STAT 6710/7710

Mathematical Statistics I

Fall Semester 2009

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@math.usu.edu`

Contents

Acknowledgements

I would like to thank my students, Hanadi B. Eltahir, Rich Madsen, and Bill Morphet, who helped during the Fall 1999 and Spring 2000 semesters in typesetting these lecture notes using LaTeX and for their suggestions how to improve some of the material presented in class. Thanks are also due to about 40 students who took Stat 6710/20 with me since the Fall 2000 semester for their valuable comments that helped to improve and correct these lecture notes.

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

The primary textbook required for this class is:

- Casella, G., and Berger, R. L. (2002): *Statistical Inference* (Second Edition), Duxbury Press/Thomson Learning, Pacific Grove, CA.

A Web page dedicated to this class is accessible at:

`http://www.math.usu.edu/~symanzik/teaching/2009_stat6710/stat6710.html`

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

- Bandelow, C. (1981): *Einführung in die Wahrscheinlichkeitstheorie*, Bibliographisches Institut, Mannheim, Germany.
- Bickel, P. J., and Doksum, K. A. (2001): *Mathematical Statistics, Basic Ideas and Selected Topics, Vol. I* (Second Edition), Prentice–Hall, Upper Saddle River, NJ.
- Casella, G., and Berger, R. L. (1990): *Statistical Inference*, Wadsworth & Brooks/Cole, Pacific Grove, CA.
- Fisz, M. (1989): *Wahrscheinlichkeitsrechnung und mathematische Statistik*, VEB Deutscher Verlag der Wissenschaften, Berlin, German Democratic Republic.
- Kelly, D. G. (1994): *Introduction to Probability*, Macmillan, New York, NY.
- Miller, I., and Miller, M. (1999): *John E. Freund's Mathematical Statistics* (Sixth Edition), Prentice–Hall, Upper Saddle River, NJ.

- Mood, A. M., and Graybill, F. A., and Boes, D. C. (1974): *Introduction to the Theory of Statistics* (Third Edition), McGraw-Hill, Singapore.

- Parzen, E. (1960): *Modern Probability Theory and Its Applications*, Wiley, New York, NY.

- Rohatgi, V. K. (1976): *An Introduction to Probability Theory and Mathematical Statistics*, John Wiley and Sons, New York, NY.

- Rohatgi, V. K., and Saleh, A. K. Md. E. (2001): *An Introduction to Probability and Statistics* (Second Edition), John Wiley and Sons, New York, NY.

- Searle, S. R. (1971): *Linear Models*, Wiley, New York, NY.

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

- Bronstein, I. N. and Semendjajew, K. A. (1985): *Taschenbuch der Mathematik* (22. Auflage), Verlag Harri Deutsch, Thun, German Democratic Republic.

- Bronstein, I. N. and Semendjajew, K. A. (1986): *Ergänzende Kapitel zu Taschenbuch der Mathematik* (4. Auflage), Verlag Harri Deutsch, Thun, German Democratic Republic.

- Sieber, H. (1980): *Mathematische Formeln — Erweiterte Ausgabe E*, Ernst Klett, Stuttgart, Germany.

Jürgen Symanzik, August 22, 2009.

1 Axioms of Probability

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

1.1 σ–Fields

Let Ω be the sample space of all possible outcomes of a chance experiment. Let $\omega \in \Omega$ (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 Ω 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 Ω.

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

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

(i) $\Omega \in L$

(ii) $A \in L \Longrightarrow A^C \in L$

(iii) $A, B \in L \Longrightarrow A \cup B \in L$

■

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

Recall *De Morgan's Laws:*

$$\bigcup_{A \in \mathcal{A}} A = ... \qquad \text{and} \quad \bigcap_{A \in \mathcal{A}} A = ... \qquad .$$

Note:
So (ii), (iii) imply (iii)': $A, B \in L \Longrightarrow A \cap B \in L$ [can replace (iii)].

Proof:

$$A, B \in L \Longrightarrow ...$$

■

Definition 1.1.2:

A class L of subsets of Ω is called a σ–**field** (Borel field, σ–algebra) if it is a field and closed under *countable* unions, i.e.,

(iv) $\{A_n\}_{n=1}^{\infty} \in L \Longrightarrow \bigcup\limits_{n=1}^{\infty} A_n \in L$. ■

Note:

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

Example 1.1.3:

For some Ω, let L contain all finite and all cofinite sets (A is *cofinite* if A^C is finite — for example, if $\Omega = I\!N$, $A = \{x \mid x \geq c\}$ is not finite but since $A^C = \{x \mid x < c\}$ is finite, A is cofinite). Then L is a field. But L is a σ–field **iff** (if and only if) Ω is finite.

For example, let $\Omega = Z$. Take $A_n = \{n\}$, each finite, so $A_n \in L$. But $\bigcup\limits_{n=1}^{\infty} A_n = Z^+ \not\in L$, since the set is not finite (it is infinite) and also not cofinite ($\left(\bigcup\limits_{n=1}^{\infty} A_n\right)^C = Z_0^-$ is infinite, too).

Question: Does this construction work for $\Omega = Z^+$??

■

Note:

The largest σ–field in Ω is the *power set* $\mathcal{P}(\Omega)$ of all subsets of Ω. The smallest σ–field is $L = \{\emptyset, \Omega\}$. ■

Terminology:

A set $A \in L$ is said to be "*measurable* L".

Note:
We often begin with a class of sets, say a, which may not be a field or a σ–field.

Definition 1.1.4:
The σ–field generated by a, $\sigma(a)$, is the smallest σ–field containing a, or the intersection of all σ–fields containing a. ■

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

Proof:
(ii) Suppose $L = \bigcap\limits_{\theta} L_\theta$. We have to show that conditions (i) and (ii) of Def. 1.1.1 and (iv) of Def. 1.1.2 are fulfilled:

■

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

What is $\sigma(a)$?

What is $\sigma(b)$?

■

If Ω is finite or countable, we will usually use $L = \mathcal{P}(\Omega)$. If $\mid \Omega \mid= n < \infty$, then $\mid L \mid= 2^n$.

If Ω is uncountable, $\mathcal{P}(\Omega)$ may be too large to be useful and we may have to use some smaller σ–field.

Definition 1.1.6:
If $\Omega = I\!\!R$, an important special case is the **Borel σ–field**, i.e., the σ–field generated from all half–open intervals of the form $(a, b]$, denoted $\mathcal{B}$ or $\mathcal{B}_1$. The sets of $\mathcal{B}$ are called **Borel sets**.

The Borel σ–field on $I\!\!R^d$ ($\mathcal{B}_d$) is the σ–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:
$\mathcal{B}$ contains all points: $\{x\} = \bigcap\limits_{n=1}^{\infty}$ ...
closed intervals: $[x, y] =$
open intervals: $(x, y) =$
and semi–infinite intervals: $(x, \infty) =$ ■

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

Definition 1.1.7: **Kolmogorov Axioms of Probability**
A **probability measure** (pm), P, on (Ω, L) is a set function $P : L \rightarrow I\!\!R$ 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 $\bigcup\limits_{n=1}^{\infty} A_n \in L$, then $P(\bigcup\limits_{n=1}^{\infty} A_n) = \sum\limits_{n=1}^{\infty} P(A_n)$.

■

Note:
$\bigcup\limits_{n=1}^{\infty} A_n \in L$ holds automatically if L is a σ–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 (Ω, L), it holds:

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

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

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

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

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

Proof:

$\blacksquare$

Theorem 1.2.2: **Principle of Inclusion–Exclusion**

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

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

Proof:

$n = 1$ is trivial

$n = 2$ is Theorem 1.2.1 (iv)

use induction for higher n (Homework) ■

Note:

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

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

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

Theorem 1.2.3: **Bonferroni's Inequality**

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

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

Proof:

■

Theorem 1.2.4: **Boole's Inequality**
Let $A, B \in L$. Then

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

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

Proof:
Homework ■

Definition 1.2.5: **Continuity of sets**
For a sequence of sets $\{A_n\}_{n=1}^{\infty}, A_n \in L$ and $A \in L$, we say

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

(ii) $A_n \downarrow A$ if $A_1 \supseteq A_2 \supseteq A_3 \supseteq \ldots$ and $A = \bigcap_{n=1}^{\infty} A_n$.

■

Example:

■

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:

■

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:

■

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 Ω are **equally likely** (or occur with uniform probability) if $P(\omega_j) = \frac{1}{n} \ \ \forall j = 1, \ldots, n$. ■

Note:

If this is true, $P(A) = \dfrac{\text{number } \omega_j \text{ in } A}{\text{number } \omega_j \text{ in } \Omega}$. 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 $= \underbrace{n_2 + n_2 + \ldots + n_2}_{n_1 \text{ times}} = n_1 \times n_2$.

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 $= (n_1 \times n_2 \times n_3 \times \ldots \times n_{k-1}) \times n_k$. ■

Definition 1.3.3:

For positive integer n, we define **n factorial** as $n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 = n \times (n-1)!$ and $0! = 1$. ■

Definition 1.3.4:

For nonnegative integers $n \geq r$, we define the **binomial coefficient** (read as n choose r) as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-r+1)}{1 \cdot 2 \cdot 3 \cdot \ldots \cdot r}.$$

■

Note:

A useful extension for the binomial coefficient for $n < r$ is

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

■

Note:

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

(i) the size of the set, n;

(ii) the size of the sample, r;

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

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

■

Theorem 1.3.5:

The number of ways to draw r elements from a set of n, if

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

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

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

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

Proof:

(i)

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

(ii)

$\blacksquare$

Theorem 1.3.6: **The Binomial Theorem**
If n is a non–negative integer, then

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

Proof: (By Induction)

Induction Base:

$\blacksquare$

Corollary 1.3.7:

For integers n, it holds:

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

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

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

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

Proof:

$\blacksquare$

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 Ω is used for a probability space (Ω, 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 (Ω, 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, (Ω, L, P_H) is a probability space.

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

$\blacksquare$

Note:

What we have done is to move to a new sample space $\mathcal{H}$ and a new σ–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 $\blacksquare$

Definition 1.4.4:

A collection of subsets $\{A_n\}_{n=1}^{\infty}$ of Ω form a **partition** of Ω if

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

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

$\blacksquare$

Theorem 1.4.5: **Law of Total Probability**

If $\{H_j\}_{j=1}^{\infty}$ is a partition of Ω, and $P(H_j) > 0 \quad \forall j$, 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

$\Longrightarrow$ the right side of Th. 1.4.5 is true.

The left side proof:

$\blacksquare$

Theorem 1.4.6: **Bayes' Rule**

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

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

Proof:

$\blacksquare$

Definition 1.4.7:

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

Note:

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

■

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 $A1, A_2$ and A_3 pairwise independent and also mutually independent?

■

Example 1.4.11: (from Rohatgi, page 37, Example 5)

- r students. 365 possible birthdays for each student that are equally likely.
- One student at a time is asked for his/her birthday.

- If one of the other students hears this birthday and it matches his/her birthday, this other student has to raise his/her hand — if at least one other student raises his/her hand, the procedure is over.

- We are interested in

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

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

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

  and

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

  where ${}_nP_r = n \cdot (n-1) \cdot \ldots \cdot (n-r+1)$.

<u>Proof:</u>
It is

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

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

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

For general p_k and restrictions on r and k see Homework. ■

2 Random Variables

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

2.1 Measurable Functions

Definition 2.1.1:

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

$\blacksquare$

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

Definition 2.1.2:
Suppose (Ω, L) and $(\mathcal{S}, \mathcal{B})$ are two measurable spaces and $X : \Omega \to \mathcal{S}$ is a mapping from Ω to $\mathcal{S}$. We say that X is **measurable** $L - \mathcal{B}$ if $X^{-1}(B) \in L$ for every set $B \in \mathcal{B}$, where $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$.

