on \([-k, k]\)

\( \implies \exists \delta = \delta(\varepsilon, k) : |X| \leq k, |X_n - X| < \delta \implies |g(X_n) - g(X)| < \varepsilon \)

Let

\[
A = \{ |X| \leq k \} = \{ \omega : |X(\omega)| \leq k \}
\]
\[
B = \{ |X_n - X| < \delta \} = \{ \omega : |X_n(\omega) - X(\omega)| < \delta \}
\]
\[
C = \{ |g(X_n) - g(X)| < \varepsilon \} = \{ \omega : |g(X_n(\omega)) - g(X(\omega))| < \varepsilon \}
\]

If \( \omega \in A \cap B \)

\( \implies \omega \in C \)

\( \implies A \cap B \subset C \)

\( \implies C^C \subset (A \cap B)^C = A^C \cup B^C \)

\( \implies P(C^C) \leq P(A^C \cup B^C) \leq P(A^C) + P(B^C) \)

**Now:**

\[
P(|g(X_n) - g(X)| \geq \varepsilon) \leq P(|X| > k) + P(|X_n - X| \geq \delta) \leq \frac{\varepsilon}{2} \text{ by 1,} \quad \leq \frac{\varepsilon}{2} \text{ for } n \geq n_0(\varepsilon, \delta, k) \text{ since } X_n \xrightarrow{p} X
\]

\[
\leq \varepsilon \quad \text{for } n \geq n_0(\varepsilon, \delta, k)
\]
Corollary 6.1.13:
Let \( X_n \overset{p}{\to} c, c \in \mathbb{R} \) and let \( g \) be a continuous function on \( \mathbb{R} \). Then \( g(X_n) \overset{p}{\to} g(c) \). □

Theorem 6.1.14:
\( X_n \overset{p}{\to} X \iff X_n \overset{d}{\to} X \).

Proof:
\( X_n \overset{p}{\to} X \iff P(|X_n - X| > \epsilon) \to 0 \) as \( n \to \infty \) \( \forall \epsilon > 0 \)

It holds:
\[
P(X \leq x - \epsilon) = P(X \leq x - \epsilon, |X_n - X| \leq \epsilon) + P(X \leq x - \epsilon, |X_n - X| > \epsilon) \\
\leq P(X_n \leq x) + P(|X_n - X| > \epsilon)
\]

\((A)\) holds since \( X \leq x - \epsilon \) and \( X_n \) within \( \epsilon \) of \( X \), thus \( X_n \leq x \).

Similarly, it holds:
\[
P(X_n \leq x) \leq P(X \leq x + \epsilon) + P(|X_n - X| > \epsilon) \\
= P(X_n \leq x, |X_n - X| \leq \epsilon) + P(X_n \leq x, |X_n - X| > \epsilon)
\]

Combining the 2 inequalities from above gives:
\[
P(X \leq x - \epsilon) - P(|X_n - X| > \epsilon) \leq P(X_n \leq x) \leq P(X \leq x + \epsilon) - P(|X_n - X| > \epsilon)
\]

Therefore,
\[
P(X \leq x - \epsilon) \leq F_n(x) \leq P(X \leq x + \epsilon) \text{ as } n \to \infty.
\]

Since the cdf’s \( F_n(\cdot) \) are not necessarily left continuous, we get the following result for \( \epsilon \downarrow 0 \):
\[
P(X < x) \leq F_n(x) \leq P(X \leq x) = F_X(x)
\]

Let \( x \) be a continuity point of \( F \). Then it holds:
\[
F(x) = P(X < x) \leq F_n(x) \leq F(x)
\]
\[
\implies F_n(x) \to F(x)
\]
\[
\implies X_n \overset{d}{\to} X
\]

□
Theorem 6.1.15:
Let \( c \in \mathbb{R} \) be a constant. Then it holds:

\[
X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c.
\]

Example 6.1.16:
In this example, we will see that

\[
X_n \xrightarrow{d} X \not\iff X_n \xrightarrow{p} X
\]

for some rv \( X \). Let \( X_n \) be identically distributed rv’s and let \((X_n, X)\) have the following joint distribution:

\[
\begin{array}{c|cc}
X_n & 0 & 1 \\
\hline
X & 0 & 1/2 & 1/2 \\
1 & 1/2 & 0 & 1/2 \\
\end{array}
\]

Obviously, \( X_n \xrightarrow{d} X \) since all have exactly the same cdf, but for any \( \epsilon \in (0, 1) \), it is

\[
P(|X_n - X| > \epsilon) = P(|X_n - X| = 1) = 1 \quad \forall n,
\]

so \( \lim_{n \to \infty} P(|X_n - X| > \epsilon) = 0 \). Therefore, \( X_n \not\xrightarrow{p} X \).

Theorem 6.1.17:
Let \( \{X_n\}_{n=1}^{\infty} \) and \( \{Y_n\}_{n=1}^{\infty} \) be sequences of rv’s and \( X \) be a rv defined on a probability space \((\Omega, L, P)\). Then it holds:

\[
Y_n \xrightarrow{d} X, |X_n - Y_n| \xrightarrow{p} 0 \implies X_n \xrightarrow{d} X.
\]

Proof:
Similar to the proof of Theorem 6.1.14. See also Rohatgi, page 253, Theorem 14.
Theorem 6.1.18: Slutsky’s Theorem
Let \((X_n)_{n=1}^{\infty}\) and \((Y_n)_{n=1}^{\infty}\) be sequences of rv’s and \(X\) be a rv defined on a probability space \((\Omega, \mathcal{F}, P)\). Let \(c \in \mathbb{R}\) be a constant. Then it holds:

(i) \(X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c \implies X_n \xrightarrow{d} X + c.\)

(ii) \(X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c \implies X_nY_n \xrightarrow{d} cX.\)

If \(c = 0\), then also \(X_nY_n \xrightarrow{p} 0.\)

(iii) \(X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c \implies \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}\) if \(c \neq 0.\)

Proof:

(i) \(Y_n \xrightarrow{p} c \xrightarrow{Th.6.1.17(i)} Y_n - c \xrightarrow{p} 0\)

\[\implies Y_n - c = Y_n + (X_n - X_n) - c = (X_n + Y_n) - (X_n + c) \xrightarrow{p} 0 \quad (A)\]

\(X_n \xrightarrow{d} X \xrightarrow{Th.6.1.9(i)} X_n + c \xrightarrow{d} X + c \quad (B)\)

