STAT 6710

Mathematical Statistics I

Fall Semester 1999

Dr. Jürgen Symanzik

Utah State University

Department of Mathematics and Statistics

3900 Old Main Hill

Logan, UT 84322–3900

Tel.: (435) 797–0696

FAX: (435) 797–1822

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

Contents

Acknowledgements

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

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

The textbook required for this class is:

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

A Web page dedicated to this class is accessible at:

`http://www.math.usu.edu/~symanzik/teaching/1999_stat6710/stat6710.html`

This course closely follows Rohatgi (1976) as described in the syllabus. Additional material originates from the lectures from Professors Hering, Trenkler, and Gather I have attended while studying at the Universität Dortmund, Germany, the collection of Masters and PhD Preliminary Exam questions from Iowa State University, Ames, Iowa, and the following textbooks:

- Bandelow, C. (1981): *Einführung in die Wahrscheinlichkeitstheorie*, Bibliographisches Institut, Mannheim, Germany.

- 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.

- 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.

- 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, January 18, 2000

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 2: Wednesday 9/1/1999	
	Scribe: Jürgen Symanzik

1 Axioms of Probability

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:*

$$\bigcup_{A \in \mathcal{A}} A = (\bigcap_{A \in \mathcal{A}} A^C)^C \text{ and } \bigcap_{A \in \mathcal{A}} A = (\bigcup_{A \in \mathcal{A}} A^C)^C.$$

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

Proof:

$$A, B \in L \stackrel{(ii)}{\Longrightarrow} A^C, B^C \in L \stackrel{(iii)}{\Longrightarrow} (A^C \cup B^C) \in L \stackrel{(ii)}{\Longrightarrow} (A^C \cup B^C)^C \in L \stackrel{DM}{\Longrightarrow} A \cap B \in L$$

■

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). Then L is a field. But L is a σ–field **iff** (if and only if) Ω is finite.

Example:
$\Omega = Z$. 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 $((\bigcup\limits_{n=1}^{\infty} A_n)^C = Z_0^-$ is infinite, too).

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

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".

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 3: Friday 9/3/1999	
	Scribe: Jürgen Symanzik

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:
(i) $\Omega \in L_\theta \ \ \forall \theta \Longrightarrow \Omega \in L$
(ii) Let $A \in L \Longrightarrow A \in L_\theta \ \ \forall \theta \Longrightarrow A^C \in L_\theta \ \ \forall \theta \Longrightarrow A^C \in L$
(iv) Let $A_n \in L \ \ \forall n \Longrightarrow A_n \in L_\theta \ \ \forall \theta \ \ \forall n \Longrightarrow \bigcup\limits_n A_n \in L_\theta \ \ \forall \theta \Longrightarrow \bigcup\limits_n A_n \in L$ ■

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

What is $\sigma(a)$?

$\sigma(a)$: must include $\Omega, \emptyset, \{0\}$

also: $\{1, 2, 3\}$ by 1.1.1 (ii)

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

What is $\sigma(b)$?

$\sigma(b)$: must include $\Omega, \emptyset, \{0\}, \{0, 1\}$

also: $\{1, 2, 3\}, \{2, 3\}$ by 1.1.1 (ii)

$\{0, 2, 3\}$ by 1.1.1 (iii)

$\{1\}$ by 1.1.1 (ii)

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

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: $\{x\} = \bigcap\limits_{n=1}^{\infty} (x - \frac{1}{n}, x]$
closed intervals: $[x, y] = (x, y] + \{x\} = (x, y] \cup \{x\}$
open intervals: $(x, y) = (x, y] - \{y\} = (x, y] \cap \{y\}^C$
and semi–infinite intervals: $(x, \infty) = \bigcup\limits_{n=1}^{\infty} (x, x + n]$ ■

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 $P : L \to I\!R$ satisfying

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

(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.*

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 4: Wednesday 9/8/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

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:

(i) $A_n \equiv \emptyset \quad \forall n \Rightarrow \bigcup\limits_{n=1}^{\infty} A_n = \emptyset \in L$.

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

$P(\emptyset) = P(\bigcup\limits_{n=1}^{\infty} A_n) \overset{Def 1.1.7(iii)}{=} \sum\limits_{n=1}^{\infty} P(A_n) = \sum\limits_{n=1}^{\infty} P(\emptyset)$

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

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

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

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

$$\begin{aligned} 1 = P(\Omega) &= P(\bigcup_{n=1}^{\infty} A_n) \\ &\overset{Def 1.1.7(iii)}{=} \sum_{n=1}^{\infty} P(A_n) \end{aligned}$$

$$
\begin{aligned}
&= P(A_1) + P(A_2) + \sum_{n=3}^{\infty} P(A_n) \\
&\overset{Th1.2.1(i)}{=} P(A_1) + P(A_2) \\
&= P(A) + P(A^C)
\end{aligned}
$$

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

(iii) By Th. 1.2.1 (ii) $P(A) = 1 - P(A^C) \Longrightarrow P(A) \leq 1 \quad \forall A \in L$ since $P(A^C) \geq 0$ by Def 1.1.7 (i)

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

$$
\begin{aligned}
\Longrightarrow P(A \cup B) &= P((A \cap B^C) \cup (A \cap B) \cup (B \cap A^C)) \\
&\overset{Def.1.1.7(iii)}{=} P(A \cap B^C) + P(A \cap B) + P(B \cap A^C) \\
&= P(A \cap B^C) + P(A \cap B) + P(B \cap A^C) + P(A \cap B) - P(A \cap B) \\
&= (P(A \cap B^C) + P(A \cap B)) + (P(B \cap A^C) + P(A \cap B)) - P(A \cap B) \\
&= P(A) + P(B) - P(A \cap B)
\end{aligned}
$$

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

$P(B) = P((B \cap A^C) \cup A) \overset{Def1.1.7(iii)}{=} P(B \cap A^C) + P(A)$

$\Longrightarrow P(A) = P(B) - P(B \cap A^C)$

$\Longrightarrow P(A) \leq P(B)$ since $P(B \cap A^C) \geq 0$ by Def 1.1.7 (i)

■

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

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

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

Proof:

$n = 1$ is trivial

$n = 2$ is Theorem 1.2.1 (iv)

use induction for higher n (Homework) ■

Theorem 1.2.3: **Bonferroni's Inequality**

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

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

Proof:

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

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

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

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

$$\begin{aligned}
P(\bigcup_{i=1}^{n+1} A_i) &= P((\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) \\
&\overset{Th.1.2.1(iv)}{=} P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) - P((\bigcup_{i=1}^{n} A_i) \cap A_{n+1}) \\
&\overset{Def.1.1.7(i)}{\leq} P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) \\
&\overset{I.B.}{\leq} \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) \\
&= \sum_{i=1}^{n+1} P(A_i)
\end{aligned}$$

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

For $n = 2$, Th. 1.2.3 left side evaluates to $P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1 \cup A_2)$, which is true by Th. 1.2.1 (iv).

For $n = 3$, Th. 1.2.3 left side evaluates to
$P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \leq P(A_1 \cup A_2 \cup A_3)$.

This holds since

$$
\begin{aligned}
&P(A_1 \cup A_2 \cup A_3) \\
&\quad = P((A_1 \cup A_2) \cup A_3) \\
&\quad \overset{Th.1.2.1(iv)}{=} P(A_1 \cup A_2) + P(A_3) - P((A_1 \cup A_2) \cap A_3) \\
&\quad = P(A_1 \cup A_2) + P(A_3) - P((A_1 \cap A_3) \cup (A_2 \cap A_3)) \\
&\quad \overset{Th.1.2.1(iv)}{=} P(A_1) + P(A_2) - P(A_1 \cap A_2) + P(A_3) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\
&\qquad\qquad + P((A_1 \cap A_3) \cap (A_2 \cap A_3)) \\
&\quad = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \\
&\quad \overset{Def1.1.7(i)}{\geq} P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3)
\end{aligned}
$$

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

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

$$
\begin{aligned}
P(\bigcup_{i=1}^{n+1} A_i) &= P((\bigcup_{i=1}^{n} A_i) \cup A_{n+1}) \\
&= P(\bigcup_{i=1}^{n} A_i) + P(A_{n+1}) - P((\bigcup_{i=1}^{n} A_i) \cap A_{n+1}) \\
&\overset{left\ I.B.}{\geq} \sum_{i=1}^{n} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) + P(A_{n+1}) - P((\bigcup_{i=1}^{n} A_i) \cap A_{n+1}) \\
&= \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) - P(\bigcup_{i=1}^{n} (A_i \cap A_{n+1})) \\
&\overset{Th.1.2.3\ right\ side}{\geq} \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n} P(A_i \cap A_j) - \sum_{i=1}^{n} P(A_i \cap A_{n+1}) \\
&= \sum_{i=1}^{n+1} P(A_i) - \sum_{i<j}^{n+1} P(A_i \cap A_j)
\end{aligned}
$$

■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
	Lecture 5: Friday 9/10/1999
	Scribe: Jürgen Symanzik, Rich Madsen

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

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

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

<u>Proof</u>:
Homework ■

<u>Definition 1.2.5:</u> **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 $A = \bigcup\limits_{n=1}^{\infty} A_n$.

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

■

<u>Theorem 1.2.6:</u>
If $\{A_n\}_{n=1}^{\infty}, A_n \in L$ and $A \in L$, then $\lim\limits_{n \to \infty} P(A_n) = P(A)$ if 1.2.5 (i) or 1.2.5 (ii) holds.

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

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

It is $A = \bigcup\limits_{n=1}^{\infty} A_n = \bigcup\limits_{n=1}^{\infty} B_n$

and also $A_n = \bigcup\limits_{i=1}^{n} A_i = \bigcup\limits_{i=1}^{n} B_i$

$$\underline{\mathrm{P(A)}} = \mathrm{P}(\bigcup_{k=1}^{\infty} B_k) \overset{\text{By Def. 1.1.7 (iii)}}{=} \sum_{k=1}^{\infty} P(B_k) = \lim_{n\to\infty}[\sum_{k=1}^{n} P(B_k)]$$

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

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

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

Then, $A_1^C \subseteq A_2^C \subseteq A_3^C \subseteq \ldots$ and $A^C = (\bigcap_{n=1}^{\infty} A_n)^C \overset{DeMorgan}{=} \bigcup_{n=1}^{\infty} A_n^C$

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

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

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

Theorem 1.2.7:

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

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

Proof:

Part (i):

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

$$0 \overset{\text{By K.A.P. (i)}}{\leq} P(\bigcup_{n=1}^{\infty} A_n) \overset{\text{By Bonferroni's Inequality}}{\leq} \sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} 0 = 0$$

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

Part (ii):

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

$$\overset{\text{by Th. 1.2.1 (ii)}}{\Longrightarrow} P(A_n^C) = 0 \;\; \forall n \overset{\text{by Th. 1.2.7 (i)}}{\Longrightarrow} P(\bigcup_{n=1}^{\infty} A_n^C) = 0 \overset{\text{De Morgan}}{\Longrightarrow} P(\bigcap_{n=1}^{\infty} A_n) = 1$$ ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 6: Monday 9/13/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

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, $P(A) = \frac{\text{number } \omega_j \text{ in } A}{\text{number } \omega_j \text{ in } \Omega}$. Therefore, to calculate such probabilities, we just need to be able to count elements accurately.

Theorem 1.3.2: **Fundamental Theorem of Counting**

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

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

ways to do it.

Proof: (By Induction)

Induction Base:

$k = 1$: trivial

$k = 2$: n_1 ways to choose a_1. For each, n_2 ways to choose a_2.

Total # of ways $= \underbrace{n_2 + n_2 + \ldots + n_2}_{n_1 \text{ times}} = n_1 \times n_2$.

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

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

■

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

(i) the size of the set, n;

(ii) the size of the sample, r;

(iii) whether the result will be **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) n choices to select 1^{st}
$n-1$ choices to select 2^{nd}
$\vdots$
$n-r+1$ choices to select r^{th}
By Theorem 1.3.2, there are $n \times (n-1) \times \ldots \times (n-r+1) = \frac{n \times (n-1) \times \ldots \times (n-r+1) \times (n-r)!}{(n-r)!} = \frac{n!}{(n-r)!}$ ways to do so.

Corollary:

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

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

n choices to select 2^{nd}

$\vdots$

n choices to select r^{th}

By Theorem 1.3.2, there are $\underbrace{n \times n \times \ldots \times n}_{r \text{ times}} = n^r$ ways to do so.

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

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

■

Theorem 1.3.6: **The Binomial Theorem**

If n is a non–negative integer, then

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

Proof: (By Induction)

Induction Base:

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

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

Induction Step:

Suppose it is true for k. We show that it is true for $k+1$.

$$\begin{aligned}
(1+x)^{k+1} &= (1+x)^k(1+x) \\
&\overset{IB}{=} \left(\sum_{r=0}^{k}\binom{k}{r}x^r\right)(1+x) \\
&= \sum_{r=0}^{k}\binom{k}{r}x^r + \sum_{r=0}^{k}\binom{k}{r}x^{r+1} \\
&= \binom{k}{0}x^0 + \sum_{r=1}^{k}\left[\binom{k}{r}+\binom{k}{r-1}\right]x^r + \binom{k}{k}x^{k+1} \\
&= \binom{k+1}{0}x^0 + \sum_{r=1}^{k}\left[\binom{k}{r}+\binom{k}{r-1}\right]x^r + \binom{k+1}{k+1}x^{k+1} \\
&\overset{(*)}{=} \binom{k+1}{0}x^0 + \sum_{r=1}^{k}\binom{k+1}{r}x^r + \binom{k+1}{k+1}x^{k+1} \\
&= \sum_{r=0}^{k+1}\binom{k+1}{r}x^r
\end{aligned}$$

(*) Here we use Theorem 1.3.8 (i). Since the proof of Theorem 1.3.8 (i) only needs algebraic transformations without using the Binomial Theorem, part (i) of Theorem 1.3.8 can be applied here. ■

Corollary 1.3.7:

For a non–negative integer n, it holds:

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

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

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

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

Proof:

Use the Binomial Theorem:

(i) Let $x=1$. Then

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

(ii) Let $x = -1$. Then

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

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

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

Substitute $x = 1$, then

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

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

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

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

$\blacksquare$

Note:

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

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

Theorem 1.3.8:

For non–negative integers, n, m, r, it holds:

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

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

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

Proof:

Homework $\blacksquare$

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 7: Wednesday 9/15/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

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.

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

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

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

$$\begin{aligned} P_H(\bigcup_{n=1}^{\infty} A_n) &\overset{Def.1.4.1}{=} \frac{P((\bigcup_{n=1}^{\infty} A_n) \cap H)}{P(H)} \\ &= \frac{P(\bigcup_{n=1}^{\infty} (A_n \cap H))}{P(H)} \end{aligned}$$

$$\overset{Def.1.1.7(iii)}{=} \frac{\sum_{n=1}^{\infty} P(A_n \cap H)}{P(H)}$$

$$= \sum_{n=1}^{\infty} (\frac{P(A_n \cap H)}{P(H)})$$

$$\overset{Def 1.4.1}{=} \sum_{n=1}^{\infty} P_H(A_n)$$

■

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 ■

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.

■

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

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

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

The left side proof:
H_j are disjoint $\Rightarrow A \cap H_j$ are disjoint

$A = A \cap \Omega \stackrel{Def 1.4.4}{=} A \cap (\bigcup_{j=1}^{\infty} H_j) = \bigcup_{j=1}^{\infty} (A \cap H_j)$

$\Longrightarrow P(A) = P(\bigcup_{j=1}^{\infty} (A \cap H_j)) \stackrel{Def 1.1.7(iii)}{=} \sum_{j=1}^{\infty} P(A \cap H_j)$ ■

Theorem 1.4.6: **Bayes' Rule**
Let $\{H_j\}_{j=1}^{\infty}$ be a partition of Ω, and $P(H_j) > 0 \ \ \forall j$. Let $A \in L$ and $P(A) > 0$. Then

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

Proof:
$P(H_j \cap A) \stackrel{Def 1.4.1}{=} P(A) \cdot P(H_j|A) = P(H_j) \cdot P(A|H_j)$

$\Longrightarrow P(H_j|A) = \frac{P(H_j) \cdot P(A|H_j)}{P(A)} \stackrel{Th. 1.4.5}{=} \frac{P(H_j) \cdot P(A|H_j)}{\sum_{n=1}^{\infty} P(H_n)P(A|H_n)}.$ ■

Definition 1.4.7:
For $A, B \in L$, A and B are **independent** iff $P(A \cap B) = P(A)P(B)$. ■

Note:

- There are no restrictions on $P(A)$ or $P(B)$.
- If A and B are independent, then $P(A|B) = P(A)$ (given that $P(B) > 0$) and $P(B|A) = P(B)$ (given that $P(A) > 0$).

- If A and B are independent, then the following events are independent as well: A and B^C; A^C and B; A^C and B^C.

■

<u>Definition 1.4.8:</u>
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 $P(A_1 \cap A_2) = P(A_1)P(A_2)$. ■

<u>Definition 1.4.9:</u>
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 $P(\bigcap\limits_{j=1}^{k} A_{i_j}) = \prod\limits_{j=1}^{k} P(A_{i_j})$.
■

<u>Note:</u>
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.

<u>Example 1.4.10:</u>
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?

$P(A_1 \cap A_2) = .25 = .5 \cdot .5 = P(A_1) \cdot P(A_2) \Rightarrow A_1, A_2$ are independent.

$P(A_1 \cap A_3) = .25 = .5 \cdot .5 = P(A_1) \cdot P(A_3) \Rightarrow A_1, A_3$ are independent.

$P(A_2 \cap A_3) = .25 = .5 \cdot .5 = P(A_2) \cdot P(A_3) \Rightarrow A_2, A_3$ are independent.

Thus, A_1, A_2, A_3 are pairwise independent.

$P(A_1 \cap A_2 \cap A_3) = 0 \neq .5 \cdot .5 \cdot .5 = P(A_1) \cdot P(A_2) \cdot P(A_3) \Rightarrow A_1, A_2, A_3$ are not mutually independent.
■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 8: Friday 9/17/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

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

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

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

It is

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

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

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}

$$\begin{aligned} \text{So } p_2 &= P(A \cap B) \\ &= P(A) \cdot P(B|A) \\ &= P(\text{no student has a matching b-day with the } 1^{\text{st}}\text{student }) \times \\ & \quad P(\text{at least one of the remaining students has a mathching b-day with the second,} \\ & \qquad \text{given that no one matched the first.)} \end{aligned}$$

$$
\begin{aligned}
&= (1-p_1)[1-P(\text{all } (r\text{-}2) \text{ students have a b-day on the remaining 363 out of 364 days})\\
&= \left(\frac{364}{365}\right)^{r-1}\left[1-\left(\frac{363}{364}\right)^{r-2}\right]\\
&= \left(\frac{365-1}{365}\right)^{r-1}\left[1-\left(\frac{363}{364}\right)^{r-2}\right]
\end{aligned}
$$

<u>Working backwards from the book</u>

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

(Same as what we found before)

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

Let A $\equiv$ No one has the same b-day as the first student
Let B $\equiv$ No one has the same b-day as the second student
Let C $\equiv$ At least one of the other $(r-3)$ has the same b-day as the third students

Now:

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

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

<u>Working backwards from the book</u>

$$\begin{aligned}
p_3 &= \left(\frac{{}_{365}P_{3-1}}{(365)^{3-1}}\right)\left(1-\frac{3-1}{365}\right)^{r-3+1}\left[1-\left(\frac{365-3}{365-3+1}\right)^{r-3}\right] \\
&= \left(\frac{{}_{365}P_{2}}{(365)^{2}}\right)\left(1-\frac{2}{365}\right)^{r-2}\left[1-\left(\frac{362}{363}\right)^{r-3}\right] \\
&= \left(\frac{(365)(364)}{(365)^{2}}\right)\left(\frac{363}{365}\right)^{r-2}\left[1-\left(\frac{362}{363}\right)^{r-3}\right] \\
&= \left(\frac{364}{365}\right)\left(\frac{363}{365}\right)^{r-2}\left[1-\left(\frac{362}{363}\right)^{r-3}\right]
\end{aligned}$$

(Same as what we found before)

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

■

2 Random Variables

2.1 Measurable Functions

Definition 2.1.1:

- A **random variable** (rv) is a set function from 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**, $\underline{X} = (X_1(\omega), X_2(\omega), \ldots, X_k(\omega))$.