$\blacksquare$

Example 2.1.3:
Record the opinion of 50 people: "yes" (y) or "no" (n).
$\Omega = \{$All 2^{50} possible sequences of y/n$\}$ — HUGE !
$L = \mathcal{P}(\Omega)$

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

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

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

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

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

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

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

Some Technical Points about Measurable Functions

2.1.5:
Suppose (Ω, 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 $I\!R$ is generated by $\{(-\infty, x] : x \in I\!R\}$.

2.1.6:
If $(\Omega, L), (\Omega', L')$, and (Ω'', 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 $f : I\!R^i \to I\!R^k$ is a continuous function, then f is measurable $\mathcal{B}^i - \mathcal{B}^k$.

2.1.8:
If $f_j : \Omega \to I\!R, j = 1, \ldots k$ and $g : I\!R^k \to I\!R$ 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 $[-\infty, \infty] = I\!R \cup \{-\infty, \infty\}$.
We say $f : \Omega \to I\!R$ is measurable $L - \mathcal{B}$ if

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

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

$\blacksquare$

<u>2.1.10:</u>

Suppose $f_1, f_2, \dots$ is a sequence of real–valued measurable functions $(\Omega, L) \to (I\!R, \mathcal{B})$. Then it holds:

(i) $\sup\limits_{n\to\infty} f_n$ (supremum), $\inf\limits_{n\to\infty} f_n$ (infimum), $\lim\limits_{n\to\infty} \sup f_n$ (limit superior), and $\lim\limits_{n\to\infty} \inf f_n$ (limit inferior) are measurable.

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

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

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

$\blacksquare$

<u>Example 2.1.11:</u>

(i) Let

$$f_n(x) = \frac{1}{x^n}, \quad x > 1.$$

It holds

- $\sup\limits_{n\to\infty} f_n(x) = \frac{1}{x}$
- $\inf\limits_{n\to\infty} f_n(x) = 0$
- $\lim\limits_{n\to\infty} \sup f_n(x) = 0$
- $\lim\limits_{n\to\infty} \inf f_n(x) = 0$
- $\lim\limits_{n\to\infty} f_n(x) = \lim\limits_{n\to\infty} \sup f_n(x) = \lim\limits_{n\to\infty} \inf f_n(x) = 0$

(ii) Let

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

It holds

- $\sup\limits_{n\to\infty} f_n(x) = \begin{cases} x^3, & x \geq -1 \\ 0, & \text{otherwise} \end{cases}$
- $\inf\limits_{n\to\infty} f_n(x) = \begin{cases} x^3, & x \leq 1 \\ 0, & \text{otherwise} \end{cases}$

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

(iii) Let

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

It holds

- $\sup_{n\to\infty} f_n(x) =| x |^3$
- $\inf_{n\to\infty} f_n(x) = - | x |^3$
- $\lim_{n\to\infty} \sup f_n(x) =| x |^3$
- $\lim_{n\to\infty} \inf f_n(x) = - | x |^3$
- $\lim_{n\to\infty} f_n(x) = \lim_{n\to\infty} \sup f_n(x) = \lim_{n\to\infty} \inf f_n(x) = 0$ if $x = 0$, but $\lim_{n\to\infty} f_n(x)$ does not exist for $x \neq 0$ since $\lim_{n\to\infty} \sup f_n(x) \neq \lim_{n\to\infty} \inf f_n(x)$ for $x \neq 0$

■

2.2 Probability Distribution of a Random Variable

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

Theorem 2.2.1:
A random variable X on (Ω, L, P) **induces** a probability measure on a space $(I\!R, \mathcal{B}, Q)$ with the probability distribution Q of X defined by

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

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

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

$\blacksquare$

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 $I\!R$. $\blacksquare$

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 $(I\!R, \mathcal{B})$. The cdf associated with P is

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

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

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

<u>Proof:</u>

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 $A_n \uparrow I\!R$. ■

Definition 2.2.4:

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

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

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

■

Note:

There are rvs that are mixtures of continuous and discrete rvs. One such example is a truncated failure time distribution. We assume a continuous distribution (e.g., exponential) up to a given truncation point x and assign the "remaining" probability to the truncation point. Thus, a single point has a probability > 0 and $F(x)$ jumps at the truncation point x. ■

Definition 2.2.5:

Two random variables X and Y are **identically distributed** iff $P_X(X \in A) = P_Y(Y \in A) \;\; \forall A \in L$. ■

Note:

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

$X = \#$ H in 3 coin tosses

$Y = \#$ T in 3 coin tosses

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

Theorem 2.2.6:

The following two statements are equivalent:

(i) X, Y are identically distributed.

(ii) $F_X(x) = F_Y(x) \quad \forall x \in I\!R$.

Proof:

(i) ⇒ (ii):

(ii) ⇒ (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 (Ω, L, P). X is **discrete** if there exists a countable set $E \subset I\!R$ 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, $p_i \geq 0, \sum\limits_{i=1}^{\infty} p_i = 1$.

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 $\{p_n\}_{n=1}^{\infty}, p_n \geq 0 \ \ \forall n \geq 1, \sum\limits_{n=1}^{\infty} p_n = 1$, $\{p_n\}_{n=1}^{\infty}$ is the pmf of some rv X.

Note:
The issue of continuous rv's and probability density functions (pdfs) is more complicated. A rv $X : \Omega \rightarrow I\!R$ 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 $x_0 \in I\!R$ 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 $x \in I\!R$. ■

Definition 2.3.3:
A real–valued function F defined on $[a, b]$ is **absolutely continuous** on $[a, b]$ iff

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

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

■

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 (Ω, 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 $a, b \in I\!R, a < b$, then

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

exists and is well defined. ■

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

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

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

Proof:

Part (i): From Definition 2.3.5 above.

Part (ii): By Fundamental Theorem of Calculus. ■

Note:

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

However, note that every cdf F can be written as

$$F(x) = aF_d(x) + (1-a)F_c(x),\ 0 \le a \le 1,$$

where F_d is the cdf of a discrete rv and F_c is a continuous (but not necessarily absolute continuous) cdf.

Some authors, such as Marek Fisz *Wahrscheinlichkeitsrechnung und mathematische Statistik*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1989, are even more specific. There it is stated that every cdf F can be written as

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

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 $F'(x)$ 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

$$F(x) = \begin{cases} 0, & x < 0 \\ 1/2, & x = 0 \\ 1/2 + x/2, & 0 < x < 1 \\ 1, & x \ge 1 \end{cases}$$

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

■

Definition 2.3.8:

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

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

An Excursion into Logic

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

Basic Operators:

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

These operators are defined as follows:

A	B	$\neg A$	$\neg B$	$A \wedge B$	$A \vee B$	$A \Rightarrow B$	$A \Leftrightarrow B$	$A \oplus B$
1	1	0	0	1	1	1	1	0
1	0	0	1	0	1	0	0	1
0	1	1	0	0	1	1	0	1
0	0	1	1	0	0	1	1	0

Implication: (A implies B)

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

A	B	$A \Rightarrow B$	$\neg A$	$\neg B$	$\neg B \Rightarrow \neg A$	$\neg A \vee B$
1	1	1				
1	0	0				
0	1	1				
0	0	1				

Equivalence: (A is equivalent to B)

$A \Leftrightarrow B$ is equivalent to $(A \Rightarrow B) \wedge (B \Rightarrow A)$ is equivalent to $(\neg A \vee B) \wedge (A \vee \neg B)$:

A	B	$A \Leftrightarrow B$	$A \Rightarrow B$	$B \Rightarrow A$	$(A \Rightarrow B) \wedge (B \Rightarrow A)$	$\neg A \vee B$	$A \vee \neg B$	$(\neg A \vee B) \wedge (A \vee \neg B)$
1	1	1						
1	0	0						
0	1	0						
0	0	1						

Negations of Quantifiers:

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

$\neg\forall x \in X \; : \; B(x)$ is equivalent to ...

$\neg\exists x \in X \; : \; B(x)$ is equivalent to ...

$\exists x \in X \; \forall y \in Y \; : \; B(x, y)$ implies ...

2.4 Transformations of Random Variables

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

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

Proof: ■

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

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 $f_X(x) = 1/2 I_{[-1,1]}(x)$, 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, ■

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 (Ω, L, P) to $(I\!R, \mathcal{B})$, the **support** of X (or P) is any set $A \in L$ for which $P(A) = 1$. For a continuous random variable X with pdf f, we can think of the support of X as $\mathcal{X} = X^{-1}(\{x : f_X(x) > 0\})$. ■

Definition 2.4.5:

Let f be a real–valued function defined on $D \subseteq I\!R, D \in \mathcal{B}$. We say:

f is **non–decreasing** if $x < y \Longrightarrow f(x) \leq f(y) \;\; \forall x, y \in D$

f is **strictly non–decreasing** (or **increasing**) if $x < y \Longrightarrow f(x) < f(y) \;\; \forall x, y \in D$

f is **non–increasing** if $x < y \Longrightarrow f(x) \geq f(y) \;\; \forall x, y \in D$

f is **strictly non–increasing** (or **decreasing**) if $x < y \Longrightarrow f(x) > f(y) \;\; \forall x, y \in D$

f is **monotonic** on D if f is either increasing or decreasing and write $f \uparrow$ or $f \downarrow$. ■

Theorem 2.4.6:

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

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

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \mid \frac{d}{dy} g^{-1}(y) \mid \cdot I_{g(\mathcal{X})}(y).$$

Proof:

■

Note:

In Theorem 2.4.6, we can also write

$$f_Y(y) = \frac{f(x)}{\mid \frac{dg(x)}{dx} \mid}\Bigg|_{x=g^{-1}(y)}, y \in g(\mathcal{X})$$

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

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

(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,

$$f_Y(y) = \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \cdot \mid \frac{d}{dy} g_i^{-1}(y) \mid \cdot I_{\mathcal{Y}}(y)$$

■

Example 2.4.8:

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

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

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

Two possible approaches:

Method 1: cdfs

For $0 < y < 1$ we have

$$F_Y(y) \quad = \quad P_Y(Y \leq y)$$

$$
\begin{aligned}
&= P_X(\sin X \le y) \\
&= P_X([0 \le X \le \sin^{-1}(y)] \text{ or } [\pi - \sin^{-1}(y) \le X \le \pi]) \\
&= F_X(\sin^{-1}(y)) + (1 - F_X(\pi - \sin^{-1}(y)))
\end{aligned}
$$

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

$$
\begin{aligned}
f_Y(y) &= F_Y'(y) \\
&= f_X(\sin^{-1}(y))\frac{1}{\sqrt{1-y^2}} + (-1)f_X(\pi - \sin^{-1}(y))\frac{-1}{\sqrt{1-y^2}} \\
&= \frac{1}{\sqrt{1-y^2}}\left(f_X(\sin^{-1}(y)) + f_X(\pi - \sin^{-1}(y))\right) \\
&= \frac{1}{\sqrt{1-y^2}}\left(\frac{2(\sin^{-1}(y))}{\pi^2} + \frac{2(\pi - \sin^{-1}(y))}{\pi^2}\right) \\
&= \frac{1}{\pi^2\sqrt{1-y^2}}\ 2\pi \\
&= \frac{2}{\pi\sqrt{1-y^2}} \cdot I_{(0,1)}(y)
\end{aligned}
$$

<u>Method 2: Use of Theorem 2.4.7</u>

■

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:

$\blacksquare$

Note:

This proof also holds if there exist multiple intervals with $x_i < x_j$ and $F_X(x_i) = F_X(x_j)$, i.e., if the support of X is split in more than just 2 disjoint intervals. $\blacksquare$

3 Moments and Generating Functions

3.1 Expectation

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

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

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

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

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

■

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:

■

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

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

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

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:

$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + (E(X))^2) \\ &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - (E(X))^2 \end{aligned}$$

■

Theorem 3.1.7:

If $Var(X)$ exists and a and b are finite constants, then $Var(aX + b)$ exists and equals $a^2Var(X)$.