Combining (A) and (B), it follows from Theorem 6.1.17:

\(X_n + Y_n \xrightarrow{d} X + c\)

(ii) Case \(c = 0\):

\(\forall \epsilon > 0 \quad \forall k > 0\), it is

\[P(\mid X_nY_n \mid > \epsilon) = P(\mid X_nY_n \mid > \epsilon, Y_n \leq \frac{\epsilon}{k}) + P(\mid X_nY_n \mid > \epsilon, Y_n > \frac{\epsilon}{k})\]

\[\leq P(\mid X_n \mid > \epsilon) + P(Y_n > \frac{\epsilon}{k})\]

\[\leq P(\mid X_n \mid > k) + P(\mid Y_n \mid > \frac{\epsilon}{k})\]

Since \(X_n \xrightarrow{d} X\) and \(Y_n \xrightarrow{p} 0\), it follows

\[\lim_{n \to \infty} P(\mid X_nY_n \mid > \epsilon) \leq P(\mid X_n \mid > k) \to 0\] as \(k \to \infty.\)

Therefore,

\[X_nY_n \xrightarrow{p} 0.\]
Case $c \neq 0$:
Since $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, it follows $X_n Y_n - cX_n = X_n (Y_n - c) \xrightarrow{p} 0$.

\[ \Rightarrow X_n Y_n \xrightarrow{p} cX_n \]

Th.6.1.14 $X_n Y_n \xrightarrow{d} cX_n$

Since $cX_n \xrightarrow{d} cX$ by Theorem 6.1.9 (ii), it follows from Theorem 6.1.17:

$X_n Y_n \xrightarrow{d} cX$

(iii) Let $Z_n \xrightarrow{p} 1$ and let $Y_n = cZ_n$.

\[ \frac{c}{Z_n} \xrightarrow{p} 1 \]

Th.6.1.11 (v, viii) $\frac{1}{Z_n} \xrightarrow{p} \frac{1}{c}$

With part (ii) above, it follows:

\[ X_n \xrightarrow{d} X \quad \text{and} \quad \frac{1}{Z_n} \xrightarrow{p} \frac{1}{c} \]

\[ \Rightarrow \frac{X_n}{Z_n} \xrightarrow{d} \frac{X}{c} \]

\[ \text{Definition 6.1.19:} \]
Let $(X_n)_{n=1}^{\infty}$ be a sequence of rv’s such that $E(|X_n|^{r}) < \infty$ for some $r > 0$. We say that $X_n$ converges in the $r^{th}$ mean to a rv $X$ ($X_n \xrightarrow{r} X$) if $E(|X|^{r}) < \infty$ and

\[ \lim_{n \to \infty} E(|X_n - X|^{r}) = 0. \]

\[ \text{Example 6.1.20:} \]
Let $(X_n)_{n=1}^{\infty}$ be a sequence of rv’s defined by $P(X_n = 0) = 1 - \frac{1}{n}$ and $P(X_n = 1) = \frac{1}{n}$.

It is $E(|X_n|^{r}) = \frac{1}{n}$ $\forall r > 0$. Therefore, $X_n \xrightarrow{r} 0$ $\forall r > 0$.

\[ \text{Note:} \]
The special cases $r = 1$ and $r = 2$ are called convergence in absolute mean for $r = 1$ ($X_n \xrightarrow{1} X$) and convergence in mean square for $r = 2$ ($X_n \xrightarrow{m^2} X$ or $X_n \xrightarrow{2} X$).

\[ \text{Theorem 6.1.21:} \]
Assume that $X_n \xrightarrow{r} X$ for some $r > 0$. Then $X_n \xrightarrow{p} X$.

\[ \text{Proof:} \]
Using Markov’s Inequality (Corollary 3.5.2), it holds for any $\epsilon > 0$:

\[ \frac{E(|X_n - X|^{r})}{\epsilon^{r}} \geq P(|X_n - X| \geq \epsilon) \]
\[ X_n \xrightarrow{p} X \implies \lim_{n \to \infty} E(\left| X_n - X \right|^r) = 0 \]

\[ \implies \lim_{n \to \infty} P(\left| X_n - X \right| \geq \epsilon) \leq \lim_{n \to \infty} \frac{E(\left| X_n - X \right|^r)}{\epsilon^r} = 0 \]

\[ \implies X_n \xrightarrow{p} X \]

\[ \text{Example 6.1.22:} \]

Let \( (X_n)_{n=1}^\infty \) be a sequence of rv's defined by \( P(X_n = 0) = 1 - \frac{1}{n^r} \) and \( P(X_n = n) = \frac{1}{n^r} \) for some \( r > 0 \).

For any \( \epsilon > 0 \), \( P(\left| X_n \right| > \epsilon) \to 0 \) as \( n \to \infty \); so \( X_n \xrightarrow{p} 0 \).

For \( 0 < s < r \), \( E(\left| X_n \right|^s) = \frac{1}{n^{r-s}} \to 0 \) as \( n \to \infty \); so \( X_n \xrightarrow{s} 0 \). But \( E(\left| X_n \right|^r) \neq 0 \) as \( n \to \infty \); so \( X_n \nrightarrow 0 \).

\[ \text{Theorem 6.1.23:} \]

If \( X_n \xrightarrow{r} X \), then it holds:

(i) \[ \lim_{n \to \infty} E(\left| X_n \right|^r) = E(\left| X \right|^r) \]; and

(ii) \( X_n \xrightarrow{s} X \) for \( 0 < s < r \).

\[ \text{Proof:} \]

(i) For \( 0 < r \leq 1 \), it holds:

\[ E(\left| X_n \right|^r) = E(\left| X_n - X + X \right|^r) \leq E(\left| X_n - X \right|^r + \left| X \right|^r) \]

\[ \implies E(\left| X_n \right|^r) - E(\left| X \right|^r) \leq E(\left| X_n - X \right|^r) \]

\[ \implies \lim_{n \to \infty} E(\left| X_n \right|^r) = \lim_{n \to \infty} E(\left| X_n - X \right|^r) = 0 \]

\[ \implies \lim_{n \to \infty} E(\left| X_n \right|^r) \leq E(\left| X \right|^r) \quad (A) \]