■

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 $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$.

■

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

$X : \Omega \to \mathcal{S} = \{$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 $B \in B, X^{-1}(B) \in L = \mathcal{P}(\Omega)$.

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

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

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 $L = \sigma(\bigcup\limits_{n=1}^{\infty} L_n)$.

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 $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, n\}$) and $X_n^{-1}(\frac{k}{n}) \in L_n \quad \forall k = 0, 1, \ldots, n$.

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). ■

Stat 6710 Mathematical Statistics I — Fall Semester 1999

Lecture 9: Monday 9/20/1999

Scribe: Jürgen Symanzik

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.

2.1.10:

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

(ii) If $f = \lim_n f_n$ exists, then f is measurable.

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

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

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 10: Wednesday 9/22/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

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

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 $\Rightarrow X^{-1}(B) = \{\omega : X(\omega) \in B\} = A \in L \ \ \forall B \in \mathcal{B}$.

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

(ii) $Q(I\!R) = P(X^{-1}(I\!R)) \overset{X\ rv}{=} P(\Omega) \overset{Def.1.1.7(ii)}{=} 1$

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

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

$(*)$ holds since $X^{-1}(\cdot)$ commutes with unions/intersections and preserves disjointedness.

■

Definition 2.2.2:
A real–valued function F on $(-\infty, \infty)$ that is non–decreasing, right–continuous, and satisfies

$$F(-\infty) = 0, F(\infty) = 1$$

is called a **cumulative distribution function** (cdf) on $I\!R$. ■

Note:
No mention of probability space or measure P in Definition 2.2.2 above.

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

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

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

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

Proof:

(i) Let $x_1 < x_2$

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

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

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

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

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

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

By Theorem 1.2.6 it follows that

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

It is

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

$$\Longrightarrow (\lim_{n\to\infty} F(x_n)) - F(x) = \lim_{n\to\infty}(F(x_n) - F(x)) = \lim_{n\to\infty} P(A_n) = 0$$
$$\Longrightarrow \lim_{n\to\infty} F(x_n) = F(x)$$
$$\Longrightarrow F(x) \text{ is right–continuous.}$$

(iii) $F(-n) \overset{Def.\_2.2.3}{=} P(\{\omega : X(\omega) \leq -n\})$
$\Longrightarrow$

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

(iv) $F(n) \overset{Def.\_2.2.3}{=} P(\{\omega : X(\omega) \leq n\})$
$\Longrightarrow$

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

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

Proof:
(i) ⇒ (ii):

$$\begin{array}{rcl} F_X(x) & = & P_X((-\infty, x]) \\ & = & P(\{\omega : X(\omega) \in (-\infty, x]\}) \\ & \stackrel{by Def. 2.2.5}{=} & P(\{\omega : Y(\omega) \in (-\infty, x]\}) \\ & = & P_Y((-\infty, x]) \\ & = & F_Y(X) \end{array}$$

(ii) ⇒ (i):
Requires extra knowledge from measure theory. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 11: Friday 9/24/1999	
	Scribe: Jürgen Symanzik

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 $x \in R$. ■

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

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

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

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

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

Here, $F_d(x)$ and $F_c(x)$ are discrete and continuous cdfs (as above). $F_s(x)$ is called a **singular** cdf. Singular means that $F_s(x)$ is continuous and its derivative $F'(x)$ 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... ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 12: Monday 9/27/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

Example 2.3.7:
Consider

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

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

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

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

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

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

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

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

■

Definition 2.3.8:
The two–valued function $I_A(x)$ is called **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.

$A \Rightarrow B$ 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	0	0	1	1
1	0	0	0	1	0	0
0	1	1	1	0	1	1
0	0	1	1	1	1	1

$A \Leftrightarrow B$ 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	1	1	1	1	1
1	0	0	0	1	0	0	1	0
0	1	0	1	0	0	1	0	0
0	0	1	1	1	1	1	1	1

Negations of Quantifiers:

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

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

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

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 13: Wednesday 9/29/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

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:

$$\begin{aligned}
F_Y(y) &= P_Y(Y \leq y) \\
&= P(\{\omega : g(X(\omega)) \leq y\}) \\
&= P(\{\omega : X(\omega) \in B_y\}) \quad \text{where } B_y = g^{-1}(-\infty, y] \in \mathcal{B} \text{ since g is Borel-measureable.} \\
&= P(X^{-1}(B_y))
\end{aligned}$$

■

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,

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

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

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

Definition 2.4.4:
For a random variable X from (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 **(strictly) non–decreasing** if $x < y \Rightarrow f(x) \ (<) \ \leq f(y) \ \ \forall x, y \in D$
f is **(strictly) non–increasing** if $x < y \Rightarrow f(x) \ (>) \ \geq f(y) \ \ \forall x, y \in D$
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:

Part (i): $g'(x) > 0 \ \forall x \in \mathcal{X}$

So g is strictly increasing and continuous.

Therefore, $x = g^{-1}(y)$ exists and it is also strictly increasing and also differentiable.

Then, from Rohatgi, page 9, Theorem 15:

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

We get $F_Y(y) = P_Y(Y \leq y) = P_Y(g(X) \leq y) = P_X(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$ for $y \in g(\mathcal{X})$ and, by differentiation,

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

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

So g is strictly decreasing and continuous.

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

Then, from Rohatgi, page 9, Theorem 15:

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

We get $F_Y(y) = P_Y(Y \leq y) = P_Y(g(X) \leq y) = P_X(X \geq g^{-1}(y)) = 1 - P_X(X \leq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$ for $y \in g(\mathcal{X})$ and, by differentiation,

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

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

Combining parts (i) and (ii), we can therefore write

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

$\blacksquare$

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lectures 14 & 15: Friday 10/1/1999 & Monday 10/4/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

Note:
In Theorem 2.4.6, we can also write

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

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

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,

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

■

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

Example 2.4.8:

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

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

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

Two possible approaches:

Method 1: cdfs

For $0 < y < 1$ we have

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

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

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

Method 2: Use of Theorem 2.4.7

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

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

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

Thus, by use of Theorem 2.4.7, we get

$$
f_Y(y) = \sum_{i=1}^{2} f_X(g_i^{-1}(y)) \cdot \mid \frac{d}{dy} g_i^{-1}(y) \mid \cdot I_{\mathcal{Y}}(y)
$$

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

Obviously, both results are identical. ■

Theorem 2.4.9:

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

Proof:

We have to consider two possible cases:

(a) F_X is strictly increasing, i.e., $F_X(x_1) < F_X(x_2)$ for $x_1 < x_2$, and

(b) F_X is non–decreasing, i.e., there exists $x_1 < x_2$ and $F_X(x_1) = F_X(x_2)$. Assume that x_1 is the infimum and x_2 the supremum of those values for which $F_X(x_1) = F_X(x_2)$ holds.

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

Without loss of generality:

$F_X^{-1}(1) = +\infty$ if $F_X(x) < 1 \quad \forall x \in \mathbb{R}$ and
$F_X^{-1}(0) = -\infty$ if $F_X(x) > 0 \quad \forall x \in \mathbb{R}$.

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

$$
\begin{aligned}
P(Y \leq y) &= P(F_X(X) \leq y) \\
&\overset{F_X^{-1}\uparrow}{=} P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(y)) \\
&\overset{(*)}{=} P(X \leq F_X^{-1}(y)) \\
&= F_X(F_X^{-1}(y)) \\
&= y
\end{aligned}
$$

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

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

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

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 16: Wednesday 10/6/1999	
	Scribe: Jürgen Symanzik

3 Moments and Generating Functions

3.1 Expectation

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

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

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

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

$$E(\mid X \mid) = \tfrac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2} dx = \frac{1}{\pi} [\log(1+x^2)]_0^{\infty} = \infty$$

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

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

Proof:
Continuous case only:

Existence:

$$\begin{aligned} E(\mid aX + b \mid) &= \int_{-\infty}^{\infty} \mid ax + b \mid f_X(x) dx \\ &\leq \int_{-\infty}^{\infty} (\mid a \mid \cdot \mid x \mid + \mid b \mid) f_X(x) dx \\ &= \mid a \mid \int_{-\infty}^{\infty} \mid x \mid f_X(x) dx + \mid b \mid \int_{-\infty}^{\infty} f_X(x) dx \end{aligned}$$

$$
\begin{aligned}
&= \mid a \mid E(\mid X \mid) + \mid b \mid \\
&< \infty
\end{aligned}
$$

Numerical Result:

$$
\begin{aligned}
E(aX+b) &= \int_{-\infty}^{\infty} (ax+b) f_X(x)dx \\
&= a\int_{-\infty}^{\infty} x f_X(x)dx + b\int_{-\infty}^{\infty} f_X(x)dx \\
&= aE(X) + b
\end{aligned}
$$

■

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} **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

$$
\begin{aligned}
& E\left(\mid ((aX+b)-E(aX+b))^2 \mid\right) \\
= \quad & E\left(((aX+b)-E(aX+b))^2\right) \\
= \quad & Var(aX+b) \\
\overset{Th.3.1.6}{=} \quad & E((aX+b)^2)-(E(aX+b))^2 \\
\overset{Th.3.1.2}{=} \quad & E(a^2X^2+2abX+b^2)-(aE(X)+b)^2 \\
\overset{Th.3.1.2}{=} \quad & a^2E(X^2)+2abE(X)+b^2-a^2(E(X))^2-2abE(X)-b^2 \\
= \quad & a^2(E(X^2)-(E(X))^2) \\
\overset{Th.3.1.6}{=} \quad & a^2Var(X) \\
< \quad & \infty \quad \text{since } Var(X) \text{ exists}
\end{aligned}
$$

■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 17: Friday 10/8/1999	
	Scribe: Jürgen Symanzik

Theorem 3.1.8:
If the t^{th} moment of a rv X exists, then all moments of order $0 < s < t$ exist.

Proof:
Continuous case only:

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

■

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

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

Proof:
Continuous case only:

$\infty > \int_{I\!R} \mid x \mid^t f_X(x)dx = \lim_{n \to \infty} \int_{|x| \leq n} \mid x \mid^t f_X(x)dx$

$\Longrightarrow \lim_{n \to \infty} \int_{|x|>n} \mid x \mid^t f_X(x)dx = 0$

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

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

$$E(X) = \int_0^\infty x f_X(x) dx \ = \lim_{n \to \infty} \int_0^n x f_X(x) dx$$

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

Let $u = x$ and $dv = f_X(x)dx$, then

$$\begin{aligned}
\int_0^n x f_X(x) dx &= (xF(x)) \mid_0^n - \int_0^n F_X(x) dx \\
&= nF_X(n) - 0F_X(0) - \int_0^n F_X(x) dx \\
&= nF_X(n) - n + n - \int_0^n F_X(x) dx \\
&= nF_X(n) - n + \int_0^n [1 - F_X(x)] dx
\end{aligned}$$

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

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

$$= -nP(X > n) + \int_0^n [1 - F_X(x)]dx$$

$$\stackrel{X \geq 0}{=} -n^1 P(|X| > n) + \int_0^n [1 - F_X(x)]dx$$

$$\Longrightarrow E(X^1) = \lim_{n \to \infty} [-n^1 P(|X| > n) + \int_0^n [1 - F_X(x)]dx]$$

$$\stackrel{Th.\ 3.1.9}{=} 0 + \lim_{n \to \infty} \int_0^n [1 - F_X(x)]dx$$

$$= \int_0^\infty [1 - F_X(x)]dx$$

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

We still have to show the converse implication:

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

$$\int_0^n x f_X(x)dx \stackrel{X \geq 0}{=} \int_0^n \mid x \mid f_X(x)dx = -n[1 - F_X(n)] + \int_0^n [1 - F_X(x)]dx$$

as seen above.

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

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

Thus,

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

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

Corollary 3.1.12:

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

Proof:

$$E(| X |^s) \overset{Lemma\ 3.1.11}{=} \int_0^\infty [1 - F_{|X|^s}(z)] dz = \int_0^\infty P(| X |^s > z) dz.$$

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

$$\begin{aligned} \int_0^\infty P(| X |^s > \ z) dz &= \int_0^\infty P(| X |^s > y^s) sy^{s-1} dy \\ &= s \int_0^\infty y^{s-1} P(| X |^s > y^s) dy \\ &\overset{monotonic\ \uparrow}{=} s \int_0^\infty y^{s-1} P(| X |> y) dy \end{aligned}$$

■

Proof (of Theorem 3.1.10):

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

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

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

$$\leq \int_0^N sy^{s-1} \cdot 1 \ dy + s \int_N^\infty y^{s-1} \frac{\epsilon}{y^t} dy$$

$$= y^s \mid_0^N + \ s\epsilon \int_N^\infty y^{s-1} \frac{1}{y^t} dy$$

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

It is $\int_N^\infty y^c dy = \begin{cases} \frac{1}{c+1} y^{c+1} \mid_N^\infty, & c \neq -1 \\ \ln y \mid_N^\infty, & c = 1 \end{cases}$

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

Thus, for $E(| \ X \ |^s) \ < \ \infty$, it must hold that $s - 1 - t \ < \ -1$, or equivalently, $s \ < \ t$. So $E(| \ X \ |^s) \ < \ \infty$, i.e., it exists, for every s with $0 \ < \ s \ < \ t$ for a rv X with a distribution such that $\lim_{n \to \infty} n^t P(| X |> n) = 0$ for some $t > 0$. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 19: Wednesday 10/13/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

Theorem 3.1.13:

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

Proof:

- For $\epsilon > 0$, we select some k_0 such that
$$\frac{P(\mid X \mid > \alpha k)}{P(\mid X \mid > k)} < \epsilon \quad \forall k \geq k_0.$$
- Select k_1 such that $P(\mid X \mid > k) < \epsilon \quad \forall k \geq k_1$.
- Select $N = \max(k_0, k_1)$.
- If we have some fixed positive integer r:
$$\frac{P(\mid X \mid > \alpha^r k)}{P(\mid X \mid > k)} = \frac{P(\mid X \mid > \alpha k)}{P(\mid X \mid > k)} \cdot \frac{P(\mid X \mid > \alpha^2 k)}{P(\mid X \mid > \alpha k)} \cdot \frac{P(\mid X \mid > \alpha^3 k)}{P(\mid X \mid > \alpha^2 k)} \cdot \ldots \cdot \frac{P(\mid X \mid > \alpha^r k)}{P(\mid X \mid > \alpha^{r-1} k)}$$
$$\Longrightarrow \frac{P(\mid X \mid > \alpha^r k)}{P(\mid X \mid > k)} = \frac{P(\mid X \mid > \alpha k)}{P(\mid X \mid > k)} \cdot \frac{P(\mid X \mid > \alpha \cdot (\alpha k))}{P(\mid X \mid > 1 \cdot (\alpha k))} \cdot \frac{P(\mid X \mid > \alpha \cdot (\alpha^2 k))}{P(\mid X \mid > 1 \cdot (\alpha^2 k))} \cdot \ldots \cdot \frac{P(\mid X \mid > \alpha \cdot (\alpha^{r-1} k))}{P(\mid X \mid > 1 \cdot (\alpha^{r-1} k))}$$
- Note: Each of these r terms on the right side is $> \epsilon$ by our original statement of selecting some k_0 such that $\frac{P(|X|>\alpha k)}{P(|X|>k)} < \epsilon \quad \forall k \geq k_0$ and since $\alpha > 1$ and therefore $\alpha^n k \geq k_0$.
- Now we get for our entire expression that $\frac{P(|X|>\alpha^r k)}{P(|X|>k)} \leq \epsilon^r$ for $k \geq N$ (since in this case also $k \geq k_0$) and $\alpha > 1$.
- Overall, we have $P(\mid X \mid > \alpha^r k) \leq \epsilon^r P(\mid X \mid > k) \leq \epsilon^{r+1}$ for $k \geq N$ (since in this case also $k \geq k_1$).
- For a fixed positive integer n:
$$E(\mid X \mid^n) \stackrel{Cor. 3.1.12}{=} n \cdot \int_0^\infty x^{n-1} P(\mid X \mid > x) dx = n \int_0^N x^{n-1} P(\mid X \mid > x) dx + n \int_N^\infty x^{n-1} P(\mid X \mid > x) dx$$
- We know that:
$$n \int_0^N x^{n-1} P(\mid X \mid > x) dx \leq \int_0^N n x^{n-1} dx = x^n \mid_0^N = N^n \quad < \infty$$

but is

$$n \int_N^\infty x^{n-1} P(|\ X\ |> x)dx \quad < \infty \quad ?$$

- To check the second part, we use:

$$\int_N^\infty x^{n-1} P(|\ X\ |> x)dx \ = \ \sum_{r=1}^\infty \int_{\alpha^{r-1}N}^{\alpha^r N} x^{n-1} P(|\ X\ |> x)dx$$

- We know that:

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

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

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

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

- Now we go back to our original inequality:

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

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

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

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 20: Friday 10/15/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

3.2 Generating Functions

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

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

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

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

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

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.

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\
&= \int_{-\infty}^{\infty} (\frac{\partial}{\partial t} e^{tx} f_X(x)) dx \\
&= \int_{-\infty}^{\infty} x e^{tx} f_X(x) dx \\
&= E(X e^{tX})
\end{aligned}$$

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

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

$$\begin{aligned}
\frac{d^n}{dt^n} M_X(t) &= \frac{d}{dt} \left(\frac{d^{n-1}}{dt^{n-1}} M_X(t) \right) \\
&= \frac{d}{dt} \left(\int_{-\infty}^{\infty} x^{n-1} e^{tx} f_X(x) dx \right) \\
&= \int_{-\infty}^{\infty} (\frac{\partial}{\partial t} x^{n-1} e^{tx} f_X(x)) dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} x^n e^{tx} f_X(x) dx \\
&= E(X^n e^{tX})
\end{aligned}
$$

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

Example 3.2.3:

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

Then,

$$
\begin{aligned}
M_X(t) &= \int_a^b \frac{e^{tx}}{b-a} dx = \frac{e^{tb} - e^{ta}}{t(b-a)} \\
M_X(0) &= \frac{0}{0} \\
&\overset{\text{L'Hospital}}{=} \frac{be^{tb} - ae^{ta}}{b-a} \\
&= 1
\end{aligned}
$$

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

$$
\begin{aligned}
M_X'(t) &= \frac{(be^{tb} - ae^{ta})t(b-a) - (e^{tb} - e^{ta})(b-a)}{t^2(b-a)^2} \\
&= \frac{t(be^{tb} - ae^{ta}) - (e^{tb} - e^{ta})}{t^2(b-a)}
\end{aligned}
$$

$$
\begin{aligned}
\Longrightarrow E(X) &= M_X'(0) = \frac{0}{0} \\
&\overset{\text{L'Hospital}}{=} \frac{be^{tb} - ae^{ta} + tb^2e^{tb} - ta^2e^{ta} - be^{tb} + ae^{ta}}{2t(b-a)} \Bigg|_{t=0} \\
&= \frac{tb^2e^{tb} - ta^2e^{ta}}{2t(b-a)} \Bigg|_{t=0} = \frac{0}{0} \\
&\overset{\text{L'Hospital}}{=} \frac{b^2e^{tb} - a^2e^{ta} + tb^3e^{tb} - ta^3e^{ta}}{2(b-a)} \Bigg|_{t=0} \\
&= \frac{b^2 - a^2}{2(b-a)} \\
&= \frac{b+a}{2}
\end{aligned}
$$

■

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 $| f_n | \leq g$ almost everywhere (i.e., except for a set of Borel–measure 0) and if $f_n \to f$ almost everywhere, then f_n and f are integrable and

$$\int f_n(x)dx \to \int f(x)dx.$$

■

Note:

If f is differentiable with respect to 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

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

■

Theorem 3.2.6:

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

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

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 $\left|\frac{\partial}{\partial\theta} f(x,\theta) \mid_{\theta=\theta_0}\right| \leq g(x,\theta) \;\; \forall x \; \forall\theta_0$ in some ϵ–neighborhood of θ, then

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

■

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 $\mid x \mid \leq e^{\mid x \mid} \quad \forall x$, we get

$$g(x,t) \leq \begin{cases} e^{(t+\delta_0+1)x} f_X(x), & x \geq 0 \\ e^{(t-\delta_0-1)x} f_X(x), & x < 0 \end{cases}$$

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

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

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

<u>Note:</u>
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. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 21: Monday 10/18/1999	
	Scribe: Jürgen Symanzik

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 Homeworks.