Proof:

Existence & Numerical Result:

$Var(aX + b) = E\left(((aX + b) - E(aX + b))^2\right)$ exists if $E\left(\mid ((aX + b) - E(aX + b))^2 \mid\right)$ exists.

It holds that

■

Theorem 3.1.8:

If the t^{th} absolute moment of a rv X exists for some $t > 0$, then all absolute moments of order $0 < s < t$ exist.

Proof:

Continuous case only:

■

Theorem 3.1.9:
If the t^{th} absolute moment of a rv X exists for some $t > 0$, then

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

Proof:
Continuous case only:

■

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

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

■

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

$\blacksquare$

Corollary 3.1.12:

$$E(\mid X \mid^s) = s \int_0^\infty y^{s-1} P(\mid X \mid > \ y) dy$$

Proof:

■

Proof (of Theorem 3.1.10):

■

Theorem 3.1.13:

Let X be a rv such that
$$\lim_{k\to\infty}\frac{P(\mid X\mid>\alpha k)}{P(\mid X\mid>k)}=0\ \ \forall\alpha>1.$$
Then, all moments of X exist.

Proof:

• For $\epsilon>0$, we select some k_0 such that
$$\frac{P(\mid X\mid>\alpha k)}{P(\mid X\mid>k)}<\epsilon\ \ \forall k\geq k_0.$$
• Select k_1 such that $P(\mid X\mid>k)<\epsilon\ \ \forall k\geq k_1$.

• Select $N=\max(k_0,k_1)$.

• If we have some fixed positive integer r:
$$\begin{aligned}\frac{P(\mid X\mid>\alpha^r k)}{P(\mid X\mid>k)} &= \frac{P(\mid X\mid>\alpha k)}{P(\mid X\mid>k)}\cdot\frac{P(\mid X\mid>\alpha^2 k)}{P(\mid X\mid>\alpha k)}\cdot\frac{P(\mid X\mid>\alpha^3 k)}{P(\mid X\mid>\alpha^2 k)}\cdot\ldots\cdot\frac{P(\mid X\mid>\alpha^r k)}{P(\mid X\mid>\alpha^{r-1}k)}\\ &= \frac{P(\mid X\mid>\alpha k)}{P(\mid X\mid>k)}\cdot\frac{P(\mid X\mid>\alpha\cdot(\alpha k))}{P(\mid X\mid>1\cdot(\alpha k))}\cdot\frac{P(\mid X\mid>\alpha\cdot(\alpha^2 k))}{P(\mid X\mid>1\cdot(\alpha^2 k))}\cdot\ldots\cdot\frac{P(\mid X\mid>\alpha\cdot(\alpha^{r-1}k))}{P(\mid X\mid>1\cdot(\alpha^{r-1}k))}\end{aligned}$$
• Note: Each of these r terms on the right side is $<\epsilon$ by our original statement of selecting some k_0 such that $\frac{P(|X|>\alpha k)}{P(|X|>k)}<\epsilon\ \ \forall k\geq k_0$ 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 $P(\mid X\mid>\alpha^r k)\leq\epsilon^r P(\mid X\mid>k)\leq\epsilon^{r+1}$ for $k\geq N$ (since in this case also $k\geq k_1$).

• For a fixed positive integer n:
$$E(\mid X\mid^n)\overset{Cor. 3.1.12}{=} n\cdot\int_0^\infty x^{n-1}P(\mid X\mid>x)dx\ =\ n\int_0^N x^{n-1}P(\mid X\mid>x)dx+n\int_N^\infty x^{n-1}P(\mid X\mid>x)dx$$
• We know that:
$$n\int_0^N x^{n-1}P(\mid X\mid>x)dx\leq\int_0^N nx^{n-1}dx=x^n\ |_0^N=N^n\quad<\infty$$
but is
$$n\int_N^\infty x^{n-1}P(\mid X\mid>x)dx\quad<\infty\quad ?$$
• To check the second part, we use:
$$\int_N^\infty x^{n-1}P(\mid X\mid>x)dx\ =\ \sum_{r=1}^{\infty}\int_{\alpha^{r-1}N}^{\alpha^r N}x^{n-1}P(\mid X\mid>x)dx$$

- We know that:

$$\int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1} P(\mid X \mid > x)dx \ \le \ \epsilon^r \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1}dx$$

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

- Since $(\alpha^{r-1}N)^{n-1} \le x^{n-1} \le (\alpha^r N)^{n-1} \ \ \forall x \in (\alpha^{r-1}N, \alpha^r N)$, we get:

$$\epsilon^r \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1}dx \ \ \le \ \ \epsilon^r(\alpha^r N)^{n-1} \int_{\alpha^{r-1}N}^{\alpha^r N} 1dx \ \ \le \ \ \epsilon^r(\alpha^r N)^{n-1}(\alpha^r N) \ \ = \ \ \epsilon^r(\alpha^r N)^n$$

- Now we go back to our original inequality:

$$\int_{N}^{\infty} x^{n-1} P(\mid X \mid > x)dx \ \le \ \sum_{r=1}^{\infty} \epsilon^r \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1}dx \ \ \le \ \ \sum_{r=1}^{\infty} \epsilon^r(\alpha^r N)^n \ \ = \ \ N^n \sum_{r=1}^{\infty} (\epsilon \cdot \alpha^n)^r$$

$$= \frac{N^n \epsilon \alpha^n}{1 - \epsilon \alpha^n} \text{ if } \epsilon\alpha^n < 1 \text{ or, equivalently, if } \epsilon < \frac{1}{\alpha^n}$$

- Since $\dfrac{N^n \epsilon \alpha^n}{1 - \epsilon \alpha^n}$ is finite, all moments $E(\mid X \mid^n)$ exist. ■

3.2 Moment Generating Functions

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

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

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

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

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

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

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.

■

Note:
We use the notation $\frac{\partial}{\partial t} f(x,t)$ for the partial derivative of f with respect to t and the notation $\frac{d}{dt} f(t)$ for the (ordinary) derivative of f with respect to t. ■

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

■

Note:
In the previous example, we made use of **L'Hospital's rule**. This rule gives conditions under which we can resolve indefinite expressions of the type "$\frac{\pm 0}{\pm 0}$" and "$\frac{\pm\infty}{\pm\infty}$".

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

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

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

(v) Indefinite expressions of the form "$0 \cdot \infty$" can be handled by rearranging them to "$\frac{0}{1/\infty}$" and $\lim\limits_{x \to -\infty} \dfrac{f(x)}{g(x)}$ can be handled by use of the rules for $\lim\limits_{x \to \infty} \dfrac{f(-x)}{g(-x)}$.

■

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 θ (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 θ.

Proof:

Uses the Fundamental Theorem of Calculus and the chain rule. ■

Theorem 3.2.5: **Lebesque's Dominated Convergence Theorem**

Let g be an integrable function such that $\int_{-\infty}^{\infty} g(x)dx < \infty$. If $\mid f_n \mid \leq g$ 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 θ, then

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

and

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

while

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

■

Theorem 3.2.6:

Let $f_n(x, \theta_0) = \frac{f(x, \theta_0 + \delta_n) - f(x, \theta_0)}{\delta_n}$ for some θ_0. Suppose there exists an integrable function $g(x)$ such that $\int_{-\infty}^{\infty} g(x)dx < \infty$ and $\mid f_n(x, \theta) \mid \leq g(x) \quad \forall x$, then

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

Usually, if f is differentiable for all θ, 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 θ. Suppose there exists an integrable function $g(x, \theta)$ such that $\int_{-\infty}^{\infty} g(x, \theta)dx < \infty$ and $\left| \frac{\partial}{\partial \theta} f(x, \theta) \mid_{\theta = \theta_0} \right| \leq g(x, \theta) \quad \forall x \; \forall \theta_0$ in some ϵ–neighborhood of θ, 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
$$\left| \frac{\partial}{\partial t} e^{tx} f_X(x) \mid_{t=t'} \right| = \mid x \mid e^{t'x} f_X(x) \text{ for } \mid t' - t \mid \leq \delta_0.$$

Choose t, δ_0 small enough such that $t + \delta_0 \in (-h, h)$ and $t - \delta_0 \in (-h, h)$. Then,
$$\left| \frac{\partial}{\partial t} e^{tx} f_X(x) \mid_{t=t'} \right| \leq g(x, t)$$

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

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

Suppose mgf $M_X(t)$ exists for $\mid t \mid \leq h$ for some $h > 1$. Then $\mid t+\delta_0+1 \mid < h$ and $\mid t-\delta_0-1 \mid < h$.

Since $\mid x \mid \leq e^{|x|} \;\; \forall x$, 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, $\displaystyle\int_0^\infty g(x,t)dx \leq M_X(t + \delta_0 + 1) < \infty$ and $\displaystyle\int_{-\infty}^0 g(x,t)dx \leq M_X(t - \delta_0 - 1) < \infty$ and, therefore, $\displaystyle\int_{-\infty}^\infty g(x)dx < \infty$.

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

If $h \leq 1$, we may need to check more carefully to see if the condition holds.

Note:
If $M_X(t)$ exists for $t \in (-h, h)$, then we have an infinite collection of moments.

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

Example 3.2.8:
Let X_1 and X_2 be rv's with pdfs
$$f_{X_1}(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x} \exp(-\frac{1}{2}(\log x)^2) \cdot I_{(0,\infty)}(x)$$

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

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

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

Theorem 3.2.9:

Let X and Y be 2 rv's with cdf's F_X and F_Y for which all moments exist.

(i) If F_X and F_Y have bounded support, then $F_X(u) = F_Y(u) \quad \forall u$ 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 $F_X(u) = F_Y(u) \quad \forall u$.

■

Note:

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

Theorem 3.2.10:

Suppose rv's $\{X_i\}_{i=1}^{\infty}$ have mgf's $M_{X_i}(t)$ and that $\lim\limits_{i \to \infty} M_{X_i}(t) = M_X(t) \quad \forall t \in (-h, h)$ 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 $\lim\limits_{i \to \infty} F_{X_i}(x) = F_X(x)$, 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:

■

3.3 Complex–Valued Random Variables and Characteristic Functions

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

Recall the following facts regarding complexd numbers:

$i^0 = +1; i = \sqrt{-1}; i^2 = -1; i^3 = -i; i^4 = +1$; etc.
in the planar Gauss'ian number plane it holds that $i = (0, 1)$

$z = a + ib = r(\cos\phi + i\sin\phi)$

$r = \mid z \mid = \sqrt{a^2 + b^2}$

$\tan\phi = \frac{b}{a}$

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

<u>Mathematical Operations on Complex Numbers:</u>

$z_1 \pm z_2 = (a_1 \pm a_2) + i(b_1 \pm b_2)$

$z_1 \cdot z_2 = r_1 r_2 e^{i(\phi_1+\phi_2)} = r_1 r_2(\cos(\phi_1 + \phi_2) + i\sin(\phi_1 + \phi_2))$

$\frac{z_1}{z_2} = \frac{r_1}{r_2} e^{i(\phi_1-\phi_2)} = \frac{r_1}{r_2}(\cos(\phi_1 - \phi_2) + i\sin(\phi_1 - \phi_2))$

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

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

$\ln z = \ln(a + ib) = \ln(\mid z \mid) + i\phi \pm ik \cdot 2\pi$ where $\phi = \arctan\frac{b}{a}$, $k = 0, \pm1, \pm2, \ldots$, and the main value is obtained for $k = 0$

<u>Note:</u>
Similar to real numbers, where we define $\sqrt{4} = 2$ while it holds that $2^2 = 4$ and $(-2)^2 = 4$, the n^{th} root and also the logarithm of complex numbers have one main value. However, if we read n^{th} root and logarithm as mappings where the inverse mappings (power and exponential function) yield the original values again, there exist additional solutions that produce the original values. For example, the main value of $\sqrt{-1}$ is i. However, it holds that $i^2 = -1$ and $(-i)^2 = (-1)^2 i^2 = i^2 = -1$. So, all solutions to $\sqrt{-1}$ are $\{i, -i\}$.