Similarly,

\[ E(\left| X \right|^r) = E(\left| X - X_n + X_n \right|^r) \leq E(\left| X - X_n \right|^r + \left| X_n \right|^r) \]

\[ \implies E(\left| X \right|^r) - E(\left| X_n \right|^r) \leq E(\left| X - X_n \right|^r) \]

\[ \implies \lim_{n \to \infty} E(\left| X \right|^r) - \lim_{n \to \infty} E(\left| X_n \right|^r) \leq \lim_{n \to \infty} E(\left| X - X_n \right|^r) = 0 \]

\[ \implies E(\left| X \right|^r) \leq \lim_{n \to \infty} E(\left| X_n \right|^r) \quad (B) \]

Combining \((A)\) and \((B)\) gives

\[ \lim_{n \to \infty} E(\left| X_n \right|^r) = E(\left| X \right|^r) \]
For $r > 1$, it follows from Minkowski’s Inequality (Theorem 4.8.3):

\[
\left[ E\left( \left| X - X_n + X_n \right| \right) \right]^\frac{1}{r} \leq \left[ E\left( \left| X - X_n \right| \right) \right]^\frac{1}{r} + \left[ E\left( \left| X_n \right| \right) \right]^\frac{1}{r}
\]

\[
\implies \left[ E\left( \left| X \right| \right) \right]^\frac{1}{r} - \left[ E\left( \left| X_n \right| \right) \right]^\frac{1}{r} \leq \left[ E\left( \left| X - X_n \right| \right) \right]^\frac{1}{r}
\]

\[
\implies \left[ E\left( \left| X \right| \right) \right]^\frac{1}{r} - \lim_{n \to \infty} \left[ E\left( \left| X_n \right| \right) \right]^\frac{1}{r} \leq \lim_{n \to \infty} \left[ E\left( \left| X_n - X \right| \right) \right]^\frac{1}{r} = 0 \quad \text{since} \quad X_n \xrightarrow{r} X
\]

\[
\implies \left[ E\left( \left| X \right| \right) \right]^\frac{1}{r} \leq \lim_{n \to \infty} \left[ E\left( \left| X_n \right| \right) \right]^\frac{1}{r} \quad (C)
\]

Similarly,

\[
\left[ E\left( \left| X_n - X + X \right| \right) \right]^\frac{1}{r} \leq \left[ E\left( \left| X_n - X \right| \right) \right]^\frac{1}{r} + \left[ E\left( \left| X \right| \right) \right]^\frac{1}{r}
\]

\[
\implies \lim_{n \to \infty} \left[ E\left( \left| X_n \right| \right) \right]^\frac{1}{r} - \lim_{n \to \infty} \left[ E\left( \left| X \right| \right) \right]^\frac{1}{r} \leq \lim_{n \to \infty} \left[ E\left( \left| X_n - X \right| \right) \right]^\frac{1}{r} = 0 \quad \text{since} \quad X_n \xrightarrow{r} X
\]

\[
\implies \lim_{n \to \infty} \left[ E\left( \left| X_n \right| \right) \right]^\frac{1}{r} \leq \left[ E\left( \left| X \right| \right) \right]^\frac{1}{r} \quad (D)
\]

Combining $(C)$ and $(D)$ gives

\[
\lim_{n \to \infty} \left[ E\left( \left| X_n \right| \right) \right]^\frac{1}{r} = \left[ E\left( \left| X \right| \right) \right]^\frac{1}{r}
\]

\[
\implies \lim_{n \to \infty} E\left( \left| X_n \right| \right) = E\left( \left| X \right| \right)
\]

(ii) For $1 \leq s < r$, it follows from Lyapunov’s Inequality (Theorem 3.5.4):

\[
\left[ E\left( \left| X_n - X \right| \right) \right]^\frac{s}{r} \leq \left[ E\left( \left| X_n - X \right| \right)^s \right]^\frac{1}{r}
\]

\[
\implies E\left( \left| X_n - X \right| \right) \leq \left[ E\left( \left| X_n - X \right| \right)^s \right]^\frac{1}{r}
\]

\[
\implies \lim_{n \to \infty} E\left( \left| X_n - X \right| \right) \leq \lim_{n \to \infty} \left[ E\left( \left| X_n - X \right| \right)^s \right]^\frac{1}{r} = 0 \quad \text{since} \quad X_n \xrightarrow{r} X
\]

\[
\implies X_n \xrightarrow{r} X
\]

An additional proof is required for $0 < s < r \leq 1$. 

\[\blacksquare\]
Definition 6.1.24:
Let \( \{X_n\}_{n=1}^{\infty} \) be a sequence of rv’s on \((\Omega, L, P)\). We say that \(X_n\) converges almost surely to a rv \(X (X_n \overset{a.s.}{\rightarrow} X)\) or \(X_n\) converges with probability 1 to \(X (X_n \overset{w.p.1}{\rightarrow} X)\) or \(X_n\) converges strongly to \(X\) iff
\[
P(\{\omega : X_n(\omega) \rightarrow X(\omega) \text{ as } n \rightarrow \infty\}) = 1.
\]

Note:
An interesting characterization of convergence with probability 1 and convergence in probability can be found in Parzen (1960) “Modern Probability Theory and Its Applications” on page 416 (see Handout).

Example 6.1.25:
Let \(\Omega = [0, 1]\) and \(P\) a uniform distribution on \(\Omega\). Let \(X_n(\omega) = \omega + \omega^n\) and \(X(\omega) = \omega\).

For \(\omega \in [0, 1]\), \(\omega^n \to 0\) as \(n \to \infty\). So \(X_n(\omega) \to X(\omega) \ \forall \omega \in [0, 1]\).

However, for \(\omega = 1\), \(X_n(1) = 2 \neq 1 = X(1) \ \forall n\), i.e., convergence fails at \(\omega = 1\).

Anyway, since \(P(\{\omega : X_n(\omega) \rightarrow X(\omega) \text{ as } n \rightarrow \infty\}) = P(\{\omega \in [0, 1]\}) = 1\), it is \(X_n \overset{a.s}{\rightarrow} X\).

Theorem 6.1.26:
\(X_n \overset{a.s}{\rightarrow} X \implies X_n \overset{p}{\rightarrow} X\).