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

Theorem 3.2.9:
Let X and Y be 2 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:

$$\begin{aligned} M_Y(t) &= E(e^{(aX+b)t}) \\ &= E(e^{aXt} e^{bt}) \\ &= e^{bt} E(e^{Xat}) \\ &= e^{bt} M_X(at) \end{aligned}$$

■

3.3 Complex–Valued Random Variables and Characteristic Functions

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

Mathematical Operations on Complex Numbers:

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

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

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

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

$\sqrt[n]{z} = \sqrt[n]{a + ib} = \sqrt[n]{r}\left(\cos(\frac{\phi + k \cdot 360^0}{n}) + i\sin(\frac{\phi + k \cdot 360^0}{n})\right)$ for $k = 0, 1, \ldots, (n-1)$ and the main value for $k = 0$

$\ln z = \ln(a + ib) = \ln(\mid z \mid) + i\phi \pm i2n\pi$ where $\phi = \arctan\frac{b}{a}$ and the main value for $n = 0$

Conjugate Complex Numbers:

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

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

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

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

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

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

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

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

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

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

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 $\mid e^{itx} \mid = 1$.

■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 22: Wednesday 10/20/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

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) $\mid \Phi_X(t) \mid \leq 1 \quad \forall t \in I\!R$.

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

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

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

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

(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:

(i) Let $\epsilon > 0$.

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

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

This third condition holds since $\mid e^{\imath 0} - 1 \mid = 0$ and the exponential function is continuous. Therefore, if we select $(t' - t)$ and x small enough, $\mid e^{\imath(t'-t)x} - 1 \mid$ will be $< \frac{\epsilon}{2}$ for a given ϵ.

Let $t, t' \in I\!R$, $t < t'$, and $t' - t < \delta$. Then,

$$\begin{aligned}
\mid \Phi_X(t') - \Phi_X(t) \mid &= \mid \int_{-\infty}^{+\infty} e^{\imath t' x} f_X(x)dx - \int_{-\infty}^{+\infty} e^{\imath t x} f_X(x)dx \mid \\
&= \mid \int_{-\infty}^{+\infty} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid \\
&= \mid \int_{-\infty}^{-a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx + \int_{-a}^{+a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx + \int_{+a}^{+\infty} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid \\
&\leq \mid \int_{-\infty}^{-a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid + \mid \int_{-a}^{+a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid + \mid \int_{+a}^{+\infty} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid
\end{aligned}$$

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

$$\begin{aligned}
\mid \int_{-\infty}^{-a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid &= \mid \int_{-\infty}^{-a} e^{\imath t' x} f_X(x)dx - \int_{-\infty}^{-a} e^{\imath t x} f_X(x)dx \mid \\
&\leq \mid \int_{-\infty}^{-a} e^{\imath t' x} f_X(x)dx \mid + \mid \int_{-\infty}^{-a} e^{\imath t x} f_X(x)dx \mid \\
&\leq \int_{-\infty}^{-a} \mid e^{\imath t' x} \mid f_X(x)dx + \int_{-\infty}^{-a} \mid e^{\imath t x} \mid f_X(x)dx \\
&\overset{(A)}{=} \int_{-\infty}^{-a} 1 f_X(x)dx + \int_{-\infty}^{-a} 1 f_X(x)dx \\
&= \int_{-\infty}^{-a} 2 f_X(x)dx.
\end{aligned}$$

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

Similarly,

$$\mid \int_{+a}^{+\infty} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid \leq \int_{+a}^{+\infty} 2 f_X(x)dx$$

Returning to the main part of the proof, we get

$$\begin{aligned}
\mid \Phi_X(t') - \Phi_X(t) \mid &\leq \int_{-\infty}^{-a} 2 f_X(x)dx + \mid \int_{-a}^{+a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid + \int_{+a}^{+\infty} 2 f_X(x)dx \\
&= 2 \left(\int_{-\infty}^{-a} f_X(x)dx + \int_{+a}^{+\infty} f_X(x)dx \right) + \mid \int_{-a}^{+a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid \\
&= 2P(\mid X \mid > a) + \mid \int_{-a}^{+a} (e^{\imath t' x} - e^{\imath t x}) f_X(x)dx \mid
\end{aligned}$$

$$\overset{Condition\ (ii)}{\leq} 2\frac{\epsilon}{4} + \mid \int_{-a}^{+a} (e^{\imath t' x} - e^{\imath t x}) f_X(x) dx \mid$$

$$= \frac{\epsilon}{2} + \mid \int_{-a}^{+a} e^{\imath t x} (e^{\imath (t'-t) x} - 1) f_X(x) dx \mid$$

$$\leq \frac{\epsilon}{2} + \int_{-a}^{+a} \mid e^{\imath t x} (e^{\imath (t'-t) x} - 1) \mid f_X(x) dx$$

$$\leq \frac{\epsilon}{2} + \int_{-a}^{+a} \mid e^{\imath t x} \mid \cdot \mid (e^{\imath (t'-t) x} - 1) \mid f_X(x) dx$$

$$\overset{(B)}{\leq} \frac{\epsilon}{2} + \int_{-a}^{+a} 1 \frac{\epsilon}{2} f_X(x) dx$$

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

$$= \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

$$= \epsilon$$

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

Theorem 3.3.4: **Bochner's Theorem**

Let $\Phi : I\!R \to \mathbb{C}$ 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 $\displaystyle\int_{-\infty}^{\infty} \mid \Phi_X(t) \mid dt < \infty$. 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 $\lim\limits_{i \to \infty} \Phi_{X_i}(t) = \Phi_X(t) \quad \forall t \in (-h, h)$ 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. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 23: Friday 10/22/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

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:

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

(ii) $\Phi_X(t) = \sum_{k=0}^{1} e^{itk} P(X = k) = e^{it0}(1-p) + e^{it1}p = 1 + p(e^{it} - 1)$

(iii) $\Phi_X(t) = \sum_{n \in I\!N_0} e^{itn} \cdot \frac{c^n}{n!} e^{-c} = e^{-c} \sum_{n=0}^{\infty} \frac{1}{n!} (c \cdot e^{it})^n = e^{-c} \cdot e^{c \cdot e^{it}} = e^{c(e^{it}-1)}$ since $\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x$

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

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

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

$\Rightarrow \Phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \cos(tx) e^{\frac{-x^2}{2}} dx$

Since $E(X)$ exists, $\Phi_X(t)$ is differentiable according to Theorem 3.3.5 and the following holds:

$$\begin{aligned}
\Phi'_X(t) &= Re(\Phi'_X(t)) \\
&= Re\left(\int_{-\infty}^{\infty} ix \underbrace{e^{itx}}_{\cos(tx)+i\sin(tx)} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx\right) \\
&= Re\left(\int_{-\infty}^{\infty} ix\cos(tx)\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx + \int_{-\infty}^{\infty} -x\sin(tx)\frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx\right) \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \underbrace{(-\sin(tx))}_{u}\underbrace{xe^{\frac{-x^2}{2}}}_{v'}dx \quad \mid\ u' = -t\cos(tx) \text{ and } v = -e^{\frac{-x^2}{2}} \\
&= \underbrace{\frac{1}{\sqrt{2\pi}}\sin(tx)e^{\frac{-x^2}{2}} \mid_{-\infty}^{\infty}}_{=0 \text{ since sin is odd}} - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(-t\cos(tx))(-e^{\frac{-x^2}{2}})dx \\
&= -t\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\cos(tx)e^{\frac{-x^2}{2}}dx \\
&= -t\Phi_X(t)
\end{aligned}$$

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

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

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

(vi) For $\sigma > 0, \mu \in I\!R$, we know that if $X \sim N(0,1)$, then $\sigma X + \mu \sim N(\mu, \sigma^2)$. By Theorem 3.3.3 (vii) we have

$$\Phi_{\sigma X+\mu}(t) = e^{it\mu}\Phi_X(\sigma t) = e^{it\mu}e^{-\frac{1}{2}\sigma^2 t^2}.$$

(vii)

$$\begin{aligned}
\Phi_X(t) &= \int_0^{\infty} e^{itx}\gamma(p,q,x)dx \\
&= \int_0^{\infty} \frac{q^p}{\Gamma(p)}x^{p-1}e^{-(q-it)x}dx \\
&= \frac{q^p}{\Gamma(p)}(q-it)^{-p}\int_0^{\infty}((q-it)x)^{p-1}e^{-(q-it)x}(q-it)dx \ \mid\ u = (q-it)x,\ du = (q-it)dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{q^p}{\Gamma(p)}(q-it)^{-p}\underbrace{\int_0^\infty (u)^{p-1}e^{-u}du}_{=\Gamma(p)} \\
&= q^p(q-it)^{-p} \\
&= \frac{(q-it)^{-p}}{q^{-p}} \\
&= \left(\frac{q-it}{q}\right)^{-p} \\
&= \left(1-\frac{it}{q}\right)^{-p}
\end{aligned}
$$

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

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

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

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

■

Example 3.3.13:

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

It is

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

■

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,

- $X \sim$ Dirac(c):

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

which is undefined.

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

$$\begin{aligned}
\int_{-\infty}^{\infty} \mid \Phi_X(t) \mid dt &= \int_{-\infty}^{\infty} \mid 1 + p(e^{it} - 1) \mid dt \\
&= \int_{-\infty}^{\infty} \mid pe^{it} - (p-1) \mid dt \\
&\geq \int_{-\infty}^{\infty} \mid \ \mid pe^{it} \mid - \mid (p-1) \mid \ \mid dt \\
&\geq \int_{-\infty}^{\infty} \mid pe^{it} \mid dt - \int_{-\infty}^{\infty} \mid (p-1) \mid dt \\
&= p \int_{-\infty}^{\infty} 1 dt - (1-p) \int_{-\infty}^{\infty} 1 \ dt \\
&= (2p-1) \int_{-\infty}^{\infty} 1 \ dt \\
&= (2p-1) x \mid_{-\infty}^{\infty}
\end{aligned}$$

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

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

$\blacksquare$

3.4 Probability Generating Functions

Definition 3.4.1:

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

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

■

Theorem 3.4.2:

$G(s)$ converges for $\mid s \mid < 1$.

Proof:

$\mid G(s) \mid \leq \sum_{k=0}^{\infty} \mid p_k s^k \mid \leq \sum_{k=0}^{\infty} \mid p_k \mid = 1$ ■

Theorem 3.4.3:

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

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

■

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:

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

■

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

3.5 Moment Inequalities

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

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

Proof:
Continuous case only:

$$\begin{aligned}
E(h(X)) &= \int_{-\infty}^{\infty} h(x) f_X(x) dx \\
&= \int_A h(x) f_X(x) dx + \int_{A^C} h(x) f_X(x) dx \quad \mid \text{ where } A = \{x : h(x) \geq \epsilon\} \\
&\geq \int_A h(x) f_X(x) dx \\
&\geq \int_A \epsilon f_X(x) dx \\
&= \epsilon P(h(X) \geq \epsilon) \quad \forall \epsilon > 0
\end{aligned}$$

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

Corollary 3.5.2: **Markov's Inequality**
Let $h(X) = \mid X \mid^r$ 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:
Since $P(\mid X \mid \geq k) = P(\mid X \mid^r \geq k^r)$ for $k > 0$, it follows using Theorem 3.5.1:

$$P(\mid X \mid \geq k) = P(\mid X \mid^r \geq k^r) \overset{Th.3.5.1}{\leq} \frac{E(\mid X \mid^r)}{k^r}$$

$\blacksquare$

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:

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

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

$$P(\mid X - \mu \mid > k\sigma) = P(\mid X - \mu \mid^2 > k^2\sigma^2) \overset{Th.3.5.1}{\leq} \frac{E(\mid X - \mu \mid^2)}{k^2\sigma^2} = \frac{Var(X)}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

$\blacksquare$

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:
Let $Q(u, v) = E\left((u \mid X \mid^{\frac{k-1}{2}} + v \mid X \mid^{\frac{k+1}{2}})^2\right)$. Obviously, $Q(u, v) \geq 0 \;\; \forall u, v \in I\!R$. Also,

$$\begin{aligned} Q(u, v) &= \int_{-\infty}^{\infty} (u \mid x \mid^{\frac{k-1}{2}} + v \mid x \mid^{\frac{k+1}{2}})^2 f_X(x) dx \\ &= u^2 \int_{-\infty}^{\infty} \mid x \mid^{k-1} f_X(x) dx + 2uv \int_{-\infty}^{\infty} \mid x \mid^k f_X(x) dx + v^2 \int_{-\infty}^{\infty} \mid x \mid^{k+1} f_X(x) dx \\ &= u^2\beta_{k-1} + 2uv\beta_k + v^2\beta_{k+1} \\ &\geq 0 \;\; \forall u, v \in I\!R \end{aligned}$$

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

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

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

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

This means that $\beta_1^2 \leq \beta_0\beta_2$, $\beta_2^4 \leq \beta_1^2\beta_3^2$, $\beta_3^6 \leq \beta_2^3\beta_4^3$, and so on. Multiplying these, we get:

$$\begin{aligned}
\prod_{j=1}^{k-1} \beta_j^{2j} &\leq \prod_{j=1}^{k-1} \beta_{j-1}^j \beta_{j+1}^j \\
&= (\beta_0\beta_2)(\beta_1^2\beta_3^2)(\beta_2^3\beta_4^3)(\beta_3^4\beta_5^4)\dots(\beta_{k-3}^{k-2}\beta_{k-1}^{k-2})(\beta_{k-2}^{k-1}\beta_k^{k-1}) \\
&= \beta_0 \beta_{k-1}^{k-2}\beta_k^{k-1} \prod_{j=1}^{k-2} \beta_j^{2j}
\end{aligned}$$

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

$$\begin{aligned}
\beta_{k-1}^{2k-2} &\leq \beta_0 \beta_k^{k-1} \beta_{k-1}^{k-2} \\
\overset{\beta_0 = 1}{\Longrightarrow} \beta_{k-1}^k &\leq \beta_k^{k-1} \\
\Longrightarrow \beta_{k-1}^1 &\leq \beta_k^{\frac{k-1}{k}} \\
\Longrightarrow \beta_{k-1}^{\frac{1}{k-1}} &\leq \beta_k^{\frac{1}{k}}
\end{aligned}$$

■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 25: Wednesday 10/27/1999	
	Scribe: Jürgen Symanzik

4 Random Vectors

4.1 Joint, Marginal, and Conditional Distributions

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

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

is the **joint cumulative distribution function** 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 $\lim\limits_{x_k \rightarrow -\infty} F(\underline{x}) = 0 \;\; \forall x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n \in I\!R$.

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 $n = 2$:

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

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

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

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

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

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** 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** of (X,Y). ■

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

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

■

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

Note:

(i)

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

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

(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 $P(X \leq x) = P(X \leq x, -\infty < Y < \infty) = F(x,\infty)$.

■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 26: Friday 10/29/1999	
	Scribe: Jürgen Symanzik

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

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

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

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

■

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:

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

and

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

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

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

and

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

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

4.2 Independent Random Variables

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:

$$\begin{aligned} P(a < X \leq b, c < Y \leq d) &= F_{X,Y}(b, d) - F_{X,Y}(a, d) - F_{X,Y}(b, c) + F_{X,Y}(a, c) \\ &= F_X(b)F_Y(d) - F_X(a)F_Y(d) - F_X(b)F_Y(c) + F_X(a)F_Y(c) \\ &= (F_X(b) - F_X(a))(F_Y(d) - F_Y(c)) \\ &= P(a < X \leq b)P(c < Y \leq d) \end{aligned}$$

■

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. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 27: Monday 11/1/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

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:

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

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

$$\begin{aligned}
\lim_{\underline{a} \uparrow \underline{b}} P(\underline{a} < \underline{X} \leq \underline{b}) &= \lim_{a_i \uparrow b_i \ \forall i \in \{1, \ldots, n\}} P(a_1 < X_1 \leq b_1, \ a_2 < X_2 \leq b_2, \ \ldots, \ a_n < X_n \leq b_n) \\
&= P(X_1 = b_1, \ X_2 = b_2, \ \ldots, \ X_n = b_n) \\
&= P(\underline{X} = \underline{b}) \\
&\overset{indep.}{=} \prod_{i=1}^{n} P(X_i = b_i)
\end{aligned}$$

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

$$
\begin{aligned}
\Longrightarrow & F(x_1., x_2., x_3.) \\
& = P(X_1 = x_{1a})F(x_2., x_3.) \ + \ P(X_1 = x_{1b})F(x_2., x_3.) \ + \ P(X_1 = x_{1c})F(x_2., x_3.) \\
& = F(x_1.)F(x_2., x_3.) \\
& = F(x_1.)[P(X_2 = x_{2a})F(x_3.) \ + \ P(X_2 = x_{2b})F(x_3.) \ + \ P(X_2 = x_{2c})F(x_3.)] \\
& = F(x_1.)[P(X_2 = x_{2a}) \ + \ P(X_2 = x_{2b}) \ + \ P(X_2 = x_{2c})]F(x_3.) \\
& = F(x_1.)F(x_2.)F(x_3.)
\end{aligned}
$$

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

$$
\begin{aligned}
F_{\underline{X}}(\underline{x}) &= \sum_{\underline{x}_i \in B} P(\underline{X} = \underline{x}_i) \\
&= \sum_{\underline{x}_i \in B} P(X_1 = x_{i1}, \ X_2 = x_{i2}, \ \ldots, \ X_n = x_{in}) \\
&\overset{indep.}{=} \sum_{\underline{x}_i \in B} P(X_1 = x_{i1}) P(X_2 = x_{i2}, \ \ldots, \ X_n = x_{in}) \\
&= \sum_{x_{i1} \leq x_1, \ \ldots, \ x_{in} \leq x_n} P(X_1 = x_{i1}) P(X_2 = x_{i2}, \ \ldots, \ X_n = x_{in}) \\
&= \sum_{x_{i1} \leq x_1} P(X_1 = x_{i1}) \sum_{x_{i2} \leq x_2, \ \ldots, \ x_{in} \leq x_n} P(X_2 = x_{i2}, \ \ldots \ X_n = x_{in}) \\
&= F_{X_1}(x_1) \sum_{x_{i2} \leq x_2, \ \ldots, \ x_{in} \leq x_n} P(X_2 = x_{i2}, \ \ldots, \ X_n = x_{in}) \\
&\overset{indep.}{=} F_{X_1}(x_1) \sum_{x_{i2} \leq x_2} P(X_2 = x_{i2}) \sum_{x_{i3} \leq x_3, \ \ldots, \ x_{in} \leq x_n} P(X_3 = x_{i3}, \ \ldots, \ X_n = x_{in}) \\
&= F_{X_1}(x_1) F_{X_2}(x_2) \sum_{x_{i3} \leq x_3, \ \ldots, \ x_{in} \leq x_n} P(X_3 = x_{i3}, \ \ldots, \ X_n = x_{in}) \\
&= \ldots \\
&= F_{X_1}(x_1) F_{X_2}(x_2) \ldots F_{X_n}(x_n) \\
&= \prod_{i=1}^{n} F_{X_i}(x_i)
\end{aligned}
$$