Conjugate Complex Numbers:

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

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

$z = \overline{z}$ iff $z \in I\!R$

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

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

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

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

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

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

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

Definition 3.3.1:
Let (Ω, L, P) be a probability space and X and Y real–valued rv's, i.e., $X, Y : (\Omega, L) \to (I\!R, \mathcal{B})$

(i) $Z = X + iY : (\Omega, L) \to (\mathbb{C}, \mathcal{B}_{\mathbb{C}})$ is called a **complex–valued random variable** ($\mathbb{C}$-rv).

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

■

Note:
$E(Z)$ exists iff $E(\mid X \mid)$ and $E(\mid Y \mid)$ exist. It also holds that if $E(Z)$ exists, then $\mid E(Z) \mid \leq E(\mid Z \mid)$ (see Homework). ■

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

Note:

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

(ii) $\Phi_X(t) = \sum_{x \in \mathcal{X}} e^{itx} P(X = x) = \sum_{x \in \mathcal{X}} \cos(tx) P(X = x) + i \sum_{x \in \mathcal{X}} \sin(tx) P(X = x)(x)$ 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 Φ_X be the characteristic function of a real–valued rv X. Then it holds:

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

(ii) $| \Phi_X(t) |\leq 1 \;\; \forall t \in I\!R$.

(iii) Φ_X is uniformly continuous, i.e., $\forall \epsilon > 0 \; \exists \delta > 0 \; \forall t_1, t_2 \in I\!R :| t_1 - t_2 |< \delta \Rightarrow| \Phi(t_1) - \Phi(t_2) |< \epsilon$.

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

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

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

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

Proof:

See Homework for parts (i), (ii), (iv), (v), (vi), and (vii).

Part (iii):

Known conditions:

■

Theorem 3.3.4: **Bochner's Theorem**

Let $\Phi : I\!R \to \mathbb{C}$ 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, Φ_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 Φ_X and let Φ_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 Φ_X. Let $a, b \in I\!R$, $a < b$. 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 Φ_X and Φ_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 Φ_X such that $\int_{-\infty}^{\infty} \mid \Phi_X(t) \mid dt < \infty$. 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 $\lim\limits_{i \to \infty} \Phi_{X_i}(t) = \Phi_X(t) \;\; \forall t \in (-h, h)$ for some $h > 0$ and $\Phi_X(t)$ is itself a characteristic function (of a rv X with cdf F_X), then $\lim\limits_{i \to \infty} F_{X_i}(x) = F_X(x)$ 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:

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

Proof:

■

Example 3.3.13:

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

It is

■

Note:

The restriction $\int_{-\infty}^{\infty} \mid \Phi_X(t) \mid dt < \infty$ 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,

- $X \sim \text{Dirac}(c)$:

$$\begin{aligned} \int_{-\infty}^{\infty} \mid \Phi_X(t) \mid dt &= \int_{-\infty}^{\infty} \mid e^{itc} \mid dt \\ &= \int_{-\infty}^{\infty} 1 dt \\ &= t \mid_{-\infty}^{\infty} \end{aligned}$$

which is undefined.

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

$$\begin{aligned} \int_{-\infty}^{\infty} \mid \Phi_X(t) \mid dt &= \int_{-\infty}^{\infty} \mid 1 + p(e^{it} - 1) \mid dt \\ &= \int_{-\infty}^{\infty} \mid pe^{it} - (p-1) \mid dt \\ &\geq \int_{-\infty}^{\infty} \mid \ \mid pe^{it} \mid - \mid (p-1) \mid \ \mid dt \\ &\geq \int_{-\infty}^{\infty} \mid pe^{it} \mid dt - \int_{-\infty}^{\infty} \mid (p-1) \mid dt \\ &= p \int_{-\infty}^{\infty} 1 dt - (1-p) \int_{-\infty}^{\infty} 1 \ dt \end{aligned}$$

$$
\begin{aligned}
&= (2p-1)\int_{-\infty}^{\infty} 1\ dt \\
&= (2p-1)t \mid_{-\infty}^{\infty}
\end{aligned}
$$

which is undefined for $p \neq 1/2$.
If $p = 1/2$, we have

$$
\begin{aligned}
\int_{-\infty}^{\infty} \mid pe^{it} - (p-1) \mid dt &= 1/2 \int_{-\infty}^{\infty} \mid e^{it} + 1 \mid dt \\
&= 1/2 \int_{-\infty}^{\infty} \mid \cos t + i \sin t + 1 \mid dt \\
&= 1/2 \int_{-\infty}^{\infty} \sqrt{(\cos t + 1)^2 + (\sin t)^2}\ dt \\
&= 1/2 \int_{-\infty}^{\infty} \sqrt{\cos^2 t + 2\cos t + 1 + \sin^2 t}\ dt \\
&= 1/2 \int_{-\infty}^{\infty} \sqrt{2 + 2\cos t}\ dt
\end{aligned}
$$

which also does not exist.

- Otherwise, $X \sim N(0,1)$:

$$
\begin{aligned}
\int_{-\infty}^{\infty} \mid \Phi_X(t) \mid dt &= \int_{-\infty}^{\infty} \exp(-t^2/2) dt \\
&= \sqrt{2\pi} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt \\
&= \sqrt{2\pi} \\
&< \infty
\end{aligned}
$$

■

3.4 Probability Generating Functions

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

Definition 3.4.1:

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

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

■

Theorem 3.4.2:

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

Proof:

■

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:

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

■

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:

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

■

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

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

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

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

Proof:
Homework ■

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

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

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

It is

■

3.5 Moment Inequalities

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

Theorem 3.5.1:
Let $h(X)$ be a non–negative Borel–measurable function of a rv X. If $E(h(X))$ exists, then it holds:

$$P(h(X) \geq \epsilon) \leq \frac{E(h(X))}{\epsilon} \quad \forall \epsilon > 0$$

Proof:
Continuous case only:

■

Corollary 3.5.2: **Markov's Inequality**
Let $h(X) = \mid X \mid^r$ and $\epsilon = k^r$ where $r > 0$ and $k > 0$. If $E(\mid X \mid^r)$ exists, then it holds:

$$P(\mid X \mid \geq k) \leq \frac{E(\mid X \mid^r)}{k^r}$$

Proof:

■

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

$$P(\mid X - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}$$

Proof:

$\blacksquare$

Note:
For $k = 2$, it follows from Corollary 3.5.3 that

$$P(\mid X - \mu \mid < 2\sigma) \geq 1 - \frac{1}{2^2} = 0.75,$$

no matter what the distribution of X is. Unfortunately, this is not very precise for many distributions, e.g., the Normal distribution, where it holds that $P(\mid X - \mu \mid < 2\sigma) \approx 0.95$. $\blacksquare$

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

$$(\beta_{k-1})^{\frac{1}{k-1}} \leq (\beta_k)^{\frac{1}{k}},$$

i.e., $(E(\mid X \mid^{k-1}))^{\frac{1}{k-1}} \leq (E(\mid X \mid^k))^{\frac{1}{k}}$.

Proof:
Continuous case only:

$\blacksquare$

<u>Note:</u>

- It follows from Theorem 3.5.4 that

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

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

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

and

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

Therefore, equality holds in Theorem 3.5.4.

$\blacksquare$

4 Random Vectors

4.1 Joint, Marginal, and Conditional Distributions

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

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

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

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

■

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

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

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

Note:

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

(ii) $\lim\limits_{\underline{x} \rightarrow \underline{\infty}} F(\underline{x}) = \lim\limits_{x_1 \rightarrow \infty, \ldots x_n \rightarrow \infty} F(\underline{x}) = 1$ and $\lim\limits_{x_k \rightarrow -\infty} F(\underline{x}) = 0 \quad \forall x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n \in I\!R$.

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 $F(\underline{x}) = F(x_1, \ldots, x_n)$ is the joint cdf of some n–rv $\underline{X}$ 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) $\forall \underline{x} \in I\!R^n \ \forall \epsilon_i > 0, i = 1, \ldots n$, the following inequality holds:

$$\begin{aligned}
F(\underline{x} + \underline{\epsilon}) \quad & - \quad \sum_{i=1}^{n} F(x_1 + \epsilon_1, \ldots, x_{i-1} + \epsilon_{i-1}, x_i, x_{i+1} + \epsilon_{i+1}, \ldots, x_n + \epsilon_n) \\
& + \quad \sum_{1 \leq i < j \leq n} F(x_1 + \epsilon_1, \ldots, x_{i-1} + \epsilon_{i-1}, x_i, x_{i+1} + \epsilon_{i+1}, \ldots, \\
& \qquad\qquad x_{j-1} + \epsilon_{j-1}, x_j, x_{j+1} + \epsilon_{j+1}, \ldots, x_n + \epsilon_n) \\
& \mp \quad \ldots \\
& + \quad (-1)^n F(\underline{x}) \\
& \geq \quad 0
\end{aligned}$$

■

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

$$\begin{aligned}
& P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = \\
& \quad P(X \leq x_2, Y \leq y_2) - P(X \leq x_1, Y \leq y_2) - P(X \leq x_2, Y \leq y_1) + P(X \leq x_1, Y \leq y_1) \geq 0
\end{aligned}$$

■

Note:
We will restrict ourselves to $n = 2$ for most of the next Definitions and Theorems but those can be easily generalized to $n > 2$. The term **bivariate rv** is often used to refer to a 2–rv and **multivariate rv** is used to refer to an n–rv, $n \geq 2$. ■

Definition 4.1.4:
A 2–rv (X, Y) is **discrete** if there exists a countable collection $\mathcal{X}$ of pairs (x_i, y_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, $\sum\limits_{i,j} p_{ij} = 1$ and $\{p_{ij}\}$ is the **joint probability mass function** (**joint pmf**) of (X, Y). ■

Definition 4.1.5:
Let (X, Y) be a discrete 2–rv with joint pmf $\{p_{ij}\}$. Define

$$p_{i\cdot} = \sum_{j=1}^{\infty} p_{ij} = \sum_{j=1}^{\infty} P(X = x_i, Y = y_j) = P(X = x_i)$$

and

$$p_{\cdot j} = \sum_{i=1}^{\infty} p_{ij} = \sum_{i=1}^{\infty} P(X = x_i, Y = y_j) = P(Y = y_j).$$

Then $\{p_{i\cdot}\}$ is called the **marginal probability mass function** (**marginal pmf**) of X and $\{p_{\cdot j}\}$ 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

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v)\ dv\ du \quad \forall (x, y) \in I\!R^2$$

where F is the joint cdf of (X, Y). We call f the **joint probability density function** (**joint pdf**) of (X, Y). ■

Note:
If F is continuous at (x, y), then

$$\frac{\partial^2 F(x, y)}{\partial x\ \partial y} = f(x, y).$$

■

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

Note:

(i)

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y) dy \right) dx = F(\infty, \infty) = 1 =$$

$$\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x, y) dx \right) dy = \int_{-\infty}^{\infty} f_Y(y) dy$$

and $f_X(x) \geq 0 \quad \forall x \in I\!R$ and $f_Y(y) \geq 0 \quad \forall y \in I\!R$.