Proof:
Choose \(\epsilon > 0\) and \(\delta > 0\). Find \(n_0 = n_0(\epsilon, \delta)\) such that
\[
P(\bigcap_{n=n_0}^{\infty} \{|X_n - X| \leq \epsilon\}) \geq 1 - \delta.
\]

Since \(\bigcap_{n=n_0}^{\infty} \{|X_n - X| \leq \epsilon\} \subseteq \{|X_n - X| \leq \epsilon\} \ \forall n \geq n_0\), it is
\[
P(\{|X_n - X| \leq \epsilon\}) \geq P(\bigcap_{n=n_0}^{\infty} \{|X_n - X| \leq \epsilon\}) \geq 1 - \delta \ \forall n \geq n_0.
\]

Therefore, \(P(\{|X_n - X| \leq \epsilon\}) \to 1\) as \(n \to \infty\). Thus, \(X_n \overset{p}{\rightarrow} X\).
Example 6.1.27:
\[ X_n \overset{p}{\to} X \nLeftrightarrow X_n \overset{a.s.}{\to} X \]

Let \( \Omega = (0, 1] \) and \( P \) a uniform distribution on \( \Omega \).
Define \( A_n \) by
\[
\begin{align*}
A_1 &= (0, \frac{1}{2}], A_2 = (\frac{1}{2}, 1] \\
A_3 &= (0, \frac{1}{4}], A_4 = (\frac{1}{4}, \frac{1}{2}], A_5 = (\frac{1}{2}, \frac{3}{4}], A_6 = (\frac{3}{4}, 1] \\
A_7 &= (0, \frac{1}{8}], A_8 = (\frac{1}{8}, \frac{1}{4}], \ldots
\end{align*}
\]

Let \( X_n(\omega) = I_{A_n}(\omega) \).
It is \( P(|X_n - 0| \geq \epsilon) \to 0 \ \forall \epsilon > 0 \) since \( X_n \) is 0 except on \( A_n \) and \( P(A_n) \downarrow 0 \). Thus \( X_n \overset{p}{\to} 0 \).

But \( P(\{\omega : X_n(\omega) \to 0\}) = 0 \) (and not 1) because any \( \omega \) keeps being in some \( A_n \) beyond any \( n_0 \), i.e., \( X_n(\omega) \) looks like 0...010...010...010..., so \( X_n \overset{a.s.}{\to} 0 \).

Example 6.1.28:
\[ X_n \overset{r}{\to} X \nLeftrightarrow X_n \overset{a.s.}{\to} X \]

Let \( X_n \) be independent rv’s such that \( P(\{X_n = 0\}) = 1 - \frac{1}{n} \) and \( P(\{X_n = 1\}) = \frac{1}{n} \).
It is \( E(|X_n - 0|) = E(|X_n|) = \frac{1}{n} \to 0 \) as \( n \to \infty \), so \( X_n \overset{r}{\to} 0 \ \forall r > 0 \).

But
\[
P(X_n = 0 \ \forall m \leq n \leq n_0) = \prod_{n=m}^{n_0} (1 - \frac{1}{n}) = (\frac{m-1}{m})(\frac{m+1}{m+2}) \cdots (\frac{n_0-2}{n_0})(\frac{n_0-1}{n_0}) = \frac{m-1}{n_0}
\]

As \( n_0 \to \infty \), it is \( P(X_n = 0 \ \forall m \leq n \leq n_0) \to 0 \ \forall m \), so \( X_n \overset{a.s.}{\to} 0 \).

Example 6.1.29:
\[ X_n \overset{a.s.}{\to} X \nLeftrightarrow X_n \overset{r}{\to} X \]

Let \( \Omega = [0, 1] \) and \( P \) a uniform distribution on \( \Omega \).

Let \( A_n = [0, \frac{1}{m^n}] \).

Let \( X_n(\omega) = nI_{A_n}(\omega) \) and \( X(\omega) = 0 \).
It holds that \( \forall \omega > 0 \ \exists n_0 : \frac{1}{m^{n_0}} < \omega \implies X_n(\omega) = 0 \ \forall n > n_0 \) and \( P(\omega = 0) = 0 \). Thus, \( X_n \overset{a.s.}{\to} 0 \).

But \( E(|X_n - 0|) = \frac{n^r}{n} \to \infty \ \forall r > 0 \), so \( X_n \overset{r}{\to} X \).
6.2 Weak Laws of Large Numbers

Theorem 6.2.1: WLLN: Version I

Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of iid rv's with mean \( E(X_i) = \mu \) and variance \( \text{Var}(X_i) = \sigma^2 < \infty \). Let \( \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \). Then it holds

\[
\lim_{n \to \infty} P(\mid \overline{X}_n - \mu \mid \geq \epsilon) = 0,
\]

i.e., \( \overline{X}_n \xrightarrow{p} \mu \).

Proof:

By Markov's Inequality, it holds for all \( \epsilon > 0 \):

\[
P(\mid \overline{X}_n - \mu \mid \geq \epsilon) \leq \frac{E((\overline{X}_n - \mu)^2)}{\epsilon^2} = \frac{\text{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \text{ as } n \to \infty.
\]

Note:

For iid rv's with finite variance, \( \overline{X}_n \) is consistent for \( \mu \).

A more general way to derive a "WLLN" follows in the next Definition.

Definition 6.2.2:

Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of rv's. Let \( T_n = \sum_{i=1}^{n} X_i \). We say that \( \{X_i\} \) obeys the WLLN with respect to a sequence of norming constants \( \{B_i\}_{i=1}^{\infty}, B_i > 0, B_i \uparrow \infty \), if there exists a sequence of centering constants \( \{A_i\}_{i=1}^{\infty} \) such that

\[
B_n^{-1}(T_n - A_n) \xrightarrow{p} 0.
\]

Theorem 6.2.3:

Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of pairwise uncorrelated rv's with \( E(X_i) = \mu_i \) and \( \text{Var}(X_i) = \sigma_i^2, i \in \mathbb{N} \).

If \( \sum_{i=1}^{n} \sigma_i^2 \to \infty \) as \( n \to \infty \), we can choose \( A_n = \sum_{i=1}^{n} \mu_i \) and \( B_n = \sum_{i=1}^{n} \sigma_i^2 \) and get

\[
\frac{\sum_{i=1}^{n} (X_i - \mu_i)}{\sum_{i=1}^{n} \sigma_i^2} \xrightarrow{p} 0,
\]

\begin{align*}
\text{(136)}
\end{align*}
Proof:
By Markov’s Inequality, it holds for all $\epsilon > 0$:

$$P\left(\left| \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i \right| > \epsilon \sum_{i=1}^{n} \sigma_i^2 \right) \leq \frac{E\left(\sum_{i=1}^{n} (X_i - \mu)^2 \right)}{\epsilon^2 \sum_{i=1}^{n} \sigma_i^2} = \frac{1}{\epsilon^2 \sum_{i=1}^{n} \sigma_i^2} \rightarrow 0 \text{ as } n \rightarrow \infty$$

Note:
To obtain Theorem 6.2.1, we choose $A_n = n\mu$ and $B_n = n\sigma^2$.