(ii) Continuous case: Homework ■

Theorem 4.2.6:

$X_1, \ldots, X_n$ are independent iff $P(X_i \in A_i,\ i = 1, \ldots, n) = \prod_{i=1}^{n} P(X_i \in A_i)\quad \forall$ 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:

$$\begin{aligned} F_{g(X_1),g(X_2),\ldots,g(X_n)}(h_1, h_2, \ldots, h_n) &= P(g(X_1) \leq h_1, g(X_2) \leq h_2, \ldots, g(X_n) \leq h_n) \\ &\overset{(*)}{=} P(X_1 \in g_1^{-1}(-\infty, h_1], \ldots, X_n \in g_n^{-1}(-\infty, h_n]) \\ &\overset{Th.\ 4.2.6}{=} \prod_{i=1}^{n} P(X_i \in g_i^{-1}(-\infty, h_i]) \\ &= \prod_{i=1}^{n} P(g_i(X_i) \leq h_i) \\ &= \prod_{i=1}^{n} F_{g_i(X_i)}(h_i) \end{aligned}$$

$(*)$ holds since $g_1^{-1}(-\infty, h_1] \in B, \ldots, g_n^{-1}(-\infty, h_n] \in B$ ■

Theorem 4.2.8:

If $X_1, \ldots, X_n$ are independent, then also every subcollection $X_{i_1}, \ldots, X_{i_k},\ k = 2, \ldots, n-1$, $1 \leq i_1 < i_2 \ldots < i_k \leq n$, is independent. ■

Definition 4.2.9:

A set (or a sequence) of rv's $\{X_n\}_{n=1}^{\infty}$ is independent iff every finite subcollection is independent. ■

Note:

Recall that X and Y are *identically distributed* iff $F_X(x) = F_Y(x)\quad \forall x \in I\!R$ 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

Theorem 4.3.1:

If X and Y are rv's on $(\Omega, L, P) \rightarrow I\!R$, 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

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

and

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

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

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

and

$$F_{MIN_n}(z) = 1 - (1 - F_X(z))^n.$$

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

$$f_{MAX_n}(z) = n \cdot F_X^{n-1}(z) \cdot f_X(z)$$

and

$$f_{MIN_n}(z) = n \cdot (1 - F_X(z))^{n-1} \cdot f_X(z)$$

for all continuity points of F_X. ■

Note:

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

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

However, note that MAX_n and MIN_n are not independent. See Rohatgi, page 129, Corollary, for more details. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 28: Wednesday 11/3/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

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

Theorem 4.3.3:
If $g : I\!R^n \to I\!R^m$ 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 $V = Y + 1$. Then $X = UV$ and $Y = V - 1$. So the joint pmf of U, V is

$$
\begin{aligned}
P(U = u, V = v) &= \binom{n}{uv} p^{uv}(1-p)^{n-uv} \binom{n}{v-1} p^{v-1}(1-p)^{n+1-v} \\
&= \binom{n}{uv}\binom{n}{v-1} p^{uv+v-1}(1-p)^{2n+1-uv-v}
\end{aligned}
$$

for $v \in \{1, 2, \ldots, n+1\}$ and $uv \in \{0, 1, \ldots, n\}$. ■

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

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

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

$$
B_{\underline{u}} = \{(u_1', \ldots, u_n') : -\infty < u_i' < u_i \quad \forall i = 1, \ldots, n\}
$$

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

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

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 $I\!R^n$ into $I\!R^n$, 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 $\underline{u} \in I\!R^n$ and
$$B_{\underline{u}} = \{(u'_1, \ldots, u'_n) : -\infty < u'_i < u_i \quad \forall i = 1, \ldots, n\}.$$
Then,
$$\begin{aligned} G_{\underline{U}}(\underline{u}) &= \underset{g^{-1}(B_{\underline{u}})}{\int \cdots \int} f_{\underline{X}}(\underline{x}) d(\underline{x}) \\ &= \underset{B_{\underline{u}}}{\int \cdots \int} f_{\underline{X}}(h_1(\underline{u}), \ldots, h_n(\underline{u})) \mid J \mid d(\underline{u}) \end{aligned}$$
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 $\mathbb{R}^n$ into $\mathbb{R}^n$.

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

(iii) Suppose that for each $\underline{u} \in B = \{\underline{u} \in \mathbb{R}^n : \underline{u} = g(\underline{x}) \text{ for some } \underline{x} \in \mathcal{X}\}$ there is a finite number $k = k(\underline{u})$ of inverses.

(iv) Suppose we can partition $\mathcal{X}$ into $\mathcal{X}_0, \mathcal{X}_1, \ldots, X_k$ 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

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

$\mathcal{X} = I\!R^2$, but U, V are not 1–to–1 mappings since $(U,V)(x,y) = (U,V)(-x,-y)$, i.e., conditions do not apply for the use of Theorem 4.3.5. Let

$$\begin{aligned}
\mathcal{X}_0 &= \{(x,y) : y = 0\} \\
\mathcal{X}_1 &= \{(x,y) : y > 0\} \\
\mathcal{X}_2 &= \{(x,y) : y < 0\}
\end{aligned}$$

Then $P((X,Y) \in \mathcal{X}_0) = 0$.

Let $B = \{(u,v) : v > 0\} = g(\mathcal{X}_1) = g(\mathcal{X}_2)$.

Inverses:

$$\begin{aligned}
B \to \mathcal{X}_1 : x &= h_{11}(u,v) = uv \\
y &= h_{12}(u,v) = v \\
B \to \mathcal{X}_2 : x &= h_{21}(u,v) = -uv \\
y &= h_{22}(u,v) = -v \\
J_1 &= \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} \Rightarrow |J_1| = \mid v \mid \\
J_2 &= \begin{vmatrix} -v & -u \\ 0 & -1 \end{vmatrix} \Rightarrow |J_2| = \mid v \mid \\
f_{X,Y}(x,y) &= \frac{1}{2\pi} e^{-x^2/2} e^{-y^2/2} \\
f_{U,V}(u,v) &= \mid v \mid \frac{1}{2\pi} e^{-(uv)^2/2} e^{-v^2/2} + \mid v \mid \frac{1}{2\pi} e^{-(-uv)^2/2} e^{-(-v)^2/2} \\
&= \frac{v}{\pi} e^{\frac{-(u^2+1)v^2}{2}}, \quad -\infty < u < \infty, \quad 0 < v < \infty
\end{aligned}$$

Marginal:

$$\begin{aligned}
f_U(u) &= \int_0^\infty \frac{v}{\pi} e^{\frac{-(u^2+1)v^2}{2}} dv \quad \mid \quad z = \frac{(u^2+1)v^2}{2}, \quad \frac{dz}{dv} = (u^2+1)v \\
&= \int_0^\infty \frac{1}{\pi(u^2+1)} e^{-z} dz
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\pi(u^2+1)}(-e^{-z})\Big|_0^\infty \\
&= \frac{1}{\pi(1+u^2)}\,, \quad -\infty < u < \infty
\end{aligned}
$$

Thus, the quotient of two iid $N(0,1)$ rv's is a rv that has a Cauchy distribution. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 29: Friday 11/5/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

4.4 Order Statistics

Definition 4.4.1:

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

Theorem 4.4.3:

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

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

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:

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

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

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

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

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

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

For $n = 3$, there are $3! = 6$ possible arrangements. In general, there are $n!$ arrangements of $X_1, \ldots, X_n$ for each $(X_{(1)}, \ldots, X_{(n)})$. This mapping is not 1–to–1. For each mapping, we have a $n \times n$ matrix J that results from an $n \times n$ identity matrix through the rearrangement or rows. Therefore, $\mid J \mid = 1$. By Theorem 4.3.6, we get

$$\begin{aligned} f_{X_{(1)},\ldots,X_{(n)}}(x_{(1)}, \ldots, x_{(n)}) &= n! f_{X_1,\ldots,X_n}(x_{(k_1)}, x_{(k_2)}, \ldots, x_{(k_n)}) \\ &= n! \prod_{i=1}^{n} f_{X_i}(x_{(k_i)}) \\ &= n! \prod_{i=1}^{n} f_X(x_i) \end{aligned}$$

■

<u>Theorem 4.4.4:</u> 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

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

if $x_j < x_k$ and 0 otherwise.

■

4.5 Multivariate Expectation

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

Definition 4.5.1:

If $n = 1$, 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

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

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:

$$\begin{aligned}
E(\mid S \mid) &= \int_{I\!R^n} \mid \sum_{i=1}^{n} a_i x_i \mid f_{\underline{X}}(\underline{x}) d\underline{x} \\
&\leq \int_{I\!R^n} \sum_{i=1}^{n} \mid a_i \mid \cdot \mid x_i \mid f_{\underline{X}}(\underline{x}) d\underline{x} \\
&= \sum_{i=1}^{n} \mid a_i \mid \int_{I\!R} \mid x_i \mid \left(\int_{I\!R^{n-1}} f_{\underline{X}}(\underline{x}) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n \right) dx_i \\
&= \sum_{i=1}^{n} \mid a_i \mid \int_{I\!R} \mid x_i \mid f_{X_i}(x_i) dx_i \\
&= \sum_{i=1}^{n} \mid a_i \mid E(\mid X_i \mid) \\
&< \infty
\end{aligned}$$

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

■

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

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

■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 30: Monday 11/8/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

<u>Theorem 4.5.3:</u>

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.

<u>Proof:</u>

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,

$$\begin{aligned}
E(\prod_{i=1}^{n} g_i(x_i)) &= \int_{\mathbb{R}^n} \prod_{i=1}^{n} g_i(x_i) f_{\underline{X}}(\underline{x}) d\underline{x} \\
&\overset{Th. 4.2.5}{=} \int_{\mathbb{R}^n} \prod_{i=1}^{n} (g_i(x_i) f_{X_i}(x_i) dx_i) \\
&= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{i=1}^{n} g_i(x_i) \prod_{i=1}^{n} f_{X_i}(x_i) \prod_{i=1}^{n} dx_i \\
&\overset{Th. 4.2.7}{=} \int_{\mathbb{R}} g_1(x_1) f_{X_1}(x_1) dx_1 \int_{\mathbb{R}} g_2(x_2) f_{X_2}(x_2) dx_2 \ldots \int_{\mathbb{R}} g_n(x_n) f_{X_n}(x_n) dx_n \\
&= \prod_{i=1}^{n} \int_{\mathbb{R}} g_i(x_i) f_{X_i}(x_i) dx_i \\
&= \prod_{i=1}^{n} E(g_i(X_i))
\end{aligned}$$

■

<u>Corollary 4.5.4:</u>

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

<u>Theorem 4.5.5:</u>

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.

<u>Proof:</u>

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

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

"$\Longleftarrow$": From Theorem 4.2.6, we know that X and Y are independent iff $P(X \in A_1,\ Y \in A_2) = P(X \in A_1)\ P(Y \in A_2)\ \ \forall$ Borel sets A_1 and A_2.

How do we relate Theorem 4.2.6 to g_1 and g_2? Let us define two Borel–measurable functions g_1 and g_2 as:

$$\begin{aligned} g_1(x) &= I_{A_1}(x) = \begin{cases} 1, & x \in A_1 \\ 0, & \text{otherwise} \end{cases} \\ g_2(y) &= I_{A_2}(y) = \begin{cases} 1, & y \in A_2 \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

Then,

$$\begin{aligned} E(g_1(X)) &= 0 \cdot P(X \in A_1^c) + 1 \cdot P(X \in A_1) = P(X \in A_1), \\ E(g_2(Y)) &= 0 \cdot P(Y \in A_2^c) + 1 \cdot P(Y \in A_2) = P(Y \in A_2) \end{aligned}$$

and

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

$\Longrightarrow P(X \in A_1, Y \in A_2) = E(g_1(X) \cdot g_2(Y)) \overset{given}{=} E(g_1(X)) \cdot E(g_2(Y)) = P(X \in A_1) P(Y \in A_2)$

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

$\Longrightarrow X,\ Y$ independent by Theorem 4.2.6. ∎

Definition 4.5.6:

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

Note:

If we set $i_r = i_s = 1$ and $i_j = 0\ \ \forall j \neq r, s$ in Definition 4.5.6, we get

$$\mu_{\;0\ldots0\ \underset{\underset{r}{\uparrow}}{1}\ 0\ldots0\ \underset{\underset{s}{\uparrow}}{1}\ 0\ldots0} = \mu_{rs} = Cov(X_r, X_s).$$

∎

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

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

(i) $Cov(X, Y)$ 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 $(\alpha, \beta) \in I\!R^2 - \{(0,0)\}$ 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:

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

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

$\Longrightarrow \mid ab \mid \ \leq \frac{a^2+b^2}{2}\ \forall\ a, b \in I\!R \quad (*)$

(i)

$$\begin{aligned}
E(\mid XY \mid) &= \int_{I\!R^2} \mid xy \mid f_{X,Y}(x,y)dx\ dy \\
&\overset{(*)}{\leq} \int_{I\!R^2} \frac{x^2+y^2}{2} f_{X,Y}(x,y)dx\ dy \\
&= \int_{I\!R^2} \frac{x^2}{2} f_{X,Y}(x,y)dx\ dy + \int_{I\!R^2} \frac{y^2}{2} f_{X,Y}(x,y)dy\ dx \\
&= \int_{I\!R} \frac{x^2}{2} f_X(x)dx + \int_{I\!R} \frac{y^2}{2} f_Y(y)dy \\
&= \frac{E(X^2)\ +\ E(Y^2)}{2} \\
&< \infty
\end{aligned}$$

$\Longrightarrow E(XY)$ exists

$\Longrightarrow Cov(X,Y) = E(XY) - E(X)E(Y)$ exists

(ii) $0 \leq E((\alpha X + \beta Y)^2) = \alpha^2 E(X^2)\ +\ 2\alpha\beta E(XY)\ +\ \beta^2 E(Y^2)\ \forall\ \alpha, \beta \in I\!R \quad (A)$

If $E(X^2) = 0$, then X has a degenerate 1–point Dirac distribution and the inequality trivially is true. Therefore, we can assume that $E(X^2) > 0$. As (A) is true for all $\alpha, \beta \in I\!R$, we can choose $\alpha = \frac{-E(XY)}{E(X^2)},\ \beta = 1$.

$\Longrightarrow \frac{(E(XY))^2}{E(X^2)} - 2\frac{(E(XY))^2}{E(X^2)} + E(Y^2) \geq 0$

$\Longrightarrow -(E(XY))^2 + E(Y^2)E(X^2) \geq 0$

$\Longrightarrow (E(XY))^2 \leq E(X^2)\ E(Y^2)$

(iii) When are the left and right sides of the inequality in (ii) equal?

Assume that $E(X^2) > 0$. $(E(XY))^2 = E(X^2)E(Y^2)$ holds iff $E((\alpha X + \beta Y)^2) = 0$. This can only happen if $P(\alpha X + \beta Y = 0) = 1$.

Otherwise, if $P(\alpha X + \beta Y = 0) = P(Y = -\frac{\alpha}{\beta}X) = 1$ for some $(\alpha, \beta) \in I\!R^2 - \{(0,0)\}$, i.e., Y is linearly dependent on X with probability 1, this implies:

$$(E(XY))^2 = (E(X \cdot \frac{-\alpha X}{\beta}))^2 = (\frac{\alpha}{\beta})^2(E(X^2))^2 = E(X^2)(\frac{\alpha}{\beta})^2 E(X^2) = E(X^2)E(Y^2)$$

■

4.6 Multivariate Generating Functions

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 $|\ \underline{t}\ | = \sqrt{\sum_{i=1}^{n} t_i^2} < h$ 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** $\Phi_{\underline{X}} : I\!R^n \to \mathbb{C}$ 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

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

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

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

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

Proof:
Rohatgi, page 162: Theorem 7, Corollary, Theorem 8, and Theorem 9 (for mmgf and the case $n = 2$). ■

Theorem 4.6.4:
Let $X_1, \ldots, X_n$ be independent rv's.

(i) If mgf's $M_{X_1}(t), \ldots, M_{X_n}(t)$ exist, then the mgf of $Y = \sum_{i=1}^{n} a_i X_i$ is

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

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

(ii) The characteristic function of $Y = \sum_{j=1}^{n} a_j X_j$ is

$$\Phi_Y(t) = \prod_{j=1}^{n} \Phi_{X_j}(a_j t)$$

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

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

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

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

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

Proof:
Homework (parts (ii) and (iv) only) ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 31: Wednesday 11/10/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

<u>Theorem 4.6.5:</u>
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 $Y = \sum\limits_{i=1}^n X_i$ is

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

<u>Proof:</u>
<u>Version 1:</u>

$$\begin{aligned}
G_{X_i}(s) &= E(s^{X_i}) \\
G_Y(s) &= E(s^Y) \\
&= E(s^{\sum_{i=1}^n X_i}) \\
&\overset{indep.}{=} \prod_{i=1}^n E(s^{X_i}) \\
&= \prod_{i=1}^n G_{X_i}(s)
\end{aligned}$$

<u>Version 2:</u> (case $n = 2$ only)

$$\begin{aligned}
G_Y(s) &= P(Y=0) + P(Y=1)s + P(Y=2)s^2 + \ldots \\
&= P(X_1 = 0, X_2 = 0) + \\
&\qquad (P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1))\, s + \\
&\qquad (P(X_1 = 2, X_2 = 0) + P(X_1 = 1, X_2 = 1) + P(X_1 = 0, X_2 = 2))\, s^2 + \ldots \\
&\overset{indep.}{=} P(X_1 = 0)P(X_2 = 0) + \\
&\qquad (P(X_1 = 1)P(X_2 = 0) + P(X_1 = 0)P(X_2 = 1))\, s + \\
&\qquad (P(X_1 = 2)P(X_2 = 0) + P(X_1 = 1)P(X_2 = 1) + P(X_1 = 0)P(X_2 = 2))\, s^2 + \ldots \\
&= \Big(P(X_1 = 0) + P(X_1 = 1)s + P(X_1 = 2)s^2 + \ldots\Big) \cdot \\
&\qquad \Big(P(X_2 = 0) + P(X_2 = 1)s + P(X_2 = 2)s^2 + \ldots\Big) \\
&= G_{X_1}(s) \cdot G_{X_2}(s)
\end{aligned}$$

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 $S_N = \sum\limits_{i=1}^{N} X_i$. The pgf of S_N is

$$G_{S_N}(s) = G_N(G_X(s)).$$

Proof:

$$\begin{aligned}
P(S_N = k) &= \sum_{n=0}^{\infty} P(S_N = k | N = n) \cdot P(N = n) \\
\Longrightarrow G_{S_N}(s) &= \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} P(S_N = k | N = n) \cdot P(N = n) \cdot s^k \\
&= \sum_{n=0}^{\infty} P(N = n) \sum_{k=0}^{\infty} P(S_N = k | N = n) \cdot s^k \\
&= \sum_{n=0}^{\infty} P(N = n) \sum_{k=0}^{\infty} P(S_n = k) \cdot s^k \\
&= \sum_{n=0}^{\infty} P(N = n) \sum_{k=0}^{\infty} P(\sum_{i=1}^{n} X_i = k) \cdot s^k \\
&\overset{Th.4.6.5}{=} \sum_{n=0}^{\infty} P(N = n) \prod_{i=1}^{n} G_{X_i}(s) \\
&\overset{iid}{=} \sum_{n=0}^{\infty} P(N = n) \cdot (G_X(s))^n \\
&= G_N(G_X(s))
\end{aligned}$$

■

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?

$$\begin{aligned}
G_X(s) &= G_N(s) = ps^2 + qs + r \\
G_{S_N}(s) &= p(ps^2 + ps + r)^2 + q(ps^2 + ps + r) + r
\end{aligned}$$

■

Theorem 4.6.8:

Let $X_1, \ldots, X_N$ be iid rv's with common mgf $M_X(t)$. Let N be a discrete rv on the non–negative integers with mgf $M_N(t)$. Let N be independent of the X_i's. Define $S_N = \sum_{i=1}^{N} X_i$. The mgf of S_N is

$$M_{S_N}(t) = M_N(\ln M_X(t)).$$

Proof:

Consider the case that the X_i's are non–negative integers:

We know that

$$\begin{aligned}
G_X(S) &= E(S^X) = E(e^{\ln S^X}) = E(e^{(\ln S) \cdot X}) = M_X(\ln S) \\
\Longrightarrow M_X(S) &= G_X(e^S) \\
\Longrightarrow M_{S_N}(t) &= G_{S_N}(e^t) \stackrel{Th.4.6.6}{=} G_N(G_X(e^t)) = G_N(M_X(t)) = M_N(\ln M_X(t))
\end{aligned}$$

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. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 32: Friday 11/12/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

4.7 Conditional Expectation

In Section 4.1, we established that the conditional pmf of X given $Y = y_j$ (for $P_Y(y_j) > 0$) is a pmf. For continuous rv's X and Y, when $f_Y(y) > 0$, $f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$, and $f_{X,Y}$ and f_Y are continuous, then $f_{X|Y}(x \mid y)$ is a pdf and it is the conditional pdf of X given $Y = y$.