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

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

■

Note:

In Definition 1.4.1, we defined conditional probability distributions in some probability space (Ω, 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 $P(Y = y_j) = p_{\cdot j} > 0$, then the **conditional probability mass function** (**conditional pmf**) of X given $Y = y_j$ (for fixed j) is defined as

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

■

Note:

For a continuous 2–rv (X, Y) with pdf f, $P(X \leq x \mid Y = y)$ is not defined. Let $\epsilon > 0$ and suppose that $P(y - \epsilon < Y \leq y + \epsilon) > 0$. For every x and every interval $(y - \epsilon, y + \epsilon]$, consider the conditional probability of $X \leq x$ given $Y \in (y - \epsilon, y + \epsilon]$. We have

$$P(X \leq x \mid y - \epsilon < Y \leq y + \epsilon) = \frac{P(X \leq x, y - \epsilon < Y \leq y + \epsilon)}{P(y - \epsilon < Y \leq y + \epsilon)}$$

which is well–defined if $P(y - \epsilon < Y \leq y + \epsilon) > 0$ holds.

So, when does

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

exist? See the next definition. ■

Definition 4.1.10:

The **conditional cumulative distribution function** (**conditional cdf**) of a rv X given that $Y = y$ is defined to be

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

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

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

■

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

$$\begin{aligned} F_{X|Y}(x \mid y) &= \lim_{\epsilon \to 0+} \frac{P(X \leq x, Y \in (y - \epsilon, y + \epsilon])}{P(Y \in (y - \epsilon, y + \epsilon])} \\ &= \lim_{\epsilon \to 0+} \left(\frac{\frac{1}{2\epsilon} \int_{-\infty}^{x} \int_{y-\epsilon}^{y+\epsilon} f_{X,Y}(u, v) dv \; du}{\frac{1}{2\epsilon} \int_{y-\epsilon}^{y+\epsilon} f_Y(v) dv} \right) \\ &= \frac{\int_{-\infty}^{x} f_{X,Y}(u, y) du}{f_Y(y)} \\ &= \int_{-\infty}^{x} \frac{f_{X,Y}(u, y)}{f_Y(y)} du. \end{aligned}$$

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:

$\blacksquare$

4.2 Independent Random Variables

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

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

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

We can show

(i) f_α 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

$$F_{X,Y}(x, y) = F_X(x) F_Y(y) \quad \forall (x, y) \in I\!R^2.$$

■

Lemma 4.2.3:
If X and Y are independent, $a, b, c, d \in I\!R$, 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: ■

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

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

■

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

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

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

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

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

where $f_{\underline{X}}$ is the joint pdf and $f_{X_1}, \ldots, f_{X_n}$ are the marginal pdfs of $\underline{X}$.

Proof:

(i) Discrete case:

■

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 $A_i \in \mathcal{B}$ (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: ■

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 $F_X(x) = F_Y(x)\ \ \forall x \in I\!\!R$ 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 defintion of independence to 2 random vectors $\underline{X}^{n \times 1}$ and $\underline{Y}^{n \times 1}$: $\underline{X}$ and $\underline{Y}$ are independent iff $F_{\underline{X},\underline{Y}}(\underline{x}, \underline{y}) = F_{\underline{X}}(\underline{x}) F_{\underline{Y}}(\underline{y}) \ \ \forall \underline{x}, \underline{y} \in I\!R^n$.

This does not mean that the components X_i of $\underline{X}$ or the components Y_i of $\underline{Y}$ 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 $f(\underline{X})$ and $g(\underline{Y})$ are independent. ■

Corollary 4.2.11: (to Factorization Theorem 4.2.5)
If X and Y are independent rv's, then

$$F_{X \mid Y}(x \mid y) = F_X(x) \ \ \forall x,$$

and

$$F_{Y \mid X}(y \mid x) = F_Y(y) \ \ \forall y.$$

■

4.3 Functions of Random Vectors

(Based on Casella/Berger, Sections 4.3 & 4.6)

Theorem 4.3.1:
If X and Y are rv's on $(\Omega, L, P) \to I\!R$, then

(i) $X \pm Y$ is a rv.

(ii) XY is a rv.

(iii) If $\{\omega : Y(\omega) = 0\} = \emptyset$, then $\frac{X}{Y}$ is a rv.

■

Theorem 4.3.2:
Let $X_1, \ldots, X_n$ be rv's on $(\Omega, L, P) \to I\!R$. 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

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

and

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

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

$$F_{MAX_n}(z) = ...$$

and

$$F_{MIN_n}(z) = ...$$

If F_X is absolutely continuous with pdf f_X, then the pdfs of MAX_n and MIN_n are

$$f_{MAX_n}(z) = ...$$

and

$$f_{MIN_n}(z) = ...$$

for all continuity points of F_X. ■

Note:
Using Theorem 4.3.2, it is easy to derive the joint cdf and pdf of MAX_n and MIN_n for iid rv's $\{X_1, \ldots, X_n\}$. For example, if the X_i's are iid with cdf F_X and pdf f_X, then the joint pdf of MAX_n and MIN_n is

$$f_{MAX_n, MIN_n}(x, y) = \begin{cases} 0, & x \leq y \\ n(n-1) \cdot (F_X(x) - F_X(y))^{n-2} \cdot f_X(x) f_X(y), & x > y \end{cases}$$

However, note that MAX_n and MIN_n are not independent. ■

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

Theorem 4.3.3:
If $g : I\!R^n \to I\!R^m$ is a Borel–measurable function (i.e., $\forall B \in \mathcal{B}^m : g^{-1}(B) \in \mathcal{B}^n$) and if $\underline{X} = (X_1, \ldots, X_n)$ is an n–rv, then $g(\underline{X})$ is an m–rv.

Proof:
If $B \in \mathcal{B}^m$, then $\{\omega : g(\underline{X}(\omega)) \in B\} = \{\omega : \underline{X}(\omega) \in g^{-1}(B)\} \in \mathcal{B}^n$. ■

Question: How do we handle more general transformations of $\underline{X}$?

Discrete Case:
Let $\underline{X} = (X_1, \ldots, X_n)$ be a discrete n–rv and $\mathcal{X} \subset I\!R^n$ be the countable support of $\underline{X}$, i.e., $P(\underline{X} \in \mathcal{X}) = 1$ and $P(\underline{X} = \underline{x}) > 0 \ \ \forall \underline{x} \in \mathcal{X}$.

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

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

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

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 $V = Y + 1$.

■

Continuous Case:
Let $\underline{X} = (X_1, \ldots, X_n)$ be a continuous n–rv with joint cdf $F_{\underline{X}}$ and joint pdf $f_{\underline{X}}$.

Let

$$\underline{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} = g(\underline{X}) = \begin{pmatrix} g_1(\underline{X}) \\ \vdots \\ g_n(\underline{X}) \end{pmatrix},$$

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

If $B \in \mathcal{B}^n$, then

$$P(\underline{U} \in B) = P(\underline{X} \in g^{-1}(B)) = \underset{g^{-1}(B)}{\int \ldots \int} f_{\underline{X}}(\underline{x}) d(\underline{x}) = \underset{g^{-1}(B)}{\int \ldots \int} f_{\underline{X}}(\underline{x}) \prod_{i=1}^{n} dx_i$$

where $g^{-1}(B) = \{\underline{x} = (x_1, \ldots, x_n) \in I\!R^n : g(\underline{x}) \in B\}$.

Suppose we define B as the half–infinite n–dimensional interval

$$B_{\underline{u}} = \{(\tilde{u}_1, \ldots, \tilde{u}_n) : -\infty < \tilde{u}_i < u_i \;\; \forall i = 1, \ldots, n\}$$

for any $\underline{u} \in I\!R^n$. Then the joint cdf of $\underline{U}$ is

$$G_{\underline{U}}(\underline{u}) = P(\underline{U} \in B_{\underline{u}}) = P(g_1(\underline{X}) \leq u_1, \ldots, g_n(\underline{X}) \leq u_n) = \underset{g^{-1}(B_{\underline{u}})}{\int \ldots \int} f_{\underline{X}}(\underline{x}) d(\underline{x}).$$

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

Under certain conditions, we can write $f_{\underline{U}}$ in terms of the original pdf $f_{\underline{X}}$ of $\underline{X}$ as stated in the next Theorem:

Theorem 4.3.5: **Multivariate Transformation**
Let $\underline{X} = (X_1, \ldots, X_n)$ be a continuous n–rv with joint pdf $f_{\underline{X}}$.

(i) Let

$$\underline{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} = g(\underline{X}) = \begin{pmatrix} g_1(\underline{X}) \\ \vdots \\ g_n(\underline{X}) \end{pmatrix},$$

(i.e., $U_i = g_i(\underline{X})$) 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 $x_i = h_i(\underline{u}) = h_i(u_1, \ldots, u_n), i = 1, \ldots, n$, over the range of the transformation g.

(ii) Assume both g and h are continuous.

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

(iv) Assume that the Jacobian of the inverse transformation

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

is different from 0 for all $\underline{u}$ in the range of g.

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

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

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

$$B_{\underline{u}} = \{(\tilde{u}_1, \ldots, \tilde{u}_n) : -\infty < \tilde{u}_i < u_i \;\; \forall i = 1, \ldots, n\}.$$

Then,

$$\begin{aligned} G_{\underline{U}}(\underline{u}) &= \underset{g^{-1}(B_{\underline{u}})}{\int \ldots \int} f_{\underline{X}}(\underline{x}) d(\underline{x}) \\ &= \underset{B_{\underline{u}}}{\int \ldots \int} f_{\underline{X}}(h_1(\underline{u}), \ldots, h_n(\underline{u})) \mid J \mid d(\underline{u}) \end{aligned}$$

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

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

(i) Let

$$\underline{U} = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix} = g(\underline{X}) = \begin{pmatrix} g_1(\underline{X}) \\ \vdots \\ g_n(\underline{X}) \end{pmatrix},$$

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

(ii) Let $\mathcal{X} = \{\underline{x} : f_{\underline{X}}(\underline{x}) > 0\}$ be the support of $\underline{X}$.

(iii) Suppose that for each $\underline{u} \in B = \{\underline{u} \in I\!R^n : \underline{u} = g(\underline{x})$ for some $\underline{x} \in \mathcal{X}\}$ there is a finite number $k = k(\underline{u})$ 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) $\underline{U} = g(\underline{X})$ is a 1–to–1–mapping from $\mathcal{X}_l$ onto B for all $l = 1, \ldots, k$, with inverse transformation $h_l(\underline{u}) = \begin{pmatrix} h_{l1}(\underline{u}) \\ \vdots \\ h_{ln}(\underline{u}) \end{pmatrix}$, $\underline{u} \in B$, i.e., for each $\underline{u} \in B$, $h_l(\underline{u})$ is the unique $\underline{x} \in \mathcal{X}_l$ such that $\underline{u} = g(\underline{x})$.

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

(vi) Assume the Jacobian of each of the inverse transformations

$$J_l = \begin{vmatrix} \frac{\partial x_1}{\partial u_1} & \ldots & \frac{\partial x_1}{\partial u_n} \\ \vdots & & \vdots \\ \frac{\partial x_n}{\partial u_1} & \ldots & \frac{\partial x_n}{\partial u_n} \end{vmatrix} = \begin{vmatrix} \frac{\partial h_{l1}}{\partial u_1} & \ldots & \frac{\partial h_{l1}}{\partial u_n} \\ \vdots & & \vdots \\ \frac{\partial h_{ln}}{\partial u_1} & \ldots & \frac{\partial h_{ln}}{\partial u_n} \end{vmatrix}, \quad l = 1, \ldots, k,$$

is different from 0 for all $\underline{u}$ in the range of g.