Theorem 6.2.4:
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of rv’s. Let $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. A necessary and sufficient condition for $\{X_i\}$ to obey the WLLN with respect to $B_n = n$ is that

$$E\left(\frac{\overline{X}_n^2}{1 + \overline{X}_n^2}\right) \rightarrow 0$$
as $n \rightarrow \infty$.

Proof:
Rohatgi, page 258, Theorem 2.

Example 6.2.5:
Let $(X_1, \ldots, X_n)$ be jointly Normal with $E(X_i) = 0$, $E(X_i^2) = 1$ for all $i$, and $Cov(X_i, X_j) = \rho$ if $|i - j| = 1$ and $Cov(X_i, X_j) = 0$ if $|i - j| > 1$. Then, $T_n \sim N(0, n + 2(n - 1)\rho) = N(0, \sigma^2)$. It is

$$E\left(\frac{\overline{X}_n^2}{1 + \overline{X}_n^2}\right) = E\left(\frac{T_n^2}{n^2 + T_n^2}\right) = \frac{2}{\sqrt{2\pi}\sigma} \int_{0}^{\infty} \frac{x^2}{n^2 + x^2} e^{-\frac{x^2}{2\sigma^2}} \, dx \quad | \quad y = \frac{x}{\sigma}, \quad dy = \frac{dx}{\sigma}$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} \frac{\sigma^2 y^2}{n^2 + \sigma^2 y^2} e^{-\frac{y^2}{2}} \, dy$$

$$= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} \frac{(n + 2(n - 1)\rho)y^2}{n^2 + (n + 2(n - 1)\rho)y^2} e^{-\frac{y^2}{2}} \, dy$$

$$\leq \frac{n + 2(n - 1)\rho}{n^2} \left[ \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} y^2 e^{-\frac{y^2}{2}} \, dy \right] = 1, \text{ since Var of } N(0,1) \text{ distribution}$$

$$\rightarrow 0 \text{ as } n \rightarrow \infty$$

$$\Rightarrow \overline{X}_n \overset{p}{\rightarrow} 0$$
Note:
We would like to have a WLLN that just depends on means but does not depend on the existence of finite variances. To approach this, we consider the following:

Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of rv’s. Let \( T_n = \sum_{i=1}^{n} X_i \). We truncate each \( X_i \) at \( c > 0 \) and get

\[
X_i^c = \begin{cases} 
X_i, & |X_i| \leq c \\
0, & \text{otherwise}
\end{cases}
\]

Let \( T_n^c = \sum_{i=1}^{n} X_i^c \) and \( m_n = \sum_{i=1}^{n} E(X_i^c) \).

Lemma 6.2.6:
For \( T_n, T_n^c \) and \( m_n \) as defined in the Note above, it holds:

\[
P(|T_n - m_n| > \epsilon) \leq P(|T_n^c - m_n| > \epsilon) + \sum_{i=1}^{n} P(|X_i| > c) \quad \forall \epsilon > 0
\]

Proof:
It holds for all \( \epsilon > 0 \):

\[
P(|T_n - m_n| > \epsilon) = P(|T_n - m_n| > \epsilon \text{ and } |X_i| \leq c \quad \forall i \in \{1, \ldots, n\}) + P(|T_n - m_n| > \epsilon \text{ and } |X_i| > c \text{ for at least one } i \in \{1, \ldots, n\}) \\
\leq P(|T_n - m_n| > \epsilon) + P(|X_i| > c \text{ for at least one } i \in \{1, \ldots, n\}) \\
\leq P(|T_n^c - m_n| > \epsilon) + \sum_{i=1}^{n} P(|X_i| > c)
\]

(*) holds since \( T_n^c = T_n \) when \( |X_i| \leq c \quad \forall i \in \{1, \ldots, n\} \).

Note:
If the \( X_i \)’s are identically distributed, then

\[
P(|T_n - m_n| > \epsilon) \leq P(|T_n^c - m_n| > \epsilon) + n P(|X_1| > c) \quad \forall \epsilon > 0.
\]

If the \( X_i \)’s are iid, then

\[
P(|T_n - m_n| > \epsilon) \leq \frac{n E((X_1^c)^2)}{\epsilon^2} + n P(|X_1| > c) \quad \forall \epsilon > 0 \quad (*).
\]

Note that \( P(|X_i| > c) = P(|X_1| > c) \quad \forall i \in \mathbb{N} \) if the \( X_i \)’s are identically distributed and that \( E((X_1^c)^2) = E((X_1^c)^2) \quad \forall i \in \mathbb{N} \) if the \( X_i \)’s are iid.
**Theorem 6.2.7: Khintchine's WLLN**

Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of iid rv's with finite mean \( E(X_i) = \mu \). Then it holds:

\[
\overline{X}_n = \frac{1}{n} T_n \xrightarrow{p} \mu
\]

**Proof:**

If we take \( c = n \) and replace \( \epsilon \) by \( n\epsilon \) in (\(*\)) in the Note above, we get

\[
P(|T_n - m_n| > n\epsilon) \leq \frac{E((X_1^n)^2)}{n\epsilon^2} + nP(|X_1| > n).
\]

Since \( E(|X_1|) < \infty \), it is \( nP(|X_1| > n) \to 0 \) as \( n \to \infty \) by Theorem 3.1.9. From Corollary 3.1.12 we know that \( E(|X|^\alpha) = \alpha \int_0^\infty x^{\alpha-1} P(|X| > x)dx \). Therefore,

\[
E((X_1^n)^2) = 2 \int_0^n xP(|X_1^n| > x)dx = 2 \int_0^A xP(|X_1^n| > x)dx + 2 \int_A^n xP(|X_1^n| > x)dx \leq K + \delta \int_A^n dx \leq K + n\delta
\]

In (\(+)\), \( A \) is chosen sufficiently large such that \( xP(|X_1^n| > x) < \frac{\delta}{2} \) \( \forall x \geq A \) for an arbitrary constant \( \delta > 0 \) and \( K > 0 \) a constant.