Definition 4.7.1:
Let X, Y be rv's on (Ω, L, P). Let h be a Borel–measurable function. Assume that $E(h(X))$ exists. Then the **conditional expectation** of $h(X)$ given Y, i.e., $E(h(X) \mid Y)$, is a rv that takes the value $E(h(X) \mid y)$. It is defined as

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

■

Note:

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

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

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

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

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

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

(iii) Moments are defined in the usual way. If $E(\mid X \mid^r) < \infty$, 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

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

So,

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

and

$$E(Y \mid x) = \int_x^1 \frac{1}{1-x}\ y\ dy = \frac{1}{1-x} \frac{y^2}{2}\bigg|_x^1 = \frac{1}{2}\frac{1-x^2}{1-x} = \frac{1+x}{2}.$$

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

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:

$$\begin{aligned} E_Y(E(h(X) \mid Y)) &= \int_{-\infty}^{\infty} E(h(x) \mid y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x) f_{X|Y}(x \mid y) f_Y(y) dx dy \\ &= \int_{-\infty}^{\infty} h(x) \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx \\ &= \int_{-\infty}^{\infty} h(x) f_X(x) dx \\ &= E(h(X)) \end{aligned}$$

■

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:

$$\begin{aligned}
Var_Y(E(X \mid Y)) + E_Y(Var(X \mid Y)) &= E_Y((E(X \mid Y))^2) - (E_Y(E(X \mid Y)))^2 \\
&\quad + E_Y(E(X^2 \mid Y) - (E(X \mid Y))^2) \\
&= E_Y((E(X \mid Y))^2) - (E(X))^2 + E(X^2) - E_Y((E(X \mid Y))^2) \\
&= E(X^2) - (E(X))^2 \\
&= Var(X)
\end{aligned}$$

■

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

For equality, it is necessary that $E_Y(Var(X \mid Y)) = E_Y((X - E(X \mid Y))^2 \mid Y) = 0$ which holds if $X = E(X \mid Y) = g(Y)$.

If X, Y are independent, $F_{X|Y}(x \mid y) = F_X(x) \;\; \forall x$. Thus, if $E(h(X))$ exists, then $E(h(X) \mid Y) = E(h(X))$. ■

4.8 Inequalities and Identities

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

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

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

Proof:
Fix b. Let

$$\begin{aligned}
g(a) &= \frac{1}{p}a^p + \frac{1}{q}b^q - ab \\
\Longrightarrow g'(a) &= a^{p-1} - b \overset{!}{=} 0 \\
\Longrightarrow b &= a^{p-1} \\
\Longrightarrow b^q &= a^{(p-1)q} = a^p \\
g''(a) &= (p-1)a^{p-2} > 0
\end{aligned}$$

Since $g''(a) > 0$, this is really a minimum. The minimum value is obtained for $b = a^{p-1}$ and it is

$$\frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p = a^p(\frac{1}{p} + \frac{1}{q} - 1) = 0.$$

Since $g''(a) > 0$, the minimum is unique and $g(a) \geq 0$. Therefore $g(a) + ab = \frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$. ■

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 $a = \frac{|X|}{(E(|X|^p))^{\frac{1}{p}}}$ and $b = \frac{|Y|}{(E(|Y|^q))^{\frac{1}{q}}}$.

$$\stackrel{Lemma\ 4.8.1}{\Longrightarrow} \frac{1}{p}\frac{|X|^p}{E(|X|^p)} + \frac{1}{q}\frac{|Y|^q}{E(|Y|^q)} \geq \frac{|XY|}{(E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}}}$$

Taking expectations on both sides of this inequality, we get

$$1 = \frac{1}{p} + \frac{1}{q} \geq \frac{E(|XY|)}{(E(|X|^p))^{\frac{1}{p}}(E(|Y|^q))^{\frac{1}{q}}}$$

The result follows immediately when multiplying both sides with $(E(\mid X \mid^p))^{\frac{1}{p}} (E(\mid Y \mid^q))^{\frac{1}{q}}$. ■

Note:

Note that Theorem 4.5.7 (ii) (Cauchy–Schwarz–Inequality) is a special case of Theorem 4.8.2 with $p = q = 2$. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 33: Monday 11/15/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

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:

$$\begin{aligned}
E(\mid X+Y \mid^p) &= E(\mid X+Y \mid \cdot \mid X+Y \mid^{p-1}) \\
&\leq E((\mid X \mid + \mid Y \mid) \cdot \mid X+Y \mid^{p-1}) \\
&= E(\mid X \mid \cdot \mid X+Y \mid^{p-1}) \ + \ E(\mid Y \mid \cdot \mid X+Y \mid^{p-1}) \\
&\overset{Th.4.8.2}{\leq} (\ E(\mid X \mid^p)\ )^{\frac{1}{p}} \cdot (\ E(\ (\mid X \ + \ Y \mid^{p-1})^q\ )\ )^{\frac{1}{q}} \\
&\quad\ + (\ E(\mid Y \mid^p)\ )^{\frac{1}{p}} \cdot (\ E(\ (\mid X \ + \ Y \mid^{p-1})^q\ )\ )^{\frac{1}{q}} \quad (A) \\
&\overset{(*)}{=} \Big((\ E(\mid X \mid^p)\ )^{\frac{1}{p}} + (\ E(\mid Y \mid^p)\ )^{\frac{1}{p}} \Big) \cdot (\ E(\mid X \ + \ Y \mid^p\ )\ )^{\frac{1}{q}}
\end{aligned}$$

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

The result on the left side is $(E(\mid X+Y \mid^p))^{1-\frac{1}{q}} = (E(\mid X+Y \mid^p))^{\frac{1}{p}}$, and the result on the right side is $(\ E(\mid X \mid^p)\ )^{\frac{1}{p}} + (\ E(\mid Y \mid^p)\ )^{\frac{1}{p}}$. Therefore, Theorem 4.8.3 holds.

In (A), we define $q = \frac{p}{p-1}$. Therefore, $\frac{1}{p} + \frac{1}{q} = 1$ and $(p-1)q = p$.

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

Definition 4.8.4:

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

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

∎

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 $x_0 = E(X)$:

$$l(x) = ax + b \text{ for some } a, b \in I\!R$$

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

$$\Longrightarrow g(x) \;\geq\; ax + b \;\; \forall x \in I\!R$$

$$\Longrightarrow E(g(X)) \geq E(aX + b) = aE(X) + b = l(E(X)) \overset{\text{tangent point}}{=} g(E(X))$$

Therefore, Theorem 4.8.5 holds. ■

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., $E(g(X)) \leq g(E(X))$.

■

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

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

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

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

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

(i) $a_A \geq a_G$:

$$\begin{aligned}
\ln(a_A) &= \ln\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) \\
&= \ln(E(X)) \\
&\overset{ln\ concave}{\geq} E(\ln(X)) \\
&= \sum_{i=1}^{n}\frac{1}{n}\ln(a_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\ln(a_i) \\
&= \frac{1}{n}\ln(\prod_{i=1}^{n} a_i) \\
&= \ln((\prod_{i=1}^{n} a_i)^{\frac{1}{n}}) \\
&= \ln(a_G)
\end{aligned}$$

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

(ii) $a_A \geq a_H$:

$$\begin{aligned}
\frac{1}{a_A} &= \frac{1}{\frac{1}{n}\sum_{i=1}^{n} a_i} \\
&= \frac{1}{E(X)} \\
&\overset{1/X\ convex}{\leq} E(\frac{1}{X}) \\
&= \sum_{i=1}^{n}\frac{1}{n}\frac{1}{a_i} \\
&= \frac{1}{n}\left(\frac{1}{a_1} + \frac{1}{a_2} + \ldots \frac{1}{a_n}\right) \\
&= \frac{1}{a_H}
\end{aligned}$$

Inverting both sides gives $a_A \geq a_H$.

(iii) $a_G \geq a_H$:

$$\begin{aligned}
-\ln(a_H) &= \ln(a_H^{-1}) \\
&= \ln(\frac{1}{a_H}) \\
&= \ln\left(\frac{1}{n}(\frac{1}{a_1} + \frac{1}{a_2} + \ldots + \frac{1}{a_n})\right) \\
&= \ln(E(\frac{1}{X})) \\
&\overset{ln\ concave}{\geq} E(\ln(\frac{1}{X})) \\
&= \sum_{i=1}^{n} \frac{1}{n} \ln(\frac{1}{a_i}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \ln(\frac{1}{a_i}) \\
&= \frac{1}{n} \sum_{i=1}^{n} - \ln a_i \\
&= -\frac{1}{n} \ln(\prod_{i=1}^{n} a_i) \\
&= -\ln(\prod_{i=1}^{n} a_i)^{\frac{1}{n}} \\
&= -\ln a_G
\end{aligned}$$

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

In summary, $a_H \leq a_G \leq a_A$. Note that it would have been sufficient to prove steps (i) and (iii) only to establish this result. However, step (ii) has been included to provide another example how to apply Theorem 4.8.5. ■

Theorem 4.8.7: **Covariance Inequality**
Let X be a rv with finite mean 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.

<u>Proof:</u>
Homework ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 34: Wednesday 11/17/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

5 Particular Distributions

5.1 Multivariate Normal Distributions

Definition 5.1.1:
A rv X has a **(univariate) Normal distribution**, i.e., $X \sim N(\mu, \sigma^2)$ with $\mu \in I\!R$ 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 $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ 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:
$E(X) = \mu_1 + a_{11}E(Z_1) + a_{12}E(Z_2) = \mu_1$ 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 $a_{i1} = a_{i2} = 0$. ■

Theorem 5.1.3:
Define $\underline{g} : I\!R^2 \to I\!R^2$ as $\underline{g}(\underline{x}) = C\underline{x} + \underline{d}$. If $\underline{X}$ is a bivariate Normal rv, then $\underline{g}(\underline{X})$ also is a bivariate Normal rv.

Proof:

$$
\begin{aligned}
g(\underline{X}) &= C\underline{X} + \underline{d} \\
&= C(A\underline{Z} + \underline{\mu}) + \underline{d} \\
&= \underbrace{(CA)}_{\text{another matrix}} \quad \underline{Z} + \underbrace{(C\underline{\mu} + \underline{d})}_{\text{another vector}} \\
&= \tilde{A}\underline{Z} + \underline{\tilde{\mu}} \quad \text{which represents another bivariate Normal distribution}
\end{aligned}
$$

■

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

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

■

Theorem 5.1.5:
Assume that $\sigma_1 > 0, \sigma_2 > 0$ and $\mid \rho \mid < 1$. 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:
The mapping $\underline{Z} \to \underline{X}$ is 1–to–1:

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

$$
\begin{aligned}
\Longrightarrow \underline{Z} &= A^{-1}(\underline{X} - \underline{\mu}) \qquad \text{(requirement is that A is invertible.)} \\
J &= |A^{-1}| = \frac{1}{|A|} \\
|A| &= \sqrt{|A|^2} = \sqrt{|A| \cdot |A^T|} = \sqrt{|AA^T|} = \sqrt{|\Sigma|} = \sqrt{\sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2} = \sigma_1\sigma_2\sqrt{1-\rho^2}
\end{aligned}
$$

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

$$
\begin{aligned}
f_{\underline{Z}}(\underline{z}) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z_1^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_2^2}{2}} \\
&= \frac{1}{2\pi} e^{-\frac{1}{2}(\underline{z}^T \underline{z})}
\end{aligned}
$$

As already stated, the mapping from $\underline{Z}$ to $\underline{X}$ is 1–to–1, so we can apply Theorem 4.3.5:

$$
\begin{aligned}
f_{\underline{X}}(\underline{x}) &= \frac{1}{2\pi\sqrt{|\Sigma|}} \exp(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \underbrace{(A^{-1})^T A^{-1}}_{\Sigma^{-1}} (\underline{x} - \underline{\mu})) \\
&\stackrel{(*)}{=} \frac{1}{2\pi\sqrt{|\Sigma|}} \exp(-\frac{1}{2}(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}))
\end{aligned}
$$

This proves the 1^{st} line of the Theorem. Step $(*)$ holds since

$$
(A^{-1})^T A^{-1} = (A^T)^{-1} A^{-1} = (AA^T)^{-1} = \Sigma^{-1}.
$$

The second line of the Theorem is based on the following transformations:

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

■

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

Theorem 5.1.6:
If (X, Y) has a non–degenerate $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ distribution, then the conditional distribution of X given $Y = y$ is

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

Proof:

Homework ■

Example 5.1.7:

Let rv's (X_1, Y_1) be $N(0,0,1,1,0)$ with pdf $f_1(x,y)$ and (X_2, Y_2) be $N(0,0,1,1,\rho)$ with pdf $f_2(x,y)$. Let (X, Y) be the rv that corresponds to the pdf

$$f_{X,Y}(x,y) = \frac{1}{2} f_1(x,y) + \frac{1}{2} f_2(x,y).$$

(X, Y) is a bivariate Normal rv iff $\rho = 0$. However, the marginal distributions of X and Y are always $N(0,1)$ distributions. See also Rohatgi, page 229, Remark 2. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 35: Friday 11/19/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

Theorem 5.1.8:
The mgf $M_{\underline{X}}(\underline{t})$ of a 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 $\underline{X} = (X, Y)$:

$$\begin{aligned}
M_X(t) &= E(\exp(tX)) \\
&= \int_{-\infty}^{\infty} \exp(tx) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}[-2\sigma^2 tx + (x-\mu)^2]\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}[-2\sigma^2 tx + (x^2 - 2\mu x + \mu^2)]\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}[x^2 - 2(\mu + \sigma^2 t)x + (\mu + t\sigma^2)^2 - (\mu + t\sigma^2)^2 + \mu^2]\right) dx \\
&= \exp\left(-\frac{1}{2\sigma^2}[-(\mu + t\sigma^2)^2 + \mu^2]\right) \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}[x - (\mu + t\sigma^2)]^2\right) dx}_{\text{pdf of } N(\mu + t\sigma^2, \sigma^2), \text{ that integrates to } 1} \\
&= \exp\left(-\frac{1}{2\sigma^2}[-\mu^2 - 2\mu t\sigma^2 - t^2\sigma^4 + \mu^2]\right) \\
&= \exp\left(\frac{-2\mu t\sigma^2 - t^2\sigma^4}{-2\sigma^2}\right) \\
&= \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)
\end{aligned}$$

Bivariate Normal mgf:

$$M_{X,Y}(t_1, t_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(t_1 x + t_2 y)\, f_{X,Y}(x, y)\, dx\, dy$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp(t_1 x)\ \exp(t_2 y)\ f_X(x)\ f_{Y|X}(y \mid x)\ dy\ dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \exp(t_2 y)\ f_{Y|X}(y \mid x)\ dy\right)\ \exp(t_1 x)\ f_X(x)\ dx \\
&\stackrel{(A)}{=} \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} \frac{\exp(t_2 y)}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}}\ \exp\left(\frac{-(y-\beta_X)^2}{2\sigma_2^2(1-\rho^2)}\right)\ dy\right)\ \exp(t_1 x)\ f_X(x)\ dx \\
&\qquad\qquad \mid \quad \beta_X = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1) \\
&\stackrel{(B)}{=} \int_{-\infty}^{\infty} \exp\left(\beta_X t_2 + \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2\right)\ \exp(t_1 x)\ f_X(x)\ dx \\
&= \int_{-\infty}^{\infty} \exp\left([\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)]t_2 + \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2 + t_1 x\right)\ f_X(x)\ dx \\
&= \int_{-\infty}^{\infty} \exp\left(\mu_2 t_2 + \rho\frac{\sigma_2}{\sigma_1}t_2 x - \rho\frac{\sigma_2}{\sigma_1}\mu_1 t_2 + \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2 + t_1 x\right)\ f_X(x)\ dx \\
&= \exp\left(\frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2 + t_2\mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1 t_2\right) \int_{-\infty}^{\infty} \exp\left((t_1 + \rho\frac{\sigma_2}{\sigma_1}t_2)x\right)\ f_X(x)\ dx \\
&\stackrel{(C)}{=} \exp\left(\frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2 + t_2\mu_2 - \rho\frac{\sigma_2}{\sigma_1}\mu_1 t_2\right) \cdot \exp\left(\mu_1(t_1 + \rho\frac{\sigma_2}{\sigma_1}t_2) + \frac{1}{2}\sigma_1^2(t_1 + \rho\frac{\sigma_2}{\sigma_1}t_2)^2\right) \\
&= \exp\left(\frac{1}{2}\sigma_2^2 t_2^2 - \frac{1}{2}\rho^2\sigma_2^2 t_2^2 + \mu_2 t_2 - \mu_1\rho\frac{\sigma_2}{\sigma_1}t_2 + \mu_1 t_1 + \mu_1\rho\frac{\sigma_2}{\sigma_1}t_2 + \frac{1}{2}\sigma_1^2 t_1^2 + \rho\sigma_1\sigma_2 t_1 t_2 + \frac{1}{2}\rho^2\sigma_2^2 t_2^2\right) \\
&= \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{\sigma_1^2 t_1^2 + \sigma_2^2 t_2^2 + 2\rho\sigma_1\sigma_2 t_1 t_2}{2}\right)
\end{aligned}
$$

(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 $A \in I\!R^{k\times k}$ be a $k \times k$ matrix, and let $\underline{\mu} \in I\!R^k$ 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

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

(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

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

■

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

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. ■

5.2 Exponential Familty of Distributions

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 $\eta = (\eta_1, \ldots, \eta_k)'$. ■

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,

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

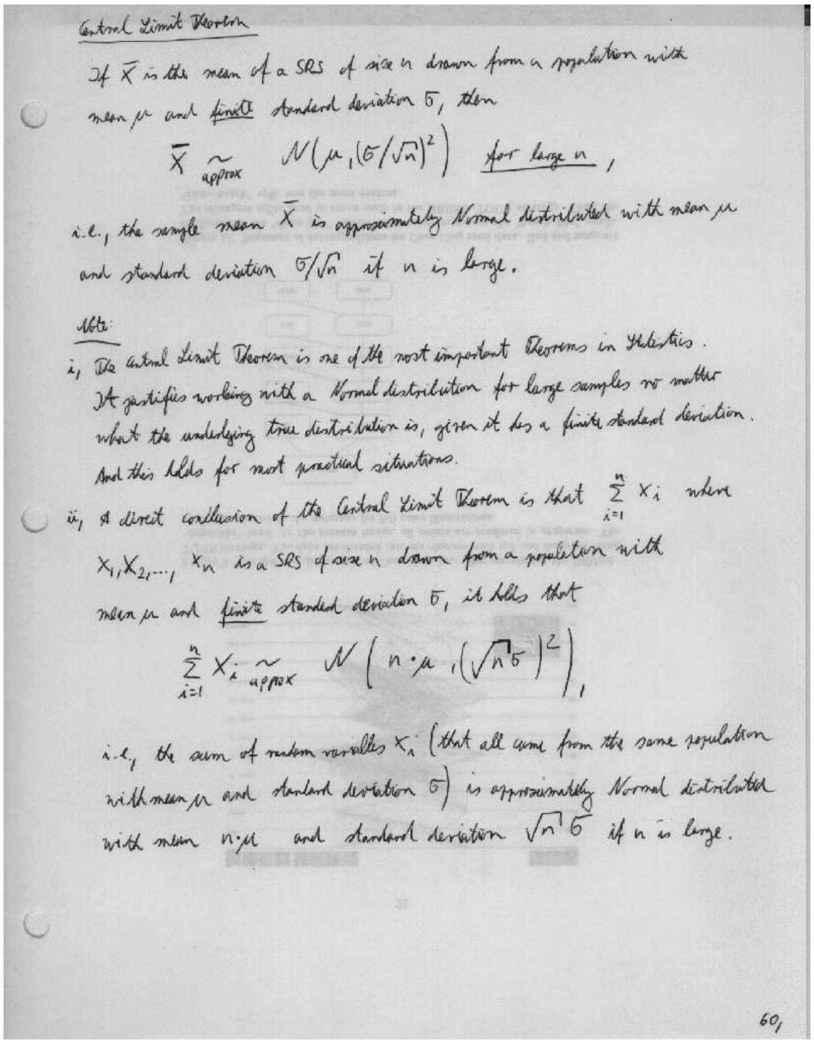
Thus, this is a 2–parameter exponential family. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 36: Monday 11/22/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

6 Limit Theorems

Motivation:

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



Central Limit Theorem

If $\overline{X}$ is the mean of a SRS of size n drawn from a population with mean μ and finite standard deviation σ, then

$$\overline{X} \underset{approx}{\sim} N(\mu, (\sigma/\sqrt{n})^2) \quad \text{for large } n,$$

i.e., the sample mean $\overline{X}$ is approximately Normal distributed with mean μ and standard deviation $\sigma/\sqrt{n}$ if n is large.