Then the joint pdf of $\underline{U}$ is given by

$$f_{\underline{U}}(\underline{u}) = \sum_{l=1}^{k} \mid J_l \mid f_{\underline{X}}(h_{l1}(\underline{u}), \ldots, h_{ln}(\underline{u})).$$

■

Example 4.3.7:

Let X, Y be iid $\sim N(0,1)$. Define

$$U = g_1(X,Y) = \begin{cases} \frac{X}{Y}, & Y \neq 0 \\ 0, & Y = 0 \end{cases}$$

and

$$V = g_2(X,Y) = \mid Y \mid .$$

■

4.4 Order Statistics

(Based on Casella/Berger, Section 5.4)

Definition 4.4.1:

Let $(X_1, \ldots, X_n)$ be an n–rv. The k^{th} **order statistic** $X_{(k)}$ is the k^{th} smallest of the $X_i's$, i.e., $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(2)} = \min\{\{X_1, \ldots, X_n\} \setminus X_{(1)}\}$, …, $X_{(n)} = \max\{X_1, \ldots, X_n\}$. It is $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ and $\{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}$ is the set of order statistics for $(X_1, \ldots, X_n)$. ■

Note:

As shown in Theorem 4.3.2, $X_{(1)}$ and $X_{(n)}$ are rv's. This result will be extended in the following Theorem:

Theorem 4.4.2:

Let $(X_1, \ldots, X_n)$ be an n–rv. Then the k^{th} order statistic $X_{(k)}$, $k = 1, \ldots, n$, is also a rv. ■

Theorem 4.4.3:

Let $X_1, \ldots, X_n$ be continuous iid rv's with pdf f_X. The joint pdf of $X_{(1)}, \ldots, X_{(n)}$ is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n! \prod_{i=1}^{n} f_X(x_i), & x_1 \leq x_2 \leq \ldots \leq x_n \\ 0, & \text{otherwise} \end{cases}$$

Proof:

For the case $n = 3$, look at the following scenario how X_1, X_2, and X_3 can be possibly ordered to yield $X_{(1)} < X_{(2)} < X_{(3)}$. Columns represent $X_{(1)}, X_{(2)}$, and $X_{(3)}$. Rows represent X_1, X_2, and X_3:

$$k = 1: \quad X_1 < X_2 < X_3 \quad : \quad \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} X_{(1)} X_{(2)} X_{(3)} \\ \left|\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right| \end{array}$$

$$k = 2: \quad X_1 < X_3 < X_2 \quad : \quad \left|\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array}\right|$$

$$k = 3: \quad X_2 < X_1 < X_3 \quad : \quad \left|\begin{array}{ccc} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{array}\right|$$

$$k = 4: \quad X_2 < X_3 < X_1 \qquad : \qquad \begin{vmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix}$$

$$k = 5: \quad X_3 < X_1 < X_2 \qquad : \qquad \begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix}$$

$$k = 6: \quad X_3 < X_2 < X_1 \qquad : \qquad \begin{vmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{vmatrix}$$

For $n = 3$, there are $3! = 6$ possible arrangements.

For example, if $k = 2$, we have

$$X_1 < X_3 < X_2$$

with corresponding inverse

$$X_1 = X_{(1)}, \quad X_2 = X_{(3)}, \quad X_3 = X_{(2)}$$

and

$$J_2 = \det \begin{pmatrix} \dfrac{\partial x_1}{\partial x_{(1)}} & \dfrac{\partial x_1}{\partial x_{(2)}} & \dfrac{\partial x_1}{\partial x_{(3)}} \\ \dfrac{\partial x_2}{\partial x_{(1)}} & \dfrac{\partial x_2}{\partial x_{(2)}} & \dfrac{\partial x_2}{\partial x_{(3)}} \\ \dfrac{\partial x_3}{\partial x_{(1)}} & \dfrac{\partial x_3}{\partial x_{(2)}} & \dfrac{\partial x_3}{\partial x_{(3)}} \end{pmatrix} = \det \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

with $\mid J_2 \mid = 1$.

$\blacksquare$

Theorem 4.4.4: Let $X_1, \ldots, X_n$ be continuous iid rv's with pdf f_X and cdf F_X. Then the following holds:

(i) The marginal pdf of $X_{(k)}$, $k = 1, \ldots, n$, is

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}(F_X(x))^{k-1}(1 - F_X(x))^{n-k} f_X(x).$$

(ii) The joint pdf of $X_{(j)}$ and $X_{(k)}$, $1 \leq j < k \leq n$, is

$$\begin{aligned} f_{X_{(j)},X_{(k)}}(x_j, x_k) &= \frac{n!}{(j-1)!(k-j-1)!(n-k)!} \times \\ & (F_X(x_j))^{j-1}(F_X(x_k) - F_X(x_j))^{k-j-1}(1 - F_X(x_k))^{n-k} f_X(x_j) f_X(x_k) \end{aligned}$$

if $x_j < x_k$ and 0 otherwise.

■

4.5 Multivariate Expectation

(Based on Casella/Berger, Sections 4.2, 4.6 & 4.7)

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

Definition 4.5.1:

If $m = 1$, i.e., g is univariate, we define the following:

(i) Let $\underline{X}$ be discrete with joint pmf $p_{i_1,\ldots,i_n} = P(X_1 = x_{i_1}, \quad \ldots, \quad X_n = x_{i_n})$. If $\sum\limits_{i_1,\ldots,i_n} p_{i_1,\ldots,i_n} \cdot \mid g(x_{i_1}, \ldots, x_{i_n}) \mid < \infty$, we define $E(g(\underline{X})) = \sum\limits_{i_1,\ldots,i_n} p_{i_1,\ldots,i_n} \cdot g(x_{i_1}, \ldots, x_{i_n})$ and this value exists.

(ii) Let $\underline{X}$ be continuous with joint pdf $f_{\underline{X}}(\underline{x})$. If $\int_{I\!R^n} \mid g(\underline{x}) \mid f_{\underline{X}}(\underline{x}) \, d\underline{x} < \infty$, we define $E(g(\underline{X})) = \int_{I\!R^n} g(\underline{x}) f_{\underline{X}}(\underline{x}) d\underline{x}$ 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 $I\!R^n \to I\!R^n$, then

$$E(\underline{X}) = \begin{pmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$$

provided that $E(\mid X_i \mid) < \infty \quad \forall i = 1, \ldots, n$.

Similarly, provided that all expectations exist, we get for the variance–covariance matrix:

$$Var(\underline{X}) = \Sigma_{\underline{X}} = E((\underline{X} - E(\underline{X})) \; (\underline{X} - E(\underline{X}))')$$

with $(i, j)^{th}$ component

$$E((X_i - E(X_i)) \; (X_j - E(X_j))) = Cov(X_i, X_j)$$

and with $(i, i)^{th}$ component

$$E((X_i - E(X_i)) \; (X_i - E(X_i))) = Var(X_i) = \sigma_i^2.$$

The correlation ρ_{ij} of X_i and X_j is defined as

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sigma_i \sigma_j}.$$

Joint higher–order moments can be defined similarly when needed. ■

Note:
We are often interested in (weighted) sums of rv's or products of rv's and their expectations. This will be addressed in the next two Theorems:

Theorem 4.5.2:
Let $X_i, i = 1, \ldots, n$, be rv's such that $E(\mid X_i \mid) < \infty$. Let $a_1, \ldots, a_n \in I\!R$ and define $S = \sum_{i=1}^{n} a_i X_i$. Then it holds that $E(\mid S \mid) < \infty$ and

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

Proof:
Continuous case only: ■

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

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

■

Theorem 4.5.3:
Let $X_i, i = 1, \ldots, n$, be independent rv's such that $E(\mid X_i \mid) < \infty$. 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, $f_{\underline{X}}(\underline{x}) = \prod_{i=1}^{n} f_{X_i}(x_i)$, and by Theorem 4.2.7, $g_i(X_i), i = 1, \ldots, n$, are also independent. Therefore, ■

Corollary 4.5.4:

If X, Y are independent, then $Cov(X, Y) = 0$. ■

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:

"$\Longrightarrow$": It follows from Theorem 4.5.3 and the independence of X and Y that

$$E(g_1(X)g_2(Y)) = E(g_1(X)) \cdot E(g_2(Y)).$$

"$\Longleftarrow$":

■

Definition 4.5.6:

The $(i_1^{th}, i_2^{th}, \ldots, i_n^{th})$ **multi–way moment** of $\underline{X} = (X_1, \ldots, X_n)$ is defined as

$$m_{i_1 i_2 \ldots i_n} = E(X_1^{i_1} X_2^{i_2} \ldots X_n^{i_n})$$

if it exists.

The $(i_1^{th}, i_2^{th}, \ldots, i_n^{th})$ **multi–way central moment** of $\underline{X} = (X_1, \ldots, X_n)$ is defined as

$$\mu_{i_1 i_2 \ldots i_n} = E(\prod_{j=1}^{n} (X_j - E(X_j))^{i_j})$$

if it exists. ■

<u>Note:</u>
If we set $i_r = i_s = 1$ and $i_j = 0 \ \ \forall j \neq r, s$ in Definition 4.5.6, we get

$$\mu_{\substack{0 \ldots 0 & 1 & 0 \ldots 0 & 1 & 0 \ldots 0 \\ & \uparrow & & \uparrow & \\ & r & & s & }} = \mu_{rs} = Cov(X_r, X_s).$$

■

Theorem 4.5.7: **Cauchy–Schwarz–Inequality**

Let X, Y be 2 rv's with finite variance. Then it holds:

(i) $Cov(X, Y)$ 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: $Var(X),\ Var(Y) < \infty$. Then also $E(X^2),\ E(X),\ E(Y^2),\ E(Y) < \infty$.

Result used in proof:

■

4.6 Multivariate Generating Functions

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

Definition 4.6.1:

Let $\underline{X}' = (X_1, \ldots, X_n)$ be an n–rv. We define the **multivariate moment generating function** (mmgf) of $\underline{X}$ as

$$M_{\underline{X}}(\underline{t}) = E(e^{\underline{t}'\underline{X}}) = E(\exp\left(\sum_{i=1}^{n} t_i X_i\right))$$

if this expectation exists for $\mid \underline{t} \mid = \sqrt{\sum_{i=1}^{n} t_i^2} < h$ for some $h > 0$. ■

Definition 4.6.2:

Let $\underline{X}' = (X_1, \ldots, X_n)$ be an n–rv. We define the **n–dimensional characteristic function** $\Phi_{\underline{X}} : \mathbb{R}^n \to \mathbb{C}$ of $\underline{X}$ as

$$\Phi_{\underline{X}}(\underline{t}) = E(e^{i\underline{t}'\underline{X}}) = E(\exp\left(i \sum_{j=1}^{n} t_j X_j\right)).$$

■

Note:

(i) $\Phi_{\underline{X}}(\underline{t})$ exists for any real–valued n–rv.

(ii) If $M_{\underline{X}}(\underline{t})$ exists, then $\Phi_{\underline{X}}(\underline{t}) = M_{\underline{X}}(i\underline{t})$.

■

Theorem 4.6.3:

(i) If $M_{\underline{X}}(\underline{t})$ exists, it is unique and uniquely determines the joint distribution of $\underline{X}$. $\Phi_{\underline{X}}(\underline{t})$ is also unique and uniquely determines the joint distribution of $\underline{X}$.

(ii) $M_{\underline{X}}(\underline{t})$ (if it exists) and $\Phi_{\underline{X}}(\underline{t})$ uniquely determine all marginal distributions of $\underline{X}$, i.e., $M_{X_i}(t_i) = M_{\underline{X}}(\underline{0}, t_i, \underline{0})$ and and $\Phi_{X_i}(t_i) = \Phi_{\underline{X}}(\underline{0}, t_i, \underline{0})$.