Therefore,

\[
\frac{E((X_1^n)^2)}{n\epsilon^2} \leq \frac{k}{n\epsilon^2} + \frac{\delta}{\epsilon^2}
\]

Since \( \delta \) is arbitrary, we can make the right hand side of this last inequality arbitrarily small for sufficiently large \( n \).

Since \( E(X_i) = \mu \) \( \forall i \), it is \( \frac{m_n}{n} \to \mu \) as \( n \to \infty \).  

**Note:**

Theorem 6.2.7 meets the previously stated goal of not having a finite variance requirement.
6.3 Strong Laws of Large Numbers

Definition 6.3.1:
Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of rv’s. Let \( T_n = \sum_{i=1}^{n} X_i \). We say that \( \{X_i\} \) obeys the SLLN with respect to a sequence of norming constants \( \{B_i\}_{i=1}^{\infty}, B_i > 0, B_i \uparrow \infty \), if there exists a sequence of centering constants \( \{A_i\}_{i=1}^{\infty} \) such that
\[
B_n^{-1}(T_n - A_n) \xrightarrow{a.s.} 0.
\]

Note:
Unless otherwise specified, we will only use the case that \( B_n = n \) in this section.

Theorem 6.3.2:
\( X_n \xrightarrow{a.s.} X \iff \lim_{n \to \infty} P(\sup_{m \geq n} |X_m - X| > \epsilon) = 0 \ \forall \epsilon > 0 \).

Proof: (see also Rohatgi, page 249, Theorem 11)
WLOG, we can assume that \( X = 0 \) since \( X_n \xrightarrow{a.s.} X \) implies \( X_n - X \xrightarrow{a.s.} 0 \). Thus, we have to prove:
\[
X_n \xrightarrow{a.s.} 0 \iff \lim_{n \to \infty} P(\sup_{m \geq n} |X_m| > \epsilon) = 0 \ \forall \epsilon > 0
\]

"\( \iff \)" :
Choose \( \epsilon > 0 \) and define
\[
A_n(\epsilon) = \{\sup_{m \geq n} |X_m| > \epsilon\}
C = \{\lim_{n \to \infty} X_n = 0\}
\]
We know that \( P(C) = 1 \) and therefore \( P(C^c) = 0 \).
Let \( B_n(\epsilon) = C \cap A_n(\epsilon) \). Note that \( B_{n+1}(\epsilon) \subseteq B_n(\epsilon) \) and for the limit set \( \bigcap_{n=1}^{\infty} B_n(\epsilon) = 0 \). It follows that
\[
\lim_{n \to \infty} P(B_n(\epsilon)) = P\left(\bigcap_{n=1}^{\infty} B_n(\epsilon)\right) = 0.
\]
We also have

\[ P(B_n(\epsilon)) = P(A_n \cap C) \]
\[ = 1 - P(C^c \cup A_n^c) \]
\[ = 1 - P(C^c) - P(A_n^c) + P(C^c \cap A_n^c) \]
\[ = P(A_n) \]

\[ \lim_{n \to \infty} P(A_n(\epsilon)) = 0 \]

"\( \Leftarrow \)"

Assume that \( \lim_{n \to \infty} P(A_n(\epsilon)) = 0 \) \( \forall \epsilon > 0 \) and define \( D(\epsilon) = \{ \lim_{n \to \infty} X_n \mid X_n \geq \epsilon \} \). Since \( D(\epsilon) \subseteq A_n(\epsilon) \) \( \forall n \in \mathbb{N} \), it follows that \( P(D(\epsilon)) = 0 \) \( \forall \epsilon > 0 \). Also,

\[ C^c = \{ \lim_{n \to \infty} X_n \neq 0 \} \subseteq \bigcup_{k=1}^{\infty} \{ \lim_{n \to \infty} X_n \geq \frac{1}{k} \}. \]

\[ \implies 1 - P(C) \leq \sum_{k=1}^{\infty} P(D(\frac{1}{k})) = 0 \]
\[ \implies X_n \overset{a.s.}{\to} 0 \]

Note:

(i) \( X_n \overset{a.s.}{\to} 0 \) implies that \( \forall \epsilon > 0 \) \( \forall \delta > 0 \) \( \exists n_0 \in \mathbb{N} : P(\sup_{n \geq n_0} |X_n| > \epsilon) < \delta. \)

(ii) Recall that for a given sequence of events \( \{A_n\}_{n=1}^{\infty} \),

\[ A = \lim_{n \to \infty} A_n = \lim_{n \to \infty} \bigcup_{k=n}^{\infty} A_k = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \]

is the event that infinitely many of the \( A_n \) occur. We write \( P(A) = P(A_n \ i.o.) \) where \( i.o. \) stands for "infinitely often".

(iii) Using the terminology defined in (ii) above, we can rewrite Theorem 6.3.2 as

\[ X_n \overset{a.s.}{\to} 0 \iff P(\sup_{n} |X_n| > \epsilon \ i.o.) = 0 \ \forall \epsilon > 0. \]
Theorem 6.3.3: Borel–Cantelli Lemma

(i) 1st BC–Lemma:

Let \( \{A_n\}_{n=1}^{\infty} \) be a sequence of events such that \( \sum_{n=1}^{\infty} P(A_n) < \infty \). Then \( P(A) = 0 \).

(ii) 2nd BC–Lemma:

Let \( \{A_n\}_{n=1}^{\infty} \) be a sequence of independent events such that \( \sum_{n=1}^{\infty} P(A_n) = \infty \). Then \( P(A) = 1 \).

Proof:
(i):

\[
P(A) = P(\lim_{n \to \infty} \bigcup_{k=n}^{\infty} A_k)
= \lim_{n \to \infty} P(\bigcup_{k=n}^{\infty} A_k)
\leq \lim_{n \to \infty} \sum_{k=n}^{\infty} P(A_k)
= \lim_{n \to \infty} \left( \sum_{k=1}^{\infty} P(A_k) - \sum_{k=1}^{n-1} P(A_k) \right)
= 0
\]

(ii): We have \( A^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c \). Therefore,

\[
P(A^c) = P(\lim_{n \to \infty} \bigcap_{k=n}^{\infty} A_k^c) = \lim_{n \to \infty} P(\bigcap_{k=n}^{\infty} A_k^c).
\]

If we choose \( n_0 > n \), it holds that

\[
\bigcap_{k=n}^{\infty} A_k^c \subseteq \bigcap_{k=n}^{n_0} A_k^c.
\]

Therefore,

\[
P(\bigcap_{k=n}^{\infty} A_k^c) \leq \lim_{n_0 \to \infty} P(\bigcap_{k=n}^{n_0} A_k^c)
= \lim_{n_0 \to \infty} \prod_{k=n}^{n_0} (1 - P(A_k))
\]

indep.

\[
\leq \lim_{n_0 \to \infty} \exp \left( - \sum_{k=n}^{n_0} P(A_k) \right)
= 0
\]

\( \implies P(A) = 1 \) \( \blacksquare \)
Example 6.3.4:
Independence is necessary for 2nd BC-Lemma:

Let \( \Omega = (0, 1) \) and \( P \) a uniform distribution on \( \Omega \).