Note:

i, The Central Limit Theorem is one of the most important Theorems in Statistics. It justifies working with a Normal distribution for large samples no matter what the underlying true distribution is, given it has a finite standard deviation. And this holds for most practical situations.

ii, A direct conclusion of the Central Limit Theorem is that $\sum_{i=1}^{n} X_i$ where $X_1, X_2, \ldots, X_n$ is a SRS of size n drawn from a population with mean μ and finite standard deviation σ, it holds that

$$\sum_{i=1}^{n} X_i \underset{approx}{\sim} N\left(n \cdot \mu, (\sqrt{n}\,\sigma)^2\right),$$

i.e., the sum of random variables X_i (that all come from the same population with mean μ and standard deviation σ) is approximately Normal distributed with mean $n \cdot \mu$ and standard deviation $\sqrt{n}\,\sigma$ if n is large.

60,

What does this mean at a more theoretical level???

6.1 Modes of Convergence

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

Note:

(i) Commonly used statistics are:

Sample Mean: $\overline{X}_n = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$

Sample Variance: $S_n^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} (X_i - \overline{X}_n)^2$

Sample Median, Order Statistics, Min, Max, etc.

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

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

■

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

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

Example 6.1.3:
Let $X_n \sim N(0, \frac{1}{n})$. Then

$$F_n(x) = \int_{-\infty}^{x} \frac{\exp\left(-\frac{1}{2} n t^2\right)}{\sqrt{\frac{2\pi}{n}}} \, dt$$

$$
\begin{aligned}
&= \int_{-\infty}^{\sqrt{n}x} \frac{\exp(-\frac{1}{2}s^2)}{\sqrt{2\pi}}\, ds \\
&= \phi(\sqrt{n}x)
\end{aligned}
$$

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

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

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

Example 6.1.4:

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

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

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

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

Example 6.1.5:

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

It is

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

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

It holds that $F_n \xrightarrow{w} F_X$ but

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

Thus, convergence in distribution does not imply convergence of moments/means. ■

Note:

Convergence in distribution does not say that the X_i's are close to each other or to X. It only means that their cdf's are (eventually) close to some cdf F. The X_i's do not even have to be defined on the same probability space. ■

Example 6.1.6:

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

Theorem 6.1.7:

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

Theorem 6.1.8:

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

Theorem 6.1.9:

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

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

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

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

Proof:

Part (iii):

Suppose that $a > 0, a_n > 0$. Let $Y_n = a_n X_n + b_n$ and $Y = aX + b$. It is

$$F_Y(y) = P(Y < y) = P(aX + b < y) = P(X < \frac{y-b}{a}) = F_X(\frac{y-b}{a}).$$

Likewise,

$$F_{Y_n}(y) = F_{X_n}(\frac{y-b_n}{a_n}).$$

If y is a continuity point of F_Y, $\frac{y-b}{a}$ is a continuity point of F_X. Since $a_n \to a, b_n \to b$ and $F_{X_n}(x) \to F_X(x)$, it follows that $F_{Y_n}(y) \to F_Y(y)$ for every continuity point y of F_Y. Thus, $a_n X_n + b_n \xrightarrow{d} aX + b$. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 37: Wednesday 11/24/1999	
	Scribe: Jürgen Symanzik, Rich Madsen

Definition 6.1.10:

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of rv's defined on a probability space (Ω, L, P). We say that X_n **converges in probability** to a rv X ($X_n \xrightarrow{p} X$, P-$\lim\limits_{n\to\infty} X_n = X$) if

$$\lim_{n\to\infty} P(\mid X_n - X \mid > \epsilon) = 0 \quad \forall \epsilon > 0.$$

■

Note:

The following are equivalent:

$\lim\limits_{n\to\infty} P(\mid X_n - X \mid > \epsilon) = 0 \Leftrightarrow \lim\limits_{n\to\infty} P(\mid X_n - X \mid \leq \epsilon) = 1 \Leftrightarrow \lim\limits_{n\to\infty} P(\{\omega : \mid X_n(\omega) - X(\omega) \mid > \epsilon\}) = 0.$

If X is degenerate, i.e., $P(X = c) = 1$, we say that X_n is **consistent** for c. For example, let X_n such that $P(X_n = 0) = 1 - \frac{1}{n}$ and $P(X_n = 1) = \frac{1}{n}$. Then

$$P(\mid X_n \mid > \epsilon) = \begin{cases} \frac{1}{n}, & 0 < \epsilon < 1 \\ 0, & \epsilon \geq 1 \end{cases}$$

Therefore, $\lim\limits_{n\to\infty} P(\mid X_n \mid > \epsilon) = 0 \ \ \forall \epsilon > 0$. So $X_n \xrightarrow{p} 0$, i.e., X_n is consistent for 0. ■

Theorem 6.1.11:

(i) $X_n \xrightarrow{p} X \Longleftrightarrow X_n - X \xrightarrow{p} 0$.

(ii) $X_n \xrightarrow{p} X, X_n \xrightarrow{p} Y \Longrightarrow P(X = Y) = 1$.

(iii) $X_n \xrightarrow{p} X, X_m \xrightarrow{p} X \Longrightarrow X_n - X_m \xrightarrow{p} 0$ as $n, m \to \infty$.

(iv) $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \Longrightarrow X_n \pm Y_n \xrightarrow{p} X \pm Y$.

(v) $X_n \xrightarrow{p} X, k \in I\!R$ a constant $\Longrightarrow kX_n \xrightarrow{p} kX$.

(vi) $X_n \xrightarrow{p} k, k \in I\!R$ a constant $\Longrightarrow X_n^r \xrightarrow{p} k^r \ \ \forall r \in I\!N$.

(vii) $X_n \xrightarrow{p} a, Y_n \xrightarrow{p} b, \ a, b \in I\!R \Longrightarrow X_n Y_n \xrightarrow{p} ab$.

(viii) $X_n \xrightarrow{p} 1 \Longrightarrow X_n^{-1} \xrightarrow{p} 1$.

(ix) $X_n \xrightarrow{p} a, Y_n \xrightarrow{p} b,\ a \in I\!R, b \in I\!R - \{0\} \Longrightarrow \frac{X_n}{Y_n} \xrightarrow{p} \frac{a}{b}$.

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

(xi) $X_n \xrightarrow{p} X, Y_n \xrightarrow{p} Y \Longrightarrow X_n Y_n \xrightarrow{p} XY$.

Proof:
See Rohatgi, page 244–245 for partial proofs. ■

Theorem 6.1.12:
Let $X_n \xrightarrow{p} X$ and let g be a continuous function on $I\!R$. Then $g(X_n) \xrightarrow{p} g(X)$.

Proof:
Preconditions:

1.) X rv $\Longrightarrow \forall \epsilon > 0 \quad \exists k = k(\epsilon) : P(|X| > k) < \frac{\epsilon}{2}$

2.) g is continuous on $I\!R$

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

$\Longrightarrow \exists \delta = \delta(\epsilon, k) : |X| \leq k, |X_n - X| < \delta \Rightarrow |g(X_n) - g(X)| < \epsilon$

Let

$$\begin{aligned}
A &= \{|X| \leq k\} = \{\omega : |X(\omega)| \leq k\} \\
B &= \{|X_n - X| < \delta\} = \{\omega : |X_n(\omega) - X(\omega)| < \delta\} \\
C &= \{|g(X_n) - g(X)| < \epsilon\} = \{\omega : |g(X_n(\omega)) - g(X(\omega))| < \epsilon\}
\end{aligned}$$

If $\omega \in A \cap B$

$\overset{2.)}{\Longrightarrow} \omega \in C$

$\Longrightarrow A \cap B \subset C$

$\Longrightarrow C^C \subset (A \cap B)^C = A^C \cup B^C$

$\Longrightarrow P(C^C) \leq P(A^C \cup B^C) \leq P(A^C) + P(B^C)$

Now:

$$\begin{aligned}
P(|g(X_n) - g(X)| \geq \epsilon) &\leq \underbrace{P(|X| > k)}_{\leq \frac{\epsilon}{2} \text{ by 1.)}} + \underbrace{P(|X_n - X| \geq \delta)}_{\leq \frac{\epsilon}{2} \text{ for } n \geq n_0(\epsilon,\delta,k) \text{ since } X_n \xrightarrow{p} X} \\
&\leq \epsilon \quad \text{for } n \geq n_0(\epsilon, \delta, k)
\end{aligned}$$

■

Corollary 6.1.13:
Let $X_n \xrightarrow{p} c, c \in I\!R$ and let g be a continuous function on $I\!R$. Then $g(X_n) \xrightarrow{p} g(c)$. ■

Theorem 6.1.14:
$X_n \xrightarrow{p} X \Longrightarrow X_n \xrightarrow{d} X$.

Proof:
$X_n \xrightarrow{p} X \Leftrightarrow P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty \quad \forall \epsilon > 0$

It holds:

$$\begin{aligned} P(X \leq x - \epsilon) &= P(X \leq x - \epsilon, |X_n - X| \leq \epsilon) + P(X \leq x - \epsilon, |X_n - X| > \epsilon) \\ &\overset{(A)}{\leq} P(X_n \leq x) + P(|X_n - X| > \epsilon) \end{aligned}$$

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

Similarly, it holds:

$$\begin{aligned} P(X_n \leq x) &\leq P(X \leq x + \epsilon) + P(|X_n - X| > \epsilon) \\ &= P(X_n \leq x, \mid X_n - X \mid \leq \epsilon) + P(X_n \leq x, \mid X_n - X \mid > \epsilon) \end{aligned}$$

Combining the 2 inequalities from above gives:

$$P(X \leq x - \epsilon) - \underbrace{P(|X_n - X| > \epsilon)}_{\to 0 \text{ as } n \to \infty} \leq \underbrace{P(X_n \leq x)}_{= F_n(x)} \leq P(X \leq x + \epsilon) + \underbrace{P(|X_n - X| > \epsilon)}_{\to 0 \text{ as } n \to \infty}$$

Therefore,

$$P(X \leq x - \epsilon) \leq F_n(x) \leq P(X \leq x + \epsilon) \text{ as } n \to \infty.$$

Since the cdf's $F_n(\cdot)$ are not necessarily left continuous, we get the following result for $\epsilon \downarrow 0$:

$$P(X < x) \leq F_n(x) \leq P(X \leq x) = F_X(x)$$

Let x be a continuity point of F. Then it holds:

$$F(x) = P(X < x) \leq F_n(x) \leq F(x)$$

$\Longrightarrow F_n(x) \to F(x)$

$\Longrightarrow X_n \xrightarrow{d} X$ ■

Theorem 6.1.15:

Let $c \in I\!R$ be a constant. Then it holds:

$$X_n \xrightarrow{d} c \Longleftrightarrow X_n \xrightarrow{p} c.$$

■

Example 6.1.16:

In this example, we will see that

$$X_n \xrightarrow{d} X \not\Longrightarrow X_n \xrightarrow{p} X$$

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

X \ X_n	0	1	
0	0	$\frac{1}{2}$	$\frac{1}{2}$
1	$\frac{1}{2}$	0	$\frac{1}{2}$
	$\frac{1}{2}$	$\frac{1}{2}$	1

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

$$P(\mid X_n - X \mid > \epsilon) = P(\mid X_n - X \mid = 1) = 1 \quad \forall n,$$

so $\lim\limits_{n \to \infty} P(\mid X_n - X \mid > \epsilon) \neq 0$. Therefore, $X_n \not\xrightarrow{p} X$. ■

Theorem 6.1.17:

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

$$Y_n \xrightarrow{d} X, \mid X_n - Y_n \mid \xrightarrow{p} 0 \Longrightarrow X_n \xrightarrow{d} X.$$

Proof:

Similar to the proof of Theorem 6.1.14. See also Rohatgi, page 253, Theorem 14. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 38: Monday 11/29/1999	
	Scribe: Jürgen Symanzik, Bill Morphet

Theorem 6.1.18: **Slutsky's Theorem**

Let $(X_n)_{n=1}^{\infty}$ and $(Y_n)_{n=1}^{\infty}$ be sequences of rv's and X be a rv defined on a probability space (Ω, L, P). Let $c \in I\!R$ be a constant. Then it holds:

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

(ii) $X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c \Longrightarrow X_n Y_n \xrightarrow{d} cX.$
If $c = 0$, then also $X_n Y_n \xrightarrow{p} 0.$

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

Proof:

(i) $Y_n \xrightarrow{p} c \overset{Th.6.1.11(i)}{\Longleftrightarrow} Y_n - c \xrightarrow{p} 0$

$$\Longrightarrow Y_n - c = Y_n + (X_n - X_n) - c = (X_n + Y_n) - (X_n + c) \xrightarrow{p} 0 \quad (A)$$

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

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

$$X_n + Y_n \xrightarrow{d} X + c$$

(ii) Case $c = 0$:
$\forall \epsilon > 0 \;\; \forall k > 0$, it is

$$\begin{aligned} P(\mid X_n Y_n \mid > \epsilon) &= P(\mid X_n Y_n \mid > \epsilon, \; Y_n \leq \frac{\epsilon}{k}) + P(\mid X_n Y_n \mid > \epsilon, \; Y_n > \frac{\epsilon}{k}) \\ &\leq P(\mid X_n \frac{\epsilon}{k} \mid > \epsilon) + P(Y_n > \frac{\epsilon}{k}) \\ &\leq P(\mid X_n \mid > k) + P(\mid Y_n \mid > \frac{\epsilon}{k}) \end{aligned}$$

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

$$\overline{\lim_{n \to \infty}} P(\mid X_n Y_n \mid > \epsilon) \leq P(\mid X_n \mid > k) \to 0 \text{ as } k \to \infty.$$

Therefore,

$$X_n Y_n \xrightarrow{p} 0.$$

Case $c \neq 0$:

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

$\Longrightarrow X_nY_n \xrightarrow{p} cX_n$

$\overset{Th.6.1.14}{\Longrightarrow} X_nY_n \xrightarrow{d} cX_n$

Since $cX_n \xrightarrow{d} cX$ by Theorem 6.1.9 (ii), it follows from Theorem 6.1.17:

$$X_nY_n \xrightarrow{d} cX$$

(iii) Let $Z_n \xrightarrow{p} 1$ and let $Y_n = cZ_n$.

$\overset{c\neq 0}{\Longrightarrow} \frac{1}{Y_n} = \frac{1}{Z_n} \cdot \frac{1}{c}$

$\overset{Th.6.1.11(v,viii)}{\Longrightarrow} \frac{1}{Y_n} \xrightarrow{p} \frac{1}{c}$

With part (ii) above, it follows:

$X_n \xrightarrow{d} X$ and $\frac{1}{Y_n} \xrightarrow{p} \frac{1}{c}$

$\Longrightarrow \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$

■

Definition 6.1.19:

Let $(X_n)_{n=1}^{\infty}$ be a sequence of rv's such that $E(\mid X_n \mid^r) < \infty$ for some $r > 0$. We say that X_n **converges in the** r^{th} **mean** to a rv X ($X_n \xrightarrow{r} X$) if $E(\mid X \mid^r) < \infty$ and

$$\lim_{n\to\infty} E(\mid X_n - X \mid^r) = 0.$$

■

Example 6.1.20:

Let $(X_n)_{n=1}^{\infty}$ be a sequence of rv's defined by $P(X_n = 0) = 1 - \frac{1}{n}$ and $P(X_n = 1) = \frac{1}{n}$.

It is $E(\mid X_n \mid^r) = \frac{1}{n} \ \forall r > 0$. Therefore, $X_n \xrightarrow{r} 0 \ \forall r > 0$. ■

Note:

The special cases $r = 1$ and $r = 2$ are called *convergence in absolute mean* for $r = 1$ ($X_n \xrightarrow{1} X$) and *convergence in mean square* for $r = 2$ ($X_n \xrightarrow{ms} X$ or $X_n \xrightarrow{2} X$). ■

Theorem 6.1.21:

Assume that $X_n \xrightarrow{r} X$ for some $r > 0$. Then $X_n \xrightarrow{p} X$.

Proof:

Using Markov's Inequality (Corollary 3.5.2), it holds for any $\epsilon > 0$:

$$\frac{E(\mid X_n - X \mid^r)}{\epsilon^r} \geq P(\mid X_n - X \mid \geq \epsilon)$$

$$X_n \xrightarrow{r} X \Longrightarrow \lim_{n\to\infty} E(\mid X_n - X \mid^r) = 0$$

$$\Longrightarrow \lim_{n\to\infty} P(\mid X_n - X \mid \geq \epsilon) \leq \lim_{n\to\infty} \frac{E(\mid X_n - X \mid^r)}{\epsilon^r} = 0$$

$$\Longrightarrow X_n \xrightarrow{p} X$$

■

Example 6.1.22:
Let $(X_n)_{n=1}^{\infty}$ be a sequence of rv's defined by $P(X_n = 0) = 1 - \frac{1}{n^r}$ and $P(X_n = n) = \frac{1}{n^r}$ for some $r > 0$.

For any $\epsilon > 0$, $P(\mid X_n \mid > \epsilon) \to 0$ as $n \to \infty$; so $X_n \xrightarrow{p} 0$.

For $0 < s < r$, $E(\mid X_n \mid^s) = \frac{1}{n^{r-s}} \to 0$ as $n \to \infty$; so $X_n \xrightarrow{s} 0$. But $E(\mid X_n \mid^r) = 1 \not\to 0$ as $n \to \infty$; so $X_n \overset{r}{\not\longrightarrow} 0$. ■

Theorem 6.1.23:
If $X_n \xrightarrow{r} X$, then it holds:

(i) $\lim_{n\to\infty} E(\mid X_n \mid^r) = E(\mid X \mid^r)$; and

(ii) $X_n \xrightarrow{s} X$ for $0 < s < r$.