(iii) Joint moments of all orders (if they exist) can be obtained as

$$m_{i_1 \ldots i_n} = \frac{\partial^{i_1+i_2+\ldots+i_n}}{\partial t_1^{i_1} \partial t_2^{i_2} \ldots \partial t_n^{i_n}} M_{\underline{X}}(\underline{t}) \bigg|_{\underline{t}=\underline{0}} = E(X_1^{i_1} X_2^{i_2} \ldots X_n^{i_n})$$

if the mmgf exists and

$$m_{i_1 \ldots i_n} = \frac{1}{i^{i_1+i_2+\ldots+i_n}} \; \frac{\partial^{i_1+i_2+\ldots+i_n}}{\partial t_1^{i_1} \partial t_2^{i_2} \ldots \partial t_n^{i_n}} \Phi_{\underline{X}}(\underline{0}) = E(X_1^{i_1} X_2^{i_2} \ldots X_n^{i_n}).$$

(iv) $X_1, \ldots, X_n$ are independent rv's iff

$$M_{\underline{X}}(t_1, \ldots, t_n) = M_{\underline{X}}(t_1, \underline{0}) \cdot M_{\underline{X}}(0, t_2, \underline{0}) \cdot \ldots \cdot M_{\underline{X}}(\underline{0}, t_n) \ \ \forall t_1, \ldots, t_n \in I\!R,$$

given that $M_{\underline{X}}(\underline{t})$ exists.

Similarly, $X_1, \ldots, X_n$ are independent rv's iff

$$\Phi_{\underline{X}}(t_1, \ldots, t_n) = \Phi_{\underline{X}}(t_1, \underline{0}) \cdot \Phi_{\underline{X}}(0, t_2, \underline{0}) \cdot \ldots \cdot \Phi_{\underline{X}}(\underline{0}, t_n) \ \ \forall t_1, \ldots, t_n \in I\!R.$$

■

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

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(a_i t) \qquad \text{[Note: } t\text{]}$$

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

$$\Phi_Y(t) = \prod_{j=1}^n \Phi_{X_j}(a_j t) \qquad \text{[Note: } t\text{]}$$

(iii) If mgf's $M_{X_1}(t), \ldots, M_{X_n}(t)$ exist, then the mmgf of $\underline{X}$ is

$$M_{\underline{X}}(\underline{t}) = \prod_{i=1}^n M_{X_i}(t_i) \qquad \text{[Note: } t_i\text{]}$$

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

(iv) The n–dimensional characteristic function of $\underline{X}$ is

$$\Phi_{\underline{X}}(\underline{t}) = \prod_{j=1}^n \Phi_{X_j}(t_j). \qquad \text{[Note: } t_j\text{]}$$

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:

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

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: ■

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? ■

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

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 proof this Theorem. ■

4.7 Conditional Expectation

(Based on Casella/Berger, Section 4.4)

In Section 4.1, we established that the conditional pmf of X given $Y = y_j$ (for $P_Y(y_j) > 0$) is a pmf. For continuous rv's X and Y, when $f_Y(y) > 0$, $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, 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 (Ω, L, P). Let h be a Borel–measurable function. 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

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

■

Note:
Depending on the source, two different definitions of the conditional expectation exist: (i) Casella and Berger (2002, p. 150), Miller and Miller (1999, p. 161), and Rohatgi and Saleh (2001, p. 165) do **not require** that $E(h(X))$ exists. (ii) Rohatgi (1976, p. 168) and Bickel and Doksum (2001, p. 479) **require** that $E(h(X))$ exists.

In case of the rv's X and Y with joint pdf

$$f_{X,Y}(x,y) = xe^{-x(y+1)} I_{[0,\infty)}(x) I_{[0,\infty)}(y),$$

it holds in case (i) that $E(Y \mid X) = \frac{1}{X}$ even though $E(Y)$ does not exist (see Rohatgi and Saleh 2001, p. 168, for details), whereas in case (ii), the conditional expectation does not exist!

■

Note:

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

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

(a) $E(c \mid Y) = c \quad \forall c \in I\!R$.

(b) $E(aX + b \mid Y) = aE(X \mid Y) + b \quad \forall a, b \in I\!R$.

(c) If g_1, g_2 are Borel–measurable functions and if $E(g_1(X)), E(g_2(X))$ exist, then $E(a_1 g_1(X) + a_2 g_2(X) \mid Y) = a_1 E(g_1(X) \mid Y) + a_2 E(g_2(X) \mid Y) \;\; \forall a_1, a_2 \in I\!R$.

(d) If $X \geq 0$, i.e., $P(X \geq 0) = 1$, then $E(X \mid Y) \geq 0$.

(e) If $X_1 \geq X_2$, i.e., $P(X_1 \geq X_2) = 1$, then $E(X_1 \mid Y) \geq E(X_2 \mid Y)$.

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

■

Example 4.7.2:

Recall Example 4.1.12:

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

The conditional pdf's $f_{Y|X}(y \mid x)$ and $f_{X|Y}(x \mid y)$ have been calculated as:

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

and

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

■

Theorem 4.7.3:

If $E(h(X))$ exists, then

$$E_Y(E_{X|Y}(h(X) \mid Y)) = E(h(X)).$$

Proof:

Continuous case only:

■

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:

$\blacksquare$

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

$$E_Y(Var(X \mid Y)) = E_Y(E((X - E(X \mid Y))^2 \mid Y)) = E_Y(E(X^2 \mid Y) - (E(X \mid Y))^2) = 0$$

which holds if $X = E(X \mid Y) = g(Y)$.

If X, Y are independent, $F_{X|Y}(x \mid y) = F_X(x) \ \ \forall x$.

Thus, if $E(h(X))$ exists, then $E(h(X) \mid Y) = E(h(X))$. $\blacksquare$

Proof: (of Theorem 4.6.8)

$\blacksquare$

4.8 Inequalities and Identities

(Based on Casella/Berger, Section 4.7)

Lemma 4.8.1:
Let a, b be positive numbers and $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$ (i.e., $pq = p + q$ and $q = \frac{p}{p-1}$). Then it holds that

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality iff $a^p = b^q$.

Proof:

■

Theorem 4.8.2: **Hölders 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

$$E(\mid XY \mid) \leq (E(\mid X \mid^p))^{\frac{1}{p}} (E(\mid Y \mid^q))^{\frac{1}{q}}.$$

Proof:
In Lemma 4.8.1, let

$$a = \frac{\mid X \mid}{(E(\mid X \mid^p))^{\frac{1}{p}}} > 0 \quad \text{and} \quad b = \frac{\mid Y \mid}{(E(\mid Y \mid^q))^{\frac{1}{q}}} > 0.$$

■

Note:

Note that Theorem 4.5.7 (ii) (Cauchy–Schwarz Inequality) is a special case of Theorem 4.8.2 with $p = q = 2$. If $X \sim Dirac(0)$ or $Y \sim Dirac(0)$ and it therefore holds that $E(\mid X \mid) = 0$ or $E(\mid Y \mid) = 0$, the inequality trivially holds. ■

Theorem 4.8.3: **Minkowski's Inequality**

Let X, Y be 2 rv's. Then it holds for $1 \leq p < \infty$ that

$$(E(\mid X + Y \mid^p))^{\frac{1}{p}} \leq (E(\mid X \mid^p))^{\frac{1}{p}} + (E(\mid Y \mid^p))^{\frac{1}{p}}.$$

Proof:

Assume $p > 1$ (the inequality trivially holds for $p = 1$):

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

■

Definition 4.8.4:

A function $g(x)$ is **convex** if

$$g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y) \quad \forall x, y \in I\!R \quad \forall 0 < \lambda < 1.$$

■

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 $x_0 = E(X)$:

$\blacksquare$

Note:

Typical convex functions g are:

(i) $g_1(x) = \mid x \mid \quad \Rightarrow \quad E(\mid X \mid) \geq \mid E(X) \mid$.

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

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

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

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

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

(vii) Does it hold that $E\left(\frac{X}{Y}\right)$ equals $\frac{E(X)}{E(Y)}$? The answer is ...

$\blacksquare$

Example 4.8.6:

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

$$\textbf{arithmetic mean} \ : \ a_A = \frac{1}{n}(a_1 + a_2 + \ldots + a_n) = \frac{1}{n}\sum_{i=1}^{n} a_i$$

$$\textbf{geometric mean} \ : \ a_G = (a_1 \cdot a_2 \cdot \ldots \cdot a_n)^{\frac{1}{n}} = \left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}}$$

$$\textbf{harmonic mean} \ : \ a_H = \frac{1}{\frac{1}{n}\left(\frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n}\right)} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i}}$$

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

(i) $a_A \geq a_G$:

(ii) $a_A \geq a_H$:

(iii) $a_G \geq a_H$:

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 μ.

(i) If $g(x)$ is non–decreasing, then
$$E(g(X)(X - \mu)) \geq 0$$
if this expectation exists.

(ii) If $g(x)$ is non–decreasing and $h(x)$ is non–increasing, then
$$E(g(X)h(X)) \leq E(g(X))E(h(X))$$
if all expectations exist.

(iii) If $g(x)$ and $h(x)$ are both non–decreasing or if $g(x)$ and $h(x)$ are both non–increasing, then
$$E(g(X)h(X)) \geq E(g(X))E(h(X))$$
if all expectations exist.

Proof:
Homework ■

Note:
Theorem 4.8.7 is called Covariance Inequality because

- (ii) implies $Cov(g(X), h(X)) \leq 0$, and
- (iii) implies $Cov(g(X), h(X)) \geq 0$.

■

5 Particular Distributions

5.1 Multivariate Normal Distributions

(Based on Casella/Berger, Exercises 4.45 through 4.50)

Definition 5.1.1:
A rv X has a **(univariate) Normal distribution**, i.e., $X \sim N(\mu, \sigma^2)$ with $\mu \in I\!R$ and $\sigma > 0$, iff it has the pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

X has a **standard Normal distribution** iff $\mu = 0$ and $\sigma^2 = 1$, i.e., $X \sim N(0, 1)$. ■

Note:
If $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$ and $Var(X) = \sigma^2$.

If $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$, X_1 and X_2 independent, and $c_1, c_2 \in I\!R$, 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 $a_{11}, a_{12}, a_{21}, a_{22}, \mu_1, \mu_2 \in I\!R$ 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

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad \underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \underline{X} = \begin{pmatrix} X \\ Y \end{pmatrix}, \quad \underline{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

then we can write

$$\underline{X} = A\underline{Z} + \underline{\mu}.$$

■

Note:

(i) Recall that for $X \sim N(\mu, \sigma^2)$, X can be defined as $X = \sigma Z + \mu$, where $Z \sim N(0, 1)$.

(ii) $E(X) = \mu_1 + a_{11}E(Z_1) + a_{12}E(Z_2) = \mu_1$ and $E(Y) = \mu_2 + a_{21}E(Z_1) + a_{22}E(Z_2) = \mu_2$. The marginal distributions are $X \sim N(\mu_1, a_{11}^2 + a_{12}^2)$ and $Y \sim N(\mu_2, a_{21}^2 + a_{22}^2)$. Thus, X and Y have (univariate) Normal marginal densities or degenerate marginal densities (which correspond to Dirac distributions) if $a_{i1} = a_{i2} = 0$.

(iii) There exists another (equivalent) formulation of the previous defintion using the joint pdf (see Rohatgi, page 227).

■

Example: Let $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim (\mu_2, \sigma_2^2)$, X and Y independent. What is the distribution of $\begin{pmatrix} X \\ Y \end{pmatrix}$?

■

Theorem 5.1.3:
Define $\underline{g} : I\!R^2 \to I\!R^2$ as $\underline{g}(\underline{x}) = C\underline{x} + \underline{d}$ with $C \in I\!R^{2\times 2}$ a 2×2 matrix and $\underline{d} \in I\!R^2$ a 2–dimensional vector. If $\underline{X}$ is a bivariate Normal rv, then $\underline{g}(\underline{X})$ also is a bivariate Normal rv.