Let \( A_n = I_{(0, \frac{1}{n})}(\omega) \). Therefore,
\[
\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.
\]

But for any \( \omega \in \Omega \), \( A_n \) occurs only for \( 1, 2, \ldots, \lfloor \frac{1}{\omega} \rfloor \), where \( \lfloor \frac{1}{\omega} \rfloor \) denotes the largest integer ("floor") that is \( \leq \frac{1}{\omega} \). Therefore, \( P(A) = P(A_n \cap \omega) = 0 \).

Lemma 6.3.5: Kolmogorov’s Inequality

Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of independent rv’s with common mean 0 and variances \( \sigma_i^2 \). Let \( T_n = \sum_{i=1}^{n} X_i \). Then it holds:
\[
P(\max_{1 \leq k \leq n} |T_k| \geq \epsilon) \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{\epsilon^2} \quad \forall \epsilon > 0
\]

Proof:
See Rohatgi, page 268, Lemma 2.

Lemma 6.3.6: Kronecker’s Lemma

If \( \sum_{i=1}^{\infty} X_i \) converges to \( s < \infty \) and \( B_n \uparrow \infty \), then it holds:
\[
\frac{1}{B_n} \sum_{i=1}^{n} B_k X_k \to 0
\]

Proof:
See Rohatgi, page 269, Lemma 3.

Theorem 6.3.7: Cauchy Criterion

\( X_n \xrightarrow{a.s.} X \iff \lim_{n \to \infty} P(\sup_{m} |X_{n+m} - X_n| \leq \epsilon) = 1 \quad \forall \epsilon > 0 \).

Proof:
See Rohatgi, page 270, Theorem 5.
Theorem 6.3.8:
If \( \sum_{n=1}^{\infty} \text{Var}(X_n) < \infty \), then \( \sum_{n=1}^{\infty} (X_n - E(X_n)) \) converges almost surely.

Proof:
See Rohatgi, page 272, Theorem 6.
Corollary 6.3.9:
Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of independent rv’s. Let \( \{B_i\}_{i=1}^{\infty}, B_i > 0, B_i \uparrow \infty \), a sequence of norming constants. Let \( T_n = \sum_{i=1}^{n} X_i \). If \( \sum_{i=1}^{\infty} \frac{\text{Var}(X_i)}{B_i^2} < \infty \) then it holds:
\[
\frac{T_n - E(T_n)}{B_n} \xrightarrow{a.s.} 0
\]

Proof:
This Corollary follows directly from Theorem 6.3.8 and Lemma 6.3.6.

Lemma 6.3.10: Equivalence Lemma
Let \( \{X_i\}_{i=1}^{\infty} \) and \( \{X'_i\}_{i=1}^{\infty} \) be sequences of rv’s. Let \( T_n = \sum_{i=1}^{n} X_i \) and \( T'_n = \sum_{i=1}^{n} X'_i \).

If the series \( \sum_{i=1}^{\infty} P(X_i \neq X'_i) < \infty \), then the series \( \{X_i\} \) and \( \{X'_i\} \) are tail-equivalent and \( T_n \) and \( T'_n \) are convergence-equivalent, i.e., for \( B_n \uparrow \infty \) the sequences \( \frac{1}{B_n} T_n \) and \( \frac{1}{B_n} T'_n \) converge on the same event and to the same limit, except for a null set.

Proof:
See Rohatgi, page 266, Lemma 1.

Lemma 6.3.11:
Let \( X \) be a rv with \( E(|X|) < \infty \). Then it holds:
\[
\sum_{n=1}^{\infty} P(|X| \geq n) \leq E(|X|) \leq 1 + \sum_{n=1}^{\infty} P(|X| \geq n)
\]

Proof:
Continuous case only:
Let \( X \) have a pdf \( f \). Then it holds:
\[
E(|X|) = \int_{-\infty}^{\infty} |x| f(x) \, dx = \sum_{k=0}^{\infty} \int_{k \leq |x| \leq k+1} |x| f(x) \, dx
\]
\[
\Rightarrow \sum_{k=0}^{\infty} k P(k \leq |X| \leq k + 1) \leq E(|X|) \leq \sum_{k=0}^{\infty} (k + 1) P(k \leq |X| \leq k + 1)
\]
It is
\[
\sum_{k=0}^{\infty} k P(k \leq |X| \leq k + 1) = \sum_{k=0}^{\infty} \sum_{n=1}^{k} P(k \leq |X| \leq k + 1)
\]
\[
= \sum_{n=1}^{\infty} \sum_{k=n}^{\infty} P(k \leq |X| \leq k + 1)
\]
\[
= \sum_{n=1}^{\infty} P(|X| \geq n)
\]

Similarly,
\[
\sum_{k=0}^{\infty} (k + 1) P(k \leq |X| \leq k + 1) = \sum_{n=1}^{\infty} P(|X| \geq n) + \sum_{k=0}^{\infty} P(k \leq |X| \leq k + 1)
\]
\[
= \sum_{n=1}^{\infty} P(|X| \geq n) + 1
\]

**Theorem 6.3.12: Kolmogorov’s SLLN**

Let \( \{X_i\}_{i=1}^{\infty} \) be a sequence of iid rv’s. Let \( T_n = \sum_{i=1}^{n} X_i \). Then it holds:
\[
\frac{T_n}{n} = \overline{X}_n \rightarrow a.s. \mu < \infty \iff E(|X|) < \infty \text{ (and then } \mu = E(X))
\]

**Proof:**

\( \Rightarrow \) :

Suppose that \( \overline{X}_n \rightarrow a.s. \mu < \infty \). It is
\[
T_n = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n = T_{n-1} + X_n.
\]
\[
\Rightarrow \frac{T_n}{n} = \frac{T_{n-1}}{n} + \frac{X_n}{n} \rightarrow a.s. \mu \rightarrow 1
\]
\[
\frac{X_n}{n} \rightarrow ^{a.s.} 0
\]

By 1st Borel–Cantelli Lemma, we must have \( \sum_{n=1}^{\infty} P(|X_n| \geq n) < \infty \).