Proof:

(i) For $0 < r \leq 1$, it holds:

$$E(\mid X_n \mid^r) = E(\mid X_n - X + X \mid^r) \leq E(\mid X_n - X \mid^r + \mid X \mid^r)$$

$$\Longrightarrow E(\mid X_n \mid^r) - E(\mid X \mid^r) \leq E(\mid X_n - X \mid^r)$$

$$\Longrightarrow \lim_{n\to\infty} E(\mid X_n \mid^r) - \lim_{n\to\infty} E(\mid X \mid^r) \leq \lim_{n\to\infty} E(\mid X_n - X \mid^r) = 0$$

$$\Longrightarrow \lim_{n\to\infty} E(\mid X_n \mid^r) \leq E(\mid X \mid^r) \quad (A)$$

Similarly,

$$E(\mid X \mid^r) = E(\mid X - X_n + X_n \mid^r) \leq E(\mid X_n - X \mid^r + \mid X_n \mid^r)$$

$$\Longrightarrow E(\mid X \mid^r) - E(\mid X_n \mid^r) \leq E(\mid X_n - X \mid^r)$$

$$\Longrightarrow \lim_{n\to\infty} E(\mid X \mid^r) - \lim_{n\to\infty} E(\mid X_n \mid^r) \leq \lim_{n\to\infty} E(\mid X_n - X \mid^r) = 0$$

$$\Longrightarrow E(\mid X \mid^r) \leq \lim_{n\to\infty} E(\mid X_n \mid^r) \quad (B)$$

Combining (A) and (B) gives

$$\lim_{n\to\infty} E(\mid X_n \mid^r) = E(\mid X \mid^r)$$

For $r > 1$, it follows from Minkowski's Inequality (Theorem 4.8.3):

$$[E(| X - X_n + X_n |^r)]^{\frac{1}{r}} \leq [E(| X - X_n |^r)]^{\frac{1}{r}} + [E(| X_n |^r)]^{\frac{1}{r}}$$

$$\Longrightarrow [E(| X |^r)]^{\frac{1}{r}} - [E(| X_n |^r)]^{\frac{1}{r}} \leq [E(| X - X_n |^r)]^{\frac{1}{r}}$$

$$\Longrightarrow [E(| X |^r)]^{\frac{1}{r}} - \lim_{n\to\infty}[E(| X_n |^r)]^{\frac{1}{r}} \leq \lim_{n\to\infty}[E(| X_n - X |^r)]^{\frac{1}{r}} = 0 \quad \text{since } X_n \xrightarrow{r} X$$

$$\Longrightarrow [E(| X |^r)]^{\frac{1}{r}} \leq \lim_{n\to\infty}[E(| X_n |^r)]^{\frac{1}{r}} \quad (C)$$

Similarly,

$$[E(| X_n - X + X |^r)]^{\frac{1}{r}} \leq [E(| X_n - X |^r)]^{\frac{1}{r}} + [E(| X |^r)]^{\frac{1}{r}}$$

$$\Longrightarrow \lim_{n\to\infty}[E(| X_n |^r)]^{\frac{1}{r}} - \lim_{n\to\infty}[E(| X |^r)]^{\frac{1}{r}} \leq \lim_{n\to\infty}[E(| X_n - X |^r)]^{\frac{1}{r}} = 0 \quad \text{since } X_n \xrightarrow{r} X$$

$$\Longrightarrow \lim_{n\to\infty}[E(| X_n |^r)]^{\frac{1}{r}} \leq [E(| X |^r)]^{\frac{1}{r}} \quad (D)$$

Combining (C) and (D) gives

$$\lim_{n\to\infty}[E(| X_n |^r)]^{\frac{1}{r}} = [E(| X |^r)]^{\frac{1}{r}}$$

$$\Longrightarrow \lim_{n\to\infty} E(| X_n |^r) = E(| X |^r)$$

(ii) For $1 \leq s < r$, it follows from Lyapunov's Inequality (Theorem 3.5.4):

$$[E(| X_n - X |^s)]^{\frac{1}{s}} \leq [E(| X_n - X |^r)]^{\frac{1}{r}}$$

$$\Longrightarrow E(| X_n - X |^s) \leq [E(| X_n - X |^r)]^{\frac{s}{r}}$$

$$\Longrightarrow \lim_{n\to\infty} E(| X_n - X |^s) \leq \lim_{n\to\infty}[E(| X_n - X |^r)]^{\frac{s}{r}} = 0 \quad \text{since } X_n \xrightarrow{r} X$$

$$\Longrightarrow X_n \xrightarrow{s} X$$

An additional proof is required for $0 < s < r \leq 1$.

■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lectures 39 & 40: Wednesday 12/1/1999 & Friday 12/3/1999	
	Scribe: Jürgen Symanzik, Hanadi B. Eltahir

Definition 6.1.24:
Let $\{X_n\}_{n=1}^\infty$ be a sequence of rv's on (Ω, L, P). We say that X_n **converges almost surely** to a rv X ($X_n \xrightarrow{a.s.} X$) or X_n **converges with probability 1** to X ($X_n \xrightarrow{w.p.1} X$) or X_n converges strongly to X iff

$$P(\{\omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = 1.$$

Note:
An interesting characterization of convergence with probability 1 and convergence in probability can be found in Parzen (1960) "Modern Probability Theory and Its Applications" on page 416 (see Handout). ■

Example 6.1.25:
Let $\Omega = [0, 1]$ and P a uniform distribution on Ω. Let $X_n(\omega) = \omega + \omega^n$ and $X(\omega) = \omega$.

For $\omega \in [0, 1)$, $\omega^n \to 0$ as $n \to \infty$. So $X_n(\omega) \to X(\omega) \ \ \forall \omega \in [0, 1)$.

However, for $\omega = 1$, $X_n(1) = 2 \neq 1 = X(1) \ \ \forall n$, i.e., convergence fails at $\omega = 1$.

Anyway, since $P(\{\omega : X_n(\omega) \to X(\omega) \text{ as } n \to \infty\}) = P(\{\omega \in [0, 1)\}) = 1$, it is $X_n \xrightarrow{a.s.} X$. ■

Theorem 6.1.26:
$X_n \xrightarrow{a.s.} X \Longrightarrow X_n \xrightarrow{p} X$.

Proof:
Choose $\epsilon > 0$ and $\delta > 0$. Find $n_0 = n_0(\epsilon, \delta)$ such that

$$P\left(\bigcap_{n=n_0}^{\infty} \{| X_n - X | \leq \epsilon\}\right) \geq 1 - \delta.$$

Since $\bigcap\limits_{n=n_0}^{\infty} \{| X_n - X | \leq \epsilon\} \subseteq \{| X_n - X | \leq \epsilon\} \quad \forall n \geq n_0$, it is

$$P\left(\{| X_n - X | \leq \epsilon\}\right) \geq P\left(\bigcap_{n=n_0}^{\infty} \{| X_n - X | \leq \epsilon\}\right) \geq 1 - \delta \ \ \forall n \geq n_0.$$

Therefore, $P(\{| X_n - X | \leq \epsilon\}) \to 1$ as $n \to \infty$. Thus, $X_n \xrightarrow{p} X$. ■

Example 6.1.27:
$X_n \xrightarrow{p} X \not\Longrightarrow X_n \xrightarrow{a.s.} X$:

Let $\Omega = (0, 1]$ and P a uniform distribution on Ω.
Define A_n by

$$A_1 = (0, \tfrac{1}{2}], A_2 = (\tfrac{1}{2}, 1]$$
$$A_3 = (0, \tfrac{1}{4}], A_4 = (\tfrac{1}{4}, \tfrac{1}{2}], A_5 = (\tfrac{1}{2}, \tfrac{3}{4}], A_6 = (\tfrac{3}{4}, 1]$$
$$A_7 = (0, \tfrac{1}{8}], A_8 = (\tfrac{1}{8}, \tfrac{1}{4}], \ldots$$

Let $X_n(\omega) = I_{A_n}(\omega)$.

It is $P(\mid X_n - 0 \mid \geq \epsilon) \to 0 \;\; \forall \epsilon > 0$ since X_n is 0 except on A_n and $P(A_n) \downarrow 0$. Thus $X_n \xrightarrow{p} 0$.

But $P(\{\omega : X_n(\omega) \to 0\}) = 0$ (and not 1) because any ω keeps being in some A_n beyond any n_0, i.e., $X_n(\omega)$ looks like $0 \ldots 010 \ldots 010 \ldots 010 \ldots$, so $X_n \overset{a.s.}{\not\longrightarrow} 0$. ∎

Example 6.1.28:
$X_n \xrightarrow{r} X \not\Longrightarrow X_n \xrightarrow{a.s.} X$:

Let X_n be independent rv's such that $P(X_n = 0) = 1 - \frac{1}{n}$ and $P(X_n = 1) = \frac{1}{n}$.

It is $E(\mid X_n - 0 \mid^r) = E(\mid X_n \mid^r) = E(\mid X_n \mid) = \frac{1}{n} \to 0$ as $n \to \infty$, so $X_n \xrightarrow{r} 0 \;\; \forall r > 0$.

But

$$P(X_n = 0 \;\; \forall m \leq n \leq n_0) = \prod_{n=m}^{n_0} (1 - \frac{1}{n}) = (\frac{m-1}{m})(\frac{m}{m+1})(\frac{m+1}{m+2}) \ldots (\frac{n_0 - 2}{n_0 - 1})(\frac{n_0 - 1}{n_0}) = \frac{m-1}{n_0}$$

As $n_0 \to \infty$, it is $P(X_n = 0 \;\; \forall m \leq n \leq n_0) \to 0 \;\; \forall m$, so $X_n \overset{a.s.}{\not\longrightarrow} 0$. ∎

Example 6.1.29:
$X_n \xrightarrow{a.s.} X \not\Longrightarrow X_n \xrightarrow{r} X$:

Let $\Omega = [0, 1]$ and P a uniform distribution on Ω.

Let $A_n = [0, \frac{1}{\ln n}]$.

Let $X_n(\omega) = n I_{A_n}(\omega)$ and $X(\omega) = 0$.

It holds that $\forall \omega > 0 \;\; \exists n_0 : \frac{1}{\ln n_0} < \omega \Longrightarrow X_n(\omega) = 0 \;\; \forall n > n_0$ and $P(\omega = 0) = 0$. Thus, $X_n \xrightarrow{a.s.} 0$.

But $E(\mid X_n - 0 \mid^r) = \frac{n^r}{\ln n} \to \infty \;\; \forall r > 0$, so $X_n \overset{r}{\not\longrightarrow} X$. ∎

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 41: Monday 12/6/1999	
	Scribe: Jürgen Symanzik

6.2 Weak Laws of Large Numbers

Theorem 6.2.1: **WLLN: Version I**

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid rv's with mean $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2 < \infty$. Let $\overline{X}_n = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$. Then it holds

$$\lim_{n \to \infty} P(\mid \overline{X}_n - \mu \mid \geq \epsilon) = 0,$$

i.e., $\overline{X}_n \xrightarrow{p} \mu$.

Proof:

By Markov's Inequality, it holds for all $\epsilon > 0$:

$$P(\mid \overline{X}_n - \mu \mid \geq \epsilon) \leq \frac{E((\overline{X}_n - \mu)^2)}{\epsilon^2} = \frac{Var(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \longrightarrow 0 \text{ as } n \to \infty$$ ■

Note:

For iid rv's with finite variance, $\overline{X}_n$ is consistent for μ.

A more general way to derive a "WLLN" follows in the next Definition. ■

Definition 6.2.2:

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of rv's. Let $T_n = \sum\limits_{i=1}^{n} X_i$. We say that $\{X_i\}$ obeys the WLLN with respect to a sequence of **norming constants** $\{B_i\}_{i=1}^{\infty}$, $B_i > 0, B_i \uparrow \infty$, if there exists a sequence of **centering constants** $\{A_i\}_{i=1}^{\infty}$ such that

$$B_n^{-1}(T_n - A_n) \xrightarrow{p} 0.$$

■

Theorem 6.2.3:

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of pairwise uncorrelated rv's with $E(X_i) = \mu_i$ and $Var(X_i) = \sigma_i^2$, $i \in I\!N$. If $\sum\limits_{i=1}^{n} \sigma_i^2 \to \infty$ as $n \to \infty$, we can choose $A_n = \sum\limits_{i=1}^{n} \mu_i$ and $B_n = \sum\limits_{i=1}^{n} \sigma_i^2$ and get

$$\frac{\sum\limits_{i=1}^{n} (X_i - \mu_i)}{\sum\limits_{i=1}^{n} \sigma_i^2} \xrightarrow{p} 0.$$

Proof:
By Markov's Inequality, it holds for all $\epsilon > 0$:

$$P(\mid \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i \mid > \epsilon \sum_{i=1}^{n} \sigma_i^2) \leq \frac{E((\sum_{i=1}^{n}(X_i - \mu))^2)}{\epsilon^2 (\sum_{i=1}^{n} \sigma_i^2)^2} = \frac{1}{\epsilon^2 \sum_{i=1}^{n} \sigma_i^2} \longrightarrow 0 \text{ as } n \to \infty \qquad \blacksquare$$

Note:
To obtain Theorem 6.2.1, we choose $A_n = n\mu$ and $B_n = n\sigma^2$. ■

Theorem 6.2.4:
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of rv's. Let $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$. A necessary and sufficient condition for $\{X_i\}$ to obey the WLLN with respect to $B_n = n$ is that

$$E\left(\frac{\overline{X}_n^2}{1 + \overline{X}_n^2}\right) \to 0$$

as $n \to \infty$.

Proof:
Rohatgi, page 258, Theorem 2. ■

Example 6.2.5:
Let $(X_1, \ldots, X_n)$ be jointly Normal with $E(X_i) = 0$, $E(X_i^2) = 1$ for all i, and $Cov(X_i, X_j) = \rho$ if $\mid i - j \mid = 1$ and $Cov(X_i, X_j) = 0$ if $\mid i - j \mid > 1$. Then, $T_n \sim N(0, n + 2(n-1)\rho) = N(0, \sigma^2)$. It is

$$\begin{aligned}
E\left(\frac{\overline{X}_n^2}{1 + \overline{X}_n^2}\right) &= E\left(\frac{T_n^2}{n^2 + T_n^2}\right) \\
&= \frac{2}{\sqrt{2\pi}\sigma} \int_0^{\infty} \frac{x^2}{n^2 + x^2} e^{-\frac{x^2}{2\sigma^2}} dx \quad \mid \quad y = \frac{x}{\sigma}, \quad dy = \frac{dx}{\sigma} \\
&= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \frac{\sigma^2 y^2}{n^2 + \sigma^2 y^2} e^{-\frac{y^2}{2}} dy \\
&= \frac{2}{\sqrt{2\pi}} \int_0^{\infty} \frac{(n + 2(n-1)\rho) y^2}{n^2 + (n + 2(n-1)\rho) y^2} e^{-\frac{y^2}{2}} dy \\
&\leq \frac{n + 2(n-1)\rho}{n^2} \qquad \underbrace{\int_0^{\infty} \frac{2}{\sqrt{2\pi}} y^2 e^{-\frac{y^2}{2}} dy}_{=1, \text{ since Var of } N(0,1) \text{ distribution}} \\
&\to 0 \qquad \text{as } n \to \infty
\end{aligned}$$

$$\Longrightarrow \quad \overline{X}_n \xrightarrow{p} 0$$

■

Note:

We would like to have a WLLN that just depends on means but does not depend on the existence of finite variances. To approach this, we consider the following:

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of rv's. Let $T_n = \sum_{i=1}^{n} X_i$. We truncate each X_i at $c > 0$ and get

$$X_i^c = \begin{cases} X_i, & | X_i | \leq c \\ 0, & \text{otherwise} \end{cases}$$

Let $T_n^c = \sum_{i=1}^{n} X_i^c$ and $m_n = \sum_{i=1}^{n} E(X_i^c)$. ■

Lemma 6.2.6:

For T_n, T_n^c and m_n as defined in the Note above, it holds:

$$P(| T_n - m_n |> \epsilon) \leq P(| T_n^c - m_n |> \epsilon) + \sum_{i=1}^{n} P(| X_i |> c) \quad \forall \epsilon > 0$$

Proof:

It holds for all $\epsilon > 0$:

$$\begin{aligned} P(| T_n - m_n |> \epsilon) &= P(| T_n - m_n |> \epsilon \text{ and } | X_i |\leq c \ \ \forall i \in \{1, \ldots, n\}) + \\ & \qquad P(| T_n - m_n |> \epsilon \text{ and } | X_i |> c \text{ for at least one } i \in \{1, \ldots, n\} \\ &\overset{(*)}{\leq} P(| T_n^c - m_n |> \epsilon) + P(| X_i |> c \text{ for at least one } i \in \{1, \ldots, n\}) \\ &\leq P(| T_n^c - m_n |> \epsilon) + \sum_{i=1}^{n} P(| X_i |> c) \end{aligned}$$

$(*)$ holds since $T_n^c = T_n$ when $| X_i |\leq c \ \ \forall i \in \{1, \ldots, n\}$. ■

Note:

If the X_i's are identically distributed, then

$$P(| T_n - m_n |> \epsilon) \leq P(| T_n^c - m_n |> \epsilon) + nP(| X_1 |> c) \quad \forall \epsilon > 0.$$

If the X_i's are iid, then

$$P(| T_n - m_n |> \epsilon) \leq \frac{nE((X_1^c)^2)}{\epsilon^2} + nP(| X_1 |> c) \quad \forall \epsilon > 0 \quad (*).$$

Note that $P(| X_i |> c) = P(| X_1 |> c) \ \ \forall i \in I\!N$ if the X_i's are identically distributed and that $E((X_i^c)^2) = E((X_1^c)^2) \ \ \forall i \in I\!N$ if the X_i's are iid. ■

Theorem 6.2.7: **Khintchine's WLLN**
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid rv's with finite mean $E(X_i) = \mu$. Then it holds:

$$\overline{X}_n = \frac{1}{n} T_n \xrightarrow{p} \mu$$

Proof:
If we take $c = n$ and replace ϵ by $n\epsilon$ in $(*)$ in the Note above, we get

$$P(\mid T_n - m_n \mid > n\epsilon) \leq \frac{E((X_1^n)^2)}{n\epsilon^2} + nP(\mid X_1 \mid > n).$$

Since $E(\mid X_1 \mid) < \infty$, it is $nP(\mid X_1 \mid > n) \to 0$ as $n \to \infty$ by Theorem 3.1.9. From Corollary 3.1.12 we know that $E(\mid X \mid^{\alpha}) = \alpha \int_0^{\infty} x^{\alpha-1} P(\mid X \mid > x) dx$. Therefore,

$$\begin{aligned} E((X_1^n)^2) &= 2 \int_0^n x P(\mid X_1^n \mid > x) dx \\ &= 2 \int_0^A x P(\mid X_1^n \mid > x) dx + 2 \int_A^n x P(\mid X_1^n \mid > x) dx \\ &\overset{(+)}{\leq} K + \delta \int_A^n dx \\ &\leq K + n\delta \end{aligned}$$

In $(+)$, A is chosen sufficiently large such that $xP(\mid X_1^n \mid > x) < \frac{\delta}{2} \ \forall x \geq A$ for an arbitrary constant $\delta > 0$ and $K > 0$ a constant.

Therefore,

$$\frac{E((X_1^n)^2)}{n\epsilon^2} \leq \frac{k}{n\epsilon^2} + \frac{\delta}{\epsilon^2}$$

Since δ is arbitrary, we can make the right hand side of this last inequality arbitrarily small for sufficiently large n.