Proof:

■

Note:

$$\begin{aligned} \rho\sigma_1\sigma_2 = Cov(X,Y) &= Cov(a_{11}Z_1 + a_{12}Z_2, a_{21}Z_1 + a_{22}Z_2) \\ &= a_{11}a_{21}Cov(Z_1, Z_1) + (a_{11}a_{22} + a_{12}a_{21})Cov(Z_1, Z_2) + a_{12}a_{22}Cov(Z_2, Z_2) \\ &= a_{11}a_{21} + a_{12}a_{22} \end{aligned}$$

since Z_1, Z_2 are iid $N(0,1)$ rv's. ■

Definition 5.1.4:
The **variance–covariance matrix** of (X, Y) is

$$\Sigma = AA^T = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{pmatrix} = \begin{pmatrix} a_{11}^2 + a_{12}^2 & a_{11}a_{21} + a_{12}a_{22} \\ a_{11}a_{21} + a_{12}a_{22} & a_{21}^2 + a_{22}^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

■

Note:

$$det(\Sigma) = \quad \mid \Sigma \mid \quad = \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 = \sigma_1^2\sigma_2^2(1-\rho^2),$$
$$\sqrt{|\Sigma|} = \sigma_1\sigma_2\sqrt{1-\rho^2}$$

and

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

■

Theorem 5.1.5:

Assume that $\sigma_1 > 0, \sigma_2 > 0$ and $\mid \rho \mid < 1$. Then the joint pdf of $\underline{X} = (X, Y) = A\underline{Z} + \underline{\mu}$ (as defined in Definition 5.1.2) is

$$\begin{aligned} f_{\underline{X}}(\underline{x}) &= \frac{1}{2\pi\sqrt{\mid \Sigma \mid}} \exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu})'\Sigma^{-1}(\underline{x}-\underline{\mu})\right) \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right)\right) \end{aligned}$$

Proof:

Since Σ is positive definite and symmetric, A is invertible.

The mapping $\underline{Z} \to \underline{X}$ is 1–to–1:

The second line of the Theorem is based on the transformations stated in the Note following Definition 5.1.4:

$$f_{\underline{X}}(\underline{x}) \quad = \quad \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right)\right)$$

■

<u>Note:</u>
In the situation of Theorem 5.1.5, we say that $(X, Y) \sim N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. ■

<u>Theorem 5.1.6:</u>
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

$$N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2), \sigma_1^2(1 - \rho^2)).$$

<u>Proof:</u>
Homework ■

<u>Example 5.1.7:</u>
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 $M_{\underline{X}}(\underline{t})$ of a non–singular bivariate Normal rv $\underline{X}' = (X, Y)$ is

$$M_{\underline{X}}(\underline{t}) = M_{X,Y}(t_1, t_2) = \exp(\underline{\mu}'\underline{t} + \frac{1}{2}\underline{t}'\Sigma\underline{t}) = \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{1}{2}(\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2\rho\sigma_1\sigma_2 t_1 t_2)\right).$$

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 $\underline{X}' = (X, Y)$:

Bivariate Normal mgf:

(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 $\underline{Z}$ 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 $\underline{\mu} \in \mathbb{R}^k$ be a k–dimensional vector. Then $\underline{X} = A\underline{Z} + \underline{\mu}$ has a **multivariate Normal distribution** with mean vector $\underline{\mu}$ and variance–covariance matrix $\Sigma = AA'$. ■

Note:

(i) If Σ is non–singular, $\underline{X}$ has the joint pdf

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{k/2}(\mid \Sigma \mid)^{1/2}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu})'\Sigma^{-1}(\underline{x} - \underline{\mu})\right).$$

If Σ is singular, the joint pdf does exist but it cannot be written explicitly.

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

(iii) If Σ is non–singular, then $\underline{X}$ has mgf

$$M_{\underline{X}}(\underline{t}) = \exp(\underline{\mu}'\underline{t} + \frac{1}{2}\underline{t}'\Sigma\underline{t}).$$

(iv) $\underline{X}$ has characteristic function

$$\Phi_{\underline{X}}(\underline{t}) = \exp(i\underline{\mu}'\underline{t} - \frac{1}{2}\underline{t}'\Sigma\underline{t})$$

(no matter whether Σ is singular or non–singular).

(v) See Searle, S. R. (1971): "Linear Models", Chapter 2.7, for more details on singular Normal distributions.

$\blacksquare$

Theorem 5.1.11:

The components $X_1, \ldots, X_k$ of a normally distributed k–rv $\underline{X}$ are independent iff $Cov(X_i, X_j) = 0 \quad \forall i, j = 1, \ldots, k,\ i \neq j$. $\blacksquare$

Theorem 5.1.12:

Let $\underline{X}' = (X_1, \ldots, X_k)$. $\underline{X}$ has a k–dimensional Normal distribution iff every linear function of $\underline{X}$, i.e., $\underline{X}'\underline{t} = t_1X_1 + t_2X_2 + \ldots + t_kX_k$, 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 $\underline{X}$, i.e., $\underline{X}'\underline{t}$, has a univariate Normal distribution.

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

$$\begin{aligned}
M_{\underline{X}'\underline{t}}(s) &= \exp\left(E(\underline{X}'\underline{t}) \cdot s + \frac{1}{2}Var(\underline{X}'\underline{t}) \cdot s^2\right) \\
&= \exp\left(\underline{\mu}'\underline{t}s + \frac{1}{2}\underline{t}'\Sigma\underline{t}s^2\right) \\
\Longrightarrow M_{\underline{X}'\underline{t}}(1) &= \exp\left(\underline{\mu}'\underline{t} + \frac{1}{2}\underline{t}'\Sigma\underline{t}\right) \\
&= M_{\underline{X}}(\underline{t})
\end{aligned}$$

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

5.2 Exponential Family of Distributions

(Based on Casella/Berger, Section 3.4)

Definition 5.2.1:

Let ϑ 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 $\{\underline{x} : f(\underline{x};\theta) > 0\}$ is independent of θ, where $\underline{x} = (x_1, \ldots, x_n)$.

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 ϑ and Borel–measurable functions $T(\underline{X})$ and $S(\underline{X})$ on $I\!R^n$ such that

$$f(\underline{x};\theta) = \exp(Q(\theta)T(\underline{x}) + D(\theta) + S(\underline{x})).$$

■

Note:

We can also write $f(\underline{x};\theta)$ as

$$f(\underline{x};\eta) = h(\underline{x})c(\eta)\exp(\eta T(\underline{x}))$$

where $h(\underline{x}) = \exp(S(\underline{x}))$, $\eta = Q(\theta)$, and $c(\eta) = \exp(D(Q^{-1}(\eta)))$, and call this the **exponential family in canonical form** for a natural parameter η. ■

Definition 5.2.2:

Let $\underline{\vartheta} \subseteq I\!R^k$ be a k–dimensional interval. Let $\{f(\cdot;\underline{\theta}) : \underline{\theta} \in \underline{\vartheta}\}$ be a family of pdf's (or pmf's). We assume that the set $\{\underline{x} : f(\underline{x};\underline{\theta}) > 0\}$ is independent of $\underline{\theta}$, where $\underline{x} = (x_1, \ldots, x_n)$.

We say that the family $\{f(\cdot;\underline{\theta}) : \underline{\theta} \in \underline{\vartheta}\}$ is a **k–parameter exponential family** if there exist real–valued functions $Q_1(\underline{\theta}), \ldots Q_k(\underline{\theta})$ and $D(\underline{\theta})$ on $\underline{\vartheta}$ and Borel–measurable functions $T_1(\underline{X}), \ldots, T_k(\underline{X})$ and $S(\underline{X})$ on $I\!R^n$ such that

$$f(\underline{x};\underline{\theta}) = \exp\left(\sum_{i=1}^{k} Q_i(\underline{\theta})T_i(\underline{x}) + D(\underline{\theta}) + S(\underline{x})\right).$$

■

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 $\underline{\eta} = (\eta_1, \ldots, \eta_k)'$ as

$$f(\underline{x};\underline{\eta}) = h(\underline{x})c(\underline{\eta})\exp\left(\sum_{i=1}^{k} \eta_i T_i(\underline{x})\right),$$

and define the **natural parameter space** as the set of points $\underline{\eta} \in W \subseteq I\!R^n$ for which the integral

$$\int_{I\!R^n} \exp\left(\sum_{i=1}^{k} \eta_i T_i(\underline{x})\right) h(\underline{x}) d\underline{x}$$

is finite. ■

Note:
If the support of the family of pdf's is some fixed interval (a, b), we can bring the expression $I_{(a,b)}(x)$ into exponential form as $\exp(\ln I_{(a,b)}(x))$ and then continue to write the pdf's as an exponential family as defined above, given this family is an exponential family. Note that for $I_{(a,b)}(x) \in \{0, 1\}$, it follows that $\ln I_{(a,b)}(x) \in \{-\infty, 0\}$ and therefore $\exp(\ln I_{(a,b)}(x)) \in \{0, 1\}$ as needed. ■

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

$$\begin{aligned} f(x; \underline{\theta}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = \exp\left(-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right) \\ \underline{\theta} &= (\mu, \sigma^2) \\ \underline{\vartheta} &= \{(\mu, \sigma^2) : \mu \in I\!R, \sigma^2 > 0\} \end{aligned}$$

Therefore,

$$\begin{aligned} Q_1(\underline{\theta}) &= -\frac{1}{2\sigma^2} \\ T_1(x) &= x^2 \\ Q_2(\underline{\theta}) &= \frac{\mu}{\sigma^2} \\ T_2(x) &= x \\ D(\underline{\theta}) &= -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \\ S(x) &= 0 \end{aligned}$$

Thus, this is a 2–parameter exponential family.

Canonical form:

$$f(x; \underline{\theta}) = \exp(Q_1(\underline{\theta})T_1(x) + Q_2(\underline{\theta})T_2(x) + D(\underline{\theta}) + S(x))$$

$$\Longrightarrow h(x) = \exp(S(x)) = \exp(0) = 1$$

$$\underline{\eta} = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} Q_1(\underline{\theta}) \\ Q_2(\underline{\theta}) \end{pmatrix} = \begin{pmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix}$$

Therefore, when we solve $\quad \eta_1 = -\dfrac{1}{2\sigma^2}, \quad \eta_2 = \dfrac{\mu}{\sigma^2} \quad$ for $\quad \sigma^2, \ \mu, \quad$ we get

$$\sigma^2 = -\frac{1}{2\eta_1}, \quad \mu = \sigma^2\eta_2 = -\frac{1}{2\eta_1}\eta_2 = -\frac{\eta_2}{2\eta_1}.$$

Thus,

$$\begin{aligned}
C(\underline{\eta}) &= \exp(D(Q^{-1}(\underline{\eta}))) \\
&= \exp\left(\frac{-\left(-\frac{\eta_2}{2\eta_1}\right)^2}{2\left(-\frac{1}{2\eta_1}\right)} - \frac{1}{2}\ln\left(2\pi\left(-\frac{1}{2\eta_1}\right)\right)\right) \\
&= \exp\left(\frac{\eta_2^2}{4\eta_1} - \frac{1}{2}\ln\left(-\frac{\pi}{\eta_1}\right)\right)
\end{aligned}$$

Therefore, $f(x;\underline{\theta})$ can be reparametrized in canonical form as

$$f(x;\underline{\eta}) = 1 \cdot \exp\left(\frac{\eta_2^2}{4\eta_1} - \frac{1}{2}\ln\left(-\frac{\pi}{\eta_1}\right)\right)\exp\left(\eta_1 x^2 + \eta_2 x\right).$$

The natural parameter space is

$$\{(\eta_1, \eta_2) \mid \eta_1 < 0, \eta_2 \in I\!R\}$$

because

$$\int_{-\infty}^{\infty} \exp\left(\eta_1 x^2 + \eta_2 x\right) \cdot 1 dx < \infty$$

for $\eta_1 < 0$ (independent from η_2), but

$$\int_{-\infty}^{\infty} \exp\left(\eta_1 x^2 + \eta_2 x\right) \cdot 1 dx \qquad \text{undefined}$$

for $\eta_1 \geq 0$ (independent from η_2). ■

To Be Continued ...