\( \Rightarrow \)

Lemma 6.3.11
\[
E(|X|) < \infty
\]

Th. 6.2.7 (WLLN)
\[\overline{X}_n \rightarrow p \mu \rightarrow E(X)\]

Since \( \overline{X}_n \rightarrow a.s. \mu \), it holds that \( \overline{X}_n \rightarrow p \mu \). Therefore, it must hold that \( \mu = E(X) \).
Let $E(|X|) < \infty$. Define truncated r.v.’s:

$$
X_k' = \begin{cases} X_k, & \text{if } |X_k| \leq k \\ 0, & \text{otherwise} \end{cases}
$$

$$
T_n' = \sum_{k=1}^n X_k'
$$

$$
\overline{X}_n = \frac{T_n'}{n}
$$

Then it holds:

$$
\sum_{k=1}^\infty P(X_k \neq X_k') = \sum_{k=1}^\infty P(|X_k| \geq k)
$$

**iid**

$$
\sum_{k=1}^\infty P(|X| \geq k)
$$

**Lemma 6.3.11**

$$
< E(|X|)
$$

By Lemma 6.3.10, it follows that $T_n$ and $T_n'$ are convergence-equivalent. Thus, it is sufficient to prove that $\overline{X}_n \overset{a.s.}{\rightarrow} E(X)$.

We now establish the conditions needed in Corollary 6.3.9. It is

$$
Var(X_n') \leq E((X_n')^2)
$$

$$
= \int_{-n}^{n} x^2 f_X(x) dx
$$

$$
= \sum_{k=0}^{n-1} \int_{k \leq |x| < k+1} x^2 f_X(x) dx
$$

$$
\leq \sum_{k=0}^{n-1} (k+1)^2 P(k \leq |X| < k+1)
$$

$$
\Rightarrow \sum_{n=1}^\infty \frac{1}{n^2} Var(X_n') \leq \sum_{n=1}^\infty \sum_{k=0}^{n} \frac{(k+1)^2}{n^2} P(k \leq |X| < k+1)
$$

$$
= \sum_{n=1}^\infty \sum_{k=1}^{n} \frac{(k+1)^2}{n^2} P(k \leq |X| < k+1) + \sum_{n=1}^\infty \frac{1}{n^2} P(0 \leq |X| < 1)
$$

$$
= \sum_{k=1}^\infty (k+1)^2 P(k \leq |X| < k+1) \left( \sum_{n=k}^\infty \frac{1}{n^2} \right) + 2P(0 \leq |X| < 1)
$$

(A)
(*) holds since \( \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.65 < 2 \) and the first two sums can be rearranged as follows:

\[
\begin{array}{c|c}\hline
n & k \\
\hline
1 & 1 \\
2 & 1, 2 \implies 2 & 2, 3, \ldots \\
3 & 1, 2, 3 & 3 & 3, \ldots \\
\vdots & \vdots & \vdots & \vdots \\
\end{array}
\]

It is

\[
\sum_{n=k}^{\infty} \frac{1}{n^2} = \frac{1}{k^2} + \frac{1}{(k+1)^2} + \frac{1}{(k+2)^2} + \cdots \\
\leq \frac{1}{k^2} + \frac{1}{k(k+1)} + \frac{1}{(k+1)(k+2)} + \cdots \\
= \frac{1}{k^2} + \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)}
\]

From Bronstein, page 30, #7, we know that

\[
1 = \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{n(n+1)} + \cdots
\]

\[
= \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \frac{1}{3 \cdot 4} + \cdots + \frac{1}{(k-1) \cdot k} + \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)}
\]

\[
\implies \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)} = 1 - \frac{1}{1 \cdot 2} - \frac{1}{2 \cdot 3} - \frac{1}{3 \cdot 4} - \cdots - \frac{1}{(k-1) \cdot k}
\]

\[
= \frac{1}{2} - \frac{1}{2 \cdot 3} - \frac{1}{3 \cdot 4} - \cdots - \frac{1}{(k-1) \cdot k}
\]

\[
= \frac{1}{3} - \frac{1}{3 \cdot 4} - \cdots - \frac{1}{(k-1) \cdot k}
\]

\[
= \frac{1}{4} - \cdots - \frac{1}{(k-1) \cdot k}
\]

\[
= \cdots
\]

\[
= \frac{1}{k}
\]

\[
\implies \sum_{n=k}^{\infty} \frac{1}{n^2} \leq \frac{1}{k^2} + \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)}
\]

\[
= \frac{1}{k^2} + \frac{1}{k}
\]

148
Using this result in (A), we get

\[
\sum_{n=1}^{\infty} \frac{1}{n^2} \text{Var}(X'_n) \leq 2 \sum_{k=1}^{\infty} \frac{(k+1)^2}{k} P(k \leq |X| < k+1) + 2P(0 \leq |X| < 1)
\]

\[
= 2 \sum_{k=0}^{\infty} kP(k \leq |X| < k+1) + 4 \sum_{k=1}^{\infty} P(k \leq |X| < k+1)
\]

\[
+ 2 \sum_{k=1}^{\infty} \frac{1}{k} P(k \leq |X| < k+1) + 2P(0 \leq |X| < 1)
\]

\[
|B| \leq 2E(|X|) + 4 + 2 + 2
\]

\[
< \infty
\]

To establish (B), we use an inequality from the Proof of Lemma 6.3.11.

Thus, the conditions needed in Corollary 6.3.9 are met. It follows that

\[
\frac{1}{n} T'_n - \frac{1}{n} E(T'_n) \overset{a.s.}{\to} 0 \quad (C)
\]

Since \( E(X'_n) \to E(X) \) as \( n \to \infty \), it follows by Kronecker’s Lemma (6.3.6) that \( \frac{1}{n} E(T'_n) \to E(X) \). Thus, when we replace \( \frac{1}{n} E(T'_n) \) by \( E(X) \) in (C), we get:

\[
\frac{1}{n} T'_n \overset{a.s.}{\to} E(X) \quad \text{Lemma 6.3.10} \quad \frac{1}{n} T_n \overset{a.s.}{\to} E(X)
\]

Merry Xmas and a Happy New Millennium!