Since $E(X_i) = \mu \ \ \forall i$, it is $\frac{m_n}{n} \to \mu$ as $n \to \infty$. ■

Note:
Theorem 6.2.7 meets the previously stated goal of not having a finite variance requirement. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 42: Wednesday 12/8/1999	
	Scribe: Jürgen Symanzik

6.3 Strong Laws of Large Numbers

Definition 6.3.1:
Let $\{X_i\}_{i=1}^\infty$ be a sequence of rv's. Let $T_n = \sum\limits_{i=1}^n X_i$. We say that $\{X_i\}$ obeys the SLLN with respect to a sequence of **norming constants** $\{B_i\}_{i=1}^\infty$, $B_i > 0, B_i \uparrow \infty$, if there exists a sequence of **centering constants** $\{A_i\}_{i=1}^\infty$ such that

$$B_n^{-1}(T_n - A_n) \xrightarrow{a.s.} 0.$$

■

Note:
Unless otherwise specified, we will only use the case that $B_n = n$ in this section. ■

Theorem 6.3.2:
$X_n \xrightarrow{a.s.} X \iff \lim\limits_{n\to\infty} P(\sup\limits_{m\geq n} \mid X_m - X \mid > \epsilon) = 0 \;\; \forall \epsilon > 0.$

Proof: (see also Rohatgi, page 249, Theorem 11)
WLOG, we can assume that $X = 0$ since $X_n \xrightarrow{a.s.} X$ implies $X_n - X \xrightarrow{a.s.} 0$. Thus, we have to prove:

$$X_n \xrightarrow{a.s.} 0 \iff \lim_{n\to\infty} P(\sup_{m\geq n} \mid X_m \mid > \epsilon) = 0 \;\; \forall \epsilon > 0$$

"$\Longrightarrow$":
Choose $\epsilon > 0$ and define

$$\begin{aligned} A_n(\epsilon) &= \{\sup_{m\geq n} \mid X_m \mid > \epsilon\} \\ C &= \{\lim_{n\to\infty} X_n = 0\} \end{aligned}$$

We know that $P(C) = 1$ and therefore $P(C^c) = 0$.

Let $B_n(\epsilon) = C \cap A_n(\epsilon)$. Note that $B_{n+1}(\epsilon) \subseteq B_n(\epsilon)$ and for the limit set $\bigcap\limits_{n=1}^\infty B_n(\epsilon) = \emptyset$. It follows that

$$\lim_{n\to\infty} P(B_n(\epsilon)) = P(\bigcap_{n=1}^\infty B_n(\epsilon)) = 0.$$

We also have

$$\begin{aligned}
P(B_n(\epsilon)) &= P(A_n \cap C) \\
&= 1 - P(C^c \cup A_n^c) \\
&= 1 - \underbrace{P(C^c)}_{=0} - P(A_n^c) + \underbrace{P(C^c \cap A_n^C)}_{=0} \\
&= P(A_n)
\end{aligned}$$

$$\Longrightarrow \lim_{n\to\infty} P(A_n(\epsilon)) = 0$$

"$\Longleftarrow$":
Assume that $\lim\limits_{n\to\infty} P(A_n(\epsilon)) = 0 \ \ \forall \epsilon > 0$ and define $D(\epsilon) = \{\overline{\lim\limits_{n\to\infty}} \mid X_n \mid > \epsilon\}$. Since $D(\epsilon) \subseteq A_n(\epsilon) \ \ \forall n \in I\!N$, it follows that $P(D(\epsilon)) = 0 \ \ \forall \epsilon > 0$. Also,

$$C^c = \{\lim_{n\to\infty} X_n \neq 0\} \subseteq \bigcup_{k=1}^{\infty} \{\overline{\lim_{n\to\infty}} \mid X_n \mid > \frac{1}{k}\}.$$

$$\Longrightarrow 1 - P(C) \leq \sum_{k=1}^{\infty} P(D(\frac{1}{k})) = 0$$

$$\Longrightarrow X_n \xrightarrow{a.s.} 0$$

■

<u>Note:</u>

(i) $X_n \xrightarrow{a.s.} 0$ implies that $\forall \epsilon > 0 \ \forall \delta > 0 \ \exists n_0 \in I\!N \ : \ P(\sup\limits_{n \geq n_0} \mid X_n \mid > \epsilon) < \delta$.

(ii) Recall that for a given sequence of events $\{A_n\}_{n=1}^{\infty}$,

$$A = \overline{\lim_{n\to\infty}} A_n = \lim_{n\to\infty} \bigcup_{k=n}^{\infty} A_k = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

is the event that infinitely many of the A_n occur. We write $P(A) = P(A_n \ i.o.)$ where $i.o.$ stands for "infinitely often".

(iii) Using the terminology defined in (ii) above, we can rewrite Theorem 6.3.2 as

$$X_n \xrightarrow{a.s.} 0 \iff P(\mid X_n \mid > \epsilon \ i.o.) = 0 \ \ \forall \epsilon > 0.$$

■

Theorem 6.3.3: **Borel–Cantelli Lemma**

(i) 1^{st} **BC–Lemma**:

Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of events such that $\sum\limits_{n=1}^{\infty} P(A_n) < \infty$. Then $P(A) = 0$.

(ii) 2^{nd} **BC–Lemma**:

Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of independent events such that $\sum\limits_{n=1}^{\infty} P(A_n) = \infty$. Then $P(A) = 1$.

Proof:

(i):

$$\begin{aligned} P(A) &= P(\lim_{n\to\infty} \bigcup_{k=n}^{\infty} A_k) \\ &= \lim_{n\to\infty} P(\bigcup_{k=n}^{\infty} A_k) \\ &\leq \lim_{n\to\infty} \sum_{k=n}^{\infty} P(A_k) \\ &= \lim_{n\to\infty} \left(\sum_{k=1}^{\infty} P(A_k) - \sum_{k=1}^{n-1} P(A_k) \right) \\ &= 0 \end{aligned}$$

(ii): We have $A^c = \bigcup\limits_{n=1}^{\infty} \bigcap\limits_{k=n}^{\infty} A_k^c$. Therefore,

$$P(A^c) = P(\lim_{n\to\infty} \bigcap_{k=n}^{\infty} A_k^c) = \lim_{n\to\infty} P(\bigcap_{k=n}^{\infty} A_k^c).$$

If we choose $n_0 > n$, it holds that

$$\bigcap_{k=n}^{\infty} A_k^c \subseteq \bigcap_{k=n}^{n_0} A_k^c.$$

Therefore,

$$\begin{aligned} P(\bigcap_{k=n}^{\infty} A_k^c) &\leq \lim_{n_0\to\infty} P(\bigcap_{k=n}^{n_0} A_k^c) \\ &= \lim_{n_0\to\infty} \prod_{k=n}^{n_0} (1 - P(A_k)) \\ &\overset{indep.}{\leq} \lim_{n_0\to\infty} \exp\left(- \sum_{k=n}^{n_0} P(A_k) \right) \\ &= 0 \end{aligned}$$

$\Longrightarrow P(A) = 1$ ■

Example 6.3.4:
Independence is necessary for 2^{nd} BC–Lemma:

Let $\Omega = (0,1)$ and P a uniform distribution on Ω.

Let $A_n = I_{(0,\frac{1}{n})}(\omega)$. Therefore,

$$\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

But for any $\omega \in \Omega$, A_n occurs only for $1, 2, \ldots, \lfloor \frac{1}{\omega} \rfloor$, where $\lfloor \frac{1}{\omega} \rfloor$ denotes the largest integer ("floor") that is $\leq \frac{1}{\omega}$. Therefore, $P(A) = P(A_n \ i.o.) = 0$. ■

Lemma 6.3.5: **Kolmogorov's Inequality**
Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent rv's with common mean 0 and variances σ_i^2. Let $T_n = \sum_{i=1}^{n} X_i$. Then it holds:

$$P(\max_{1 \leq k \leq n} \mid T_k \mid \geq \epsilon) \leq \frac{\sum_{i=1}^{n} \sigma_i^2}{\epsilon^2} \quad \forall \epsilon > 0$$

Proof:
See Rohatgi, page 268, Lemma 2. ■

Lemma 6.3.6: **Kronecker's Lemma**
If $\sum_{i=1}^{\infty} X_i$ converges to $s < \infty$ and $B_n \uparrow \infty$, then it holds:

$$\frac{1}{B_n} \sum_{i=1}^{n} B_k X_k \rightarrow 0$$

Proof:
See Rohatgi, page 269, Lemma 3. ■

Theorem 6.3.7: **Cauchy Criterion**
$X_n \xrightarrow{a.s.} X \iff \lim_{n \to \infty} P(\sup_m \mid X_{n+m} - X_n \mid \leq \epsilon) = 1 \ \ \forall \epsilon > 0.$

Proof:
See Rohatgi, page 270, Theorem 5. ■

Theorem 6.3.8:

If $\sum\limits_{n=1}^{\infty} Var(X_n) < \infty$, then $\sum\limits_{n=1}^{\infty} (X_n - E(X_n))$ converges almost surely.

Proof:

See Rohatgi, page 272, Theorem 6. ■

Stat 6710 Mathematical Statistics I	Fall Semester 1999
Lecture 43: Friday 12/10/1999	
	Scribe: Jürgen Symanzik

Corollary 6.3.9:
Let $\{X_i\}_{i=1}^\infty$ be a sequence of independent rv's. Let $\{B_i\}_{i=1}^\infty$, $B_i > 0, B_i \uparrow \infty$, a sequence of norming constants. Let $T_n = \sum\limits_{i=1}^n X_i$. If $\sum\limits_{i=1}^\infty \frac{Var(X_i)}{B_i^2} < \infty$ then it holds:

$$\frac{T_n - E(T_n)}{B_n} \xrightarrow{a.s.} 0$$

Proof:
This Corollary follows directly from Theorem 6.3.8 and Lemma 6.3.6. ■

Lemma 6.3.10: **Equivalence Lemma**
Let $\{X_i\}_{i=1}^\infty$ and $\{X_i'\}_{i=1}^\infty$ be sequences of rv's. Let $T_n = \sum\limits_{i=1}^n X_i$ and $T_n' = \sum\limits_{i=1}^n X_i'$.

If the series $\sum\limits_{i=1}^\infty P(X_i \neq X_i') < \infty$, then the series $\{X_i\}$ and $\{X_i'\}$ are **tail–equivalent** and T_n and T_n' are **convergence–equivalent**, i.e., for $B_n \uparrow \infty$ the sequences $\frac{1}{B_n} T_n$ and $\frac{1}{B_n} T_n'$ converge on the same event and to the same limit, except for a null set.

Proof:
See Rohatgi, page 266, Lemma 1. ■

Lemma 6.3.11:
Let X be a rv with $E(\mid X \mid) < \infty$. Then it holds:

$$\sum_{n=1}^\infty P(\mid X \mid \geq n) \leq E(\mid X \mid) \leq 1 + \sum_{n=1}^\infty P(\mid X \mid \geq n)$$

Proof:
Continuous case only:
Let X have a pdf f. Then it holds:

$$E(\mid X \mid) = \int_{-\infty}^\infty \mid x \mid f(x) dx = \sum_{k=0}^\infty \int_{k \leq \mid x \mid \leq k+1} \mid x \mid f(x) dx$$

$$\Longrightarrow \sum_{k=0}^\infty k P(k \leq \mid X \mid \leq k+1) \leq E(\mid X \mid) \leq \sum_{k=0}^\infty (k+1) P(k \leq \mid X \mid \leq k+1)$$

It is

$$\begin{aligned}
\sum_{k=0}^{\infty} kP(k \leq\mid X \mid\leq k+1) &= \sum_{k=0}^{\infty}\sum_{n=1}^{k} P(k \leq\mid X \mid\leq k+1) \\
&= \sum_{n=1}^{\infty}\sum_{k=n}^{\infty} P(k \leq\mid X \mid\leq k+1) \\
&= \sum_{n=1}^{\infty} P(\mid X \mid\geq n)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\sum_{k=0}^{\infty}(k+1)P(k \leq\mid X \mid\leq k+1) &= \sum_{n=1}^{\infty} P(\mid X \mid\geq n) + \sum_{k=0}^{\infty} P(k \leq\mid X \mid\leq k+1) \\
&= \sum_{n=1}^{\infty} P(\mid X \mid\geq n) + 1
\end{aligned}$$

■

<u>Theorem 6.3.12:</u> **Kolmogorov's SLLN**

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of iid rv's. Let $T_n = \sum\limits_{i=1}^{n} X_i$. Then it holds:

$$\frac{T_n}{n} = \overline{X}_n \xrightarrow{a.s.} \mu < \infty \iff E(\mid X \mid) < \infty \text{ (and then } \mu = E(X))$$

<u>Proof:</u>

"$\Longrightarrow$":

Suppose that $\overline{X}_n \xrightarrow{a.s.} \mu < \infty$. It is

$$T_n = \sum_{i=1}^{n} X_i = \sum_{i=1}^{n-1} X_i + X_n = T_{n-1} + X_n.$$

$$\Longrightarrow \tfrac{X_n}{n} = \underbrace{\frac{T_n}{n}}_{\xrightarrow{a.s.}\mu} - \underbrace{\frac{n-1}{n}}_{\to 1}\ \underbrace{\frac{T_{n-1}}{n-1}}_{\xrightarrow{a.s.}\mu} \xrightarrow{a.s.} 0$$

By 1^{st} Borel–Cantelli Lemma, we must have $\sum\limits_{n=1}^{\infty} P(\mid X_n \mid\geq n) < \infty$.

$$\overset{Lemma\ 6.3.11}{\Longrightarrow} E(\mid X \mid) < \infty$$

$$\overset{Th.\ 6.2.7\ (WLLN)}{\Longrightarrow} \overline{X}_n \xrightarrow{p} E(X)$$

Since $\overline{X}_n \xrightarrow{a.s.} \mu$, it holds that $\overline{X}_n \xrightarrow{p} \mu$. Therefore, it must hold that $\mu = E(X)$.

"$\Longleftarrow$":

Let $E(\mid X \mid) < \infty$. Define truncated rv's:

$$\begin{aligned}
X'_k &= \begin{cases} X_k, & \text{if } \mid X_k \mid \leq k \\ 0, & \text{otherwise} \end{cases} \\
T'_n &= \sum_{k=1}^{n} X'_k \\
\overline{X}'_n &= \frac{T'_n}{n}
\end{aligned}$$

Then it holds:

$$\begin{aligned}
\sum_{k=1}^{\infty} P(X_k \neq X'_k) &= \sum_{k=1}^{\infty} P(\mid X_k \mid \geq k) \\
&\overset{iid}{=} \sum_{k=1}^{\infty} P(\mid X \mid \geq k) \\
&\overset{Lemma\ 6.3.11}{<} E(\mid X \mid) \\
&< \infty
\end{aligned}$$

By Lemma 6.3.10, it follows that T_n and T'_n are convergence–equivalent. Thus, it is sufficient to prove that $\overline{X}'_n \xrightarrow{a.s.} E(X)$.

We now establish the conditions needed in Corollary 6.3.9. It is

$$\begin{aligned}
Var(X'_n) &\leq E((X'_n)^2) \\
&= \int_{-n}^{n} x^2 f_X(x) dx \\
&= \sum_{k=0}^{n-1} \int_{k \leq |x| < k+1} x^2 f_X(x) dx \\
&\leq \sum_{k=0}^{n-1} (k+1)^2 P(k \leq \mid X \mid < k+1) \\
\Longrightarrow \sum_{n=1}^{\infty} \frac{1}{n^2} Var(X'_n) &\leq \sum_{n=1}^{\infty} \sum_{k=0}^{n} \frac{(k+1)^2}{n^2} P(k \leq \mid X \mid < k+1) \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{(k+1)^2}{n^2} P(k \leq \mid X \mid < k+1) + \sum_{n=1}^{\infty} \frac{1}{n^2} P(0 \leq \mid X \mid < 1) \\
&\overset{(*)}{=} \sum_{k=1}^{\infty} (k+1)^2 P(k \leq \mid X \mid < k+1) \left(\sum_{n=k}^{\infty} \frac{1}{n^2} \right) + 2P(0 \leq \mid X \mid < 1) \quad (A)
\end{aligned}$$

$(*)$ holds since $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.65 < 2$ and the first two sums can be rearranged as follows:

$$\begin{array}{c|c} n & k \\ \hline 1 & 1 \\ 2 & 1,2 \\ 3 & 1,2,3 \\ \vdots & \vdots \end{array} \quad \Longrightarrow \quad \begin{array}{c|c} k & n \\ \hline 1 & 1,2,3,\ldots \\ 2 & 2,3,\ldots \\ 3 & 3,\ldots \\ \vdots & \vdots \end{array}$$

It is

$$\begin{aligned} \sum_{n=k}^{\infty} \frac{1}{n^2} &= \frac{1}{k^2} + \frac{1}{(k+1)^2} + \frac{1}{(k+2)^2} + \ldots \\ &\leq \frac{1}{k^2} + \frac{1}{k(k+1)} + \frac{1}{(k+1)(k+2)} + \ldots \\ &= \frac{1}{k^2} + \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)} \end{aligned}$$

From Bronstein, page 30, # 7, we know that

$$\begin{aligned} 1 &= \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \ldots + \frac{1}{n(n+1)} + \ldots \\ &= \frac{1}{1\cdot 2} + \frac{1}{2\cdot 3} + \frac{1}{3\cdot 4} + \ldots + \frac{1}{(k-1)\cdot k} + \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)} \end{aligned}$$

$$\begin{aligned} \Longrightarrow \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)} &= 1 - \frac{1}{1\cdot 2} - \frac{1}{2\cdot 3} - \frac{1}{3\cdot 4} - \ldots - \frac{1}{(k-1)\cdot k} \\ &= \frac{1}{2} - \frac{1}{2\cdot 3} - \frac{1}{3\cdot 4} - \ldots - \frac{1}{(k-1)\cdot k} \\ &= \frac{1}{3} - \frac{1}{3\cdot 4} - \ldots - \frac{1}{(k-1)\cdot k} \\ &= \frac{1}{4} - \ldots - \frac{1}{(k-1)\cdot k} \\ &= \ldots \\ &= \frac{1}{k} \end{aligned}$$

$$\begin{aligned} \Longrightarrow \sum_{n=k}^{\infty} \frac{1}{n^2} &\leq \frac{1}{k^2} + \sum_{n=k+1}^{\infty} \frac{1}{n(n-1)} \\ &= \frac{1}{k^2} + \frac{1}{k} \end{aligned}$$

$$\leq \frac{2}{k}$$

Using this result in (A), we get

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{1}{n^2} Var(X_n') &\leq 2 \sum_{k=1}^{\infty} \frac{(k+1)^2}{k} P(k \leq \mid X \mid < k+1) + 2P(0 \leq \mid X \mid < 1) \\
&= 2 \sum_{k=0}^{\infty} k P(k \leq \mid X \mid < k+1) + 4 \sum_{k=1}^{\infty} P(k \leq \mid X \mid < k+1) \\
&\quad + 2 \sum_{k=1}^{\infty} \frac{1}{k} P(k \leq \mid X \mid < k+1) + 2P(0 \leq \mid X \mid < 1) \\
&\overset{(B)}{\leq} 2E(\mid X \mid) + 4 + 2 + 2 \\
&< \infty
\end{aligned}$$

To establish (B), we use an inequality from the Proof of Lemma 6.3.11.

Thus, the conditions needed in Corollary 6.3.9 are met. It follows that

$$\frac{1}{n} T_n' - \frac{1}{n} E(T_n') \xrightarrow{a.s.} 0 \quad (C)$$

Since $E(X_n') \rightarrow E(X)$ as $n \rightarrow \infty$, it follows by Kronecker's Lemma (6.3.6) that $\frac{1}{n} E(T_n') \rightarrow E(X)$. Thus, when we replace $\frac{1}{n} E(T_n')$ by $E(X)$ in (C), we get:

$$\frac{1}{n} T_n' \xrightarrow{a.s.} E(X) \quad \overset{Lemma\ 6.3.10}{\Longrightarrow} \quad \frac{1}{n} T_n \xrightarrow{a.s.} E(X)$$

■

Merry Xmas and a Happy New Millennium!